Automatic Image Annotation Exploiting Textual and Visual Saliency

Yun Gu[1], Haoyang Xue[1], Jie Yang*[1], Nikola Kasabov[2], and Zhenhong Jia[3]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China.
[2] Knowledge Engineering and Discovery Research Institute, Auckland University of Technology Auckland, New Zealand.
[3] School of Information Science and Engineering, Xinjiang University, Urumqi, China.
{geron762,xuehaoyangde,jieyang}@sjtu.edu.cn,
nkasabov@aut.ac.nz,jzhh@xju.edu.cn

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 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 NUS-WIDE dataset 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. Search-based methods like [5, 6] and learning-based methods like [11, 9] 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.[2] learns the probability distribution of a semantic class from images with weakly labeled information. In [7], the images are coded with sparse features via over-segmenatation for label-to-region annotation. In this paper, we focus on a combined task which provides better visual representation and annotation performance simultaenously.

Evidence from visual cognition researchers demonstrates that people are usually attracted with the salient object standing out from the rest of the scene[13].

* Corresponding author: Jie Yang, jieyang@sjtu.edu.cn

Then, the rest of the scene will be recognized via the 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.

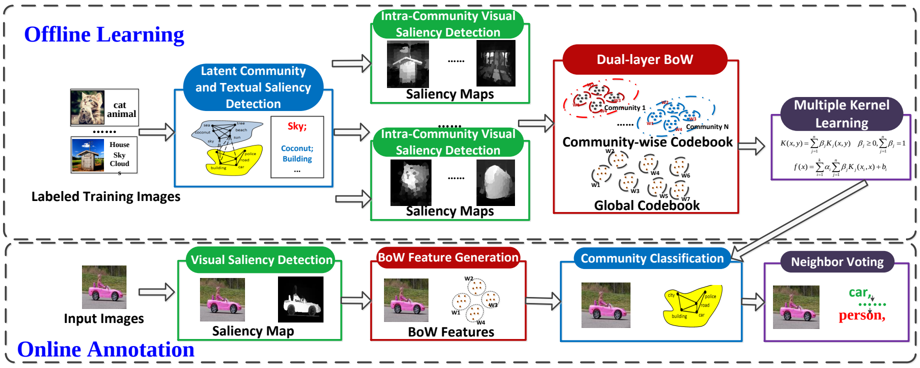

Fig. 1. The main framework of TVSA

In todays 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 is 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 an "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.

Offline Learning: Given the labeled training samples, a concept graph is firstly established by exploiting the association between the concepts. Then concept communities and latent textual saliency are detected from concept graph. 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.

Online Annotation: The DL-BoW feature is firstly generated for the unlabeled image. Then, corresponding community of the image is determined by the community classifier. Finally, neighbor-voting annotation is performed with training samples according to the result of community classification.

The rest of our paper is organized as follows: The main details of TVSA are described in Section 2; In Section 3, we evaluate the performance of TVSA with some other approaches. Finally, the conclusion is presented in Section 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 $C = \{c_1, c_2, , c_m\}$. 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. If an untagged sample is allocated to specific community, the concepts in this community are likely to be candidating labels for the image. In this paper, a fast unfolding algorithm [1] is applied to realize the latent community detection. It is proved a promising algorithm to generate proper communities under optimal time-complexity.

After latent community detection, each tag is assigned with the corresponding community. We define the correlation between tag (c_i) and community (COM_k) as follows:

$$Corr(c_i, COM_k) = \frac{1}{N_{COM_k}} \sum_{c_j \in COM_k} w_{c_j,c_i} = \frac{1}{N_{COM_k}} \sum_{c_j \in COM_k} \frac{N(c_i,c_j)}{N(c_i)} \tag{2}$$

where N_{COM_k} denotes the number of concepts in COM_k. The textual saliency of tag c_i assigned with COM_k is defined as:

$$Sal(c_i) = \frac{Corr(c_i, COM_k)}{\sum_{m=1}^{N_{COM}} Corr(c_i, COM_m)} \tag{3}$$

where N_{COM} denotes the number of communties. $Salc_i$ indicates the intra-community correlatation and inter-community discrimination. With larger $Salc_i$, the tag c_i is likely to be asscociated only with COM_k,i.e. a salient tag. Given a textual saliency threshold T_{txt}, tags are divided into two sets with high saliency and low saliency respectively. Noted that we will assign the training samples with the corresponding community by voting on the number of salient tags.

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[10] is applied to compute saliency values of pixels, with the values normalized to a range between 0 and 1. The higher the saliency value is, the more attractive an image pixel would be. As reported in [13], 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.

2.3 Dual-layer Bag of Salient Words

In our work, SIFT 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 the specific community,a $M \times N$ image I_k is featured with a saliency map $\{M_{k,m \times n}\}$, $m \leq M, n{\leq}N$ 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:

$$d_{i,j} = \|D_{k,i} - D_{k,j}\| \exp^{\frac{\|M_{k,i} - M_{k,j}\|}{\sigma}} \tag{4}$$

where $M_{k,i}$ is the saliency-level of the SIFT descriptor defined by the saliency map. The codebook can be generated by clustering based on the distance measured as Eq.4. However, for non-salient regions, we directly use $\|D_{k,i} - D_{kj}\|$ for similarity measurement since the saliency value are quite closed for them.As a result, the community-wise codebook consisting of visual words for salient and non-salient region is obtained.

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. The DL-BoW features of image are generated according to the global codebook for salient and non-salient regions.

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).

Bag-of-Salient-Words $\theta_{sal}^T\phi_{sal}(I)$: 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.

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

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

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:

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

As reported in previous work, multiple-kernel SVM shows better performance than conventional SVM learnt with combined features. We solve this problem via SimpleMKL[8].

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 initial annotation in each community based on the Euclidean distance between the visual features of the untagged image and the ones in the community. Noted that 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:

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

where w_{sal} and w_{global} are kernel weight obtained in (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, replace "sal" with "unsal" in (11). The relevance between the unlabeled image and non-salient tags are determined as:

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

The final tagging information of the image is a combination of salient and non-salient tags.

3 Experiments

In this section, some experiments are conducted to evaluate the performance of the proposed method on NUS-WIDE[3] dataset which contains 27807 images in training parts and 27808 images in testing parts. All images are tagged with labels from 81 Ground Truth. The comparison between TVSA and state-of-the-art methods MLKNN[12], MLNB[11], RLVT[6], RANK[6],NBVT[5] and LCMKL[4] is also presented to show the proposed method progresses towards better performance. All of the experiments are executed on a PC with Intel 2.4GHz CPU and 10GB RAM on MATLAB.

For TVSA, we use [10] to extract saliency map and detect 500D BoW feature for salient and non-salient regions respectively. Global features including Color Moments(225D) and Color Histogram (64D) are 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 0.4 while for the visual saliency (T_{vis})is the mean-value of image's saliency map. The number of neighbours for neighbot-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:

$$\begin{aligned} Precision(c_i) &= \frac{N_{corr}}{N_{tagged}}; Recall(c_i) = \frac{N_{corr}}{N_{all}} \\ F_1 - score(c_i) &= 2\frac{Precision(c_i) \times Recall(c_i)}{Precision(c_i) + Recall(c_i)} \end{aligned} \tag{13}$$

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 Precison, 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. Table 1 shows the performance of image annotation on NUS-WIDE:

Table 1. The performance comparison on NUS-WIDE 81 tags

Method	MLKNN	MLNB	RLVT	RANK	NBVT	LCMKL	TVSA
Precision	0.122	0.110	0.192	0.181	0.127	0.237	0.263
Recall	0.210	0.302	0.186	0.187	0.177	0.233	0.282
F1-score	0.154	0.161	0.187	0.184	0.148	0.235	0.272

As shown in Table 1, we observe that the proposed method outperforms the compared method on Avg. Precsion, Avg. Recall and Avg. F1-score with the top five relevant tags.

Finally, we also discuss the selection of key paramters of TVSA including threshold of visual saliency(T_{txt}) and textual saliency (T_{vis}). 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 choise. For textual saliency, the threshold is seleted by cross-validation among $\{0.1, 0.2, ..., 0.9\}$. We found that $T_{txt} = 0.4$ achieves the best performance.

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 sematic intra-community correlation of concepts. The dual-layer Bag-of-Words provide a good visual representatiopn based on local features and salienct regions. The robust multiple-kernel SVM is applied for community classification. Experiments on NUS-WIDE dataset demonstrate that the proposed method outperforms other state-of-the-art approaches.

Acknowledgments. This research is partly supported by NSFC, China (No: 6127325861105001) Ph.D. Programs Foundation of Ministry of Education of China (No.20120073110018).

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10), P10008 (2008)
2. Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(3), 394–410 (March 2007)
3. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. p. 48. ACM (2009)
4. Li, Q., Gu, Y., Qian, X.: Lcmkl: latent-community and multi-kernel learning based image annotation. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 1469–1472. ACM (2013)
5. Li, X., Snoek, C.G., Worring, M.: Learning social tag relevance by neighbor voting. Multimedia, IEEE Transactions on 11(7), 1310–1322 (2009)
6. Liu, D., Hua, X.S., Yang, L., Wang, M., Zhang, H.J.: Tag ranking. In: Proceedings of the 18th international conference on World wide web. pp. 351–360. ACM (2009)
7. Liu, X., Cheng, B., Yan, S., Tang, J., Chua, T.S., Jin, H.: Label to region by bi-layer sparsity priors. In: Proceedings of the 17th ACM international conference on Multimedia. pp. 115–124. ACM (2009)
8. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large scale multiple kernel learning. The Journal of Machine Learning Research 7, 1531–1565 (2006)
9. Yan, R., Natsev, A., Campbell, M.: A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
10. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Computer Vision and Pattern Recognition, 2013. CVPR 2013. IEEE Conference on. pp. 3166–3173 (2013)
11. Zhang, M.L., Peña, J.M., Robles, V.: Feature selection for multi-label naive bayes classification. Information Sciences 179(19), 3218–3229 (2009)
12. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition 40(7), 2038–2048 (2007)
13. Zhu, G., Wang, Q., Yuan, Y.: Tag-saliency: Combining bottom-up and top-down information for saliency detection. Computer Vision and Image Understanding 118, 40–49 (2014)