Saliency-Based Image Object Indexing and Retrieval

Yat Hong Jacky Lam(✉) and Sule Yildirim Yayilgan

Norwegian University of Science and Technology, 2815 Gjøvik, Norway
yathl@ntnu.no

Abstract. We suggest a novel approach to combine visual saliency model and object recognition to provide a more semantic description of an image based on human attention priority. The idea is to index and retrieve semantically more relevant images utilizing human saliency. Based on that, we developed a content-based image indexing and retrieval system. The resultant indexing and retrieval system works, though there is room for improvement in performance. We suggest the reasons and the possibilities for further improvements to develop a practical CBIR system.

1 Introduction

Content-based image retrieval (CBIR) is a long studied topic. While there is lot of existing research on low-level features, researchers are still struggling to get a better understanding on how to represent and obtain mid-level features and high-level features. The semantic difference between low and high level features representation is commonly called as the 'semantic gap' and it is a challenge to fill this gap.

To achieve a semantically meaningful description of an image's content, one of the most important steps to do is to determine the region of interest (ROI). In this research, *we study the feasibility to combine saliency prediction model with object recognition to identify key objects in an image.* The computational image saliency model is used to predict the focus points in the image. By using the object recognition algorithm, the content of an image is generated by combining the object and the saliency information with the low-level description of the image, we obtain a high level description of it. Later, the combined description is used to index the image.

This paper includes a literature review in Sect. 2, our framework in semantic CBIR is presented in Sect. 3. Evaluation is presented in Sect. 4. Future work and conclusions are suggested in Sect. 5.

2 Literature Review

2.1 Object Recognition and Deep Learning in CBIR

Deep learning refers to a collection of machine learning techniques where information is processed in multiple layers in hierarchical architectures. Since the

A. Campilho et al. (Eds.): ICIAR 2018, LNCS 10882, pp. 269–277, 2018.
https://doi.org/10.1007/978-3-319-93000-8_31

successful use of the Deep Convolution Neural Networks in the image classification task in ILSVRC1-2012 [1], deep learning became state-of-the-art in computer vision, including tasks such as image classification and object recognition. There is various research based on using it, including MobileNet-SSD [2], Faster-RCNN [3] and R-FCN [4] etc. Deep learning requires a large database for training. With the development of the vast amount of multimedia source on the Internet, large amount of images annotated by users are available for the training purpose. There are datasets for competitions in object recognition such as MS-COCO [5] and Kitti [6]. Those databases also cover a wide range of themes that allow to train and test the CBIR system on various images.

Object recognition provides an effective way towards higher level description of image. Image captioning is a popular application of the technique. Combining object recognition and natural language processing, it can provide a semantic description of image, including the class and characteristics of objects in the image, and also the action of animals and human beings in the image. More attention was put on instance-level image retrieval [7]. Many of them use deep learning and user-generated data on the internet.

2.2 Image Saliency in CBIR

Image saliency is the visual attention that a human observer puts on a certain position in an image. It can be used as a measure of relative importance or meaningfulness of that position in the image such that object containing that region should be given higher weighting in the indexing process. Image saliency map is introduced as a metric for CBIR by many researchers [8]. The saliency prediction can be done using a bottom-up approach using low-level features or top-down approach by including external information such as heuristics.

3 New Model to Combine Saliency Prediction and Object Recognition

3.1 Framework

In this paper, we propose a framework (Fig. 1) for combining saliency prediction and object recognition in addition to using low level features. After preprocessing, a vector presentation of the image is generated and stored in the database. For an image query, the procedure is similar: after generation of the feature vector, the query feature vector is compared with the feature vectors stored in the database using a similarity measure. The results are sorted by similarity and top results are retrieved.

While it is similar to most of the other CBIR algorithms, the main contribution of this study is the introduction of a new model for representing feature vectors and defining the similarity measure.

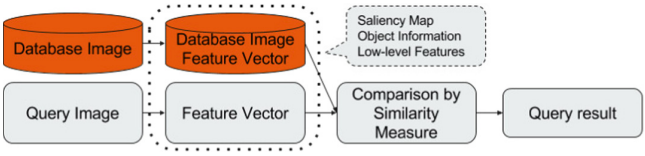


Fig. 1. The basic framework of the proposed content based image retrieval model

3.2 Object Recognition and Bounding Box

Most of the current object recognition algorithms provide bounding boxes as output. Detection is represented by a rectangular area enclosing the object. An example is shown in Fig. 2(a). Each bounding box contains three types of information:

1. Classification of object: object is represented by an integer index such as 'person' and 'toothbrush' as shown in the figure.
2. Position of the bounding box: the position is indicated by four values and also reflects the size of the object.
3. Score of detection: it is a floating point value indicating the confidence of the object recognition algorithm on the classification of the object. The classification with a low score should be rejected.

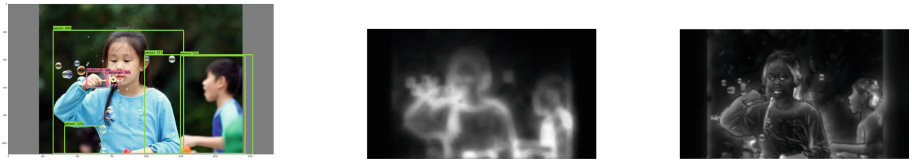
Notice that a pixel can belong to multiple bounding boxes or does not belong to any of them. Besides, due to the nature of the rectangular bounding box, some irrelevant parts of the image or the background is also included in the bounding box. In this research, object recognition is done through using a deep learning network model "Single Shot Multibox Detector (SSD) with MobileNet" [9] which is pretrained with MS-COCO data [5]. It is used due to its lightweight and speed, which is crucial for the speed performance, but the accuracy is sacrificed. The current algorithm can classify 90 different categories of objects. The number of class categories depends on the data and annotation used in the training process.

3.3 Saliency Prediction

Saliency prediction model aims to predict relative intensity of human visual attention and output as a heatmap (Fig. 2). The bright region indicates the highly salient region of the image. The Itti-Koch approach is used in our research. It is a classical model based on low level features [10]. Lower level features like colour, intensity and orientations are used to derive the saliency map. When there are alternative saliency models, especially those deep learning based approach, then they are more time consuming and hence not implemented in this study.

3.4 Feature Vector

In Fig. 3, the overall architecture for forming the feature vector is summarized:



(a) (b) (c)

Fig. 2. (a) object recognition with bounding boxes, (b) visual attention in experiment and (c) visual attention prediction by the Koch-Itti model. Image from CAT2000 database. [11]

1. The target image is processed through the object recognition algorithm. Object bounding box (OBB) and classification of objects with high confidence ($\geq t$) are obtained.
2. The saliency map of the image is calculated based on the Itti-Koch algorithm.
3. The average saliency intensity within each bounding box is calculated.
4. The list of detected objects is sorted by the average saliency.
5. The top five objects in saliency intensity is retrieved.

To determine the value of t, there has to be a balance between the number of candidate objects and also the confidence level of the object detection result. Moreover, t value varies due to the content of the image. An iterative approach is used to determine t. By setting the initial value $t = 0.1$, the list of candidate objects is generated. If the number of OBBs in the image is less than the desired value ($N = 5$ in our model), the threshold value is reduced by half until enough number of objects are detected using the model.

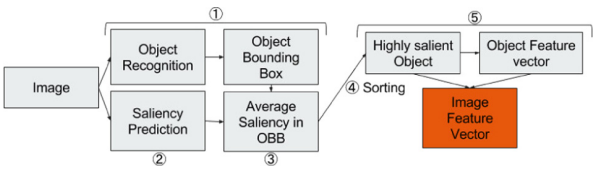


Fig. 3. Calculation of image feature vector

In our model, we have decided to choose the five most salient objects to form the primary feature vector. After the retrieval of the five highest salient objects, their category indices are used to form the primary feature vector, which provides the semantic representation of the image. The order of object in the primary feature vector is important as it shows the relative importance of the objects. However, ordering requirement can be too restrictive and it will be discussed in Sect. 5.

The object recognition algorithm provides semantic interpretation of the object. However, it misses out some important low level description of the object,

including colour and size. Low level description is extracted within each bounding box to describe low level features of objects.

1. Colour: the RGB value of the pixel in the image is converted into CIELAB colour space. The average value of all pixels in the three channels gives the average L, A and B value of the object.
2. Size: the height and the width of the bounding box determine the relative size of the object in the image.

After all, there are five feature elements for each object (L,A,B, height and width). Combining with the primary feature vector (5 elements) and 5 object feature vectors, there is a 30-element image feature vector. The graphical representation is shown in Fig. 4. These feature vectors are generated both for the query and the database images.

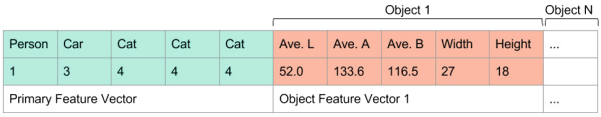


Fig. 4. Graphical representation of the image feature vector

3.5 Similarity Measure

To retrieve similar images, we have to compare the candidate image feature vector v_c and query feature vector v_q. A distance metric $D(v_c, v_q)$ is defined and the comparison is done in our model in three steps:

1. Comparison of primary vector: check if the object category agrees between the candidate and the query. Notice the order of object matter in this case. For example, the comparison between the primary vectors $[1, 2, 3, 4, 5]$ and $[2, 3, 4, 5, 1]$ results in element-wise disagreement for all elements. For each element-wise disagreement between the primary feature vectors, a penalty P is added to D. The sum of the penalty is denoted as $P(v_c, v_q)$.
2. Comparison of object feature vector: For each element-wise matching between the primary feature vectors, the distance metric of that matching is calculated with:
$$D\left(o_c^i, o_q^i\right) = \Delta L^2 + \Delta a^2 + \Delta b^2 + \alpha\left(\Delta W^2 + \Delta H^2\right)$$
where o_c^i and o_q^i are the i^{th} object feature vectors of the candidate and the query images respectively
ΔL is the difference in L channel
Δa is the difference in A channel
Δb is the difference in B channel
ΔW and ΔH is the difference of width and height respectively.

3. Adding step 2 and 3 then we get the distance metric:

$$D\left(v_c, v_q\right) = \sum_i D\left(o_c^i, o_q^i\right) + P\left(v_i, v_2\right)$$

As we can see from the definition, the most similar image with the query image is the query image itself as $D\left(v_q, v_q\right) = 0$. The result is sorted by the distance metric values. The value of P and α are empirically set to 10000 and 0.001.

3.6 Implementation

The implementation of the model is done in Python 3.5 with Tensorflow. Here is an example retrieval result shown in Fig. 5.

About the indexing time, we can observe that the indexing time is linearly increasing with the number of images as shown in Fig. 6 except for one outlier. The existence of outliers may be due to the threshold adjustment process. Although the image comes from the same database, there are variation of computation complexity between different catalogues.

4 Evaluation

To study the performance of the system, a subset of COREL image database [12], with 10 image categories, each containing 100 images, was used to evaluate the

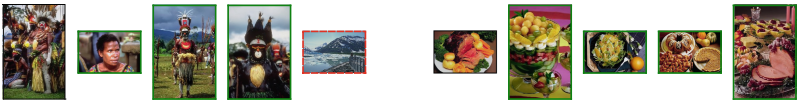


Fig. 5. Example of the retrieval result in COREL database. The first image (leftmost) is the query image, which is always the first retrieval result. The others are the 2nd–5th retrieval results. Green solid bordered image represents a successful retrieval when the wrong retrieval are bounded by red dashed border.

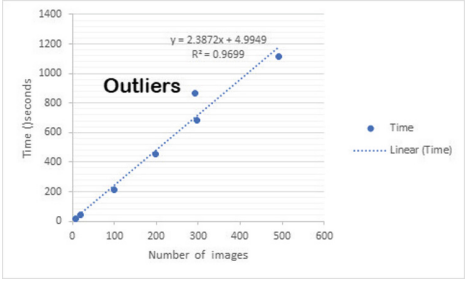


Fig. 6. Number of images vs Indexing time

system. The themes of the different categories were distinct to prevent ambiguity between categories. A retrieved image is considered as successful only if it belongs to the same category as the query. The mean average precision [13] is shown in Table 1. For comparison, we also include the average precision in 100 retrievals in our model and previous result with SIFT-LBP [14]. We also show two of the results using the ROC and precision-recall curve shown in Fig.7. Overall speaking, the results are worse than the previous result, especially the recall of the system is not very good. The precision is also quite low. This can be explained by the following reasons:

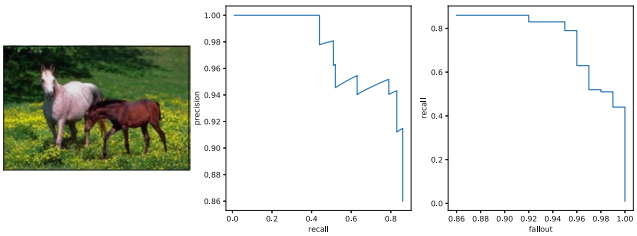
1. Limitation of object recognition algorithm: The object recognition algorithm plays an important role in the pipeline. If the class of object is not correctly determined, a heavy penalty is imposed, disregarding the similarity in low level features. For example, the category 'Horses' provides a much better result compared with other categories. In those categories, the main objects (the horses) are usually correctly determined. On the other hand, the category 'Mountains and glaciers' gives very poor result as most objects fail to be recognized. The quality of retrieval depends strongly on the object recognition algorithm, the training images and also the annotation used in the training.
2. Naive matching in primary feature vector: if an object exists in both query and candidate primary vector but in different ranking, they are still considered as mismatches. This causes high amount of mismatch in the comparison. In fact many of the images fail to match the query primary vector at all, resulting in maximum penalty (50000). These image results cannot be ranked so they are meaningless and rejected in the retrieval process. That is the reason that the maximum number of retrievals is very low in certain cases.
3. Parameters: There are three important parameters in the study: the number of bounding box, the penalty parameter P and α. These parameters are just empirically obtained and not optimized. Moreover, their optimal values are subject to various factors, including the size and context of the images.

5 Future Work and Conclusions

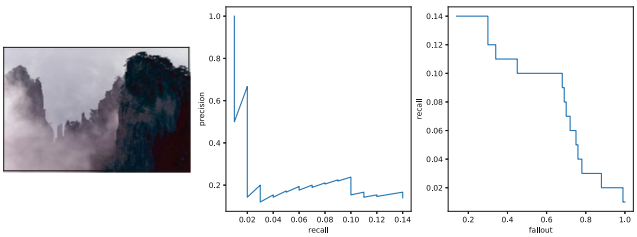
The first improvement is to allow for a matching of the object in different ordering. Objects in the same category in the candidate primary feature vector and the query primary feature vector should be matched by reordering. When there are multiple possible matches, the match should be chosen to minimize $D(v_c, v_q)$. Though this increases the complexity of the problem, it can improve the results. Besides, semantic segmentation [15] is a developing field in computer vision. It can be applied in our framework instead of the object recognition algorithm. The main advantage is to avoid the overlap area between bounding boxes of the objects, and it can avoid also multiple counting on the same object (for example, multiple bounding box on the same person). Thus the accuracy of the system can be improved.

Table 1. Mean average precision in COREL subset data and average precision comparison with SIFT-LBP approach

Categories	mAP (k=10)	mAP (k=100)	AP(k=100)	SIFT-LBP [14]
Africa people and villages	0.3486	0.1657	0.3540	0.57
Beach	0.3094	0.0849	0.1923	0.58
Buildings	0.2463	0.0482	0.1223	0.43
Buses	0.4970	0.2250	0.4029	0.93
Dinosaurs	0.5340	0.1885	0.2680	0.98
Elephants	0.6822	0.4926	0.5393	0.58
Flowers	0.6254	0.2782	0.3282	0.83
Horses	0.8163	0.5586	0.5979	0.68
Mountains and glaciers	0.2026	0.0391	0.1264	0.46
Food	0.6632	0.1840	0.2408	0.53
Overall	0.4925	0.2265	0.3172	0.637



(a)



(b)

Fig. 7. (a)Retrieval example with query image from dataset 'Horses'. (b)Retrieval example with query image from dataset 'Mountains and glaciers'.

Thus, in this research, we explored the possibility to combine object recognition algorithm and image saliency prediction model with low-level features in CBIR. Although there is a significant drawback in the current technology, this framework can shed light on the direction for developing CBIR algorithms that are more semantically meaningful.

References

1. Russakovsky, O., Deng, J., Hao, S., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
2. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
4. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp. 379–387 (2016)
5. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
6. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the kitti dataset. Int. J. Robot. Res. (IJRR) **32**(11), 1231–1237 (2013)
7. Andrej, K., Li, F.-F.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
8. Papushoy, A., Bors, A.G.: Visual attention for content based image retrieval. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 971–975, September 2015
9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
10. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
11. Borji, A., Itti, L.: Cat 2000: a large scale fixation dataset for boosting saliency research. In: CVPR 2015 Workshop on "Future of Datasets". arXiv preprint arXiv:1505.03581 (2015)
12. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001)
13. Zhou, W., Li, H., Tian, O.: Recent advance in content-based image retrieval: A literature survey. arXiv preprint arXiv:1706.06064 (2017)
14. Yuan, X., Yu, J., Qin, Z., Wan, T.: A sift-LBP image retrieval model based on bag of features. In: IEEE International Conference on Image Processing (2011)
15. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 (2015)