# LEARNING BASED THUMBNAIL CROPPING

*Xin Li* *Haibin Ling*

Center for Information Science & Technology, Computer & Information Science Department
Temple University, Philadelphia, PA 19122
{xin.li, hbling}@temple.edu

## ABSTRACT

Thumbnail cropping helps improve thumbnail readability by cropping images before shrinking them. In this paper we propose a learning based method for automatic thumbnail cropping. To this end, we use a support vector machine to learn a discriminative model that simultaneously captures the saliency distribution and spatial priors. The model is then used to determine the best cropping rectangle. The proposed approach improves traditional saliency based cropping techniques by introducing the spatial priors, which is automatically learned through learning process. The new method is tested on images from the PASCAL08 dataset, where it outperforms previous saliency based cropping.

***Index Terms***— Thumbnail cropping, support vector machine, visual saliency

## 1. INTRODUCTION

Image summarization has applications in many fields, such as multimedia retrieval, image browsing, human computer interaction, etc. Traditionally, a thumbnail is created by shrinking from an input image. However, as demonstrated in [1], these methods are usually less effective especially for very small thumbnails that are particularly important for small screen devices such as PDA and mobile phone. For this reason, intelligent image summarization has recently attracted many research efforts. These methods usually analyze the importance of each pixel in a given image, and then discard less important regions (or pixels).

A group of approaches, namely thumbnail cropping, crop images before shrinking them into thumbnails. The effectiveness of thumbnail cropping is shown in Suh et al. [1]. A key issue is to decide the cropping rectangle. Visual attention models [2] serves as a natural criterion to measure the importance of each pixel. For example, in [1] the optimal rectangle is defined as a trade-off between its size and the internal saliency value. Similar idea is also used in [3]. Another group of approaches eliminate unimportant image pixels while trying to keep important image structures that usually have high level semantic meanings. For example, Samadani et al. [4] analyze the sharp and blurry originals to present sharp in the thumbnails, thus provide quick and natural thumbnails. Avidan and Shamir [5] present the seam carving, which is sometimes referred to as content aware image resizing (primarily shrinking). This method shrinks the low-gradient pixels on image to reshape the whole image. However it meets problems if the important content appears uniformly on the whole image. Simakov et al. [6] presents another nice way for image shrinking that overcomes, to some degree, the problems suffered by seam carving. Other related work can be found in [7, 8].

In this paper, we follow the thumbnail cropping framework in [1]. While visual saliency provides an importance measure for image locations, it does not take into account the spatial prior of the true region of interest (ROI). For example, the center of an image is much more likely to be in ROI than the margin of an image. This effect is actually handled ad hoc in [3]. Furthermore, the saliency values [2] are defined based on low level image statistics and therefore not always consistent with the high level image semantics. To alleviate these problems, we propose learning a discriminative model that automatically detects the best cropping rectangles.

Our contribution is mainly two-fold. First, we propose using the learning-based approach that provides a general framework for thumbnail cropping. A support vector machine (SVM) [9] is used in our study, but other approaches (e.g., boosting [10]) can be easily applied in the framework. Second, we propose combining low level saliency feature and spatial statistics for better ROI detection. In addition, we will make our annotation public to research studies.

The rest of the paper is organized as follows. Section 2 discusses the thumbnail cropping framework using SVM. Then, Section 3 describes experimental results on Pascal 2008 [11]. Finally Section 4 gives the conclusion and discussion.

## 2. LEARNING BASED THUMBNAIL CROPPING

Given an input image $I$, the task of thumbnail cropping is to find a rectangle $\hat{\mathbf{r}}$ that is optimal according to some criterion. The problem can be formulated as

$$\hat{\mathbf{r}} = \arg\max_{\mathbf{r}\in\mathcal{R}} f(\mathbf{r}; I) \ , \qquad (1)$$

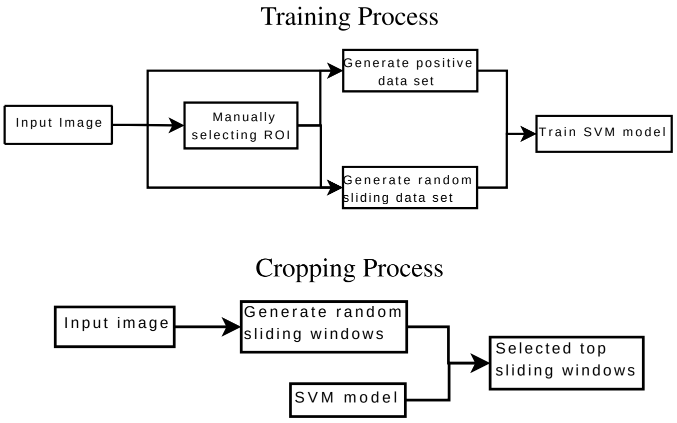
Training Process
Input Image
Manually selecting ROI
Generate positive data set
Generate random sliding data set
Train SVM model
Cropping Process
Input image
Generate random sliding windows
SVM model
Selected top sliding windows

**Fig. 1**. Flowchart of training and cropping process.

where $\mathcal{R}$ is the set of all rectangles, $\mathbf{r} = (x, y, w, h)$ contains the center of a rectangle $(x, y)$ as well as width $w$ and height $h$, and $f(\mathbf{r}; I)$ is the criterion function that depends on image $I$ (or, features extracted from image $I$).

The problem now is to find a proper criterion function $f(\mathbf{r}, I)$. In [1], the visual saliency [2] is used for this task, and the optimization is done with a greedy search strategy. In this paper, we propose an alternative solution that use learning framework for this task. Specifically, we learn a discriminative model from human annotation and then apply the model to unseen images. We use support vector machine (SVM) for the criterion function, and use saliency and rectangle location for features. The framework of training and cropping is illustrated in Figure 1.

### 2.1. Combination of saliency and spatial prior

The features used in $f(\mathbf{r}, I)$ contains two parts, the visual saliency and the spatial localization of $\mathbf{r}$.

We use the saliency map based on [2], same as in [1]. The basic idea is to use local contrasts as a measure of saliency. The contrasts are averaged across three channels: color, intensity, and orientation. Instead of using all saliency values in a candidate rectangle $\mathbf{r}$, we use a normalized saliency, which is defined as

$$s(\mathbf{r}, I) = \frac{S(\mathbf{r}, I)}{S(\Omega, I)} , \qquad (2)$$

where $S(\mathbf{r}, I)$ is the sum of all saliency values of image $I$ in domain $\mathbf{r}$, and $\Omega$ denotes the whole image region.

For spatial information, we use normalized rectangle information of $\mathbf{r}$. Specifically, let $W$ and $H$ be the width and height of the input image $I$, a normalized rectangle of $\mathbf{r}$ is given by $(x', y', w', h') = (x/W, y/H, w/W, h/H)$.

Finally, the feature vector $\mathbf{x}$ of rectangle $\mathbf{r}$ is represented as $\mathbf{x} = \mathbf{x}(\mathbf{r}, I) = (x', y', w', h', s(\mathbf{r}, I))^\top \in [0..1]^5$. The feature vector will be used for support vector machine.

### 2.2. Detecting cropping rectangle using SVM

Support vector machine (SVM) [9] is a popular machine learning technique and has been demonstrated excellent performance in many applications. Given $n$ training samples $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i$ are features and $y_i \in \{-1, +1\}$ are labels, the training of SVM maximizes the linear separation margin. Kernel techniques are used for implicitly handling the high dimensional nonlinear feature spaces. The classification boundary can be written as

$$\sum_{i=1}^{N_s} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + b = 0 , \qquad (3)$$

where $K(\mathbf{s}_i, \mathbf{x})$ is the kernel function, $\mathbf{s}_1, \ldots, \mathbf{s}_{N_s}$ are the $N_s$ support vectors. The task of learning is to select support vectors $\mathbf{s}_i$ and estimate associated parameters $\alpha_i$.

The above model originally aims at two-class classification tasks. Adjust it to our task, where a criterion function is desired, we have

$$f(\mathbf{r}, I) = \sum_{i=1}^{N_s} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}(\mathbf{r}, I)) + b . \qquad (4)$$

Now the criterion function $f(\mathbf{r}, I)$ is ready, we use the sliding window to search for the best rectangle, i.e., to find the cropping rectangle according to (1).

For implementation, we use the LibSVM package [12] and choose a Gaussian kernel defined as

$$K(\mathbf{s}_i, \mathbf{x}) = \exp\left(-\frac{||\mathbf{s}_i - \mathbf{x}||^2}{2\sigma^2}\right) . \qquad (5)$$

## 3. EXPERIMENT AND RESULT

### 3.1. Proposed Experiment

Our experiments follow the flow chart in Fig.1. We used the latest Visual Object Classes Challenge 2008 (Pascal2008) database [11]. Specifically, we randomly chose 50 images for training and another 50 for testing. We manually annotated both training and testing images with groundtruth cropping rectangles.

The positive and negative training samples (rectangles) are derived from the annotation of training data. Given a training image, its positive samples contains the manually selected cropping rectangle and several rectangles perturbed from it. One example image is shown in Fig. 2, where the ground truth is shown in blue and perturbed samples in red. For negative samples, we first generated a large number of rectangles, then we pick those rectangles who has less than 30% overlap with the ground truth bounding box. The whole positive and negative samples will be used to train the SVM model.

For SVM training process, the Gaussian Kernel with $\sigma = 0.1$ is used. This has given satisfactory results (the training error is 92.53%) for training process.

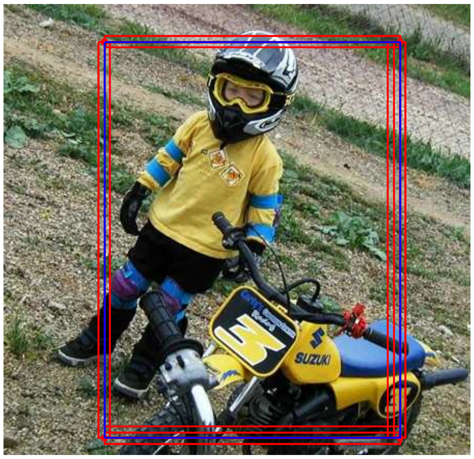

**Fig. 2**. An example image used in our experiment. The blue box is the annotated ground truth, and red boxes are the perturbed positive training samples.

In the cropping process, the sample rectangles are exhaustively generated over the whole image. For each window (bounding box), we use the center position, the width, height and saliency ratio to define the image feature inside the window. During the cropping process, the rectangles are tested by the learned SVM model. SVM model outputs the confidences of the testing samples. The higher confidence a bounding box has, the better it represents the semantic meaning of the image. Finally, we use the overlapping rate to evaluate the cropping results. In this experiment, the overlapping rate between the detected bounding box and the ground truth is based on this following function:

$$OL = \frac{Area(\mathbf{r}_{true} \bigcap \mathbf{r}_{det})}{Area(\mathbf{r}_{true} \bigcup \mathbf{r}_{det})} \tag{6}$$

where OL is Overlapping Rate, $Area.$ computes the area of given region, $\mathbf{r}_{true}$ is the ground truth rectangle and $\mathbf{r}_{det}$ is the detected cropping rectangle.

### 3.2. Experiment Result

We tested the learned model on 50 images randomly selected from the PASCAL 2008 dataset. For comparison, we also applied the saliency based approach (named auto-thumbnail) in [1] on the same dataset. The average overlapping rates for both approaches are listed in Table 1. The results show that, in average, the proposed approach outperforms previous saliency based thumbnail corpping method. Several example results are included in Figure 3.

**Table 1**. Average overlapping rates by auto-thumbnail[1] and by the proposed approach

| Case | AutoThumbnail[1] | SVM model result |
|---|---|---|
| Average | 51.10 | 56.84 |

From these results, we see that saliency information only sometimes is not enough for cropping tasks. These methods usually work fine for images with relatively condensed saliency distribution, but often meet problems for complex images, such as images from PASCAL 2008. In comparison, the propose approach compensates the saliency information by implicitly using the spatial statistics of cropping bounding boxes.

Figure 4 shows a detailed output form [1], which verifies our explanation. The cropped image region is based on the saliency map. The original image is on the upper left; the saliency map is on the lower left; the detected region according to saliency map is shown on lower right. The cropped region is shown on the upper right. This figure shows that the low level feature based saliency map can sometimes mislead the understanding, due to the lack of global context information.

## 4. CONCLUSION AND DISCUSSION

This paper has presented an approach for cropping semantic meaningful region from the image. First we propose the robust feature to generate the training data set, then SVM model is trained on manually labeled datasets. During the cropping process, exhaustive search ranks all sliding windows and picks the best one according the learnt model. The effectiveness of the proposed approach is demonstrated in a public database, compared favorably with the previous saliency based greed cropping.

In the future, we plan to further investigate the learning based image summarization. We believe that both the learning part and representation part have large rooms for improvement. In addition, we plan to use a larger database for more solid studies.

## 5. REFERENCES

[1] Bongwon Suh, Haibin Ling, Benjamin B. Bederson, and David W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in *Proceedings of the 16th annual ACM symposium on User interface software and technology (UIST)*, 2003, pp. 95–104.

[2] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[3] Li-Qun Chen, Xing Xie, Xin Fan, Wei-Ying Ma, Hong-Jiang Zhang, and He-Qin Zhou, "A visual attention model for adapting images on small displays," *ACM*

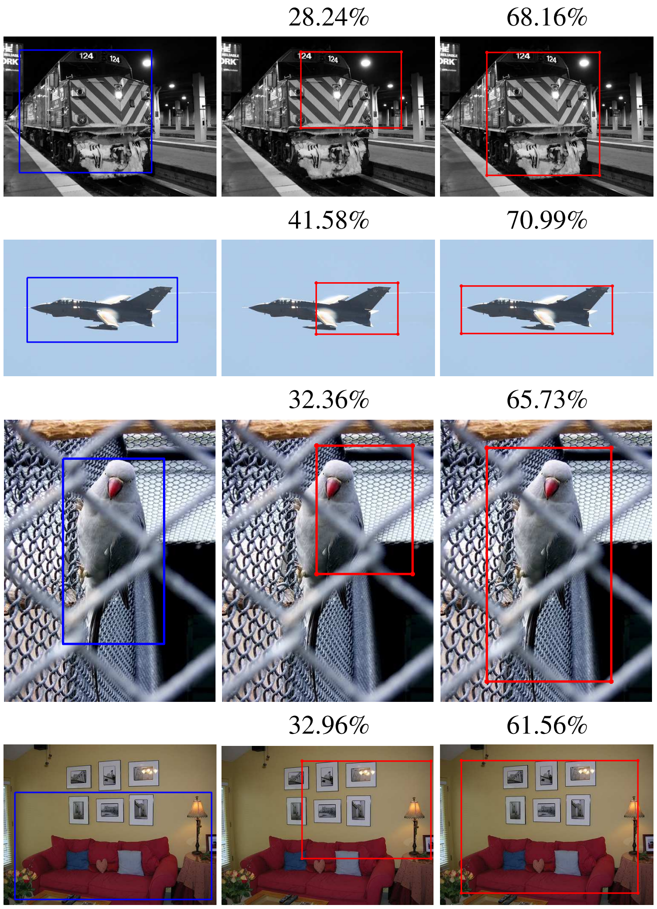

**Fig. 3**. Some experimental results. Left: images with the ground truth annotated. Middle: cropping results from [1]. Right: results from the proposed approach. The overlapping rates of result bounding boxes are given above the corresponding figure.

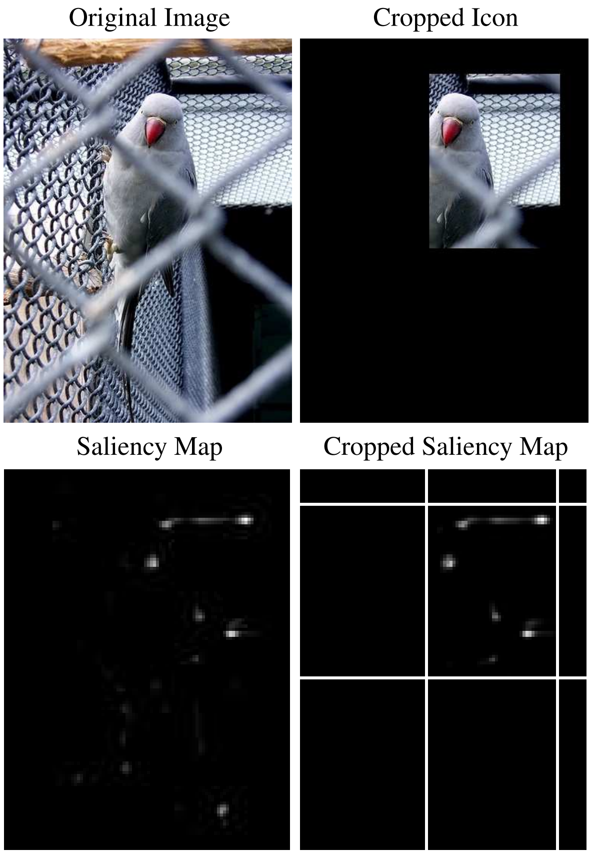

**Fig. 4**. result from AutoThumbnail [1]

*Multimedia Systems Journal*, vol. 9, no. 4, pp. 353–364, 2003.

[4] Ramin Samadani, Suk Hwan Lim, and Dan Tretter, "Representative image thumbnails for good browsing," in *Proceedings*. ICIP, 2007.

[5] Shai Avidan and Ariel Shamir, "Seam carving for content-aware image resizing," in *Proceedings*. 2007 SIGGRAPH conference, 07,2007, vol. 26.

[6] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani, "Summarizing visual data using bidirectional similarity," in *Proceedings*. IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[7] Feng Liu and Michael Gleicher, "Automatic image retargeting with fisheye-view warping," in *Proceedings of ACM symposium on User interface software and technology (UIST)*, 2005, pp. 153–162.

[8] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake, "Digital tapestry," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 589–596.

[9] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, 1995.

[10] Yoav Freund and Robert E. Schapire, "Experiments with a new boosting algorithm," in *ICML*, 1996, pp. 148–156.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[12] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.