Canonical Image Selection by Visual Context Learning

Wengang Zhou[1], Yijuan Lu[2], Houqiang Li[1], Qi Tian[3]
Dept. of EEIS, University of Science and Technology of China[1], Hefei, P.R. China
Dept. of Computer Science, Texas State University at San Marcos[2], Texas, TX 78666
Dept. of Computer Science, University of Texas at San Antonio[3], Texas, TX 78249
zhwg@mail.ustc.edu.cn[1], yl12@txstate.edu[2], lihq@ustc.edu.cn[1], qitian@cs.utsa.edu[3]

Abstract

Canonical image selection is to select a subset of photos that best summarize a photo collection. In this paper, we define the canonical image as those that contain most important and distinctive visual words. We propose to use visual context learning to discover visual word significance and develop Weighted Set Coverage algorithm to select canonical images containing distinctive visual words. Experiments with web image datasets demonstrate that the canonical images selected by our approach are not only representatives of the collected photos, but also exhibit a diverse set of views with minimal redundancy.

1. Introduction

The last decade has witnessed the proliferation of images on the web, which cast a great challenge on exploring gigantic amount of images. Currently, most image search engines simply retrieve all images relevant to a text keyword search and organize them into pages of thumbnails, leading to a tedious browsing experience for a user. Recently, many research efforts have been made to select canonical images, a subset of photos that best summarize the image collections.

Jaffe *et al*. [1] used a pure metadata-driven approach to select canonical views for a region with both geo-locations and several heuristics on metadata. Since no visual content information is considered, the selected views are sensitive to noise.

Some other research works use clustering-based approaches to discover canonical images. Raguram and Lazebnik [3] proposed to compute iconic views for a collection of images of any concept with joint clustering based on visual and textual features to extract subsets of images. Iconic views were chosen from images with the highest visual quality from each subset. Kennedy *et al*. [6] leveraged both metadata and visual features to form a hybrid approach for canonical view selection. They discovered landmark related images and performed *k*-means clustering based on global color and texture features. Then canonical views were selected as top-ranked images from top-ranked clusters. These clustering based methods unavoidably suffer from the weakness of determining the number of clusters, which greatly constrains their flexibilities. Simon *et al*. [5] proposed a pure vision-based approach. The original photo collection was partitioned into several non-overlapping subsets. Then, a greedy *k*-means clustering was adopted to select canonical views from each subset. However, like other clustering methods, the greedy *k*-means is still sensitive to outliers, which are prevalent in web image collections.

Yang *et al*. [4] proposed a canonical view selection method based on online search results. They analyzed the distribution of visual words [8] and computed a coverage score for each photograph. 200 distinctive visual words were selected by *wc-tf-idf* measurement and a greedy scheme was proposed to iteratively select a few canonical views. However, *wc-tf-idf* cannot necessarily capture the visual word latent similarity relationship, hence, cannot identify distinctive visual words very well.

In this paper, we define the representative photos as those that contain many most important and distinctive visual words. We propose to use latent visual context learning to measure visual words significance and discover distinctive visual words. Then we select the canonical images via set coverage strategy -- greedily discover photos that contain those important visual words from a candidate image pool. In order to construct a good candidate image pool and filter some noisy images, we also propose an image link graph to

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.210

rank all images and select the top ones for canonical image selection.

Our approach has two major advantages. First, the proposed visual context learning can capture the latent similarity relationship between visual words, which is very important and useful to determine the significance of visual words and images. The distinctive visual words for set coverage are selected adaptively, which ensure the flexibility of our Weighted Set Coverage approach. Second, unlike *k*-means clustering, the canonical images are selected in a greedy fashion. There is no need to know the number of canonical photo sets beforehand.

The rest of the paper is organized as follows. Section 2 and Section 3 discuss visual context learning for visual words and images, respectively. In Section 4, we describe weighted set coverage analysis to form canonical set. The experiments are given in Section 5 and conclusion is made in Section 6.

2. Visual word significance discovery

We define the representative photos as the photos that contain many important and distinctive visual words. The key step is to measure visual word significance. We formulate the visual word similarity from the perspective of visual word transition behavior, incorporating the visual word occurrence relationships in local regions. Then the significance of visual words is discovered with the idea of PageRank [9]. Visual words with significance value above the mean will be regarded as distinctive visual words.

2.1. Visual word similarity formulation

Visual word is a kind of visual concept atom and intrinsically related through images. Usually, visual concept is composed of a set of visual atoms under some geometric constraints. Therefore, it is necessary to incorporate the local geometric relationship among visual words to analyze the link context between visual words. Maximally stable extremal regions (MSER)[2] can serve for this task. Usually, MSER detector generates some regions per image with high repeatability. Such region intrinsically imposes local geometric constraints for its contained visual words.

A four-layer directed graph is constructed (Fig. 1), which contains two intermediate layers: image layer and MSER layer. A visual word V_i first transits to an image containing V_i, then to the MSER region detected in the image, and finally to another visual word V_j sharing the same MSER region. The intuition behind is that if a user is viewing an image, he or she is most likely attracted by some local features, and other features in the neighborhood may also be of interest.

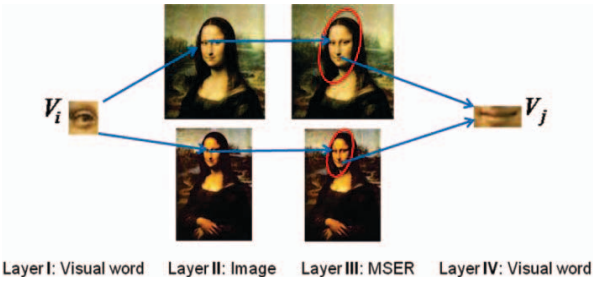

Fig. 1. An illustration of four layers between two nodes in the visual word link graph. The red ellipses in layer three denote the detected MSER regions.

Based on the visual word graph, we essentially define the visual word similarity with a propagation behavior in terms of probability as follows,

$$W^{VW}(i,j) = P(V_j \mid V_i) = \sum_I P(V_j \mid I)P(I \mid V_i) = \sum_I \sum_M P(V_j \mid M) \cdot P(M \mid I) \cdot P(I \mid V_i) \quad (1)$$

where M denotes the set of visual words in a MSER region, $P(V_j \mid M)$ denotes the normalized term-frequency of the visual word V_j in M, $P(M \mid I)$ denotes the normalized MSER-frequency of M in the image I, and $P(I \mid V_i)$ denotes the inverse image frequency of visual word V_i for image I.

2.2. Visual word rank

After obtaining visual word similarity, similar to PageRank [9], the significance vector of all visual words, R, is iteratively defined as follows,

$$R = d \cdot W^* \cdot R + (1-d) \cdot p \text{, where } p = \left[\frac{1}{n}\right]_{n \times 1} \quad (2)$$

where W^* is the column-normalized version of the transposition of W^{VW}, p is a distracting vector for random walk behavior and d is a constant damping factor. Initially, each entry in R is set the same as p. The iteration of Eq. (2) is considered converged when the change of R is small enough or a maximal iteration number is reached. Finally, we can obtain a significance value for each visual word. We define distinctive visual words as those with significance value greater than the average significance value.

3. Image rank analysis

To remove some noisy images and select a good image candidate pool, we also propose to measure image similarity by an image link graph and employ random walk re-ranking method [9] to calculate image significance.

3.1 Image similarity formulation

Generally, images are related through intermediate medium of visual words, which work as visual hyperlinks. And, different visual words will pose different weights to the image that contains them. These context relationships can be represented with an oriented graph of three-layer, as illustrated in Fig. 2. Instead of direct transition, an image first transmits to a visual word contained in it, and then to another image that shares the same visual word.

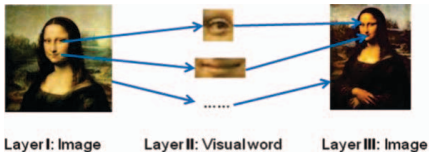

Fig. 2. An illustration of three layers between two nodes in the image link graph.

Based on the above discussion, the image transition probability can be defined as follows,

$$P(I_j \mid I_i) = \frac{1}{N_i} \sum_{V} P(I_j \mid V) \cdot P(V \mid I_i) \cdot f(R(V)) \tag{3}$$

where $P(I_j \mid V)$ is the inverse image frequency of visual word V for image I_j, $P(V \mid I_i)$ denotes the normalized term frequency of visual work V in image I_i, $R(V)$ is the significance value of visual word V obtained by Eq. (2), $f(\cdot)$ is a non-decreasing function, such as $f(x) = \exp(x)$, N_i is a normalization factor.

To avoid the images with too many features have large probability propagated from other images, a regularization term should be included for penalizing. We formulate the image visual similarity as follows,

$$W^{Img}(i,j) = Prob(I_j \mid I_i) \cdot \tau(I_j) \tag{4}$$

where $\tau(I_j)$ is the regularization term defined as

$$\tau(I_j) = \frac{1}{\sqrt{N(I_j) + \varepsilon}} \tag{5}$$

where $N(I_j)$ denotes the number of features in the *j-th* image and ε is a constant.

3.2 Image rank

Similar to visual word rank discussed in Section 2.2, the significance of image S^{Img} is also iteratively defined as follows,

$$S^{Img} = d \cdot U \cdot S^{Img} + (1-d) \cdot p \text{ ,where } p = [1/n]_{n\times 1} \tag{6}$$

where U is the column-normalized version of the transposition of W^{Img}, p and d are the same as that in Eq. (2). Finally, all images are re-ranked according to their significance values and the top-ranked ones are selected into the candidate pool of canonical images.

4. Canonical image selection

With the selected distinctive visual words and good candidate images, we can perform canonical image selection. Ideally, the canonical images should be representative to the query, and exhibit a diverse set of views. We adopt a Weighted Set Coverage (WSC) scheme to select multiple search-related canonical images, considering orthogonality and coverage [4].

We denote the candidate image pool generated in Section 3 as $S = \{I_1, I_2, \ldots, I_m\}$ and the set of distinctive visual word elements contained in these selected images as $X = \{v_1, v_2, \ldots, v_n\}$. Each visual word v_i has a weight $w(v_i)$, and each image I_i has an incremental weight c_i and an importance weight u_i. u_i is defined to be equal to $VR^{img}(I_i)$, the significance value of I_i, obtained from Eq. (6). The definitions of $w(v_i)$ and c_i are described as follows.

$$w(v_i) = \begin{cases} \sqrt{VR^{vw}(v_i)} & if \;\; VR^{vw}(v_i) > \bar{s} \\ 0 & else \end{cases} \tag{7}$$

where $\bar{s}$ is the average of all visual word significance value $VR^{vw}(v_i), (i = 1,2,..n)$.

$$c_i = \sqrt{\frac{\#(VW(I_i \setminus G) \cap X)}{\#(VW(I_i))}} \tag{8}$$

where G is a subset of S, representing the selected canonical images. $VW(I_i \setminus G)$ denotes the set of distinctive visual words only covered by image I_i but not covered by any image in G, and $\#(VW(I_i))$ denotes the total number of visual words covered by I_i.

Algorithm 1: Weighted Set Coverage (G, S, X)
Input: S, a set of images; X, a set of visual word elements.
Output: G, a subset of S with maximum coverage of good visual words.
Procedure:
1. $G = \Phi$.
2. Repeat
(1) Select $I_i \in S$, that maximize $TW_i \cdot c_i \cdot u_i$;
(2) $G \leftarrow G \cup \{I_i\}$, $S \leftarrow S \setminus \{I_i\}$
(3) If all Good visual words are covered by G, stop.

Further, we define TW_i as the total sum of weight of the distinctive visual words in $VW(I_i \setminus G)$,

$$TW_i = \sum_{v_i \in VW(I_i \setminus G)} w(v_i) \tag{9}$$

Finally, our Weighted Set Coverage is performed as Algorithm 1.

5. Experiments

We use real-world query results to test our method. For data preparation, we use Google Image search to retrieve images with 30 landmark text queries and crawl the first 500 returned images for each query. Typical queries include Colosseum, Eiffel Tower, and Golden Gate Bridge *etc.* For image representation, a standard implementation of SIFT [7] is used. To quantize local features to visual words, visual vocabulary tree [8] is adopted.

In our experiments, we select the top-ranked 150 images to construct the candidate pool for canonical photos selection. Two baseline methods are used for comparison. The first one is the top-*K* retrieved images from the text-based search results. The second baseline employs Affinity Propagation (AP) clustering on the same top 150 images. Clusters are ranked by the inter-cluster distance, and the cluster exemplars are selected as canonical images.

Fig. 3 shows three sample results obtained with our approach. To quantitatively evaluate the performance, motivated by [4], we implement a subjective evaluation of the top five results for each method and ask 5 users to answer the following three evaluation questions: (1) *Representative*: How many photos are representative to the query (0–5)? (2) *Redundant*: How many photos are redundant (0–5)? (3) *1st place voted*: Which canonical image set (among the three methods) is the most satisfactory? The results for each question are averaged over all users and 30 queries.

Table 1 shows a summary of the subjective evaluation results. As for representative score, WSC shares very similar performance to the other two baselines. However, WSC performs much better in terms of redundancy and 1st place voted scores. The consideration of orthogonality and coverage helps WSC obtain the smallest redundant score. It shows that our method can select canonical images, which are not only representatives of the collected photos, but also exhibit a diverse set of views with minimal redundancy.

6. Conclusion

In this paper, we propose a new scheme for canonical image selection with latent visual context learning and weighted set coverage. Our method can select a small set of representative images with diversity. Experiments with landmark datasets demonstrate the effectiveness of the proposed approach.

In the future, we will incorporate topic model to enrich the visual context learning and combine other image metadata for canonical image selection. Furthermore, more comprehensive user-experience evaluations will also be conducted.

Fig. 3. Top five representative images for three queries, "Colosseum", "The Sphinx", and "Mount Rushmore National Memorial", obtained by WSC. Each row corresponds to one query.

Table 1: Comparison of variant representative image selection methods by subjective evaluation.

method	Representative	Redundant	1st place voted
Top-*K*	4.60	2.14	21.3%
AP	4.47	2.33	16.0%
WSC	4.72	1.14	62.7%

7. Acknowledgement

This work was supported in part by NSFC under contract No. 60632040, Program for New Century Excellent Talents in University (NCET), Research Enhancement Program (REP) and start-up funding from the Texas State University.

References

[1] A. Jaffe and M. Naaman, "Generating summaries and visualization for large collections of geo-referenced photographs," *ACM MIR*, pp. 89-98, 2006.

[2] J. Matas, *et al*. "Robust wide baseline stereo from maximally stable extremal regions," BMVC, 2002.

[3] R. Raguram and S. Lazebnik, "Computing Iconic Summaries for General Visual Concepts," *CVPR*, 2008.

[4] Y.-H. Yang, *et al*, "ContextSeer: Context Search and Recommendation at Query Time for Shared Consumer Photos," *ACM Multimedia*, 2008.

[5] I. Simon, N. Snavely, *et al*, "Scene Summarization for Online Image Collections," In Proc. *ICCV*, pp. 1-8, 2007.

[6] L. Kennedy, et al, "Generating diverse and representative image search results for landmarks," In *WWW*, 2008.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 60(2), pp. 91-110, 2004.

[8] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," In *Proc. CVPR*, pp.2161-21682006.

[9] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems,* 30(1-7), pp. 107-117, 1998.