

# Combining global, regional and contextual features for automatic image annotation

Yong Wang<sup>a,\*</sup>, Tao Mei<sup>b</sup>, Shaogang Gong<sup>a</sup>, Xian-Sheng Hua<sup>b</sup>

<sup>a</sup>Department of Computer Science, Queen Mary, University of London, London E1 4NS, UK

<sup>b</sup>Microsoft Research Asia, Beijing 100190, PR China

## ARTICLE INFO

### Article history:

Received 24 December 2007

Received in revised form 10 April 2008

Accepted 7 May 2008

### Keywords:

Global and regional features

Textual context

Cross media relevance model

Latent semantic analysis

Image annotation

## ABSTRACT

This paper presents a novel approach to automatic image annotation which combines *global*, *regional*, and *contextual* features by an extended cross-media relevance model. Unlike typical image annotation methods which use either global or regional features exclusively, as well as neglect the textual context information among the annotated words, the proposed approach incorporates the three kinds of information which are helpful to describe image semantics to annotate images by estimating their joint probability. Specifically, we describe the *global* features as a distribution vector of visual topics and model the textual context as a multinomial distribution. The global features provide the global distribution of visual topics over an image, while the textual context relaxes the assumption of mutual independence among annotated words which is commonly adopted in most existing methods. Both the global features and textual context are learned by a probability latent semantic analysis approach from the training data. The experiments over 5k Corel images have shown that combining these three kinds of information is beneficial in image annotation.

© 2008 Elsevier Ltd. All rights reserved.

## 1. Introduction

With the prevalence of digital imaging devices such as webcams, phone cameras and digital cameras, image data accessible to users are now explosively increased. An emerging issue is how to browse and retrieve this daunting volume of data. A possible solution is content based image retrieval, in which the query is usually given as a sample image or descriptions of visual properties [1–3]. However, such kind of query is not user-friendly enough, because in many cases a user's intent cannot be described only by an image or any low-level visual properties. Another approach is to annotate images and then retrieve these images by their associated textual keywords. If all the images are annotated, image retrieval can be solved effectively and efficiently by the well-developed techniques in text retrieval. Automatic image annotation is a process to automatically generate textual words to describe the content of a given image.

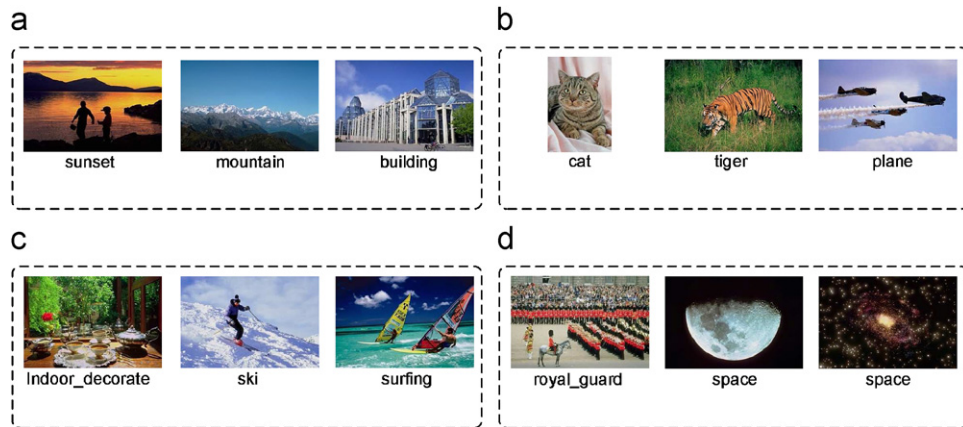
The research on automatic image annotation has proceeded along two categories. The first category poses image annotation as a supervised classification problem. Specifically, each word is viewed as a unique class. Binary classifiers for each class or a multiclass classifier is trained independently to predict the annotations of new images [4,5]. The second category represents the words and visual tokens in each image as features in different modalities. Image annotation is

then formalized by modeling the joint distribution of visual and textual features on the training data and predicting the missing textual features for a new image. The works for modeling this joint distribution include translation language model [6], cross-media relevance model (CMRM) [7], multiple Bernoulli relevance model (MBRM) [8], hidden conditional random fields (HCRF) [9], semantic distance [10], and so on.

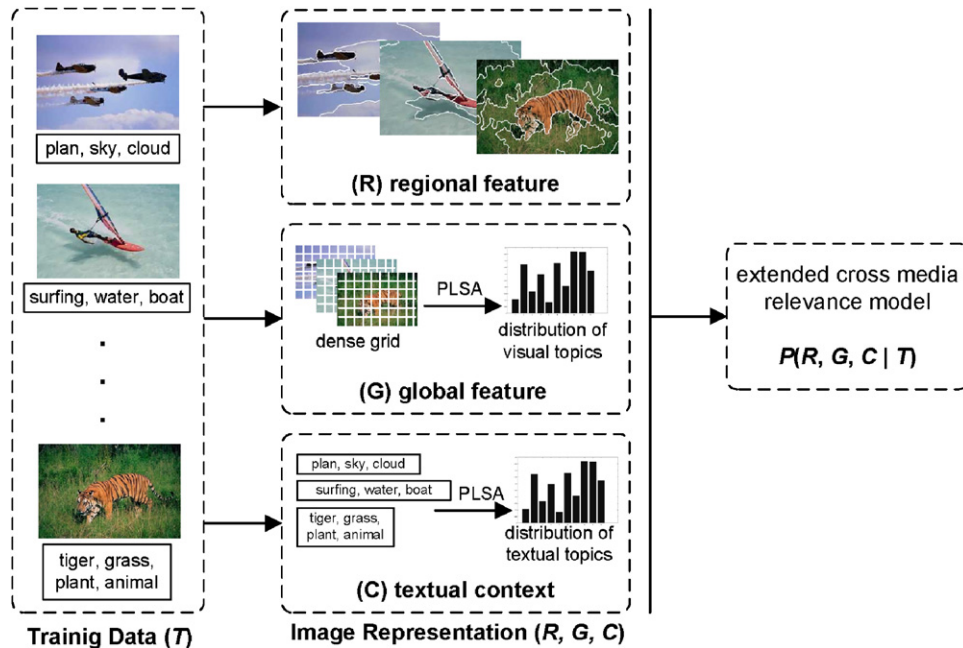
Both of the above two categories have shown good performances over different data sets. However, there still exist two challenging problems. First, most existing algorithms have taken one of two approaches, using either regional features [6,7,9,11] or global features [4,5,10] exclusively. In the approaches using global features [4,5,10], a global feature vector is extracted from an image, such as color histogram, color correlogram, edge direction histogram, and so on. The global features are advantageous in classifying simple scene categories such as "sunset," "mountain," "building," etc. In the approaches using regional features [6,7,9,11], an image is segmented into several regions and represented as a set of visual feature vectors, each of which represents one homogenous region. The underlying motivation of region-based image representation is that the appearance of many objects such as "cat," "tiger," and "plane" usually appear at a small portion of image. If a satisfying segmentation could be achieved, i.e., each object can be segmented as a homogenous and distinctive region, then region-based representation would be very meaningful. Since each of these two feature representations provides different kinds of information, they have their own advantages in classifying some certain categories. On the other hand, there are many situations where the annotation of images should be judged

\* Corresponding author.

E-mail addresses: [ywang@dcsc.qmul.ac.uk](mailto:ywang@dcsc.qmul.ac.uk) (Y. Wang), [tmei@microsoft.com](mailto:tmei@microsoft.com) (T. Mei), [sgg@dcsc.qmul.ac.uk](mailto:sgg@dcsc.qmul.ac.uk) (S. Gong), [xshua@microsoft.com](mailto:xshua@microsoft.com) (X.-S. Hua).



**Fig. 1.** The correspondence between semantics and feature representations. The semantic of an image is perceptually more relevant to: (a) global features, (b) regional features, (c) global and regional feature together, and (d) unclear. The images and their annotations are from Corel 5000 date set.



**Fig. 2.** Approach to image annotation by the extended cross-media relevance model. The extension include: (1) the global, regional, and contextual features are combined this extended model. (2) Instead of predicting the single word probability in CMRM, we first predict the context distribution and then predict the words distribution from the context distributions.

based on the combination of global and regional features. For example, the complex scene and events categories, e.g., "indoor\_decorate," "ski," and "surfing" are more suitable to be represented by both global and regional features. Fig. 1 shows some sample images. In addition, there exist some cases that it is not clear whether global or regional features are more perceptually used for annotation. We believe that the combination of these two types of features is beneficial in annotating images with such as diversity of categories.

Second, conventional approaches treat each word separately without considering the textual context relation. In other word, annotating one word to an image is independent from annotating another word to the same image. As a result, the textual context relations among annotation words have been ignored. By *textual context*, we refer to co-occurrence relationship among words. Such kind of contextual knowledge is actually embedded in the manual annotations, since human usually annotates an image with a set of

words with coherent semantic meaning as a whole entity rather than annotates each word one by one. For example, in "outdoor" images, if we have annotated "clouds," then the *a priori* probability of annotating "sky" would be higher than that of annotating "street" before examining the detailed visual content of the image. We believe that mining such kind of textual contexts from the training data is helpful for image annotations.

To address the above problems, we propose a novel extended CMRM for automatic image annotation. Unlike the typical CMRM [7] which only takes regional features to describe an image, the proposed extended CMRM incorporates both regional and global features, as well as textual context to annotate images by estimating their joint probability. Specifically, we describe the global features as a distribution vector of visual topics, and model the textual context among annotated words as a multinomial distribution of words. The global features provide the global distribution of visual topics

over an image, while the textual context relaxes the assumption of mutual independence among annotated words which is commonly adopted in most existing approaches. Both the global features and textual context are learned by a probability latent semantic analysis (PLSA) [12] approach from the training data. We should point out that Jin et al. [13] also proposed to model the textual contexts among words as multinomial distributions of words. However, the learning of textual contexts in their approach is achieved simultaneously in the learning of the joint distribution of visual words and textual words by an expectation–maximization (EM) algorithm. In our approach, the learning of the textual contexts is independent from the visual features. Moreover, Jin et al. [13] use only regional features, while our approach considers both regional and global features. Although Lisin et al. [14] also investigated combining local and global features for object recognition, these two kinds of features are actually used independently. Furthermore, the context information is not considered.

The framework is shown in Fig. 2. The images are represented by global, regional, and contextual features (i.e., G, R, and C). An extended CMRM model is learned from the training data based on the three kinds of features. Given a new image  $I$  to be annotated, we first compute the textual context  $P(C|I)$  of image  $I$  based on the extended CMRM model, and then the textual context distribution is fused with the words distribution obtain the annotation words  $w$ , i.e.,  $P(w|I)$ .

The rest of this paper is organized as follows. Section 2 introduces the original CMRM. In Section 3 we discuss how to learn textual contexts and visual topics, as well as how these three kinds of features are integrated into an extended CMRM. Section 4 give experimental results, followed by conclusion in Section 5.

## 2. Cross-media relevance model

The CMRM [7] is a non-parametric model for image annotation. CMRM represents an image as a bag of regions and annotates a test image  $I$  by estimating the joint probability of a word  $w$  and its visual blobs, i.e.,

$$P(w, b_1, \dots, b_m) = \sum_{J \in \mathcal{T}} P(J) P(w, b_1, \dots, b_m | J) \quad (1)$$

where  $(b_1, \dots, b_m)$  are the blobs of the test image  $I$ ,  $J$  represents a training image.  $P(J)$  is kept uniform over all images in  $\mathcal{T}$ . CMRM assumes that given an image, the events of observing a word  $w$  and  $b_1, \dots, b_m$  are mutually independent, so that  $P(w, b_1, \dots, b_m | J)$  can be simplified as

$$P(w, b_1, \dots, b_m | J) = P(w | J) \prod_{i=1}^m P(b_i | J) \quad (2)$$

where  $P(w | J)$  and  $P(b | J)$  are given by the following smoothed probabilities:

$$P(w | J) = (1 - \alpha) \frac{\#(w, J)}{|J|} + \alpha \frac{\#(w, \mathcal{T})}{|\mathcal{T}|} \quad (3)$$

$$P(b | J) = (1 - \beta) \frac{\#(b, J)}{|J|} + \beta \frac{\#(b, \mathcal{T})}{|\mathcal{T}|} \quad (4)$$

where  $\#(w, J)$  represents the number of occurrence of word  $w$  in  $J$ ,  $\#(w, \mathcal{T})$  represents the number of occurrence of  $w$  in the whole training set.  $|J|$  and  $|\mathcal{T}|$  represent the number of aggregated blobs and words for one image  $J$  and for the whole training set  $\mathcal{T}$  respectively. The notation of  $\#(b, J)$  and  $\#(b, \mathcal{T})$  are similar to  $\#(w, J)$  and  $\#(w, \mathcal{T})$ .  $\alpha$  and  $\beta$  are two smooth parameters.

## 3. Extended CMRM

The conventional CMRM considers only one representation of images, i.e., a bag of blobs. To deal with images which are not suitable to be represented as a bag of blobs, we need to consider other representation as well. In this Section, we propose to use visual topics as another representation. This new image representation is combined with the visual blob representation. Moreover, from Eq. (1) we know that CMRM annotates keywords individually without considering the joint distribution of different keywords. To remedy this problem, we propose textual context to model the joint distribution of different keywords. To annotate an image with multiple keywords, we first annotate the image with textual contexts and then compose the keywords from the typical distribution of keywords under each textual contexts.

### 3.1. Learning textual contexts from prior annotations

Put a simple way, *textual context* is the co-occurrence relationship among different keywords. This relationship may come from the habit in which people use a language, for example, "united" and "nation." It may also come from the correlation in semantic meaning of different keywords, for example, "sky" and "clouds." Generally, it is hard to build a universal mathematical model for textual context. Here we simplify a particular textual context as a multinomial distribution of textual keywords. Specifically, we assume there are a limited number of textual contexts and the keyword distribution  $P(w_j | J)$  of an image  $J$  is governed by the hidden conditional distribution of textual context. The learning textual contexts is based on the probabilistic latent semantic analysis (PLSA) developed by Hofmann [12]. PLSA is proposed to automatically learn topics from text documents. Suppose we are given a set of text documents  $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$ , each of which is represented by a term frequency vector, i.e.,

$$d_i = [n(d_i, w_1), n(d_i, w_2), \dots, n(d_i, w_m)] \quad (5)$$

where  $n(d_i, w_j)$  is the number of occurrence of word  $w_j$  in document  $d_i$ , and  $m$  is the vocabulary size. PLSA assumes that each word in a document is generated by a specific *hidden topic*  $z_k$ , where  $z_k \in \mathcal{Z}$  and  $\mathcal{Z}$  is the vocabulary of hidden topics. Since  $z_k$  is a hidden variable, the conditional probability of a word  $w_j$  given document  $d_i$  is a marginalization over the topics, i.e.,

$$P(w_j | d_i) = \sum_k^K P(w_j | z_k, d_i) P(z_k | d_i) \quad (6)$$

where  $K$  is the number of hidden topics,  $P(w_j | z_k, d_i)$  is the conditional probability of a word  $w_j$  given topic  $z_k$  and the document  $d_i$ ,  $P(z_k | d_i)$  is the conditional probability of topic  $z_k$  given  $d_i$ . Furthermore, PLSA assumes that the conditional probability of generating a word by a specific topic is independent from the document, i.e.,

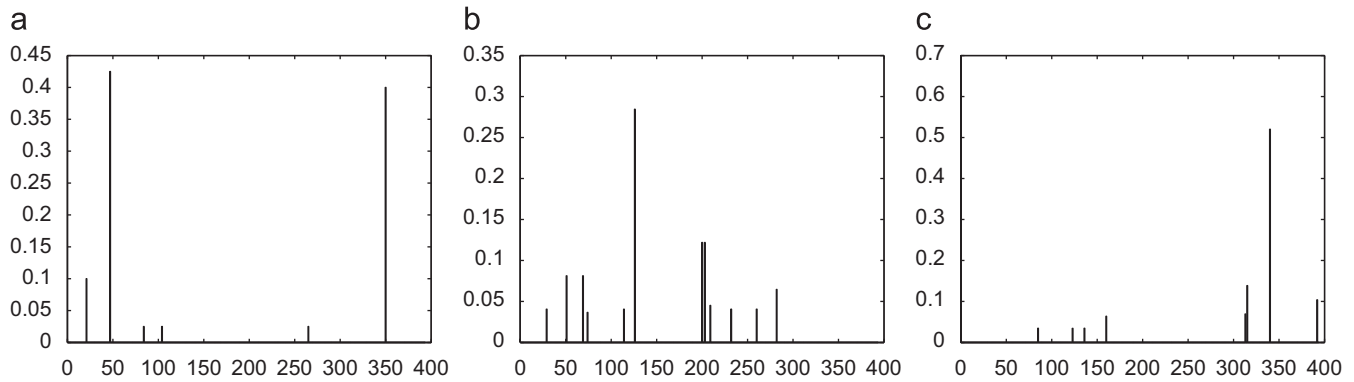
$$P(w_j | z_k, d_i) = P(w_j | z_k) \quad (7)$$

Therefore, Eq. (6) can be simplified as

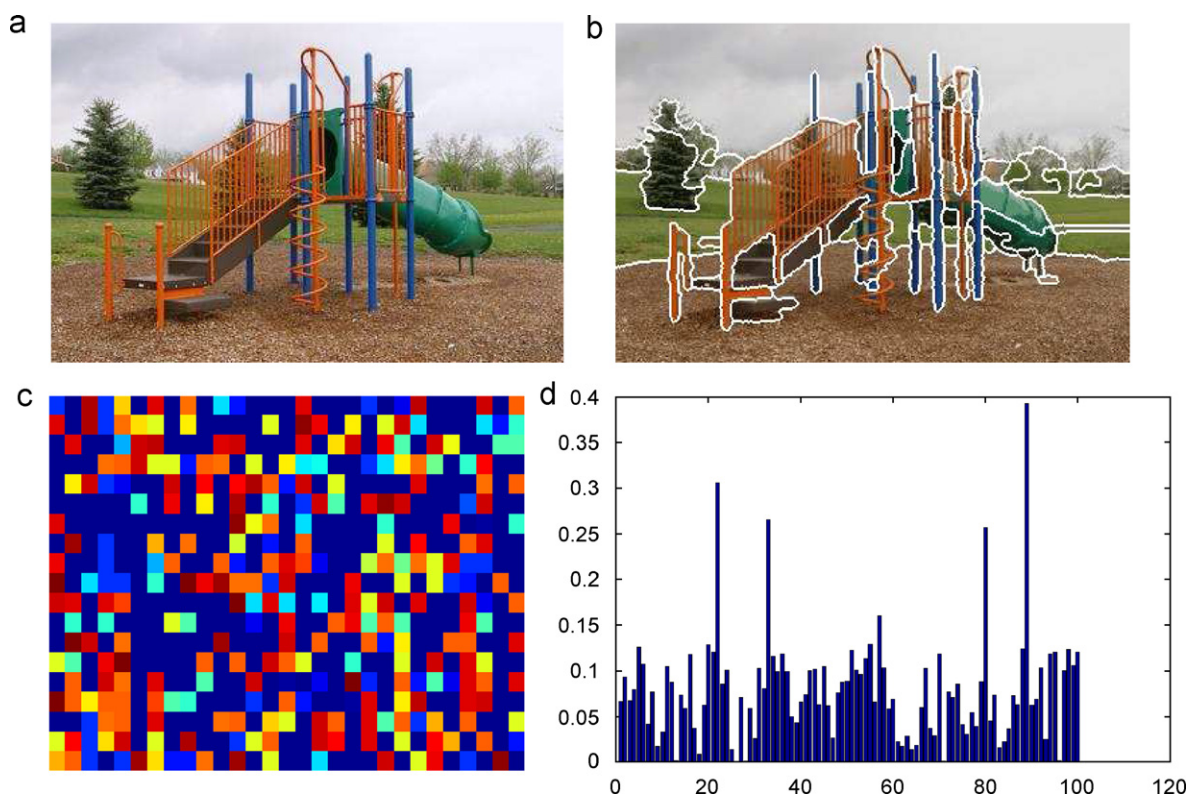
$$P(w_j | d_i) = \sum_k^K P(w_j | z_k) P(z_k | d_i) \quad (8)$$

The model parameters  $P(w_j | z_k)$  and  $P(z_k | d_i)$  can be learned by an EM algorithm [12]. Given the learned model parameters and a new document  $d$ , the topic distribution  $\{P(z_k | d)\}_{k=1}^K$  of  $d$  can be estimated by an EM algorithm similar to the training process [12].

To apply PLSA to the annotated words associated with the training images, we take each group of keywords annotated to a trainin



**Fig. 3.** Illustration of three textual contexts. The three histogram plots show the words distribution of each textual contexts. The words under each plot are the three words with largest probabilities under the corresponding textual context. (a) cat, tiger, Bengal; (b) food, market, Maui; (c) street, sign, writing.



**Fig. 4.** Illustration of two types of features: (a) original image; (b) regions by unsupervised segmentation and (d) global distribution of visual topics. The segmentation of each image is independent of each other while the global distribution of visual topic is learned automatically from the whole set of training images. The patches in (c) belonging to the same topic are indicated by the same color. Each patch corresponds to a region of  $13 \times 13$  pixel.

image as a short text document and learn a number of hidden topics from this collection of short text documents. We call these topics, which are described by multinomial distributions of keywords, as textual contexts. Fig. 3 shows three samples of textual contexts learned from the training set of Corel images. We find that textual context is able to group the words into different semantic meaning, such as "tiger" and "Bengal."

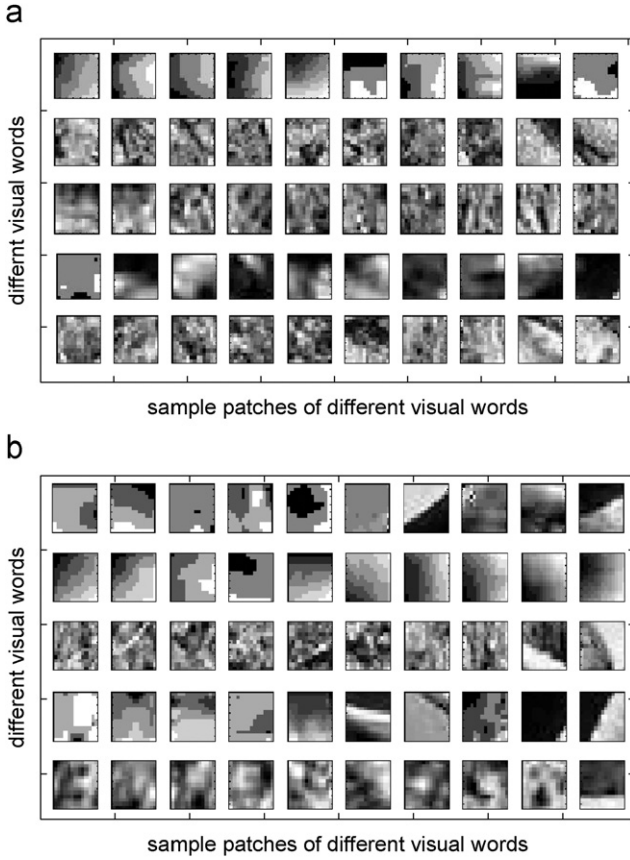
Specifically, the textual context will be used in the following way. (1) From the training data, we extracted a number of textual contexts from the manual annotation on the training images. (2) Given a test image, we "annotated" this image with textual contexts, i.e., on a test image, the textual contexts were derived from the visual features rather than from the pre-annotated keywords. (3) Compose the keywords from the textual contexts distri-

bution on the test image. We will give more details in the following sections.

### 3.2. Learning visual topics from images

Similar to learning topics from text documents, we can also learn visual topics from a collection of images. The main point is representing an image as a bag of "words," similar to the vector representation of text documents. In details, we partition an image by a regular grid and take it as an unordered set of image patches. Then we extract a 128-D SIFT descriptor [15] and vector-quantize each image patch by clustering a subset of patches from the training images, which has proved effective for object recognition [16]. We call the set of cluster centers as visual vocabulary. We can then trans-





**Fig. 5.** Illustration of two different visual topics. Visual topic can group patches with different visual appearance by their co-occurrence relationship. For each visual topic, we show the top five visual words with largest probabilities. The patches on each line is some sample patches belonging to the corresponding visual words.

form an image into a bag of visual words by assigning a visual word label to each image patch. Given this bag of visual words representation, it is then straightforward to apply PLSA to learn a set of visual topics, each of which is characterized by a multinomial distribution of visual words.

It is worth comparing visual topics to image regions. From the probabilistic model of PLSA we can find that PLSA models an image as an *unordered* set of image patches and a visual topic groups image patches by *co-occurrence* relationship. The *unorderness* suggests that the image patches grouped in a topic do not have spatial agglomeration property. The *co-occurrence* suggests that the image patches grouped in a topic do not necessarily have visual consistency. This is rather different from image regions obtained by image segmentation since image segmentation groups pixels by its visual property and spatial location. Fig. 4 illustrates the image regions and the global distribution of visual topics of an image. In Fig. 5 we show some sample patches in two different visual topics. Since the visual topics and image regions by segmentation focus on different aspects of an image, they are complementary to each other and a combination of them is expected to achieve better performance.

### 3.3. Combining global, regional, and contextual features

Our method annotates a test image  $I$  by estimating the joint probability of a textual context  $c$  learned in Section 3.1, its visual blobs (regions)  $R = (b_1, \dots, b_m)$  obtained by image segmentation, and the visual topic distribution  $H(I)$  learned in Section 3.2, i.e.,

$$P(c, R, H(I)) = \sum_{j \in \mathcal{T}} P(j) P(c, b_1, \dots, b_m, H(I) | j) \quad (9)$$

Comparing Eq. (9) with Eq. (1), there are two points of difference to elaborate. First, the original CMRM in Eq. (1) annotates an image using only the regional features  $R = (b_1, \dots, b_m)$ . However, our extended model in Eq. (9) uses both the regional features  $R$  and the global features  $H(I)$ , which represent the global distribution of visual topics in image  $I$ . This suggests our new model combines the global features and regional features. Second, Eq. (1) predicts the probability of a single word  $w$  directly, while Eq. (9) predicts the probability of a textual context  $c$ . This indicates our model does not assume the mutual independence between words given the image. Thus, the extended CMRM incorporates the textual context from the training data.

We assume the mutual independence between a textual context  $c$ , image blobs and the visual topic distribution, so that  $P(c, b_1, \dots, b_m, H(I) | j)$  can be simplified as

$$P(c, b_1, \dots, b_m, H(I) | j) = P(c | j) P(H(I) | j) \prod_{i=1}^m P(b_i | j)$$

$P(b | j)$  is estimated as the same as that in Eq. (3).  $P(c | j)$  is readily available after learning textual contexts on the manual annotations as described in Section 3.1.  $P(H(I) | j)$  is defined as the Kullback–Leibler divergence between the visual topic distribution of  $I$  and  $j$ , i.e.,

$$P(H(I) | j) = D_{KL}(H(I) | H(j)) = \sum_{i=1}^Q P(q_i | I) \log \frac{P(q_i | I)}{P(q_i | j)} \quad (10)$$

where  $D_{KL}$  is the Kullback–Leibler divergence between two distributions.

From the Bayesian theory, we know that,

$$P(c | I) = \frac{P(c, I)}{P(I)} = \frac{P(c, b_1, \dots, b_m, H(I))}{P(I)} \quad (11)$$

Therefore, a normalization on  $P(c, b_1, \dots, b_m, H(I))$  will give us the conditional distribution of textual contexts  $P(c | I)$ . The conditional keyword distribution  $P(w_j | I)$  of  $I$  is obtained by fusing the keyword distribution of all the textual contexts, i.e.,

$$P(w_j | I) = \sum_i^S P(w_j | c_i) P(c_i | I) \quad (12)$$

## 4. Experiments

To evaluate the performance of our approach, we test it on a 5460 Corel image data set, among which 3100 images is a subset of the 5000 Corel image data set [7] and another 2360 images are downloaded from Ref. [17]. We have not used the original 5000 image data set [7] mainly because we do not have all the original images. Although the blob features of the 5000 Corel image data set [7] can be

**Table 1**

The average precision, recall, and  $F_1$  values of Experiment I

Parameters	Metric	R	R + C	R + G	R + C + G
Q = 80, S = 100	Precision	0.2317	0.2386	0.2642	0.2665
	Recall	0.2786	0.2715	0.2782	0.2950
	$F_1$	0.2530	0.2540	0.2710	0.2800
Q = 100, S = 80	Precision	0.2264	0.2586	0.2577	0.2849
	Recall	0.2866	0.2959	0.3238	0.3472
	$F_1$	0.2530	0.2760	0.2870	0.3130
Q = 100, S = 120	Precision	0.2447	0.2571	0.2669	0.2837
	Recall	0.2619	0.2734	0.3104	0.3321
	$F_1$	0.2530	0.2650	0.2870	0.3060

In this experiment, some images have only one or two annotation words. Q represents the number of visual topics. S represents the number of textual contexts.

downloaded from the Internet, we still have to use the original images extract the global features. The number of different keywords on this data set is 393. Since the textual contexts represent the correlation among keywords, they are only meaningful when there are

multiple keywords in the annotations. In our 5460 Corel images, there are a lot of images with only one keyword annotated. So we exclude images with very few annotation keywords to fairly evaluate the influence of textual contexts. Thus we have designed two experiments as follows.

**Table 2**









The average precision, recall, and  $F_1$  values of Experiment II

Parameters	Metric	$R$	$R + C$	$R + G$	$R + C + G$
$Q = 80, S = 100$	Precision	0.3027	0.3508	0.3472	0.3553
	Recall	0.3486	0.3593	0.3673	0.3882
	$F_1$	0.3240	0.3550	0.3570	0.3710
$Q = 100, S = 80$	Precision	0.2881	0.3438	0.3623	0.3624
	Recall	0.3702	0.3845	0.3780	0.4060
	$F_1$	0.3240	0.3630	0.3700	0.3830
$Q = 100, S = 120$	Precision	0.2938	0.3538	0.3471	0.3607
	Recall	0.3611	0.3812	0.3961	0.3949
	$F_1$	0.3240	0.3670	0.3700	0.3770

In this experiment, all the images have at least three annotation words.  $Q$  represents the number of visual topics.  $S$  represents the number of textual contexts.

- *Experiment I*: We partition the whole data set into a training set and a test set, keeping the ratio between the training size and testing size as 9/1. This is equivalent to 4914 training images and 546 testing images. This experiment is to evaluate the performance of incorporating global features.
- *Experiment II*: We select only images with more than three keywords in their manual annotations. This is equivalent to 3192 training images and 355 test images. This experiment is to evaluate the performance of incorporating textual contexts and global features.

For the region features, we use the JSEG algorithm [18] to segment each image into 1–11 regions. Image regions with area less than 1/25 of the whole image are discarded. In average there are five

Image	Ground truth	CMRM	Our approach
	flowers, leaf, petals, stems	flowers, grass, field, tree, buildings	flowers, leaf, garden, grass, tree
	castle, shrubs, sky, water	sky, water, grass, tree, stone	sea, water, sky, tree, mountain
	beach, people, sand, water	beach, tree, water, mountain, snow	beach, water, sand, sea, mountain
	cars, prototype, tracks, turn	cars, stone, sky, field, sea	cars, tracks, turn, grass, stone
	cars, prototype, tracks, turn	cars, stone, sky, field, sea	cars, tracks, turn, grass, stone
	flowers, garden, house, window	flowers, grass, sky, field, people	flowers, garden, grass, house, stone
	mosque, pillar, stone, temple	stone, building, sand, beach, sky	stone, temple, pillar, ruins, sky
	frost, ice, sky, tree	sea, water, tree, grass, sky	sky, tree, forst, ice, grass

**Fig. 6.** Sample images and the annotations by CMRM and the proposed approach.

image regions per image. Each image region is represented by a 49-D feature vector including 1-D relative region area, 9-D color moment feature, 3-D shape descriptor and 36-D color correlogram feature [19,20]. For the dense grid, we sample  $13 \times 13$  pixels image patches without overlapping [16]. The average number of image patches per image is around 550. The image regions are clustered into 500 image blobs. Similarly, the SIFT descriptors of image patches are clustered into 500 centers. We experiment with different number of visual topics  $Q$  and different number of textual contexts  $S$  and compare the performances of the following approaches.

- (1)  $R$ : The original CMRM which annotates images using regional features and without textual contexts. Its performance is not affected by the number of visual topics and number of textual contexts.
- (2)  $R + C$ : The extended CMRM which annotates images using the regional features with textual contexts. Its performance is not affected by the number of visual topics.
- (3)  $R + G$ : The extended CMRM which annotates images using regional features and global features. Its performance is not affected by the number of textual context.
- (4)  $R + C + G$ : The final extended CMRM proposed in this paper which annotates images by fusing global features and regional features and incorporating textual contexts.

For each method, we take the top five words as the final annotation [5]. We evaluate the annotation performance by the average precision, recall, and  $F_1$  values over all testing images. The  $F_1$  value is derived from precision and recall values as in Eq. (13):

$$\begin{aligned} \text{precision}(w) &= \frac{\# \text{ of images correctly annotated with } w}{\# \text{ of images automatically annotated with } w} \\ \text{recall}(w) &= \frac{\# \text{ of images correctly annotated with } w}{\# \text{ of images manually annotated with } w} \\ F_1(w) &= \frac{2 \times \text{precision}(w) \times \text{recall}(w)}{\text{precision}(w) + \text{recall}(w)} \end{aligned} \quad (13)$$

The results of Experiment I are shown in Table 1. When choosing the number of visual topics  $Q$  as 80 and the number of textual contexts  $S$  as 100, the results show that, compared to the original CMRM ( $R$ ), incorporating global features ( $R + G$ ) improved the  $F_1$  value from 0.253 to 0.271. However, incorporating textual contexts only ( $R + C$ ) does not show much improvement (from 0.253 to 0.254). This is because in the Experiment I, many images have only one or two annotation words, making the textual contexts insignificant. The best performance is achieved by incorporating both textual contexts and global features ( $R + C + G$ ), which improves the performance from 0.253 to 0.280. The other settings of  $Q$  and  $S$  show the similar behavior. The results of Experiment II are shown in Table 2. For  $Q = 80$  and  $S = 100$ , incorporating textual contexts only ( $R + C$ ) improve the  $F_1$  value from 0.324 to 0.355. Compared to Experiment I, the improvement is more significant. Incorporating global features only ( $R + G$ ) produce the similar improvement as ( $R + C$ ). Again, the best performance is achieved by incorporating both global features and textual contexts ( $R + C + G$ ), which improved the  $F_1$  value from 0.324 to 0.371. In summary, incorporating both textual contexts and global features can improve the annotation performance compared with the original CMRM. The performance gain contributed by textual contexts is more significant in Experiment II, because each image has more than three keywords annotated.

Some test images with the annotations generated by the CMRM and the proposed approach are shown in Fig. 6. It is observed that the proposed approach can yield better annotations than CMRM, especially when the test images are associated with multiple words. The annotations from the proposed approach have better coherent

semantic due to the integration of textual context, global, and regional appearances into CMRM.

## 5. Discussion

In this paper, we have proposed a method for automatic image annotation which is extended from typical cross-media relevance model. The proposed approach takes into account both the regional features and the global features, as well as textual context. To obtain a more representative global features, we learn a number of visual topics from the training image set by the probabilistic latent semantic analysis (PLSA). Moreover, PLSA is used to model the textual contexts between annotation words. We tested the proposed approach on a 5460 Corel image data set. The experimental results on a 5460 Corel image data set show that, in the general case of annotating images with few words (i.e., less than five words), the combination of the regional features and the global features can improve the annotation performance. In the case of annotating images with multiple words (at least three words), the incorporation of textual contexts can significantly improve the annotation performance.

As we have mentioned in Fig. 1 that different features have different contributions to a specific word. In the future work, we will aim to investigate the different influence of these three features for image annotation, especially for specific categories. Furthermore, we will exploit the correlations of words in a multilabel setting.

## References

- [1] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, *ACM Comput. Surv.* 40 (65) (2008).
- [2] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (12) (2000) 1349–1380.
- [3] D. Tao, X. Li, S.J. Maybank, Negative samples analysis in relevance feedback, *IEEE Trans. Knowl. Data Eng.* 19 (4) (2007) 568–580.
- [4] E. Chang, K. Goh, G. Sychay, G. Wu, CBSA: CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003, pp. 26–38.
- [5] J. Li, J.Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (9) (2003) 1075–1088.
- [6] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, M.I. Jordan, Matching words and pictures, *J. Mach. Learn. Res.* 3 (2003) 1107–1135.
- [7] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in: *Proceedings of SIGIR*, 2003, pp. 119–126.
- [8] S.L. Feng, R. Manmatha, V. Lavrenko, Multiple bernoulli relevance models for image and video annotation, in: *Proceedings of CVPR*, June 2004, pp. 1002–1009.
- [9] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, Z. Wang, Joint multi-label multi-instance learning for image classification, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [10] T. Mei, Y. Wang, X.-S. Hua, S. Gong, S. Li, Coherent image annotation by learning semantic distance, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [11] D.M. Blei, M.I. Jordan, Modeling annotated data, in: *Proceedings of SIGIR*, 2003, pp. 127–134.
- [12] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, *Mach. Learn.* 42 (1–2) (2001) 177–196.
- [13] R. Jin, J.Y. Chai, L. Si, Effective automatic image annotation via a coherent language model and active learning, in: *Proceedings of ACM Multimedia*, 2004, pp. 892–899.
- [14] D.A. Lusin, M.A. Mattar, M.B. Blaschko, E.G. Learned-Miller, M.C. Benfield, Combining local and global image features for object class recognition, in: *Workshop on Learning in Computer Vision and Pattern Recognition at IEEE CVPR*, 2005.
- [15] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision* 60 (2) (2004) 91–110.
- [16] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524–531.
- [17] J. Li (<http://www.stat.psu.edu/~jiali/index.download.html>).
- [18] Y. Deng, B.S. Manjunath, Unsupervised segmentation of color-texture regions in images and video, *IEEE Trans. PAMI* 23 (8) (2001) 800–810.
- [19] Y. Chen, J.Z. Wang, Image categorization by learning and reasoning with regions, *J. Mach. Learn. Res.* 5 (2004) 913–939.
- [20] G.-J. Qi, X.-S. Hua, Y. Rui, T. Mei, J. Tang, H.-J. Zhang, Multiple instance learning for image categorization, in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

**About the Author**—YONG WANG is currently a Ph.D. student in computer vision at the Department of Computer Science in Queen Mary, University of London. He received his B.E. degree in automation from the University of Science and Technology of China (USTC) in 2001 and his M.Sc. in computer engineering from the National University of Singapore in 2003.

**About the Author**—TAO MEI received the B.E. degree in automation and Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively.

He joined Microsoft Research Asia as an Associate Researcher, Beijing, China, in 2006. His current research interests include image and video content analysis, computer vision, pattern recognition, and online multimedia applications such as multimedia search, advertising, recommendation, and presentation. He has authored over 40 publications in these areas and has more than 10 filed patents or pending applications. He received the ACM SIGMM 2007 Best Paper and Best Demonstration Awards. Dr. Mei is a member of the ACM and IEEE.

**About the Author**—SHAOGANG GONG is Professor of Visual Computation at the Department of Computer Science, Queen Mary, University of London, and a Member of the UK Computing Research Committee. He heads the Queen Mary Computer Vision Group and has worked in computer vision and pattern recognition for over 20 years, published over 170 papers and a monograph. He twice won the Best Science Prize (1999, 2001) of British Machine Vision Conferences, the Best Paper Award (2001) of IEEE International Workshops on Recognition and Tracking of Faces and Gestures, and the Best Paper Award (2005) of IEE International Symposium on Imaging for Crime Detection and Prevention. He was a recipient of a Queens Research Scientist Award (1987), Royal Society Research Fellow (1987, 1988), a GEC-Oxford Fellow (1989), a visiting scientist at Microsoft Research (2001) and Samsung (2003).

**About the Author**—XIAN-SHENG HUA received the B.S. and Ph.D. degrees from Beijing University, Beijing, China, in 1996 and 2001, respectively, both in Applied Mathematics.

When he was in Peking University, his major research interests were in the areas of image processing and multimedia watermarking. Since 2001, he has been with Microsoft Research Asia, Beijing, where he is currently a Lead Researcher with the internet media group. His current interests are in the areas of video content analysis, multimedia search, management, authoring, sharing, and advertising. He has authored more than 100 publications in these areas and has 30 filed patents or pending applications. He serves in the Editorial Boards of Multimedia Tools and Applications and IEEE Transactions on Multimedia. Dr. Hua is a member of the Association for Computing Machinery and IEEE.