Uncovering the Effect of Visual Saliency on Image Retrieval

Qinjie Zheng[1,2,3], Shikui Wei[1,2,3](✉), Jia Li[1,2,3], Fei Yang[1,2,3], and Yao Zhao[1,2,3]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
shkwei@bjtu.edu.cn

[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China

[3] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Abstract. Visual saliency modeling has achieved impressive performance for boosting vision-related systems. Intuitively, it should be beneficial to content-based image retrieval task, since the users' query attention is heavily related to the region of interests (ROI) in query image. Although some approaches have been proposed to combine image retrieval systems with visual saliency models, no a comprehensive and systematic study is made to discover the effect of different saliency models on image retrieval in a qualitative and quantitative manner. In this paper, we attempt to concretely investigate the diversity of visual saliency models on image retrieval by making extensive experiments based on nine popular saliency models. To cooperatively mining the complementary information from different models, we also propose a novel approach to effectively involve visual saliency into image retrieval systems by a learning process. Extensive experiments on a generally used image benchmark demonstrate that the new image retrieval system remarkably outperforms the original one and other traditional ones.

Keywords: Visual saliency · Image retrieval · Evaluation

1 Introduction

Visual saliency estimation is a mechanism simulating human vision system to detect the conspicuous content in an image, and lots of excellent saliency models have been proposed [14,15,18,22,28]. Due to good properties of visual saliency, it has been widely used in many fields, such as abstract extraction, classification, compression, monitoring [4]. In recent years, many researchers devote to introduce visual saliency into image retrieval to improve the searching accuracy. Formally, Content-Based Image Retrieval (CBIR) aims at effectively and efficiently finding out the needed images from a large-scale image database, which has achieved great progress in past two decades. Generally speaking, most of

J. Yang et al. (Eds.): CCCV 2017, Part II, CCIS 772, pp. 170–179, 2017.
https://doi.org/10.1007/978-981-10-7302-1_15

existing image retrieval methods attempt to improve image retrieval performance from the following three aspects: (1) constructing discriminative image features [6,12,26]; (2) designing good similarity estimation schemes [19,30]; (3) handling large-scale issues [2,5,13,21,31]. Since the user's query attention is heavily related to the specific regions in query image, the visual saliency is considered to be good information for boosting image retrieval accuracy. For example, Acharya *et al.* [1] directly employed Itti saliency model [11] to generate saliency map and then extracted feature vectors from the saliency map for image retrieval. Similarly, some researchers [3] exploited an image segmentation and color histogram based saliency model to extract saliency map and then extracted image features from the map. Papushoy *et al.* [23] employed GBVS visual attention model [10] to extract saliency map and introduced the salient information into region-level based image retrieval system. Similar approaches are also reported in [9,27]. In their work, the salient information is involved by weighting different regions according to their perceived saliency. In [20,25], a histogram of saliency map is extracted as separated image feature, and it is integrated into original similarity measure of image retrieval system. Although these approaches have been proposed to combine image retrieval systems with visual saliency models, no a comprehensive and systematic study is made to discover the effect of different saliency models on image retrieval in a qualitative and quantitative manner.

To concretely investigate the diversity of visual saliency models on image retrieval, we conduct extensive experiments based on nine popular saliency models. To cooperatively employing the complementary information from different models, we also propose a novel approach to effectively involve visual saliency into image retrieval systems by a learning process.

Our main contributions can be summarized as follows:

1. **Extensive Experimental Studies:** Through the experimental studies based on nine classic saliency models, we explicitly evaluate the effect of visual saliency on image retrieval and discover some effective manners of involving the salient information, which will be beneficial to many computer vision problems.
2. **A Novel Learning-based Saliency Involving Approach:** We propose a novel learning-based approach to optimally involve visual saliency into image retrieval systems. From the leaning perspective, we provide a new framework for optimally involving salient information.

2 Extensive Experimental Studies

In this section, we conduct extensive experiments based on nine popular saliency models to discover the relationship between visual saliency and image retrieval. More specifically, a popular image retrieval framework based on local image features is employed to implement baseline image retrieval systems. Under this framework, two reasonable schemes are designed to involve salient information into image retrieval process and the optimal combination of saliency model and involving schemes is identified experimentally.

2.1 The Image Retrieval Framework

We employ the popular BoW-based image retrieval framework. As a kind of local based framework, it can not only better balance the effectiveness and efficiency but also provide more flexibility for involving salient information. In Fig. 1, we sketch key steps of the framework. In particular, a visual codebook is first constructed by employing some clustering methods [12,26]. Based on the visual codebook, an orderless collection can be built for one image by replacing the image's local features with their nearest visual words. After inserting each visual word in image database and its corresponding image ID into the inverted table, we can perform image retrieval process after given a query image. We call this scheme BoW model. If Hamming embedding code, which encodes the quantization error between local feature and its visual word, is inserted into the inverted table with image ID, we call it BoW+HE scheme. In our experimental studies, both schemes are exploited. Based on the image retrieval framework, two saliency involving schemes are employed to introduce salient information into image retrieval process.

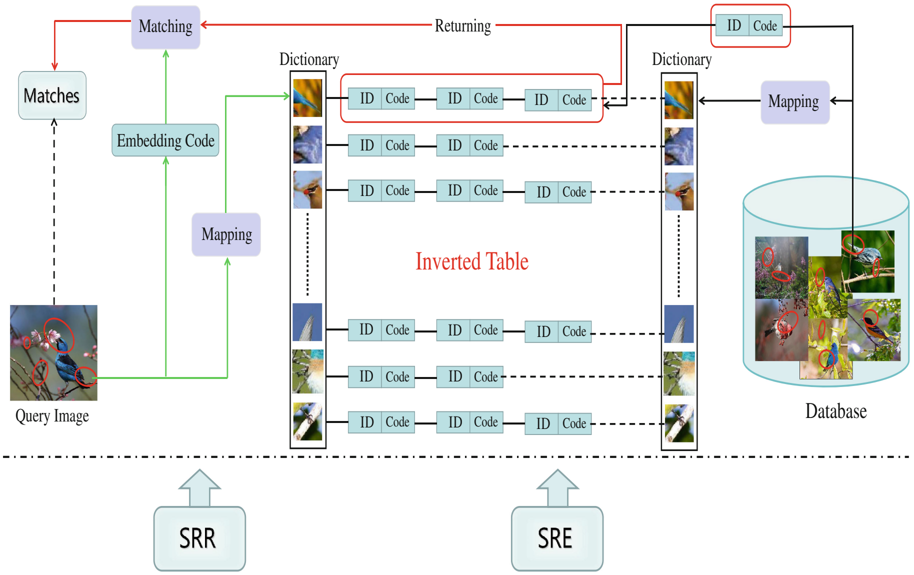

Fig. 1. The overall framework. In order to build an inverted indexing structure, each key point in database images is first mapped to the nearest visual word, and then its image ID without (BoW) or with (BoW+Embedding) embedding code is inserted into the list corresponding to the visual word. In the online query stage, each key point in the query image is also mapped to the nearest visual word, and the items in the corresponding list are returned as matches. If embedding codes are employed, the returned list will be further refined and only top n items whose are most similar to query key point in distances among their codes are returned as matches.

2.2 Dataset and Saliency Involving Schemes

INRIA Holidays Dataset [12]**:** It is a commonly used image benchmark in image retrieval area, and it contains 500 image groups with different scenes or objects. For each group, there are several images, and the total number of images in all groups is 1491. To evaluate image retrieval task, the first image in each group is treated as query, which results in a query set with 500 images. The other 991 images are treated as database images.

As indicated in Fig. 1, there are two manners to introduce visual salient information into the local-based image retrieval framework. They are listed as follows:

- **Saliency Region Representation Embedding (SRE):** A 4-dimensional vector is extracted for each image based on the whole saliency map. In particular, the first component is the average value of H channel values of all salient points, and the rest three components correspond to S channel, I channel, and sum of three channels, respectively. All the key points in an image are associated with the same vector. Given any key point in the query image, all its similar items returned from the inverted table is sorted in ascending order by the distances between the 4D vector of the query key point and the 4D vectors of database key points. Only top n items are selected and assigned to a big weight.
- **Saliency Region Representation Re-ranking (SRR):** Given a query image, the returned database images are re-ranked by the distances between the 4D vector of the query image and the 4D vectors of database images. Different from SRE, SRR is a kind of global saliency involving scheme.

2.3 A Quantitative Study on Saliency Involving Schemes

As discussed above, different saliency models will result in quite different saliency maps. To evaluate their effectiveness on improving the image retrieval quality, we conduct extensive experiments by individually employing 9 state-of-the-art saliency models (*i.e.*, AWS [8], BMS [29], GBVS [10], HFT [17], ITTI [11], LDS [7], RARE [24], SP [16], SSD [15]). The parameters for all the experiments are optimal. The experimental results are illustrated in Fig. 2. The first row is the results from BoW model with different combinations of saliency models and involving methods, and the second row is from the BoW+Embedding method. In order to clearly show the effect of different saliency models, the cases without any salient information (BL) also illustrated as baselines.

For BOW model, whatever saliency involving methods we adopt, almost all the saliency models outperform the baseline system. This means that the saliency maps extracted from existing models indeed reflect the true human salient information more or less. For these saliency models, LDS achieves the best performance, compared all the other models with any saliency involving scheme. The highest MAP is achieved at the combination of LDS and SRE schemes, which is up to 0.540. It is higher than Baseline (0.456) by 8.4% points.

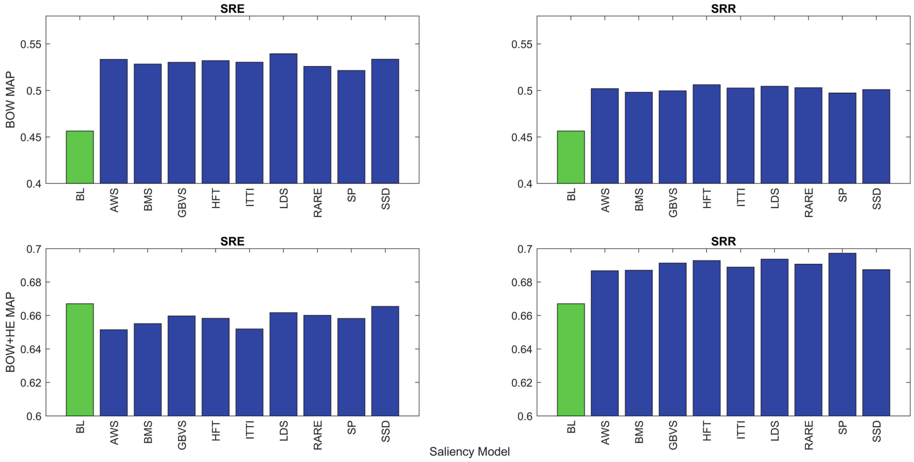

Fig. 2. Evaluation on various combinations of saliency models and saliency involving schemes in two image retrieval methods. The green lines in the first row denote the performance of BoW baseline system, whose MAP is 0.456. The green lines in the second row denote the performance of BoW+HE baseline system, whose MAP is 0.667. The yellow lines mean that the saliency involving schemes are combined with salient information labeled manually. (Color figure online)

For BOW+HE model, SRE involving scheme cannot provide positive effect on the baseline system. That is, BOW+HE model heavily depends on the saliency involving schemes. When we introduce Hamming embedding codes into image retrieval system, the searching accuracy has been improved significantly, compared BoW baseline (0.456) with BoW+HE baseline (0.667). In this situation, if saliency maps are not accurate enough, it will remarkably degrade the positive effect of Hamming embedding codes. According to our experiments, only the SRR scheme can play a positive role in image retrieval boosting. For saliency models, SP model performs the best. The highest MAP value is up to 0.697 and is higher than the baseline (0.667) by 3% points. In fact, LDS model still works well, which achieves nearly the same performance with SP.

From the 4 sub figures in Fig. 2, SRR saliency involving approach provides stable and consistent improvement for all saliency models in both image retrieval methods. That is, saliency involving approach plays an important role when introducing salient information into the image retrieval framework.

3 Learning-Based Saliency Involving Approach

In real-world image retrieval scenario, it is impossible to manually obtain saliency maps for a large-scale image database. Therefore, most of existing methods directly employed one of saliency models to extract saliency map to approximate the human vision system. However, different saliency models will result in quite different saliency maps, which can be treated as different approximations of true saliency map. In addition, the performance of different saliency

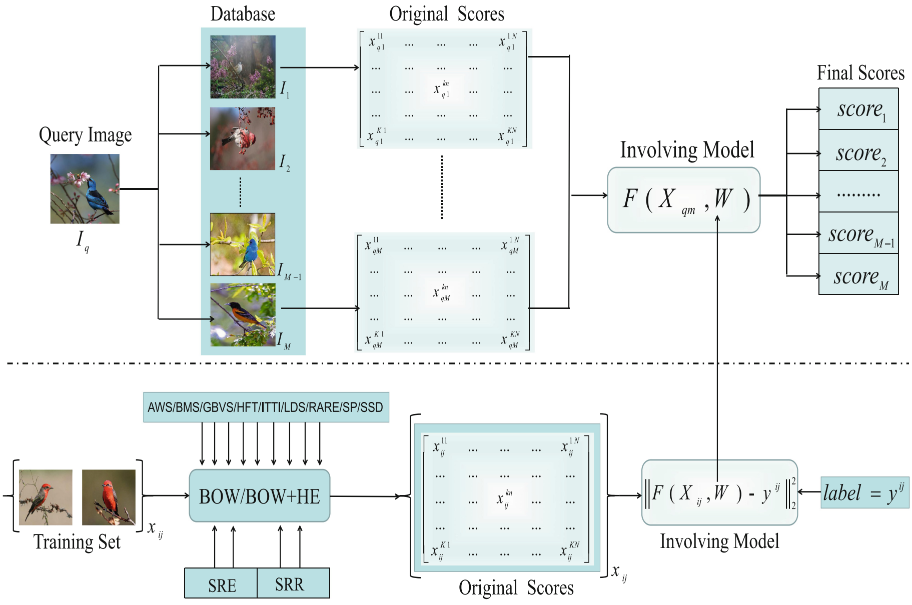

Fig. 3. Illustration of the proposed learning-based saliency involving approaches. It is divided into two parts, *i.e.* online involving and offline learning. In the offline learning stage, the similarity scores of image pairs in training set are calculated by employing all possible combinations of saliency model and involving schemes. Then, they are utilized with pair labels to train an optimal involving model. In the online stage, all scores between query image and any database image are first estimated and then are employed to obtain the a final score by involving model.

involving schemes also varies with different image retrieval schemes. If we can find a saliency involving approach that obtains an optimal complement of various saliency models and involving schemes, we can better improve the image retrieval performance of original search engines. Toward this end, we propose a novel learning-based saliency involving approach. Figure 3 illustrates the key idea of the proposed approach.

Formally, we suppose that there are N different saliency models $M = \{M_1, M_2, ...M_N\}$ and K saliency involving methods $\Phi = \{\Phi_1, \Phi_2, ...\Phi_K\}$. Given an image pair (x_i, x_j), we can get their similarity score matrix $\mathcal{X}_{ij}$ by employing different combinations of saliency models and saliency involving methods, which can be formulated as follows:

$$f_{KN}(x_i, x_j) = \mathcal{X}_{ij} = \begin{bmatrix} x_{ij}^{11} & \dots & \dots & \dots & x_{ij}^{1N} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & x_{ij}^{kn} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_{ij}^{K1} & \dots & \dots & \dots & x_{ij}^{KN} \end{bmatrix} \tag{1}$$

where x_{ij}^{kn} is the two images' similarity score obtained by combining the n^{th} saliency model and the k^{th} saliency involving method.

Our aim is to learn a weight matrix W, which can optimally involve the salient information from all combinations and provide a more reliable similarity estimation between two images. Toward this end, we propose a new similarity estimation function, which is formulated as follows:

$$\mathcal{F}(\mathcal{X}_{ij}, W) = tr(W^T \mathcal{X}_{ij})$$

$$W = \begin{bmatrix} w^{11} & ... & ... & ... & w^{1N} \\ ... & ... & ... & ... & ... \\ ... & ... & w^{kn} & ... & ... \\ ... & ... & ... & ... & ... \\ w^{K1} & ... & ... & ... & w^{KN} \end{bmatrix} \tag{2}$$

where $\mathcal{F}(\mathcal{X}_{ij}, W)$ can be treated as the new similarity score between two images x_i and x_j, and w^{kn} is the weight of similarity score x_{ij}^{kn}.

To learn the weight matrix W, we must construct a training set of triplets $\{(x_i, x_j; y^{ij})\}_{i,j=0}^{L}$, where y^{ij} is one if images x_i and x_j are truly relevant, otherwise zero. Our approach attempts to approximate the score $\mathcal{F}(\mathcal{X}_{ij}, W)$ to the relevant label y^{ij}, which can be formulated by minimizing the following approximation error:

$$\mathcal{L}(x_i, x_j; y^{ij}) = \|\mathcal{F}(\mathcal{X}_{ij}, W) - y^{ij}\|_2^2 \tag{3}$$

The overall objective function is defined as follows:

$$\min_{\mathbf{W}} \sum_{i=1}^{L} \sum_{j=1}^{L} \mathcal{L}(x_i, x_j; y^{ij}) + \lambda \|W\|_F^2 \tag{4}$$

where L is the number of training images and λ is the parameter controlling the sparse term.

4 Experiments

4.1 Experimental Setup

The INRIA Holidays dataset is employed for evaluation. For each image, 9 state-of-the-art saliency models are employed individually to extract saliency maps. In addition, 2 saliency involving methods above-mentioned are exploited to evaluate the performance of different combinations. To further show the scalability of the proposed learning-based saliency involving approach, a large-scale distracted image dataset, *i.e.*, Flickr1M dataset, is employed in our large-scale image retrieval experiments.

4.2 Evaluation on Learning-Based Saliency Involving Scheme

To obtain an optimal complement of various saliency models and involving approaches, we propose a learning-based scheme. In this section, we conduct some experiments to evaluate its effectiveness. To facilitate the experiments, the saliency involving approach is fixed to re-ranking scheme (*i.e.*, SRR) due to its stable performance. In addition, we only employ the BoW+HE framework, since boosting its performance is more challenging.

Table 1. Performance of 9 saliency models and the learning-based scheme

Method	BL	AWS	BMS	GBVS	HFT	ITTI	LDS	RARE	SP	SSD	**Proposed**
mAP	0.667	0.687	0.687	0.691	0.693	0.689	0.694	0.691	0.697	0.687	**0.701**

The experimental results are shown in Table 1. Clearly, the proposed learning-based scheme outperforms the best performance of all traditional methods. The possible reason lies in that the optimal process generates a more complete saliency map than GND saliency map. This means that the learning-based scheme can even complement the labeling error from human labelers.

4.3 Large-Scale Image Retrieval Experiments

To evaluate the scalability of the proposed learning-based scheme, we conduct some experiments on large-scale image database. In our experiments, we only employ BoW+HE framework, and all the parameters are set to be optimal. The experimental results are demonstrated in Table 2. As expected, after introducing the salient information by using the proposed scheme, the final retrieval accuracy is remarkably improved comparing the baseline system. This means that the proposed scheme can work well on large-scale image retrieval tasks.

Table 2. Performance on large-scale image retrieval scenario

Method	BOW+HE	**BOW+HE+Learning**
mAP	0.257	**0.296**

5 Conclusion

In this paper, we make comprehensive and systematic study to discover the essential relation between image retrieval and visual saliency. Specially, we explicitly discover the effect of visual saliency on image retrieval in a quantitative manner. The key finding is that salient information indeed has positive effect on image retrieval and the manner of introducing salient information play an important role on performance boosting. According to the finding, we propose a novel approach to effectively involve visual saliency into image retrieval systems by a learning process. Extensive experiments on a generally used image benchmark demonstrate that the new image retrieval system remarkably outperforms the original one and the learning-based visual saliency involving approach is also better than the traditional ones. In addition, large-scale experiments show good scalability of the proposed approach.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (No. 61572065, No. 61532005, No. 61370113), National Key Research and Development of China (No. 2016YFB0800404), Joint Fund of Ministry of Education of China and China Mobile (No. MCM20160102).

References

1. Acharya, S., Devi, M.R.V.: Image retrieval based on visual attention model. Procedia Eng. **30**, 542–545 (2012)
2. Babenko, A., Lempitsky, V.: Efficient indexing of billion-scale datasets of deep descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2055–2063 (2016)
3. Boato, G., Dang-Nguyen, D.T., Muratov, O., Alajlan, N., Natale, F.G.B.D.: Exploiting visual saliency for increasing diversity of image retrieval results. Multimedia Tools Appl. **75**(10), 1–22 (2015)
4. Chiappino, S., Mazzu, A., Marcenaro, L., Regazzoni, C.S.: A bio-inspired logical process for saliency detections in cognitive crowd monitoring. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2110–2114 (2015)
5. Dai, Q., Li, J., Wang, J., Jiang, Y.G.: Binary optimized hashing. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 1247–1256. ACM (2016)
6. Duan, L., Ma, W., Miao, J., Zhang, X.: Visual saliency based bag of phrases for image retrival. In: The ACM Siggraph International Conference, pp. 243–246 (2014)
7. Fang, S., Li, J., Tian, Y., Huang, T., Chen, X.: Learning discriminative subspaces on random contrasts for image saliency analysis. IEEE Trans. Neural Netw. Learn. Syst. **PP**(99), 1–14 (2016)
8. Garciadiaz, A., Leborn, V., Fdezvidal, X.R., Pardo, X.M.: On the relationship between optical variability, visual saliency, and eye fixations: a computational approach. J. Vis. **12**(6), 17 (2012)
9. Giouvanakis, E., Kotropoulos, C.: Saliency map driven image retrieval combining the bag-of-words model and PLSA. In: 19th International Conference on Digital Signal Processing, pp. 280–285 (2014)
10. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems, pp. 545–552 (2007)
11. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Patt. Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
12. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_24

13. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. ArXiv Preprint ArXiv:1702.08734 (2017)
14. Lee, G., Tai, Y.W., Kim, J.: Deep saliency with encoded low level distance map and high level features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 660–668 (2016)
15. Li, J., Duan, L.Y., Chen, X., Huang, T., Tian, Y.: Finding the secret of image saliency in the frequency domain. IEEE Trans. Patt. Anal. Mach. Intell. **37**(12), 1 (2015)
16. Li, J., Tian, Y., Huang, T.: Visual saliency with statistical priors. Int. J. Comput. Vis. **107**(3), 239–253 (2014)
17. Li, J., Levine, M.D., An, X., Xu, X., He, H.: Visual saliency based on scale-space analysis in the frequency domain. IEEE Trans. Patt. Anal. Mach. Intell. **35**(4), 996–1010 (2013)
18. Liang, X., Xu, C., Shen, X., Yang, J., Tang, J., Lin, L., Yan, S.: Human parsing with contextualized convolutional neural network. IEEE Trans. Patt. Anal. Mach. Intell. **39**(1), 115–127 (2017)
19. Liao, L., Wei, S., Zhao, Y., Gu, G.: Improving the similarity estimation via score distribution. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2016)
20. Liu, G.H., Yang, J.Y., Li, Z.Y.: Content-based image retrieval using computational visual attention model. Patt. Recogn. **48**(8), 2554–2566 (2015)
21. Liu, R., Zhao, Y., Wei, S., Zhu, Z., Liao, L., Qiu, S.: Indexing of CNN features for large scale image search. ArXiv Preprint ArXiv:1508.00217 (2015)
22. Liu, Y.J., Yu, C.C., Yu, M.J., He, Y.: Manifold SLIC: a fast method to compute content-sensitive superpixels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 651–659 (2016)
23. Papushoy, A., Bors, A.G.: Image retrieval based on query by saliency content. Dig. Sig. Process. **36**, 156–173 (2015)
24. Riche, N., Mancas, M., Gosselin, B., Dutoit, T.: Rare: a new bottom-up saliency model. In: 19th IEEE International Conference on Image Processing, pp. 641–644. IEEE (2012)
25. Wan, S., Jin, P., Yue, L.: An approach for image retrieval based on visual saliency. In: International Conference on Image Analysis and Signal Processing, pp. 172–175 (2009)
26. Wei, S., Xu, D., Li, X., Zhao, Y.: Joint optimization toward effective and efficient image search. IEEE Trans. Cybern. **43**(6), 2216–2227 (2013)
27. Wen, Z., Gao, J., Luo, R., Wu, H.: Image Retrieval Based on Saliency Attention. Springer, Heidelberg (2014)
28. Zhang, D., Han, J., Li, C., Wang, J., Li, X.: Detection of co-salient objects by looking deep and wide. Int. J. Comput. Vis. **120**(2), 215–232 (2016)
29. Zhang, J., Sclaroff, S.: Exploiting surroundedness for saliency detection: a boolean map approach. IEEE Trans. Patt. Anal. Mach. Intell. **38**(5), 889–902 (2016)
30. Zheng, L., Wang, S., Wang, J., Tian, Q.: Accurate image search with multi-scale contextual evidences. Int. J. Comput. Vis. **120**(1), 1–13 (2016)
31. Zhou, W., Li, H., Sun, J., Tian, Q.: Collaborative index embedding for image retrieval. IEEE Trans. Patt. Anal. Mach. Intell. **PP**, 1 (2017)