Saliency-GD: A TF-IDF Analogy for Landmark Image Mining

Wei Li, Jianmin Li(✉), and Bo Zhang

State Key Laboratory of Intelligent Technology and Systems, TNList, Department of Computer Science and Technology, Tsinghua University, Beijing, China
lijianmin@mail.tsinghua.edu.cn

Abstract. In this paper we address the problem of unsupervised landmark mining, which is to automatically discover frequently appearing landmarks from an unstructured image dataset. Landmark mining often suffers from false matches resulted from cluttered backgrounds and foregrounds, inter-class similarities, and so on. Analogous to TF-IDF in image retrieval, we propose the Saliency-GD weighting scheme of visual words, which can be easily integrated into state-of-the-art local-feature-based visual instance mining frameworks. Saliency detection provides feature weighting in image space from the attention perspective, and in feature space, the knowledge of geographic density (GD) transferred from a separate training dataset gives a multimodal selection of meaningful visual words. Experiments on public landmark datasets show that Saliency-GD weighting scheme greatly improves the landmark mining performance with increasing discrimination power of visual features.

Keywords: Landmark mining · Saliency · Geographic Density Weighting scheme

1 Introduction

This paper addresses the problem of unsupervised landmark image mining. Its goal is to automatically discover all frequently appearing landmarks from an unstructured image dataset. The mining problem differs from image retrieval [15] and duplicate image detection [21], since the latter two problems have a query as entry point and output only one cluster. It is also different from [10,11,23], as there is no additional information (such as annotations or captions) available in the mining dataset. Figure 1 illustrates some landmark images.

Landmark mining often suffers from false matches resulted from cluttered backgrounds/foregrounds, inter-class similarities, and so on. Appropriate weighting strategy can be effective in selecting discriminative local features. However, related works [2,3,12,22] usually use simple TF-IDF or ignore this issue. TF-IDF can be viewed as a weighting scheme of visual words in two dimensions, the image space and the feature space. However, in image space, we find that

B. Zeng et al. (Eds.): PCM 2017, LNCS 10735, pp. 477–486, 2018.
https://doi.org/10.1007/978-3-319-77380-3_45

Fig. 1. Example images of the landmarks in Oxford [15] (first row) and Paris [16] (second row) datasets

most visual words only appear once in each image. Moreover, those visual words which do appear multiple times in an image often represent meaningless repetitive shapes. These facts restrict the usefulness of TF in landmark mining. In feature space, although document frequency proved to be a good feature selection strategy, we ask a more interesting question: is it possible to learn better weighting from another image dataset with multimodal information?

In this work we propose Saliency-GD weighting scheme, a TF-IDF analogy for landmark image mining. In image space we leverage saliency detection to estimate the importance of each local feature from the attention perspective, while in feature space we train a novel geographic density (GD) measure to select meaningful visual words in a multimodal way, on an Internet image dataset which contains both visual and geographic information. In the literature saliency detection is mostly used in image manipulation [6], image retrieval [8], and image segmentation [18], and is rarely applied to unsupervised settings. Geographic information, on the contrary, has shown its effectiveness in some mining problems such as [4,5,17]. But our formulation differs from all these works: we do not require the *mining* dataset to contain geographic data. Instead we transfer the geographic knowledge from another separate *training* dataset.

With our elegantly designed parameterless weight mapping, the Saliency-GD weighting scheme can be seamlessly integrated into existing local-feature-based mining frameworks. Taking the state-of-the-art instance graph (IG) framework [12] as an example, experiments on two public landmark datasets show that our novel weighting scheme brings superior mining performance with increasing discrimination power of visual features.

The rest of the paper is organized as follows. Section 2 introduces the Saliency-GD weighting scheme in detail. The integration into the instance graph framework is discussed in Sect. 3. Section 4 presents the experimental results, followed by conclusions and future work in Sect. 5.

2 The Saliency-GD Weighting Scheme

Similar to TF-IDF, our scheme gives weights to visual words in two dimensions. Saliency and geographic density (GD) correspond to image space and feature space, respectively.

2.1 The Image Space: Saliency

One source of false matches in landmark mining is the cluttered environment, in which lies many irrelevant visual elements, such as trees, roads, and pedestrains. We observe that for most landmark images, the most salient object is the main body of the landmark. Therefore we leverage saliency detection algorithm to effectively differ the landmark from distracting environment elements.

We adopt the state-of-the-art salient object detection algorithm [1] to compute saliency map for each image in the mining dataset. In order to convert the pixel-level saliency given by the algorithm to feature-level saliency, we compute the weighted average of all saliency values in the local feature region. Specifically, the saliency S_f of a local feature f at location (x, y) with radius r is defined as

$$S_f = \frac{\sum_{(x',y')\in N(x,y)} e^{-\frac{2\sqrt{(x'-x)^2+(y'-y)^2}}{r}} S(x', y')}{\sum_{(x',y')\in N(x,y)} e^{-\frac{2\sqrt{(x'-x)^2+(y'-y)^2}}{r}}}, \tag{1}$$

where $N(x, y) = \{(x', y')|(x' - x)^2 + (y' - y)^2 \leq r^2\}$ is the set of points that covered by the local feature and $S(x', y')$ is the saliency value at coordinate (x', y').

In practice this computation could be slow. Instead we use a sample of points in $N(x, y)$ to approximate the above computation. In the experiments we select the central point (x, y), the four points in each direction with distance $r/2$ to the central point, and the four points in each direction with distance r to the central point.

2.2 The Feature Space: Geographic Density

In feature space, our preliminary experiments find that there are similar patterns between document frequencies trained on the mining dataset and those trained a separate Internet image dataset, and they have similar contributions to the improvement of landmark mining performance. This shows that knowledge of local features can be transferred across datasets. We are interested in whether we can transfer more knowledge about visual words if more information is available in the training dataset. In this work we answer this question with the usage of geographic information. Among the multimodal information available for Internet images, we select geographic information because it is widely available and relatively accurate compared to other forms of data such as tags, as new smart phones and cameras are mostly equipped with GPS modules. We argue that for unsupervised or weakly-supervised tasks, the availability and accuracy of these data matter the most.

Given a training dataset with geographic data, we design a novel measure to combine the document frequency and the geographic information, called geographic density (GD), that gives each visual word a weight. We first compute two simpler measures: the traditional document frequency (DF) and our proposed geographic frequency (GF). DF, as in TF-IDF, is the number of images where the visual word appears. To compute GF, we partition the latitude-longitude coordinate system of earth surface by a 5-degree-by-5-degree grid, and count the number of cells where the visual word appears. The intuition is that the less cells a visual word appears in, the more geographically discriminative it is. GD is then defined as DF divided by GF, thus captures both information. We call it *density* because this value represents the average number of visual word occurrences in geographic cells (ignoring cells with no occurrence of the specific word). In Sect. 4 we compare all three measures (DF, GF, and GD) and verify the effectiveness of this combination. Details of the training dataset are also discussed in Sect. 4.

3 Integration into the IG Framework

We first have a brief review of the instance graph framework. It first extracts local SIFT [13] features on detected Hessian-Affine regions [14], quantizes them [19], and augments each local feature with Hamming Embedding (HE) [9] and its neighboring features. Then it builds a weighted undirected graph with images as vertices. The matching scores between augmented local features, which combine the HE similarity [9] and the Jaccard similarity of neighboring visual word sets, contribute to the weights of edges. Finally instance clusters are efficiently discovered by a greedy breadth-first search algorithm on the sparse graph.

We focus on its computation of similarity score s_{ij} between two augmented local features f_i and f_j. In IG framework, it only reflects the *matching* score, but ignores the *discrimination* score. Our Saliency-GD weighting scheme can introduce the latter by replacing the similarity measure s_{ij} with $s'_{ij} = s_{ij}d_{ij}$, where d_{ij} represents the discrimination power of the two corresponding central features.

The remaining problem is how to compute d_{ij}. A good weighting scheme should have the following properties: (a) it does not bring large impact to the parameters of the existing framework; (b) it does not introduce any new parameters; and (c) it can effectively filter out very indiscriminative features while does not weigh too much on top discriminative ones. As the raw weights computed in Sect. 2 have different ranges of possible values, we propose a unified parameterless weight mapping which have all above properties. In general, for a feature f_k, there will be a mapped image space weight u_k and a mapped feature space weight v_k, both in the range $[0, 2)$. More specifically, in image space, the saliency map value (ranging from 0 to 255) is linearly mapped to $[0, 2)$. In feature space, we first sort the original weights of all visual words and take their rankings as an intermediate measure, then we map this ranking, also linearly, to $[0, 2)$. Note that there may be multiple visual words having the same original weight, and

in this case we set their rankings to the same value, i.e., the average of their rankings. Finally the discrimination score is computed by $d_{ij} = u_i u_j v_i v_j$, taking both features and both spaces into consideration.

Note that although in this work we take the IG framework as an example to illustrate our weighting scheme, with the above weight mapping, Saliency-GD can also work in other frameworks, for example thread of features [22].

4 Experiments

In this section we present our experimental results to demonstrate the effectiveness of the Saliency-GD weighting scheme.

4.1 Datasets and Evaluation Measures

We evaluate our proposal by mining on two public landmark datasets, Oxford [15] and Paris [16], both with ground truth for image retrieval tasks. Similar to [22], for each landmark, we take the union of all query images and result lists as a ground truth cluster for our mining problem. Statistics of these two datasets are shown in Table 1 and example landmark images are demonstrated in Fig. 1.

Table 1. Statistics on the two landmark datasets

Dataset	# images	# clusters	# images per cluster
Oxford	5062	11	78.8 ± 95.2
Paris	6392	11	312.7 ± 170.2

For the GD training dataset, we choose MediaEval13 [7], which contains a large number of Flickr images and their high-accuracy GPS locations and user tags. However, we discover that in the original dataset there are many images irrelevant to landmark mining, for example selfies and indoor activity images. Therefore we propose two steps to filter this dataset to make it more appropriate for our landmark mining task. First we adopt the Haar cascade face detection algorithm [20] to filter out those images that have faces in them. Second we filter out images with tags that describe certain indoor activities, such as birthday, party, wedding, etc. For fair comparison in the experiments, we also remove images taken in the areas of Oxford and Paris, to avoid possible duplicate images or appearance of the same landmark. After filtering we randomly sample 60k images to form our final training dataset. The size of the dataset is set empirically, as we observe that a larger value will not bring further improvement to the result.

Same as [12,22], we adopt pair measures (P_{pr}, R_{pr}, and F_{pr}) for quantitative evaluation. However, there are a number of other visual instances existing in

Oxford and Paris datasets, and most of them are not landmarks. To focus our evaluation on landmark mining, instead of using existing fully-annotated ground truth [22], we simply ignore image pairs consisting of two out-of-ground-truth images. This can be viewed as a sampling of the result. We conducted separate experiments to compare this sampling strategy with the use of full ground truth on Oxford dataset, and validated that the sampling can give a reasonable evaluation for landmark mining. To differ between fragmented clusters and a single big cluster in the results, we also compute the cluster measures (P_{cr}, R_{cr}, and F_{cr}) proposed in [12].

4.2 Quantitative Comparison

We take the original instance graph framework as the baseline (Baseline) and compare some variants of our weighting scheme with it, including image space saliency (Saliency), feature space document frequency (DF), geographic frequency (GF), and geographic density (GD), and also the full weighting scheme (Saliency-GD). Quantitative comparison of all these methods is summarized in Table 2.

Table 2. Quantitative comparison of different methods with best parameters. The best results are marked with bold font

Dataset	Method	P_{pr}	R_{pr}	F_{pr}	P_{cr}	R_{cr}	F_{cr}	# search hits
Oxford	Baseline	0.393	0.358	0.374	0.866	0.384	0.533	49144
Oxford	Saliency	0.458	0.381	0.416	0.959	0.392	0.556	51762
Oxford	DF	0.451	0.416	0.433	0.906	0.395	0.550	49819
Oxford	GF	0.427	0.418	0.422	0.881	0.399	0.549	53709
Oxford	GD	0.450	0.448	0.449	0.902	0.420	**0.574**	54950
Oxford	Saliency-GD	0.560	0.417	**0.478**	0.910	0.384	0.540	**42559**
Paris	Baseline	0.367	0.199	0.258	0.905	0.295	0.445	196087
Paris	Saliency	0.373	0.226	0.281	0.896	0.304	0.454	154732
Paris	DF	0.435	0.203	0.277	0.904	0.297	0.447	154924
Paris	GF	0.388	0.210	0.273	0.853	0.311	0.456	168156
Paris	GD	0.426	0.215	0.286	0.891	0.318	**0.469**	175394
Paris	Saliency-GD	0.483	0.210	**0.293**	0.905	0.314	0.467	**126641**

It is obvious that all weighting schemes, even the simplest ones, result in improvement of landmark mining performance in both datasets. For the pair measures, we observe that GD usually have a better effect than saliency, and the combination of them achieves the best result of all. This proves that knowledge from image space and feature space is complementary. Furthermore, in feature space, we find that GD has superior performance over the simpler DF

and GF, showing the power of multimodal information. Note that, because the IG framework naturally considers TF during local region matching, the results of DF in the table are actually those of the TF-IDF weighting scheme, which proved the superiority of Saliency-GD over TF-IDF. Another observation is that the Saliency-GD weighting scheme brings a larger performance boost in Oxford dataset than in Paris dataset (27.8% versus 13.6% relative improvement, respectively). This is because that Oxford dataset contains similar buildings and false match is a key reason to its low precision, which is addressed well by our weighting scheme. However, for Paris dataset, most landmarks have unique appearance, and a more dominant factor is its large cluster size, causing low recall and preventing further performance gain. For the cluster measures, GD gives the best result. This is likely due to the fact that saliency filters a large number of less informative features, leading to a high precision but having negative impact on recall, which results in more fragmented clusters.

The last column of Table 2 shows the number of search hits during the GBFS instance clustering step, a measure of redundancy in result clusters. It is clear that besides superior performance, the Saliency-GD weighting scheme also successfully reduces the redundancy, which is crucial to real-life applications.

Figure 2 plots the Pair F-Measure curves of the two datasets with varying Hamming threshold, the most sensitive parameter in the original framework. For most cases, the adoption of a weighting scheme requires a less strict Hamming threshold to achieve its best performance. This is consistent with our expectation, as a larger threshold allows more candidates to be inspected, while the precision is ensured by the weighting scheme. For Oxford dataset, due to its moderate cluster sizes and large number of false matches, weighting schemes can also reduce its parameter sensitivity, especially for large Hamming thresholds.

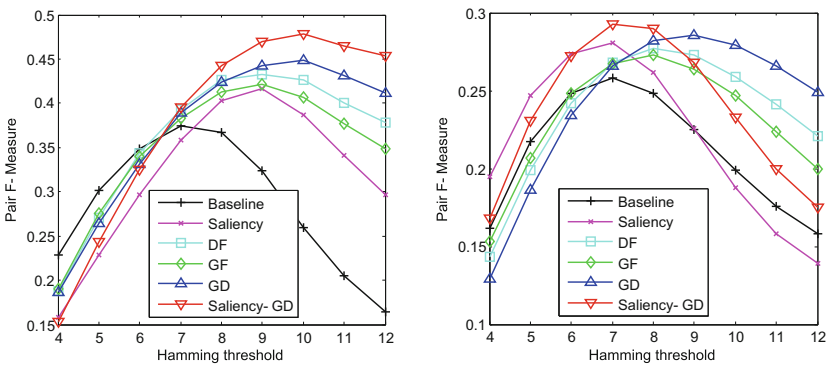

Fig. 2. The Pair F-Measure curves of (a) Oxford and (b) Paris datasets with varying Hamming threshold

4.3 Qualitative Evaluation

Figure 3 visualizes different weighting schemes. From the examples we can see that saliency effectively differs the landmarks from the cluttered environments, and GD selects meaningful regions, usually complex structures. Full Saliency-GD

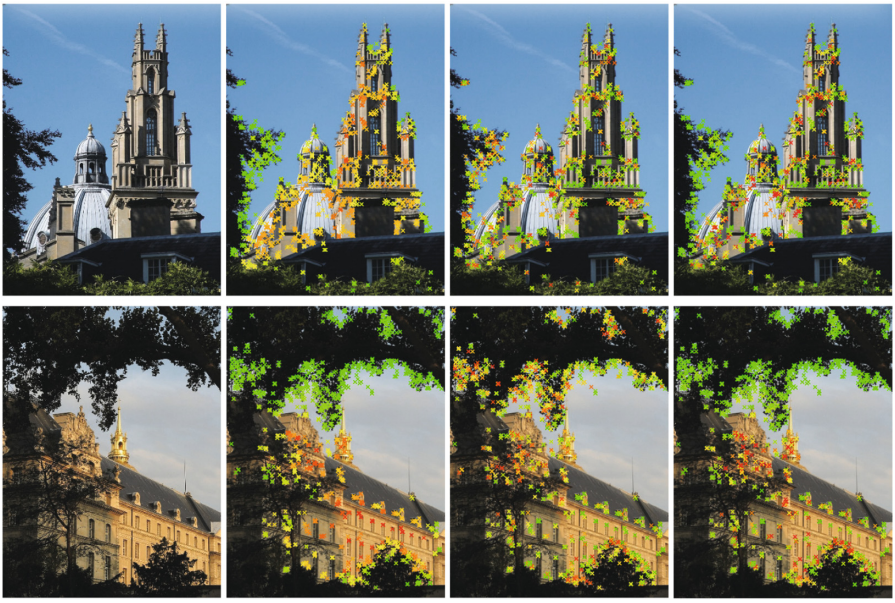

Fig. 3. Visualization of different weighting schemes on two example images. From left to right, the four columns show the original image, saliency only weighting scheme, GD only weighting scheme, and full Saliency-GD weighting scheme, respectively. Each marker shows a local feature and its color demonstrates the weight (red represents a large weight while green means a small one) (Color figure online)

Fig. 4. Example false matches in the baseline. The similarity scores corresponding to these image pairs are all reduced to less than 0.01 with our weighting scheme

weighting scheme combines the advantages of both: low weights are assigned to trees, windows, and other common components of landmarks, while discriminative local features are given more importance during matching.

Figure 4 demonstrates the elimination of false matches, including similar components of landmarks (for example, windows), similar low-level patterns and structures, and texts. For each image pair, the baseline gives a high matching score, but with the Saliency-GD weighting scheme, these false matches are effectively filtered. All shown image pairs receive weighted similarity scores below 0.01.

5 Conclusions

This work proposes the Saliency-GD weighting scheme of visual words for unsupervised landmark mining, in order to cope with the challenges of false matches resulted from cluttered backgrounds/foregrounds, inter-class similarities, as well as other reasons. Analogous to TF-IDF, our weighting scheme works in both image and feature space. In image space saliency provides feature selection from the attention perspective, and in feature space geographic density (GD) gives a multimodal filtering of meaningful visual words by transferring knowledge from a separate training dataset. This scheme can be easily integrated into existing local-feature-based visual instance mining frameworks. Experiments on two landmark datasets show that it brings superior landmark mining performance with increasing discrimination power of visual features.

Besides geographic information, the integration of more multimodal data, such as tags and user profiles, is a promising direction for future works. Some other possible extensions of this work include the construction of a high-quality training dataset, weighting in a higher-level (for example feature group or image patch level), and adaption of the weighting scheme to deep features.

Acknowledgment. This work was supported by the National Basic Research Program (973 Program) of China (No. 2013CB329403), and the National Natural Science Foundation of China (Nos. 61332007, 91420201 and 61620106010).

References

1. Cheng, M.M., Mitra, N., Huang, X., Torr, P., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2015)
2. Chum, O., Matas, J.: Large scale discovery of spatially related images. IEEE Trans. Pattern Anal. Mach. Intell. **32**(2), 371–377 (2010)
3. Chum, O., Perdoch, M., Matas, J.: Geometric min-hashing: finding a (thick) needle in a haystack. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17–24 (2009)
4. Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: World Wide Web Conference (WWW), pp. 761–770 (2009)

5. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes Paris look like Paris. ACM Trans. Graph. **31**(4), 101:1–101:9 (2012)
6. Goldberg, C., Chen, T., Zhang, F.L., Shamir, A., Hu, S.M.: Data-driven object manipulation in images. Comput. Graph. Forum **31**(2), 265–274 (2012)
7. Hauff, C., Thomee, B., Trevisiol, M.: Working notes for the placing task at MediaEval 2013. In: MediaEval Workshop (2013)
8. He, J., Feng, J., Liu, X., Cheng, T., Lin, T.H., Chung, H., Chang, S.F.: Mobile product search with bag of hash bits and boundary reranking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3005–3012 (2012)
9. Jegou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. Int. J. Comput. Vis. **87**(3), 316–336 (2010)
10. Li, H.: Multimodal visual pattern mining with convolutional neural networks. In: ACM International Conference on Multimedia Retrieval (ICMR), pp. 427–430 (2016)
11. Li, H., Ellis, J., Ji, H., Chang, S.F.: Event specific multimodal pattern mining for knowledge base construction. In: ACM International Conference on Multimedia, pp. 821–830 (2016)
12. Li, W., Wang, C., Zhang, L., Rui, Y., Zhang, B.: Scalable visual instance mining with instance graph. In: British Machine Vision Conference (BMVC), pp. 98:1–98:11 (2015)
13. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
14. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. Int. J. Comput. Vis. **60**(1), 63–86 (2004)
15. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
16. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
17. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: ACM International Conference on Image and Video Retrieval (CIVR), pp. 47–56 (2008)
18. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1939–1946 (2013)
19. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision (ICCV), pp. 1470–1477 (2003)
20. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2001)
21. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 25–32 (2009)
22. Zhang, W., Li, H., Ngo, C.W., Chang, S.F.: Scalable visual instance mining with threads of features. In: ACM International Conference on Multimedia, pp. 297–306 (2014)
23. Zhu, Z., Xu, C.: Organizing photographs with geospatial and image semantics. Multimed. Syst., 1–9 (2016)