Integrating Visual Saliency and Consistency for Re-Ranking Image Search Results

Jun Huang, Xiaokang Yang, *Senior Member, IEEE*, Xiangzhong Fang, Weiyao Lin, and Rui Zhang

***Abstract*—In this paper, we propose a new algorithm for image re-ranking in web image search applications. The proposed method focuses on investigating the following two mechanisms: 1) Visual consistency. In most web image search cases, the images that closely related to the search query are visually similar. These visually consistent images which occur most frequently in the first few web pages will be given higher ranks. 2) Visual saliency. From visual aspect, it is obvious that salient images would be easier to catch users' eyes, and it is observed that these visually salient images in the front pages are often relevant to the user's query. By integrating the above two mechanisms, our method can efficiently re-rank the images from search engines and obtain a more satisfactory search result. Experimental results on a real-world web image dataset demonstrate that our approach can effectively improve the performance of image retrieval.**

***Index Terms*—Random walk, re-ranking, visual consistency, visual saliency.**

I. Introduction

IMAGE search on the Web is of increasing importance in our daily life. Currently, many search engines have been developed to provide image search services on the Internet [1], [2]. However, since most of these search engines are mainly built on text-based search, many of the image research results are unsatisfactory or even irrelevant to the query. Although recently some search engines such as Google and Bing have introduced content-based retrieval, it is only served as a complement to textual search and the performance improvement is still limited. Therefore, it is necessary to develop new algorithms to refine (or re-rank) the resulting images from the existing search engines so that more satisfactory search results can be obtained.

Several researches have been done in image-based search result refinement. Some methods try to introduce visual information of images to refine textual search results. Fergus *et al.* [3] proposed to use the object class model to filter the output of image search engines when searching object categories. Berg *et al.* [4] developed a visual-information-based system to collect a large number of animal pictures from the web. However, these methods require a specific model for the corresponding query or concept in advance, so they are impractical for large-scale applications.

Other works consider the visual consistency of images and emphasize images which occur frequently in the search results from the search engine [5]–[7]. These approaches are based on the observation that the images related to the query are often visually similar while images that are unrelated to the search query usually look different from each other. Although the idea of considering visual consistency is pretty reasonable, the problems of how to define image similarity and how to efficiently incorporate image consistency are still challenging problems. Furthermore, since visual consistency still has its limitations in some scenarios, using visual consistency alone may not be enough and other mechanisms need to be introduced for obtaining satisfactory search results. Therefore, in this paper, we propose a new framework which can efficiently integrate image consistency as well as other mechanisms. We also propose a new random-walk-based method to integrate the different mechanisms for obtaining the final refined results.

As one of the most important phenomena in biological vision, visual attention mechanism has been studied by researchers in physiology, psychology, and computer vision [8]–[10]. Recently, some visual attention models have been applied in improving the performance of image retrieval [11], [12]. These approaches use visual attention maps to extract regions of interest (ROI) from the image. In our study, what we are concerned with are not the salient regions in one image, but salient images in a group of images. When users browse the result pages returned by an image search engine, they are more likely attracted by the thumbnails which have salient object(s), or high contrast region(s) in color and intensity. Such an assumption is extended from the basic principles of human visual attention: regions that have distinctive colors or patterns should be of high saliency, which are supported by psychological evidences [13], [14]. To the best of our understanding, there is still no work that introduces visual attention models into image re-ranking. Therefore, in this paper, we propose a new visual attention model and incorporate it into image re-ranking.

The contributions of this paper can be summarized as follows: 1) We propose a new framework which integrates visual saliency and visual consistency for image re-ranking. 2) We introduce visual attention into image re-ranking and develop a new model for detecting salient images (i.e., images with more visual attention). 3) A new random-walk-based method is proposed to integrate the re-ranking results from different mechanisms for obtaining the final refined results.

Manuscript received September 06, 2010; revised December 26, 2010 and February 24, 2011; accepted February 24, 2011. Date of publication March 14, 2011; date of current version July 20, 2011. This paper was supported in part by NSFC (61025005, 60828001, 61001146, 61071155), 973 Program (2010CB731401, 2010CB731406), and the 111 Project (B07022). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jia Li.

The authors are with the Institute of Image Communication and Information Processing, Shanghai Key lab of Digital Media Processing and Transmission, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jun.huang@sjtu.edu.cn; xkyang@sjtu.edu.cn; xzfang@sjtu.edu.cn; wylin@sjtu.edu.cn; zhang_rui@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2127463

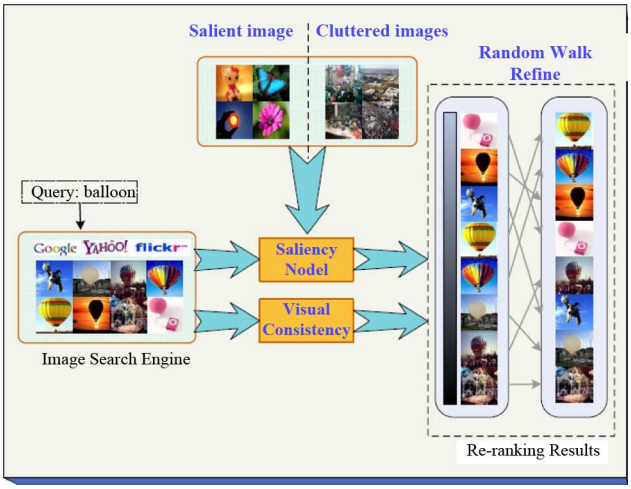


Fig. 1. Proposed re-ranking process based on visual saliency and visual consistency.

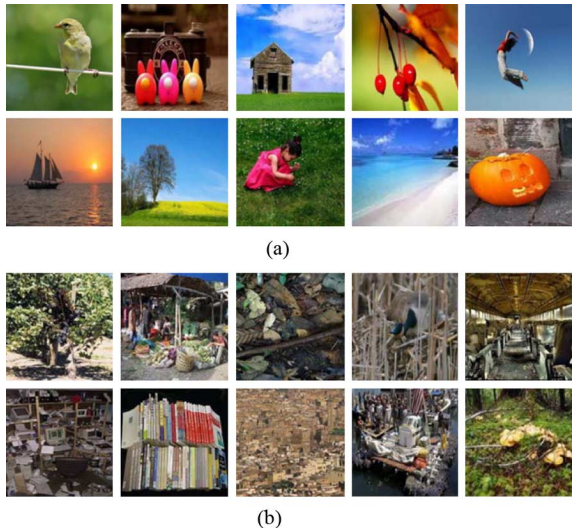


Fig. 2. Some sample images in our image database. Usually, images in (a) are more likely to attract humans' attention than images in (b). We call the former salient images and the latter cluttered images. (a) Salient images. (b) Cluttered images.

The rest of the paper is organized as follows: Section II illustrates the basic idea and the framework of our re-ranking algorithm. Section III describes the visual saliency model and the application of this model for re-ranking images. Section IV describes the image similarity model and the application of this model for re-ranking image. Section V describes the random-walk-based method for integrating different mechanisms and obtaining final results. Section VI shows the experimental results and Section VII concludes the paper.

II. Basic Idea and Framework of the Algorithm

The framework of our re-ranking method is shown in Fig. 1. In Fig. 1, a query is first submitted to one of the existing search engines. The search engine will return the resulting images, some of which are unsatisfactory to the users. The target of the proposed algorithm is to refine or re-rank these resulting images so that more relevant images are displayed first and less relevant images will be moved to the end or discarded.

In our algorithm, the images returned from the search engines are first examined by our proposed saliency model (i.e., the visuals attention model) which is trained by some datasets. Images with different saliencies will be given different "relevance scores" for later re-ranking. The introduction of saliency mechanism (or visual attention) is based on the following observation: When users browse the resulting images, they are more likely attracted by the thumbnails which have a salient object, distinctive foreground, or a region with high contrasts in a clearly visual way. Furthermore, these visually salient images in the front pages are often more relevant to the user's query. Therefore, these images should have high ranks in the refined result.

At the same time, the similarities among images from the search engine are also calculated. This is based on the observation that saliency-based re-ranking results may contain some "noises" (i.e., some salient but irrelevant images). Introducing image similarity (i.e., visual consistency) can efficiently filter these noise images by emphasizing the more frequently occurred images and disregarding less frequently occurred images. Finally, a random-walk-based method is used to obtain the final refined results.

In the following sections, we will first describe the saliency model for examining images and then describe the application of visual consistency for filtering the noise images.

III. Multiscale Visual Saliency Model

In this section, we describe our proposed saliency model to examine images. Each image examined by the saliency model will be given a "relevance score" which reflects the relevance of the image from the saliency point of view. That is, a relevance score can measure how good a retrieved result is with regard to the information needed. It encompasses topical relevance and other concerns of the user such as low-level visual stimulus. Given an image x_i, its relevance score can be calculated by the saliency model as in (1):

$$r_{sal}(x_i) = \ln \frac{P(l_{sal}|\pi_{xi})}{P(l_{clut}|\pi_{xi})} = \ln \frac{P(\pi_{xi}|l_{sal})P(l_{sal})}{P(\pi_{xi}|l_{clut})P(l_{clut})} \quad (1)$$

where $r_{sal}(x_i)$ is the relevance score for image x_i, π_{xi} are the features for x_i, l_{sal} and l_{clut} represent the labels of two image classes that we defined and $P(l_{sal}|\pi_{xi})$ and $P(l_{clut}|\pi_{xi})$ represent the probability that image x_i belongs to the class l_{sal} and l_{clut}, respectively. l_{sal} and l_{clut} are described as follows.

Salient Image Class (l_{sal}): Images in this class contain salient object(s) or high contrast region(s) in color and intensity. Some example images are shown in Fig. 2(a).

Cluttered Image Class (l_{clut}): Images in this class are of lower quality in that they may have major occlusion, serious noise, background clutter, or some other faults (i.e., it is hard to separate the main object from the background). Some example images are shown in Fig. 2(b).

Normally, if an image has higher probability of belonging to the salient image class, it will have larger relevance value. Similarly, an image with larger probability of belonging to the cluttered image class will have smaller relevance value. By differentiating these two classes, we can have an efficient model

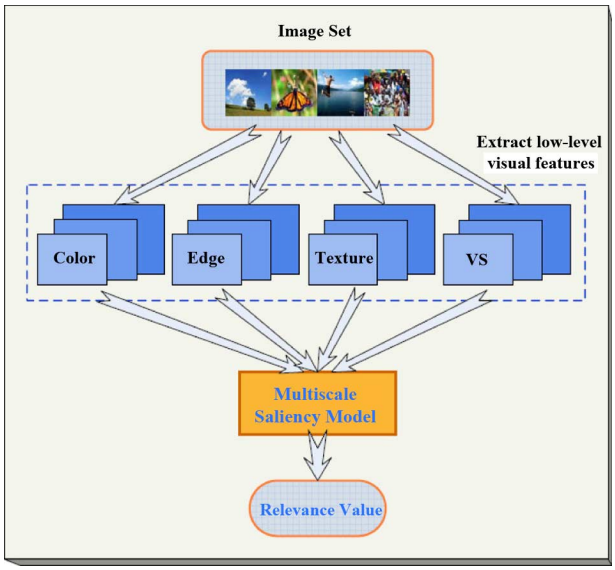


Fig. 3. Using the multiscale saliency model for calculating the relevance value.

for evaluating the image relevance. Furthermore, since we assume the dataset contains equal numbers of salient images and cluttered ones, $P(l_{sal})$ and $P(l_{clut})$ are the same and they can be dropped from equations. In the implementation, the numbers of salient image and cluttered image in our datasets are equal. Therefore, (1) can be re-written as

$$r_{sal}(x_i) = \ln \frac{P(\pi_{xi}|l_{sal})}{P(\pi_{xi}|l_{clut})}. \quad (2)$$

Fig. 3 shows the process of calculating the relevance value. In Fig. 3, the features π_{xi} of an image are first extracted. These features are separated into different sets based on four visual features and three scale levels. Then each feature set is processed separately by the models that are trained from the pre-labeled salient/cluttered image training set. Finally, these separately-processed feature set results are fused together to create the final relevance value $P(\pi_{xi}|l_{sal})$. In the following, we will describe our selection of features and the calculation of $P(\pi_{xi}|l_{sal})$ in detail.

A. Features for Describing Images

As mentioned, in our saliency model, features of an image are separated into different sets based on different visual features and scale levels. That is, an image is represented by multiple images with different scales and in each scale level, multiple feature sets with four visual features are extracted. The multiscale representation as well as the feature selection in each scale level are described in the following.

There are two reasons for using the multiscale representation: 1) Multiscale representation can give a more precise description of the image content. 2) In most situations, people judge the relevance of search results by the thumbnails and then click the thumbnails to see if the corresponding images are relevant. Therefore, normally both the small-size thumbnail images and their corresponding large-size regular images need to be considered when evaluating relevance scores. In this paper, we use Gaussian pyramidal representation [15] as the multiscale representation. However, it should be noted that the framework of our algorithm is general and other multiscale representations can also be easily used. The Gaussian pyramidal representation can be described as

$$I_{q+1}(k,l) = 2N \sum_{a=-N}^{N} \sum_{b=-N}^{N} w(a,b) I_q(2k+a, 2l+b) \quad (3)$$

where $I_0(k,l)$ is the original image and $I_q(k,l)$ is the qth scale level of the pyramid. $w(a,b)$ is the Gaussian kernel function. In our implementation, N is set to be 512 and three scale levels are created using Gaussian pyramids.

In each scale level, we extract the same features. In this paper, we extract the following four feature sets:

1) Color feature set. From Fig. 2, we note that the color spatial distributions of salient images are more concentrated than those of cluttered ones. In a salient image, colors of the object are less likely to be found in the background, while colors in a cluttered image are often scattered. Therefore, the feature of global color spatial distribution can be used to distinguish salient image class and cluttered image class. Color moments [16] is a useful and convenient feature in describing the color distribution of an image. We calculate three moments for each of the three channels in $L^*a^*b^*$ color space and aggregate the features into one feature vector.
2) Edge feature set. Since salient images often have clear backgrounds, the objects are placed in sharp edges. Therefore, we expect the edges in salient images to be clustered near the center of the image, where the objects are usually found. Edge direction histogram [17] is a simple yet effective way to characterize shape information of an object. Thus, it is used as another feature set for describing images.
3) Texture feature set. Similarly, we also expect the texture information to be meaningful enough to differentiate a clear background from a cluttered one. In the implementation, we use the local binary pattern representation [18].
4) Visual saliency (VS) feature set. For the feature of VS, we mainly take into account the contrasts in color and intensity. Adopting Itti's visual attention model [14], we get three groups of maps for one intensity channel and two color channels, respectively. Since people usually pay more attention to the regions near the center of the image, each map is covered by G_w, which is a normalized Gaussian template. The final VS vector is formed by concatenating the features of all contrast maps.

After obtaining the multiscale representation of features, we can assume that the four feature sets are independent (no matter within the same scale level or in different levels). Then the relevance value can be rewritten as

$$r_{sal}(x_i) = \ln \frac{P(\pi_{xi}|l_{sal})}{P(\pi_{xi}|l_{clut})} = \sum_{m=1}^{M} \sum_{n=1}^{N} \ln \frac{P(\pi_{mn}|l_{sal})}{P(\pi_{mn}|l_{clut})} \quad (4)$$

where π_{mn} denotes the mth feature set in the n's scale.

B. Calculating the Final Relevance Value

We can further extend (4) to a more generalized form, as in (5):

$$r_{sal}(x_i) = \sum_{m=1}^{M} \sum_{n=1}^{N} w_{mn} \cdot C(\pi_{mn}) \tag{5}$$

where $C(\pi_{mn})$ is the relevance value from the mth feature set in the n's scale and w_{mn} is the weight for fusing different $C(\pi_{mn})$. For example, in the case of (4), $w_{mn} = 1$ and $C(\pi_{mn}) = \ln P(\pi_{mn}|l_{sal})/P(\pi_{mn}|l_{clut})$. However, it should be noted that $C(\pi_{mn})$ and w_{mn} are not limited to the above values and other forms can also be incorporated to calculate the relevance value $r_{sal}(x_i)$. In the experiments, $r_{sal}(x_i)$ is calculated by the following equation:

$$r_{sal}(x_i) = \sum_{m=1}^{M} w_m \cdot C(\pi_m) \tag{6}$$

where $C(\pi_m)$ is the relevance value for the mth feature set from all scale levels (i.e., $\pi_m = [\pi_{m1}, \pi_{m2} \ldots, \pi_{mn}]$) and w_m is the weight for π_m. $C(\pi_m)$ is the confidence value which is calculated by radial basis function support vector machine (SVM). We train these SVMs on the pre-labeled salient/cluttered image, as in Fig. 3. The fusion parameter w_m is set as different values to balance the importance between different feature sets. The values of w_m are determined by cross-validation [19]. Actually, this process can also be viewed as late fusion process [20], where $C(\pi_m)$ is first calculated for different feature sets and then fused by the weighting factor w_m.

In this section, the image relevance has been calculated from the saliency point of view. Although visual saliency can provide an effective way to measure the image relevance, it still has limitations and may wrongly evaluate some irrelevant image as high-relevant images. Therefore, it is also necessary to introduce other mechanisms for providing a more satisfactory result. In the next section, we will describe the mechanism of using visual consistencies for measuring image relevance.

IV. Visual Consistency Measuring

The visual consistency mechanism is based on the similarity measure between images. In order to calculate the similarities, one popular way is to concatenate various feature sets into a long feature vector and then calculate the distance accordingly. However, the high dimension vector will cause time-consuming calculation. In this paper, we adopt a dynamic late fusion strategy [21] for similarity measuring. The method weights the importance of different features based on the variance of image similarities. For each feature, the variance is achieved by all image similarities within the image set. The method is favorable to assign larger weights to features which are good at discriminating images. This assumption is similar to the basic principle of linear discriminant analysis (LDA) [22] which treats low-variance intra-class features as important features. Moreover, since we calculate the similarity values based on features with different distributions and ranges, the variance is used as a weighting and normalizing factor. The proposed similarity measure between the images x_i and x_j is calculated as follows:

$$s(x_i, x_j) = \frac{1}{K} \sum_{k=0}^{K} \frac{1}{\sigma_k^2} s_k(x_i, x_j) \tag{7}$$

where K is the total number of features, σ_k^2 is the similarity variance of all images for the kth feature within this image set. $s_k(x_i, x_j)$ is the similarity between x_i and x_j for the kth feature, achieved by computing their Euclidean distance. When calculating the similarity in (7), we extract four global features used in Section III and two local features: histogram of edge orientations gradients (HoG) [23] and Principal component Analysis of Census Transform histograms (PACT) [24].

Based on the relevance value from the visual saliency and the similarity measure from the visual consistency, we can integrate these two mechanisms and develop a new image re-ranking algorithm. The proposed random-walk-based method for integrating the two mechanisms is described in the following section.

In this paper, the visual saliency and similarity are integrated for improving the performance of image retrieval. For comparing the results, we also do the experiments with visual similarity alone. We use the feature density estimation [25] for evaluating the relevance of similarity. The relevance can be written as

$$r_{sim}(x_i) = \frac{1}{Z_c} \sum_{x_t \in c, t \neq i} s(x_t, x_i) \tag{8}$$

where Z_c is a normalization factor and $s(x_t, x_i)$ is the similarity between x_t and x_i. The category c represents the whole set of images returned as a result of a specific query. The images within each category c will be re-ranked in descending order according to their relevance value.

V. Random-Walk-Based Integrating Method

Since the values achieved by visual saliency and consistency represent the relevance of image at different granularity, it can be expected that their combination can provide a more comprehensive description. A straightforward strategy to combine these two mechanisms is to fuse both measures by a linear model [26]. However, since they reflect image relevance from different points of view, using the linear model may not be able to obtain satisfactory results. Therefore, in this paper, we propose to use the random walk [6] for integrating the two mechanisms. Our intuition is that the random walk method can effectively balance different aspects through the iteration process.

Algorithm 1: The entire process of the proposed re-ranking method

Input: The N images $\{x_1, x_2, \ldots, x_N\}$ returned by image search engines for a certain query.

Output: Re-ranking results.

1) For the set of images, extract features for different scales;

2) Calculate the relevance score $\vec{r}_{sal}$ for input images;

3) Calculate the similarity measure between any two images;

4) Let $P_{sim}^{(k)}$ be the similarity matrix of the kth feature in the relevance graph, $P_{sim} = 1/K \sum_{k=0}^{K} 1/\sigma_k^2 P_{sim}^{(k)}$;

5) Initialize $\vec{r}_{all} = \vec{r}_{sal}$;

6) **repeat**

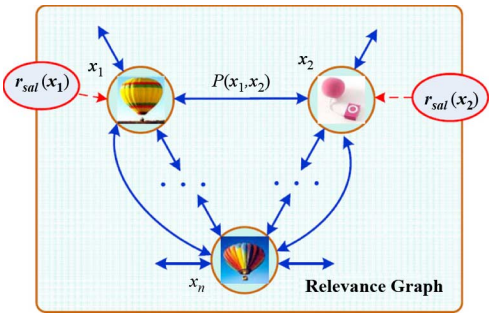


Fig. 4. Relevance graph of random walk process to refine saliency-based re-ranking results.

7) $\vec{r}_{all} = \alpha P_{sim}\vec{r}_{all} + (1-\alpha)\vec{r}_{sal}$

8) **until** $\vec{r}_{all}$ converge

9) Re-rank the images in descending order by $\vec{r}_{all}$ estimated above.

In our method, the random walk process is performed over a relevance graph to boost the performance of image re-ranking. Given the N images $\{x_1, x_2, \ldots, x_N\}$ returned by image search engines, we construct a graph with nodes being the image relevance values from the saliency mechanism and edges being the similarity measures from the consistency mechanism, as Fig. 4 illustrates. It is assumed that the graph has n nodes, each node corresponds to one image in the search result set, and the value of each node is its initial saliency relevance score. Transition matrix P_{sim} is to govern the transition of random walk process. Its element $p(x_i, x_j)$ denotes the transition probability from node x_i to node x_j. In addition, P_{sim} should be row normalized to 1:

$$p(x_i, x_j) = \frac{s(x_i, x_j)}{\sum_k s(x_i, x_k)} \tag{9}$$

where $s(x_i, x_j)$ is the similarity measure between images x_i and x_j from the consistency mechanism.

Then the proposed random-walk-based integrating method can be described as

$$r_k(x_j) = \alpha \sum_i r_{k-1}(x_i) p(x_i, x_j) + (1-\alpha) r_{sal}(x_j) \tag{10}$$

where $r_k(x_j)$ indicates the integrated score of node x_j at iteration k, $r_{sal}(x_j)$ is the relevance score from the saliency mechanism in (5), and α is the trade-off parameter with the aim to balance the saliency and similarity relevance. In the experiments, the value of α is set to 0.5.

The first term in (10) includes the state probabilities of node x_j's neighbors and their corresponding transition probabilities. The second term is the initial saliency score for node x_j. We update (10) recursively until all nodes in the graph converge. The stationary state probability of the random walk process is regarded as the final relevance score for the image. Re-ranking results are obtained by sorting the images according to their relevance scores in descending order.

TABLE I
SIXTY QUERY KEYWORDS

airplane	ant	beach	bee	bicycle	bird
cat	cliff	climbing	comet	conch	deer
desert	dish	diving	eagle	elephant	eyeglasses
fish	giraffe	grassland	guitar	helicopter	horse
iceberg	island	jumping	kayak	knife	ladybug
lake	mountain	mouse	ocean	owl	penguin
pearl	pigeon	plains	pyramid	riding	road
running	sailing	seashore	shark	sheep	ship
skating	swimming	teapot	tomato	train	turtle
umbrella	violin	volcano	walking	water lily	windmill

TABLE II
FUSION PARAMETER OF w_m

	Color	Edge	Texture	VS
w_m	0.256	0.231	0.245	0.268

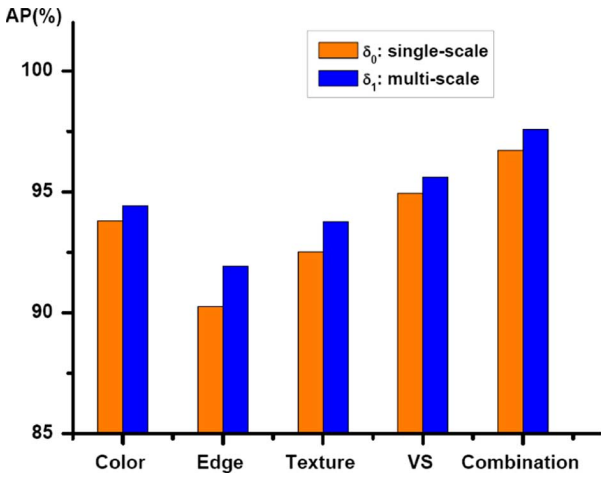


Fig. 5. Detection performance (AP) of four visual features and the combination of them. Orange denotes the AP values in the single-scale and blue shows the fusion results of three scale levels. It is noted that the AP of combination is much better than those of the four independent features.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present experimental results of the proposed image re-ranking algorithm. We will first show results of the saliency model and then show results of the entire re-ranking algorithm which integrate both the saliency model and the consistency model. For these experiments, we build two datasets:

1) *Image Set A*: In order to demonstrate the effectiveness of the saliency model, we construct a database of 6000 images, in which 3000 are for training, 1500 for validating, and 1500 for testing. These images are downloaded from a variety of sources, mostly from web photo albums and image search engines.
2) *Image Set B*: We have also collected a set of 38 274 images using Google and Yahoo Image Search on 60 query keywords. To facilitate the performance evaluation, we only use non-ambiguous concepts. The keywords of these queries, which include objects, scenes, and actions, are listed in Table I.

A. Experimental Results of the Multiscale Saliency Model

As in (1), since the saliency model is based on two classes, salient image class and cluttered image class, the image classifi-

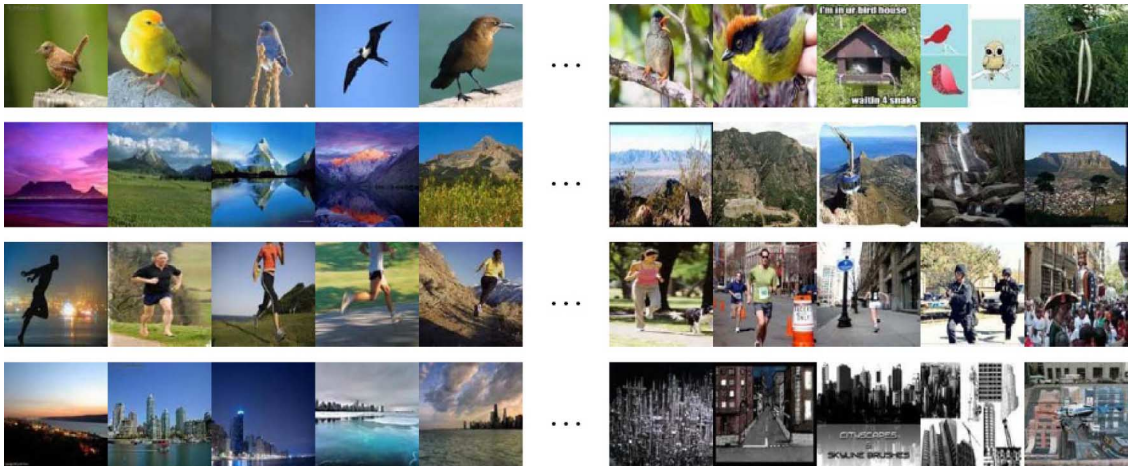


Fig. 6. The re-ranking results of the four web image search queries on "bird", "mountain", "running", and "cityscape". The first five images of each query (left) are more salient than the last five pictures (right).

cation results will be a good way to test the effectiveness of the model. This is because if an image is classified as a salient one with higher confidence, it normally has higher relevance score and higher ranks. Therefore, we manually label all the images of *Image Set A* into one of the two classes and compare the classification results on the testing set.

The classification can be performed as

$$x_i \; belongs \; to \begin{cases} \text{salient image class}, & \text{if } C(\pi_m) > th \\ \text{cluttered image class}, & \text{if } C(\pi_m) \leq th \end{cases} \tag{11}$$

where $C(\pi_m)$ is the same as in (6) and th is a threshold. In our experiments, th is set to be 0. The fusion parameter w_m in (6) is listed in Table II. w_m is determined by cross validation for balancing the importance among different feature sets.

As mentioned, our saliency model is based on the combination of four feature sets: color feature set, edge feature set, texture feature set, and VS feature set. In order to show the effect of each feature set, we perform classifications based on these feature sets, respectively, and compare them with the result which combines all the four feature sets. The average precision is compared to evaluate the classification results. The results are shown in Fig. 5. We can see that all of the feature sets have good precision results while the combination of all features obviously has the best result. This demonstrates that our proposed saliency model is effective in differentiating saliency of the image.

Furthermore, since the saliency model can examine each image and give it a saliency score, this score can also be used for image re-ranking with higher score images ranked at the top. This re-ranking result can be an effective way to evaluate the efficiency of our saliency model. Therefore, we perform another experiment. In this experiment, we directly use the saliency model to re-rank the images retrieved from an image search engine. We type the keywords "bird", "mountain", "running", and "cityscape" in Google image search engine and re-rank the first 50 returned images using our approach. Fig. 6 shows the first five images of each query (left) and the last five pictures (right) after re-ranking. It is noticeable that more salient images are ranked at the top with our saliency model.

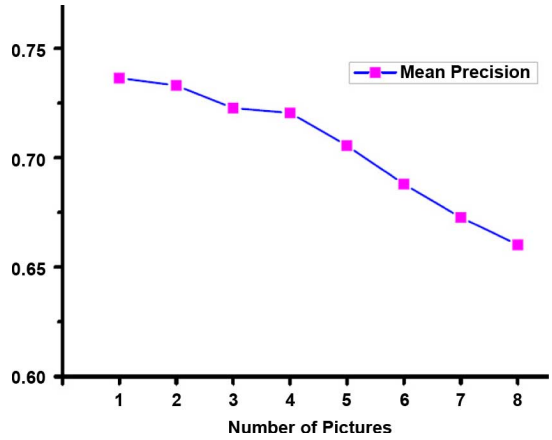


Fig. 7. Mean precision of the first N images.

Finally, we perform an eye-tracking experiment to verify our assumption in Section I and to demonstrate the effectiveness of the saliency model. Eye movements are recorded by Tobii T120 Eye Tracker. The eye-tracker records the position and duration of eye movements when navigating the images. In this experiment we use 20 queries in *Image Set B*. For each query, the first 200 images are re-ranked based on their saliency. In our algorithm, the images with higher saliency value will be ranked at the top. We put the first ten images after re-ranking as salient images and the last ten images as cluttered images. So for each query we obtain 20 images, which are resized to 128×128 (close to the size of thumbnails returned by image search engines). Then these 20 images are put randomly in a picture of 4×5 arrays.

A total of 22 participants took part in the experiment, including 14 males and 8 females. They have normal vision and have no knowledge of the experimental purpose. They are presented with a sequence of the 20 pictures for 4 s each, separated by displays of a blank screen for 3 s. The participants are not informed of what category will be displayed and are not given any specific tasks except being asked simply to look at the images. The eye-tracker records the first N images (N is from 1 to 8) they look at. We compute the proportion of salient images in the first N images and the results (mean precision) are showed

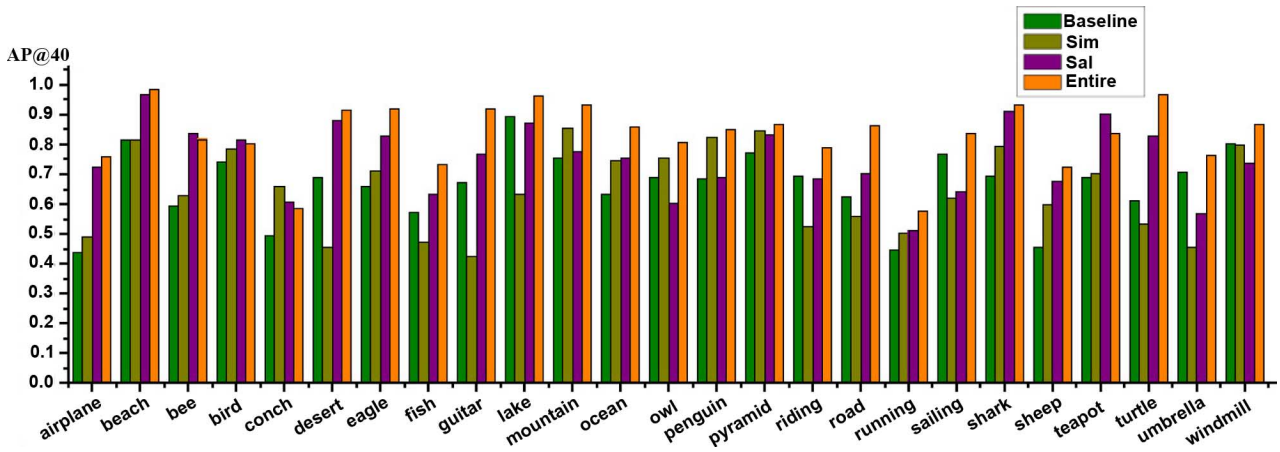


Fig. 8. Comparison of original image search results (Baseline), similarity-based re-ranking (Sim), saliency-based re-ranking (Sal), and our fusion results (Entire) for each query keyword.

in Fig. 7. It is noticed that the mean precisions are above 70% when N is smaller than 6. We believe such a result verifies the assumption that when users browse the result pages returned by an image search engine, they are more likely attracted by salient images than by cluttered ones.

B. Experimental Results for Integrating Both Saliency and Consistency Mechanisms

In the previous experiment, only the saliency mechanism is evaluated. In this section, we will evaluate the performance of our entire algorithm which integrates both visual saliency and consistency.

The entire re-ranking method is evaluated on a diversified dataset *(Image Set B)*. These images are first categorized into three classes: "Good", "Intermediate", and "Irrelevant" according to the judgment of three independent participants. Then we evaluate the re-ranking results by labeling the images into one of these predefined classes and calculate their average precision. These classes are defined as follows [5].

Good images: good examples which are related to the query and contain salient object(s) or high contrast region(s).

Intermediate images: images that are in some way related to the query, but of lower quality in that they may have major occlusion, serious noise, background clutter, or some other faults.

Irrelevant images: images that are totally unrelated to the query.

As the purpose of re-ranking is to give higher ranks for good images, when computing AP, good images are put as positive class, while intermediate and irrelevant images are put as negative classes. To reduce labeling inconsistency, we combine the results of the three participants and choose the median of the judged relevance as the ground truth label. Four methods are compared:

1) the original ranking given by the image search engines. (*Baseline* in Fig. 8 and Table III);
2) re-ranking the images using the similarity alone. (*Sim* in Fig. 8 and Table III);
3) re-ranking the images only using the saliency model. (*Sal* in Fig. 8 and Table III);
4) re-ranking the images by our entire algorithm which integrates both the saliency and consistency mechanisms. (*Entire* in Fig. 8 and Table III).

TABLE III
MEAN AP OVER ALL CONCEPTS

	MAP@10	MAP@20	MAP@30	MAP@40
$Baseline$	0.737	0.682	0.658	0.639
$Sim(200)$	0.693	0.635	0.613	0.605
$Sim(100)$	0.758	0.704	0.688	0.676
$Sal(200)$	0.779	0.745	0.733	0.725
$Sal(100)$	0.807	0.778	0.763	0.752
$Entire(200)$	0.855	0.830	0.809	0.784
$Entire(100)$	**0.872**	**0.843**	**0.816**	**0.797**

Since most users only focus their attention on the first few pages of the returned results, we choose the top 200 images of each query for the experiments. The average precision value in top 40 re-ranking results $(AP@40)$ is used to evaluate the performance. As space is limited, we show part of the results in Fig. 8. Since there are objects, scenes, and actions in the 60 queries, we arbitrarily select the queries from each of the three classes. From Fig. 8, we can note that both our saliency-based method and our entire method can give better results than the baseline while the entire method obviously performs the best.

Furthermore, several quantitative comparison results are shown in Table III. $Sal(200)$ and $Entire(200)$ denote using the saliency-based and the entire method, respectively, in the top 200 images of each query. $Sal(100)$ and $Entire(100)$ denote choosing the top 100 images for the experiments. Table III indicates that integrating the visual consistency is very effective in removing the noise images from the saliency-based results and further improves the results. Fig. 9 presents some examples of the original results from the search engine as well the re-ranked results by our entire method. The effectiveness of our algorithm can be obviously observed in Fig. 9.

We build the web image dataset from Google and Yahoo Image Search on 60 queries. It can be seen that the values of the baseline decrease gradually from $MAP@10$ to $MAP@40$ in Table III. This indicate an increase of negative examples when

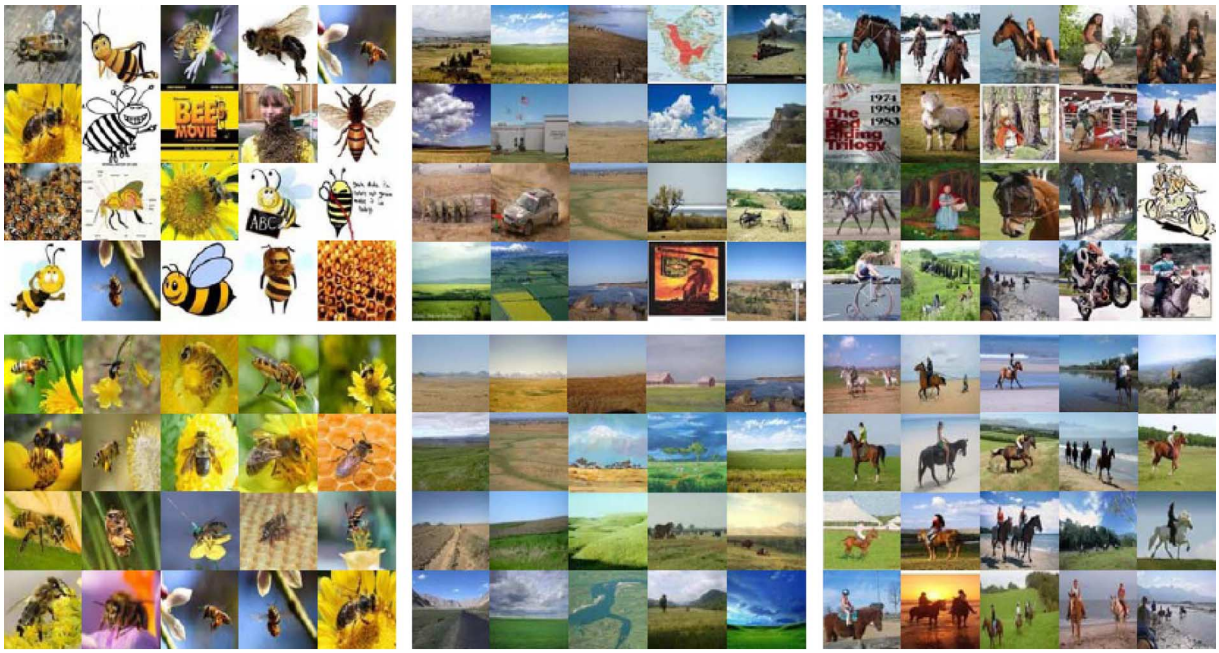


Fig. 9. Comparison of the original image search results (top) and the re-ranked results (bottom) by our entire algorithm. The search keywords for (a), (b), and (c) are "bee", "plains", and "riding", respectively.

more images are returned by these commercial search engines. It is also noticed that the values of $Sim(200)$, $Sal(200)$, and $Entire(200)$ are lower than those of $Sim(100)$, $Sal(100)$, and $Entire(100)$. We choose the top 200 images of each query for the experiments, the performance of which is significantly lower than that of the top 100 images.

As is shown in Fig. 8, the saliency-based re-ranking results are better than the baseline on most queries except for a few queries such as "owl" and "umbrella". The lower performances on these queries are because the search results contain lots of "noisy" images (i.e., salient but irrelevant images). However, by using our proposed method which integrates both the saliency and the consistency mechanisms, the results for these queries can be further improved.

In Table III, we also list the experimental results by only using visual similarity and without using visual saliency. It shows that we cannot get satisfactory re-ranking result by using visual similarity alone. Although there is slight improvement in $Sim(100)$ than the baseline, $Sim(200)$ is much worse than the baseline. The reasons may be as follows: the images in real-world web image dataset have large intra-class variations and there may be lots of noisy images. The visual similarity is effective in our proposed new framework which integrates visual saliency and visual consistency. However, if we only use visual similarity to re-rank the images, it would be too simple and not robust to deal with the large intra-class variances. Since such re-ranking on the basis of similarity is to some extent similar to clustering, it is difficult to control the center of the clusters by only using visual similarity. As shown in Fig. 8, for some queries such as "desert" and "riding", since the noise cluster are ranked at the top, the re-ranking results will be severely affected in the visual-similarity-based method. Therefore, it further verifies the effectiveness of including our proposed saliency model for re-ranking.

VII. Conclusion

In this paper, we propose a new re-ranking approach which integrates visual saliency and visual consistency mechanisms. The experimental results on a real-world web image dataset show that our approach can effectively detect the visually salient and consistent images and greatly improve user experiences. It is worth noting that in order to facilitate performance evaluation, we have only tested non-ambiguous concepts. In the future, we will extend the proposed method to deal with ambiguous concepts. Besides, the perceptual visual quality should be taken into account for image re-ranking. The images with higher saliency and higher quality will be ranked at the top.

References

[1] [Online]. Available: http://images.google.com.
[2] [Online]. Available: http://cn.bing.com/images.
[3] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for Google images," in *Proc. 8th Eur. Conf. Computer Vision (ECCV)*, 2004.
[4] T. Berg and D. Forsyth, "Animals on the web," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006.
[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's images search," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2005.
[6] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Reranking methods for visual search," *IEEE Multimedia*, vol. 14, pp. 14–22, Jul. 2007.
[7] N. Ben-Haim, B. Babenko, and S. Belongie, "Improving web-based image search via content based clustering," in *Proc. CVPR Workshop, SLAM*, 2006.
[8] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, pp. 219–227, 1985.
[9] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2004.
[10] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artif. Intell.*, vol. 146, pp. 77–123, May 2003.
[11] A. Bamidele, F. W. M. Stentiford, and J. Morphett, "An attention-based approach to content-based image retrieval," *BT Technol. J.*, vol. 22, pp. 151–160, Jul. 2004.

[12] H. Fu and Z. Chi, "Attention-driven image interpretation with application to image retrieval," *Pattern Recognit.*, vol. 39, pp. 1604–1621, Sep. 2006.
[13] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, pp. 97–136, 1980.
[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.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[15] P. J. Burt, "Fast filter transforms for image processing," *Comput. Vis., Graph., Image Process.*, vol. 16, no. 1, pp. 20–51, May 1981.
[16] M. Stricker and M. Orengo, "Similarity of color images," in *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1995.
[17] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognit.*, vol. 29, pp. 1233–1244, Aug. 1996.
[18] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray scale and rotation invariant texture analysis with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[19] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines, Software, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm.
[20] B. Tseng, C.-Y. Lin, M. Naphade, A. Natsev, and J. Smith, "Normalized classifier fusion for semantic visual concept detection," in *Proc. Int. Conf. Image Processing*, 2003.
[21] R. H. van Leuken, L. G. Pueyo, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proc. WWW*, Apr. 2009.
[22] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell., Special Issue on Face Recognition*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005.
[24] J. Wu and J. M. Rehg, "Where am I: Place instance and category recognition using spatial pact," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008.
[25] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. ACM Multimedia*, 2006.
[26] L. Wang, L. J. Yang, and X. M. Tian, "Query aware visual similarity propagation for image search reranking," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009.

Jun Huang is currently pursuing the Ph.D. degree at the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, China.

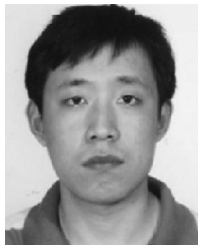
His current research interests include media analysis and retrieval, computer vision, and pattern recognition.

Xiaokang Yang (SM'04) received the B.Sci. degree from Xiamen University, Xiamen, China, in 1994, the M.Eng. degree from the Chinese Academy of Sciences, Shanghai, China, in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2000.

He is currently a Professor with the Institute of Image Communication and Information Processing, Department of Electronic Engineering, Shanghai Jiao Tong University. From April 2002 to October 2004, he was a Research Scientist in the Institute for Infocomm Research, Singapore. His current research interests include scalable video coding, video transmission over networks, video quality assessment, digital television, and pattern recognition.

Xiangzhong Fang received the M.S. degree from Shanghai University of Science and Technology, Shanghai, China, in 1993.

In 1996, he was a Senior Specialist in the National HDTV Technical Executive Experts Group (TEEG), Beijing, China. Since 2000, he has been a Professor at Shanghai Jiao Tong University. His main interests and work areas are image processing, multimedia communication, and video compression technology.



Weiyao Lin received the B.E. and M.E. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2005, respectively, and the Ph.D. degree from the University of Washington, Seattle, in 2010, all in electrical engineering.

Since 2010, he has been an Assistant Professor at the Institute of Image Communication and Information Processing, Department of Electronic Engineering, Shanghai Jiao Tong University. His research interests include video processing, machine learning, computer vision, and video coding and compression.

Rui Zhang received the B.S. and M.S. degrees from Hefei University of technology, Hefei, China, in 1995 and 1999, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2008.

Since 1999, she has worked at the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. Her research interests include image communication, image processing, and DTV.