# Explainability for Content-Based Image Retrieval

Bo Dong
Kitware Inc.
dongshuhao12@gmail.com

Roddy Collins
Kitware Inc.
roddy.collins@kitware.com

Anthony Hoogs
Kitware Inc.
anthony.hoogs@kitware.com

## 1. Introduction

We discuss how the concept of "explainability" may be applied to Content-Based Image Retrieval (CBIR) systems. CBIR typically transforms an image into a feature representation for which a similarity distance metric may be computed; recent systems have improved performance by using features from deep learning networks [11, 6, 3]. However, as these representations have no direct semantic interpretability, the behavior of the system can be difficult for the user to understand in terms of semantically significant objects in the scene which may have no significant presence in the feature representation. Conversely, the similarity metric for two images may be dominated by pixel content which is not the semantic focus of the images, such as the background. We propose *Similarity Based Saliency Maps* (SBSM) to illustrate which areas in an image the CBIR system uses when retrieving and ranking results; the SBSM thus serves to "explain" the CBIR's decisions to the user. We have implemented SBSMs in our open-source Social Media Query Toolkit (SMQTK) [4], and have conducted preliminary user studies to demonstrate that SBSMs allow the user to more efficiently retrieve images.

(a) Query image (b) Retrieved images (c) SBSM results

Figure 1: (a). A query image containing a dog and a car mirror. (b) Two retrieved images; both containing dogs and car mirrors. (c) SBSM maps indicating relevance of image regions to the match.

Figure 3 illustrates examples of SBSMs computed on the query and results from Figure 1. The query and results all contain both a dog and a car mirror; however, the system has no mechanism to communicate if the results were chosen due to the presence of the dog, or the mirror, or something else entirely. Informally, the SBSM is a heatmap; "hotter" regions contribute more to the match score with the query, while "cooler" areas have less impact.

In this paper, we describe SBSMs and their integration into SMQTK, and describe some preliminary user studies conducted using Amazon Mechanical Turk (AMT). Our findings suggest that SBSMs can help a CBIR user increase the precision of their search.

## 2. Approach

Our SBSM is a variant image region pertubation saliency maps, used to indicate importance of regions against some criteria [2, 9, 12, 8]. In the context of CBIR, a saliency map should indicate how a particular region on the retrieved image impacts the similarity. However, a classification-based saliency map indicates how image regions impact the classification probability, which is irrelevant to the image similarity. Our SBSM instead measures how result regions contribute to the distance metric used by the CBIR when computing similarity.

We perturb a retrieved image by applying a binary mask to block out the region of interest. Inside the binary mask, the region of interest has value 0; all other pixels have value 1. In general, the region of interest can be of any shape. In our setup, we simply use a $b$ x $b$ square block. By sliding the square block over the retrieval image by a stride step $s$, we are able to show the importance of the blocked areas on impacting the similarity. In order to leverage parallel computing resources (*e.g.* GPUs), we generate a set of binary masks $M$, each of which represents a state of sliding the square block.

Mathematically, given a query image $Q$, a retrieval image $A$ and a binary mask $m_i \in M$, the importance of the region blocked out by $m_i$ is estimated as follows:

$$K(Q, A, m_i) = \max(D' - D, 0)(\mathbb{1} - m_i), \quad (1)$$

$$D' = \|f(Q), f(A \odot m_i)\|, \quad (2)$$

$$D = \|(f(Q), f(A)\|, \quad (3)$$

where, $f : I \rightarrow \mathbb{R}^n$ is a black-box model, which maps a input image $I$ to a $n$ dimensional vector; $\odot$ denotes element-wise multiplication; $\|\vec{v}_1, \vec{v}_2\|$ is the similarity between the two vectors $\vec{v}_1$ and $\vec{v}_2$ based on a user-defined

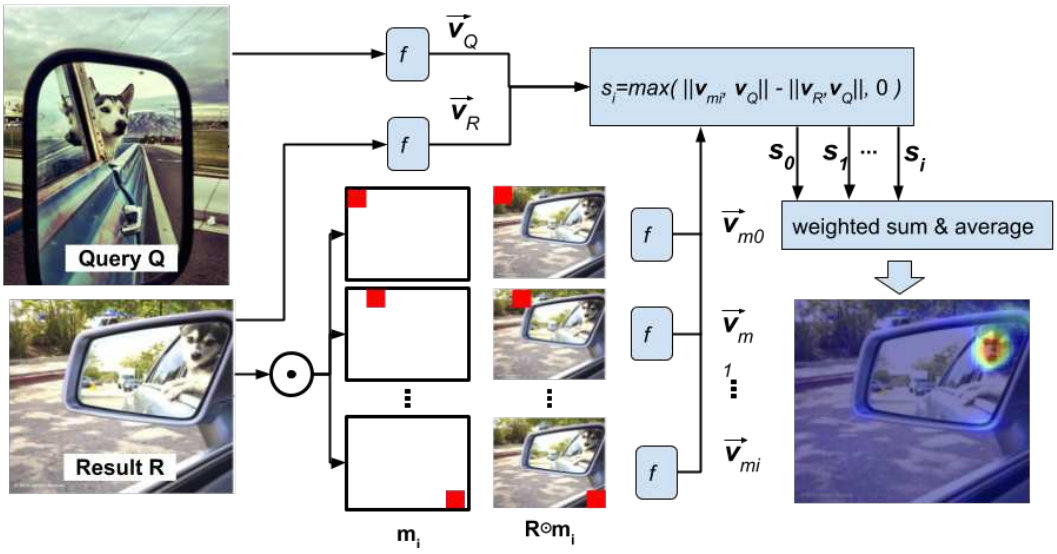



Figure 2: SBSM generation. Result image $R$ is masked by $m_i$ and run through feature extractor $f$; the distance between this new feature vector and the original query/result distance is how relevant masked region $m_i$ is in the final SBSM.

distance metric (*e.g.* $L_2$ distance); $\mathbb{I}$ is a matrix with all entries are 1 and the same shape as $m_i$. Given a binary mask set $M$ with $N$ binary masks, the SBSM is calculated as:

$$\text{SBSM}(Q, A, M) = \sum_i^N K(Q, A, m_i) \odot \frac{1}{\sum_i^N (\mathbb{I} - m_i)}. \tag{4}$$

Intuitively, Eq. 4 says that when a region is overlapped by multiple masks, the mean value is used to express the importance of the region. The overview of the proposed SBSM approach is illustrated in Figure 2.

We have experimented with various popular CNNs and distance metrics; Figures 3a-3d illustrate SBSMs generated using the features from $avgPool$ layer of the the ImageNet pretrained ResNet50 [7] and VGG19 [7] networks, and both the histogram intersection distance [10] and $L_2$ norm to measure distance between feature vectors.

## 3. User study

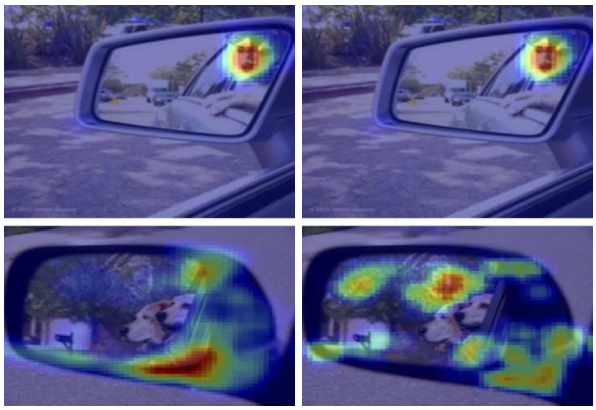
Our user study aims to show that SBSMs effectively and intuitively convey how the similarity metric affects the retrieved images, and that the user can leverage this to increase their search efficiency. We used SMQTK [4] as our base CBIR system; one of its main features is *Interactive Query Refinement*, or IQR, which allows a user to provide relevance feedback by interacting directly with the GUI to mark particular images "relevant" or "not relevant", as seen in Figure 5. This feedback trains an ad-hoc support vector machine (SVM) classifier which is used to re-rank the result set so that higher-ranked results are more likely to be relevant.

The archive image corpus was chosen from the training split of the COCO 2017 dataset [5]. For the query image, we randomly choose 12 images from the same dataset. Each query image contains only two different classes; one class is selected as the query target. A query task is then defined as retrieving images which contain the query target, using only the query image. The retrieval accuracy is defined as:

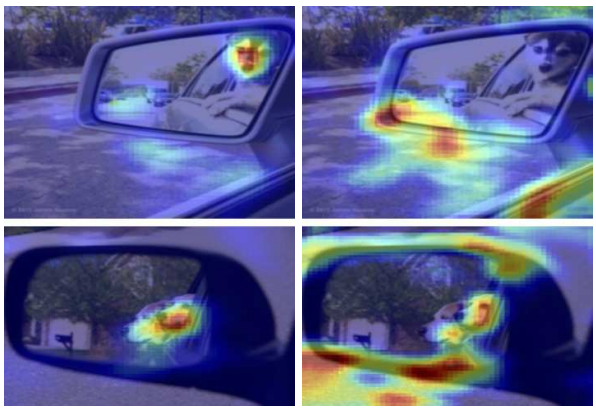
$$\frac{\sum_i^N \mathbb{1}_{A_i}(l)}{N}, \text{ and } \mathbb{1}_{A_i}(l) := \begin{cases} 1 & \text{if } l \in A_i, \\ 0 & \text{if } l \notin A_i, \end{cases} \tag{5}$$

where, $l$ is the query target label, $A_i$ is the set of annotations of the retrieved image $i$, and $N$ is the top $N$ retrieval images returned by the CBIR. Our user study used $N = 50$, the $avgPool$ layer of ImageNet pretrained ResNet50 [7] CNN, and the histogram intersection distance [10] to measure the similarity distance. The binary mask $m_i \in M$ block size is set to 20x20 and the stride step is set to 4. In the SVM, Histogram Intersection Kernel (HIK) [1] is used to define the hyperplane between positive and negative feedback.

We carried out our user study on Amazon Mechanical Turk (AMT). Each query task was tested with and without SBSM; without SBSM is regarded as a baseline. Each MTurk worker was given four unique Human Intelligence Task (HITs) (two with-SBSM, two without) to reduce query task bias and to expose the worker to both setups. For a fair comparison, each worker must give feedback to all the retrieved top-20 images in each IQR round, and to conduct two rounds of IQR. Here, three different MTurks are assigned for each four unique HITs.



(a) ResNet50,L2 (b) ResNet50,HID



(c) VGG19,L2 (d) VGG,HID

Figure 3: Similarity Based Saliency Maps (SBSMs) computed between the result and the query with various feature represenations and metrics.

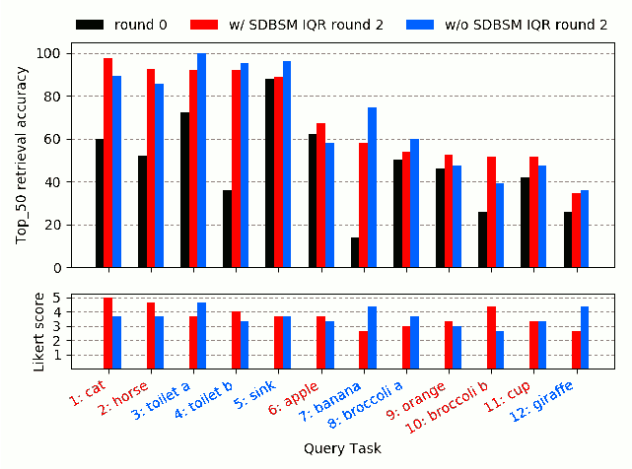
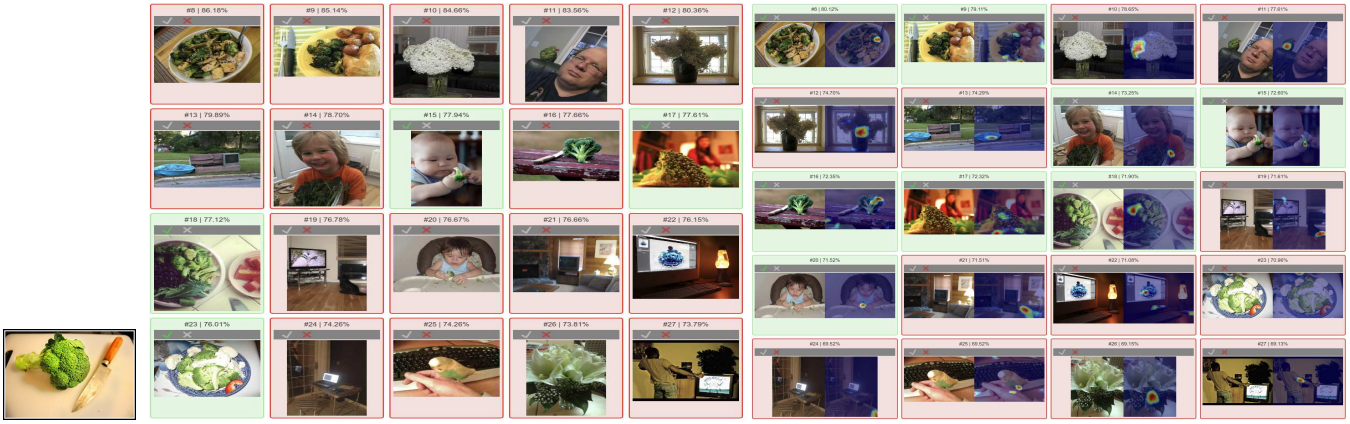



Figure 4: Top graph: retrieval accuracy of the target label within the top-50 returned results after the initial query (black bar), and after two rounds of IQR with (red) and without (blue) SBSM. Bottom graph: average Likert score; each label was processed by three unique workers with SBSM and three unique workers without. The x-axis labels are the query target; red indicates SBSM outperformed baseline for retrieval accuracty; blue indicates the opposite.

To understand the user's perspective on effectiveness, we asked the question "*I have high confidence the correct object in the retrieved image is matched to the target object in the query image*", at the end of each IQR round; the answer was a 5-point Likert scale (i.e., Strongly agree = 5, Strongly disagree=1). The hypothesis is if the SBSM perfectly overlays the query target in most retrieved images, we expect the worker would vote strongly agree (*i.e.* score 5). For the baseline, without the assistance of the overlay illustrating how the CBIR operated, we would expect a lower score.

Figure 4 illustrates two patterns. Firstly, IQR always improves retrieval accuracy for both with- and without-SBSM. The results are well aligned with other relevance feedback approaches. Secondly, the relative ordering of the Likert scores for (with, without) SBSM are strongly correlated to the relative ordering of the (with, without) retrieval scores across the labels. For query tasks with red color labels in Figure 4 (*i.e.* 1, 2, 6, 9, 10, 11), with-SBSM gives higher retrieval accuracy. Correspondingly, the Likert scores with-SBSM are also higher than the one without, which means the SBSMs overlay the query object in more retrieved images. As shown in Figure 5c, only one broccoli is not overlaid by the corresponding SBSM, and the Likert score is 4. In this case, a user is more likely to mark the corresponding images as positive feedback. However, without SBSM, the semantic gap leaves an AMT worker uncertain how to give feedback (as shown in Figure 5b). Hence, compared to the ones with SBSM, the Likert score is lower, and the retrieval accuracy is also lower.

The query tasks with blue color labels in Figure 4 (*i.e.* 3, 4, 5, 7, 8, 12) show the same correlation but in an opposite direction; with SBSM, they have lower Likert scores and lower retrieval accuracy except for the fourth query. The lower Likert score indicates the SBSM does not overlay the query object on the majority of retrieved images; the non-overlaid area (*e.g.* other objects, background) has more impact on the similarity distance, making the image less critical for query target training. During the user study, an AMT worker, most likely, votes these images as negative feedback since a worker must give feedback for every retrieved image. However, less critical does not mean it should be a negative feedback. In the case without SBSM, a worker's feedback is mainly based on his/her semantic sense of the relationship between the query and retrieved image. Therefore, as long as the query target is present in an image, they are likely to give positive feedback to the image. We hypothesize this is why retrieval accuracy with SBSM is lower



(a) Broccoli query

(b) Feedback w/o SBSM; LS=3; (36%→44%)

(c) Feedback w/ SBSM; LS=4; (36%→52%)

Figure 5: SMQTK IQR GUI and user's relevance feedback. (a): query image. In (b), (c), red background means the user marked the image "not relevant"; green indicates "relevant". LS is Likert Score; (x%→y%) means that after applying relevance feedback, the top-50 retrieval accuracy changes from x% to y%.

than without.

We observe a strong correlation between the Likert score and the retrieval accuracy, indicating the SBSM connects the user's perception of the image and the underlying feature vectors driving the CBIR. Note that the main purpose of the proposed SBSM is not to improve the CBIR performance, but to give the user insight into **why** a certain image is retrieved. However, bridging this semantic gap does not mean the base model and a user agree on semantic correctness; one person understanding another's semantic concept does not mean the person must think the semantic concept is correct. Therefore, in this work, the correlation between the Likert score and retrieval accuracy is more important.

## 4. Conclusion

We have identified two semantic barriers inhibiting efficient use of CBIR systems: first, that between image features used to represent images in the CBIR and the semantic concepts in the viewer's perception; secondly, the similarity metric used to retrieve images cannot be directly explained to the user. We propose the *Similarity Based Saliency Map* (SBSM), which visually explains the feature distance between query and result images, illustrating precisely which areas of the result most affect the distance. Our method is label-free and can be implemented using any feature source. We have implemented our method on top of an open-source CBIR system (SMQTK), and leveraged SMQTK's feedback-based interactive query refinement study to conduct a user study which highlights how SBSMs allow the user to more efficiently give feedback, leading to higher retrieval scores.

## References


[1] A. Barla, F. Odone, and A. Verri. Histogram intersection kernel for image classification. In *ICIP (3)*, 2003.

[2] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *CoRR*, abs/1704.03296, 2017.

[3] A. Gordo and D. Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *2017 IEEE CVPR 2017*

[4] Kitware. Social Media Query Toolkit (SMQTK). `https://github.com/Kitware/SMQTK`.

[5] T. Lin, M. Maire, et al. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[6] L. Mai, H. Jin, Z. L. Lin, C. Fang, J. Brandt, and F. Liu. Spatial-semantic image search by visual feature synthesis. In *2017 IEEE CVPR*

[7] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[8] V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. *arXiv:1806.07421*, June 2018.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[10] M. J. Swain and D. H. Ballard. Color indexing. *Int. J. Comput. Vision*, 7(1):11–32, Nov. 1991.

[11] F. Yang, J. Li, S. Wei, Q. Zheng, T. Liu, and Y. Zhao. Two-stream attentive cnns for image retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17

[12] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.