Automatic Thumbnail Generation Based on Visual Representativeness and Foreground Recognizability

Jingwei Huang[1,2,⋆], Huarong Chen[1,2,⋆], Bin Wang[1,2], Stephen Lin[3]
[1]School of Software, Tsinghua University [2]Tsinghua National Laboratory for Information Science and Technology [3]Microsoft Research
⋆Huarong Chen and Jingwei Huang are joint first authors. This work was done while they were visiting students at Microsoft Research.

Abstract

We present an automatic thumbnail generation technique based on two essential considerations: how well they visually represent the original photograph, and how well the foreground can be recognized after the cropping and downsizing steps of thumbnailing. These factors, while important for the image indexing purpose of thumbnails, have largely been ignored in previous methods, which instead are designed to highlight salient content while disregarding the effects of downsizing. We propose a set of image features for modeling these two considerations of thumbnails, and learn how to balance their relative effects on thumbnail generation through training on image pairs composed of photographs and their corresponding thumbnails created by an expert photographer. Experiments show the effectiveness of this approach on a variety of images, as well as its advantages over related techniques.

1. Introduction

For efficient browsing of photo collections, a set of images is typically presented as an array of thumbnails, which are reduced-size versions of the photographs. The reduction in size is usually quite significant to allow for many thumbnails to be viewed at a time, and the thumbnails are generally fixed to a uniform aspect ratio and size to facilitate orderly arrangement. Thumbnail creation involves a combination of cropping and rescaling of the original image as illustrated in Figure 1. Manually producing thumbnails for large image collections can be both time-consuming and tedious, as care is needed to ensure that each thumbnail provides an effective visual representation of the original photo. The practical significance of this problem has led to much research on automatic thumbnail generation.

Previous work has focused primarily on the cropping step of thumbnail generation. Many of them operate by extracting a rectangular region that contains the most visually salient part of a photograph. These saliency-based methods [3, 9, 23, 27, 31, 33] are effective at highlighting foreground content. Other methods based on aesthetic quality [25, 38] instead seek a crop that is visually pleasing according to compositional assessment metrics. It has been shown that aesthetics-based approaches often produce cropping results that are preferred by users over saliency-based crops [38].

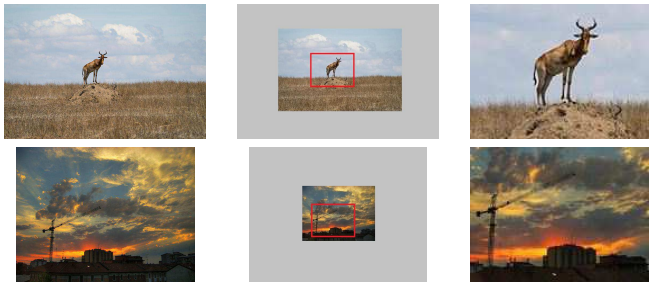


Figure 1. Image thumbnail generation. (a) Original images (viewed at low resolution). (b) Cropping (red box) and rescaling to produce thumbnails. (c) Thumbnails viewed at actual resolution.

Although these methods produce excellent results for image cropping, they share critical shortcomings for the task of thumbnail generation. One is that they do not consider how well the resulting image represents the original. Unlike a general image crop, a thumbnail serves a specific purpose as an index that should provide the viewer an accurate impression of what the original photo looks like. If the thumbnails of a vacation photo album exclude most of the background, different photographs would be more difficult to distinguish from each other based on their thumbnails. Another shortcoming is that previous methods do not account for the effects of rescaling. The utility of a thumbnail can be heavily affected by the amount of rescaling, since important subjects in an image may become difficult to recognize after too much reduction in size. A proper balance of cropping and rescaling is essential for decreasing image size in an effective way.

In this paper, we propose an image thumbnail method

(a) (b) (c)

Figure 2. Thumbnail considerations. (a) Original images (viewed at low resolution). (b) Low-quality thumbnails. (c) Our thumbnails. The first row illustrates our first consideration, that a thumbnail should give an accurate visual representation of the original image. Cutting out the mountains and sky in (b) results in a thumbnail that does not give a true impression of what the original image looks like. The second row illustrates the second consideration, that the foreground should be recognizable. Very little cropping and much rescaling in (b) leads to a thumbnail in which it is hard to identify the flowers in the foreground.

that is guided by two essential considerations on the utility of thumbnails as an image index. The first is the *visual representativeness* of the thumbnail with respect to the original image. A more visually representative thumbnail should better reflect the appearance of the actual photograph, thus providing a more effective index. We model this with various appearance features that have been used for comparing images. The second consideration is *foreground recognizability* in the thumbnail. The usefulness of a thumbnail diminishes as it becomes more difficult for the viewer to recognize the foreground subject after cropping and rescaling, as exemplified in Figure 2. To model this effect, we adapt image features commonly used for content-based image retrieval (CBIR) [29] and object recognition [19], as they serve a similar purpose in identifying and distinguishing elements.

These two factors are designed to balance each other. If only *visual representativeness* is considered, then there would be no cropping at all, since any cropping would reduce representativeness. On the other hand, considering only *foreground recognizability* would result in a tight crop around the foreground object. Neither of these factors would be appropriate to use by itself. However, they are effective when employed together, since the competing aims of the two terms can be balanced.

The relative influence of features used to model the two factors is learned through training on a set of image pairs, consisting of original photos and thumbnails created from them by an expert photographer. By accounting for the two factors, our technique produces thumbnails that are preferable to those of related methods according to quantitative comparisons and user studies.

2. Related Work

To display a photograph in limited space, prior works typically highlight the image areas of greatest saliency [14] while removing parts of the photo that would command less attention. In [35], a group of pictures is arranged into a collage of overlapping images, with the overlaps used to occlude regions of low saliency. Another way to remove less salient image content is through image retargeting [4, 26, 36, 28], which downsizes images through operations such as seam carving, local image warping, and rearranging spatial structure. Such operations, however, can introduce artifacts and image distortions that significantly reduce the appeal of results.

Image distortions can be avoided by restricting image manipulations to only cropping and rescaling, the two standard operations in thumbnail generation. For cropping, most algorithms are also driven by saliency, computed through a visual attention model [14], density of interest points [2], gaze distributions [27], correlations to images with manually annotated saliency [23], or scale and object aware saliency [33]. Based on a saliency map, these methods compute a crop window that encloses regions of high saliency [3, 9, 23, 27, 31, 33]. Saliency-driven techniques are effective at preserving foreground content, but tend to discard much contextual background information that is needed for image indexing. The work in [16] proposes a learning based thumbnail cropping method that combines saliency features and a spatial prior, but does not preserve visual representativeness well since the position and size of crops are analyzed statistically without considering image content. The recent work in [13] proposes the concept of context-aware saliency, which may assign high saliency values to background areas surrounding the foreground. Incorporating context-aware saliency into these cropping works would address the visual representativeness issue only to some degree, and it would not deal with foreground recognizability at all.

Several methods utilize aesthetics metrics instead of saliency values to guide image cropping and/or rearrange objects in images [18]. Aesthetics metrics are designed to assess the visual quality of a photograph based on low-level [22] and/or high-level [15, 21] image composition features. Based on such metrics, classifiers have been used to evaluate the aesthetic quality of crops [25]. In [38], the relationship between images before and after cropping is also taken into account. As with the cropping methods based on saliency, these works based on aesthetics do not consider how well the result visually represents the original image or the effect of rescaling the cropping result to thumbnail size.

Methods that specifically aim to generate thumbnails or render photos on small displays largely treat rescaling as an afterthought or do not explicitly discuss the rescaling step [23]. In [9, 33], crops are computed without rescal-

ing in mind, and the crop result is simply rescaled to the target size. By contrast, our work seeks to balance cropping and rescaling in a manner that preserves visual representativeness and foreground recognizability.

3. Approach

In this section, we present our thumbnail approach based on our two major considerations. Details on training set selection, the extracted image features, the training procedure, and thumbnail generation are described. Algorithmic overviews of the training and thumbnail generation methods are provided in Algorithm 1 and Algorithm 2, respectively. In both algorithms, we extract various features based on image or region properties. The features are then employed within a support vector machine (SVM) used for evaluating thumbnails within a thumbnail generation procedure.

Algorithm 1: Training(Images, Crops)

```
for i = 1 to Images.size do
    Im ← Images(i)
    GoodCrop ← Crops(i)
    Sa ← DetectSaliency(Im)
    Fg ← ExtractForeground(Im,Sa)
    Segs ← SegmentImage(Im)
    CropSet ← SampleCrops(Im.size, GoodCrop)
    for j = 1 to CropSet.size do
        Tn ← Tn +1
        TF(Tn).x =
        CalcThumbFeature(CropSet(j),Segs,Fg,Sa)
        TF(Tn).y = (CropSet(j) == GoodCrop)
    end
end
ThumbModel ← SVM_Train(TF)
return ThumbModel
```

Algorithm 2: Thumbnail Generation(Im)

```
Sa ← DetectSaliency(Im)
Fg ← ExtractForeground(Im,Sa)
Segs ← SegmentImage(Im)
CropSet ← SampleCrops(Im.size, Im.BoundingBox)
for j = 1 to CropSet.size do
    TF ← CalcThumbFeature(CropSet(j),Segs,Fg,Sa)
    ThumbScore(j) ← SVM_Predict(ThumbModel,TF)
end
Find index j with maximum ThumbScore(j)
return CropSet(j)
```

3.1. Training set

We build the training set using 600 photos selected from the MIRFLICKR-25000 dataset [1]. The photos span a diverse range of categories including landscape, sunset, night, painting, architecture, plant, animal, man-made objects and other complex scenes. The photos also vary in texture complexity, intensity distribution, and sharpness. Each image is manually cropped and scaled by an expert photographer into a thumbnail of size 160×120.

3.2. Image features

We utilize several image features to model the properties of expertly-created thumbnails in relation to their original images. These features are specifically selected to measure how well the thumbnail visually represents the original photo and how easily the foreground in the thumbnail can be recognized.

3.2.1 Visual representativeness

Our features for visual representativeness model in various respects how similar of a visual impression the thumbnail gives to the actual photograph. This notion of visual representativeness differs from that in works like bidirectional similarity [28] that measure the summarization quality of an output image rather than aiming to convey the actual undistorted appearance of the original image, which helps the user to identify a photo based on its thumbnail. Some of the features are computed with respect to foreground or salient regions, while others are derived from the image as a whole. These representational features are calculated between the cropped image and the original, as they intend to model the change in image content that results from the cropping of the thumbnail process.

Color Similarity The first feature reflects how representative the crop is in terms of color properties. To model this at a finer scale, we compute color similarity at the level of regions instead of globally over the image. If a crop has removed a region or enough of a region to alter its color properties, then the crop is less representative of the original image. We describe the color properties of a region Ω by the three central color moments $v_c(\Omega)$ of its RGB distribution [32]. The color similarity between a region Ω_a in the crop and its corresponding region Ω_b in the original image is then expressed as the normalized inner product between their color moment vectors:

$$f_{cs}(\Omega_a, \Omega_b) = \frac{v_c(\Omega_a) \cdot v_c(\Omega_b)}{||v_c(\Omega_a)|| \cdot ||v_c(\Omega_b)||}. \quad (1)$$

We aggregate this value over all the regions in the original image, which is segmented using the graph-based technique in [12]. The color similarity feature for a crop is thus calculated as

$$E_{cs}(C) = \frac{1}{\sum_{i=1}^{n} S_i} \sum_{i=1}^{n} [S_i * f_{cs}(C \cap \Omega_i, \Omega_i)] \quad (2)$$

where C denotes the area within the crop, and S_i is a weight computed as the proportion of saliency [7] in region Ω_i with respect to the whole image. The saliency weight puts greater emphasis on salient regions, whose color properties are more critical to preserve. Note that if a region is removed completely by the crop, then f_{cs} is equal to zero. A higher value of E_{cs} indicates greater color similarity.

Texture Similarity In addition to color, the similarity of texture between a crop and the original image is also included as a representational feature. We calculate a texture vector $v_t(\Omega)$ of each region using the HOG descriptor [10], and compute the texture similarity between a region Ω_a in the crop and its corresponding region Ω_b in the original image as

$$f_{ts}(\Omega_a, \Omega_b) = \frac{v_t(\Omega_a) \cdot v_t(\Omega_b)}{||v_t(\Omega_a)|| \cdot ||v_t(\Omega_b)||}. \quad (3)$$

This quantity is aggregated over all the regions in the same manner as for color similarity in Equation 2 to yield the texture similarity feature E_{ts}.

Saliency Ratio A thumbnail is more representative if it contains more of the salient content of the original photo. We model this feature by taking the ratio of summed saliency within the cropping window to the summed saliency over the whole photograph.

Edge Ratio Edges are an important low-level shape representation of images [24], so we additionally account for edge preservation in the cropped image. We detect edges in the original photo using the Canny edge detector [5], and formulate an edge ratio feature as the number of edge pixels within the crop box divided by the total number of edge pixels over the entire photograph.

Contrast Ratios The general visual impression of a photo depends greatly on how much its appearance features vary. The contrast in these appearance properties additionally affects visual elements such as how much the foreground stands out in an image. To measure how closely the cropped image adheres to the contrasts of the original photo, we compute the standard deviation of saliency, intensity, and edge strength [24] in the crop and the original image, and then take their ratios. Edge strength is computed perpendicularly to edges detected with the Canny edge detector [5]. An example of these contrast ratios is shown in Figure 3, where a more visually representative thumbnail has standard deviations of saliency, intensity and edge strength that are closer to those of the original image.

Foreground Shift Another factor that influences the perception of an image is the position of the foreground, which is a major consideration in photographic composition as seen from the common application of the rule of thirds. A significant shift in foreground position between the crop and photograph may weaken the thumbnail's visual representation quality, so we record this feature as the distance between their foreground centers after mapping the photo and crop to a [0,1] $\times$ [0,1] square. The foreground is extracted using the method of [7] incorporated with a human face detector [37].

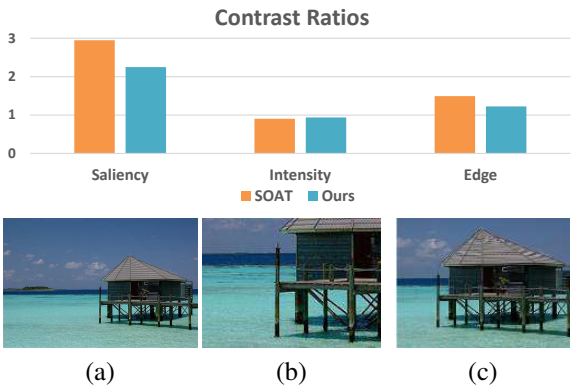


Figure 3. Contrast ratio comparison. (a) Original Image. (b) Thumbnail generated by SOAT, a state-of-the-art saliency-based cropping method [33]. (c) Our thumbnail. From the bar chart, it can be seen that the thumbnail in (c) has contrast ratios closer to one, indicating that its contrast properties are more similar to those of the original image.

3.2.2 Foreground recognizability

The thumbnail with the greatest visual representativeness is the one generated without cropping the photograph at all. However, an uncropped image would require a maximum amount of rescaling to reach thumbnail size, which may lead to foreground regions becoming less recognizable in the thumbnail. To account for this issue, we incorporate features that reflect how easily the foreground in a thumbnail can be recognized.

To model foreground recognizability in thumbnails with respect to the original image, we take advantage of features used in content-based image retrieval (CBIR) [29] and object recognition [19], which aim to identify images or objects similar to a given target. In our case, the target is the foreground in the photograph, and we model how well it can be recognized in the thumbnail based on its similarity in terms of CBIR and recognition features. Since these features are particularly intended to measure the effect of rescaling on foreground recognizability, feature comparisons are done between the foreground in the thumbnail and the foreground in the cropped image.

In CBIR, images are abstracted into feature vectors containing descriptors for color, texture, shape and/or high-level semantics [11]. The similarity between images is then determined according to distances between their feature vectors. Since a thumbnail is directly scaled from the original crop, its color properties remain the same. Here, we assess recognizability via shape and texture features, to-

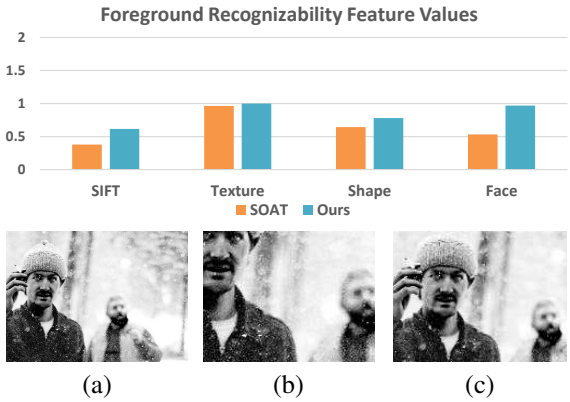
gether with features for object recognition and a measure derived from human face detection.

Shape Preservation Ratio A shape representation commonly employed in CBIR is edge information packed into a vector, such as a polar raster edge sampling signature [24]. We utilize the Canny edge detector [5] to detect edges in both the cropped image and the thumbnail. Instead of retrieval from a large dataset, our concern is on how much the shape features in the original image are retained in the thumbnail. So rather than packing the edges, we compute what proportion of edge pixels in the cropped image are also detected as edges at the corresponding pixels in the thumbnail. This ratio of preserved edges is used as a shape preservation feature.

Directional Texture Similarity In CBIR, texture is often represented in terms of six properties: coarseness, contrast, directionality, linelikeness, regularity, and roughness [34, 6]. Among the first three properties, coarseness and contrast are not closely related to recognizability of a rescaled object. Moreover, linelikeness, regularity and roughness are highly correlated with the former three properties. The remaining property, texture directionality [17], may change after rescaling a crop into a thumbnail. So the similarity of this property between the cropped image and the thumbnail is used as a recognizability feature. Texture directionality is determined by gradients computed after filtering the foreground region with the Sobel operator [30]. The gradients are then expressed as a vector after quantization into six bins of 30° width from 0° to 180°. We measure similarity as the normalized dot product of the two vectors, similar to Equations 1 and 3.

SIFT Descriptor Similarity SIFT descriptors [20] are a popular feature for object recognition. A standard use of SIFT descriptors for recognition is to first extract SIFT points and their corresponding descriptors from an object and a reference, then match pairs of SIFT points between them based on minimum descriptor distance. The object is recognized as the reference if most pairs of SIFT points are consistent with respect to a transformation model [19]. In our case, the transformation model is a known change in scale. Based on this, we measure ease of recognition based on similarity of SIFT descriptors for corresponding SIFT points with respect to the transformation. If a SIFT point computed in the cropped image does not have a corresponding SIFT point computed in the thumbnail, it is failed to be recognized. Otherwise, the similarity is measured by the normalized inner product of the corresponding SIFT descriptors, each a 128-dimensional vector. The similarity for the foreground regions in a crop and thumbnail is computed by aggregating the similarity of SIFT descriptors weighted by SIFT point saliency:

$$E_s(a,b) = \frac{\sum_{q \in SIFT(b)} s_q \cdot M(q, C_{a,b}(q))}{\sum_{q \in SIFT(b)} s_q} \qquad (4)$$



(a) (b) (c)

Figure 4. Foreground recognizability comparison. (a) Original image (displayed at low resolution). (b) Result from SOAT. (c) Our thumbnail. SIFT refers to SIFT Feature Similarity. Texture indicates Directional Texture Similarity. Shape refers to Shape Preservation Ratio, and Face indicates Face Preservation Ratio. The values of foreground recognizability features decrease with greater rescaling.

where $E_s(a,b)$ denotes SIFT descriptor similarity between cropped image b and the thumbnail a rescaled from it. s_q is the saliency value of pixel q. $SIFT(\Omega)$ represents the set of SIFT points detected in the domain Ω. $C_{a,b}(q)$ is the point in $SIFT(a)$ which has the minimum coordinate distance from the corresponding pixel of q in a. If the minimal distance is larger than a certain threshold (5 in our implementation), the corresponding SIFT point is considered not to be found after the rescaling, in which case $M(p, C_{a,b}(q))$ is set to zero. Otherwise, $M(p,q)$ is set to the normalized product of p and q's SIFT descriptors.

Face Preservation Ratio Faces are often the most important component in an image and deserve special treatment. One way to handle faces in the foreground region is to determine whether they are recognized as having the same identity after rescaling to the thumbnail. We instead use a simpler measure based on confidence values from a face detector [37]. The sums of confidence values are computed for the faces detected separately in the thumbnail and in the original crop, and then their ratio is taken as the face preservation feature. If there is no face detected in the original crop, the ratio is set to one. As detector confidence decreases with greater thumbnail rescaling, the value of this feature is reduced as well.

Area Ratio Finally, we include a feature that represents the degree of rescaling as the ratio of area in the thumbnail and the cropped image.

Figure 4 illustrates the effect of rescaling on our foreground recognizability features. Greater rescaling generally leads to both less recognizability and lower feature values.

3.3. Training and Thumbnail Generation

To balance the various features for thumbnail generation, we learn an SVM model from positive and negative thumbnail examples for the photographs in our training set (Section 3.1). The SVM model that we employ is a kernel SVM with radial basis functions, which is capable of learning the influence of all the proposed features. For each photo, we consider the thumbnail created by the expert photographer as a positive sample, and generate negative examples for it by sampling crop coordinates that are different from it. The negative examples are generated by first sampling at 30-pixel intervals the coordinates of the upper-left crop corner and the x-coordinate of the lower-right crop corner. The y-coordinate of the lower right crop corner is then determined according to the thumbnail aspect ratio (4:3 in our work). Among these samples, we keep only those that are different enough from the positive sample according to

$$C = \{(x_1, y_1, x_2, y_2) | \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{||t-t_g||^2}{2\sigma^2}} < \tau\} \quad (5)$$

as done in [38]. Here, $t = (x_1, y_1, x_2, y_2)^T$ and $t_g = (x_1^g, y_1^g, x_2^g, y_2^g)^T$ are the cropping coordinates of the negative and positive examples, with the first two coordinates for the upper-left corner, and the last two for the lower-right corner. The threshold τ controls the minimum degree of offset, and σ is a Gaussian parameter. After cropping, the negative sample is rescaled to the targeted thumbnail size.

After SVM training, our method predicts a good thumbnail for a given image by first generating a set of candidates. The candidate set is produced by exhaustively sampling crop windows of the target aspect ratio at 10-pixel intervals for the upper-left corner and x-coordinate of the lower-right corner, then rescaling them to thumbnail size. The candidates are each evaluated by the SVM to obtain an energy. The thumbnail with maximum energy is taken as our result. With our unoptimized implementation, the computation time is about 60 seconds for an 800×600 image on a 3.4GHz Intel Core i7-2600 CPU.

4. Evaluation

For evaluation of our thumbnail generation method, we present some results on various scenes, report a cross-validation experiment using thumbnails from an expert photographer as ground truth, and compare to related techniques in a user study.

4.1. Results

Several examples of our thumbnail results are displayed together with the original images in Figure 5. Our method seeks a tradeoff between visual representativeness with respect to the original image and ease of foreground recognition. In some cases such as (a), (b) and (c), a significant amount of cropping is applied in order to facilitate recognition of the foreground. In other cases such as (d), relatively little cropping is employed since a result obtained mainly from rescaling is deemed to yield good representativeness and foreground recognizability. For many other images such as (e) and (f), a more intermediate mixture of cropping and rescaling is applied in providing a balance between the two thumbnail considerations, with the placement of crop windows determined in a manner that aims to preserve the visual impression of the original image.

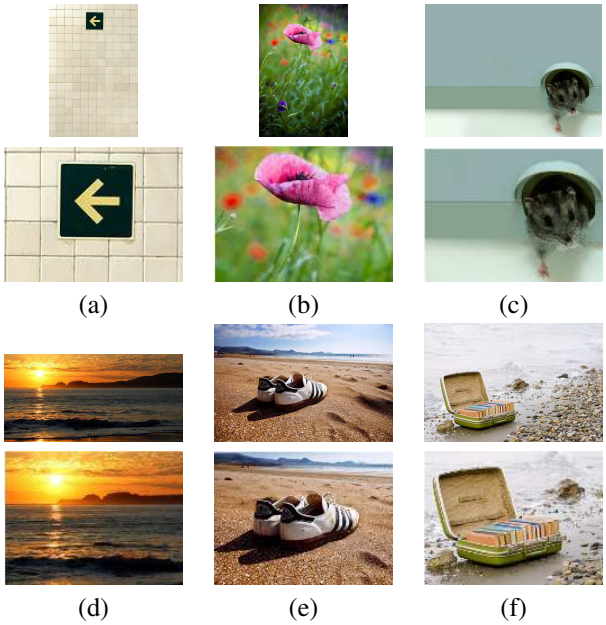


Figure 5. Image thumbnail results. For each example, the upper image is the original photograph displayed at low resolution, and the lower image is the thumbnail. Our method aims to strike a balance between how well the thumbnail visually represents the original photo and how easily the foreground can be recognized.

4.2. Features

Our method utilizes 13-dimensional feature vectors whose elements were described in 3.2. To examine whether each feature element contributes to the overall performance, we conducted experiments comparing performance with and without each individual element in the feature vector.

The tests were run using 200 images different from those used for SVM training. Thumbnails of these images were created by our expert photographer and taken as 200 positive examples with a label 1. Additionally, 6024 negative examples with a label 0 were generated using the method in Sec. 3.3. Over this set of test examples, T, we compute the following energy with and without each of the features:

$$E_F = \Sigma_{t \in T} |\hat{l}_t(F) - l_t| \quad (6)$$

where t is a thumbnail example in the test data T, $\hat{l}_t(F)$ is

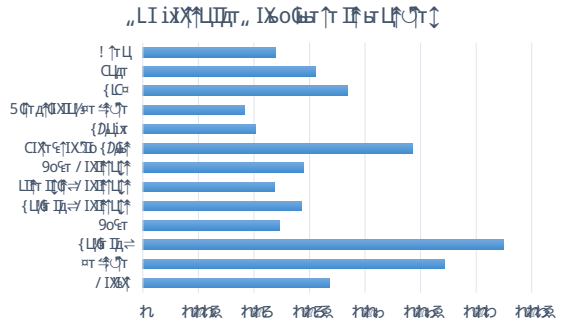


Figure 6. Importance indicator of different features. The values are normalized by their sum.

	offset	scale	h_r	b_r
Saliency-based	47.7	1.30	71.3%	47.3%
Aesthetics-based	47.5	1.46	96.3%	132.9%
Direct-downsizing	50.8	1.51	95.5%	150.8%
Ours	41.9	1.29	90.0%	85.5%

Table 1. Cross-validation comparisons.

the SVM-predicted label of test example t using features F, and l_t is the actual label of t.

The degree of each feature's importance is reflected by the difference in E_F with and without each feature f:

$$Importance_f = E_F - E_{F-\{f\}} \tag{7}$$

These values are exhibited in Figure 6. It can be seen that each of the features contributes to the overall performance, and that the saliency and texture features have a relatively larger impact.

4.3. Cross-Validation

For a quantitative evaluation of our method, we conducted a cross-validation experiment using the 200 test images with thumbnails from Sec. 4.2. The expertly-produced thumbnails are treated as ground truth in this cross-validation. We also compare to three other techniques. One is a saliency-based approach [33] that incorporates scale and objectness information by searching for the crop window that maximizes scale-dependent saliency. The second technique computes crop windows using the aesthetics-based method of [38], which accounts for relationships between the original and cropped image. Solutions from these two methods are constrained to the target aspect ratio and are rescaled to thumbnail size. The third comparison method directly downsizes the original photograph by finding the largest and most central crop window of the target aspect ratio and then rescaling it to the thumbnail size.

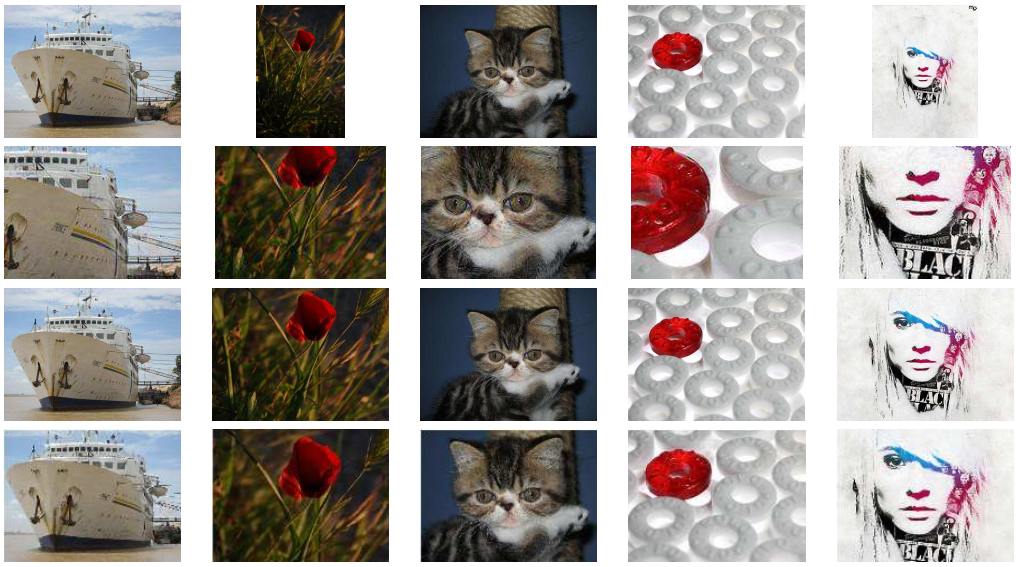
We utilize two difference measures between thumbnail results and ground truth. The first is the offset, computed as the distance between the centers of their two corresponding crops in the original photograph. The second is the ratio of their rescaling factors, calculated as $\max(\frac{s_r}{s_g}, \frac{s_g}{s_r})$ where s_r and s_g denote the rescaling factor for a thumbnail result and the ground truth, respectively. We additionally examine two other metrics that have been used for image comparison, namely hit ratio and background ratio [8]. The hit ratio measures how much of the ground truth area is captured by the thumbnail result, and is computed as $h_r = \frac{|GroundTruth \cap Result|}{|GroundTruth|}$. The background ratio represents how much area outside the ground truth thumbnail is included in the thumbnail result. It is calculated as $b_r = \frac{|Result| - |Result \cap GroundTruth|}{|GroundTruth|}$. A higher hit ratio and lower background ratio jointly indicate a result closer to the ground truth.

The average values for these evaluation metrics over the 200 images are listed in Table 1. The results of our method give the closest match to ground truth in terms of offset and rescaling factor. The aesthetics-based and direct-downsizing methods have a high hit ratio and high background ratio, since they tend to crop relatively little from the photograph. The direct-downsizing method only crops enough to satisfy the target aspect ratio, while we observe that the aesthetics-based method often crops conservatively. The saliency-based method instead tends to crop the original image substantially, which leads to a low hit ratio and low background ratio. By contrast, the amount of cropping in our method varies in a manner that balances recognizability and representativeness. It is found in this experiment to have a hit ratio that is high and a foreground ratio that is relatively low.

We note that though the images for this evaluation are different from those used for SVM training, they were created by the same expert photographer. This might give our method an advantage over the other techniques, since if there are any idiosyncracies in the photographer's thumbnail generation method, they may be captured in our SVM. Also, the images for this evaluation and those used for SVM training are from the same dataset, MIRFLICKR-25000 [1]. For an unbiased evaluation, we also conducted a user study that includes other datasets.

4.4. User Study

In the user study, each participant was presented a sequence of ten images randomly selected from a combined dataset with the 200 images from Sec. 4.2 and 490 images from the dataset of [33]. With each image they were also shown the four thumbnails from the compared techniques in a random order. They were instructed to select the "most useful" thumbnail for each given image. A total of 411 people participated in this study, most of whom completed all ten selections, and each image received either 5 or 6 votes. The results are exhibited in Figure 7. Among the



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

Figure 8. Thumbnails generated by different methods. First row: original images shown at low resolution. Second row: thumbnails from the saliency-based method (SOAT) [33]. Third row: thumbnails from the aesthetics-based method [38]. Fourth row: thumbnails from our method. The original images of (b)(d)(e) are from our dataset, while (a)(c) are from the dataset of [33].

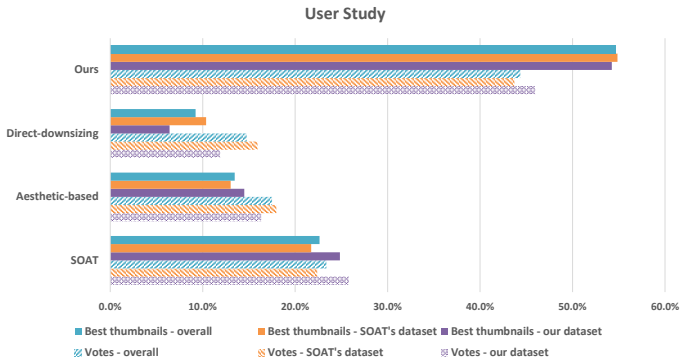


Figure 7. User study results. The bars with solid fill indicate the percentage of images for which a method was voted the best. The bars with pattern fill represent the percentage of overall votes that were cast for each method.

four methods, ours received the most overall votes (44.4%, with 46.0% among the images from our dataset and 43.7% among the images from [33]). Our method also collected the most votes on 377.3 of the images (54.7% overall, with 54.3% for our dataset, and 54.9% for the dataset in [33]). There were 77 images for which two or more methods tied for the most votes. In each of these cases, the n methods each received credit for $1/n$ image.

In Figure 8, we show some examples of thumbnails generated by different methods. It can be seen that the saliency-based method (SOAT) discards less salient parts of images, but may also remove important contextual information, making the thumbnails less suitable as an image index. SOAT may also be affected by salient background regions. The aesthetics-based method tends to produce thumbnails that visually represent the original image well. However, its limited cropping leads to considerable rescaling to reach thumbnail size, and this sometimes results in foregrounds more difficult to see. Our method generally exhibits a good tradeoff between representativeness and recognizability by determining a proper size and location of the crop window. The full set of 690 photos with thumbnails generated by the different methods is provided as supplemental material.

5. Conclusion

We presented a method for thumbnail generation that is guided by two major considerations for a useful image index. Thumbnail features were proposed to model these considerations, and their relative importance in thumbnail evaluation is learned with an SVM model trained on pairs of photos and expertly-created thumbnails. By learning from examples, our method can effectively position the crop window and balance the competing goals of visual representativeness and foreground recognizability.

Our method relies on techniques for saliency and foreground estimation. Errors in either of these will degrade the quality of our results, as well as those of other thumbnail methods. In some cases, such as photographs with multiple foreground regions that are small and separated, it would be difficult for our method to generate a thumbnail without significant sacrifices in representativeness and/or recognizability. Such photos would also be a challenge for photographers to handle.

Acknowledgments: This work was partially supported by National Science Foundation of China (61373071).

References

[1] The mirflickr retrieval evaluation. http://press.liacs.nl/mirflickr/.

[2] E. Ardizzone, A. Bruno, and G. Mazzola. Visual saliency by keypoints distribution analysis. In *Image Analysis and Processing*, pages 691–699. 2011.

[3] E. Ardizzone, A. Bruno, and G. Mazzola. Saliency based image cropping. In *Image Analysis and Processing*, pages 773–782. 2013.

[4] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. In *ACM Trans. Graph.*, volume 26, 2007.

[5] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, (6):679–698, 1986.

[6] V. Castelli and L. D. Bergman. Image databases. *Jon Wiley & Sons*, 2002.

[7] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *IEEE Computer Vision and Pattern Recognition*, pages 409–416, 2011.

[8] M. Cho, Y. M. Shin, and K. M. Lee. Co-recognition of image pairs by data-driven monte carlo image exploration. In *European Conf. on Computer Vision*, pages 144–157. 2008.

[9] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini. Self-adaptive image cropping for small displays. *IEEE Trans. Consumer Electronics*, 53(4):1622–1627, 2007.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[11] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *ACM SIGMM Workshop on Multimedia Information Retrieval*, pages 253–262, 2005.

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int'l Journal of Computer Vision*, 59(2):167–181, 2004.

[13] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):1915–1926, 2012.

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[15] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE Computer Vision and Pattern Recognition*, volume 1, pages 419–426, 2006.

[16] X. Li and H. Ling. Learning based thumbnail cropping. In *ICME*, pages 558–561, 2009.

[17] F. Liu and R. W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(7):722–733, 1996.

[18] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. In *Computer Graphics Forum*, volume 29, pages 469–478. Wiley Online Library, 2010.

[19] D. G. Lowe. Object recognition from local scale-invariant features. In *Int'l Conf. on Computer Vision*, volume 2, pages 1150–1157, 1999.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l Journal of Computer Vision*, 60(2):91–110, 2004.

[21] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *Int'l Conf. on Computer Vision*, pages 2206–2213, 2011.

[22] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *European Conf. on Computer Vision*, pages 386–399. 2008.

[23] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Int'l Conf. on Computer Vision*, pages 2232–2239, 2009.

[24] S. P. Mathew, V. E. Balas, K. Zachariah, and P. Samuel. A content-based image retrieval system based on polar raster edge sampling signature. *Acta Polytechnica Hungarica*, 11(3), 2014.

[25] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. In *ACM Multimedia*, pages 669–672, 2009.

[26] M. Rubinstein, A. Shamir, and S. Avidan. Multi-operator media retargeting. In *ACM Trans. Graph.*, volume 28, page 23, 2009.

[27] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *ACM SIGCHI*, pages 771–780, 2006.

[28] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *IEEE Computer Vision and Pattern Recognition*, 2008.

[29] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[30] I. Sobel. History and definition of the sobel operator. 2014.

[31] F. Stentiford. Attention based auto image cropping. In *Int. Conf. Computer Vision Systems*, 2007.

[32] M. A. Stricker and M. Orengo. Similarity of color images. In *IS&T/SPIE Symp. Electronic Imaging: Science & Technology*, pages 381–392, 1995.

[33] J. Sun and H. Ling. Scale and object aware image thumbnailing. *Int'l Journal of Computer Vision*, 104(2):135–153, 2013.

[34] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Trans. Systems, Man and Cybernetics*, 8(6):460–473, 1978.

[35] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum. Picture collage. In *IEEE Computer Vision and Pattern Recognition*, volume 1, pages 347–354, 2006.

[36] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee. Optimized scale-and-stretch for image resizing. *ACM Trans. Graph.*, 27(5), 2008.

[37] R. Xiao, H. Zhu, H. Sun, and X. Tang. Dynamic cascades for face detection. In *Int'l Conf. on Computer Vision*, pages 1–8, 2007.

[38] J. Yan, S. Lin, S. B. Kang, and X. Tang. Learning the change for automatic image cropping. In *IEEE Computer Vision and Pattern Recognition*, pages 971–978, 2013.