Image Retrieval Based on Query by Saliency Content

Adrian G. Bors and Alex Papushoy

Abstract This chapter outlines a content based image retrieval (CBIR) methodology that takes into account the saliency in images. Natural images are depictions of real-life objects and scenes, usually set in cluttered environments. The performance of image retrieval in these scenarios may suffer because there is no way of knowing which parts of the image are of interest to the user. The human visual system provides a clue to what would be of interest in the image, by involuntarily shifting the focus of attention to salient image areas. The application of computational models of selective visual attention to image understanding can produce better, unsupervised retrieval results by identifying perceptually important areas of the image that usually correspond to its semantic meaning, whilst discarding irrelevant information. This chapter explores the construction of a retrieval system incorporating a visual attention model and proposes a new method for selecting salient image regions, as well as embedding an improved representation for salient image edges for determining global image saliency.

1 Introduction

Visual information retrieval is one of basic pursuits required by people in the current technology driven society. Whether from mobile devices or whilst browsing the web, people search for information and a significant part of such information is visual. Many image retrieval approaches use collateral information, such as keywords which may be or not associated with the images. Content-based image retrieval (CBIR) considers a user-provided image as a query, whose visual information is processed and then used in a content-based search [11, 38]. CBIR is based on the notion that visual similarity implies semantic similarity, which is not always the case, but is in general a valid assumption. Due to the ambiguous nature of images, for a given query, a set of candidate retrieval images are sorted based on their relevance/similarity to the query.

A.G. Bors (✉) • A. Papushoy
Department of Computer Science, University of York, York YO10 5GH, UK
e-mail: adrian.bors@york.ac.uk

J. Benois-Pineau, P. Le Callet (eds.), *Visual Content Indexing and Retrieval with Psycho-Visual Models*, Multimedia Systems and Applications,
DOI 10.1007/978-3-319-57687-9_8

The main challenge in CBIR systems is the ambiguity in the high-level (semantic) concepts extracted from the low-level (pixels) features of the image [5, 43, 44]. The second obstacle is the sensory gap which can be interpreted as the incompleteness of the object information captured by an imaging device. The problem stems from the fact that the same object, photographed under different illumination conditions, different view angles, located at various depths or which may be occluded by other objects, appears differently due to changes in its acquisition context [38]. Whilst the semantic concept remains unchanged, the visual information results in a different interpretation that may negatively affect the performance of a CBIR system. Moreover, there is ambiguity within the user's intent itself. Generally, it is difficult for image retrieval systems to search for broad semantic concepts because it is hard to limit the feature space without broadening the semantic gap.

The majority of research studies during the early years of CBIR research have focused on the extraction and succinct representation of the visual information that facilitates effective retrieval. Narrow image domains usually contain domain-specific images such as medical scans or illustrations, where the set of semantic concepts is restricted and the variability of each concept is rather small. On the other hand, broad domains, such as natural images on the web, contain a large set of semantic concepts with significant variabilities within them. Producing a system that can cope well with a broad image domain is much more challenging than one for the narrow domain [38]. Images are ambiguous and the user of an image retrieval system is usually only interested in specific regions or objects of interest and not the background. Early works extracted a single signature based on the global features of the image, but the background concealed the true intent. In the later approaches, in order to capture the finer detail, the images were segmented into regions from which signatures were extracted.

There are four categories of CBIR methods, [11]: bottom-up, top-down, relevance feedback and those based on image classification. Those that rely purely on the information contained in the image are bottom-up approaches such as [33], while top-down approaches consider the prior knowledge. In image classification approaches, the system is presented with training data from which it learns a query [8]. Systems involving the user in the retrieval process via relevance feedback mechanisms are a mixture of bottom-up and top-down approaches [35].

Some of the earliest examples of CBIR systems is QBIC (Query by Image Content) [2, 13] developed at IBM, and Blobworld [5]. Images are represented as scenes or objects, and videos are converted into small clips from which motion features are extracted. These distinctions enable the system to be queried in several ways: the user may search for objects (such as round and red), scenes (by defining colour proportions), shots (defined by some type of motion), a combination of the above, and based on user-defined sketches and query images. In order to query by objects, the user must use a mask indicating the objects in the image. Image segmentation was used for the automatic extraction of the object boundaries in simpler images, but user tools are also provided for manual and semi-automatic extraction. The downside of such systems is the use of global colour representations (histograms), which preserve the colour distributions but have the tendency to hide information relating to smaller areas of interest that may carry a lot of importance. In addition, in order to take full advantage of the retrieval by object, the user is heavily involved in the database population. The later versions of QBIC included the automatic segmentation of the foreground and background in order to improve the retrieval.

CANDID (Comparison Algorithm for Navigating Digital Image Databases) [25] image retrieval represented the global colour distribution in the image as a probability density function (pdf) modelled as a Gaussian mixture model (GMM). The idea originated in text document retrieval systems, where the similarity measure was simply the dot product of two feature vectors [39]. For images, the local features such as colour, texture and shape were computed for every pixel and then clustered with the k-means algorithm which defined the GMM's components and parameters. The similarity measure was then based on the dot product, representing the cosine of the angle between the two vectors. The background was considered as another pdf which was subtracted from each signature during the similarity computation. This method was applied in narrow image domains such as for retrieving aerial data and medical greyscale images.

The Chabot [27] system combined the use of keywords and simple histograms for the retrieval task. The system was highly interactive and utilized a relational database that would eventually store around 500,000 images. For the retrieval performance, the RGB colour histograms were quantised to 20 colours, which was sufficient for qualitative colour definition during query with the keywords as the primary search method. The MIT Photobook [32] system took an entirely different approach to the retrieval of faces, shapes and textures. The system performed the Karhunen-Loeve transform (KLT) on the covariance matrix of image differences from the mean image of a given training set, while extracting the eigenvectors corresponding to the largest eigenvalues. These vectors would represent the prototypical appearance of the object category and images can be efficiently represented as a small set of coefficients. The similarity between objects is computed as an Euclidean distance in the eigenspaces of the image representations. The VisualSEEk [37] system combines image colour feature-based querying and spatial layouts. The spatial organisation of objects and their relationships in an image are important descriptive features that are ignored by simple colour histogram representation methods. VisualSEEk identifies areas in a candidate image, whose colour histogram is similar to that of the query.

Certain top-down, CBIR approaches employ machine learning techniques for the relevance feedback such as the support vector machine (SVM) [4] or multiple instance learning [33]. Image ranking for retrieval systems has been performed by using integrated region matching (IRM) [44] and the Earth Mover's Distance (EMD) [23, 34]. Deep learning, emerged lately as a successful machine learning approach to a variety of vision problems. This application of deep learning to CBIR was discussed in [42, 45].

The focus of this work is to analyze and evaluate the effectiveness of a bottom-up CBIR system that employs saliency in order to define the regions of interest in order to perform localised retrieval in the broad image domain. Visual saliency was considered for CBIR in [14, 30] as well. This chapter is organized as in the following. In Sect. 2 we introduce the modelling framework for visual attention while in Sect. 3 we present the initial processing stages for the Query by Saliency Content Retrieval (QSCR) methodology. The way how saliency is taken into account by QSCR is explained in Sect. 4. The ranking of images based on their content is outlined in Sect. 5. The experimental results are provided in Sect. 6 and the conclusions of this study in Sect. 7.

2 Modelling Visual Attention

The process of meaningful information processing from images by the human brain is very complex and it is not fully understood. Human reaction to the perceived information from images takes into account previous experiences and memories, as well as eye contact and fixation. The human visual system aims to focus on interesting regions in images, which coincide with the fixation points chosen by saccades, corresponding to random eye movements, at the pre-attentive stage for foveation, representing conscious acquisition of detail. These regions are characterized by local discontinuities, features and parts of images that attract the visual attention determining them to stand out from the rest. Such salient regions tend to correspond to important semantic concepts and are useful for image understanding while the rest of image content is ignored. In Fig. 1 we present some examples of visual saliency in images.

Salient regions can be defined in two ways. Bottom-up attention is instinctive and involuntary. It is entirely driven by the image, usually by specific features, such as colour, size, orientation, position, motion or their scene context. This approach is almost a reflex and corresponds to the instinctive type of attention to a salient region. Top-down attention, on the other hand, is driven by memory and prior experiences.

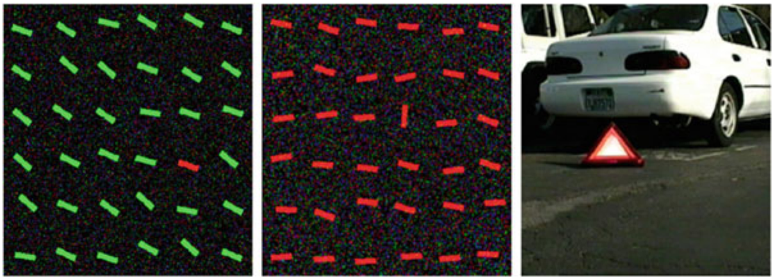


Fig. 1 Examples of visual saliency

Looking for a specific object of interest amidst many others, such as a book on a shelf or a key of a keyboard, may be defined by the previous knowledge of the title or authors of that book for example. Top-down attention driven by memories may even suppress bottom-up attention in order to reduce distraction by salient regions. Recently, memorisation studies have been undertaken in order to identify the reasoning behind the visual search [40, 41].

Visual attention is a diverse field of research and there are several models that have been proposed. Visual attention can be defined as either space-based or object-based. Spatial-based attention selects continuous spatial areas of interest, whereas object-based attention considers whole objects as driving the human attention. Object-based attention aims to address some of the disadvantages of spatial models such as their imprecision in selecting certain non-salient areas. Spatial models may select different parts of the same object as salient which means that the attention focus is shifted from one point to another in the image, whereas object-based attention considers a compact area of the image as the focus of attention. Applications of spatial-based attention to Content Based Image Retrieval tasks have been prevalent whilst those of object-based attention have not received a similar attention from the Image Retrieval community.

One of the main computer vision tasks consists of image understanding which leads to attempting to model or simulate the processing used by the human brain. Computational attention models aim to produce saliency maps that identify salient regions of the image. A saliency map relies on firstly finding the saliency value for each pixel. Salient region identification approaches fall into two main categories. The first category is based on purely computational principles such as the detection of interest points. These are detected using corner detectors and are robust under some image transformations, but are sensitive to image texture and thus would generalize poorly. Textured regions contain more corners but there are not necessarily more salient. Other computational approaches are based on image complexity assuming that homogeneous regions have lower complexity than regions of high variance. Some computational methods use either the coefficients of the wavelet transform or the entropy of local intensities [24]. Once again, such approaches assume that textured regions are more salient than others, which is not necessarily true. A spectral approach was used in [19], while [17] proposed a bottom-up model based on the maximisation of mutual information. A top-down classification method was proposed in [17] by employing the classification into either interesting or non-interesting areas.

The biologically influenced computational models of attention represent the second category of saliency models. This category further splits into two subcategories: biologically plausible and biologically inspired. Biologically plausible models are based on actual neurological processes occurring in the brain, whereas biologically inspired models do not necessarily conform to the neurological model. Generally, these models consist of three phases: the feature extraction, the computation of activation maps, and the normalization and recombination of the feature maps into a single saliency map [20].

The Itti-Koch saliency model [21, 22] is a well-known biologically plausible method modelling rapid changes of visual attention in scene which is based on neurological processes occurring in the brains of primates. The algorithm analyses colour, intensity and orientation information within nine different scale spaces by using dyadic Gaussian pyramids calculating center-surround differences using Difference of Gaussians (DoG) in order to detect local spatial discontinuities. From these, feature conspicuity maps (CM) are produced by recombining the multi-scale images and normalizing. A further linear combination produces the final saliency map. Among the salient regions, some are more salient than others. When the human brain is presented with the fixation points defining salient regions, the order in which it chooses the focus of attention (FOA) is determined by the saliency of a specific point. This principle is modelled by the algorithm by assigning each pixel in the saliency map to an input neuron in a feed-forward winner-take-all (WTA) neural network. In simulated time, the voltage in the network is increased until one of the input neurons fires, moving the FOA to the salient location represented by that neuron. After firing, the network is reset and the inhibition of return (IOR) is applied for a certain time in order to prevent the previous winner neuron from firing repeatedly. This mechanism produces a sequence of attended locations, where the order is driven by their saliency. The luminance image is produced from the average of the result for the red, green, blue image components. Orientation features are obtained from filtering the image with a bank of Gabor filters at different orientations. Image scales represent the image from original size down to 1/256th of the original image. During across-scale map combinations, low-resolution feature maps are upscaled and the final saliency map is downscaled to 1/256th of the original image. Given the amount of rescaling and Gaussian filtering occurring during the process, the saliency map produced by this model removes 99% of the high frequencies, [1]. This produces blurred edges of the salient regions after the map is upscaled to the original image size. The map only shows the peaks in saliency having high precision at low recall, which quickly drops off, [1]. Other criticism is directed at the lack of a clear optimisation objective of the system. The research study from [15] used different centre-surround difference calculations in order to optimise the Itti-Koch framework. The method proposed in [18] uses the biologically plausible model of Itti-Koch but applies a graph-based method for producing feature activation maps followed by normalisation.

Another biologically inspired hybrid method is the SUN model proposed in [47]. This model relies on a Bayesian framework based on the statistics of natural images collected off-line. The training set of natural images is decomposed though independent component analysis (ICA), yielding 326 filters, which are convolved with the image feature maps to produce the activations. This approach was shown to outperform the DoG when computing activation maps, albeit at the increase of the computational cost, as 326 filters are used in convolutions instead of 12.

3 Content Based Image Retrieval Framework

Content based image retrieval (CBIR) involves using an image as a model or query in order to search for similar images from a given pool of images. CBIR relies on the image content as a base of information for search, whilst defining image similarity remains a challenge in the context of human intent. In bottom-up computational analysis of images, the content is considered as being represented by statistics of image features. In this chapter we explain the Query by Saliency Content Retrieval (QSCR) method, which considers that the visual attention is a determinant factor which should be considered when initiating the image search. Firstly, we have a training stage in which characteristic image features are extracted from image regions corresponding to various categories of images from a training set. In the retrieval stage we rank the images, which are available from a database, according to a similarity measure. The scheme of the proposed QSCR system is provided in Fig. 2. The main parts of the QSCR system consists of image segmentation, feature extraction, saliency modelling and evaluating the distance in the feature space between a query image and a sample image from the given pool of images [29].

3.1 Image Segmentation

The mean shift segmentation algorithm is a well known clustering algorithm relying on kernel density estimation. This algorithm is a density mode-finding algorithm [9, 10] without the need to estimate explicitly the probability density. A typical kernel density estimator is given by

$$f(\mathbf{x}) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{1}$$

where n is the number of data points, h is the bandwidth parameter, d is the number of dimensions, $c_{k,d}$ is a normalizing constant and $K(\mathbf{x})$ is the kernel. The multivariate Gaussian function is considered as a kernel in this study. The mode of this density estimate is defined by $\nabla f(\mathbf{x}) = 0$. A density gradient estimator can be obtained by taking the gradient of the density estimator. In case of multivariate Gaussian estimator, it will be

$$\nabla f(\mathbf{x}) = \frac{2c_{k,d}}{nh^d} \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_i) K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h^{d+2}}\right) \tag{2}$$

and then derive a center updating vector called the mean shift vector:

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)} - \mathbf{x} \tag{3}$$

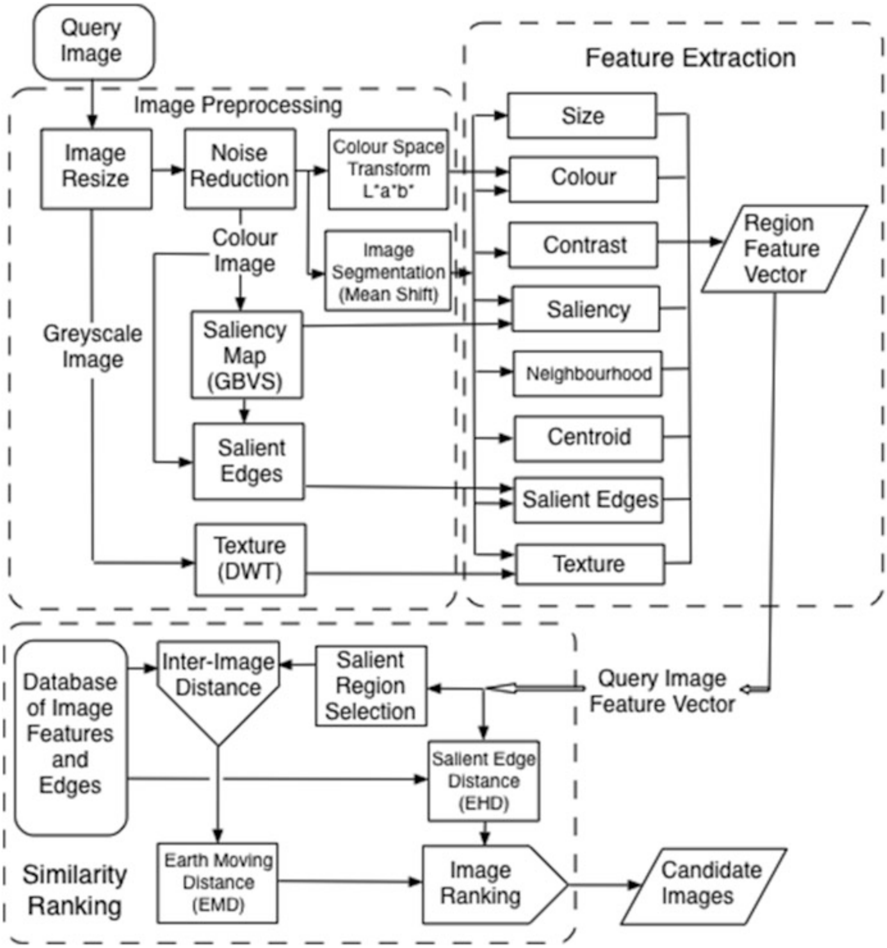


Fig. 2 The query by saliency content retrieval (QSCR) system using visual attention

Rearranging, yields the mean shift vector as:

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{h^2 c}{2} \frac{\nabla f(\mathbf{x})}{f(\mathbf{x})} \tag{4}$$

The mean shift algorithm stops when the mean shift becomes zero and consequently there is no change in the cluster center defining the mode. In the case when the algorithm starts with too many initial clusters, several of these would converge to the same mode and consequently all, but the ones corresponding to the real modes, can be removed.

In order to segment images, colour is transformed to the perceptually uniform CIELUV space. The two pixel coordinates defining their spatial location, and the colour values are combined into a single 5D input vector, which are then clustered using the mean-shift algorithm. Clusters, defining image regions correspond to the set of points that fall within the basin of attraction of a mode. The number of clusters, each characterizing a segmented region, is selected automatically by the algorithm based on the data depending only on the bandwidth parameter *h* [3].

3.2 Image Features Used for Retrieval

Each image is resized and then segmented into regions as described in the previous section. For each image region a characteristic feature vector is calculated, with entries representing statistics of colour, contrast, texture information, the region neighbourhood information and region's centroid.

Firstly, six entries characterizing the colour are represented by the median values as well as the standard deviations for the L*a*b* colour components calculated from the segmented regions. The median estimator is well known as a robust statistical estimator, whilst the variance represents the variation of that feature in the image region. The L*a*b* is well known as a colour space defining the human perception of colours. The Daubechies 4-tap filter (Db4) is used as a Discrete Wavelet Transform (DWT) [26] function for characterizing texture in images. 4 from Db4 indicates the number of coefficients used for describing the filter having two vanishing points. A larger numbers of coefficients would be useful when analysing signals with fractal properties which are also characterized by self-similarity. Db4 wavelets are chosen due to their good localisation properties, very good texture classification performance [6], high compactness, low complexity, and efficient separation between image regions of high and low frequency. Moreover Daubechie wavelet functions are able to capture smooth transitions and gradients much better than the original Haar wavelets, which are not continuous and are sensitive to noise. The lower level decompositions are up-scaled to the size of the image by using bicubic interpolation and then by averaging the pixel values across the three scales and for each direction. Three entries represent the texture energy measured as the average of the absolute values of the DWT coefficients of the region in the horizontal, vertical and oblique directions across the three image scales, [6].

The human visual system is more sensitive to contrast than to absolute brightness. Generally, the contrast is defined as the ratio between the difference in local brightness and the average brightness in a region. In order to increase its robustness, the contrast is computed as the ratio between the inter-quartile range and the median of the L* (luminance) component for each segmented region. The locations of the centers for each region are calculated as the averages of pixel locations from inside each compactly segmented region. These values are normalised by the image dimension in order to obtain values in the interval [0,1]. By giving more importance to the centroid locations, candidate images that best satisfy the spatial layout of the query image can be retrieved. However, placing too much importance on the location may actually decrease the precision and this is the reason why this feature is considered, together with the region neighbourhood, as a secondary feature, characterised by a lower weight in the calculation of the similarity measure. The region neighbourhood can provide very useful information about the image context. The neighbourhood consistency is represented by the differences between the L*, a* and b* values of the given region and those of its most important neighbouring regions located above, below, left and right, where the neighbouring significance is indicated by the size of the boundary between two regions, [33].

4 Defining the Saliency in Images

Based on the assumption that salient regions capture semantic concepts of an image, the goal of computing visual saliency is to detect such regions of interest so that they can be used as a search query. Saliency maps must concisely represent salient objects or regions of the image. In Fig. 3 we present an example of retrieving the Translucent Bowl (TB) image in SIVAL database without visual attention models compared to when visual attention models is used, assuming identical image features. As it can be observed, when using visual attention models, all first six retrieved images and the eight out of the total of nine correspond to the TB category, while when not using the visual attention models only the seventh image is from the correct category but none of the other eight images.

Saliency is used to identify which image regions attract the human visual attention and consequently should be considered in the image retrieval. Saliency is defined in two ways: at local and at the global image level, [29]. The former is defined by finding salient regions, while the latter is defined by the salient edges in the entire images. The regions which are salient would have higher weights

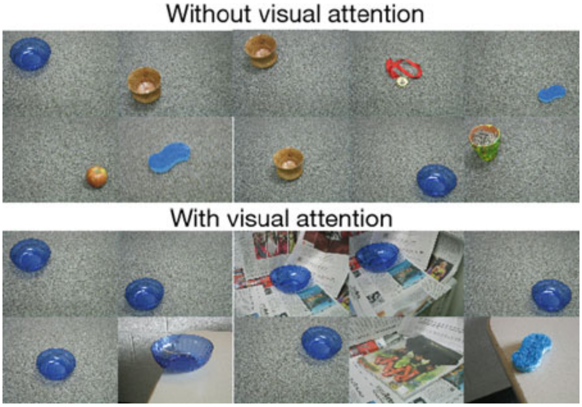


Fig. 3 Retrieving images with salient object when not using visual attention (*top image*) and when using the visual attention (*bottom images*) from SIVAL database. The query image is located at the *top left image* in each case

when considering their importance for retrieval while the salient edges are used as a constraint for evaluating the similarity of the query image to those from the given pool of images as shown in Fig. 2.

4.1 Salient Edges

In order to capture the global salient properties of a given image we consider the salient edges as in [14]. Firstly, the image is split into 16 × 16 pixels blocks, called sub-images. Salient edges are represented by means of the MPEG-7 Edge Histogram Descriptor (EHD) which is translation invariant. This represents the distribution along four main directions as well as the non-directional edges occurring in the image. Edges corresponding to each of these directions are firstly identified and then their density is evaluated for each sub-image region. The EHD histogram is represented by five values representing the mean of the bin counts for each edge type across the 16 sub-images. Each value represents the evaluation of the statistics for each of the edge orientations: vertical, horizontal, the two diagonal directions at 45 and 135 deg and the non-directional. The bin counts correspond to a specific directional edge energy and consequently the mean is an estimate that would capture it without any specific image location constraint.

4.2 Graph Based Visual Saliency

Known computational models of visual saliency are the Itti-Koch (IK) [22], Graph-Based Visual Saliency (GBVS) [18], which is the graph-based normalisation of the Itti-Koch model, the Saliency Using Natural statistics (SUN) [47], and the Frequency-Tuned Saliency (FTS) [1]. The first three methods produce low-resolution saliency blur maps that do not provide clear salient region boundaries. FTS, on the other hand, produces full resolution maps with clear boundaries, however, unlike the first three methods it only uses colour information, so it may fail to identify any salient regions when all objects in the image have the same colour.

The Graph-Based Visual Saliency (GBVS) method [18] was chosen due to its results for modelling saliency in images. The GBVS saliency extraction method is a computational approach to visual saliency based on the Itti-Koch model, but it takes a different approach to the creation of activation maps and their normalisation. Unlike the Itti-Koch model, which computes activation maps by center-surround differences of image features [22], GBVS applies a graph-based approach [18]. Generally, saliency maps are created in three steps: feature vectors are extracted for every pixel to create feature maps, then activation maps are computed, and finally the activation maps are normalized and combined. The image is converted into a representation suitable for the computation of the feature contrasts. Feature dyadic Gaussian pyramids are produced at three image scales of 2:1, 3:1, and 4:1. Gaussian pyramids are created for each channel of physiologically based DKL colour space [12], which has similar properties to L*a*b*. Orientation maps are then produced after applying Gabor filters at the orientations of $\{0, \pi/4, \pi/2, 3\pi/4\}$ degrees for every scale of each colour channel. The outputs of these Gabor filters represent the features which are then used as inputs in the GBVS algorithm.

In the first level of representation in GBVS, adjacency matrices are constructed by connecting each pixel of the map to all the other pixels, excluding itself, by using the following similarity function $\phi_1(\mathbf{M_x}, \mathbf{M_y})$ between feature vectors corresponding to the pixels located at $\mathbf{x}$ and $\mathbf{y}$:

$$\phi_1(\mathbf{M_x}, \mathbf{M_y}) = \left| \log \frac{\mathbf{M_x}}{\mathbf{M_y}} \right| \exp\left[-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right] \tag{5}$$

where $\sigma \in [0.1, 0, 2]D$, D representing the given map width. A Markov chain is defined over this adjacency matrix, where the weights of outbound edges are normalized to $[0, 1]$, by assuming that graph nodes are states, and edges are transition probabilities. By computing the equilibrium distribution yields an activation map, where large values are concentrated in areas of high activation and thus indicate the saliency in the image. The resulting activation map is smoothed and normalized. A new graph is constructed onto this activation map, with each node connected to all the others including itself and which has the edge weights given by:

$$\phi_2(\mathbf{M_x}, \mathbf{M_y}) = A(\mathbf{x}) \exp\left[-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right] \tag{6}$$

where $A(\mathbf{x})$ corresponds to the activation map value at location $\mathbf{x}$. The normalization of the activation maps leads to emphasizing the areas of true dissimilarity, while suppressing non-salient regions. The resulting. saliency map for the entire image is denoted as $S(\mathbf{x})$, for each location $\mathbf{x}$ and represents the sum of the normalized activation maps for each colour and each local orientation channel as provided by the Gabor filters.

In Fig. 4 we show a comparison of saliency maps produced by four saliency algorithms: Itti-Koch (IK) [22], Graph-Based Visual Saliency (GBVS) [18], Saliency using Natural Statistics (SUN) [47], and Frequency-Tuned Saliency (FTS) [1]. It can be seen that IK produces small highly focused peaks in saliency that tend to concentrate on small areas of the object. The peaks are also spread across the image spatially. This is because the Itti-Koch model was designed to detect areas to which the focus of attention would be diverted. Because the peaks do no capture the whole object, but rather small areas of it, it is insufficient for representing the semantic concept of the image and is not suitable for retrieval purposes.

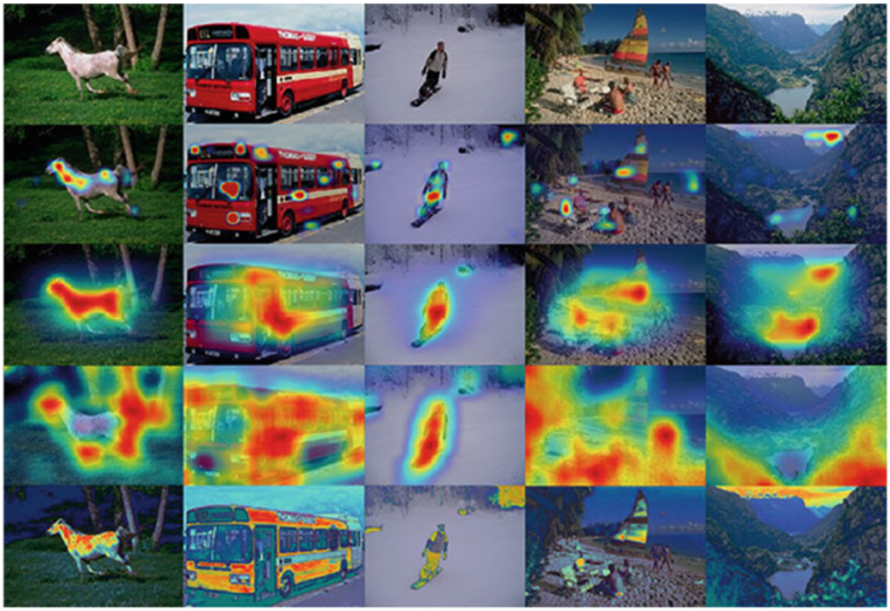


Fig. 4 Evaluation of saliency performance. Original images are in the *first row*, Itti-Koch saliency maps are in the *second row*, GBVS maps are in the *third row*, SUN maps are in the *fourth row*, and FTS maps are in the *fifth row*. Saliency maps are overlaid on the original images

The image selections produced by GBVS provides a good coverage of the salient object by correctly evaluating the saliency. It has a good balance between coverage and accuracy the results sit in between IK and SUN. Unlike the other three methods, GBVS provides a high-level understanding of the whole image and its environment, in the sense that it does not get distracted by the local details, which may result in false positives. It is able to achieve this because it models dissimilarity as a transition probability between nodes of a graph, which means that most of the time, saliency is provided by the nodes with the highest transition probability. It can be seen from the mountain landscape image in the last column that saliency is correctly indicated at the lake and sky, despite not having an evident object in that region of the image. In the second image where the red bus fills most of the image, GBVS recognises the area surrounding the door as most salient, compared to SUN algorithm, which considers the whole image as salient. It appears that the SUN algorithm only works well with the simplest of images such is the third image showing a snowboarder on snow. In the first image, given the image of the horse, which is only slightly more difficult, the SUN algorithm correctly identifies the head of horse, its legs, and tail. However, it also selects the trees, which are not that relevant for the retrieval of such images. The large amount of false positives, apparent bias, and lack of precision makes SUN an unsuitable choice for retrieval in the broad image domain, but perhaps it could prove itself useful in specialised applications. FTS algorithm, which represents a simple colour difference, only works well when there is a salient colour object in the image, and the image itself has little colour variation, such that the average colour value is close to that of the background. As it uses no other features than the colour, it lacks the robustness of other methods, but works extremely well when its conditions are met. As seen with the bus in the second image, its downside is that it does not cover salient objects when there is a lot of colour variation within, hence failing to capture the semantic concept. One of the problems with local contrast-based saliency algorithms is that they may misinterpret the negative space around true salient objects as the salient object.

GBVS is chosen for saliency computation in this study because of its robustness, accuracy, and coverage. One downside is that it does not produce full resolution saliency maps due to its computational complexity. During the up-scaling, blurred boundaries are produced, which means that saliency spills into adjacent regions and so marks them as salient, albeit to a smaller extent.

4.3 Salient Region Selection

In this study we consider that we segment the images using the mean-shift algorithm described in Sect. 3.1 and we aim to identify which of the segmented regions are salient. The purpose of the saliency maps is to select those regions that correspond to the salient areas of the image, which are to be given a higher importance in the querying procedure. For distinctive objects present in images, this represents selecting the object's regions, whereas for distinctive scenes it would come down to selecting the object and its neighbouring regions. Several approaches have been attempted to select an optimal threshold on the saliency energy of a region as the sum of the saliencies of all its component pixels. An optimal threshold would be the one that maximises the precision of retrieval rather than the one that accurately selects regions corresponding to salient objects. This is somewhat counter-intuitive, as one would think that specifying well-defined salient objects would improve the precision, but due to the semantic gap, this is actually not always the case. In the Blobworld image retrieval method [5], images are categorized as distinctive scenes or distinctive objects, or both. However, it was remarked that when considering CBIR in some image categories it would be useful to include additional contextual information and not just the salient object.

Firstly, we consider selecting regions that contain a certain percentage of salient pixels, where salient pixels are those defined by $S(\mathbf{x}) > \theta_p$. The average region saliency is calculated from the saliency of its component pixels as:

$$S(r_i) = \sum_{\mathbf{x} \in r_i} \frac{S(\mathbf{x})}{N_r} \tag{7}$$

where $\mathbf{x}$ is a pixel in region r_i, $i = 1, \ldots, R$, where R represents all segmented regions in the image, $S(\mathbf{x})$ is the value of the saliency for $\mathbf{x}$, and N_r is the number of pixels in the region r. In a different approach we can consider a saliency cut-off, given by $S(r) > \theta_R$, which was set at a value that would remove most of the regions defined by a small saliency. A third method of salient region selection is the one adopted in [14], where a threshold that maximizes the entropy between the two region partitions by the saliency threshold, was adopted. In [14] they set two cut-offs. The first cut-off was at 10% of the average region saliency value, calculated using the cumulative distribution function (CDF) of all region saliencies $S(r)$, as in Eq. (7) across the whole image database. This first cut-off was used to filter out large regions characterized by low saliency. The second cut-off was based on the total region saliency, representing the sum of all saliency values in the region, and was used to filter out very small regions characterised by very high saliency.

Another approach to select the salient regions consists in finding the average region saliency value corresponding to the minimum probability density in a non-monotonically decreasing pdf. This works well when there is a clear break between the saliency values in the pdf of salient regions and produces a good cut-off candidate. However, this method fails when the saliency pdf is monotonically-decreasing as the smallest saliency value is usually too high to select any regions. An adaptive method was attempted by using the density-based method for non-monotonically decreasing pdfs and the percentile-based cut-off otherwise. If the density-based method sets a cut-off that is too high, the retrieval performance is likely to decrease, so it is applied only if the first half of saliency values is non-monotonically decreasing. Another method that provided a suitable threshold was the one proposed in [28], which was shown to capture well the salient regions in several CBIR studies.

In the following experiments we consider that salient regions capture semantic concepts of an image. By computing the visual saliency we detect salient regions of interest in order to be used as a search query. In Fig. 5 we compare the saliency maps produced in 12 different images from diverse image categories of COREL 1000 database, by using different cut-off selection methods. Using Otsu's method to select the cut-off produces maps that discard the lower saliency values associated with the background preserving the medium to high saliency values. However, this method tends to produce large areas which results in too many background regions being included, which makes it less suitable when querying for distinctive objects. An example of this is in the second image from Fig. 5, representing a snow-covered landscape area. The cumulative distribution function of pixel saliency values corresponding to COREL 1000 image database is produced as shown in Fig. 6a. From this plot we can observe that almost 40% of the data have saliency values less than 0.1 and only 10% have a value above 0.62. The other half of the data (between 40th and 90th percentiles), has uniform probability as the gradient of the curve is approximately constant. Hence, we devise two salient region selection methods which set their cut-offs at the 60th and 80th percentiles of the image's saliency values, corresponding to approximately 0.3 and 0.5 cut-off

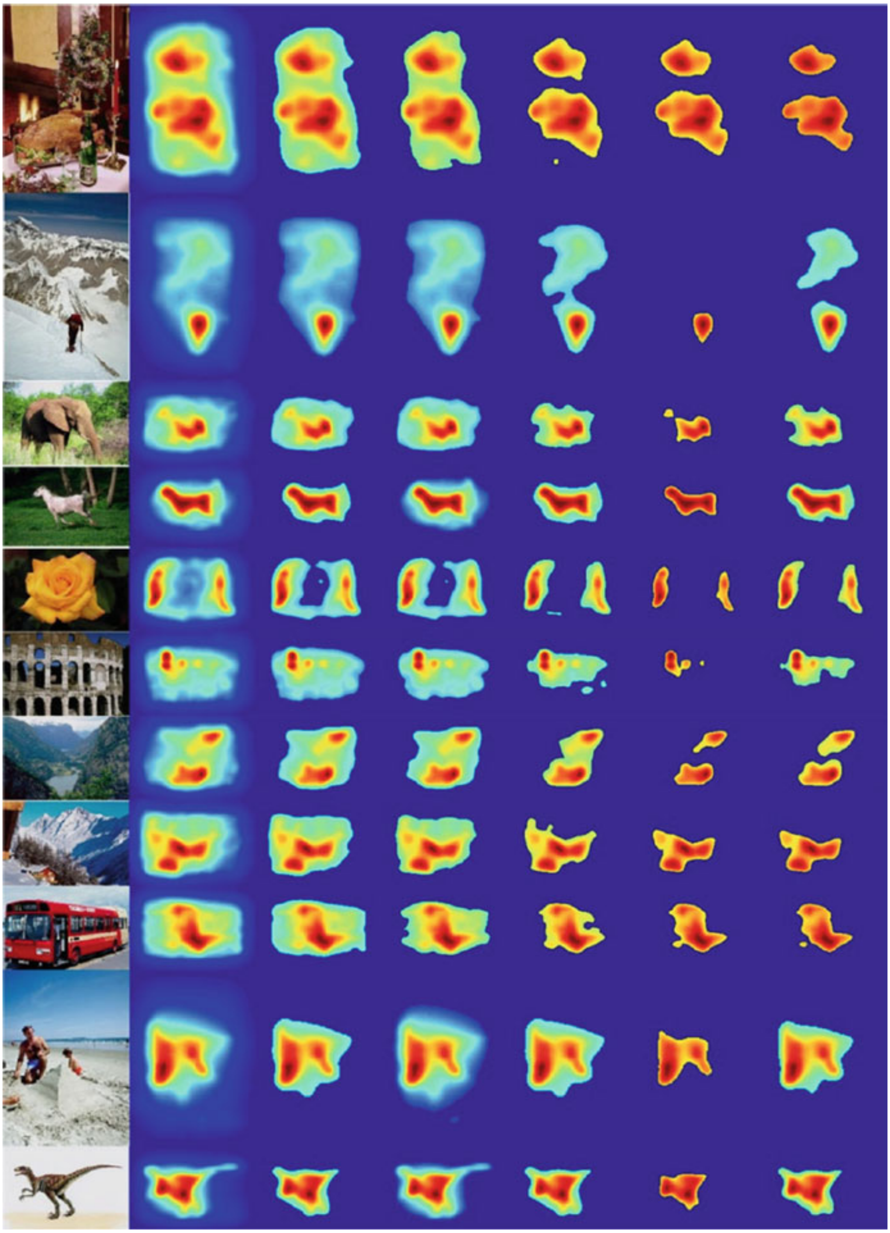


Fig. 5 Comparison of saliency cut-offs (*1*) Original image, (*2*) GBVS saliency map (SM), (*3*) Otsu's method, (*4*) Top 40%, (*5*) Top 20%, (*6*) Cut-off at 0.61, (*7*) Cut-off at twice the average saliency as in [1]

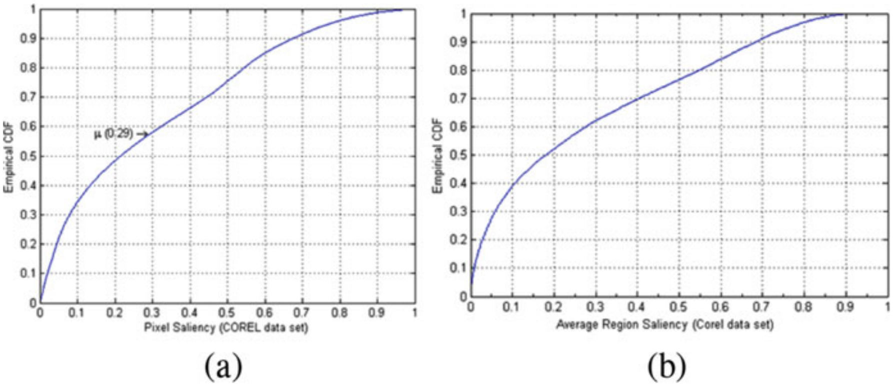


Fig. 6 Empirical cumulative distribution functions (CDF) calculated for all images from COREL 1000 database. (**a**) Pixel saliency. (**b**) Salient regions

points, respectively. The 60th percentile produces similar results to Otsu's method, but on some occasions includes too many background pixels as seen in the Beach and Dinosaur categories images, shown in the last two images from Fig. 5. The 80th percentile, representing the selection of the top 20% of data, shows a good balance between the two criteria where it selects a smaller subset of the pixels identified by Otsu's method as in the case of the Elephant and Bus categories, from the third and ninth images and at the same time captures a good amount of background pixels as in the Architecture and Beach category images from the sixth and tenth images. Obviously, it is impossible to guarantee that it will capture background information for distinctive scenes and just the object for distinctive object images, and vice versa, but at least this method is sufficiently flexible to do so. The next method is a simple fixed cut-off value set at the pixel value of 155, which corresponds to 60% precision and 40% recall. By looking at the CDF of salient pixels from COREL 1000 database, shown in Fig. 6a, this value corresponds to the 90th percentile of saliency values and so selects only the top 10% of the data. Only small portions of the image are selected and in many cases this fails to capture the background regions, resulting in lower performance. An example of this is seen in the image of the walker in the snow-covered landscape image, the Horse and the Architecture category images from second, fourth and sixth images from Fig. 5. In all these images, the most salient object has very little semantic information differentiating it from the others. For example, the walker is mostly black and very little useful information is actually contained within that region; the horse is mostly white and this is insufficient to close the semantic gap. Achieving a balance is difficult because a method that selects the regions of distinctive objects may fail when the image is both a distinctive object and a distinctive scene. An example of such a situation is the Horse category from the fourth image, where selecting the white horse by itself is too ambiguous as there are many similarly coloured regions, but by adding several background regions improves performance greatly. On the other hand, the performance would be reduced by including background regions when the image is in the category of distinctive objects. The last method, which was used in [1], sets the threshold at twice the average saliency for the image. This approximately corresponds to the top 15% of salient pixels from the empirical cumulative distribution for COREL 1000. This produces similar maps to the selection of regions with 20% salient pixels, except that it captures fewer surrounding pixels.

In the following we evaluate the segmented image region saliency by considering only those salient pixels which are among the top 20% salient pixels, which provides best results, according to the study from [29]. By considering a hard threshold for selecting salient pixels, the saliency for the rest of pixels is considered as zero for further processing. We then apply the region mask to the saliency map and consider the region saliency as given by the percentage of its salient pixels. Next, we use the saliency characteristic from all regions in the image database to construct an empirical CDF of salient regions, considering the mean-shift for the image segmentation, as explained in Sect. 3.1. The empirical CDF of the salient regions for the COREL 1000 database is shown in Fig. 6b. Now, we propose to select the most salient regions by setting the second threshold at the first point of inflexion in the CDF curve. This corresponds to the point where the gradient of the CDF curve begins to decrease. We observe that this roughly corresponds to the 35th percentile and thus our method considers the top 65% of most salient regions in the given database. We have observed that this saliency region selection threshold removes most of the regions with little saliency, while still considering some background regions containing the background information necessary for the retrieval of images of distinctive scenes. Such regions are suitable for describing the contextual semantic information.

The methods discussed above focus on selecting representative salient query regions. In the QSCR system we would segment the query image and would assume that all candidate images had been previously segmented as well. The saliency of each region in both the candidate images and the query one would then be evaluated. Once the salient query regions are determined, they could be matched with all the regions in the candidate images. Another approach could evaluate the saliency in both the query and the candidate images and the matching would be performed only with the salient regions from the candidate images. This constrains the search by reducing the number of relevant images because the query regions are searched by using only the information from the salient regions of the candidate images. Theoretically, this should improve both retrieval precision and computational speed, but in practice, the results will depend on the distinctiveness of the salient regions because the semantic gap would be stronger due to a lack of context.

5 Similarity Ranking

Given a query image, we rank all the available candidate images according to their similarity with the query image. The aim here is to combine the inter-region distance matrix with the salient edge information to rank the images by their similarity while taking into account their saliency as well. The processing stages of image segmentation, feature extraction and saliency evaluation, described in the previous section, and shown in the chart from Fig. 2, are applied initially to all images from a given database. Each region $I_j, j = 1, \ldots, N_I$, from every I image is characterized by a feature vector, and by its saliency, evaluated as described in the previous Section. Meanwhile, the energy of salient edges is evaluated for entire images. The same processing stages are applied on the query image Q, which is segmented into several regions Q_i, $i = 1, \ldots, M$ as described in Sect. 3.1. The similarity ranking becomes a many-to-many region matching problem which takes into account the saliency as well. Examples of many-to-many region matching algorithms are the Integrated Region Matching (IRM) which was used in [44] for the SIMPLIcity image retrieval algorithm and the Earth Mover's Distances (EMD), [34]. The EMD algorithm was chosen in this study due to its properties of optimising many-to-many matches, and this section of the algorithm is outlined in the lower part of the diagram from Fig. 2.

In the EMD algorithm, each image becomes a signature of feature vectors characterising each region. A saliency driven similarity measure is used between the query image Q and a given image I, represented as the weighted sum of the EMD matching cost function, considering the local saliency, and the global image saliency measure driven by the salient edges, [29]:

$$\mathscr{S}(Q, I) = W_{EMD} \, \frac{\text{EMD}(Q, I)}{\alpha_{EMD}} + W_{EHD} \, \frac{\sum\limits_{\theta} |\text{EHD}(\theta, Q) - \text{EHD}(\theta, I)|}{5 \, \alpha_E} \tag{8}$$

where $\text{EMD}(Q, I)$ is the EMD metric between images Q and I, $\text{EHD}(\theta, Q)$ represents the average salient edge energy, in five different directions of $\theta = \{0, \pi/2, \pi, 3\pi/4, \text{non-dir}\}$ for the image Q, derived as described in Sect. 4.1. The weights, found empirically, for the local region-to-region matching EMD component is $W_{EMD} = 0.7$, while for the global image component EHD, is $W_{EHD} = 0.3$. These choices indicate a higher weight for the localized saliency indicated by EMD than for the global image saliency, given by EHD, as observed in psychological studies of human visual attention. α_{EMD} and α_E represent the robust normalization factors which are set as the 95th percentile of the cumulative distribution function of the EMD and the EHD measures, respectively, calculated using a statistically significant image sample set. These robustness factors are used for normalizing the data and removing the outliers, by taking into account that the data distributions characterizing both EMD and EHD are log-normal.

EMD is an optimization algorithm which assumes a signature vector for each image, either the query image Q or any of the candidate images, I, from the database. The signature assigned to each image consists of a collection of regions, with each region represented by a feature vector and its saliency. EMD calculates a distance between the representations provided by the image signatures by transforming the matching problem into providing a solution for a known transport distribution calculation, which is solved by linear programming. The problem is a matter of transporting goods from a set of suppliers to a set of consumers by the least-cost route, defined by a flow. The intuitive idea of EMD is to assume that a certain quantity of earth is used to fill up a number of holes in the ground, [34]. The query image is associated to a specific quantity of earth, grouped on heaps, while each candidate image for retrieval is assimilated with a number of holes in the ground. Each heap and each hole correspond to a region, either in the query or in the retrieved image, respectively, while the earth corresponds to the image region feature vectors and their characteristic saliency.

We consider the distance between the two sets of features, corresponding to the regions of the query and any candidate images, as a dissimilarity cost function $D(Q_i, I_j)$:

$$D(Q_i, I_j) = \psi(S_{Q_i}, S_{I_j}) \sqrt{\beta_P(\lambda_{cl} d_{cl}^2(i,j) + \lambda_{te} d_{te}^2(i,j) + \lambda_{co} d_{co}^2(i,j)) + \beta_S(\lambda_{nn} d_{nn}^2(i,j) + \lambda_{cd} d_{cd}^2(i,j))} \tag{9}$$

where Q_i, $i = 1, \ldots, M$ from the query image Q and each region I_j, $j = 1, \ldots, N$ from the candidate retrieval image I. $\psi(S_{Q_i}, S_{I_j})$ denotes the joint saliency weight for Q_i and I_j. d_{cl}, d_{te} and d_{co} are the Euclidean distances between the primary features, weighted by β_P, corresponding to the colour, texture and contrast vectors, respectively. Meanwhile, d_{nn} and d_{cd} are the Euclidean distances between the secondary features, weighted by β_S, characterizing the colours of the nearest neighbouring regions and the centroid locations of the regions Q_i and I_j, respectively. Each feature distance component is normalized to the interval $[0, 1]$ and is weighted according to its significance for retrieval by the global weights β_P, β_S, modulating the significance for each category of features, and the individual weights, weighting the contribution of each feature as: λ_{cl}, λ_{te}, λ_{co}, λ_{nn} and λ_{cd}. The selection of primary and secondary features $\beta_P > \beta_S$, where $\beta_P + \beta_S = 1$ was performed based on computational visual attention studies [16, 46] and following extensive empirical experimentation.

The feature modelling for each segmented region is described in Sect. 3.2 and distances are calculated between vectors of features characterizing image regions from the database and those of the query image. The CIEDE2000 colour distance was chosen for the colour components of the two vectors, because it provides a better colour discrimination according to the CIE minimal perceptual colour difference, [36]. The colour feature distance d_{cl} is calculated as:

$$d_{cl}^2 = \frac{1}{6}\left[3\left(\frac{\Delta E_{00}(i,j)}{C_{Dis}}\right)^2 + \left(\frac{\sigma_{i,L*} - \sigma_{j,L*}}{\alpha_{L*}}\right)^2 + \left(\frac{\sigma_{i,a*} - \sigma_{j,a*}}{\alpha_{a*}}\right)^2 + \left(\frac{\sigma_{i,b*} - \sigma_{j,b*}}{\alpha_{b*}}\right)^2\right] \tag{10}$$

where $\Delta E_{00}(i,j)$ represents the CIEDE2000 colour difference [36], calculated between the median estimates of L*, a*, b* colour components, normalized by the largest color distance C_{Dis}, while $\{\sigma_{x,c} | x \in \{i,j\}, c \in \{L*, a*, b*\}\}$ represent the standard deviations for each colour component, and $\{\alpha_c | c \in \{L*, a*, b*\}\}$ are their corresponding 95th percentiles, calculated across the entire image database, and are used as robust normalization factors. These values are used for normalization because the cumulative distributions of these features, extracted from segmented image regions, can be modelled by log-normal functions.

The texture distance d_{te} corresponds to the Euclidean distance between the average of the absolute values of DWT coefficients corresponding to the horizontal, vertical and oblique directions for the regions Q_i and I_j, divided by their corresponding standard deviations calculated across the entire image database. The contrast difference d_{co} is represented by the normalized Euclidean distance of the contrast features for each region from I and Q, with respect to their neighbouring regions. For the sake of robust normalization, the distances corresponding to the texture features as well as those representing local contrast are divided by the 95th percentiles of the empirical cumulative distribution function of their features, computed from a representative image set. The neighbourhood characteristic difference d_{nn} is calculated as the average of the resulting 12 colour space distances to the four nearest neighbouring regions from above, below, left and right, selected such that they maximize the joint boundary in their respective direction. The centroid distance d_{cd} is the Euclidean distance between the coordinates of the regions centers.

The weight corresponding to the saliency, weighting the inter-region distances between two image regions from Q_i and I_j, from (9), is given by:

$$\psi(S_{Q_i}, S_{I_j}) = \max\left(1 - \frac{S_{Q_i} + S_{I_j}}{2}, 0.1\right) \tag{11}$$

where S_{Q_i} and S_{I_j} represent the saliency of the query image region Q_i and that of the candidate retrieval image region I_j, where the saliency of each region is calculated, following the analysis from Sect. 4.3, and represents the ratios of salient pixels from each region. It can be observed that the distance $D(Q_i, I_j)$ is smaller when the two regions Q_i and I_j are both salient. Eventually, for all regions from Q and I, it results a similarity matrix $D(Q_i, I_j)$ which defines a set of inter-region distances between each region Q_i, $i = 1, \ldots, M$ from the query image Q and each region I_j, $j = 1, \ldots, N$ from the candidate retrieval image I. The resulting inter-region similarity matrix acts as the ground distance matrix for the EMD algorithm.

The distance matrix $D(Q_i, I_j)$ represents the cost of moving the earth energy associated with the image regions from Q to fitting the gaps of energy, represented by the image regions from I. A set of weights $\{w_{Q,i} | i = 1, \ldots, M\}$ is associated with the amount of energy corresponding to a region in the query image, while $\{w_{I,j} | j = 1, \ldots, N\}$ are the weights corresponding to the candidate image, representing the size of an energy gap. All these weights represent the ratios of each segmented region from the entire image. A unit of flow is defined as the transportation of a unit of energy across a unit of ground distance. The EMD algorithm is an optimization algorithm which minimizes the cost required for transporting the energy to a specific energy gap, [34]:

$$\min(\sum_{i=1}^{M} \sum_{j=1}^{N} f_{ij} D(Q_i, I_j)) \tag{12}$$

which is subject to the following constraints:

$$f_{ij} > 0, \; i = 1, \ldots, M, j = 1, \ldots, N \tag{13}$$

$$\textstyle\sum_{j=1}^{N} f_{ij} \le w_{Q,i}, \; i = 1, \ldots, M \tag{14}$$

$$\textstyle\sum_{i=1}^{M} f_{ij} \le w_{I,j}, \; j = 1, \ldots, N \tag{15}$$

$$\textstyle\sum_{i=1}^{M} \sum_{j=1}^{N} f_{ij} = \min\left(\sum_{i=1}^{M} w_{Q,i}, \sum_{j=1}^{N} f_{ij} w_{I,j}\right) \tag{16}$$

The goal of the optimization procedure is to find the flow f_{ij} between the regions Q_i and I_j such that the cost of matching the energy from a surplus area to a deficit of energy area is minimized.

After solving this system by using linear programming, the EMD distance from (8) is calculated by normalizing the cost required:

$$\text{EMD}(Q, I) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} f_{ij} D(Q_i, I_j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} f_{ij}} \tag{17}$$

This represents the normalized cost of matching the query image signature with that of the most appropriate candidate retrieval image. The weights add up to unity only when all image regions are used. We are removing non-salient image regions, and consequently the weights would add up to a value less than one. Such signatures enable partial matching which is essential for image retrieval where there is a high likelihood of occlusion in the salient regions. The computational complexity of the proposed QSCR is contained mostly in the feature extraction stage for the given image database which is performed off-line. The computational complexity of the optimization algorithm can be substantially reduced when thresholding the distances calculated by EMD, as in [31].

6 Experimental Results

6.1 Image Retrieval Databases

In the following we apply the methodology described in the previous sections, to three different databases: COREL 1000, SIVAL and Flickr. COREL 1000 is well known for its medium-to-high image complexity and its size makes it a good choice for the development of retrieval algorithms. The database consists of ten semantic categories of natural scenes, each containing 100 images. The categories from COREL 1000 are: Africa and its people (AFR), tropical seaside and beaches (BEA), Greek and Roman architecture (ARC), buses and coaches (BUS), dinosaur illustrations (DIN), elephants in an African environment (ELE), close-ups of flowers and bouquets (FLO), brown and white horses in a natural setting (HOR), mountain landscapes and glaciers (LAN) and food and cuisine (FOO). The SIVAL (Spatially Independent, Variable Area, and Lighting) database was designed for localized image retrieval [33], by containing a large number of similar images that only differ in the salient object. It consists of 1500 images in 25 categories with 60 images per category (10 scenes, 6 photos per scene). The SIVAL categories are: chequered scarf (CS), gold medal (GM), fabric softener box (FS), coke can (CC), Julie's pot (JP), green tea box (GT), translucent bowl (TB), blue scrunge (BS), glazed wood pot (GW), felt flower rug (FF), WD40 can (WD), smiley face doll (SF), data mining book (DM), Ajax orange (AO), Sprite can (SC), apple (AP), dirty running shoe (DS), banana (BA), striped notebook (SN), candle with holder (CH), cardboard box (CB), wood rolling pin (WP), dirty work gloves (DG), rap book (RB) and large spoon (LS). The Flickr database consists of 20 categories with 100 highly diverse images in each, and 2000 images with no specific concept. The following categories are part of this database: Mexico city taxi (MC), American flag (US), New York taxi (NY), snow boarding (SB), Pepsi can (PC), fire and flames (FF), sushi (SU), orchard (OR), fireworks (FI), Persian rug (PR), waterfall (WA), Coca Cola can (CC), Canadian mounted police (MO), ostrich (OS), boat (BO), keyboard (KE), honey bee (HB), cat (CA), samurai helmet (SH) and Irish flag (IF).

6.2 Image Retrieval Performance Measures

The basic retrieval assessment is provided by precision and recall. Precision represents the number of relevant images retrieved over the total number of retrieved images, while the recall represents the number of relevant images retrieved divided by the number of relevant images in database. A precision-recall (PR) curve can be plotted by classifying all candidate images as relevant/irrelevant according to their ground truth category and then by assigning a confidence value for the decision of that classification. In this study, the confidence value is the reciprocal of the dissimilarity measure, i.e. lower dissimilarity implies more confidence. Another statistical assessment measure is the Receiver Operating Characteristic (ROC) which plots the true positive rate versus the false positive rate (false alarm) by changing a decision threshold, and can be used to select the optimal number of images to be retrieved such that both measures are maximized. The area under the ROC curve (AUC) , which corresponds to the Wilcoxon-Mann-Whitney statistic [33], is a reliable image retrieval assessment measure. This can be interpreted as the probability that a randomly chosen positive image will be ranked higher than a randomly chosen negative image. A value above 0.5 means that the image retrieval method is more likely to choose a positive image, while a value below 0.5 means that the system is more likely to choose negative images which is worse than guessing.

In this study, images are ranked based on their similarity to the query, thus producing an ordered set of results. The rank-weighted average precision (WPR) is given by, Wang et al. [44]:

$$\mathrm{WPR} = \frac{1}{N}\sum_{k=1}^{N}\frac{n_k}{k} \tag{18}$$

where N is the number of all retrieved images and n_k is the number of matches in the first k retrieved images. This measure gives more weight to matched items occurring closer to the top of the list and takes into account both precision and ranks. Ranks can be equated to recall because a higher WPR value means that relevant images are closer to the top of the list, therefore the precision would be high at lower recall values because more relevant images are retrieved. However the WPR measure is sensitive to the ratio of positive and negative examples in the database, i.e. the total number of relevant images out of the total number of candidate images.

Quantitative tests are performed by evaluating the average performance of the proposed methodology across the whole databases considering 300 queries for COREL 1000, 600 queries for Flickr, and 750 for SIVAL. Across the graph legends in this study, μ indicates the mean value for the measure represented, calculated across all categories and followed by a $\pm\sigma$ which denotes the average of the spreads.

6.3 Visual Attention Models

Following the analysis of various image saliency selection algorithms from Sect. 4.2 we use Graph-Based Visual Saliency (GBVS) algorithm for selecting the saliency in the context of the optimization algorithm, as described in Sect. 5. Using saliency as a weight for the Euclidean distances of the feature vectors is compared against the case when salience is not used at all. The Area under the ROC curve (AUC) results for COREL 1000 database are provided in Fig. 7. From this figure it can be observed that saliency improves the performance in categories where salient objects are prominent in the image such as Flowers, Horses, Dinosaurs, and

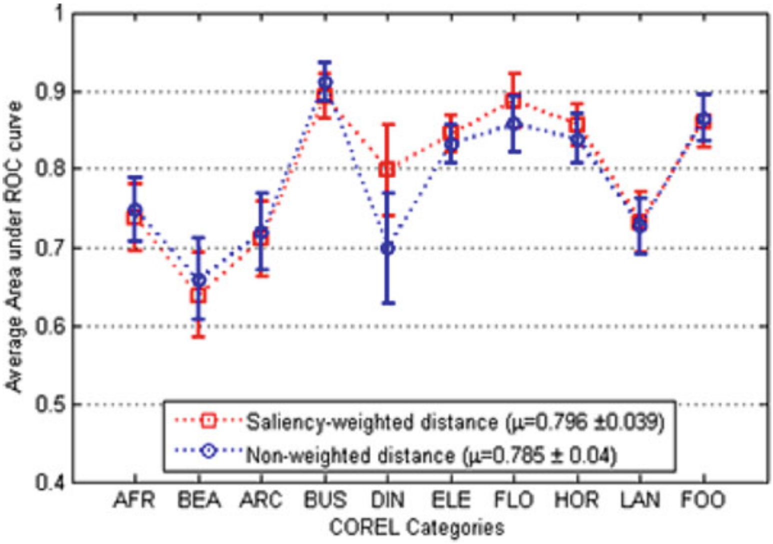


Fig. 7 Comparison between saliency weighted distances and non-weighted distances

decreases in categories where salient objects are difficult to identify, such as Beaches, Architecture, Africa, and Cuisine. This happens because in the former categories, the saliency weight gives preference to genuine candidate salient regions that correspond to a salient object well represented in that category, rather than the latter cases, where salient regions are specific in each image. Statistically, on the entire database, the mean (μ) of AUC, provided in Fig. 7, indicate that saliency is useful when used as a weighting of the distance measure.

6.4 Selecting Salient Regions from Images

In Fig. 8 we present the rank-weighted average precision results in ten image categories from COREL 1000 database when selecting the top 20% salient image data. This is compared with the case of using a fixed cut-off threshold of 0.607, which corresponds approximately to selecting the top 15% salient image data.

In Fig. 9 we provide a comparative study for the retrieval results when considering as salient regions only those corresponding to the top 65% of all salient regions from the CDF of salient regions, computed as described above. In Fig. 9a we compare the rank-weighted average precision results for the proposed image saliency region selection approach when compared to the approach which considers only the 50% most salient regions. In Fig. 9b the comparison is with a method using the maximization of entropy for the average region saliency values, proposed in [14], when using 100 saliency levels. The method based on the maximization of entropy

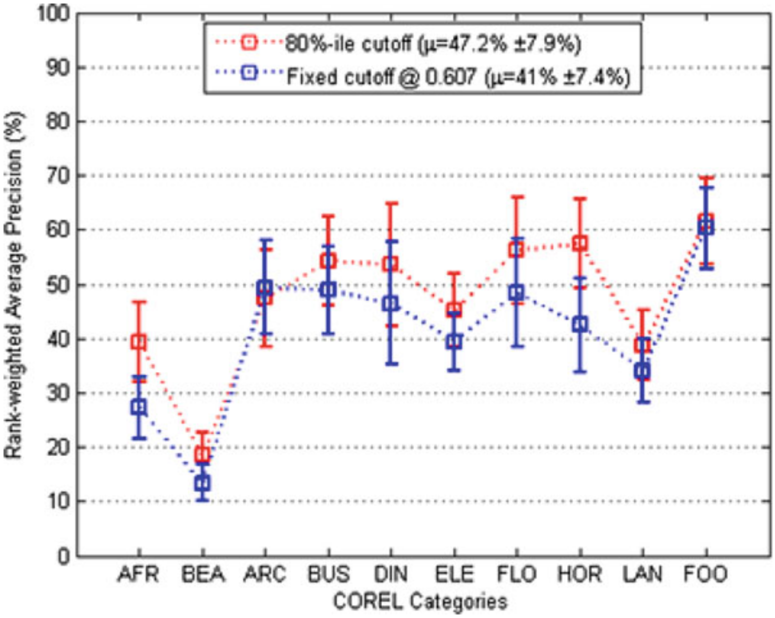


Fig. 8 Image retrieval when considering different saliency map thresholds. μ indicates the average rank-weighted precision followed by the average of the corresponding standard deviations after the ± sign. Standard deviations are indicated for each image category in the plot as well

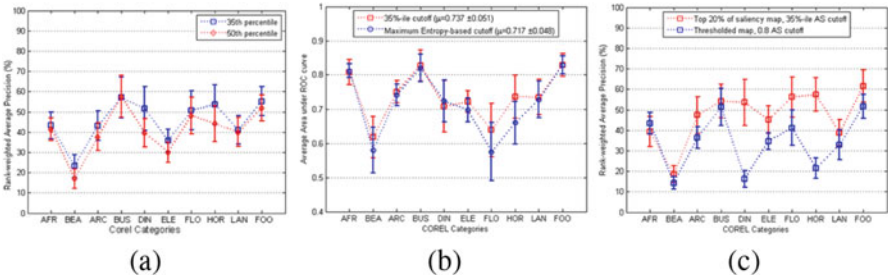


Fig. 9 Comparisons for various ways of selecting salient regions. (**a**) Rank-weighted average precision (WPR) when selecting salient regions based on the percentile of salient pixels. (**b**) Area under the ROC curve (AUC) when selecting salient regions based on the maximization of the saliency entropy. (**c**) WPR when salient regions are extracted using a thresholded saliency map

for the average region saliency values is not suitable for retrieving the images from Flower and Horse categories because it does not select enough background regions to differentiate the red/yellow flowers from buses. In both of these plots it can be observed that by selecting the top 65% salient regions outperforms the other approaches. Another method for selecting salient regions consists of binarising the saliency map using Otsu's threshold proposed in [28], then choosing the salient regions as those which have at least 80% of their pixels as salient. Figure 9c shows that this method underperforms greatly when categories have a well-defined salient object. This happens because this method selects just the salient object without including any background regions, and since those categories are classified as distinctive scenes, confusion occurs due to the semantic gap. On the other hand, the proposed QSCR method considers only the top 65% salient regions, and this was shown to be efficient in general-purpose image data sets, such as Corel and Flickr databases. However, in the case of SIVAL database, which consists entirely of distinctive objects with no semantic link to their backgrounds, salient regions are considered when they are part of the top 40% most salient regions, due to the fact that in this case the inclusion of background regions introduces false positives.

6.5 Similarity Ranking

Salient edges are extracted as explained in Sect. 4.1 and are used in the final image ranking evaluation measure from (8). The idea is that the region-to-region matching EMD distance gives a localized representation of the image while the salient edges provide a global view. Unlike in SEHD algorithm of [14], the QSCR method decouples the edge histogram from its spatial domain by considering the edge energy, corresponding to specific image feature orientations, calculated from the entire image. SEHD uses a different image segmentation algorithm and different selection of salient regions while performing the image ranking as in [7]. In Fig. 10 we compare the proposed salient edge retrieval approach, considering only the global image saliency and not the local saliency, and SEHD image retrieval method used in [14], using the average area under the ROC curve (AUC) as the comparison criterion. The categories in which the proposed approach outperforms SEHD are

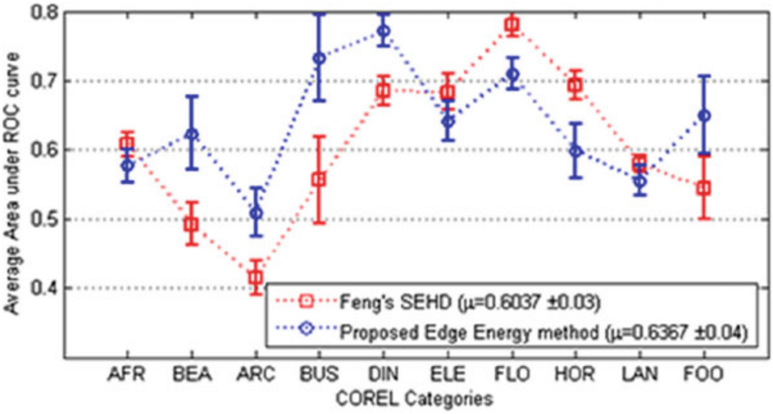


Fig. 10 Retrieval by salient edges: proposed salient edge representation compared with the Feng's SEHD approach

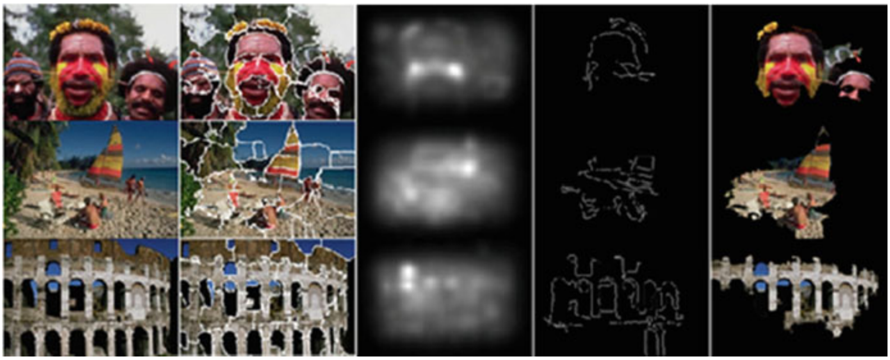


Fig. 11 Examples of extracted query information: (1) Original, (2) Image segments, (3) Saliency map, (4) Salient edges, (5) Selected salient regions

those where there is a significant amount of variation in the spatial positions of edges within the images, such as beaches and buses. The mean AUC value for SEHD and for the proposed method are 0.6037 and 0.6367, respectively. Thus, performing a two-tailed Students t-test at the highly significant 1% level with 598 degrees of freedom yields a p-value of 0.0022 which shows that the difference is statistically significant.

In Fig. 11 we provide the results for three images from three distinct image categories of COREL 1000 database after segmentation, saliency map estimation, salient edge extraction and salient region selection. These images are quite challenging due to their textured context. Examples of images from the other seven image categories from COREL database are shown in Fig. 12. It can be observed that the selected salient regions include contextual information such as in the second image from Fig. 11 and in the first, third, fifth and seventh images from Fig. 12. The salient object context is very important for image retrieval as shown in the full database results. Moreover, in the second image from Fig. 12, contextual regions are not selected since in this case they are not relevant because the main salient object is not related to its background. We have observed that the mean-shift algorithm leads to over-segmentation in some cases. However, this does not affect the salient region selection which is mainly driven by the saliency content and by the salient region selection procedure described in Sect. 4.3. Since the salient region selection is based on relative statistical measures, similar results would be obtained when using a different image segmentation algorithm.

Images are ranked according to the similarity measure $\mathscr{S}(Q, I)$ from (8), between the query image Q and a candidate retrieval image I. In Fig. 13a we present the retrieval results of 30 images for two different image categories from COREL 1000 database. In the query image, which is part of the Architecture category, from Fig. 13a it can be observed that the core of the salient region is a false positive, because the true object takes most of the image, and the most dissimilar area is a patch of sky in the middle. Nevertheless, the retrieval succeeds because the region

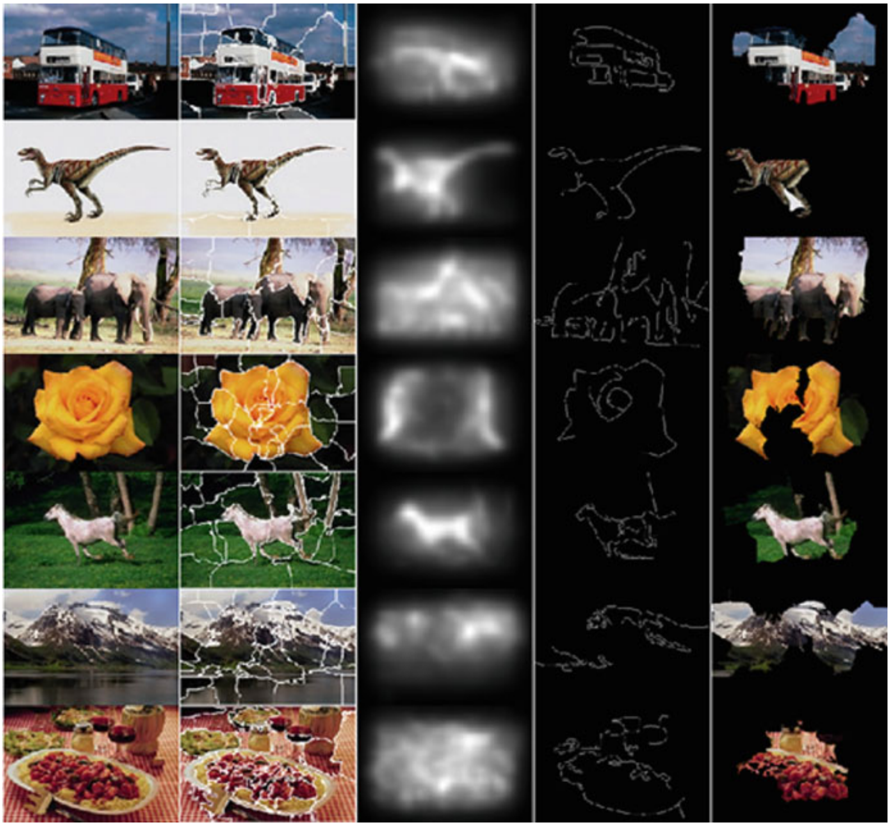


Fig. 12 Extracting query information from images for seven image categories from COREL database. The image columns indicate *from left to right*: original image, segmented regions, the GBVS saliency map, salient edges and the salient regions

selection method includes the surrounding regions in the query. The precision-recall (PR) curve corresponding to the query images is shown in Fig. 13b. Figure 14 shows a scenario where the number of positive examples in the category is much smaller, and yet the AUC value is high, as it can be observed from Fig. 14b. This means that if more positive examples were added to the database, then the precision would improve. Because all images in the category are considered relevant and the true number of positive examples is much lower, the curve underestimates the true retrieval performance. The semantic gap is evident in the retrieval of this image as the query regions contain ambiguous colours, resulting in a series of Horse and Food category images as close matches. The results when retrieving the white horse in natural habitat surroundings from Fig. 15 produces no false positives for the first 10 retrieved images, but after that creates some confusion with Africa (which basically represents people), as well as with Flowers and Elephant categories.

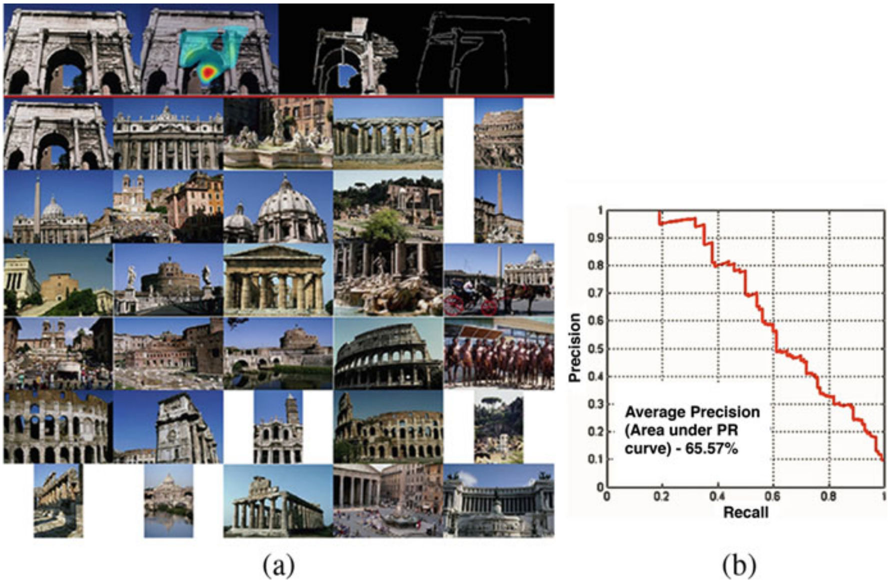


Fig. 13 Retrieval performance for Architecture category from COREL 1000 database. (**a**) The *first line* shows the query image, its saliency, selected salient regions and salient edge images while the subsequent lines displays the retrieved images in their order. In the *next six lines* are shown 30 retrieved images. (**b**) Precision-recall curve

A variety of good retrieval results are provided in Fig. 16a, b for the Bus and Flower categories from COREL 1000 database, while Fig. 16c, d shows the results for images from the Pepsi Can and Checkered Scarf categories from SIVAL database. The last two examples of more specific image categories from SIVAL database indicate very limited salient object confusion in the retrieval results.

6.6 Results for Entire Databases

Figure 17 compares the results for the proposed query by saliency content retrieval (QSCR) algorithm with SIMPLIcity from [44] when applied to COREL 1000 database. The comparison uses the same performance measures as in [44], respectively the average precision, average rank and average standard deviation of rank. As it can be observed from Fig. 17, QSCR provides better results in 4 image categories and worse in the other 6, according to the measures used. This is due to the fact that SIMPLIcity uses very selective features which are appropriate for these 6 image categories.

Figure 18 compares the results of QSCR, with the two retrieval methods proposed in [33], on Flickr database when using AUC. The results of QSCR and ACCIO are

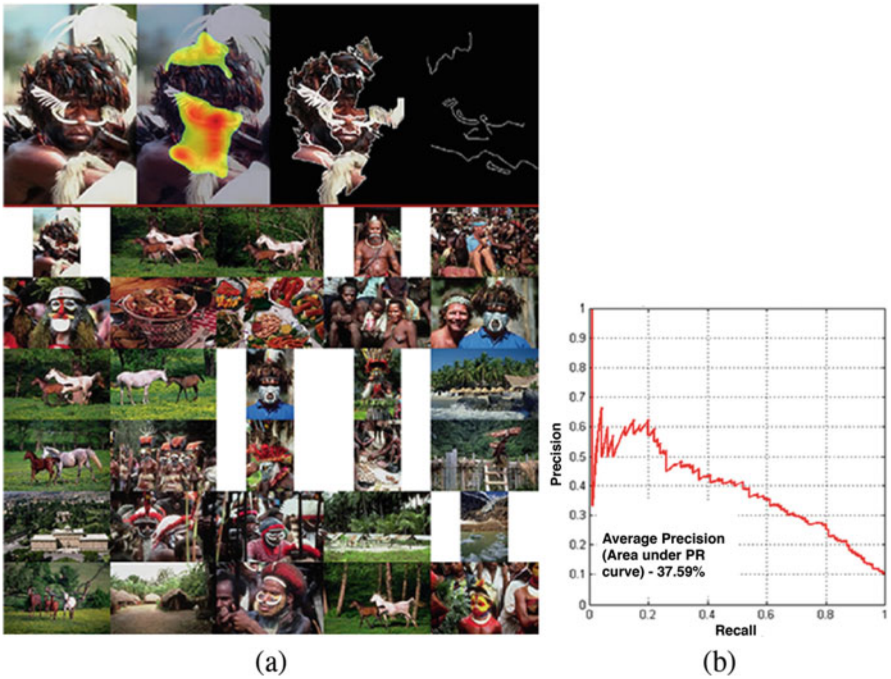
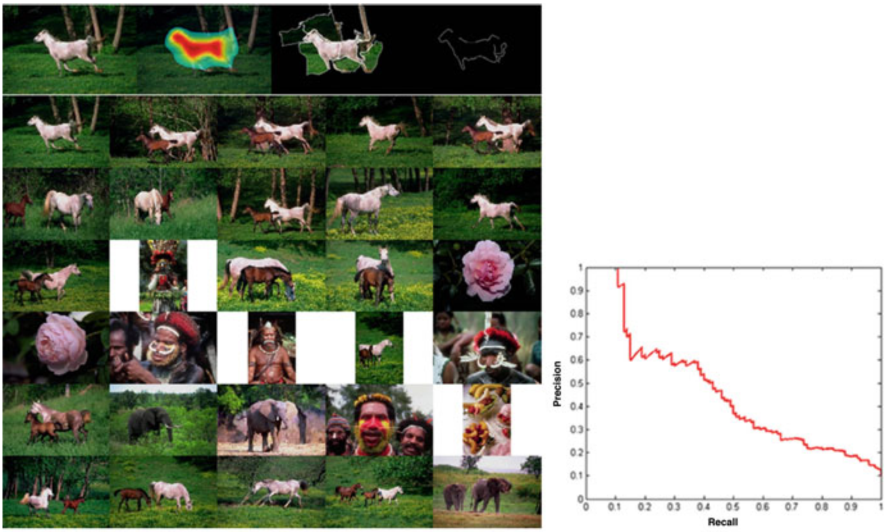


Fig. 14 Retrieval performance for Africa category from COREL 1000 database. (**a**) The *first line* shows the query image, its saliency, selected salient regions and salient edge images while the subsequent lines displays the retrieved images in their order. In the *next six lines* are shown 30 retrieved images. (**b**) Precision-recall curve

broadly similar and vary from one image category to another. However, ACCIO involves human intervention by acknowledging or not the retrieved images, while the approach described in this chapter is completely automatic. The salient edges improve the performance when the image of a certain category contain salient objects which are neither distinctive or diverse enough. This is the case with the SB category, where most of the photos depict people as salient objects, set in a snowy environment, HB, FF and FI categories, where the images are mostly close-ups, defined by mostly vertical edges.

Figure 19 provides the assessment of the retrieval results using AUC on SIVAL database when considering the retrieval of five images for each category. In this database, the objects have simple backgrounds and the saliency should highlight the main object while excluding the background which is the same for other image categories. Unlike in COREL 1000 and Flickr databases, the inclusion of the background is detrimental to the retrieval performance in this database. In the case of the images from SIVAL database we consider as salient those regions whose saliency corresponds to the top 40% of salient regions in the image instead of 35% which was used for the other two databases.



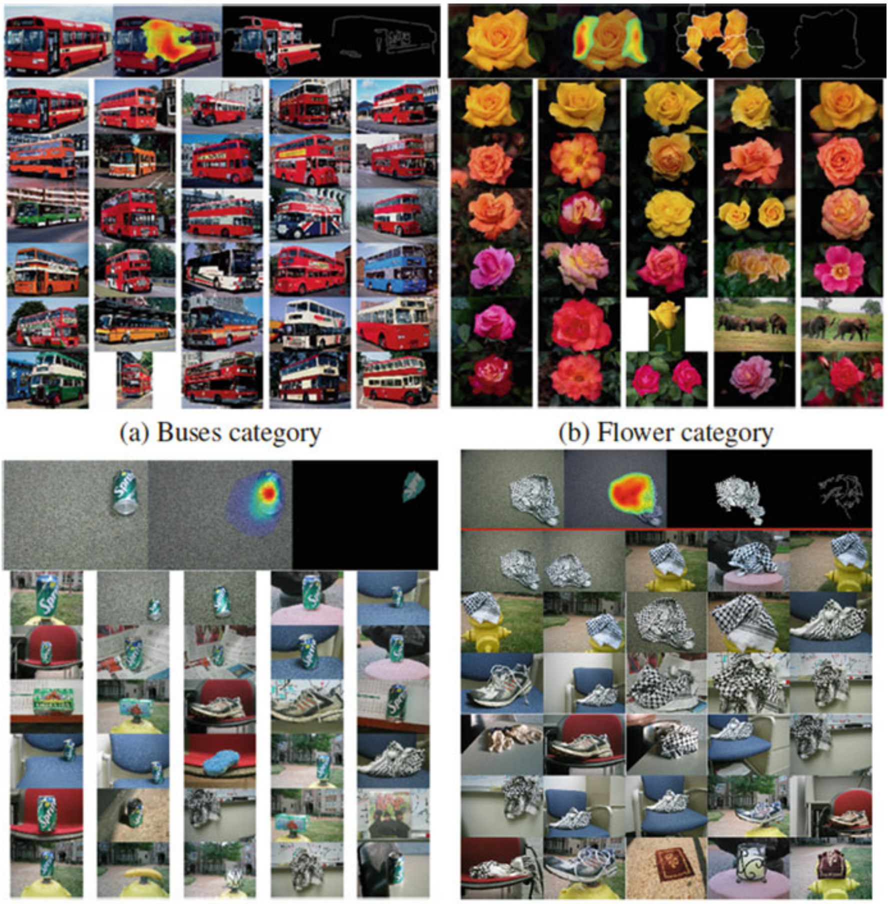
(a) (b)

Fig. 15 Retrieval performance for Horse category from COREL 1000 database. (**a**) The *first line* shows the query image, its saliency, selected salient regions and salient edge images while the subsequent lines displays the retrieved images in their order. In the *next six lines* are shown 30 retrieved images. (**b**) Precision-recall curve

6.7 Discussion

Ideally, a larger database of millions of images should be used for assessing the image retrieval. The image segmentation currently takes about 90% of the feature extraction time. Tuning of feature weights usually came down to a decision regarding the trade-off between specificity and generality. As it can be seen from the results from Fig. 16, it is possible to obtain images that visually are highly similar to the query, in terms of colour, orientation, and position, at the cost of lower recall, since only a fraction of the category has those exact images. This may be a bad thing for the retrieval of images in general, but if the user were looking for images in a specific image sequence, then this would be the best way to achieve that goal. Qualitative tests for certain features are sensitive to the query image because some images will satisfy the criterion under evaluation and hence return better results for one specific category and at the same time reduce effectiveness in another.

Corel database is well balanced in terms of objects and scenes. Thus, maximising the average performance across all categories should produce a robust system. Nevertheless, in a few cases, the changes that improved the retrieval on the Corel database, reduced the performance on the Flickr database. Due to the varied nature of the Flickr images within each category, application of distance normalisation uncovered the true, large distances between features of the images within the category, which would otherwise have a negligible impact on ground distance



(a) Buses category

(b) Flower category

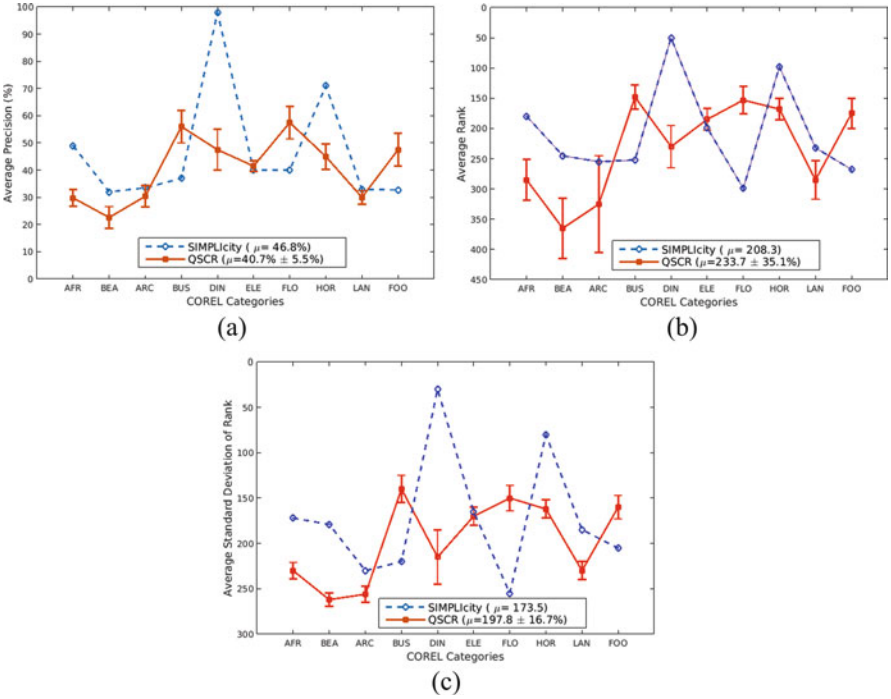
(c) Pepsi Can category

(d) Checkered Scarf category

Fig. 16 Retrieval performances for images from Corel database in (**a**) and (**b**) and from SIVAL database in (**c**) and (**d**)

because the domain of the feature values is very small. The semantic gap is most evident in this database because its images and ground truths were obtained by performing keyword search on Flickr. In addition, most of the images contain multiple salient areas, which combined with the deficiencies of computational visual attention models, result in several strong responses, which ultimately end up confusing the CBIR system.

The Corel database has also weaknesses. The categories are not entirely disjoint and it is sometimes unclear how to judge the retrieval results. When attempting to retrieve horses we may retrieve elephants as well because they are both animals and have similar relationships with their surroundings. At the lowest semantic level, this is incorrect as the retrieval is too general. Without keywords, if the user wished



(a) (b)

(c)

Fig. 17 Comparison results with SIMPLIcity

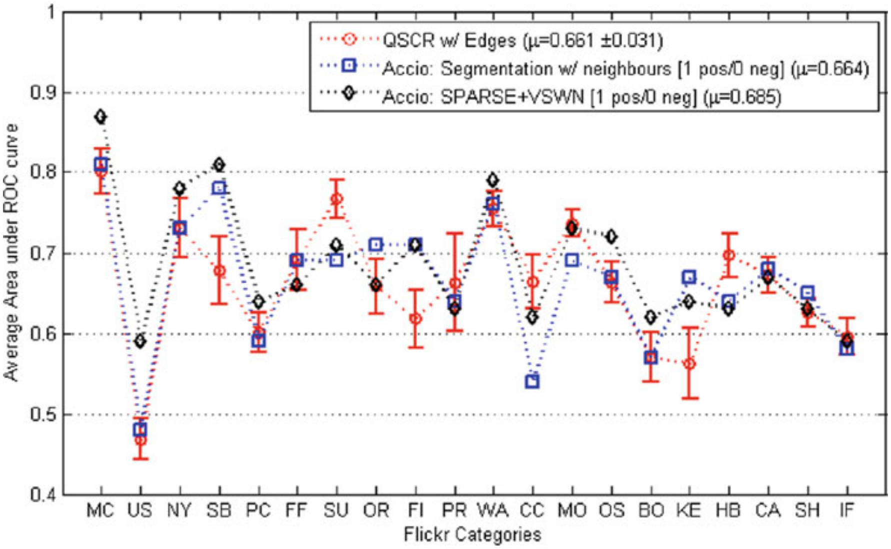


Fig. 18 Retrieval results on Flickr database and comparison with ACCIO

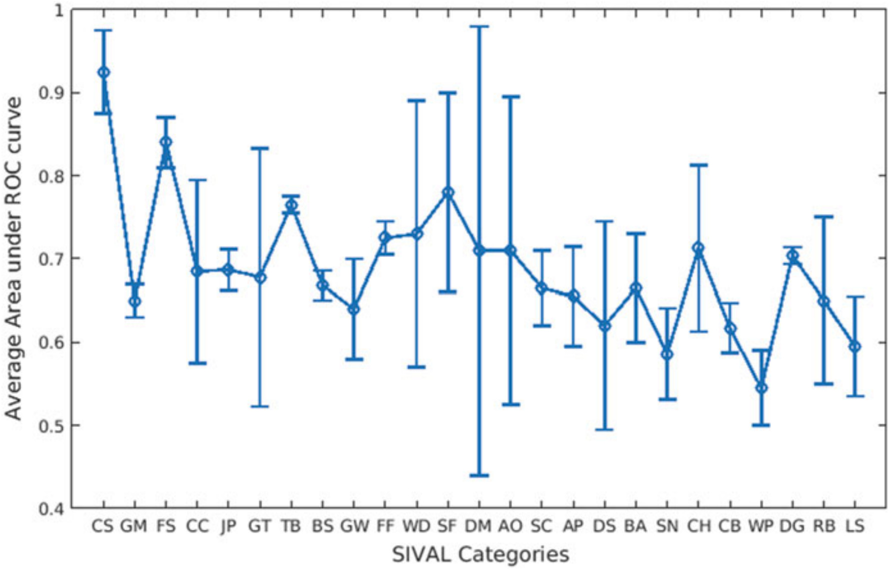


Fig. 19 Retrieval results on the SIVAL database

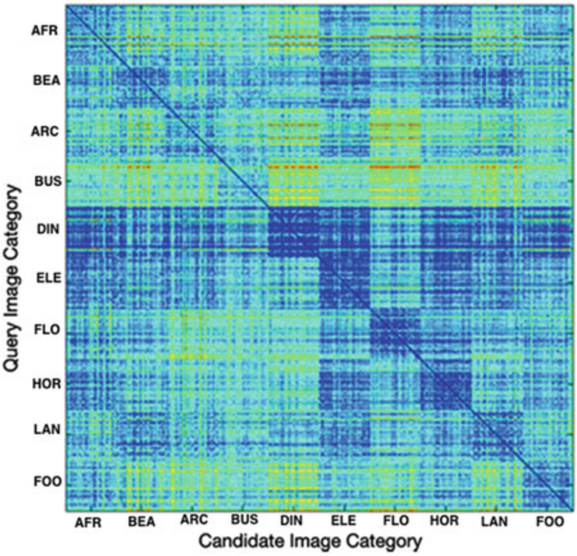
to search for animals, it would not be possible to specify such a query because "animal" is an abstract term. Such retrieval is only possible if the images are loosely clustered.

By considering distances using (8) between each pair of images for half of COREL 1000 database, an image classification matrix is produced, shown in Fig. 20. It shows the categories where the semantic gap is most prominent. It can be seen that Beach images are likely to get confused with Elephants, Landscapes, and Horses, whereas Elephants get mostly confused with Horses and to a lesser extent with Africa, Beaches, Architecture and Landscapes.

7 Conclusions

In this chapter we describe a visual saliency-driven retrieval system employing both local and global image saliency. Several visual attention models have been compared based on their ability to emphasise semantically meaningful areas of an image. The use of second-order moments of a region's colour distribution has been shown to improve performance considerably on occasions where the semantic gap would otherwise have a negative effect. The contrast feature was shown to provide a small boost in performance, indicating a potential for discriminative power in some types of images. The use of salient edges was shown to improve results where there

Fig. 20 Dissimilarity matrix applied on the Corel database, where *darker texture* denotes higher similarity



is little spatial variation of the salient object within images. A new salient region selection method, that uses the cumulative distribution of the saliency values in the database to select an appropriate threshold, was discussed in this chapter. An ideal CBIR solution would incorporate a variety of search mechanisms and would select the best choice dynamically, thus maximising its performance. The use of visual attention models would be one of the mechanisms that a CBIR solution should employ because it is indispensable in highly localized scenarios such as those found in the SIVAL database, a global ranking method would fail in SIVAL, regardless of the choices of features and distance metrics. This implies that systems must be able to distinguish between images of distinctive objects or distinctive scenes, leading to the thought of using the visual attention when searching for image content. Little work has been done before on such semantics-sensitive approaches to the retrieval task and it would be of great benefit to future CBIR systems. In their current state, computational models of visual attention are still basic because they operate on the notion of contrasting features, so they cannot accurately identify salient objects in complex images that are commonplace. Therefore, saliency models are the limiting factor for the concept of retrieval by visually salient objects and image features. In the future more reliable models of the human intent, such as those involving human memorisation processes, should be considered for CBIR systems in order to provide better retrieval results.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1597–1604 (2009)
2. Ashley, J.J., Barber, R., Flickner, M., Hafner, J.L., Lee, D., Niblack, C.W., Petkovic, D.: Automatic and semiautomatic methods for image annotation and retrieval in query by image content (QBIC). In: Proceedings of SPIE 2420, Storage and Retrieval for Image and Video Databases III, San Jose, CA, pp. 24–35 (1995)
3. Bors, A.G., Nasios, N.: Kernel bandwidth estimation for nonparametric modelling. IEEE Trans. Syst. Man Cybern. B Cybern. **39**(6), 1543–1555 (2009)
4. Bi J., Chen, Y., Wang, J.: A sparse support vector machine approach to region-based image categorization. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, vol. 19, pp. 1121–1128 (2005)
5. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8), 1026–1038 (2002)
6. Chang, T., Kuo, C.-C.J.: Texture analysis and classification with tree-structured wavelet transform. IEEE Trans. Image Process. **2**(4), 429–441 (1993)
7. Chang, R., Liao, C.C.C.: Region-based image retrieval using edgeflow segmentation and region adjacency graph. In: Proceedings of International Conference on Multimedia and Expo, pp. 1883–1886 (2004)
8. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with images. J. Mach. Learn. Res. **5**, 913–939 (2004)
9. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. **17**(8), 790–799 (1995)
10. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
11. Datta, R., Joshi, D., Li, J., Wan, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), 5:1–5:60 (2008)
12. Derrington, A.M., Krauskopf, J., Lennie, P.: Chromatic mechanisms in lateral geniculate. Nucleus Macaque J. Physiol. **357**, 241–265 (1984)
13. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and effective querying by image content. J. Intell. Inf. Syst. **3**, 231–262 (1994)
14. Feng, S., Xu, D., Yang, X.: Attention-driven salient edge(s) and region(s) extraction with application to CBIR. Signal Process. **90**(1), 1–15 (2010)
15. Frintrop, S., Klodt, M., Rome, E.: International Conference on Computer Vision Systems (2007)
16. Frintrop, S., Rome, E., Christensen, H.: Computational visual attention systems and their cognitive foundations. A survey. ACM Trans. Appl. Percept. **7**(1), 6.1–6.46 (2010)
17. Gao, D., Han, S., Vasconcelos, N.: Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(6), 989–1005 (2009)
18. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), vol. 19, pp. 545–552 (2007)
19. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
20. Islam, M.M., Zhang, D., Lu, G.: Comparison of retrieval effectiveness of different region based image representations. In: International Conference on Information, Communications and Signal Processing, pp. 1–6 (2007)

21. Itti, L., Ullman, S.: Shifts in selective visual attention, Towards the underlying neural circuitry. Hum. Neurobiol. **4**(4), 219–227 (1985)
22. 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)
23. Jing, F., Li, M., Zhang, H.-J., Zhang, B.: An efficient and effective region-based image retrieval framework. IEEE Trans. Image Process. **13**(5), 699–709 (2004)
24. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vis. **45**(2), 83–105 (2001)
25. Kelly, P.M., Cannon, M., Hush, D.R.: Query by image example: the CANDID approach. In: Proceedings of SPIE 2420, San Jose, CA, pp. 238–248 (1995)
26. Mallat, S.G.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic, New York (2009)
27. Ogle, V.E., Stonebraker, M.: Chabot: retrieval from a relational database of images. IEEE Comput. **28**(9), 40–48 (1995)
28. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
29. Papushoy, A., Bors, A.G.: Image retrieval based on query by saliency content. Digital Signal Process. **36**(1), 156–173 (2015)
30. Papushoy, A., Bors, A.G.: Visual attention for content based image retrieval. In: Proceedings of IEEE International Conference on Image Processing (ICIP), Quebec City, pp. 971–975 (2015)
31. Pele, O., Taskar, B.: The tangent earth mover's distance. In: Proceedings of Geometric Science of Information. Lecture Notes on Computer Science, vol. 8085, pp. 397–404 (2013)
32. Pentland, A., Picard, R.W., Sclaroff, S.: Photobook: content-based manipulation of image databases. Int. J. Comput. Vis. **18**(3), 233–254 (1995)
33. Rahmani, R., Goldman, S.A., Zhang, H., Cholleti, S.R., Fritts, J.E.: Localized content-based image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1902–1912 (2008)
34. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover distance as a metric for image retrieval. Int. J. Comput. Vis. **402**, 99–121 (2000)
35. Rui, Y., Huang, T., Ortega, M., Mehrotra, S.: Relevance feedback: a power tool in interactive content-based image retrieval. IEEE Trans. Circuits Syst. Video Technol. **8**(5), 644–655 (1998)
36. Sharma, G., Wu, W., Dalal, E.N.: The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. Color Res. Appl. **30**(1), 1–24 (2005)
37. Smith, J.R., Chang, S.-F.: VisualSEEk: a fully automated content-based image query system. In: Proceedings of the ACM International Conference on Multimedia, Boston, pp. 87–98 (1996)
38. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
39. Stentiford, F.W.M.: A visual attention estimator applied to image subject enhancement and colour and grey level compression. In: 17th International Conference on Pattern Recognition, vol. 3, Cambridge, pp. 638–641 (2004)
40. Tas, A.C., Luck, S.J., Hollingworth, A.: The relationship between visual attention and visual working memory encoding: a dissociation between covert and overt orienting. J. Exp. Psychol. Hum. Percept. Perform. **42**, 1121–1138 (2016)
41. Vondrick, C., Khosla, A., Pirsiavash, H., Malisiewicz, T., Torralba, A.: Visualizing object detection features. Int. J. Comput. Vis. **119**(2), 145–158 (2016)
42. Wan, J., Wang, D., Hoi, S., Wu, P., Zhu, J., Li, J.: Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of ACM International Conference on Multimedia, pp. 157–266 (2014)
43. Wang, J.Z., Wiederhold, G., Firschein, O., Sha, X.W.: Content-based image indexing and searching using Daubechies' wavelets. Int. J. Digit. Libr. **1**(4), 311–328 (1998)

44. Wang, J., Li, J., Wiederhold, G.: Simplicity: semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001)
45. Wang, H., Cai,Y., Zhang, Y., Pan, H., Lv, W., Han, H.: Deep learning for image retrieval: what works and what doesn't. In: Proceedings of IEEE International Conference on Data Mining Workshop, pp. 1576–1583 (2015)
46. Wolfe, J.M.: Visual search. In: Pashler, H. (ed.) Attention. Psychology Press, Hove, East Sussex (1998)
47. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN, A Bayesian framework for saliency using natural statistics. J. Vis. **8**(7), 32.1–32.20 (2008)