Automatic Image Cropping Algorithm Based on Inference of The Photographer’s Intention

Paween Khoenkaw and Punpiti Piamsa-nga
Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Jatujak, Bangkok 10900 Thailand, g4885027@ku.ac.th, pp@ku.ac.th

Abstract

Photo cropping is a method to cut area of interest from a photograph. Mostly, it is done automatically by using salient-based techniques to mark cropped areas which are about detecting attractive contents to the audience. Instead in this paper, we propose an automatic photograph cropping algorithm based on the intention of the photographer. Usually, focal position and human faces are common evidences to indicate the photographer’s location of interest. Therefore, autofocus detection and face detection algorithms are proposed to determine importance map of photographs. In order to select a cropped area, we used a genetic algorithm to determine a location in the importance map that has maximum information per selected area. Experiments were performed to evaluate user satisfaction with cropped images, by allowing users to compare results of several cropping algorithms used by Google’s Picasa, applied to a collection of high quality photo images. The results indicate that users prefer results from this algorithm to all algorithms of Picasa, especially the issues that are related to photo composition.

Keywords: Photo cropping, Importance map, Genetic algorithm, Face detection, Focal point detection

1. Introduction

To display high-resolution images on a small device, the image must be scaled down to match the device’s aspect ratio and resolution; as a result, important contents are sometimes too small to see. Image cropping is an alternative for selecting important visual information to show on an arbitrary aspect-ratio display. Photo cropping is widely used in photo album software, such as Google Picasa 3 [1] and Cropp.me [2]; it can suggest choices of cropped results by different algorithms to the users in order to ask them for their favors on a specific display or printer. However, there are many devices whose display areas have varieties of sizes, resolutions, and aspect ratios. A single version of cropped photographs cannot be optimally appropriate for every type of output device. Multiple versions of manual cropping on a single photograph is also a very tedious task since the number of photographs owned by a person is too high to process each for many types of displays. Rather than just being a recommender for cropping, photograph cropping algorithm is required to be automatic. In this research, we aim to determine the best cropped version of image by using an automatic method.

The methodology of automatic cropping algorithms is about 1) to find is a quantitative construct used to identify the most important area, called “importance map” [3] and 2) to determine the best cropped area from the importance map. Importance map can be generated by two approaches: psychological model and object recognition model.

The psychological model is about to find the most attractive part of image by mimicking psychology principles on how human eyes interact onto photographs [4]. Most of the visual perception area in human eyes are low resolution, only two degrees of visual angle in human is perception in high resolution by small area called “fovea centralis”. In order to get high resolution, human brain must move fovea across the whole image to get a snapshot of the priority part of image and the brain then fill the rest. This model was developed based on the human’s attentions at his first sight of the image using eye tracker to mimic this phenomenon. The computable version of this model is generally called "visual attention model" [4]. Low-level features such as intensity, edge, color and orientation are extracted, weighted and merged into a single map, called “saliency map”. Many automatic image cropping models [5-8] use this approach to create importance map. This approach is computationally intensive [9] and human vision is a complex process that we still do not know how it exactly works.
For object recognition model, this approach is to use a specific object recognition algorithm to directly locate the object of interest. High level features, such as human faces, text in image, geometry shape detection, were also proposed for importance map generation in cropping application [10]. The applications that utilize this model are mostly used in a specific domain, such as face thumbnail generator and surveillance application. Mostly, this model is based on a classifier that must be trained with object collection and it cannot operate well on other application domains.

Since the object recognition approach works best in only specific domain, many researchers proposed the solution to generalize them by combining it with a psychological model [6-8, 11]. However, it also inherits the complexities of psychological model and huge amount of data are needed to train the classifier.

Photograph contains information that should attract the audience. Previously, cropping algorithms are designed to detect then crop the most informative area. This method is appropriate for photographs taken by users who do not have knowledge about photography composition. However, for a skill photographer, a picture is a medium for delivering his/her ideas to the audience, which is sometimes not necessary to be fit with a general information detection algorithm. We have an assumption that cropping a picture by detecting intention of photographer is better than detecting a location that has high information of low-level image features. In this research, instead of creating a visual attention model from scratch by mimicking psychology of human perception, we shortcut this process by looking for a clue given by photographer.

Because most digital cameras have autofocus mechanism as a standard feature and human face detection is an option, focal point and human face become evidence about where the attentive locations defined by photographers are. For focal point detection, we investigated research on camera’s autofocus system and used the best filter to detect the location of focal points in a picture. Results of the filter are called “focus map”. For human face detection, any algorithms such as [12, 13], can be used to generate a “face map”.

Both focus and face maps are combined to generate an importance map, which indicates amount of information in each area in a picture. Genetic algorithms are proposed to find an optimal cropped area, which should have maximum information per area. We used pictures from National Geographic Photos of the day website [14] to create a dataset. We evaluated satisfaction by 340 users by comparing our results with commercial products. The results are better than all cropping algorithms in Picasa 3 by 19.6% with confidence of 0.05.

Our paper presents as follows. Related work is described in section 2. Section 3 has details of proposed automatic cropping algorithm. Experiments and their results are described in section 4. Results and discussions are in section 5. Finally, we have concluded in section 6.

2. Camera’s autofocus system

If a particular object in an image is in focus, objects at different distances from the lens are blurred (Figure 1) [15]. Autofocus algorithms use the fact that an in-focus object should have a sharper image than out-of-focus objects to deduce which object is of primarily interested.

![Figure 1. The same scene taken at different focal points.](image)

The objective of the automatic focus system is to adjust the lens position in order to maximize focus quality in the photographer's area of interest. An autofocus system is composed of four components: lens position control, image capture device, digital filter/transformer, and focus quality measurement. The focus quality can be determined by echo time of arrival [16], using laser [17] or image processing.
An automatic focus system is depicted in Figure 2. Light passing through the lens forms an image on the capturing device, such as a charge-coupled device (CCD). Then, features related to focus quality of the digitized image are extracted. Finally, the focus quality measurement uses these extracted features to assess the focus quality of the target objects and finally adjust the lens position. The process is repeated until the highest focus quality is obtained [18].

![Figure 2. A model of camera’s auto focusing system](image)

We are interested in image filters of the autofocus system because we can use them to detect the location of focal points. Digital autofocus algorithms use different types of filters, such as edge detection and gradient [19] [20] [21] [22] [23], discrete cosine transform [24] [25] [26], Laplacian mask [27], discrete fourier transform [28], wavelet [29], entropy [30], and Laplace filter[31, 32]. All these filters are designed to extract the high-frequency components with the assumption that the focus location is composed of high-frequency components. The study of image filter used in autofocus systems by [33] concluded that the Laplace filter is the most accurate.

3. Proposed algorithm

The concept of proposed algorithm is explained in following steps: 1) find the focus area and an available face area to generate a focus map and a face map, respectively; 2) combine both feature maps to create an importance map; and 3) determine an optimal cropped area based on a given aspect-ratio using a genetic algorithm. The process is shown in Figure 3. The importance map generator and optimizer are described next.

![Figure 3. Proposed automatic photo cropping model](image)

3.1. Importance map generator

The importance map generator is composed of two feature maps: focus map and face map. The face map is generated by “Viola and Jones” detector [34] that is implemented in OpenCV [35] to detect human faces. In this research, human face is an additional feature used to improve accuracy in images where human faces are detected. However, the main contribution of this work in importance map generator is the focus map generator since it is a clue given by photographer.
Focus map is generated based on the same principle of autofocus system by applying the same filter used in a camera’s autofocus system as the filter for finding focus location in an image. Laplace filter, one of the most accurate filters for autofocus camera system [33], is adopted into this research. Let \( I \) be an input image and \( w(s, t) \) be a Laplace kernel. In this research, we used \((s, t) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \).

A focus map \( I_f \) is defined by equation (1).

\[
I_f(x, y) = \sum_{a=-\infty}^{a=\infty} \sum_{b=-\infty}^{b=\infty} w(s, t)I(x + s, y + t)
\]  

(1)

Figure 4 depicts focus maps of images in Figure 1. The images are taken from the same location but with different focal points. By observation, highlights appeared on importance maps are focused area in original images.

If all objects are located at the hyperfocal distance (focus at infinity) then every part of image is in-focus, hence the focal map will not have any clues to indicate a potential cropping area. For example, in images of stars, every star is in-focus no matter how far a star is from the others. Therefore, we require an alternative feature for this case. Face is also an important feature for cell-phone photographers, since a cell-phone camera cannot usually create depth of field. Therefore, we created an importance map \( I_g \) as a linear combination of the focus and face maps, as should in equation (2) where \( I_f \) is the face map; and \( w_1 \) and \( w_2 \) are weights.

\[
I_g = w_1I_h + w_2I_f
\]  

(2)

### 3.2. Genetic algorithm of pan-and-scan method

In many cases, results from the two maps suggest totally different cropping areas; therefore, it is hard to make a decision. For example in Figure 5, the focus map \( I_h \) (Figure 5B) shows that a focus area of Figure 5A is mainly at the microphone and the face map (Figure 5C) indicates that the important area is on the girl's face. Cropped versions using only focus map and only face map are shown in Figure 5D and F, respectively. However, the suggested results should be Figure 5F, where the surrounding area outside the feature maps must be investigated.

Selecting a cropped area is about finding the area that has maximal information per area. There are two approaches to image cropping: “pan and scan” and “automatic zoom”. The pan-and-scan method crops an image to fit the aspect ratio of the display or printing devices, such as printing a widescreen image on A4 paper. Automatic zoom selects any area in the image in order to display that particular area very clearly. For example, not only to match aspect ratio of the cell phone display an image must be cropped, it also must be zoomed around the important object to see it easily on a small display.

Pan-and-scan method selects a cropping window of the required size from the source image that has maximal information. The best cropping location should produce the maximum sum of \( I_g \), in equation (3), where \( (x', y') \) are coordinates of one corner of a cropping window of size \( w' \) by \( h' \). This equation is solved using a genetic algorithm. An example of automatic zoom with face map is shown in Figure 5F. An example of pan-and-scan result without face map is shown in Figure 6.

\[
[x', y'] = \text{ArgMax}_{x', y'} \sum_{x=x'}^{x'+w'} \sum_{y=y'}^{y'+h'} I_g(x, y)
\]  

(3)
3.3. Automatic zoom with genetic algorithm

In principle, we can just add a zoom factor as a new parameter along with aspect ratio to the pan-and-scan algorithm; however, it is not efficient since this problem is NP-complete and requires a brute force method to find the exact solution. In this research, we proposed an automatic zoom algorithm by creating a wrapper module on top of the pan-and-scan algorithm. The wrapper takes a zoom rate \(d; 0 < d < 1\), a given aspect ratio \((\theta)\), a zoom limit \(\phi\) and threshold \(t\) on the amount of information allowed to be lost. The concept of the wrapper is to reduce the size of cropping window until we find the smallest area that can keep information more than the given threshold. Algorithm 1 shows pseudo code of the proposed automatic zoom algorithm. The algorithm starts by creating candidate target window which is \(d\) times smaller than the original size of image (line 5). Then the algorithm uses this window to crop the source image using proposed “pan and scan” method (line 12) to create the candidate result. After that, it will try to reduce size of cropping window further to find a new candidate. If the loss of a newer candidate is over the threshold, the current candidate becomes the final result and the process is ended.

Algorithm 1. Automatic zoom

Input: importance-map \((I_g)\), zoom rate \((d)\), information-loss threshold \((t)\), zoom limit \((\phi)\) and target aspect ratio \((\theta)\)

Output: Rectangular coordinate \([x, y, w, h]\)

Step 1: \(z = \frac{log(\theta^{-1})}{log(d)}\)

Step 2: \(x_0' = 1, y_0' = 1\)

Step 3: \(h' \_0 = h(y) \_0, w' \_0 = w(x) \_0\)

Step 4: FOR \(i=1\) to \(z\)

Step 5: IF \(h' \_0 \geq w' \_0\) THEN

Step 6: \(w' \_i = w' \_i-1 \times d\)

Step 7: \(h' \_i = h'(x) \_i \times \theta^{-1}\)

Step 8: ELSE

Step 9: \(h' \_i = h' \_i-1 \times d\)

Step 10: \(w' \_i = h' \_i \times \theta\)

Step 11: ENDIF

Step 12: \([x' \_i, y' \_i] = \text{ArgMin} \_x, y \_i \sum_{x=x_{i-1}}^{x_{i+1}} \sum_{y=y_{i-1}}^{y_{i+1}} I_g(x, y)\)

Step 13: \(I_{g_i} = \sum_{x=x_{i-1}}^{x_{i+1}} \sum_{y=y_{i-1}}^{y_{i+1}} I_g(x, y)\)

Step 14: IF \(i > 2\) AND \(\frac{I_{g_{i-1}}}{I_{g_i}} > t + 1\) THEN
An example of a result from Algorithm 1 is depicted in Figure 7. A green window in Figure 7A indicates result of automatic zoom algorithm. Figure 7B is importance map of Figure 7A. In Figure 7B, all candidates of cropping window are shown as red rectangles; white rectangle indicates the first candidate that loss information is more than allowance threshold; and the green rectangle is the final decision of cropping widow.

The genetic algorithm in the optimizer module (line 12) is the stochastic process that can create slightly different results in every time it performs. In every configuration of our experiment, we run 20 times to determine the average result.

Figure 7. An example result of automatic zoom algorithm: (A) the original image and cropping window determined by the proposed algorithm; and (B) the candidate results (red box) of each iteration, the first candidate that losses information over the threshold (white box), and final decision (green box).

4. Experiments and Results

Our proposed algorithm is evaluated by comparing satisfaction score. Cropped results from our proposed algorithm are compared with results from all cropping algorithms of Picasa 3.

4.1. Dataset and algorithm implementation

Dataset in the experiment is composed of 100 images randomly downloaded from the National Geographic’s “Photo of the day” website [14] whose reputation is linked to high quality photography. By observation, there are varieties of image shooting such as macro shot, close-up shot, medium shot, and long shot. Images in dataset have various sizes ranged from 372x745 to 553x732 pixels (portrait mode) and 470x325 to 1024x750 pixels (landscape mode).

There are two important parameters required by this algorithm: zoom rate ($d$) and information-loss threshold ($t$). By preliminary experiment, we recommend $d > 0.9$ and $t=0.04$; otherwise, recall rate of the system will drop rapidly. In the experiments, $d$ is set as 0.959 and $t$ is set as 0.04 ($t=0.04$). If application does not require large magnification ratio, or it needs only changing aspect ratio we recommend using a small value of $t$.

The face detection was implemented using OpenCV version 0.9.7.1. Importance map generation and automatic-zoom algorithm was implemented using Matlab. Weights of face map ($w_1$) and focus map ($w_2$) are set equally as 0.5. To maximize possible loss and prevent the effect of rotation in some algorithms, the value of target aspect ratio is set as 1:1 ($\theta=1$). The equation 3 was solved by genetic algorithm module in Matlab's optimization toolbox (optimtool). Both variables $x'$ and $y'$ are encoded as floating-point chromosomes and processed by uniform mutation function with mutation rate at 0.02 and scattered crossover function with crossover rate is set at 0.8. Population size was limited to 20. The algorithm stops when fitness change is less than $10^{-8}$. The initial populations are located around the top-left of the image ($x' \approx 0$ and $y' \approx 0$).
In Figure 8, the experimental results show that this algorithm can solve the optimization within 50 generations. That also shows the best fitness value and average fitness values of every population are rapidly converging on the stable state (fitness value is not changed). By complexity analysis, it is 70 times as fast as a brute-force method, when window size is determined by using iterative pan-and-scan method at different zoom factor.

![Figure 8](image)

**Figure 8.** Best and mean fitness scores (high is good) of each generation while solving one zoom iteration.

4.2. Cropping function of Picasa 3

Picasa 3 (version 3.9.0) [1] is Google's photograph management application. It has a function for semi-automatic image cropping. It has been done by proposing three cropped versions from three classified algorithms as shown in Figure 9. User must select one of three algorithms or crop the picture manually. We use all three cropped images generated by Picasa to compare with results from this algorithm. However, from preliminary experiment, we found that if the data loss in cropping process is less than 20%, differences among results from any methods are hard to be noticed. In order to make results easy to compare, we forced every algorithm to drop large amounts of data by using 1:1 as target aspect ratio. We also can ignore a situation that any algorithm may avoid data loss by rotate aspect ratio.

![Figure 9](image)

**Figure 9.** The interface of Google Picasa. 3 cropped versions (A, B, and C) and original image (D)
4.3. Qualitative evaluation method

To evaluate performance of each algorithm, we develop a questionnaire website to collect opinions. The questionnaire was designed using a four-level Likert scale (Reject, Poor, Good, and Excellent). All images in the dataset were used to evaluate qualitative results of our proposed algorithm by comparing with result of three Picasa algorithms. An example of questionnaire is shown in Figure 10.

Figure 10. User interface of questionnaire website for qualitative evaluations

4.4. Experimental results

Qualitative evaluations are performed by 340 volunteers, which are a group of photographers and general Internet users. 80% of them are male, and other 20% are female. 59% have knowledge in photographic composition. Ages are between 10 and 60 years old and average age is 23.8 years old. Each volunteer is asked to evaluate cropped results from 20 original images. Totally, 3780 cases are evaluated. Results from a four-level Likert scale are converted to percentage and shown in Table 1 and they show that the user group preferred this algorithm 30% higher than Picasa 1, 23% higher than Picasa 2, and 6% higher than Picasa 3. The two-sample t-test confirmed that the mean differences of all results were significant ($p<0.05$).

Table 1. Evaluation results

<table>
<thead>
<tr>
<th></th>
<th>This algorithm</th>
<th>Picasa 1</th>
<th>Picasa 2</th>
<th>Picasa 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>27.38%</td>
<td>7.83%</td>
<td>11.66%</td>
<td>22.27%</td>
</tr>
<tr>
<td>Good</td>
<td>45.18%</td>
<td>23.43%</td>
<td>28.80%</td>
<td>42.30%</td>
</tr>
<tr>
<td>Poor</td>
<td>21.40%</td>
<td>32.67%</td>
<td>32.75%</td>
<td>24.39%</td>
</tr>
<tr>
<td>Reject</td>
<td>6.03%</td>
<td>36.05%</td>
<td>26.77%</td>
<td>11.03%</td>
</tr>
<tr>
<td>Average</td>
<td>64.67%</td>
<td>34.33%</td>
<td>41.33%</td>
<td>58.33%</td>
</tr>
<tr>
<td>Std.dev</td>
<td>28.39</td>
<td>31.75</td>
<td>32.63</td>
<td>30.73</td>
</tr>
</tbody>
</table>

However, different algorithms produced results with different losses of information since each algorithm can be configured to drop smaller amount of data in order to gain more user satisfaction. To make the comparison as fair as possible, we investigate the effect of image loss on each algorithm by analyzing the result of each method. The distribution of data loss of each algorithm is shown in Figure 11. User satisfactions of each method of all pictures and percentage of data loss are plotted on scatter chart in Figure 12. We found that Picasa 1 and Picasa 2 generated results by allowing huge loss of data; therefore, the cropped image is relatively small. Both Picasa 1 and Picasa 2 methods allow 60% loss of
data on average. User satisfactions of those two algorithms are 41.6% and 34.3% (in between poor and reject). Picasa3, the best method among Picasa’s cropping functions, generated results with lower loss allowance at 32% on average, where it is ranged from 24% to 57% and get the satisfaction score 58.6% (in between good and excellent). This algorithm generated results which the average loss is about 39% and ranged from 17% to almost 97%. Therefore, we can conclude that our algorithm dropped larger amount of image data and it still gained 6% more satisfaction than Picasa3 (and all other Picasa’s).

Figure 11. Data loss distribution of four cropping algorithms

Figure 12. Relationship between user satisfaction scores (0=Reject; 1=Poor; 2=Good; 3=Excellent) and percent data loss of four cropping algorithms

We also investigated whether the demography of the evaluators had an effect on the results. A one-way within subjects ANOVA was conducted to compare the effect of age and gender on this algorithm, Picasa 1, Picasa 2 and Picasa 3, the analysis result shown that there was not a significant effect of age and gender on every result ($p > 0.05$). However, evaluators with and without background in photographic composition had different opinions on the results. The t-test results indicate that evaluators without photographic backgrounds cannot distinguish differences between results from Picasa3 and this algorithm. The results show that ours is 3.3% better than Picasa 3 but t-test concluded that has no significant difference ($p > 0.5$). However, evaluators who have background in photography...
strongly ranked this algorithm 7.3% higher than Picasa3 ($p<0.5$). Details of evaluation scores are shown in Table 2. We also conclude that, photographers prefer our proposed algorithm and general users cannot decide the best between this algorithm and Picasa 3; however, this algorithm allows higher loss of information.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Satisfaction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>This algorithm</td>
<td>Picasa3</td>
</tr>
<tr>
<td>Evaluators have background in photography</td>
<td>63.6</td>
</tr>
<tr>
<td>Evaluators do not have backgrounds in photography</td>
<td>67.3</td>
</tr>
<tr>
<td>All</td>
<td>64.6</td>
</tr>
</tbody>
</table>

5. Discussion

By observation, we found that this algorithm did not well perform in two situations: low-contrast image and noisy image. When the image has low contrast, such as a foggy scene or motion blur, it affects the sharpness of the image. Both focus and face maps are unclear and the suggested cropping area tended to be too large; therefore, it cannot allow high amounts of loss. For noisy image, general face detection algorithms try to find all possible faces and it is sometimes too sensitive to detect false-positive objects. We are investigating how to compensate weights based on sensitivity of importance map generator.

6. Conclusions

We proposed an automatic image cropping algorithm based on inference of the photographer’s intention. The focal-point and face location are used as clues about the location in the picture that is the most interesting to the photographer. These features are used to generate an importance map. An automatic algorithm is designed based on the pan-and-scan method where a genetic algorithm selects the location that has maximal information per unit area from the importance map. To zoom in on the target area, a wrapper for the pan-and-scan-based algorithm is applied with successively smaller cropping size until a data-loss threshold is reached. Experiments were performed to compare user satisfaction with the cropped images proposed by this algorithm as compared with cropped images from three cropping algorithms used in Google’s commercial products, using images from the National Geographic website. The result shows that this algorithm is 30.3%, 23.0% and 6.0% better than Picasa1, Picasa2, and Picasa3 algorithms, respectively. Users without photography backgrounds prefer this algorithm to Picasa1 and Picasa2 and equally satisfied with cropping results from algorithm and Picasa 3, even though this algorithm allows 7% higher data loss. On the other hand, users with background in photograph prefer images cropped by this algorithm to images from any of Picasa’s at the same level of data loss.

We are currently extending the algorithm for video cropping. For instance, we analyze camera motion to identify the important part of a frame and optimize the algorithm for live feed applications. The preliminary results are encouraging.

7. Acknowledgement

This research is partially supported by Kasetsart University Research and Development Institute. Authors would like to thank Dr. James E. Brucker of Kasetsart University for his suggestion on writing and his assistance on proof reading.

8. References


