Automatic Thumbnail Cropping and its Effectiveness

Bongwon Suh*, Haibin Ling, Benjamin B. Bederson*, David W. Jacobs
Department of Computer Science
*Human-Computer Interaction Laboratory
University of Maryland
College Park, MD 20742 USA
+1 301-405-2764
{sbw, hbling, bederson, djacobs}@cs.umd.edu

ABSTRACT
Thumbnail images provide users of image retrieval and browsing systems with a method for quickly scanning large numbers of images. Recognizing the objects in an image is important in many retrieval tasks, but thumbnails generated by shrinking the original image often render objects illegible. We study the ability of computer vision systems to detect key components of images so that automated cropping, prior to shrinking, can render objects more recognizable. We evaluate automatic cropping techniques 1) based on a general method that detects salient portions of images, and 2) based on automatic face detection. Our user study shows that these methods result in small thumbnails that are substantially more recognizable and easier to find in the context of visual search.

Keywords
Saliency map, thumbnail, image cropping, face detection, usability study, visual search, zoomable user interfaces

INTRODUCTION
Thumbnail images are now widely used for visualizing large numbers of images given limited screen real estate. The QBIC system developed by Flickner *et al.* [10] is a notable image database example. A zoomable image browser, PhotoMesa [3], lays out thumbnails in a zoomable space and lets users move through the space of images with a simple set of navigation functions. PhotoFinder applied thumbnails as a visualization method for personal photo collections [14]. Popular commercial products such as Adobe Photoshop Album [2] and ACDSee [1] also use thumbnails to represent image files in their interfaces.

Current systems generate thumbnails by shrinking the original image. This method is simple. However, thumbnails generated this way can be difficult to recognize, especially when the thumbnails are very small. This phenomenon is not unexpected, since shrinking an image causes detailed information to be lost. An intuitive solution is to keep the more informative part of the image and cut less informative regions before shrinking. Some commercial products allow users to manually crop and shrink images [20]. Burton *et al.* [4] proposed and compared several image simplification methods to enhance the full-size images before subsampling. They chose edge-detecting smoothing, lossy image compression, and self-organizing feature map as three different techniques in their work.

In quite a different context, DeCarlo and Santella [8] tracked a user's eye movements to determine interesting portions of images, and generated non-photorealistic, painterly images that enhanced the most salient parts of the image. Chen *et al*. [5] use a visual attention model as a cue to conduct image adaptation for small displays.

In this paper, we study the effectiveness of saliency based cropping methods for preserving the recognizability of important objects in thumbnails. Our first method is a general cropping method based on the saliency map of Itti and Koch that models human visual attention [12][13]. A saliency map of a given image describes the importance of each position in the image. In our method, we use the saliency map directly as an indication of how much information each position in images contains. The merit of this method is that the saliency map is built up from low-level features only, so it can be applied to general images. We then select the most informative portion of the image.

Although this saliency based method is useful, it does not consider semantic information in images. We show that semantic information can be used to further improve thumbnail cropping, using automatic face detection. We choose this domain because a great many pictures of interest show human faces, and also because face detection methods have begun to achieve high accuracy and efficiency [22].

In this paper we describe saliency based cropping and face detection based cropping after first discussing related work from the field of visual attention. We then explain the

UIST '03, Vancouver, BC, Canada

design of a user study that evaluates the thumbnail methods. This paper concludes with a discussion of our findings and future work.

RELATED WORK

Visual attention is the ability of biological visual systems to detect interesting parts of the visual input [12][13][16][17] [21]. The saliency map of an image describes the degree of saliency of each position in the image. The saliency map is a matrix corresponding to the input image that describes the degree of saliency of each position in the input image.

Itti and Koch [12][13] provided an approach to compute a saliency map for images. Their method first uses pyramid technology to compute three feature maps for three low level features: color, intensity, and orientation. For each feature, saliency is detected when a portion of an image differs in that feature from neighboring regions. Then these feature maps are combined together to form a single saliency map. After this, in a series of iterations, salient pixels suppress the saliency of their neighbors, to concentrate saliency in a few key points.

Chen *et al*. [5] proposed using semantic models together with the saliency model of Itti and Koch to identify important portions of an image, prior to cropping. Their method is based on an attention model that uses attention objects as the basic elements. The overall attention value of each attention object is calculated by combining attention values from different models. For semantic attention models they use a face detection technique [15] and a text detection technique [6] to compute two different attention values. The method provides a way to combine semantic information with low-level features. However, when combining the different values, their method uses heuristic weights that are different for five different predefined image types. Images need to be manually categorized into these five categories prior to applying their method. Furthermore, it heavily relies on semantic extraction techniques. When the corresponding semantic technique is not available or when the technique fails to provide a good result (e.g. no face found in the image), it is hard to expect a good result from the method. On the other hand, our algorithm is totally automatic and works well without manual intervention or any assumptions about the image types.

THUMBNAIL CROPPING

Problem Definition

We define the thumbnail cropping problem as follows: Given an image I, the goal of thumbnail cropping is to find a rectangle R_C, containing a subset of the image I_C so that the main objects in the image are visible in the subimage. We then shrink I_C to a thumbnail. In the rest of this paper, we use the word "cropping" to indicate thumbnail cropping.

In the next subsection, we propose a general cropping method, which is based on the saliency map and can be applied to general images. Next, a face detection based cropping method is introduced for images with faces.

A General Cropping Method Based on the Saliency Map

In this method, we use the saliency value to evaluate the degree of informativeness of different positions in the image I. The cropping rectangle R_C should satisfy two conditions: having a small size and containing most of the salient parts of the image. These two conditions generally conflict with each other. Our goal is to find the optimal rectangle to balance these two conditions.

An example saliency map is given in Figure 1:

Figure 1: left: original image, right: saliency map of the image shown left

Find Cropping Rectangle with Fixed Threshold using Brute Force Algorithm

We use Itti and Koch's saliency algorithm because their method is based on low-level features and hence independent of semantic information in images. We choose Itti and Koch's model also because it is one of the most practical algorithms on real images.

Once the saliency map S_I is ready, our goal is to find the crop rectangle R_C that is expected to contain the most informative part of the image. Since the saliency map is used as the criteria of importance, the sum of saliency within R_C should contain most of the saliency value in S_I. Based on this idea, we can find R_C as the smallest rectangle containing a fixed fraction of saliency. To illustrate this formally, we define candidates set $\Re(\lambda)$ for R_C and the fraction threshold λ as

$$\Re(\lambda) = \left\{ r : \frac{\sum_{(x,y)\in r} S_I(x,y)}{\sum_{(x,y)} S_I(x,y)} > \lambda \right\}$$

Then R_C is given by

$$R_C = \arg\min_{r\in\Re(\lambda)}(area(r))$$

R_C denotes the minimum rectangle that satisfies the threshold defined above. A brute force algorithm was developed to compute R_C.

Find Cropping Rectangle with Fixed Threshold using Greedy Algorithm

The brute force method works, however, it is not time efficient. Two main factors slow down the computation. First, the algorithm to compute the saliency map involves several series of iterations. Some of the iterations involve convolutions using very large filter templates (on the order of the size of the saliency map). These convolutions make the computation very time consuming.

Second, the brute force algorithm basically searches all sub-rectangles exhaustively. While techniques exist to speed up this exhaustive search, it still takes a lot of time.

We found that we can achieve basically the same results much more efficiently by: 1) using fewer iterations and smaller filter templates during the saliency map calculation; 2) squaring the saliency to enhance it; 3) using a greedy search instead of brute force method by only considering rectangles that include the peaks of the saliency.

```
Rectangle GREEDY_CROPPING (S, λ)
thresholdSum ← λ * Total saliency value in S
R_c ← the center of S
currentSaliencySum ← saliency value of R_c
WHILE currentSaliencySum < thresholdSum DO
  P ← Maximum saliency point outside R_c
  R' ← Small rectangle centered at P
  R_c ← UNION(R_c, R')
  UPDATE currentSaliencySum with new region R_c
ENDWHILE
RETURN R_c
```

Figure 2: Algorithm to find cropping rectangle with fixed saliency threshold. S is the input saliency map and λ is the threshold.

Figure 2 shows the algorithm GREEDY_CROPPING to find the cropping rectangle with fixed saliency threshold λ . The greedy algorithm calculates R_C by incrementally including the next most salient peak point P. Also when including a salient point P in R_C, we union R_C with a small rectangle centered at P. This is because if P is within the foreground object, it is expected that a small region surrounding P would also contain the object.

This algorithm can be modified to satisfy further requirements. For example, the UNION function in Figure 2 can be altered when the cropped rectangle should have the same aspect ratio as the original image. Rather than just merging two rectangles, UNION needs to calculate the minimum surrounding bounds that have the same aspect ratio as the original image. As another example, the initial value of R_C can be set to either the center of image, S, or the most salient point or any other point. Since the initial point always falls in the result thumbnail, it can be regarded as a point with extremely large saliency. When the most salient point is selected as an initial point, the result can be optimized to have the minimum size. But, we found that to begin the algorithm with the center of images gives more robust and faster results even though it might increase the size of the result thumbnail especially when all salient points are skewed to one side of an image.

Find Cropping Rectangle with Dynamic Threshold

Experience shows that the most effective threshold varies from image to image. We therefore have developed a method for adaptively determining the threshold λ .

Intuitively, we want to choose a threshold at a point of diminishing returns, where adding small amounts of additional saliency requires a large increase in the rectangle. We use an area-threshold graph to visualize this. The X axis indicates the threshold (fraction of saliency) while the Y axis shows the normalized area of the cropping rectangle as the result of the greedy algorithm mentioned above. Here the normalized area has a value between 0 and 1. The solid curve in Figure 3 gives an example of an area-threshold graph.

A natural solution is to use the threshold with maximum gradient in the area-threshold graph. We approximate this using a binary search method to find the threshold in three steps: First, we calculate the area-threshold graph for the given image. Second, we use a binary search method to find the threshold where the graph goes up quickly. Third, the threshold is tuned back to the position where a local maximum gradient exists. The dotted lines in Figure 3 demonstrate the process of finding the threshold for the image given in Figure 1.

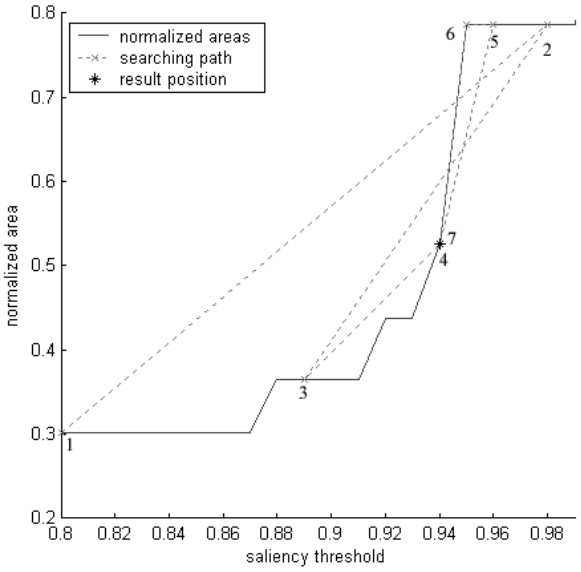

Figure 3: The solid line represents the area-threshold graph. The dotted lines show the process of searching for the best threshold. The numbers indicate the sequence of searching

Examples of Saliency Map Based Cropping

After getting R_C, we can directly crop the input image I. Thumbnails of the image given in Figure 1 are shown in Figure 4. It is clear from Figure 4 that the cropped thumbnail can be more easily recognized than the thumbnail without cropping.

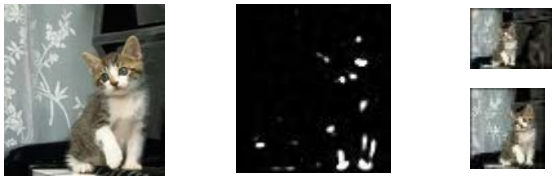

Figure 4 (left): the image cropped based on the saliency map; (middle): the cropping rectangle which contains most of the saliency parts; (right top): a thumbnail subsampled from the original image; (right bottom): a thumbnail subsampled from the cropped image (left part of this figure).

Figure 5 shows the result of an image whose salient parts are more scattered. Photos focusing primarily on the subject and without much background information often have this property. A merit of our algorithm is that it is not sensitive to this.

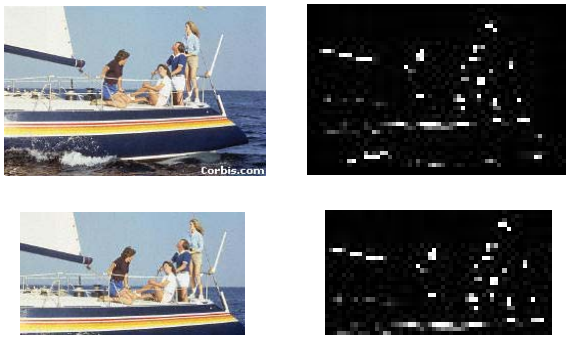

Figure 5 (left top): the original image (courtesy of Corbis [7]); (right top): the saliency map; (left bottom): the cropped image; (right bottom): the cropped saliency map which contains most of the salienct parts.

Face Detection Based Cropping

In the above section, we proposed a general method for thumbnail cropping. The method relies only on low-level features. However, if our goal is to make the objects of interest in an image more recognizable, we can clearly do this more effectively when we are able to automatically detect the position of these objects.

Images of people are essential in a lot of research and application areas. At the same time, face processing is a rapidly expanding area and has attracted a lot of research effort in recent years. Face detection is one of the most important problems in the area. [22] surveys the numerous methods proposed for face detection.

For human image thumbnails, we claim that recognizability will increase if we crop the image to contain only the face region. Based on this claim, we designed a thumbnail cropping approach based on face detection. First, we identify faces by applying CMU's on-line face detection [9][19] to the given images. Then, the cropping rectangle R_C is computed as containing all the detected faces. After that, the thumbnail is generated from the image cropped from the original image by R_C.

Figure 6 (left): the original image; (middle): the face detection result from CMU's online face detection [9]; (right): the cropped image based on the face detection result.

Figure 6 shows an example image, its face detection result and the cropped image. Figure 7 shows the three thumbnails generated via three different methods. In this example, we can see that face detection based cropping method is a very effective way to create thumbnails, while saliency based cropping produces little improvement because the original image has few non-salient regions to cut.

Figure 7: Thumbnails generated by the three different methods. (left): without cropping; (middle): saliency based cropping; (right): face detection based cropping.

USER STUDY

We ran a controlled empirical study to examine the effect of different thumbnail generation methods on the ability of users to recognize objects in images. The experiment is divided into two parts. First, we measured how recognition rates change depending on thumbnail size and thumbnail generation techniques. Participants were asked to recognize objects in small thumbnails (Recognition Task). Second, we measured how the thumbnail generation technique affects search performance (Visual Search Task). Participants were asked to find images that match given descriptions.

Design of Study

The recognition tasks were designed to measure the successful recognition rate of thumbnail images as three conditions varied: image set, thumbnail technique, and thumbnail size. We measured the correctness as a dependent variable.

The visual search task conditions were designed to measure the effectiveness of image search with thumbnails generated with different techniques. The experiment employed a 3x3 within-subjects factorial design, with image set and thumbnail technique as independent variables. We measured search time as a dependant variable. But, since the face-detection clipping is not applicable to the Animal Set and the Corbis Set, we omitted the visual search tasks with those conditions as in Figure 8. The total duration of the experiment for each participant was about 45 minutes.

Thumbnail Technique	Animal Set	Corbis Set	Face Set
Plain shrunken thumbnail	√	√	√
Saliency based cropping	√	√	√
Face detection based cropping	X	X	√

Figure 8: Visual search task design. Checkmarks (√) show which image sets were tested with which image cropping techniques.

Participants

There were 20 participants in this study. Participants were college or graduate students at the University of Maryland at College Park recruited on the campus. All participants were familiar with computers. Before the tasks began, all participants were asked to pick ten familiar persons out of fifteen candidates. Two participants had difficulty with choosing them. Since the participants must recognize the people whose images are used for identification, the results from those two participants were excluded from the analysis.

Image Sets

We used three image sets for the experiment. We also used filler images as distracters to minimize the duplicate exposure of images in the visual search tasks. There were 500 filler images and images were randomly chosen from this set as needed. These images were carefully chosen so that none of them were similar to images in the three test image sets.

Animal Set (AS)

The "Animal Set" includes images of ten different animals and there are five images per animal. All images were gathered from various sources on the Web. The reason we chose animals as target images was to test recognition and visual search performance with familiar objects. The basic criteria of choosing animals were 1) that the animals should be very familiar so that participants can recognize them without prior learning; and 2) they should be easily distinguishable from each other. As an example, donkeys and horses are too similar to each other. To prevent confusion, we only used horses.

Corbis Set (CS)

Corbis is a well known source for digital images and provides various types of tailored digital photos [7]. Its images are professionally taken and manually cropped. The goal of this set is to represent images already in the best possible shape. We randomly selected 100 images out of 10,000 images. We used only 10 images as search targets for visual search tasks to reduce the experimental errors. But during the experiment, we found that one task was problematic because there were very similar images in the fillers and sometimes participants picked unintended images as an answer. Therefore we discarded the result from the task. A total of five observations were discarded due to this condition.

Face Set (FS)

This set includes images of fifteen well known people who are either politicians or entertainers. Five images per person were used for this experiment. All images were gathered from the Web. We used this set to test the effectiveness of face detection based cropping technique and to see how the participants' recognition rate varies with different types of images.

Some images in this set contained more than one face. In this case, we cropped the image so that the resulting image contains all the faces in the original image. Out of 75 images, multiple faces were detected in 25 images. We found that 13 of them contained erratic detections. All erroneously detected faces were included in the cropped thumbnail sets since we intended to test our cropping method with available face detection techniques, which are not perfect.

Thumbnail Techniques

Plain shrinking without cropping

The images were scaled down to smaller dimensions. We prepared ten levels of thumbnails from 32 to 68 pixels in the larger dimension. The thumbnail size was increased by four pixels per level. But, for the Face Set images, we increased the number of levels to twelve because we found that some faces are not identifiable even in a 68 pixel thumbnail.

Cropping Technique and Image Set		Ratio	Variance
Saliency based cropping	Corbis Set	61.3%	0.110
	Animal Set	53.9%	0.127
	Face Set	54.3%	0.128
	All	57.6%	0.124
Face detection based cropping (Face Set)		16.1%	0.120

Figure 9: Ratio of cropped to original image size.

Saliency based cropping

By using the saliency based cropping algorithms described above, we cropped out background of the images. Then we shrunk cropped images to ten sizes of thumbnails. Figure 9 shows how much area was cropped for each technique.

Face detection based cropping

Faces were detected by CMU's algorithm as described above. If there were multiple faces detected, we chose the bounding region that contains all detected faces. Then twelve levels of thumbnails from 36 to 80 pixels were prepared for the experiment.

Recognition Task

We used the "Animal Set" and the "Face Set" images to measure how accurately participants could recognize objects in small thumbnails. First, users were asked to identify animals in thumbnails. The thumbnails in this task were chosen randomly from all levels of the Animal Set images. This task was repeated 50 times.

When the user clicked the "Next" button, a thumbnail was shown as in Figure 10 for two seconds. Since we intended to measure pure recognizability of thumbnails, we limited the time thumbnails were shown. According to our pilot user study, users tended to guess answers even though they could not clearly identify objects in thumbnails when they saw them for a long time. To discourage participants' from guessing, the interface was designed to make thumbnails disappear after a short period of time, two seconds. For the same reason, we introduced more animals in the answer list. Although we used only ten animals in this experiment, we listed 30 animals as possible answers as seen in Figure 10, to limit the subject's ability to guess identity based on crude cues. In this way, participants were prevented from choosing similarly shaped animals by guess. For example, when participants think that they saw a bird-ish animal, they would select swan if it is the only avian animal. By having multiple birds in the candidate list, we could prevent those undesired behaviors.

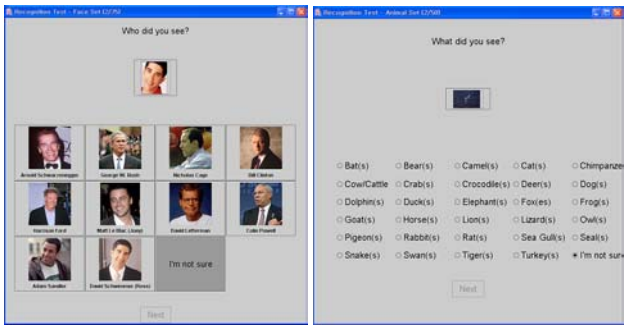

Figure 10: Recognition task interfaces. Participants were asked to click what they saw or "I'm not sure" button. Left: Face Set recognition interface, Right: Animal Set recognition interface

After the Animal Set recognition task, users were asked to identify a person in the same way. This Face Set recognition task was repeated 75 times. In this session, the candidates were shown as portraits in addition to names as seen in Figure 10.

Visual Search Task

For each testing condition in Figure 8, participants were given two tasks. Thus, for each visual search session, fourteen search tasks were assigned per participant. The order of tasks was randomized to reduce learning effects.

As shown in Figure 11, participants were asked to find one image among 100 images. For the visual search task, it was important to provide equal search conditions for each task and participant. To ensure fairness, we designed the search condition carefully. We suppressed the duplicate occurrences of images and manipulated the locations of the target images.

For the Animal Set search tasks, we randomly chose one target image out of 50 Animal Set images. Then we carefully selected 25 non-similar looking animal images. After that we mixed them with 49 more images randomly chosen from the filler set as distracters. For the Face Set and Corbis Set tasks, we prepared the task image sets in the same way.

The tasks were given as verbal descriptions for the Animal Set and Corbis set tasks. For the Face Set tasks, a portrait of a target person was given as well as the person's name. The given portraits were separately chosen from an independent collection so that they were not duplicated with images used for the tasks.

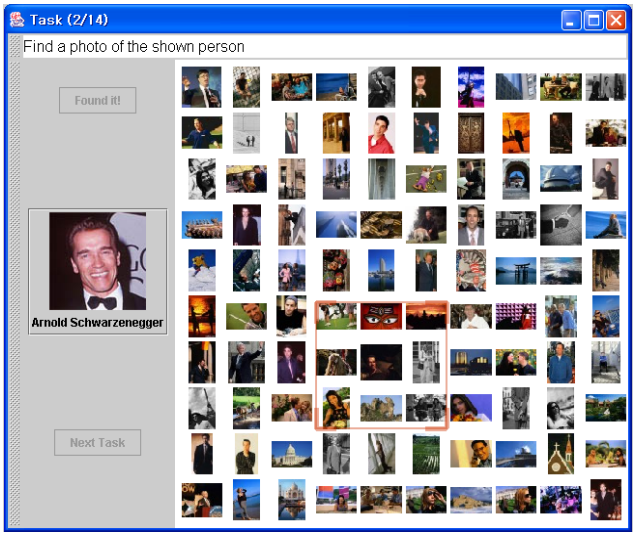

Figure 11: Visual search task interface. Participant were asked to find an image that matches a given task description. Users can zoom in, zoom out, and pan freely until they find the right image.

We used a custom-made image browser based on PhotoMesa [3] as our visual search interface. PhotoMesa provides a zooming environment for image navigation with a simple set of control functions. Users click the left mouse button to zoom into a group of images (as indicated by a red rectangle) to see the images in detail and click the right mouse button to zoom out to see more images to overview. Panning is supported either by mouse dragging or arrow keys. The animation between zooming helps user to remember where things fit together based on spatial relationships. PhotoMesa can display a large number of thumbnails in groups on the screen at the same time. Since this user study was intended to test pure visual search, all images were presented in a single cluster as in Figure 11.

Participants were allowed to zoom in, zoom out and pan freely for navigation. When users identify the target image, they were asked to zoom into the full scale of the image and click the "Found it" button located on the upper left corner of the interface to finish the task. Before the visual search session, they were given as much time as they wanted until they found it comfortable to use the zoomable interface. Most participants found it very easy to navigate and reported no problem with the navigation during the session.

RECOGNITION TASK RESULTS

Figure 12 shows the results from the recognition tasks. The horizontal axis represents the size of thumbnails and the vertical axis denotes the recognition accuracy. Each data point in the graph denotes the successful recognition rate of the thumbnails at that level. As shown, the bigger the thumbnails are, the more accurately participants recognize objects in the thumbnails. And this fits well with our intuition. But the interesting point here is that the automatic cropping techniques perform significantly better than the original thumbnails.

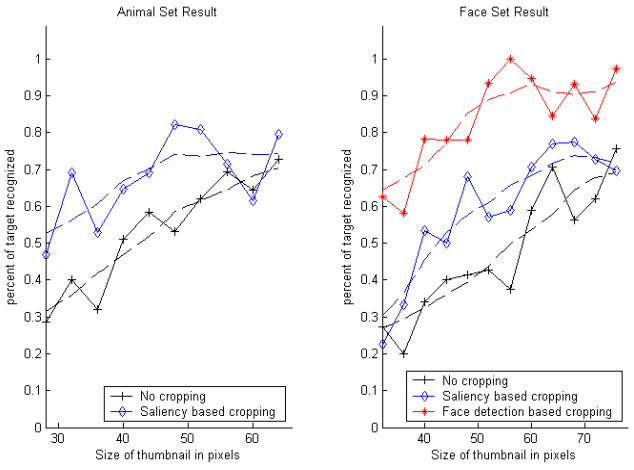

Figure 12: Recognition Task Results. Dashed lines are interpolated from jagged data points.

There were clear correlations in the results. Participants recognized objects in bigger thumbnails more accurately regardless of the thumbnail techniques. Therefore, we used Paired T-test (two tailed) to analyze the results. The results are shown in Figure 13.

The first graph shows the results from the "Animal Set" with two different thumbnail techniques, no cropping and saliency based cropping. As clearly shown, users were able to recognize objects more accurately with saliency based cropped thumbnails than with plain thumbnails with no cropping. One of the major reasons for the difference can be attributed to the fact that the effective portion of images is drawn relatively larger in saliency based cropped images. But, if the main object region is cropped out, this would not be true. In this case, the users would see more non-core parts of images and the recognition rate of the cropped thumbnails would be less than that of plain thumbnails. The goal of this test is to measure if saliency based cropping cut out the right part of images. The recognition test result shows that participants recognize objects better with saliency based thumbnails than plain thumbnails. Therefore, we can say that saliency based cropping cut out the right part of images.

Condition	*t*-Value	P value
No cropping vs. Saliency based cropping on Animal Set	4.33	0.002
No cropping vs. Saliency based cropping on Face Set	4.16	0.002
No cropping vs. Face Detection based cropping on Face Set	9.56	< 0.001
Saliency based cropping vs. Face detection based cropping on Face Set	7.34	< 0.001
Animal Set vs. Face Set with no cropping	5.00	0.001
Animal Set vs. Face Set with saliency based cropping	3.08	0.005

Figure 13: Analysis results of Recognition Task (Paired T-Test). Every curve in Figure 12 is significantly different from each other.

During the experiment, participants mentioned that the background sometimes helped with recognition. For example, when they saw blue background, they immediately suspected that the images would be about sea animals. Similarly, the camel was well identified in every thumbnail technique even in very small scale thumbnails because the images have unique desert backgrounds (4 out of 5 images).

Since saliency based cropping cuts out large portion of background (42.4%), we suspected that this might harm recognition. But the result shows that it is not true. Users performed better with cropped images. Even when

background was cut out, users still could see some of the background and they got sufficient help from this information. It implies that the saliency based cropping is well balanced. The cropped image shows the main objects bigger while giving enough background information.

The second graph shows results similar to the first. The second graph represents the results from the "Face Set" with three different types of thumbnail techniques, no cropping, saliency based cropping, and face detection based cropping. As seen in the graph, participants perform much better with face detection based thumbnails. It is not surprising that users can identify a person more easily with images with bigger faces.

Compared to the Animal Set result, the Face Set images are less accurately identified. This is because humans have similar visual characteristics while animals have more distinguishing features. In other words, animals can be identified with overall shapes and colors but humans cannot be distinguished easily with those features. The main feature that distinguishes humans is the face. The experimental results clearly show that participants recognized persons better with face detection based thumbnails.

The results also show that saliency cropped thumbnails are useful for recognizing humans as well as animals. We found that saliency based cropped images include persons in the photos so that persons in the images can be presented larger in cropped images. The test results show that the saliency based cropping does increase the recognition rate.

In this study, we used two types of image sets and three different thumbnail techniques. To achieve a higher recognition rate, it is important to show major distinguishing features. If well cropped, a small sized thumbnail would be sufficient to represent the whole image. Face detection based cropping shows benefits when this type of feature extraction is possible. But, in a real image browsing task, it is not always possible to know users' searching intention. For the same image, users' focus might be different for browsing purposes. For example, users might want to find a person at some point, but the next time, they would like to focus on costumes only. We believe that the saliency based cropping technique can be applied in most cases when semantic object detection is not available or users' search behavior is not known.

In addition, the recognition rate is not the same for different types of images. This implies that the minimum recognizable size should be different depending on image types.

VISUAL SEARCH TASK RESULTS

Figure 14 shows the result of the visual search tasks. Most participants were able to finish the tasks within the 120 second timeout (15 timeouts out of 231 tasks) and also chose the desired answer (5 wrong answers out of 231 tasks). Wrong answers and timed out tasks were excluded from the analysis.

A two way analysis of variance (ANOVA) was conducted on the search time for two conditions, thumbnail technique and image sets. As shown, participants found the answer images faster with cropped thumbnails. Overall, there was a strong difference for visual search performance depending to thumbnail techniques, $F(2, 219) = 5.58$, $p = 0.004$.

Since we did not look at face detection cropping for the Animal Set and the Corbis Set, we did another analysis with the two thumbnail techniques (plain thumbnail, saliency based cropped thumbnail) to see if the saliency based algorithm is better. The result shows a significant improvement on visual search with saliency based cropping, $F(1, 190) = 3.823$, $p = 0.05$. We therefore believe that the proposed saliency based cropping algorithm make a significant contribution to visual search.

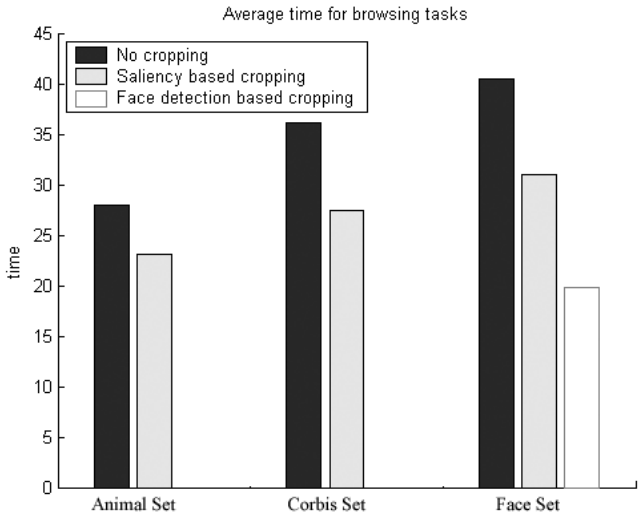

Figure 14: Visual search task results.

Condition	F value	P value
Thumbnail techniques on three sets	5.58	0.004
Thumbnail techniques on Face Set	4.56	0.013
No cropping vs. Saliency based thumbnail on three image sets	3.82	0.052
Three image sets regardless of thumbnail techniques	2.44	0.089

Figure 15 List of ANOVA results from the visual search task

When the results from the Face Set alone were analyzed by one way ANOVA with three thumbnail technique conditions, there also was a significant effect, $F(2, 87)=4.56$, $p = 0.013$. But for the Animal Set and the Corbis

Set, there was only a borderline significant effect over different techniques. We think that this is due to the small number of observations. We believe those results would also be significant if there were more participants because there was a clear trend showing an improvement of 18% on the Animal Set and 24% on the Corbis Set. Lack of significance can also be attributed to the fact that the search task itself has large variances by its nature. We found that the location of answer images affects the visual search performance. Users begin to look for images from anywhere in the image space (Figure 11). Participants scanned the image space from the upper-left corner, from the lower-right corner, or sometimes randomly. If the answer image is located in the initial position of users' attention, it would be found much earlier. Since we could not control users' behavior, we randomized the location of the answer images. But as a result, there was large variance.

Before the experiment, we were afraid that the cropped thumbnails of the Corbis Set images would affect the search result negatively since the images in the Corbis Set are already in good shape and we were concerned that cutting off their background would harm participants' visual search. But according to our result, saliency based cropped thumbnails does not harm users' visual search. Rather, it showed a tendency to increase participants' search performance. We think that this is because the saliency based cropping algorithm cut the right amount of information without removing core information in the images. At least, we can conclude that it did not make visual search worse to use the cropped thumbnails.

Another interesting thing we found is that the visual search task with the Animal Set tends to take less time than with the Corbis Set and the Face Set, $F(2, 219) = 2.44$, $p = 0.089$. This might be because the given Corbis Set and Face Set tasks were harder than the Animal Set. But we think there is another interesting factor. During the experiment, when he found the answer image after a while, one participant said that "Oh… This is not what I expected. I expected blue background when I'm supposed to find an airplane." Since one of the authors was observing the experiment session, it was observed that the participant passed over the correct answer image during the search even though he saw the image at reasonably big scale. Since all of the visual search tasks except finding faces were given as verbal descriptions, users did not have any information about what the answer images would be like. We think that this verbal description was one of the factors in performance differences between image sets. We found that animals are easier to find by guessing background than other image sets.

DISSCUSSION AND CONCLUSION

We developed and evaluated two automatic cropping methods. A general thumbnail cropping method based on a saliency model finds the informative portion of images and cuts out the non-core part of images. Thumbnail images generated from the cropped part of images increases users' recognition and helps users in visual search. This technique is general and can be used without any prior assumption about images since it uses only low level features. Furthermore, it also can be used for images already in good shape. Since it dynamically decides how much to cut away, it can prevent cutting out too much.

The face detection based cropping technique shows how semantic information can be used to enhance thumbnail cropping. With a face detection technique, we created more effective thumbnails, which significantly increased users' recognizing and finding performance.

Our study shows strong empirical evidence that the more salient a portion of image, the more informative it is. We also showed that using more recognizable thumbnails increases visual search performance.

Another finding of interest is that users tend to have mental models about search targets. Users tend to develop a model about what a target will look like by guessing its color and shape. We observed that they spent a long time searching or even skipped the correct answer when their guesses were wrong or they were unable to guess. It is known that humans have an "attentional control setting" – a mental setting about what they are (and are not) looking for while performing a given task. Interestingly, it is also known that humans have difficulty in switching their attentional control setting instantaneously [11]. This theory explains our observation. We think that this phenomenon should be regarded in designing image browsing interfaces especially in situations where users need to skim a large number of images.

There are several interesting directions for future research. One direction involves determining how to apply these techniques to other browsing environments. In our study, we used a zoomable interface for visual search. We believe that the image cropping techniques presented in this paper can benefit other types of interfaces that deal with a large number of images as well. While our research confirms that well cropped thumbnails can increase users' visual search performance, we did not try to build a model about recognition, attention and its relationship on image browsing. Further research about human's attention and perception model [18][21] would help designing a better image browsing system.

Another interesting direction would be to combine image adaptation techniques (i.e. saliency based smoothing) with the image cropping techniques. This would allow faster thumbnail processing and delivery for thumbnail-based retrieval systems.

ACKNOWLEDGMENTS

We would like to acknowledge the face group at Carnegie Mellon University for providing resources for face detection processing.

REFERENCES

1. ACDSee, ACD Systems, http://www.adsystems.com
2. Adobe Photoshop Album, Adobe Systems Inc., http://www.adobe.com/products/photoshopalbum/
3. Bederson, B. B. PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps. *UIST 2001, ACM Symposium on User Interface Software and Technology, CHI Letters*, *3*(2), pp. 71-80. 2001.
4. Burton, C., Johnston, L., and Sonenberg, E. Case Study: An Empirical Investigation of Thumbnail Image Recognition, *In Proc. of Information Visualization, Atlanta, Georgia,* pp115-121, 1995.
5. Chen, L., Xie, X., Fan, X., Ma, W., Zhang, H., and Zhou, H. A Visual attention model for adapting images on small displays, *ACM Multimedia Systems Journal,* To appear in Fall 2003.
6. Chen, X., and Zhang, H. Text Area Detection from Video Frames. *In Proc. Of 2nd IEEE Pacific-Rim Conference on Multimedia (PCM2001)*, Beijing, China, pp. 222-228, October 2001.
7. Corbis, http://www.corbis.com
8. DeCarlo, D., and Santella, A. Stylization and Abstraction of Photographs, *In Proc. on ACM SIGGRAPH 2002*, pp. 769-776, 2002.
9. Face Detection Demonstration. Robotics Institute, Carnegie Mellon University http://www.vasc.ri.cmu.edu/cgi-bin/demos/findface.cgi
10. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. Query by Image and Video Content: The QBIC System, *IEEE Computer*, Volume: 28, Issue: 9 , pp.23 -32, Sept. 1995.
11. Folk, C.L., Remington, R.W., and Johnston, J.C. Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: HP&P*, 18:1030-44, 1992.
12. Itti, L., and Koch, C. A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems, *SPIE human vision and electronic imaging IV(HVEI'99),* San Jose, CA, pp. 473-482, 1999.
13. Itti, L., Koch, C., and Niebur, E., A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20(11), pp. 1254-9, 1998.
14. Kang, H., and Shneiderman, B. Visualization Methods for Personal Photo Collections: Browsing and Searching in the PhotoFinder, *In Proc. of IEEE International Conference on Multimedia and Expo (ICME2000)* New York: IEEE, pp. 1539-1542, 2000.
15. Li, S., Zhu, L., Zhang, Z., Blake, A., Zhang, H., and Shum, H. Statistical Learning of Multi-view Face Detection. *In Proc. of European Conference on Computer Vision (ECCV) 2002*, Vol. 4, pp. 67-81, 2002.
16. Milanese, R., Wechsler H., Gil S., Bost J., and Pun T. Integration of Bottom-Up and Top-Down Cues for Visual Attention Using Non-Linear Relaxation, *In proc. of Computer Vision and Pattern Recognition, IEEE*, pp. 781-785, 1994.
17. Milanese, R. Detecting Salient Regions in an Image: from Biological Evidence to Computer Implementation, *Ph.D. thesis, Univ. of Geneva,* 1993.
18. Palmer, J. Ames, C.T., Lindsey, D.T. Measuring the Effect of Attention on Simple Visual Search. *Journal of Experimental Psychology: Human Perception & Performance*, 19, pp. 108-130, 1993.
19. Schneiderman, H., and Kanade, T. A Statistical Model for 3D Object Detection Applied to Faces and Cars. *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, June, 2000.
20. Vimas Technologies. http://www.vimas.com
21. Wolfe, J.M. Guided Search 2.0: A Revised Model of Visual Search, *Psychonomic Buttletin and Review,* Vol. 1, No. 2, pp. 202-238, 1994.
22. Yang, M., Kriegman, D., and Ahuja, N. Detecting Faces in Images: A Survey, *IEEE Transactions on Pattern Analysis and Mach Intelligence,* 24(1), pp. 34-58, 2002.