Visual Saliency Fusion Based Multi-feature for Semantic Image Retrieval

Jianan Chen[1], Cong Bai[1(✉)], Ling Huang[1], Zhi Liu[2], and Shengyong Chen[1]

[1] College of Computer Science, Zhejiang University of Technology, Hangzhou, China
congbai@zjut.edu.cn
[2] School of Communication and Information Engineering,
Shanghai University, Shanghai, China

Abstract. In this paper, a saliency fusion based content-based image retrieval method is proposed. Different saliency detection methods were conducted firstly and the output saliency maps were fused by double low rank matrix recovery method. Then the images were segmented into foreground and background according to the fusion result. As the foreground and background had the different impacts on the semantic understanding of the image, different features represented in the form of histogram were extracted. Finally, a fusion of z-score normalized Chi-Square distance is adopted as the similarity measurement. This proposal has been implemented on three widely used benchmark databases and the results evaluated in terms of mean Average Precision (mAP), precision, recall, and $F1$-measure show that our proposal outperforms the referred state-of-the-art approaches.

Keywords: Saliency fusion · Bag of words
Semantic image retrieval · Chi-square

1 Introduction

With the development of visual technology, more attention have been put on image retrieval both in the industry and the research community. Image retrieval techniques could be widely classified into two categories: text-based and content-based. The text-based approach index images in database by key-words, which came from manually added annotation. However, manual annotation is an imprecise and time consuming job. Content-based image retrieval (CBIR) searches images by their own visual contents, which has been presented in the early 1990s [15]. However, how to extract meaningful features from the large collections of image data is still a challenging problem due to the deviation of semantic understanding between human and computer [9]. Semantic gap exists between low-level handcrafted features and high-level human perception [8]. The reason is that a highly evolved human brain could transform the visual signals into concrete subject, while the computer couldn't do that with high accuracy up to now.

© Springer Nature Singapore Pte Ltd. 2017
J. Yang et al. (Eds.): CCCV 2017, Part II, CCIS 772, pp. 126–136, 2017.
https://doi.org/10.1007/978-981-10-7302-1_11

Many researches have been dedicated into semantic image retrieval, comprehensive reviews could be found in [12,15]. Bag-of-words (BOW) framework, which is initially proposed in [17], is the most famous one among them. The key idea of BoW is to quantify each local feature into one or more so-called visual words, and each image is represented as a set of unordered visual words. Our work also used BOW partially. Furthermore, we noted that the semantic understanding of the image could be divided into two parts: the foreground object and the background regions in general. That means that retrieval results possibly met the high level perception could be got if foreground and background were represented individually by different features. In our previous work [1], image segementation based on RC-saliency [3] was used to segment the foreground and background in the image, and got a better performance in semantic image retrieval. Although RC-saliency has a good effect in the direction of color division, it has some limitations in texture and shape. That means it could only perform well in special kinds of images. So it is far from the broadness of human vision. Thus segmentation based on one visual saliency model could have the limitations on its universality, which could decreases the performance of the retrieval. Obviously, better retrieval results could be got if the performance of segmentation improved. So a saliency fusion based multi-feature model is take into consideration in our proposal.

In this paper, a novel semantic image retrieval method named saliency fusion based multi-feature (SFMF) is proposed. Firstly, we figure out seven saliency maps generated by different methods, and fused them by double low rank matrix recovery method (DLRMR) [7]. Secondly, SaliencyCut [3] based on the fused saliency map is used to segment images into foreground objects and background regions. Finally, local features and global features extracted from both foreground objects and background regions with different weights in the similarity fusion are used for retrieval.

The remaining of this paper is organized as follows. Section 2 describes SFMF in details. Experiment results and analysis are given in Sect. 3, followed by conclusion in Sect. 4.

2 The Proposed SFMF

There are three phrases in the SFMF: fusion stage, offline processing and online retrieval. In the fusion stage, segmentation based on DLRMR saliency fusion is performed on each database image, and thus each image is divided into two parts: foreground object and background region. After the segmentation by fusion saliency maps, images are represented by a multi-feature representation. Furthermore, between the foreground objects and background regions, different features and different weights in the similarity are considered in the retrieval.

2.1 Fusion Stage

The framework of the segmentation based on saliency fusion is illustrated in Fig. 1. Different from the onefold saliency map, saliency fusion intended to

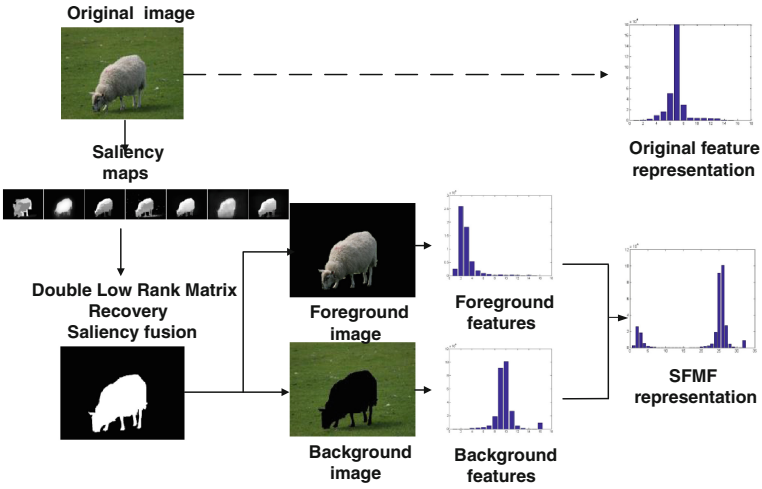

Fig. 1. Segmentation based on saliency fusion. Segmentation is executed by SaliencyCut using saliency map generated by saliency fusion by double low rank matrix recovery. After segmentation, background region and foreground object are obtained and represented by different features.

combine various saliency detection methods makes the fusion results better than each individual saliency detection methods. It highlights the advantages of several algorithms and avoids the weakness of a few algorithms so that the final saliency map obtained as a result of fusion outperforms each of them. For the reasons mentioned above, double low rank matrix recovery(DLRMR)[7] is used to cast object and background decomposition problem. Furthermore, different features are used to represent foreground objects and background regions, which makes the SFMF representation to be more characteristic than the solo feature representation of the whole image.

First of all, seven saliency detection methods: AMC [6], BL [18], BSCA [16], HC [3], MR [20], MS [19], ST [13] have been chosen to obtain saliency maps $\{S_k|1 \leq k \leq 7\}$.

Then mean shift algorithm [4] segmented the image into regions $\{P_i\}_{i=1,...,n}$, where n is the number of super-pixels. The saliency map S_k could be represented using an n-dimensional vector $\mathbf{X}_k = [x_{1k}, x_{2k}, ..., x_{nk}]^T$, the i^{th} element of the vector corresponds to the mean of the saliency values of pixels in the super-pixel P_i. By arranging $\mathbf{X}_k$ into a matrix, we get the combined matrix representation of individual saliency maps as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_7]$.$\mathbf{X} \in R^{n \times 7}$. With super-pixel instead of pixel as the smallest unit to calculate more in line with content semantics visual.

Generally, a natural image $\mathbf{I}$ could be decomposed as:

$$\Phi(\mathbf{I}) = \mathbf{A} + \mathbf{E}, \tag{1}$$

where Φ indicates a certain transformation, $\mathbf{A}$ and $\mathbf{E}$ denote matrices corresponding to background and foreground. Then treat matrix $\mathbf{X}$ as a feature representation of the image $\mathbf{I}$ in the saliency feature space, with each row representing a super-pixel feature vector [7]. Equation (1) could be rewritten as:

$$\mathbf{X} = \mathbf{A} + \mathbf{E}, \tag{2}$$

Therefore, saliency fusion could be cast as a low rank affinity pursuit. Given matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_7]$. $\mathbf{X} \in R^{n \times 7}$, the low rank matrix recovery problem could then be formulated as:

$$\min_{\mathbf{A},\mathbf{E}} rank(\mathbf{A}) + \lambda(rank(\mathbf{E})) \quad s.t.\ \mathbf{X} = \mathbf{A} + \mathbf{E}, \tag{3}$$

where parameter $\lambda > 0$ balances the effects between two ranks [7].

Through the alternating direction method of multipliers(ADMM) [2], the final low rank $\mathbf{E}$ measures the contribution of each saliency method and learns an adaptive combination of the maps. It counts the value of E in every region on each saliency map and separate the object area from the background with a suitable threshold, which is recommended 1.4 times around average valuer.

Some comparison examples of the segmentation using the saliency maps fused by DLRMR and generated by one single visual saliency model, such as RC-saliency are shown in Fig. 2. From these examples, we could observe that saliency map fusion could get better segmentation performance than single visual saliency model.

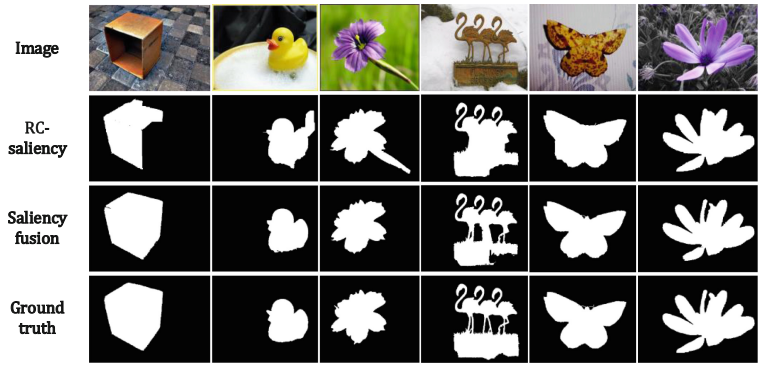

Fig. 2. Given the input image and the ground truth, the segmentation using the saliency maps generated by DLRMR are better than by one saliency model such as RC-saliency, the method used in our previous work.

2.2 Offline Processing

After the images have been segmented into foreground objects and background regions, different feature will be extracted considering their different characters, which is distinct from traditional image retrieval methods.

For the background regions, features are extracted in HSV color space. As they had large areas of similar colors and textures in general, local binary patterns (LBP) in V channel and color histograms in H and S channel were extracted as color and texture features. We choose these two features not only they are simple but also efficient.

For the foreground objects, beside the texture and color features as extracted from the background, local features should also be considered. So the Scale-invariant feature transform (SIFT) [14] is chosen as an instinct choice for its successful achievements in object retrieval task. SIFT feature is packed in the BOW framework for retrieval. That means SIFT features extracted from the images will be compared with a visual words vocabulary clustered by K-means and the frequency of the visual words appearance in the image will be used as the representation of the image.

The features extracted from the images are defined as formula (4),

$$F \begin{cases} F_f = (H_h, H_s, LBP_v, SIFT_g), \\ F_b = (H_h, H_s, LBP_v), \end{cases} \tag{4}$$

where F_f is features of foreground image, and F_b is features of background image, H_h and H_s are histogram features in hue (H) and saturation (S) channel of HSV color space, LBP_v is histogram of local binary patterns statistics in value (V) channel in HSV color space, where $SIFT_g$ is the histogram of visual words in gray level space. We set the parameters in advance, the weight of F_b is less than F_f.

2.3 Online Retrieval

The input query image is also needed to segment based on saliency fusion and represented by different features in the way mentioned in Section 2.2. A fusion of the z-score normalized chi-square distances is proposed to measure the similarity between the query and the images in the database.

The chi-square distance between the histogram of query image H_Q and the histogram of image from the database H_I is defined as:

$$DS(Q, I) = \sum_{j=1}^{K} \frac{(H_Q(j) - H_I(j))^2}{H_Q(j) + H_I(j)}, \tag{5}$$

where K is the number of bins in the histogram.

Since different histograms are constructed for an image, the similarity between the images is measured by fusing the distances of histograms with different weights. However, normalization before fusion is necessary because each histogram is composed of different feature vectors. Additionally, in order to avoid errors introduced by outliers, distances are normalized through the following ways: given a query image, by calculating distances on one type of histogram between this query and all images from the database, one set of distances

$\{DS_A(Q, I_i)\}$ obtained, where $i = \{1, 2, ..., P\}$ and P is the number of images in the database. Thus the normalized distance is defined as:

$$DS_A^N(Q, I_i) = \frac{DS_A(Q, I_i) - \mu_{A_Q}}{\sigma_{A_Q}}, \tag{6}$$

where μ_{A_Q} and σ_{A_Q} are the mean value and the standard deviation of the distances set $\{DS_A(Q, I_i)\}$ respectively.

Finally, all kinds of distances are fused as one distance to decide the similarly between the images and the query. In the stage of distances fusion, not all of the distances have the same weight, more weights on the LBPs distances of background and foreground are given. In particularly, 3 times on foreground and 2 times on background is the better choice as an experimental choice.

The SFMF algorithm framework is summarized in Algorithm 1.

Algorithm 1. SFMF framework

Input: A query image

1: Conduct individual saliency detection methods and obtain S_k.

2: Compute saliency fusion map by DLRMR and segment the image into foreground and background.

3: Extract visual features F_f from foreground object.

4: Extract visual features F_b from background regions.

5: Combine F_f and F_b to obtain F.

6: Measure the distance between query image F and images in dataset by z-score normalized chi-square distance.

Output:

Image retrieval results ordered by the distance.

3 Experiments

3.1 Image Dataset

The experiments were executed on three publicly available and widely used benchmark database: Corel 5k [10], VOC 2006 [5] and Corel 10k [11].

Corel 5k contains 50 themes with 100 images for each of size $192 * 128$ or $128 * 192$ in JPEG format. 8 themes were selected as the compared methods did, which have an obvious object in the image. The selected themes including "bear", "pyramid", "building", "plane", "snowberg", "horse", "tiger" and "train". Among those images, 200 images are used to train the vision dictionary, and 800 images for testing.

In Pascal VOC 2006, ten themes were selected: "sheep", "motorbike", "cow", "horse", "dog", "bus", "car", "person" and "bicycle" and 1000 images are selected randomly covered all these ten themes. Same with the Corel 5k, 200 images are used to train the visual words dictionary, and the rest of them are used for testing.

To prove the versatility of the algorithm, a large scale image database, Corel 10 is chosen. Corel 10k has expanded the number of pictures in the B set to reach 10000 images in total, and contains 100 categories from diverse contents such as sunset, beach, flower, building, car, horses, mountains, fish, food, door, etc.

Some examples of queries from these databases and its retrieval results are demonstrated in Fig. 3. From these examples, we could see that the proposed algorithm could get very good retrieval results obviously for the images containing salient objects with clear background as the salient objects were firstly segmented from the background regions, such as 'Poker' and 'Gun' in the figure. Furthermore, the proposed algorithm could also get good results for the images containing salient objects with complicated background, as 'Hippo' and 'Terraces' in the figure.

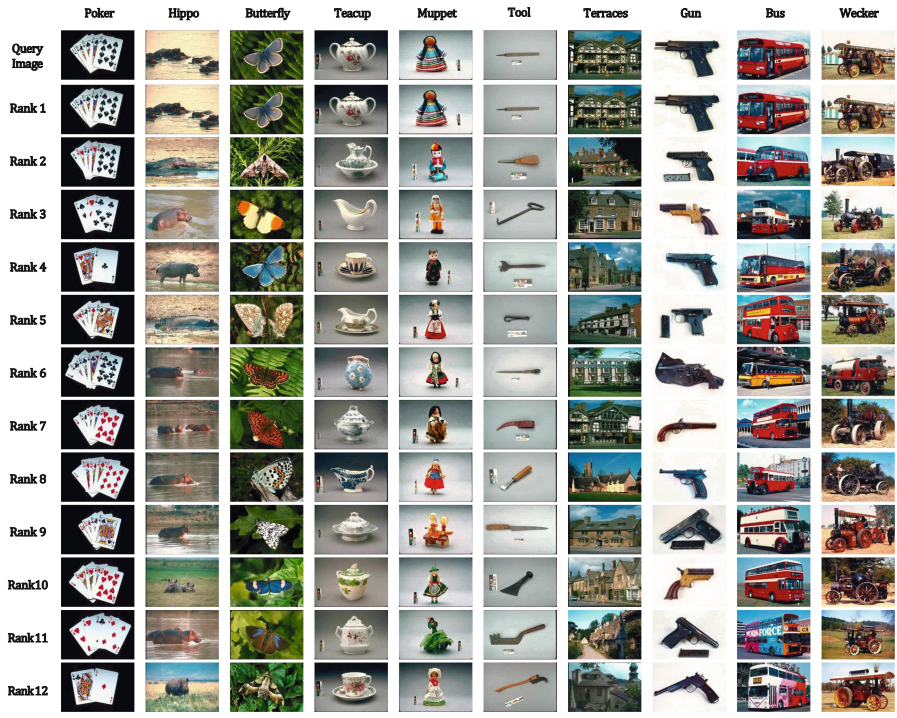

Fig. 3. Examples of query and its top 12 retrieved images

3.2 Experimental Measurement

We compute the bounded mean Average Precision (mAP) to count how many positive images at top K relevant results. The mAP at bound K is defined as follows:

$$mAP_K = \frac{\sum_{i=1}^{N} AP_K}{N}, \qquad 1 \leqslant K \leqslant N, \tag{7}$$

where N is the number of query topics, K is the top K retrieval results considered. And AP at bound K is defined as follows:

$$AP_K = \frac{\sum_{i=1}^{K} \rho_i P_i}{N_{GT}}, \tag{8}$$

where K is computing depth of AP, ρ_i is a boolean function defined as follows:

$$\rho_i = \begin{cases} 1, \ the\ i-th\ query\ result\ is\ correct, \\ 0, \ the\ i-th\ query\ result\ is\ incorrect, \end{cases} \tag{9}$$

and P_i is precision of top i results, N_{GT} is the number of positive samples in top K query results from Ground Truth [21].

Other two metrics are precision and recall. These two metrics are often combined as the weighted harmonic mean, namely F-measure, and it is an overall performance measure [11]. It could be defined as follows:

$$\begin{cases} F = \frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R}, \\ P = \frac{I_N}{N}, \\ R = \frac{I_N}{M}, \end{cases} \tag{10}$$

In the experiments of image retrieval, $precision(P)$ is the ratio of the number of retrieved similar images to the number of retrieved images, while $recall(R)$ is the ratio of the number of retrieved similar images to the total number of similar images. Where I_N is the number of retrieved similar images, N is the total number of images retrieved and M is the total number of similar images in the database. The coefficient β allows one to weight either precision or recall more heavily, and they are balanced when $\beta = 1$. If there is no particular reason to favor precision or recall, $\beta = 1$ is commonly used to image retrieval or information retrieval. Parameters are set as the same as [11], $N = 12$; $M = 100$ and $\beta = 1$ on Corel-10k dataset, Thus, F-measure is so called F_1-measure.

3.3 Comparison

BOW could be seen as the baseline of content-based image retrieval, especially for the semantic image retrieval. RoI-BOW [21] is a recently reported method that improved the BOW model with the region segmentation. With these segmentation, images could be seen as the combination of the regions of interest (RoI) and the regions of Non-RoI. SSH and MSD are also recently reported by Liu [10,11]. These algorithms used saliency model also and got a good results in image retrieval field. SBMF is our earlier work [1], thus we selected the BOW, RoI-BOW, SSH, MSD and SBMF for comparisons.

Tables 1 and 2 show the mAP at top ten and top twenty retrieved results for each theme in Corel 5K and VOC 2006 database respectively. From the table, we could conduct the conclusion that the performance of the proposed SFMF outperforms in most of the themes selected and archives the best performance overall.

Table 1. Comparison of the mAP on Corel 5K

Themes	Top 10 (%)				Top 20 (%)			
	SFMF	SBMF	ROI	BOW	SFMF	SBMF	ROI	BOW
Pyramid	**98.57**	97.78	94.08	81.13	**96.91**	96.12	90.33	70.76
Bear	**90.79**	82.44	83.78	73.57	**83.55**	69.71	66.71	60.67
Building	94.21	**94.25**	87.11	76.76	89.55	**89.76**	78.50	63.83
Horse	**99.69**	97.60	98.61	74.54	**98.94**	94.72	95.69	60.44
Plane	**93.14**	90.14	89.26	86.47	**87.96**	84.00	79.79	78.93
Snowberg	**90.54**	89.22	88.17	75.94	**84.79**	82.82	79.68	59.30
Tiger	**95.95**	90.93	94.21	77.22	**92.00**	83.92	86.34	62.99
Train	**90.36**	89.20	90.86	73.73	**85.13**	82.30	86.34	56.69
Average	**94.16**	91.45	90.76	77.42	**89.85**	83.14	82.92	64.20

Table 2. Comparison of the mAP on VOC 2006

Themes	Top10 (%)				Top20 (%)			
	SFMF	SBMF	ROI	BOW	SFMF	SBMF	ROI	BOW
Bicycle	**97.85**	91.80	86.70	85.44	**95.92**	85.07	74.78	73.62
Bus	**92.40**	85.12	83.47	81.90	**86.48**	76.15	70.34	63.65
Car	**99.64**	97.39	84.47	79.42	**99.02**	95.24	69.92	62.17
Cat	**79.56**	77.30	71.82	75.97	**66.13**	61.87	55.21	58.31
Cow	**90.00**	89.62	81.82	81.84	79.59	**81.40**	71.72	61.94
Dog	82.80	**82.98**	76.51	74.99	**66.13**	65.15	60.64	59.90
Horse	**82.19**	76.27	73.19	77.06	**69.20**	61.06	56.43	58.00
Motorbike	**82.06**	78.68	77.72	77.57	**70.41**	66.21	59.33	60.62
Person	80.56	**80.71**	81.43	79.15	**66.20**	65.28	62.48	58.18
Sheep	**96.62**	93.82	84.48	77.06	**93.46**	89.70	68.39	57.82
Average	**88.38**	85.37	80.16	79.04	**79.25**	74.71	64.92	61.42

To further verify the improvement, experiments are conducted on Corel 10k database also, the results are shown in Table 3. From this table, we could see that the proposed SFMF algorithm achieves the better performances than the BOW baseline, MSD, SSH and SBMF.

3.4 Running Efficiency

Although the proposed SFMF uses multiple features in retrieval, this proposal scheme has high efficiency in terms of vector space and runtime. The length of the feature vector of the foreground objects is 366: 256 of LBP, 30 of SIFT, 80 of color histogram; and the length of the feature vector of the background regions is

Table 3. The precision, recall, F-measure of five methods on Corel 10k

	precision(%)	recall(%)	F-measure(%)
BOW	30.36	3.64	6.51
MSD	45.62	5.48	9.78
SSH	54.88	6.58	11.76
SBMF	59.98	7.20	12.85
SFMF	**74.59**	**8.95**	**15.98**

336: 256 of LBP, 80 of color histogram. So the total length of the feature vector for the retrieval is 702.

The running time evaluation experiments are implemented in Matlab2013b on the laptop with a Core I7-4720 processor with 8 GB memory. The running time of segmenting and training visual words reaches about 2.6 s per image and 1.4 s per image respectively, and the running time of retrieval is about 0.5 s per query. It should be noted that the segmentation and training could be done off-line prior to retrieval.

4 Conclusion and Perspectives

A novel semantic image retrieval method named SFMF was proposed, which integrated saliency fusion method and traditional image feature representation. Images were segmented into foreground objects and background regions by saliency maps generated by visual saliency fusion, then different features were extracted in consideration of different characteristics of foreground and background. Experiments implemented on three widely used databases have proved the amelioration of the segmentation based on visual saliency fusion could lead the better performance.

However, it also should be noted that the calculation load of the segmentation and visual saliency fusion is still heavy. Furthermore, more efficient features vector needs to be discovered in the future as the size of LBP constitutes the large proportion of the feature vector. These two issues could be discovered perspectively.

Acknowledgments. This work is supported by Natural Science Foundation of China under Grant No. 61502424, 61471230, U1509207 and 61325019, Zhejiang Provincial Natural Science Foundation of China under Grant No. LY15F020028. The author would like to thanks Dr. Junxia Li and Prof. Jian Yang for providing the source code of DLRMR [7].

References

1. Bai, C., Chen, J.N., Huang, L., Kplama, K., Chen, S.: Saliency based multi-feature modeling for semantic image retrieval. J. Vis. Commun. Image Representation (Accepted)
2. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. **3**(1), 1–122 (2011)
3. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE Tran. Pattern Anal. Mach. Intel. **37**(3), 569–582 (2015)
4. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intel. **24**(5), 603–619 (2002)
5. Everingham, M., Zisserman, A., Williams, C.K.I., Van Gool, L.: The PASCAL Visual Object Classes Challenge 2006 (VOC 2006). http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf
6. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.H.: Saliency detection via absorbing markov chain. In: ICCV, pp. 1665–1672 (2013)
7. Li, J., Lei, L., Zhang, F., Jian, Y., Rajan, D.: Double low rank matrix recovery for saliency fusion. IEEE Trans. Image Process. **25**(9), 4421–4432 (2016)
8. Li, Z., Tang, J.: Weakly supervised deep matrix factorization for social image understanding. IEEE Trans. Image Process. **26**(1), 276–288 (2017)
9. Li, Z., Tang, J., He, X.: Robust structured nonnegative matrix factorization for image representation. IEEE Trans. Neural Networks Learn. Syst. **PP**(99), 1–14 (2017)
10. Liu, G.H., Li, Z.Y., Zhang, L., Xu, Y.: Image retrieval based on micro-structure descriptor. Pattern Recogn. **44**(9), 2123–2133 (2011)
11. Liu, G.H., Yang, J.Y., Li, Z.Y.: Content-based image retrieval using computational visual attention model. Pattern Recogn. **48**(8), 2554–2566 (2015)
12. Liu, Y., Zhang, D., Lu, G., Ma, W.: A survey of content-based image retrieval with high-level semantics. Pattern Recogn. **40**(1), 262–282 (2007)
13. Liu, Z., Zou, W., Le, M.O.: Saliency tree: a novel saliency detection framework. IEEE Trans. Image Process. **23**(5), 1937–1952 (2014)
14. Lowe, D.G., Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
15. Mei, T., Rui, Y., Li, S., Tian, Q.: Multimedia search reranking. ACM Comput. Surv. **46**(3), 1–38 (2014)
16. Qin, Y., Lu, H., Xu, Y., Wang, H.: Saliency detection via cellular automata. In: Computer Vision and Pattern Recognition, pp. 110–119 (2015)
17. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, p. 1470 (2003)
18. Tong, N., Lu, H., Xiang, R., Yang, M.H.: Salient object detection via bootstrap learning. In: Computer Vision and Pattern Recognition, pp. 1884–1892 (2015)
19. Tong, N., Lu, H., Zhang, L., Xiang, R.: Saliency detection with multi-scale superpixels. IEEE Sig. Process. Lett. **21**(9), 1035–1039 (2014)
20. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)
21. Zhang, J., Li, D., Zhao, Y., Chen, Z., Yuan, Y.: Representation of image content based on roi-bow. J. Vis. Commun. Image Representation **26**, 37–49 (2015)