Multimodal Content-Aware Image Thumbnailing

Kohei Yamamoto† Hayato Kobayashi‡ Yukihiro Tagami‡ Hideki Nakayama†
†The University of Tokyo ‡Yahoo Japan Corporation
{yamamoto, nakayama}@nlab.ci.i.u-tokyo.ac.jp {hakobaya, yutagami}@yahoo-corp.jp

ABSTRACT

News article recommendation has the key problem of needing to eliminate the redundant information in a ranked list in order to provide more relevant information within a limited time and space. In this study, we tackle this problem by using image thumbnailing, which can be regarded as the summarization of news images. We propose a multimodal image thumbnailing method considering news text as well as images themselves. We evaluate this approach on a real data set based on news articles that appeared on Yahoo! JAPAN. Experimental results demonstrate the effectiveness of our proposed method.

1. INTRODUCTION

Image thumbnailing is a technique for creating reduced-size versions of images to improve their *visibility*, which means the capability of allowing users to easily recognize their content. This technique is one of the most important factors in enhancing the user experience of applications displaying images, especially for mobile devices. There are two approaches to effectively improving visibility. One is *image cropping*, which means cutting out important parts expressing content in an image, and the other is *image retargeting*, which means reconstructing a new image including such parts. We focus on the cropping approach, since it would be more appropriate for our target application, i.e., a news curation service. In fact, most content holders prohibit the use of reconstructed thumbnails because they carry the risk of misleading users.

An interesting challenge of image thumbnailing is how to define the content of an image. There have been several studies on image thumbnailing defining *visual saliency* as important content in an image [5]. However, in the case of news articles, the content of a news image can vary depending on the corresponding text. For example, let us consider an image of a person holding a fish. The person should be the focus when the corresponding text is "A famous person went fishing", while the fish should be when the text is "A rare fish has been caught."

In this paper, we propose a multimodal image thumbnailing method considering both images and text. The proposed method generates an energy map expressing content by aligning image fractions and words via multimodal neural networks, and we can crop an appropriate region with respect to the corresponding text by using the energy map. To the best of our knowledge, there is no study directly considering multimodal image thumbnailing.

2. METHOD

The goal of our method is to generate thumbnails reflecting the content of corresponding text. We first briefly describe a model that aligns text to the visual regions through multimodal embeddings. We then treat these alignment scores as energy scores to generate multimodal energy maps. Final thumbnail regions are determined from these energy maps.

2.1 Learning multimodal alignment

Following the method of Karpathy et al. [3], we learn the alignment between words of sentences and the regions of the images.

Image representations. We detect candidate objects in every image with a Region Convolutional Neural Network (RCNN) [1] and VGGNet [4] pre-trained with ImageNet. We then use the top 10 detected bounding boxes and compute the representations on the basis of the pixels inside each bounding box $\{I_i \mid i = 1, ..., 10\}$ as follows:

$$v_i = W_{\mathrm{m}}[\mathrm{CNN}(I_i)] \qquad (1)$$

where $\mathrm{CNN}(I_i)$ transforms the pixels inside bounding box I_i into 4,096-dimensional activations of the fully connected layer immediately before the classifier. W_{m} has $h \times 4,096$ dimensions. Every image is thus represented as a set of h-dimensional vectors v_i.

Sentence representations. We use a Bidirectional Recurrent Neural Network (BRNN) to compute the word representations as follows:

$$s_t = \mathrm{BRNN}(\mathbb{I}_t) \qquad (2)$$

Here, index $t = 1, ..., N$ denotes the position of a word in a sentence and $\mathbb{I}_t$ is an indicator column vector that has a single one at the index of the t-th word in a word vocabulary. $\mathrm{BRNN}(\mathbb{I}_t)$ takes a sequence of N words and transforms each one into an h-dimensional vector s_t.

Image-sentence alignments. Following [3], we interpret the dot product $v_i^T s_t$ between the i-th region and t-th word as a measure of similarity and use it to define the score between image k and sentence l as follows:

$$S_{kl} = \textstyle\sum_{i \in g_k} \max_{t \in h_l} v_i^T s_t \qquad (3)$$

Here, g_k is the set of image regions in image k, and h_l is the set of words in sentence l. Assuming that S_{kk} expresses the

WWW'16 Companion, April 11–15, 2016, Montreal, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889413.

Table 1: Experimental results.

	Accuracy
`Saliency Map`	0.7067
`RCNN-based`	0.7533
`Multimodal`	0.7633
`Saliency Map + Multimodal`	**0.7967**

score of corresponding image-sentence pair, and we optimize the following ranking loss:

$$C(\theta) = \textstyle\sum_k \Big[\sum_l \max(0, S_{kl} - S_{kk} + 1) + \sum_l \max(0, S_{lk} - S_{kk} + 1) \Big] \quad (4)$$

This objective encourages aligned image-sentence pairs to have a higher score. We use Stochastic Gradient Descent to optimize the model. We cross-validate learning rate and weight decay due to overfitting concerns. During the test, we compute image region-sentence scores from this model.

2.2 Generate thumbnails via multimodal energy maps

Multimodal content-aware energy map. We first get the top 100 detected locations and their detection scores in every image with an RCNN. We then generate the energy map that represent the existence of objects by accumulating the detection scores for every pixel in the corresponding locations in every image, and we call it `RCNN-based` energy map. We then generate a multimodal energy map by accumulating the image region-sentence scores mentioned above on the `RCNN-based` energy map in the same way. We call it `Multimodal` energy map. This `Multimodal` energy map enhances the locations reflecting the content of the corresponding sentence.

Find thumbnail region. Once we get the energy map, our goal is to find the final crop region R_{C} expected to contain the most important content. Since we use this energy value as the criteria of importance, the sum of energy values within R_{C} should become as high as possible. Based on this idea, we can find R_{C} from the following thresholded candidates set $\mathfrak{R}(\lambda)$ that is a subset of all candidate set $\mathcal{R}$ that satisfy the required aspect ratio.

$$\mathfrak{R}(\lambda) = \left\{ r \mid \frac{\sum_{(x,y)\in r} E(x,y)}{\sum_{(x,y)\in P} E(x,y)} > \lambda \right\} \quad (5)$$

Here, P denotes a set of all pixels in a given image. r denotes a set of all pixels in a candidate region. $E(x,y)$ denotes the energy value of (x,y). λ denotes the fraction threshold. Then, final region R_{C} is determined as follows:

$$R_{\mathrm{C}} = \begin{cases} \underset{r\in\mathcal{R}}{\arg\max} \sum_{(x,y)\in r} E(x,y) & (\mathfrak{R}(\lambda) = \emptyset) \\ \underset{r\in\mathfrak{R}(\lambda)}{\arg\min} A_r & (otherwise) \end{cases} \quad (6)$$

Here, A_r denotes the area of the region r.

3. EXPERIMENT

Dataset. We evaluated our method with a dataset based on news articles and images used on Yahoo! JAPAN. This dataset contained 2,954 news articles, and each had an original image and a thumbnail with an aspect ratio of 242 × 100 that a professional editor cut out by manual operation for the mobile news application. We used these thumbnails as ground truth regions. In the dataset, we used 300 for testing and the rest for training and validation.

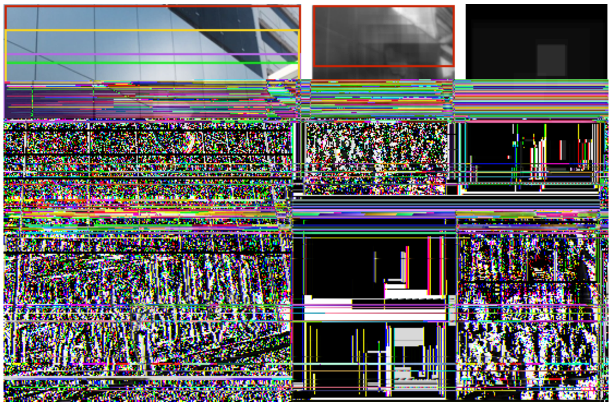

Figure 1: Left: Original image. Green rectangle is ground truth. Right: Top left is Saliency Map, top right is RCNN-based, bottom left is Multimodal, bottom right is Saliency Map+Multimodal. Article: "Apple showed the patent... "[1]

Evaluation. As an evaluation, we calculated the intersection over union (IOU) value as follow:

$$\mathrm{IOU} = \frac{R_{\mathrm{C}} \cap R_{\mathrm{GT}}}{R_{\mathrm{C}} \cup R_{\mathrm{GT}}} \quad (7)$$

Here, R_{C} denotes the predicted region and R_{GT} denotes the ground truth region. We assumed that the ratio of samples satisfied IOU > 0.5 for *Accuracy*. We adopted the `Saliency Map` [2] as a baseline. We combined the `Saliency Map` with the `Multimodal` energy map with early-fusion. We then searched for the best combination ratio of each energy map using cross-validation. The experimental results are summarized in Table 1. The bolded number indicates the best performance. The `Multimodal` model achieved better results than only visual information models. Figure 1 shows an example of cropping regions in the results of Table 1. We can see that the `Saliency Map` wrongly recognized the upper shadow region as the important content and that `RCNN-based` tended to focus on people since this is basically important in generic object detection. The `Multimodal` appropriately cropped the Apple logo. This implies that our method could reflect the content of the sentence. In this example, the result of `Saliency Map + Multimodal` was a little worse than `Multimodal`, but the overall accuracy of this combination method is the best in Table 1. We conclude that the `Saliency Map` and `Multimodal` play a complementary role to each other.

4. REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[2] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on PAMI*, 1998.

[3] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] B. Suh, H. Ling, B. B. Bederson, and J. D. W. Automatic thumbnail cropping and its effectiveness. In *UIST*, 2015.

[1] http://headlines.yahoo.co.jp/hl?a=20151025-00010001-newswitch-sci [Access: 5 February, 2016]