Gaze-Based Interaction for Semi-Automatic Photo Cropping

Anthony Santella[*] **Maneesh Agrawala**[†] **Doug DeCarlo**[*] **David Salesin**[‡] **Michael Cohen**[§]

ABSTRACT

We present an interactive method for cropping photographs given minimal information about the location of important content, provided by eye tracking. Cropping is formulated in a general optimization framework that facilitates adding new composition rules, as well as adapting the system to particular applications. Our system uses fixation data to identify important content and compute the best crop for any given aspect ratio or size, enabling applications such as automatic snapshot recomposition, adaptive documents, and thumbnailing. We validate our approach with studies in which users compare our crops to ones produced by hand and by a completely automatic approach. Experiments show that viewers prefer our gaze-based crops to uncropped images and fully automatic crops.

Author Keywords

cropping, photography, composition, evaluation, eye tracking, visual perception

ACM Classification Keywords

H5.2. Information interfaces and presentation: User Interfaces.

INTRODUCTION

In art, there is a common saying: what you leave out is as important as what you put in. Cropping photographs is an art that consists entirely of leaving out. Successful crops alter the composition of an image to emphasize the most important image content by framing and enlarging it, while simultaneously removing distracting elements. As shown in Figure 1, an effective crop focuses the viewer's attention on the subject of the image, while a poor crop is distracting.

In standard photo-editing applications, a designer directly draws a crop rectangle around the important content. Producing a pleasing crop is not in itself very difficult. However, even a professional designer usually needs a period of focused effort to generate a single crop. Hand-crafting attractive crops for large collections of images, or creating crops with multiple aspect ratios—necessary for adaptive documents or different standard size prints—is time-consuming and burdensome. As a result, many photographers do not crop their photographs.

As an alternative, we present an implicit, attentive interface [31] for cropping. Users simply look at each image for a few seconds, while the system records their eye movements. Our system uses these recordings to identify the important image content and can then automatically generate crops of any size or aspect ratio. Beyond cropping, accurate identification of relevant image content without explicit interaction is an important problem. It can enable applications that monitor or respond to user gaze in images like video and image transmission, analysis and quantification of viewing behavior over images, and region-of-interest (ROI) selection in image editing.

Our approach builds on the work of Suh et al. [29] and Chen et al. [5], who have developed fully automated salience-based image-cropping techniques requiring no user input. Both of these systems identify important image areas using a bottom-up computational model of visual salience based on low-level contrast measures [16], and an image-based face detection system [20]. However, because these techniques consider only low-level features and faces, they often miss other important features in an image and can generate extremely awkward crops. In contrast, our system relies on the empirical salience revealed by the eye movements of a viewer. We can more reliably identify a photo's subject, and therefore generate more effective crops. Although our system incurs the cost of requiring some user input, this cost is minor if eye movements are recorded surreptitiously; users, or viewers, of images need to look at the images almost by definition. Once gaze data is collected, the system can generate families of crops at any desired size and aspect ratio.

[*]Rutgers University, Computer Science and Cognitive Science
110 Frelinghuysen Road, Piscataway, NJ 08854-8019
{asantell,decarlo}@cs.rutgers.edu
[†]UC Berkeley Computer Science
387 Soda Hall, Berkeley, CA 94720-1776
maneesh@cs.berkeley.edu
[‡]Adobe Systems and University of Washington
801 North 34th St., Seattle, Washington 98103
salesin@adobe.com
[§]Microsoft Research, One Microsoft Way, Redmond WA 98052
mcohen@microsoft.com

CHI 2006, April 22-28, 2006, Montréal, Québec, Canada.

(a) original (b) gaze-based (c) automatic

Figure 1. Cropping can drastically change the impact of a photo. Compared to the original snapshot, a good crop as produced by our gaze-based system can improve an image. A poorly chosen crop, as produced by a fully automatic method [29], can be distracting. Automatic methods using computational models of visual salience confuse prominent but unimportant features with relevant content. Here the yellow light in the upper left background is visually prominent but does not contribute to the meaning of the image.

Our system treats cropping as an optimization problem. It searches the space of possible crops to find one that respects the interest of the viewer and abides by a few basic rules of composition. Our goal is to create aesthetically pleasing crops without explicit interaction. Accordingly, we validate our approach with forced-choice experiments in which subjects compare the aesthetics of crops made using gaze-based interaction to automatic and handmade crops. Gaze-based crops were judged superior to all but handmade crops.

The specific contributions of our work include:

- A novel general algorithm for quantifying the importance of image content based on recorded eye movements.
- Identification and implementation of a set of compositional rules that allow the quantitative evaluation of a crop.
- User studies validating the appropriateness and effectiveness of our approach in comparison to previous techniques.

We first review relevant background on composition and identifying important image content. Then we describe and evaluate our gaze-based cropping algorithm.

BACKGROUND AND RELATED WORK

Composition in psychology and photography

Psychology and art history suggest that good composition is an objective quality that is amenable to computational assessment. Art historians have proposed that pleasing or dynamically balanced composition is a perceptual phenomenon that arises spontaneously from the interaction of "visual forces" across an image [1, 2]. This view has been substantiated by experimental research [23, 24, 26] in cognitive psychology. Eye movements may play an important role in judgments about composition (see Locher [22] for a review). A similar note is struck in critical discussions of composition. Graham [13] for example describes how to compose an image by placing objects that lead the viewer's eye on a pleasing path through the scene's various centers of interest. Unfortunately, these investigations are largely qualitative; no experimentally based model of composition exists.

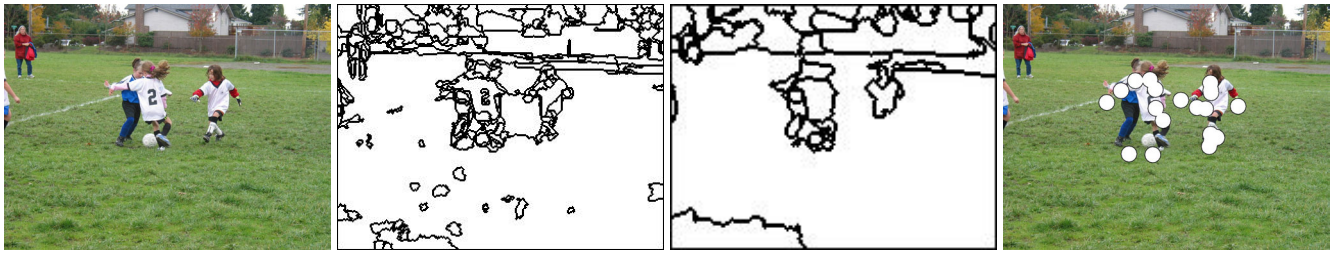
A key aspect of practical composition, emphasized by photography manuals (e.g., [14, 27]) is being aware of what is in the viewfinder. Peterson [27] specifically suggests scanning all four sides of the viewfinder to ensure that only relevant content is included, and that no distracting fragments of outside objects intrude. He also suggests "filling the frame" so the subject takes up most of the available space. Both techniques are presented as solutions to the tendency of casual photographers to create disorganized, cluttered images.

In addition to these rules, practical discussions of composition often mention placement of subject matter according to geometrical criteria such as centering, the rule of thirds (or fifths), and the golden ratio. All of these compositional formulas postulate that a composition is pleasing when subject matter is placed in specific locations. However, it is important to recall that all of these "rules" are intended as rules of thumb: not laws, but heuristics that are as often broken as obeyed. Nevertheless, these rules are worth investigating because of their long history of successful use.

Computational approaches to composition

Previous work in automated composition has focused on implementing simple rules for subject placement. The rule of thirds has been used to position automatically detected features of interest in an automatic robot camera [4], and in prototype on-camera composition assistance [3]. The same kind of approach, using the rules of thirds and fifths, has been used to place silhouette edges in automated view selection of 3D models [12]. Another compositional heuristic, that features should be balanced from left to right, has been used to arrange images and text objects in a window [25].

Maximizing the area of an image devoted to subject matter is an alternative standard of composition. Subjects that fill the frame of a picture are usually considered to have greater impact. In thumbnailing, tight crops also maximize the size of important features. Tight thumbnails have been created by cropping out nearly featureless areas of photographs using salience techniques [5, 29]. The same goal has been pursued by cutting out important scene elements (identified via salience and face detection) and pasting them, more closely packed, into an in-filled background [28]. Both of these tech-



(a) (b) (c) (d)

Figure 2. (a) Photo I; (b) Fine segmentation S_{fine}; (c) Coarse segmentation S_{coarse}; (d) Fixation locations

niques are concerned primarily with compact presentation, not aesthetics or preserving the meaning of an image.

Minimal interaction for identifying important content

Interaction to identify important content can be avoided by assuming that some easily extractable feature is the subject. Important objects like faces are strong candidates, but face detection is far from robust. Prominent low-level features are an alternative. Vision scientists have had some success modeling the low-level features that attract the eye [15, 16]. These models can be used to make predictions about important content under the assumption that something distinctive must therefore be important [28, 29]. Prominence, however, is not consistently related to importance. As a result, this importance estimate can produce awkward images when used in tasks like cropping (see Figure 1).

Computational salience models predict where people look; an alternative, interactive approach is to record where a viewer actually looks. A person must interact with the system, but the interaction is minimal. Viewers naturally examine important features based on the task at hand. In most situations the task-relevant elements are the important content or subject matter of the image. A number of gaze-based applications and evaluation techniques use eye tracking to identify what the user thinks is important [9, 10, 30]. Although collecting eye movement data requires special equipment, for the user it is relatively quick and effortless. Our system does not require conscious pointing with one's eyes, which presents difficulties [17]. Rather, fixation information is collected surreptitiously as a measure of user attention. This mode of interaction has been termed an implicit attentive interface [31].

CROPPING

To define a good crop we first need a model that explicitly represents important image content. We build such a representation using segmentation and eye tracking. Next, our system chooses the best crop from a large set of possible crops. We have created an objective function that assigns a score to each potential crop based on general rules for creating attractive crops. We minimize this objective function over all possible crops to identify the best one.

Photo representation

To enable cropping we need to identify meaningful elements in a photo. We begin by segmenting the photo at two scales: one fine, and the other coarse. See Figure 2(b,c). A segmentation represents the image as a collection of contiguous regions of pixels that have similar color. It provides information about image regions in a way that typically correlates with the physical features in the visual scene.

The fine- and coarse-scale segmentations, S_{fine} and S_{coarse}, are produced by downsampling the photograph by a factor of 4 and 8, respectively, and performing a segmentation using EDISON [6].[1] In configuring EDISON we use a spatial bandwidth of 7 pixels and a color bandwidth of 6.5 in $L^*u^*v^*$ space for S_{fine}; these parameters are 10 and 9, respectively, for S_{coarse}.

Content identification

To identify important image content we identify segmentation regions that correspond to highly examined parts of the image. To measure how much each region is examined we cannot simply count fixations that rest in each region. Current segmentation techniques cannot perfectly capture all scene boundaries, so regions may not represent complete features. Furthermore, current eye tracking systems are not very precise, so fixations may only be near, not on, their target. Accordingly, we make a soft assignment between fixations and nearby regions, to conservatively estimate where the viewer looked.

Input to content identification is a set of fixation locations $\mathbf{x}_k$ for $k \in [1 \ldots N]$, and a corresponding set of durations t_k. An example set of fixations are displayed in Figure 2(d). Each dot marks a fixation location; durations are not indicated. A fixation in the vicinity of a region raises the importance of that region, but only to the degree that the entire region was examined. For example, we would not want to significantly raise the importance of a large background area, such as the sky, simply because an object of interest was located nearby. To avoid this problem, we make the importance contributed by a fixation fall off sharply with distance, and average the contribution of all fixations over the entire region. We compute the average distance D between an input point $\mathbf{x}_k$ and a region $R \in S_{\text{fine}}$ over all the pixels in the region:

$$D(k,R) = \frac{1}{\|R\|} \sum_{i \in R} \|\mathbf{x}_k - i\| \qquad (1)$$

[1] EDISON is available at http://www.caip.rutgers.edu/riul.

Figure 3. Left is the initial labeling of foreground regions (white), background regions (black), and unlabeled regions (gray). Middle is the final binary labeling of foreground and background regions. Right is the final content map M.

We then compute the total interest in this region using a Gaussian-weighted sum of the corresponding fixation durations t_k. This importance estimate gives the relative time the viewer spent examining a particular region R:

$$m(R) = \sum_{k \in [1...N]} e^{\frac{-D(k,R)^2}{2\sigma^2}} t_k \tag{2}$$

Where σ controls the falloff of fixation influence, and in our experiments is set to a degree of visual angle, or about 20 pixels. This approach spreads the interest represented by a fixation but only to nearby regions of relatively small area.

Extracting complete objects

We want to extract entire examined areas even if these consist of multiple segmentation regions. Thus we use our importance estimate to guide the extraction of foreground objects using the "lazy snapping" [21] approach. This approach takes an oversegmentation of the image, and a partial labeling of foreground and background regions (hand labeled in the "lazy snapping" work). We use S_{fine} as the base segmentation, and assign foreground labels to the top 10th percentile of regions based on $m(R)$ scores. Regions in the bottom 50th percentile are labeled as background. A graph-cut optimization then assigns labels to the remaining regions. The content map is formed by coloring these foreground regions by their importance estimate $m(R)$. Regions identified as background remain black.

We assume that important content requires some space around it: the top of a person's head, for example, should not be up against the edge of a crop. Therefore, we extend each region's importance beyond its segment borders by averaging the importance image with a dilated version of itself. The dilation is performed with a flattened hemispherical kernel (radius 75 pixels). Finally, the result is scaled so that its maximum value is one—this yields the content map M, an example of which is shown in Figure 3.

Measures of crop quality

Given the segmentations and the content map, we define measures that evaluate the quality of potential crops. Our objective function is designed to consider three basic properties of a good photograph:

- A photo should include an entire subject and some context around it.
- Awkward intersections should be avoided between the edge of the crop and objects. In other words, the crop edges should pass through featureless areas when possible.
- The area of the photo covered by subject matter should be maximized to increase clarity.

We will consider each of these criteria in turn.

Including the subject: A crop must contain the important content. Thus, we define a term T_{subj} to be the percentage of the "mass" of the content map that is omitted. More precisely, given an image I and crop rectangle Ω, the content term is defined as

$$T_{\text{subj}}(\Omega) = 1 - \frac{\sum_{i \in \Omega} M(i)}{\sum_{i \in I} M(i)} \tag{3}$$

This term evaluates to 0 when the crop contains all of the important content in M, and to 1 when it contains none of it. Extreme violations of this rule—cutting out all the content, for example—are qualitatively worse than minor ones. We also include the square of this term, $T^2_{\text{subj}}(\Omega)$, in our optimization. This squared term has a strong influence only when much of the content is lost, resulting in extremely poor scores and preventing the important content from being completely cropped out.

Finally, to discourage cutting through parts of the subject, we penalize crops that pass through regions identified as content in M. The term T_{whole} is the average of all the pixels in the content map M through which the boundary of the crop rectangle passes:

$$T_{\text{whole}}(\Omega) = \frac{1}{\|\partial\Omega\|} \sum_{i \in \partial\Omega} M(i) \tag{4}$$

where $\partial\Omega$ is the set of pixels on the boundary of the rectangle Ω. This term encourages crops to cleanly include or exclude examined regions.

Avoiding cuts through background objects: We want our crop rectangle, when possible, to pass through homogeneous regions. If the crop cuts through prominent background features they can also appear awkward and draw attention.

Segmentation boundaries break the image into homogeneous areas. Therefore, we count the number of segmentation borders crossed by the edge of the crop rectangle $\partial\Omega$ in the coarse segmentation S_{coarse}—we call this $B(\partial\Omega)$. We count only crossings between regions with significantly different average colors (those more than a distance of 35 apart in $L^*u^*v^*$ space [11]). The term T_{cut} is:

$$T_{\text{cut}}(\Omega) = B(\partial\Omega) \big/ \|\partial\Omega\| \tag{5}$$

The number of crossings is normalized by the perimeter of the crop rectangle, in order to compute the density of noticeable region cuts along the crop border. Normalization makes this measure independent of the size of the crop.

Maximizing content area: To fill the crop rectangle with our content, we include the term T_{size} which penalizes the size of the crop. It is computed as the percentage of the area

(a) (b) (c) (d)

Figure 4. Positioning content: (a) a photo; (b) default crop; (c) crop using thirds rule; (d) centered.

of the original image that the crop includes:

$$T_{\text{size}}(\Omega) = \|\Omega\| / \|I\| \tag{6}$$

where $\|\Omega\|$ and $\|I\|$ are the areas of the crop and original image respectively.

Placing content

We also create rules for placing content at particular locations in the crop. Our approach is to find the "center" of the content, and then include an objective term that prefers crops that place this center at a particular location.

We begin by thresholding the content map M at 75% of its maximum in order to extract the most prominent content. We then calculate the centroids of each contiguous area of content. Our goal is to frame the crop so that these centers rest in particular locations. For each connected component C_i of content we calculate the distance $d(C_i)$ between its centroid and the closest target location. The objective is an area weighted average over the penalty for each component. To center the content, for example, the target location is the center of the crop rectangle $(\frac{1}{2}, \frac{1}{2})$ in normalized coordinates. For the rule of thirds, target locations are $(\frac{1}{3}, \frac{1}{3})$, $(\frac{1}{3}, \frac{2}{3})$, $(\frac{2}{3}, \frac{1}{3})$, and $(\frac{2}{3}, \frac{2}{3})$. The content placement term is then:

$$T_{\text{placement}}(\Omega) = \frac{1}{\sum_i A(C_i)} \sum_i A(C_i) \frac{d(C_i)}{d_{\max}} \tag{7}$$

where $d_{\max}$ is the distance to the target location from the furthest point in that crop and $A(C_i)$ is the area of the connected component. This formulation succeeds in positioning content (see Figure 4). Subjectively, however, we did not feel that it improved the appearance of most crops. It was therefore not used in our evaluation study.

Building the objective function

To form the objective function, the terms described in the previous sections are collected into a feature vector $\mathbf{T}$:

$$\mathbf{T}(\Omega) = \left[\; T_{\text{subj}}(\Omega) \;\; T^2_{\text{subj}}(\Omega) \;\; T_{\text{whole}}(\Omega) \;\; T_{\text{cut}}(\Omega) \;\; T_{\text{size}}(\Omega) \; \right]^\top \tag{8}$$

The objective function we minimize is a weighted sum of these terms:

$$\mathbf{T}(\Omega) \cdot \mathbf{w} \tag{9}$$

where $\mathbf{w}$ controls the relative importance of each term in the final objective function. The weights used to generate all our crops were $\mathbf{w} = [1, 1, 0.75, 0.3, 0.15]^\top$.

Performance and optimization

All times are reported for a 3GHz Pentium IV PC. Preprocessing to segment the photo takes about 10 seconds. The graph-cut resegmentation that identifies foreground features takes about 30 seconds in a mixed MATLAB and native implementation, dominated by MATLAB code that sets up the problem. Graph-cut optimization is performed using code from Andrew Goldberg's Network Optimization library.[2] Summed area tables [7] are used to speed up the calculation of $T_{\text{subj}}(\Omega)$.

The space of potential crops for an image is four dimensional, consisting of the location, width and height of the crop rectangle. One could search over this whole space for an arbitrary crop, or specify a target aspect ratio or size. For our results we fixed the aspect ratio, which leaves a three dimensional search space. At a granularity of 25 pixels, searching this space takes about 30 seconds for a 1280×960 image in a MATLAB implementation. Finer grained searches did not produce significantly different results. A uniform native implementation and coarse-to-fine search (which evaluates the objective function at widely sampled points in the parameter space and then refines search only around minima) would run significantly faster.

RESULTS

Eye tracking procedure

Fifty images were viewed to collect eye tracking data. The images were selected by the authors as photos that could benefit from cropping. The photos included various images of people, street scenes, and still life compositions, though the majority of images were of people, as consumer snapshots usually are.

Images were broken into two groups, each of which were viewed by four different viewers. Two of these eight viewers were secondary authors. Naive viewers knew that their eye movements were going to be used to subsequently crop the image. However, they were not told how the algorithm worked in any detail. Viewers' eye movements were recorded for 10 seconds over each image with a Tobii x50

[2] available at http://www.avglab.com/andrew/soft.html

Figure 5. Fixations (marked as white circles) made (on the left) by a viewer following instructions to "identify important content in the photo," and (on the right) by another viewer instructed to insert and adjust the contrast of a photo. Fixated locations are similar.

dual eye tracker. The values for both eyes were averaged to produce a single location, and fixations were identified.

In an actual application, eye tracking data could be collected passively whenever a user views a photo. This viewing could occur during a variety of tasks, browsing images, adjusting image properties, selecting an image to illustrate some text, or choosing images that require cropping. It was impractical however for us to record eye movements under realistic conditions for the variety of tasks that might be of interest. Our approach to validation observed that many tasks share an element of implicit visual search for important content. In typical photos there is overlap in what is "important" between most tasks. We chose to evaluate our approach using this generic search task, recording eye movements of subjects told to "find the important subject matter in the photo."

Viewers were not told to *look* at the important content, merely to identify it. These instructions were important, however; they resulted in a more focused pattern of viewing than sometimes occurred in initial tests when viewers were instructed to simply "look at the image." In general, a task has a strong effect on eye movements, and without clear instructions, viewers tend to make up their own differing interpretations of the task. Even when clear instructions are given, brief fixations in unimportant outlying regions sometimes occur. These fixations could be filtered out, but our algorithm is sufficiently insensitive to them that we have not found filtering necessary.

Eye movements during actual tasks

Given that our eye tracking data was not collected during a real task, we present an informal experiment demonstrating that fixations made during one real task are similar to those found using the protocol described above. Our intuition is that most real-world tasks involving images implicitly require a viewer to identify important content.

Three naive subjects and one author followed instructions to insert an image into a word-processor document, and then adjust the color and contrast of the image until it appeared optimal. During this process, the subjects' eye movements were recorded. This procedure was repeated four times with different images. An informal analysis indicates that fixations made during this task were always located near those made by viewers in our initial data collection. An example of data collected in our initial protocol and during an insertion and adjustment task is shown in Figure 5. This experiment is not a full evaluation of gaze-based interaction for cropping under field conditions. However, it does suggest that real tasks involving images are functionally similar to our artificial task of identifying important content. We leave deeper investigation of this issue as future work.

Discussion

A representative gallery of crop results is presented in Figure 6. Crops at two aspect ratios are illustrated: the first is the aspect ratio of the original image, the second is the reciprocal of that ratio (which is generally more challenging). In some cases no clean crop of a particular ratio is possible, which is reflected in a poor score for the best crop.

Occasional mistakes in identifying foreground regions can result in awkward crops. The same problem is more severe in automatic approaches where the main subject can be entirely cut out if prominent content is not meaningful (see the sixth row of results where the automatic technique cuts off the girl's head). Though not perfect, our approach rarely fails dramatically because fixations provide a better measure of important features than salience methods.

Adaptive documents

The adaptive document layout (ADL) system [18] is designed to change the layout of a document in response to the size of the display window. In the original ADL system, text boxes could continuously adapt in aspect ratio, but images had to jump discretely between a few hand-selected aspect ratios.

By including our cropping system in the ADL pipeline, the combined system can select the appropriate crop for any aspect ratio and thereby allow adaptive documents the freedom to continuously adjust the aspect ratio of their images. We can crop images to different aspect ratios without explicit user intervention and switch between them as necessary. Moreover, a desired aspect ratio that does not fit the content can be identified by its poor objective function score, and the closest ratio with an acceptable score can be used instead. The objective term weights given above assume images that need cropping. In adapting well-composed pictures to different ratios we remove the size term from the objective function as it is not necessary to crop the image if it will fit in the target region. We tested our application's ability to crop well-composed images by cropping about 25 pictures to several standard aspect ratios (see Figure 7 for some examples).

EXPERIMENTAL VALIDATION

We validated our results by comparing viewers' assessment of the aesthetic appeal of gaze-based and salience-based crops. Four types of crops were included in our evaluation:

- Original: the uncropped photo.
- Salience: fully automatic crops [29].
- Gaze: crops generated via our system.
- Hand: crops made by a professional designer.

Figure 6. Results for a set of representative images. (a) Original image; (b) fully automatic crop [Suh et al. 2003]; (c) gaze-based content map; (d,e) gaze-based crops to horizontal and vertical aspect ratios.

Figure 7. Original well-composed images (left), adapted to two different aspect ratios. An ADL document (right) using our crops. If eye movements are collected passively during document construction, our approach allows adaptation of images to arbitrary aspect ratios with no explicit user effort.

All of these crops were constrained to have the same aspect ratio as the original photograph. The salience and gaze-based methods were also separately compared using crops made to the reciprocal of the original image's aspect ratio. We hypothesized that having an accurate representation of the content of an image would be particularly critical when making large changes in crop aspect ratio, and therefore our approach would perform relatively well in this situation.

Task and stimuli

We compared cropping techniques using a forced-choice paradigm. A subject is shown two crops, and decides which one "looks better." A forced-choice task is simple, and so results are more likely to be consistent (compared for example to assigning numerical ratings to isolated images).

Fifty images in need of cropping were used in these experiments. For each image, eye-tracking data from one of the viewers was randomly selected and used to generate the gaze-based crop; this same crop was used in all trials. All pairs of cropping techniques were compared to each other for each photo, 350 trials per subject. Each pair was displayed side by side on a 19-inch CRT. All images were displayed downsampled to 600x400 pixels. Order and position (left or right) were randomized. Subjects were told to pick the image they thought looked better even if both appeared similar. A total of 8 subjects completed this experiment, which took about 25 minutes on average. The majority of subjects were graduate students, about half had a background in graphics. One subject (a secondary author) was also an eye-tracked viewer; one other author also participated in the experiment. Author responses did not differ significantly from those of naive subjects.

Preferences and response times were collected for each subject. The overall mean response time was 3.9 seconds. Trials with response times exceeding 10 seconds were discarded as outliers. Rejected data represented about 10 percent of the trials. Rejecting long trials did not have any qualitative effect on statistical significances. It is worth noting that subjects in rejected trials tended to prefer the overall less popular condition.

	Original	Salience	Gaze	Hand
Original	–	.5109	.4393*	.2659**
Salience	.4891	–	.4160**	.3389**
Gaze	.5607*	**.5840****	–	.3250**
Hand	.7341**	.6611**	.6750**	–

Figure 8. Preference results for each condition pair across all viewers. Each entry gives the fraction of trials in which the condition in the row was preferred over the condition in the column. Entries marked with * differ significantly from chance at p <.05, those marked with ** are significant at p <.005. Other entries are not significantly different.

	Salience flipped	Gaze flipped
Salience flipped	–	.4063**
Gaze flipped	.5937**	–

Figure 9. Preferences for flipped aspect ratio.

Results

Preferences were analyzed with pairwise sign tests for each condition pair. Most condition pairs were significantly different; see Figures 8 and 9. Most importantly, our gaze-based approach is preferred to salience-based cropping in 58.4% of trials. Response times were broken down by condition pair and by which condition of the pair was preferred. This data was analyzed with a two-way ANOVA. There was a significant global effect of the condition pair (i.e., decisions were harder for some pairs, p <.05) and an interaction between condition pair and which element of the pair was preferred (within some pairs judging a condition to be bad was easier than judging it to be good, p <.001). Multi-comparisons revealed response times were faster when flipped-aspect-ratio gaze-based crops were preferred to corresponding salience-based crops and when hand crops were preferred to salience-based crops and originals (p <.05).

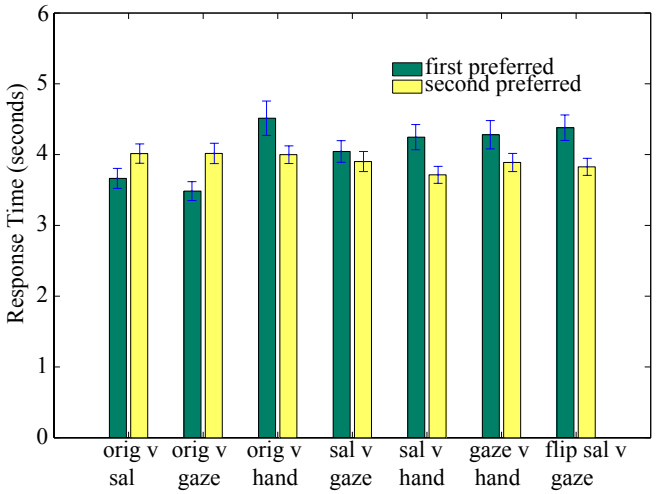


Figure 10. Response times for each pair of conditions.

Kendall analysis

An alternate analysis of forced choice data presented by Kendall [19] allows assessment of consistency internal to and between subjects (see David [8] for a detailed explanation). Analysis confirmed that subjects were internally consistent and agreed with each other ($p < .01$). This analysis also allows us to distill responses into a single measure of the quality of each condition compared to all others (essentially the number of occasions two subjects agreed about the superiority of the condition). We can test if differences in this value between two conditions are significant. Analysis shows that all conditions are significantly different at $p < .01$. Salience-based crops were better than uncropped images, gaze-based crops were better than salience-based, and hand crops were best of all.

Discussion

Preference results demonstrate that our gaze-based approach was significantly preferred to uncropped images and automatic techniques, though the quality of our crops was still short of handmade crops. The original photos in our experiment were picked because they appeared to need cropping, so we would expect any cropping to improve them. Fully automatic crops were, however, judged worse than originals, though the difference is not significant. Kendall analysis indicates fully automatic salience-based crops performed better than uncropped images overall.

Response times are indicative of the difficulty of a choice. Very obvious bad crops should be rejected with short response times; long decisions are likely made on the basis of subtler, harder-to-judge differences. For example, viewers can quickly notice when a crop removes a person's head.

In our tests (Figure 10), response times reinforce our subjective impression that salience-based crops often failed in dramatic ways. Response times are fast in the common situation where flipped gaze-based crops are judged superior to flipped salience-based crops, but they are longer in the rarer cases where gaze-based crops are judged to be worse (see Figure 10 column 7). Because flipping aspect ratio removes more of the image, it highlights mistakes in content identification. Response times suggest that automatic salience methods mistakenly crop out obviously important parts of the image at altered aspect ratios. In contrast, our eye tracking approach, when it fails (which preference data shows it does less frequently), does so in subtler ways that take longer to assess.

CONCLUSION AND FUTURE WORK

Experimental results and the crops themselves suggest that gaze-based cropping is successful. Our approach produces crops that are judged more attractive and that remove important content less often than automatic crops. Fixations provide sufficient information to robustly identify content in most images. Gaze-based crops should be useful for minimally interactive, high-volume cropping.

The success of our approach for identifying subject matter suggests it should also be useful in other implicit gaze-based interactions with photos where it is important to determine accurate regions of interest without explicit interaction like selective image enhancement, transmission, or editing. We hope that this work will spur interest in the HCI community regarding implicit interfaces for use with images.

Our work also serves as a jumping off point for better computational models of composition and further studies of the link between eye movements and composition. Given we know what the subject is, can eye movements tell us something about how good the composition is, and perhaps how to improve it? A reoccurring theme in qualitative art literature (in Graham [13] for example) is that cyclical or inward spiral patterns of "eye movement" (to our knowledge, these theories are based on intuition and never verified with actual eye-movement recordings) are good, while compositions that lead the viewer's eye to the edge of an image are displeasing and unbalanced. This hypothesized relationship could be assessed experimentally, and may yield a diagnostic measure of compositional quality specific to eye movements that could guide photo cropping.

We implemented centering and the rule of thirds, but our subjective assessment was that positioning features using these rules did not consistently improve crops. However, content placement is important. If Arnheim's thesis [2] is correct that a formally balanced composition is a purely perceptual phenomenon based on balancing the visual "weight" of objects, perceptual studies should allow us to model this and assess balance computationally. Ultimately, low-level or formal properties and high-level or subject matter issues need to be combined to create a fuller model of composition.

Fairly little psychological research has been conducted that would allow us to build such a complete computational model of what it means for an image to be well composed. However, further psychological investigation paired with computational modeling may allow for much more flex-

ible and complete definitions of a well-composed image, and accordingly better automatic cropping.

ACKNOWLEDGMENTS

Deep thanks to B. Suh, H. Ling, B. Bederson and D. Jacobs for access to their salience cropping system. Thanks also to Chuck Jacobs, Eileen Kowler, Mary Czerwinski, Ed Cutrell and Patrick Ledda. This research is partially supported by the NSF through grant HLC 0308121.

REFERENCES

1. Arnheim, R. *Art and Visual Perception*. University of California Press, 1974.
2. Arnheim, R. *The Power of the Center*. University of California Press, 1988.
3. Banerjee, S. *Composition Guided Image Aquisition*. PhD thesis, University of Texas at Austin, 2004.
4. Byers, Z., Dixon, M., Smart, W. D., and Grimm, C. M. Say cheese!: Experiences with a robot photographer. *Proceedings of the Fifteenth Innovative Applications of Artificial Intelligence Conference (IAAI-03), Acapulco, Mexico*.
5. Chen, L., Xie, X., Fan, X., Ma, W., Shang, H., and Zhou, H. A visual attention mode for adapting images on small displays. *MSR-TR-2002-125 Microsoft Research, Redmond, WA* (2002).
6. Christoudias, C., Georgescu, B., and Meer, P. Synergism in low level vision. *Proceedings ICPR 2002*.
7. Crow, F. Summed-area tables for texture mapping. *Siggraph '84*. 207–212.
8. David, H. A. *The method of paired comparisons*. Charles Griffin and Company, London, 1969.
9. DeCarlo, D., and Santella, A. Stylization and abstraction of photographs. *Proceedings of ACM SIGGRAPH 2002*. 769–776.
10. Duchowski, A. Acuity-matching resolution degradation through wavelet coefficient scaling. *IEEE Trans. on Image Processing 9*, 8 (2000), 1437–1440.
11. Foley, J., van Dam, A., Feiner, S., and Hughes, J. *Computer Graphics: Principles and Practice, 2nd edition*. Addison Wesley, 1997.
12. Gooch, B., Reinhard, E., Moulding, C., and Shirley, P. Artistic composition for image creation. *Proceedings of the 12th Eurographics workshop on Rendering Technique*. 83–88.
13. Graham, D. *Composing Pictures*. Van Nostrand Reinhold, 1970.
14. Grill, T., and Scanlon, M. *Photographic Composition Guidelines for total image control through effective design*. AMPHOTO, 1988.
15. Itti, L., and Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40* (2000), 1489–1506.
16. 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* (1998), 1254–1259.
17. Jacob, R. J. Eye-movement-based human-computer interaction techniques: Toward non-command interfaces. 151–190.
18. Jacobs, C., Li, W., Schrier, E., Bargeron, D., and Salesin, D. Adaptive grid-based document layout. *ACM Trans. Graph. 22*, 3 (2003), 838–847.
19. Kendall, M. G. On the method of paired comparisons. *Biometrika 31* (1940), 324 –345.
20. Li, S. Z., Zhu, L., Zhang, Z. Q., Blake, A., Zhang, H. J., and Shum, H. Statistical learning of multi-view face detection. *Proceedings 7th European Conference on Computer Vision(ECCV 2002) 4* (2002), 67 –81.
21. Li, Y., Sun, J., Tang, C.-K., and Shum, H.-Y. Lazy snapping. *ACM Trans. Graph. 23*, 3 (2004), 303–308.
22. Locher, P. J. The contribution of eye-movement research to an understanding of the nature of pictorial balance perception: a review of the literature. *Empirical Studies of the Arts 14*, 2 (1996), 146–163.
23. Locher, P. J., Stappers, P. J., and Overbeeke, K. The role of balance as an organizing design principle underlying adults' compositional strategies for creating visual displays. *Acta Psychologica 99* (1998), 141–161.
24. Locher, P. J., Stappers, P. J., and Overbeeke, K. An empirical evaluation of the visual rightness theory of pictorial composition. *Acta Psychologica 103* (1999), 261–280.
25. Lok, S., Feiner, S., and Ngai, G. Evaluation of visual balance for automated layout. *Proceedings of the 9th international conference on Intelligent user interface*. 101–106.
26. McManus, I., Edmondson, D., and Rodgers, J. Balance in pictures. *British Journal of Psychology 76* (1985), 73–94.
27. Peterson, B. *Learning to see Creatively: How to compose great photographs*. AMPHOTO, 1988.
28. Setlur, V., Takagi, S., Raskar, R., Gleicher, M., and Gooch, B. Automatic image retargeting. *ACM SIGGRAPH 2004 Technical Sketch*.
29. Suh, B., Ling, H., Bederson, B. B., and Jacobs, D. W. Automatic thumbnail cropping and its effectivness. *ACM Conference on User Interface and Software Technolgy (UIST 2003)* (2003), 95–104.
30. Vertegaal, R. The gaze groupware system: Mediating joint attention in mutiparty communication and collaboration. *Proceedings CHI '99*. 294–301.
31. Vertegaal, R. Designing attentive interfaces. *Proceedings of the Eye Tracking Research and Applications (ETRA) Symposium 2002*. 23–30.