Neural Process Lett (2017) 45:777–789
DOI 10.1007/s11063-016-9511-4

Cross-Modal Saliency Correlation for Image Annotation

Yun Gu[1] · **Haoyang Xue**[2] · **Jie Yang**[2]

Published online: 2 March 2016

Abstract Automatic image annotation is an attractive service for users and administrators of online photo sharing websites. In this paper, we propose an image annotation approach exploiting the crossmodal saliency correlation including visual and textual saliency. For textual saliency, a concept graph is firstly established based on the association between the labels. Then semantic communities and latent textual saliency are detected; For visual saliency, we adopt a dual-layer BoW (DL-BoW) model integrated with the local features and salient regions of the image. Experiments on MIRFlickr and IAPR TC-12 datasets demonstrate that the proposed method outperforms other state-of-the-art approaches.

Keywords Image annotation · Visual saliency · Textual saliency

1 Introduction

With the explosive growth of web images, image annotation has drawn wide attentions in recent years. Given an image, the goal of image annotation is to analyze its visual content and assign the labels to it. Numerous approaches have been proposed for automatic image annotation. In recent years, great research effort has been devoted to automatic image annotation [11,16,19,20,23,24,26–28,30,32,33]. In general, approaches for image annotation can be classified into two categories: learning-based and search-based annotation [16]. In search-

✉ Jie Yang
jieyang@sjtu.edu.cn

Yun Gu
geron762@sjtu.edu.cn

Haoyang Xue
xuehaoyangde@sjtu.edu.cn

[1] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

[2] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China

based annotation, the labels are directly provided and annotated by utilizing images in the database. The k-nearest neighbor (KNN) search (including the extended algorithms) is widely used because of its simplicity and good performance with large scale data [16,23,26,30]. For learning-based methods, the annotation problem can be considered a multi-class classification that predicts one label from a set of exclusive labels, or a binary classification that makes a binary decision on each label independently. In previous work, researchers applied machine learning methods such as the support vector machine (SVM) to the annotation problem [4,5,7,21] and showed its good performance with high dimensional data. In traditional image annotation problems, the number of classes or labels is always limited and samples of each class are often uniform. This can be considered as a classification problem. However, there are more than hundreds of labels (even millions) in an online image dataset like Flickr. Since each image can be tagged with many labels, this problem is no longer compatible with a traditional classification model. Both search-based methods and learning-based methods are demonstrated with good performance on state-of-art datasets. However, most of them focus on learning with pre-extracted features while some works are dealing with the visual representation. Some works try to construct robust model [3] which learns the probability distribution of a semantic class from images with weakly labeled information. In [15], the images are coded with sparse features via over-segmentation for label-to-region annotation. In most recent works, [22] proposed an image annotation approach which selects some relevant tags with diverse semantics. Wu et al. [29] focus on missing label problems which add two regularizations to keep inter and intra class smoothness. Lu and Wang [18] learns the mapping between bag-of-visual-salient features and noise labels via non-negative matrix factorization. Fu et al. [6] transfers the latent information from large scale web image datasets to image annotation task using Deep Learning framework. Cao et al. [2] learns more discriminative features which aims to reduce the intra-class variations.

In this paper, we focus on a combined task which provides better visual representation and annotation performance simultaneously. Evidence from visual cognition researchers demonstrates that people are usually attracted with the salient object standing out from the rest of the scene [35]. Then, the rest of the scene will be recognized via its visual features and concept correlation with the salient object. It naturally leads to the adoption of visual saliency model for image annotation. However, the number of images with region-wise labels is quite limited. In most cases, we can only get the images with some tags. Although the salient region can be extracted by some saliency detection methods, the corresponding ‘salient’ tag is not easy to obtain.

In today’s image annotation, the number of labels (i.e. concepts / tags) is quite large and label concurrence is pretty common. Intuitively, the non-salient objects, i.e. background scene, are likely to occur with the salient objects in various scenes. For instance, the tag ‘sky’ may appear in urban views which are often associated with ‘road’, etc. However, ‘sky’ can also appear in outdoor scenes with ‘dog’ and ‘trees’, etc. Since these two scenes are quite different, we can infer that the label ‘sky’ is a ‘background’ (i.e. non-salient) tag. Therefore, the coherence of the label concurrence may reveal the textual saliency.

In this paper, a Textual-Visual Saliency based Annotation (TVSA) method is proposed for image annotation by learning training sample based on visual and textual saliency. Figure 1 illustrates our framework, which consists of two parts: offline learning and online annotation.

1.1 Offline Learning

Given the labeled training samples, a concept graph is firstly established by exploiting the association between the concepts. Then concept communities are detected from concept

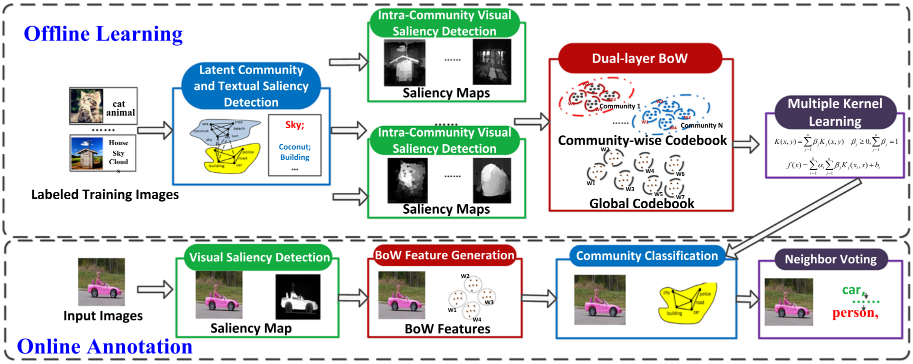


Fig. 1 The main framework of TVSA including offline learning and online annotation

graph. The textual saliency of each tag is measured by the concept removal gain (CRG) from its community. In each community, the salient region of images are detected which is used for dual-layer Bag-of-Words (DL-BoW) generation. The community classifiers are trained with Multiple-Kernel SVM based on the local features (DL-BoW) and global features of training samples in each concept community.

1.2 Online Annotation

First, the DL-BoW feature is firstly generated based on salient and non-salient visual words. Then, the corresponding community of the untagged image is determined by the community classifier. After that, intra-community annotation is performed with training samples according to the result of the community classification which assigns the salient tags to the image. The non-salient tags are determined by both intra and inter-community annotation.

Compared with our previous work in [10], we give better textual saliency representation and conduct more experiments on synthetic datasets to evaluate the performance of TVSA. The rest of our paper is organized as follows: The main details of TVSA are described in Sect. 2; In Sect. 3, we evaluate the performance of TVSA with some other approaches. Finally, the conclusion is presented in Sect. 4.

2 Methodology

2.1 Textual Saliency Detection

The first step of TVSA is to construct a concept graph based on the tagged images. In this paper, we construct a directed-weighted graph $G = \{V, E\}$. The elements of vertex set V are tags from concept set $C = \{c_1, c_2, \ldots, c_m\}$. The concept c_i is connected with c_j by a directed edge e_{ij} if an image in training set is tagged with c_i and c_j at the same time. Let w_{ij} denote the weight of e_{ij} which implies the semantic correlation between two concepts and determined as follows:

$$w_{c_i,c_j} = P(c_j|c_i) = \frac{N(c_i, c_j)}{N(c_i)} \tag{1}$$

where $P(c_j|c_i)$ is the conditional probability of concept c_j given c_i, $N(c_i)$ stands for the number of images tagged with concept c_i in the image collection and $N(c_i, c_j)$ stands for the number of images tagged with concept c_i and c_j simultaneously.

Concepts which often appear in the same scene or have similar semantic characteristics are likely to be grouped into the same community[1]. If an untagged sample is allocated to specific community, the concepts in this community are likely to be candidates for the image. The quality of the community detection, which is critical, is often measured by the modularity of the partition. Given a concept graph $G = \{V, E\}$ partitioned into M communities, denoted as $S = \{s_1, s_2, \dots, s_M\}$, modularity Q is defined as the sum of the community allocation status between concepts given as:

$$Q = \frac{1}{g} \sum_{1 \le i,j \le |C|} \left\{ \left[w_{i,j} - \frac{d_i d_j}{g} \right] \delta_1(c_i, c_j) \right\}, \quad g = \sum_{i,j} w_{i,j} \tag{2}$$

where $w_{i,j}$ denotes the directed weight of the links between concepts c_i and c_j, $d_i = \sum_j w_{i,j}$ is the sum of weights of the links attached to concept c_i, δ-function $\delta_1(c_i, c_j)$ is 1 if concepts c_i and c_j are assigned the same community and 0 otherwise, $g = \sum_{i,j} w_{i,j}$ is the sum of all weights, and $|C|$ represents the number of concepts (usually $|C| \ge M$, i.e., the number of concepts is greater than the number of communities). Higher modularity of communities leads to better partition quality, which is the objective function that needs to be optimized in community detection algorithms. In this paper, a fast unfolding algorithm [1] is applied to realize latent community detection. This algorithm has proved promising in generating proper communities with optimal time complexity.

After latent community detection, each tag is assigned with the corresponding community. However, the contribution of each concept to its community can be different. Some concepts are linked only within the concepts in the same community while some of them may have association with other communities according to the connection status based on the concept graph. If we remove a concept from specific community, the modularity of community is likely to be changed. We define the concept removal gain (CRG) between tag (c_l) and community (COM_k) as follows:

$$\begin{aligned} CRG(c_l, COM_k) &= Q_{COM_k} - Q_{COM_k/c_l} \\ Q_C &= \frac{1}{g_C} \sum_{c_i, c_j \in C} \left[w_{i,j} - \frac{d_i d_j}{g_C} \right] \\ g_C &= \sum_{c_i, c_j \in C} w_{i,j} \end{aligned} \tag{3}$$

where Q_C measures the intra-community modularity with the concepts belonging to set C which is exactly a special case of the modularity defined in Eq. (2) where $\delta_1(c_i, c_j) = 1$. $CRG(c_l, COM_k)$ shows the gain when concept c_l is removed from community COM_k .With larger $CRG(c_l, COM_k)$, the tag c_l is likely to be associated only with COM_k, i.e. a salient tag. In our previous work [10], the textual saliency is measured by the weight of tags in sum of pair-wise correlation between tags. This strategy works well but cannot fully reflect the contribution of a specific tag to the corresponding community. CRG in this paper measures the textual saliency based on the removal gain of the tags from its community which is more reasonable. Given a textual saliency threshold T_{txt}, tags are divided into two sets with high saliency and low saliency respectively. We will assign the training samples with the corresponding community by voting on the number of salient tags.

[1] The term ‘community’ comes from research field of networks which is similar to ‘clique’ in graph-cut problems but not identical.

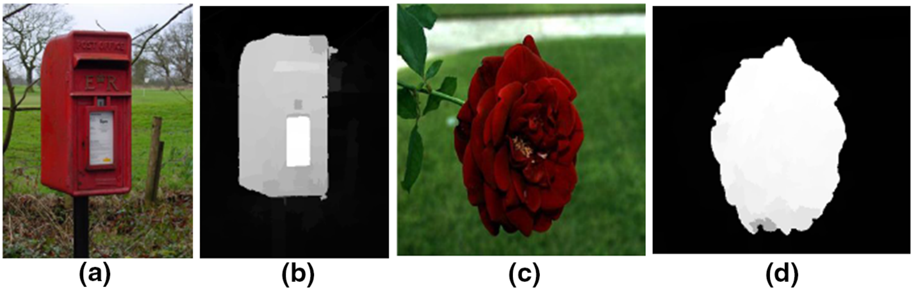


Fig. 2 Example of visual saliency detection. **a** and **c** are original images; **b** and **d** are corresponding saliency map. We can find that the lighter regions in saliency map refer to the salient objects in original images

2.2 Visual Saliency Detection

In each community, the visual saliency of a pixel refers to its relative attractiveness with respect to the whole image. To generate a saliency map for each image, a MATLAB implementation of Manifold Ranking-Based Visual Saliency [31] is applied to compute saliency values of pixels, with the values normalized to a range between 0 and 1. As shown in Fig. 2, the higher the saliency value is, the more attractive an image pixel would be. According to the result in [35], the salient portions often correspond to semantic objects in an image. Given a saliency value threshold T_{vis}, we can divide an image into two disjoint regions, one of high saliency and the other of low saliency. They will both be used to extract the visual words indicating the saliency-level. In this paper, we use the mean of the saliency values of image as the visual saliency threshold which is relatively adaptive to image variance.

2.3 Dual-layer Bag of Salient Words

Given an image, we try to use local or block-wise features for visual representation. In this paper, SIFT [17] is adopted to extract the local features in training images. Firstly, we extract visual words according to region saliency in each community. Then, the global codebook is generated according to the community-wise codebook.

In specific community, a $M\times N$ image I_k is featured with a saliency map $\{M_{k,m\times n}\}$, $m \le M$, $n\le N$ and n_k SIFT descriptors $\{D_{k,j}\}$, $j = 1...n_k$. We generate the intra-community codebook with the SIFT features and the corresponding value of the saliency map for high and low salient regions respectively. For instance, the distance between two SIFT descriptors $D_{k,i}$ and $D_{k,j}$in salient region is defined as:

$$dist(i, j) = dist(D_{k,i}, D_{k,j}) \times dist(M_{k,i}, M_{k,j}) \tag{4}$$

where $M_{k,i}$ is the saliency-level of the SIFT descriptor defined by the saliency map. The value of saliency-level is determined by the description region based on the position and scale factor. As shown in Fig. 3, the yellow corners refer to the local interesting points with different scales. The saliency-level of SIFT point is the mean value of the saliency value in the 4 × 4 grid around the interesting point. According to the origin of SIFT, each descriptor is generated based on its scale in 4×4 neighborhoods. Compared with our previous work [10] which measures the visual saliency based on single pixel, the proposed strategy fully considers the characteristics of SIFT descriptors. The codebook can be generated by clustering based on the distance measured as Eq. (4). However, for non-salient regions, we

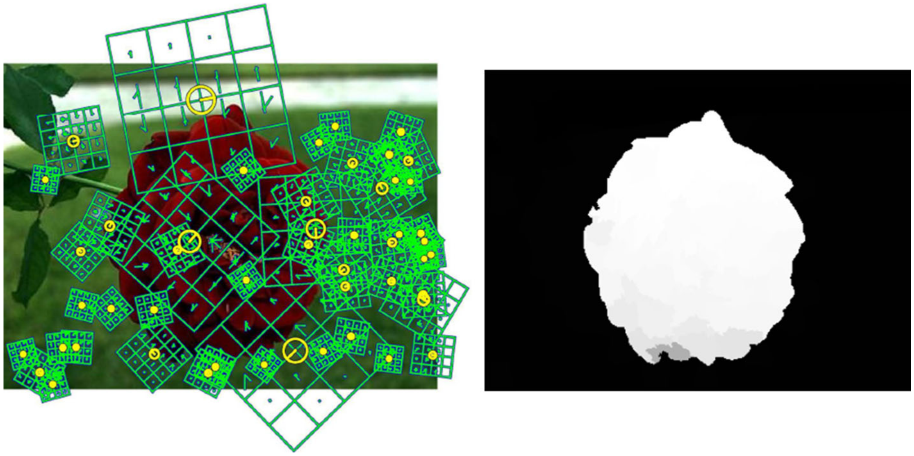


Fig. 3 The image on the left shows local interesting points with different scales. The saliency-level of the point is determined by the mean value of saliency in the 4 × 4 grid around the interesting point. For the grids cropped by the boundary of image, only the saliency values within the image are considered

directly use $dist(D_{k,i}, D_{k,j})$ for similarity measurement since the saliency value are quite closed for them. For distance function $dist(\cdot, \cdot)$, we directly use the Euclidean distance based on the normalized feature.

According to the distance between local features, the community-wise codebook consisting of visual words for salient and non-salient region is obtained via K-means clustering algorithm. Based on the community-wise codebook, we can obtain the global codebook by clustering the visual words from all communities for salient and non-salient regions. Images are quantized into Bag-of-Words features based on the global codebook.

2.4 Community Classifier: Learning and Inference

We define the score of interpreting an image I with the corresponding community as :

$$F(I) = \Theta^T \Phi(I) = \theta^T \phi_{sal}(I) + \eta^T \phi_{unsal}(I) + \beta^T \omega(I) \tag{5}$$

In the following, we describe in detail each term in Eq. (5).

2.4.1 Bag-of-Salient-Words $\theta^T \phi_{sal}(I)$

For an unlabeled image I, we can extract the local feature based on salient visual words. θ_i can be weight associated with the similarity between each training samples I_k and the unlabeled image.Therefore, we can parameterize this potential function as:

$$\theta^T \phi_{sal}(I) = \sum_{I_k \in I_{Com}} \theta_k K_{sal}(I, I_k) \tag{6}$$

where $K_{sal}(I, I_k)$ is a similarity function,I_{Com} denote the images in specific community.

2.4.2 Bag-of-non-salient-Words $\eta^T \phi_{unsal}(I)$

This potential function captures the similarity on non-salient words between each training samples I_k and the unlabeled image. As shown above, we can parameterize it as:

$$\eta^T \phi_{unsal}(I) = \sum_{I_k \in I_{Com}} \eta_k K_{unsal}(I, I_k) \tag{7}$$

Global features $\beta^T \omega(I)$: This part indicates how likely the image I assigned with this community based on global features of I. It is shown as:

$$\beta^T \omega(I) = \sum_{I_k \in I_{Com}} \beta_k K_{global}(I, I_k) \tag{8}$$

We learn our model in a multiple-kernel learning SVM framework. The multiple-kernel SVM model can be trained with adaptively-weighted combined kernels and each kernel is in accordance with a specific type of visual feature. The decision function is defined as follows:

$$\begin{aligned} F(I) &= \sum_{I_k \in I_{Com}} \theta_i K_{sal}(I, I_k) + \eta_i K_{unsal}(I, I_k) + \beta_i K_{global}(I, I_k) \\ &= \sum_{I_k \in I_{Com}} \alpha_k \left\{ \frac{\theta_k}{\alpha_k} K_{sal}(I, I_k) + \frac{\eta_k}{\alpha_k} K_{unsal}(I, I_k) + \frac{\beta_k}{\alpha_k} K_{global}(I, I_k) \right\} \\ &= \sum_{I_k \in I_{Com}} \alpha_k \sum_m w_m K_m(I, I_k) = \sum_{I_k \in I_{Com}} \alpha_k K(I, I_k) \end{aligned} \tag{9}$$

where $K(\cdot)$ is the combined kernel, $K_m(\cdot)$ is the sub-kernel of m_{th} visual feature and w_m is the weight for sub-kernel to be learnt. In order to get a sparse solution, we add the $l_1$norm constraints and the learning problem is shown as follows:

$$\begin{aligned} &\min \frac{1}{2} \|F\| + C \sum_{I_k \in I_{Com}} \xi_k \\ &s.t. F(I) = \sum_{I_k \in I_{Com}} \alpha_k K(I, I_k) \\ &K(I, I_k) = \sum_m w_m K_m(I, I_k), \; w_m \geq 0, \sum_m w_m = 1 \\ &\xi_k \geq 0, \; y_k F(I_k) \geq 1 - \xi_k \end{aligned} \tag{10}$$

where ξ_k is the relaxation variables in SVM and y_k is the two-class label of samples. The community classifier is exactly a multi-class SVM. However, it has to be converted in several two-class problems to learn. In this paper, we adopt the widely-used one-versus-all strategy. As reported in previous work, multiple-kernel SVM shows better performance than conventional SVM learnt with combined features. We solve this problem via SimpleMKL [25].

2.5 Labeling: Neighbor-Voting in Communities

The corresponding communities of an untagged image can be determined by the trained community classifiers. A naive KNN search is carried out to realize the tag assignment in each community based on the Euclidean distance between the visual features of the untagged image and the ones in the community. We will firstly tag the image with the salient tags.

The non-salient tag is assigned based both on the correlation of salient tag and the visual feature. Let $r(I, r_{c_i}^{sal})$ denote the relevance between image I and salient tag c_i. $r(I, r_{c_i}^{sal})$ is determined by the K-nearest-neighbors measured with Bag-of-Salient-Words feature and global features:

$$r(I, r_{c_i}^{sal}) = \frac{1}{K}\left\{\sum_{I_j \in \mathcal{N}_K^{sal}(I)} w_{sal} r(I_j, r_{c_i}^{sal}) + \sum_{I_j \in \mathcal{N}_K^{global}(I)} w_{global} r(I_j, r_{c_i}^{sal})\right\} \tag{11}$$

where w_{sal} and w_{global} are kernel weight obtained in Eq. (10); $\mathcal{N}_K^{sal}(I)$ is the K-nearest-neighbors measured with salient word feature; $\mathcal{N}_K^{global}(I)$ is the K-nearest-neighbors measured with global feature which can reduce the impact of false/miss salient regions.

Similarly, the relevance between the unlabeled image and non-salient tags based on visual features are firstly determined by global and non-salient features:

$$r(I, r_{c_i}^{unsal}) = \frac{1}{K}\left\{\sum_{I_j \in \mathcal{N}_K^{unsal}(I)} w_{unsal} r(I_j, r_{c_i}^{unsal}) + \sum_{I_j \in \mathcal{N}_K^{global}(I)} w_{global} r(I_j, r_{c_i}^{unsal})\right\} \tag{12}$$

Inspired by the association ability of human beings, the non-salient tags are also complemented by the salient tags as follows:

$$\begin{aligned} r^*(I, r_{c_i}^{unsal}) &= r(I, r_{c_i}^{unsal}) + \sum_k r(I, r_{c_k}^{sal}) \times w(c_k, c_i) \\ &= r(I, r_{c_i}^{unsal}) + \sum_k r(I, r_{c_k}^{sal}) \times P(c_i|c_k) \end{aligned} \tag{13}$$

where c_k is the salient concept assigned to image by TVSA. The final tagging information of the image is a combination of salient and non-salient tags.

2.6 Time Complexity Analysis

In this section, we analyze the computational complexity of TVSA. In offline learning, we firstly detect the semantic communities on concept graph. The time complexity of constructing concept graph is $O(m^2)$ where m is the number of semantic concepts and the space complexity is $O(m^2)$. The complexity of community detection is $O(m^2 log_2 m)$ since the iteration of community detection is similar to a hierarchical clustering process. The space complexity of community detection is also $O(m^2)$. The textual saliency is measured by CRG values in Eq. (3). The time complexity of calculating CRG is $O(m)$ since the saliency of each tag should be obtained. After textual saliency detection, all training samples are assigned to specific community whose time complexity is $O(nmlog_2 m)$ where n is the number of training samples and $n \gg m$. Therefore, it can be considered linear to the number of training samples. In visual saliency detection , we first generate the saliency map of images which is related to the specific algorithms adopted. Yang et al. [31] used in this paper is actually a graph-based propagation method which is quite efficient. The most-time consuming process is to generate the dual-layer codebook by KMeans algorithm. When given n_d SIFT descriptors, the time-complexity is $O(kn_d)$ where k is the maximum number of iteration. Then, we train the multiple kernel SVM based on the samples with multiple features. Since we adopt one-versus-all? strategy for learning community classifiers, we have to train n_c MKL-SVMs where n_c is the number of communities.

In online annotation, the salient, non-salient and global features are extracted. The time-complexity for extracting salient and non-salient feature is linear to the size of codebook. Then, the untagged samples are firstly assigned to the most relevant communities by MKL-SVM. After that, we will give the initial tags via intra-community annotation which is actually a KNN process in top M relevant communities. If KNN is boosted by KD-Tree, the time complexity of building for all communities is $O(n_c n_{cs} log_2 n_c)$ where n_c is the number of communities and n_{cs} is the number of training samples in each community. The time complexity of annotation is $O(Mn_t log_2 n_{cs})$ where M is the number of candidate communities and n_t is the number of untagged images. Therefore, the total time complexity of intra-annotation is $O(n_c n_{cs} log_2 n_c + Mn_t log_2 n_{cs})$.

3 Experiments

3.1 Pre-settings and Evaluation Measures

In this section, some experiments are conducted to evaluate the performance of the proposed method on MIRFlickr [12] and IAPR TC-12 [8] datasets.

The annotation model is trained using the training part while the evaluation of the model is based on the testing part. All visual features are deployed for the compared methods. The comparison between TVSA and state-of-the-art methods MLKNN [34], RLVT [14], RANK [14], NBVT [13] and LCMKL [9] is also presented to show the proposed method progresses towards better performance. NBVT is a neighbor voting method for tag relevance estimation. RLVT takes the relevance between tags into consideration based on the Google distance combined with low-level visual features. RANK is an extension of RLVT using a random walk. ML-KNN is derived from the k-nearest method exploiting Bayesian rules. LCMKL is a general framework using community detection and multiple kernel learning for image annotation. In order to present the improvement compared with our previous works in [10], we denote the result based on [10] with TVSA-prev. The method proposed in this paper is represented by TVSA-cur. All of the experiments are executed on a PC with Intel 2.4GHz CPU and 10GB RAM on MATLAB.

For TVSA, we use [31] to extract saliency map and detect 500D BoW feature for salient and non-salient regions respectively. Global feature ,i.e. Color Histogram (64D), is also adopted as visual representation. For the baseline methods, a 1000D BoW feature and the global features mentioned above are deployed. The parameter settings for TVSA are listed as follows: The threshold of textual saliency (T_{txt})is set to 0.5 while for the visual saliency (T_{vis})is the mean-value of image's saliency map. The number of neighbors for neighbor-voting is 100. The scaling factor σ in Eq. (4) is 10.

In this paper, Precision, Recall and F1-score are used to measure the performance of image annotation. For concept c_i, they are determined as follows:

$$Precision(c_i) = \frac{N_{corr}}{N_{tagged}}; Recall(c_i) = \frac{N_{corr}}{N_{all}}$$
$$F_1 - score(c_i) = 2\frac{Precision(c_i) \times Recall(c_i)}{Precision(c_i) + Recall(c_i)} \tag{14}$$

where N_{tagged} denotes the number of images tagged with a specific concept c_i, in testing part by image annotation, N_{corr} denotes the number of images tagged correctly according to the original tagging information and N_{all} denotes the number of images tagged with c_i in training part. For each concept, we can obtain Precision, Recall and F1-score respectively. The global performance is obtained via averaging over all concepts. To make fair comparisons, the top five relevant concepts of the image are selected for annotation.

3.2 Experiments on IAPR TC-12

IAPR TC-12 dataset was used for the ImageClef Challenge from 2006 to 2008. It consists of still natural images taken from locations around the world and comprising an assorted cross-section of still natural images. The numbers of training and test samples are 17,665 and 1962, respectively. The number of tags is 291.

Table 1 shows the performance of image annotation on IAPR TC-12 dataset. We observe that the proposed method outperforms the compared method on Avg. Precision, Avg. Recall and Avg. F1-score with the top 10 relevant tags. Since TVSA focus on tagging the visual and textual salient objects, the recall of TVSA is relatively higher than other methods. Some exemplars are shown in Fig. 4.

3.3 Experiments on MIRFlickr

MIRFlickr consists of 25,000 images that were downloaded from the social photography site Flickr.com through its public API. The color images are representative of a generic domain

Table 1 The performance comparison on IAPR TC-12 dataset measured by Precision, Recall and F1-score

Method	MLKNN	NBVT	RANK	RLVT	LCMKL	TVSA-prev	TVSA-cur
Recall	0.162	0.195	0.179	0.106	0.287	0.295	0.321
Precision	0.112	0.091	0.191	0.254	0.121	0.131	0.128
F1-score	0.132	0.124	0.134	0.149	0.170	0.181	0.184

Images	GT	MLKNN	NBVT	RANK	RLVT	LCMKL	TVSA-prev	TVSA-prev
	desert dune	cloud tree mountain sky man	sky cloud landscape horizon sea	bird cloud formation horizon jungle	bird cloud formation horizon jungle	desert dune photo bus airplane	desert sky dune cloud tree	desert dune sky airplane photo
	cloud mountain	sky tree rock landscape wall	sky tree cloud slope snow	jungle restaurant portrait hammock fern	jungle restaurant portrait hammock fern	mountain cloud slope snow sea	cloud mountain snow slope sea	cloud mountain snow slope sea
	man	sky wall people desert tree	sky tree cloud slope snow	shoe shelf sock terrain neck	shelf shoe sock couch people	man woman hair cap girl	man people girl hair cap	man people hair terrain cap
	tree	mountain front tree team range	tree mountain view bush front	waterfall shop classroom pavement stand	shop waterfall classroom pavement desk	tree sky people front building	tree bush front view building	tree bush front view building

Fig. 4 Tagging exemplars from IAPR TC-12 dataset. 'GT' denotes the ground truth. We select only top five relevant tags generated by each method. The salient regions like 'man' and 'tree' are ranked higher than the non-salient objects by TVSA

Table 2 The performance comparison on MIRFlickr dataset measured by Precision, Recall and F1-score

Method	MLKNN	NBVT	RANK	RLVT	LCMKL	TVSA-prev	TVSA-cur
Recall	0.210	0.211	0.208	0.209	0.200	0.201	0.223
Precision	0.091	0.123	0.119	0.121	0.118	0.122	0.140
F1-score	0.126	0.155	0.151	0.153	0.148	0.152	0.173

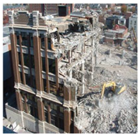
Images	GT	MLKNN	NBVT	RANK	RLVT	LCMKL	TVSA-prev	TVSA-cur
	plant_life	people people _r1 structures plant_life indoor	structures plant_life people indoor male	river_r1 bird_r1 sea_r1 dog_r1 dog	river_r1 sea_r1 bird_r1 dog_r1 dog	sky plant_life water tree clouds	plant_life sky tree river clouds	plant_life sky tree water clouds
	animals bird bird_r1	people bird plant_life people_r1 indoor	people indoor animals bird sky	female bird people dog_r1 dog	female indoor people bird female_r1	animals baby baby_r1 bird bird_r1	animals bird people baby baby_r1	animals bird bird_r1 baby baby_r1
	plant_life sky tree	people structures plant_life sky indoor	plant_life people structures indoor female	animals baby baby_r1 bird bird_r1	animals baby baby_r1 bird bird_r1	plant_life sky tree clouds water	plant_life sky tree bird water	plant_life sky tree clouds water
	structures transport	people structures plant_life sky indoor	indoor people plant_life structures sky	sea_r1 river_r1 dog_r1 dog bird_r1	dog_r1 sea_r1 river_r1 dog bird_r1	car Sky transport night night_r1	structures transport sky night night_r1	structures transport night car night_r1

Fig. 5 Tagging exemplars from MIRFlickr dataset. 'GT' denotes the ground truth. Some textual concepts are duplicated in MIRFlickr like 'bird' and 'bird_r1'

and are of high quality. The numbers of training and test samples are both 12,500. The number of tags is 38.

Table 2 shows the performance of image annotation on MIRFlickr dataset. Similar to the result of IAPR TC-12 dataset, we observe that the proposed method outperforms the compared method on Avg. Precision, Avg. Recall and Avg. F1-score with the top 5 relevant tags. However, the improvement gained by TVSA is not as obvious as on IAPR TC-12 dataset. The main reason is that the number of concepts from MIRFlickr is only 38 and some of them are actually duplicated like 'bird_r1' and 'bird'. Since the number of textual concepts and the tag co-occurrence are both limited, the detection of textual saliency cannot provide better performance. Some exemplars are shown in Fig. 5.

3.4 Discussions

Most of the previous works for image annotation follows two basic strategies: search-based methods which assign tags for images based on the similarity between images; learning-based methods which consider image annotation as a classification problem (Neural Networks/SVM). They did not fully take the structure of semantic labels and the cross-modal correlations. Our method is based on two assumptions: the co-occurrence of labels is a key factor in image annotation and the salient visual features have strong associations with salient textual labels. Therefore, our approach has relatively clear target in annotation and introduce promising performance.

We also discuss the selection of key parameters of TVSA in this section including threshold of visual saliency(T_{txt}), textual saliency (T_{vis}) and the size of bag of visual words.

For visual saliency, it is not appropriate to set a fixed threshold since the distribution of saliency map varies in different images. The mean-value of image's saliency map is a relative simple and good choice. For textual saliency, the threshold is selected by cross-validation among $\{0.1, 0.2, \ldots, 0.9\}$. We found that $T_{txt} = 0.5$ achieves the best performance.

For the size of bag of visual words, we find that with larger size of salient visual words, the performance of TVSA can be improved. When using 1000D Bag-of-salient-words, the F1-score of TVSA is 0.220 on IAPR TC-12 and 0.192 on MIRFlickr. However, larger size of non-salient visual words cannot provide better performance since the non-salient regions often refers to simple concepts like 'sky' or the objects without salient semantic information.

4 Conclusion

In this paper, a Textual-Visual Saliency based framework for image annotation is proposed. Our work integrates the textual saliency on labels and visual saliency on images. A concept graph is constructed which implies a dense semantic intra-community correlation of concepts. The dual-layer Bag-of-Words provide a good visual representation based on local features and salient regions. The robust multiple-kernel SVM is applied for community classification. Experiments on IAPR TC-12 and MIRFlickr datasets demonstrate that the proposed method outperforms other state-of-the-art approaches.

Acknowledgements This work is partly supported by NSFC China (No: 61572315), and 973 Plan, China (No. 2015CB856004).

References

1. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech 2008(10):P10008
2. Cao X, Zhang H, Guo X, Liu S, Meng D (2015) Sled: semantic label embedding dictionary representation for multi-label image annotation. IEEE Trans Image Process 24:2746
3. Carneiro G, Chan A, Moreno P, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans Pattern Anal Mach Intell 29(3):394–410. doi:10.1109/TPAMI.2007.61
4. Chapelle O, Haffner P, Vapnik VN (1999) Support vector machines for histogram-based image classification. IEEE Trans Neural Netw 10(5):1055–1064
5. Cusano C, Ciocca G, Schettini R (2003) Image annotation using SVM. In: Electronic imaging 2004. International Society for Optics and Photonics, pp. 330–338
6. Fu J, Mei T, Yang K, Lu H, Rui Y (2015) Tagging personal photos with transfer deep learning. In: Proceedings of the 24th international conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp 344–354
7. Goh KS, Chang EY, Li B (2005) Using one-class and two-class svms for multiclass image annotation. IEEE Trans Knowl Data Eng 17(10):1333–1346
8. Grubinger M, Clough P, M U Ller H, Deselaers T (2006) The iapr tc-12 benchmark: A new evaluation resource for visual information systems. Int Workshop OntoImage 5:10
9. Gu Y, Qian X, Li Q, Wang M, Hong R, Tian Q (2015) Image annotation by latent community detection and multikernel learning. IEEE Trans Image Process 24(11):3450–3463. doi:10.1109/TIP.2015.2443501

10. Gu Y, Xue H, Yang J, Jia Z (2014) Automatic image annotation exploiting textual and visual saliency. In: Neural information processing. Springer, Berlin, pp 95–102
11. Han Y, Wu F, Tian Q, Zhuang Y (2012) Image annotation by input-output structural grouping sparsity. IEEE Trans Image Process 21(6):3066–3079
12. Huiskes MJ, Lew MS (2008) The mir flickr retrieval evaluation. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval, ACM, pp 39–43
13. Li X, Snoek CG, Worring M (2009) Learning social tag relevance by neighbor voting. IEEE Trans Multimed 11(7):1310–1322
14. Liu D, Hua XS, Yang L, Wang M, Zhang HJ (2009) Tag ranking. In: Proceedings of the 18th international conference on World wide web, ACM, pp. 351–360
15. Liu X, Cheng B, Yan S, Tang J, Chua TS, Jin H (2009) Label to region by bi-layer sparsity priors. In: Proceedings of the 17th ACM international conference on Multimedia, ACM, pp. 115–124
16. Liu X, Liu R, Li F, Cao Q (2012) Graph-based dimensionality reduction for knn-based image annotation. In: ICPR, pp. 1253–1256
17. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
18. Lu Z, Wang L (2015) Learning descriptive visual representation for image classification and annotation. Pattern Recognit 48(2):498–508
19. Ma Z, Nie F, Yang Y, Uijlings JR, Sebe N (2012) Web image annotation via subspace-sparsity collaborated feature selection. IEEE Trans Multimed 14(4):1021–1030
20. Makadia A, Pavlovic V, Kumar S (2010) Baselines for image annotation. Int J Comput Vis 90(1):88–105
21. Qi X, Han Y (2007) Incorporating multiple SVMs for automatic image annotation. Pattern Recognit 40(2):728–741
22. Qian X, Hua XS, Tang YY, Mei T (2014) Social image tagging with diverse semantics. IEEE Transactions on Cybern 44(12):2493–2508. doi:10.1109/TCYB.2014.2309593
23. Qian X, Liu X, Zheng C, Du Y, Hou X (2013) Tagging photos using users' vocabularies. Neurocomputing 111:144–153
24. Saito P, de Rezende PJ, Falc A O AX, Suzuki CT, Gomes JF (2013) A data reduction and organization approach for efficient image annotation. In: Proceedings of the 28th annual ACM symposium on applied computing, pp. 53–57
25. Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B (2006) Large scale multiple kernel learning. J Mach Learn Res 7:1531–1565
26. Tang J, Hong R, Yan S, Chua TS, Qi GJ, Jain R (2011) Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. ACM Trans Intell Syst Technol (TIST) 2(2):14
27. Tang J, Yan S, Zhao C, Chua TS, Jain R (2013) Label-specific training set construction from web resource for image annotation. Signal Process 93(8):2199–2204
28. Tang J, Zha ZJ, Tao D, Chua TS (2012) Semantic-gap-oriented active learning for multilabel image annotation. IEEE Trans Image Process 21(4):2354–2360
29. Wu B, Lyu S, Hu BG, Ji Q (2015) Multi-label learning with missing labels for image annotation and facial action unit recognition. Pattern Recognit 48:2279
30. Yan R, Natsev A, Campbell M (2008) A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In: CVPR, pp. 1–8
31. Yang C, Zhang L, Lu H, Ruan X, Yang MH (2013) Saliency detection via graph-based manifold ranking. In: Computer vision and pattern recognition, 2013. IEEE Conference on CVPR 2013, pp. 3166–3173
32. Yang Y, Wu F, Nie F, Shen HT, Zhuang Y, Hauptmann AG (2012) Web and personal image annotation by mining label correlation with relaxed visual graph embedding. IEEE Trans Image Process 21(3):1339–1351
33. Zhang D, Islam MM, Lu G (2012) A review on automatic image annotation techniques. Pattern Recognit 45(1):346–362
34. Zhang ML, Zhou ZH (2007) Ml-knn: a lazy learning approach to multi-label learning. Pattern Recognit 40(7):2038–2048
35. Zhu G, Wang Q, Yuan Y (2014) Tag-saliency: combining bottom-up and top-down information for saliency detection. Comput Vis Image Underst 118:40–49