Electronic Letters on Computer Vision and Image Analysis 20(1):1-15; 2021

Saliency-Based Image Retrieval as a Refinement to Content-Based Image Retrieval

Mohammad Al-Azawi

Department of Computer Science and MIS, Oman College of Management and Technology
Muscat, Oman

Received 05th of October 2020; accepted 18th of December 2020

Abstract

Due to the importance of searching for an image in a database in various applications, many algorithms have been proposed to identify the contents of the image. Algorithms that identify the content of the image as a whole can offer good results in some applications and fail to produce satisfactory results in other applications. Therefore, searching for an object inside the image was used to overcome the limitations of identifying the image as a whole. Hence, studies focused on segmenting the image into small sub-images and identified their contents. In this paper, we introduce a new algorithm inspired by human attention and utilises the saliency principles to identify the contents of an image and search for similar objects in the images stored in a database. We also demonstrate that the use of salient objects produces better and more accurate results in the image retrieval process. A new retrieval algorithm is therefore presented here, focused on identifying the objects extracted from the salient regions. To assess the efficiency of the proposed algorithm, a new evaluation method is also proposed which considers the order of the retrieved image in assessing the efficiency of the algorithm.

Keywords-component; Image Saliency, Human Attention, Image Retrieval, Image Features, CBIR, SBIR

1 Introduction

Content-based image retrieval (CBIR) is one of the hot topics that attract the researchers to investigate due to the immense increase in the use of multimedia in various applications such as medical applications and machine vision. Traditional image recognition and retrieval algorithms identify the contents of the image as a whole based on features extracted from the image such as colour distribution, texture, and shapes. Using such kind of recognition may not produce satisfactory results because some images with different contents may have similar colour distributions or texture. Therefore, considering the image as a whole gives irrelevant results sometimes. Recognition by part (RBP) was used to reduce the effects of the aforementioned problem and improve the results. In this approach, image segmentation is used to divide the image into segments (regions) and each segment goes through a recognition process which describes it. The descriptions of all segments together form the overall description of the image. The main issue with RBP is the over-segmentation, i.e. segmenting the object itself into smaller chunks since it relies on the visual contents, which makes the identification process more complicated. In addition to the above problem, unimportant segments are considered for identification as well.

Inspired by the human visual system and human attention principles, we are presenting an approach that identifies the contents of the salient regions only and neglects other regions. Salient regions, which are the regions that attract human attention, can be used to identify the important parts of a scene. These regions are

Correspondence to: <mohd.alazawi@omancollege.edu.om>

Recommended for acceptance by Angel D Sappa
https://doi.org/10.5565/rev/elcvia.1325
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

extracted from the image using a saliency extraction algorithm. In this paper, we introduce a refined retrieval technique that utilises the principles of saliency in image retrieval. In the proposed technique, the salient objects in the query image are compared with the salient objects in the images stored in the database, which means that the algorithm does not match the whole image but only certain regions of it. This would produce better results as, in some cases, the unimportant regions might be dominant, and the effect of the salient region is negligible. For example, consider the case in which one needs to search for a ball in a field using colour histogram features. In this case, the effect of the surrounding environment on the histogram is much greater than the ball, so the images retrieved would be more related to the green grass than the ball. The same thing would happen with a bird or an aeroplane in the sky where the sky would be the dominant part.

The remainder of the paper will be structured as follows; the necessary background and reviews are provided in section 2. A brief overview of the proposed algorithm is given in section 3 and a review of the experimental results and the evaluation of the proposed technique is provided in section 4. Finally, the conclusions are presented in section 5.

2 Background and Review

In this section, the necessary background and review for the proposed algorithm are presented. Since the purpose of this paper is to present a new retrieval algorithm that utilises the concept of saliency, two key topics, the principles of image retrieval and saliency, need to be addressed here.

2.1 Features and Image Retrieval

Image retrieval systems are the systems that use descriptors extracted from low-level features (LLF) or high-level features (HLF) to recognise and describe images based on their content. Low-level features, such as colour and texture, are widely used for such purposes where metrics can be extracted from them. On the other hand, some researchers proposed using high-level or semantic features, such as shape and borders, by deriving them from low-level features.

One of the frequently used LLF in describing image contents is the colour histogram, which shows the frequency of occurrences of luminance values in an image. Measures, such as statistical measures, can be extracted from the histogram and used to identify the contents of an image. Many algorithms used colour histogram feature because it can provide a good description of the contents and does not get affected by the rotation, scaling, or transformation of the image [1].

The methods of comparing the histograms of two images are divided into two types, Bin-by-Bin Distance (BBD) and Cross-Bins Distance (CBD). In BBD, Minkowski distances, which find the distance between two histograms bin by bin, are widely used such as in ref [1], [2], and [3]. The intersection between histograms is another well-known similarity indicator, in which, if there are intersections between histograms, the similarity between the corresponding images is possible. This measure sums the minimum of the corresponding components in the two histograms, which is high if the intersection between the histograms is high, i.e. if the corresponding components in the two histograms are close to each other. Empirical experiments showed that the average accuracy of applying the BBD techniques is extremely low, in our test it was around 40%. This low accuracy is due to the low likelihood of corresponding bins to having similar values even if the images are similar but not identical. This is because BBD is highly affected by imaging environment such as lighting conditions that may shift the histogram. Therefore, we need a better measure that can find the relation between the histograms regardless of their shift or scale. One of the good approaches to achieving this is Cross-Bin Distance (CBD).

In CBD methods, the histogram itself is not used in the comparison process, instead, metrics such as statistical measures are used. Statistical measures such as expectation values, standard deviation and skewness are quite common. The results obtained from applying CBD are much more accurate than BBD since all bins are involved in calculating the metrics values.

The texture is the second significant low-level feature that is widely used in comparing the contents of the images. To describe the texture of an image or a region, Gabor filter, wavelet transform, or local statistics measures may be used. Tamura features, which are coarseness, directionality, regularity, contrast, line-likeness, and roughness, were used to describe the texture in many publications. Among these features, the first three are more important than others, other features are related to the previous three and do not add much information to the description. Another well-known statistical approach is the Grey Level Cooccurrence Matrices (GLCM), which is the most commonly used method to describe the texture [4].

High-level features, such as borders and shapes, can be derived from low-level features using various methods and mechanisms. For example, low-level points feature can be used to construct high-level features such as objects. Intrinsic structure finding is one of the methods that can convert LLF into HLF where the points are described in accordance with the structure. Border tracking is the second important technique for getting high-level features from low-level features. In border tracking, the border of the set of points is traced to give the external shape of the object. In the same way, other HLFs such as geometric structure and shape can be extracted using a variety of techniques which are beyond the scope of this research.

2.2 Attention and Saliency

The Human Vision System (HVS) functions as a passive selector or a bypass filter that acknowledges certain stimuli and reject others [5]; this feature is known as attention. Attention is one of the characteristics of human brain's information processing that is used to minimise the amount of information it needs to handle by choosing a subset of the available information that the brain focuses on and processes. Human attention comprises three aspects, which are orienting, filtering, and searching aspects; the information processing goes sequentially through these aspects [6],[7].

Saliency is the computer representation or simulation of human attention in which one can define the salient object as the object that captures the attention or attracts the human vision system. HVS uses two stages to identify the objects, pre-attentive and attentive stages [8]. In the pre-attentive stage, regions that present spatial discontinuity (pop-out features) of an image are detected, whereas the relationships and associations among these regions are found and grouped in the attentive stage. According to the above proposition, both low-level and high-level features are required to identify the salient object. Furthermore, both local features of the object and global details of the image are required for the same task. Local features, which are correspondent to the pre-attentive phase in human, give visual importance to the object locally, while global image details are correspondent to the attentive recognition of the HVS. The salient objects are defined using local features then the global details of the image are used to assign the importance or significance to them.

2.3 Salient Points and Regions Extraction

Literature classifies saliency extraction methods mainly into Bottom-Up (BU) and Top-Down (TD) saliency extraction classes [9]. In BU, the techniques utilise low-level features that can be derived from the regions such as colour, texture, and luminance to identify the saliency level of a region. In the second class (TD), a knowledge database is constructed first, then comes the search for interesting points according to this knowledge database. Another broad classification of the salient points extraction techniques was proposed by Toet [10], in which he classified the techniques as biologically-based, purely-computational, or a combination of both.

Several approaches were proposed to extract the salient points or regions from an image both in spatial and in frequency domains. Wavelet, which is a multiresolution representation that expresses image variations at different scales, was one of the techniques used for saliency identification. This approach was adopted in several researches such as [11], [12], [13], [14], and [15]. Geometric features such as corners were, also, used to identify the saliency of a point, they were used as a measure of saliency firstly by Schmid and Mohr [9], [16]. Comer detectors, such as Harris and the modified approaches like Moravec and SUZAN, were designed to be used with robotics vision and shape recognition; therefore, they have an excellent feature that they are constant with scaling, shifting, and rotation.

Saliency map is one of the well-known methods of extracting the salient regions. Early work in this field was done by Koch and Ullman [17] and Itti *et al.* [18]. Itti *et al.* developed a model that uses image hierarchy to identify the regions of attention. In their model, the images are reduced in size and resolution using Gaussian pyramid. The mentioned approach used low-level features such as intensity, colour, and orientation to highlight salient regions. Forty-two feature maps were generated from these features and the saliency map is extracted by combining the above maps. The inhibition of return was used to prohibit the algorithm from considering the same salient object more than once.

Frequency domain was used to extract the saliency of an object in many literatures such as [19], [20], [21], [22], [23], [24], and [25]. The idea behind using the frequency domain is that salient points are usually of high change in frequency domain both in magnitude and in orientation. Bruce et al. [19] is an example of such approaches, in which the authors suggested using the magnitude to define the salient regions in an image. They divided the image into sub-images and used Fourier Transform to convert the sub-images from spatial domain to frequency domain, then they calculated the magnitude of the image from the real and imaginary parts of the spectral image and consider the regions with high frequency as a salient region.

The strengths and weaknesses of the mentioned methods are beyond the scope of this paper, a sufficient discussion is found in [26] and [27], which also includes a new method of extracting the saliency based on the irregularity of the intensity of the region. In this method, some statistical measures are used to measure the irregularity of the region. Regions with high irregularity measure are considered to be salient regions. In this research, we shall adopt Irregularity-Based Saliency given in [26] and [27] as a measure to extract the salient regions of an image.

2.4 **Image retrieval and Saliency**

Due to their significance in information processing and the tremendous growth of multimedia use in different applications, image retrieval systems have been the focus of numerous research in the last few decades. Content-based image retrieval is the core of image retrieval systems since it focuses on the image content whether in matching it with other image content or in auto-labelling the image. As discussed earlier, image retrieval as a whole does not provide accurate results because of the domination of the background on the extracted features. Therefore, dividing the image into regions and using these regions in the recognition may produce better results, but due to the irregularity of the shape of the object, the effect of the background is still an issue that needs to be addressed. Shrivastava and Tyagi [28] is an example of region-based image retrieval systems. They used selected regions of interest in the matching process, where the image is divided into 3×3 or 5×5 regions and each region is given a 4-bit binary code. These binary codes are compared with the regions in another image. A similar approach is used by Wu et al. [29].

To improve the retrieval results, some researchers tried to link the image retrieval with saliency such as [30], where the authors attempted to develop an image retrieval system focused on the integration of target-driven with stimulus-driven visual saliency discrimination. In [31], the author tried to develop a new schema of CBIR based on SIFT salient points. The author divided the image into salient and non-salient regions based on the distribution of salient points and then used different colour descriptors to describe them. They firstly used SIFT to extract non-isolated salient points, and then built an index of these salient points and their neighbouring area according to their colour features. In [32], the authors used saliency in medical images using SVM and LBP features extracted from the salient regions. Alaei et al. used saliency in retrieving document images. The author proposed an appearance-based document image retrieval system using image saliency maps depending on human visual attention [33]. In [34], the authors propose a region merging strategy to solve the problem of background effect on the retrieval results. In their approach, boundary super-pixels are clustered to generate the initial saliency map, then adjacent regions are merged to get the final saliency maps and finally they use these maps in the image retrieval process. Many other studies in the same field have been published such as [35], [36], [37], [38] and [39].

Almost all studies share the same idea which is identifying the salient part of the image using various saliency detection approaches. Most of the retrieval algorithms used WANG dataset which is a subset of 1,000

images of the Corel stock photo database. In this dataset, the images are manually classified into 10 classes with 100 images each. The main limitation of the dataset is that the retrieval ratio is always high as the images in each class are very similar to each other and differ from other classes. Developing algorithms for special purposes and using specialised datasets such as medical applications and document identification is another limitation of the algorithms as the images are almost similar and identifying the saliency is an easy task.

3 Saliency Based Image Retrievals

In this work, the regions in the image are divided into two sets, salient regions set, and non-salient regions set. Based on that, we shall define the mapping ($\boldsymbol{\zeta}$) from the image space $\boldsymbol{I}$ into the regions space $\mathbf{R}$ such that:

$$\begin{aligned} \boldsymbol{\zeta}&: \boldsymbol{I} \rightarrow \boldsymbol{R^M} \\ \boldsymbol{R} &= \{\boldsymbol{r_i}|\ \boldsymbol{i} = \mathbf{0}, \mathbf{1}, \dots, \boldsymbol{M}\} \end{aligned} \tag{1}$$

where $\mathbf{R}$ is the set of regions that can be extracted from the image $\mathbf{I}$ and $\mathbf{M}$ is the total number of regions**.**

Furthermore, the sets $\mathbf{R_S}$ and $\mathbf{R_N}$ are defined as the sets of salient (important) and non-salient (unimportant) regions in the image respectively. The sets $\mathbf{R_S}$ and $\mathbf{R_N}$ are subsets of $\mathbf{R}$ and can be expressed as given in equation (2).

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{R_S} \cup \boldsymbol{R_N} \\ \boldsymbol{R_S} &= \{\boldsymbol{r_{s_i}}|\ \boldsymbol{i} = \mathbf{0}, \mathbf{1}, \dots, \boldsymbol{N}\} \\ \boldsymbol{R_N} &= \{\boldsymbol{r_{n_i}}|\ \boldsymbol{i} = \mathbf{0}, \mathbf{1}, \dots, \boldsymbol{M} - \boldsymbol{N}\} \end{aligned} \tag{2}$$

where $\boldsymbol{r_{s_i}}$ is the $\boldsymbol{i^{th}}$ salient regions, $\mathbf{N}$ is the number of salient regions in the image and $\boldsymbol{r_{n_i}}$ is the $\boldsymbol{i^{th}}$ non-salient or unimportant region. For each image, a set of salient regions is extracted and each member of this set ($\boldsymbol{r_{s_i}}$) is compared and matched with the regions of the images stored in the database.

Extracting the salient object from the salient region is another refinement that significantly improves the result. Therefore, the mapping ($\boldsymbol{\varpi}$) from the set of salient regions into the set of salient objects is defined as follows:

$$\boldsymbol{\varpi}: \boldsymbol{R_s} \rightarrow \boldsymbol{R_o} \tag{3}$$

where $\mathbf{R_o}$ is the set of all salient objects extracted from the salient regions. Usually, the mapping $\boldsymbol{\varpi}$ is an injective one-to-one mapping, which means there is only one salient object in each salient region.

The features set ($\boldsymbol{H}$) is divided into two parts, a set of salient regions' features (important) ($\boldsymbol{H_S}$) and a set of non-salient regions' features (unimportant) ($\boldsymbol{H_N}$) such that:

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{H_S} \cup \boldsymbol{H_N} \\ \boldsymbol{H} &= \{\boldsymbol{h_i}: \boldsymbol{i} = \mathbf{0}, \dots, \boldsymbol{M}\} \\ \boldsymbol{H_S} &= \{\boldsymbol{h_{s_i}}: \boldsymbol{i} = \mathbf{0}, \dots, \boldsymbol{N}\} \\ \boldsymbol{H_N} &= \{\boldsymbol{h_{n_i}}: \boldsymbol{i} = \mathbf{0}, \dots, \boldsymbol{M} - \boldsymbol{N}\} \end{aligned} \tag{4}$$

where $\boldsymbol{h_i}$ is the features extracted from any region in the image, $\boldsymbol{h_{s_i}}$ is the features extracted from the salient regions (important) and $\boldsymbol{h_{n_i}}$ is the features extracted from the non-salient or unimportant regions. The important information is usually contained in the object (salient region), while most of the unimportant information is in the background.

The mapping φ from the object space to the feature space is defined as follows:

$$\boldsymbol{\varphi}: \boldsymbol{R_o} \rightarrow \boldsymbol{H} \tag{5}$$

This mapping maps the object $\boldsymbol{r_{oi}} \in \boldsymbol{R_o}$ to the feature $\boldsymbol{h_i} \in \boldsymbol{H}$.

Finally, the mapping $\boldsymbol{\nu}$ is defined as the mapping that converts the feature $\boldsymbol{h_i} \in \boldsymbol{H}$ into a set of metrics represented by an ordered n-tuple of real numbers $\hat{\mathbb{r}}$ in $\mathbb{R}$.

$$\boldsymbol{\nu}: \boldsymbol{H} \rightarrow \mathbb{R}^{\boldsymbol{n}} \tag{6}$$

For instance, consider the following example:

$\boldsymbol{R} = \{\boldsymbol{r_1}, \boldsymbol{r_2}, \boldsymbol{r_3}, \boldsymbol{r_4}, \boldsymbol{r_5}, \boldsymbol{r_6}\}$

Assume that the set of salient regions is $\boldsymbol{R_S} = \{\boldsymbol{r_1}, \boldsymbol{r_3}\}$, then the set of non-salient regions is $\mathbf{R_N} = \{\mathbf{r_2}, \mathbf{r_4}, \mathbf{r_5}, \mathbf{r_6}\}$. In other words, $\boldsymbol{r_{s_1}} = \boldsymbol{r_1}$ and $\boldsymbol{r_{s_2}} = \boldsymbol{r_3}$ as the first and third regions have been identified as salient regions. Therefore, in this case, $\boldsymbol{M} = |\boldsymbol{R}| = \mathbf{6}$, $\boldsymbol{N} = |\boldsymbol{R_S}| = \mathbf{2}$ and $|\boldsymbol{R_N}| = \mathbf{4}$.

Next, the salient objects are extracted from the salient regions using the mapping $\boldsymbol{\varpi}$ which produces the set $\boldsymbol{R_0} = \{\boldsymbol{r_{o1}}, \boldsymbol{r_{o2}}\}$.

If we assume that the feature extraction mapping φ is a histogram extractor, then:

$\boldsymbol{\varphi}(\boldsymbol{r_{o1}}) = \boldsymbol{h_1}$ and $\boldsymbol{\varphi}(\boldsymbol{r_{o2}}) = \boldsymbol{h_2}$, where $\boldsymbol{h_1}$ is the histogram of the object $\boldsymbol{r_{o1}}$ and $\boldsymbol{h_2}$ is the histogram of the object $\boldsymbol{r_{o2}}$. Other regions' histograms are not needed to be extracted as they were marked as non-salient or unimportant.

Let us further assume that the mapping $\boldsymbol{\nu}$ will extract some descriptive measures from the histogram such as mean ($\boldsymbol{\mu}$), standard deviation ($\boldsymbol{\sigma}$), and skewness ($\boldsymbol{\varsigma}$). Therefore, the result of the mapping $\boldsymbol{\nu}$ will be as follows:

$\nu(h_1) = \widehat{\mathbb{r}_1} = <\mu_1, \sigma_1, \varsigma_1)$ and $\nu(h_2) = \widehat{\mathbb{r}_2} = <\mu_2, \sigma_2, \varsigma_2)$.

Figure 1 shows the above process graphically.

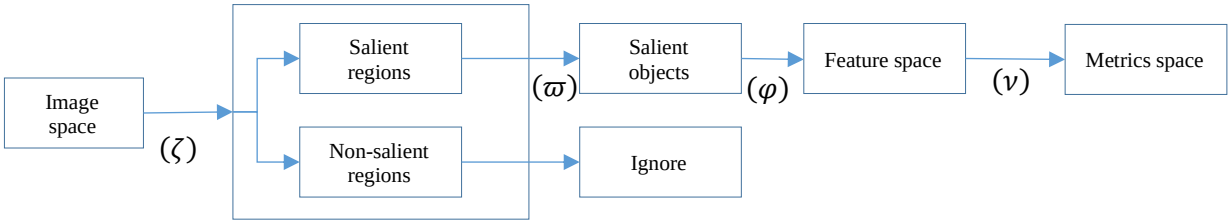


Figure 1: Graphical representation of the mappings mentioned in the above discussion

3.1 Information overlap

The overlap between the information content of the background (non-salient) and that of the object (salient) is a severe problem that affects the accuracy of the retrieval process as a whole. In other words, if we assume that the colour values in an image are represented by the set $\boldsymbol{\Omega}$, which represents the universal set, such that $\boldsymbol{\Omega} = \{\boldsymbol{x_i} : \mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{W} \times \boldsymbol{H}\}$, where W and H are the width and the height of the image respectively and $\mathbf{x_i}$ is the value of the pixel colour of the image considering that the image pixels are arranged to be a one-dimensional vector i.e. $\boldsymbol{x_k} = \boldsymbol{I}(\boldsymbol{i}, \boldsymbol{j})$ and $\boldsymbol{k} = (\boldsymbol{i} - \mathbf{1}) \times \boldsymbol{W} + \boldsymbol{j}$. The set of the pixels is divided into two components, object $\boldsymbol{\Omega_S}$ and background $\boldsymbol{\Omega_B}$ as shown in equation (7).

$$\begin{aligned} \boldsymbol{\Omega} &= \boldsymbol{\Omega_S} \cup \boldsymbol{\Omega_B} \\ \boldsymbol{\Omega_S} &= \{\boldsymbol{x_{Si}}\} \\ \boldsymbol{\Omega_B} &= \{\boldsymbol{x_{Bi}}\} \end{aligned} \tag{7}$$

If we consider that the object is isolated from the background and the borders are well-defined, then the cardinality of the universal set is equal to the sum of the cardinalities of the two components, i.e.:

$$\begin{aligned} |\boldsymbol{\Omega}| &= |\boldsymbol{\Omega_S}| + |\boldsymbol{\Omega_B}| - |\boldsymbol{\Omega_S} \cap \boldsymbol{\Omega_B}| \\ \boldsymbol{\Omega_S} \cap \boldsymbol{\Omega_B} &= \emptyset \\ |\boldsymbol{\Omega_S} \cap \boldsymbol{\Omega_B}| &= \mathbf{0} \\ |\boldsymbol{\Omega}| &= |\boldsymbol{\Omega_S}| + |\boldsymbol{\Omega_B}| \end{aligned} \tag{8}$$

Based on the histogram definition, the function H(x) can be defined as the number of occurrence of the variable $\boldsymbol{x}$ i.e. $\boldsymbol{H}(\boldsymbol{x_i}) = \boldsymbol{P}(\boldsymbol{x_i})$ with $\boldsymbol{P}(\boldsymbol{x_i})$ is the probability of occurrence of the bin $\boldsymbol{x_i}$. For instance, let us assume that the occurrence of the pixel in the background is the event **B** and the occurrence of it in the object is the event **O**, then the ideal case in which best outcome can be achieved is when the two events are

independent, i.e. $\boldsymbol{P(O|B) = P(O)}$. This can be achieved only if there is no intersection or overlap between the object and the background. Without well-defined borders and distinction between the object and the background, the background may affect the result because the features extracted are not for the object only, i.e. $\boldsymbol{\varphi(\mathbf{R_I} \cup \mathbf{R_U}) = \varphi(\mathbf{R_I}) + \varphi(\mathbf{R_U}) + \varphi(\mathbf{R_I} \cap \mathbf{R_U})}$.

To elaborate on this discussion consider Figure 2, which shows the two components of the information contents represented by the grey histogram of the image. It is clear from Figure 2 (a) that the background component has higher values for the bins; this will lead to giving the background component more significance than the object component in the extracted features or measures. Figure 2 (b) and (c) show the histogram of the object $\boldsymbol{\varphi(R_I)}$ and the background $\boldsymbol{\varphi(R_U)}$ respectively. In (b) and (c), the overlap was not considered i.e. $\boldsymbol{\varphi(R_I \cap R_U) = \emptyset}$ which is a common drawback of many regular thresholding processes. To overcome this problem the information contents should be extracted after segmentation. Figure 2 (d) shows the histogram of the object including the effect of the background on the grey level bins and (e) in the same figure shows the histogram of the object without this effect. Finally, (f) shows the difference between the two histograms i.e. $\boldsymbol{\varphi(R_I \cap R_U)}$.

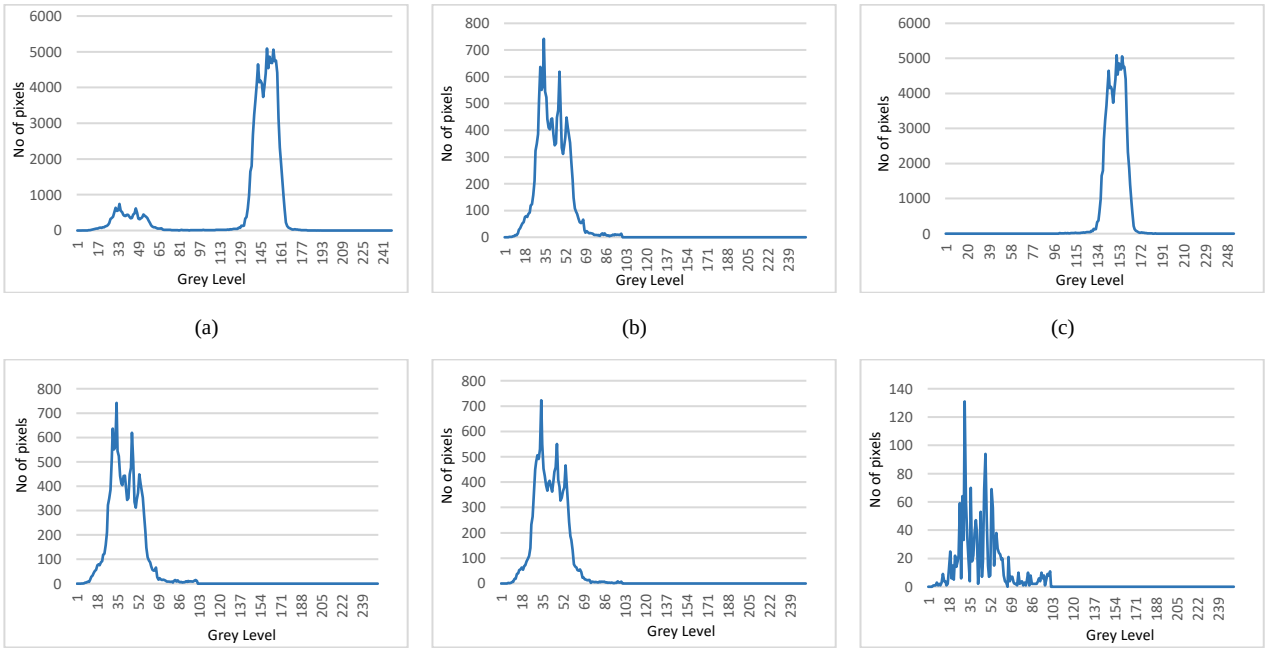


Figure 2: Information contents represented by image histogram, (a) whole image histogram, (b) object histogram, (c) background histogram, (d) object's histogram with background effect, (e) object's histogram alone, (f) difference between the histograms in (d) and (e)

3.2 Comparison Measure

Similarity or dissimilarity measures may be used to compare images based on their metrics values. Similar images should have a maximum similarity measure or a minimum dissimilarity measure. Let us assume that the sets of metrics of the query image and the i^{th} image in the database are $\mathbb{r}^q$ and $\mathbb{r}_i^d$ respectively and are defined as follows:

$$\begin{aligned} \mathbb{r}^q &= \{\boldsymbol{m_1^q}, \quad \boldsymbol{m_2^q}, \ldots, \boldsymbol{m_n^q}\} \\ \mathbb{r}_i^d &= \{\boldsymbol{m_{i1}^d}, \quad \boldsymbol{m_{i2}^d}, \ldots, \boldsymbol{m_{in}^d}\} \end{aligned} \tag{9}$$

where m_1^q is the metric number 1 of the query image, m_{i1}^d is the metric number 1 of the i^{th} image in the database and n is the number of metrics used in the matching process.

Furthermore, we shall define the similarity measure $S\left(\mathbb{r}^q, \mathbb{r}_i^d\right)$ that indicates how similar two images are. Images are said to be similar if the value of the similarity measure is high and larger than a prespecified threshold. Dissimilarity or distance measure $D\left(\mathbb{r}^q, \mathbb{r}_i^d\right)$ is another way to find the similarity between two images where the smaller the distance is, the more similar the images are. Due to its simplicity and accuracy, Minkowski distance (MD) is widely used in publications. The distance between the two sets of metrics given in equation (9) is calculated as shown below in equation (10).

$$\begin{aligned} \boldsymbol{D}\left(\mathbb{r}^{\boldsymbol{q}}, \mathbb{r}_{\boldsymbol{i}}^{\boldsymbol{d}}\right) &= \left|\left|\mathbb{r}^{\boldsymbol{q}} - \mathbb{r}_{\boldsymbol{i}}^{\boldsymbol{d}}\right|\right| \\ \boldsymbol{D}\left(\mathbb{r}^{\boldsymbol{q}}, \mathbb{r}_{\boldsymbol{i}}^{\boldsymbol{d}}\right) &= \sqrt[r]{\sum_{i=1}^{n} (\Delta \boldsymbol{m_i})^r} \\ \boldsymbol{D}\left(\mathbb{r}^{\boldsymbol{q}}, \mathbb{r}_{\boldsymbol{i}}^{\boldsymbol{d}}\right) &= \sqrt[r]{(\Delta \boldsymbol{m_1})^r + (\Delta \boldsymbol{m_2})^r + \cdots (\Delta \boldsymbol{m_n})^r} \\ \boldsymbol{D}\left(\mathbb{r}^{\boldsymbol{q}}, \mathbb{r}_{\boldsymbol{i}}^{\boldsymbol{d}}\right) &= \sqrt[r]{\left(\boldsymbol{m_1^q} - \boldsymbol{m_{i1}^d}\right)^r + \left(\boldsymbol{m_2^q} - \boldsymbol{m_{i2}^d}\right)^r + \cdots \left(\boldsymbol{m_n^q} - \boldsymbol{m_{in}^d}\right)^r} \end{aligned} \tag{10}$$

where $\boldsymbol{r}$ is an integer number and is commonly equal to two, the distance is then known as Euclidian distance.

3.3 System Flowchart

The basic diagram of the Saliency-Based Image Retrieval System (SBIR) is shown in Figure 3. In this system, the image at first goes through a saliency region extraction phase to produce a set of salient regions. For each salient region, the salient object is extracted for which a set of features is extracted using an appropriate feature extractor. Every object in the query image is then compared with all objects in the images stored in the database. Images with the best matching features are retrieved as relevant images.

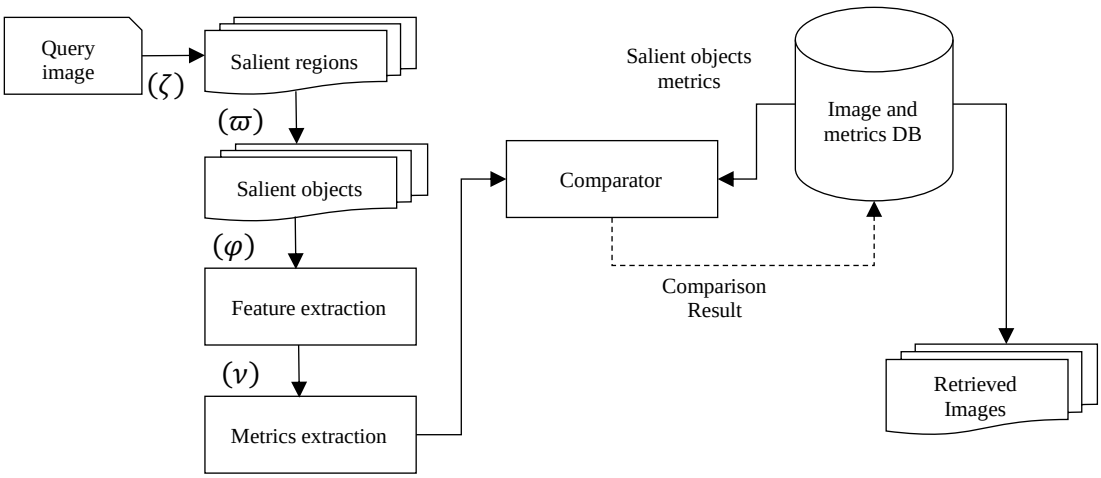


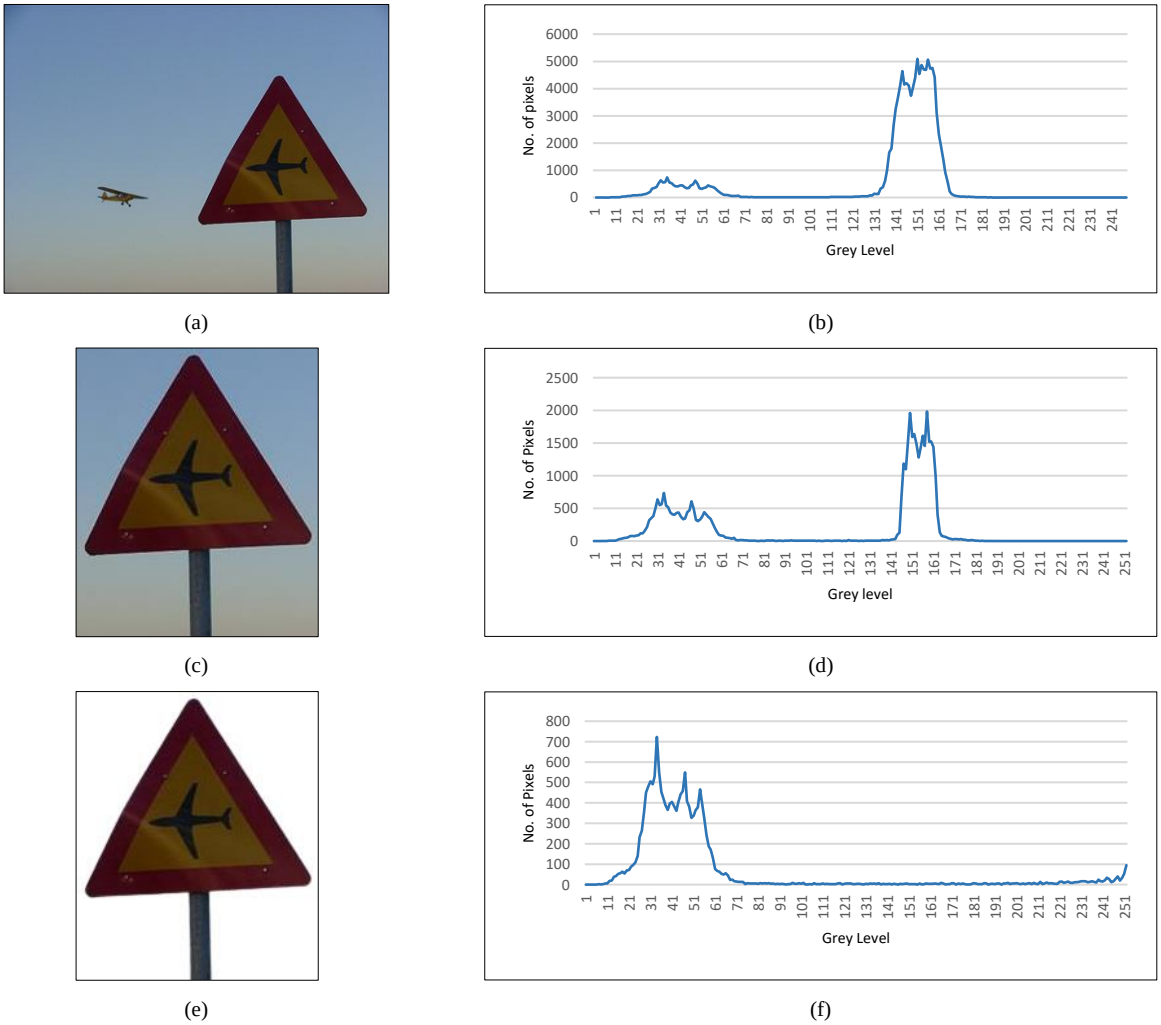
Figure 3: Proposed approach (SBIR) Diagram

4 Results and Discussion

4.1 The Effect of the Background

In this work, we considered three different forms of matching, which are: (i) Whole Image Identification (WII), (ii) Region-Based Identification (RBI), and (iii) Object-Based Identification (OBI). An illustration of the three forms of identification is given in Figure 4. In this figure, (a) and (b) show the image and the information content used in WII, in which the entire image is considered in the retrieval process. In this case,

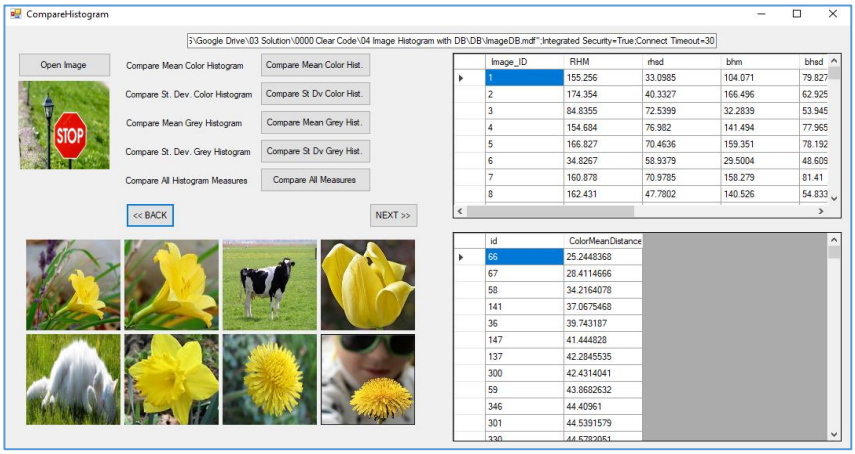
it is clear that the background effect is very high and the accuracy of the result is very low. The region and the information it contains used in RBI are shown in (c) and (d) where the salient region is extracted from the original image using a saliency identification process. In this case, the effect of the background is very much smaller than WII, but it still affects the accuracy and the result of the retrieving process. Finally, (e) and (f) show the salient object and its information content used in OBI matching, in which the effect of the background is almost zero.



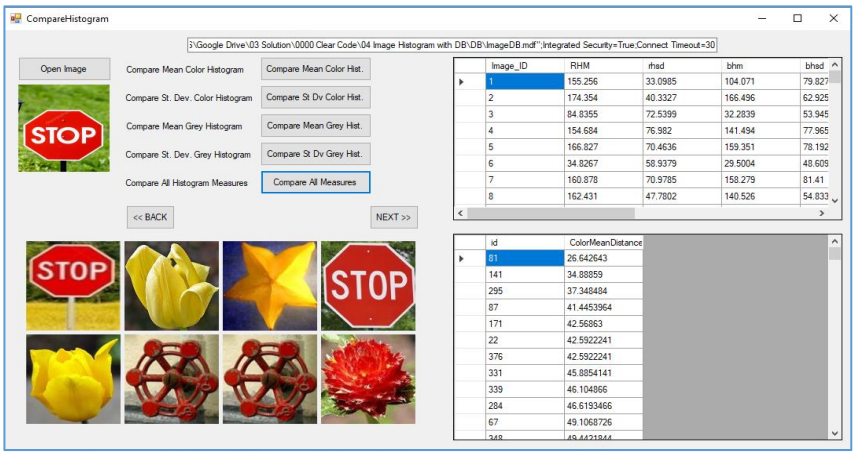
(a) (b) (c) (d) (e) (f)

Figure 4: Information contents in an image, (a) the original image, (b) the histogram of the original image, (c) the salient region, (d) the histogram of the salient region, (e) the salient object, (f) the histogram of the salient object

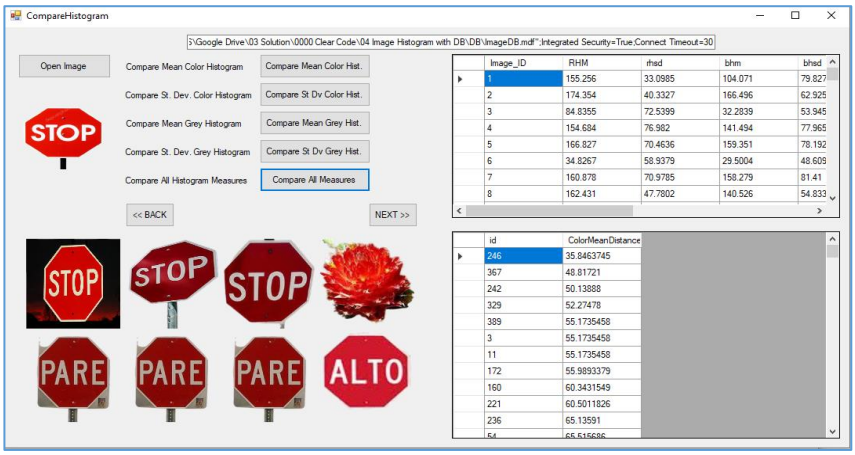
To study the efficiency of the three forms mentioned above, we undertook a comparison among them using measures extracted from the histogram for simplicity. Firstly, consider Figure 5 (a) which shows a search for the query image as a whole. From the retrieved image, it is clear that the background has more effect on the results than the object itself. Secondly, when using regions, as shown in Figure 5 (b), the retrieved images are more relevant to the query image, but still, the results are not satisfactory as there are a lot of irrelevant images retrieved. Finally, when using OBI, the retrieved images are quite similar to the query image and highly relevant as shown in Figure 5 (c).



(a)



(b)



(c)

Figure 5: Illustration of the three types of image content identification, (a) using WII, (b) using RBI, (c) using OBI

4.2 **Dataset and Implementation**

Many CBIR studies use WANG dataset for benchmarking and results validation. WANG dataset is a subset of 1,000 images of the Corel stock photo database, which are manually selected and classified into 10 classes with 100 images each [40]. Although WANG dataset is a standard dataset adopted by a lot of researchers, it

has an inherent drawback in the fact that it was organised to give a high retrieval rate as it contains many similar images. Therefore, a new dataset has been constructed considering this limitation and to fulfil the aim of our research. The new dataset is a collection of images from the following datasets:

1- WANG dataset [40].
2- MIRFLICKR dataset [41].
3- Linköping University, Computer vision lab dataset [42].
4- CIFAR dataset [43].
5- MSRA10 salient object dataset [44].

In addition to images offered by Flickr under Creative Commons copyright licenses.

The images in the dataset were selected carefully to satisfy the requirement of the discussion in this work such as containing similar objects with different backgrounds or similar backgrounds with different objects. The new dataset and the source code are available online at https://azawi.odoo.com/sbir/. The algorithm was implemented with C# language and SQL server for the database using Microsoft visual studio.

4.3 Evaluation of Image Retrieval Techniques

Several evaluation measures can be used to assess the results of the three forms of image retrieval mentioned above, one of which is the Precision and Recall Curves. Precision-Recall Curve provides a good visual representation of the result obtained by extracting the precision and recall values and drawing the relationship between them. The precision and recall can be calculated as given below in equation (11).

$$Precision = \frac{No, of\ Relevant\ Image\ Retrieved}{Total\ No. of\ Image\ Retrieved}$$
$$Recall = \frac{No. of\ Relevant\ Image\ Retrieved}{Total\ No. of\ relevant\ Image\ in\ database} \quad (11)$$

The precision versus recall graph is shown in Figure 6, from which one can notice that due to the effect of the background which dominates the features of the object itself, WII gave some irrelevant results. Similar results might be obtained from applying RBI but with less effect of the background. The best results are obtained by applying the OBI in which only the object will be compared.

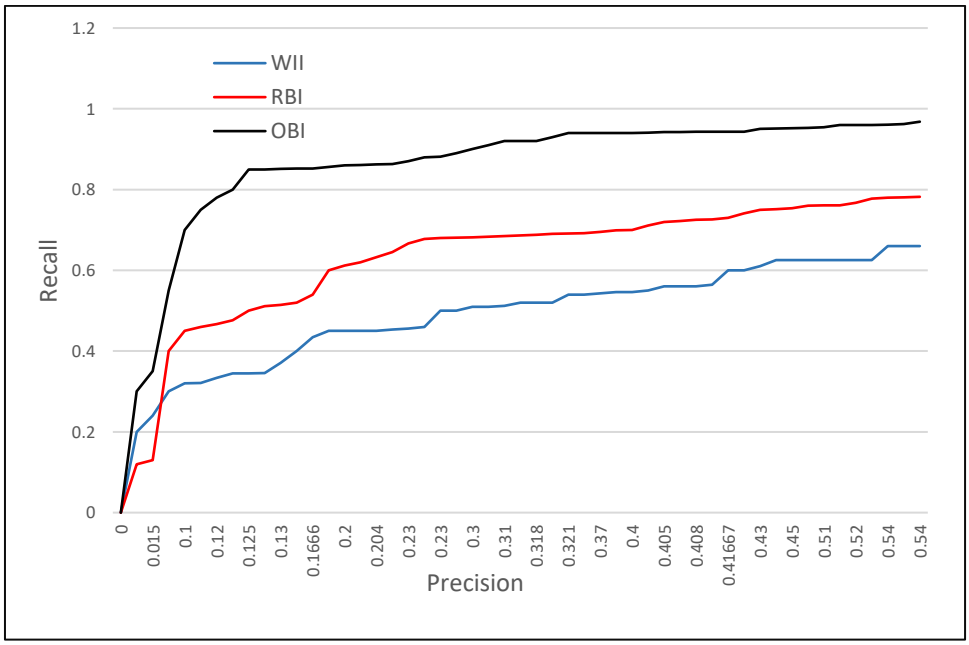


Figure 6: Precision-Recall curve of the three retrieval ways

In order to make the evaluation more reasonable, we shall present the weighted efficiency evaluation measure (WEEM). In this method, the order of the images retrieved is taken into account in the evaluation. A higher weight is assigned to the images retrieved first than those retrieved later.

The efficiency evaluation measure EEM can be calculated by dividing the number of relevant retrieved images by the number of similar images in the database, i.e.

$$\boldsymbol{EEM = \frac{N_{RR}}{N_T}} \tag{12}$$

where EEM is the Efficiency Evaluation Measure, N_{RR} is the number of relevant retrieved images, and N_T is the number of retrieved images in the ideal case in which all the images retrieved are correct. By considering the order of the retrieved images, Equation (12) is modified by multiplying a weight value by the number of retrieved images based on the retrieval order, i.e.

$$\begin{aligned} \boldsymbol{N_{RR}} &= \boldsymbol{\sum_{i=1}^{N} K(N-i+1)} \\ \boldsymbol{K} &= \begin{cases} \boldsymbol{0} & \boldsymbol{for\ irrelivant\ images} \\ \boldsymbol{1} & \boldsymbol{for\ relivant\ images} \end{cases} \\ \boldsymbol{WEEM} &= \boldsymbol{\frac{\sum_{i=1}^{N} K(N-i+1)}{\sum_{i=1}^{N}(N-i+1)}} \end{aligned} \tag{13}$$

By applying the measure given in Equation (13), the efficiency of the retrieval process is given in Figure 7. In the figure, it is shown that the efficiency of the OBI is far better than the results obtained from EII and RBI.

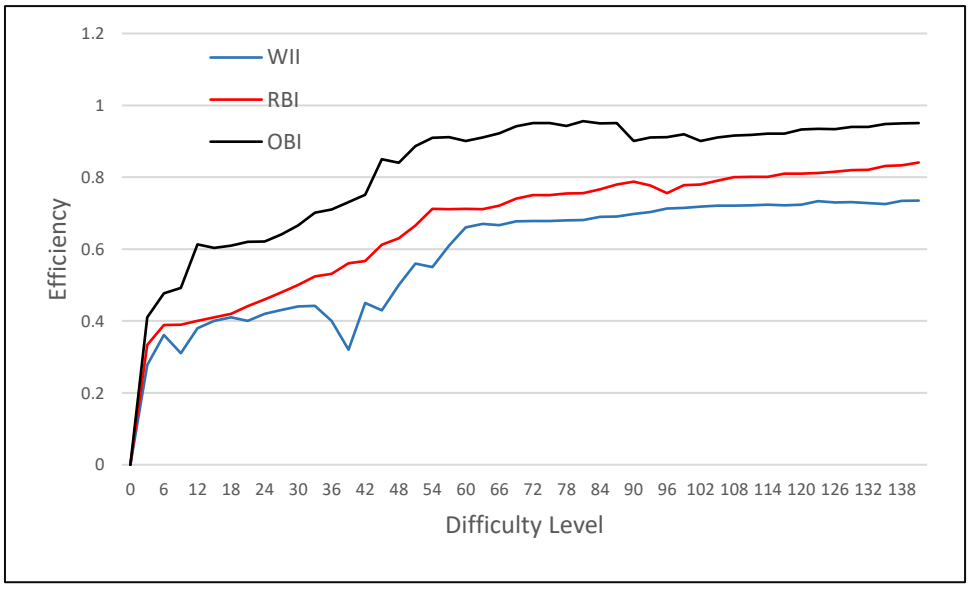


Figure 7: Retrieving Evaluation using WEEM

5 Conclusions

In this study, we presented a retrieval system that utilises the principles of human attention and saliency in the query image to retrieve similar images or more precisely, images that contain similar objects. The similarity between images is not based on the contents of the image as a whole but on the objects contained in the image, which ensures that the retrieval process has semantic properties. The experiments showed that SBIR worked better and gave better results than standard CBIR methods. The histogram has been selected as the image information descriptor in this paper to simplify the discussion, however, other features can be used as well. The results obtained showed that when applying OBI, the relevance of the retrieved images was drastically

improved due to removing the effect of the background in the extracted metrics. In addition, we have developed a new evaluation method in which the order of the retrieved image is considered in the evaluation process, and this is reasonable since it is important to have images that are similar to the query image retrieved first.

References

[1] S. Wang, "A Robust CBIR Approach Using Local Color Histograms," *Univ. Alberta / Dept. Comput. Sci. Alberta*, no. October, 2001, [Online]. Available: http://cis.temple.edu/~lakamper/courses/cis595_2004/papers/wang2001.pdf.

[2] Y. An, M. Riaz, and J. Park, "CBIR based on adaptive segmentation of HSV color space," *UKSim2010 - UKSim 12th Int. Conf. Comput. Model. Simul.*, pp. 248–251, 2010, doi: 10.1109/UKSIM.2010.53.

[3] W. T. Chen, W. C. Liu, and M. S. Chen, "Adaptive color feature extraction based on image color distributions," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2005–2016, 2010, doi: 10.1109/TIP.2010.2051753.

[4] B. Imran, "Content-Based Image Retrieval Based on Texture and Color Combinations Using Tamura Texture Features and Gabor Texture Methods," vol. 5, no. 1, pp. 23–27, 2019, doi: 10.11648/j.ajnna.20190501.14.

[5] D. Le Meur, Olivier; Callet, Patrick Le; Barba, Dominique; Thoreau, O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A Coherent Computational Approach to Model Bottom-up Visual Attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, p. 802, 2006.

[6] M. Al-Azawi, "Human Visual Attention and Machine Vision Computational Saliency," in *IEEE International Conference on Electrical, Electronic, Computer, Mechanical and Computing, EECCMC*, 2018.

[7] M. Al-azawi, "Saliency Identification as a Computational Model of Human Visual Attention," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 3, pp. 3348–3360, 2020, [Online]. Available: http://sersc.org/journals/index.php/IJAST/article/view/4780.

[8] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, 2001, doi: 10.1023/A:1012460413855.

[9] P. Kapsalas, K. Rapantzikos, A. Sofou, and Y. Avrithis, "Regions of interest for accurate object detection," in *2008 International Workshop on Content-Based Multimedia Indexing, CBMI 2008, Conference Proceedings*, 2008, pp. 147–154, doi: 10.1109/CBMI.2008.4564940.

[10] A. Toet, "Computational versus Psychophysical Bottom-Up Image Saliency: {A} Comparative Evaluation Study," *{IEEE} Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, 2011.

[11] E. Loupias, N. Sebe, S. Bres, and J.-M. Jolion, "Wavelet-based salient points for image retrieval," in *International Conference on Image Processing*, 2002, pp. 518–521 vol.2, doi: 10.1109/icip.2000.899469.

[12] Q. Tian, N. Sebe, M. S. Lew, E. Loupias, and T. S. Huang, "Content-based image retrieval using wavelet-based salient points," in *Storage and Retrieval for Media Databases 2001*, 2001, vol. 4315, pp. 425–436, doi: 10.1117/12.410953.

[13] H. Song, B. Li, and L. Zhang, "Color salient points detection using wavelet," *Proc. World Congr. Intell. Control Autom.*, vol. 2, pp. 10298–10301, 2006, doi: 10.1109/WCICA.2006.1714018.

[14] S.-H. Lin, Dong-Woei; Yang, "Wavelet-Based Salient Region Extraction," in *Advances in Multimedia Information Processing – PCM 2007. Vol. 4810*, Hong Kong: Springer, 2007, pp. 389–392.

[15] R. N. Arivazhagan, S.; Shebiah, S. Arivazhagan, and R. N. Shebiah, "Object Recognition using Wavelet Based Salient Points," *Open Signal Process. J.*, vol. 2, no. 2000, pp. 14–20, 2009.

[16] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 530–535, 1997, doi: 10.1109/34.589215.

[17] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry,"

Hum. Neurobiol., vol. 4, no. 4, pp. 219–227, 1985, doi: 10.1007/978-94-009-3833-5_5.

[18] M. B. Vinay and K. S. Rekha, "A model of saliency-based visual attention for rapid scene analysis," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 412–415, 2019.

[19] N. D. B. Bruce, D. P. Loach, and J. K. Tsotsos, "VISUAL CORRELATES OF FIXATION SELECTION : A LOOK AT THE SPATIAL FREQUENCY DOMAIN Department of Computer Science and Centre for Vision Research 4700 Keele Street , Toronto , Ontario , Canada M3J 1P3," *Spectrum*, pp. 289–292, 2007.

[20] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007, doi: 10.1109/CVPR.2007.383267.

[21] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, 2013, doi: 10.1109/TPAMI.2012.147.

[22] S. Liu and J. Hu, "Visual saliency based on frequency domain analysis and spatial information," *Multimed. Tools Appl.*, vol. 75, no. 23, pp. 16699–16711, 2016, doi: 10.1007/s11042-016-3903-3.

[23] R. Achantay, S. Hemamiz, F. Estraday, and S. Süsstrunky, "Frequency-tuned salient region detection," *2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work. 2009*, vol. 2009 IEEE, no. Ic, pp. 1597–1604, 2009, doi: 10.1109/CVPRW.2009.5206596.

[24] L. Zhou, Bolei ; Hou, Xiaodi ; Zhang, "A phase discrepancy analysis of object motion," in *Proceedings of the 10th Asian conference on Computer vision ACCV'10*, 2010.

[25] C.-W. Fang, Yuming; Lin, Weisi ; Lee, Bu-Sung ; Lau, Chiew-Tong ; Chen, Zhenzhong ; Lin *et al.*, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Trans. Multimed.*, vol. 14, no. 1, pp. 187–198, 2012, doi: 10.1109/TMM.2011.2169775.

[26] M. Al-Azawi, Y. Yang, and H. Istance, "Irregularity-based saliency identification and evaluation," *IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2013*, 2013, doi: 10.1109/ICCIC.2013.6724128.

[27] M. Al-Azawi, Y. Yang, and H. Istance, "Irregularity-based image regions saliency identification and evaluation," *Multimed. Tools Appl.*, vol. 75, no. 1, pp. 25–48, 2016, doi: 10.1007/s11042-014-2248-z.

[28] N. Shrivastava and V. Tyagi, "Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching," *Inf. Sci. (Ny).*, vol. 259, pp. 212–224, 2014, doi: 10.1016/j.ins.2013.08.043.

[29] J. Wu, "A Novel Image Retrieval Method with Saliency Feature Vector," *Int. J. Performability Eng.*, vol. 14, no. 2, pp. 223–231, 2018, doi: 10.23940/ijpe.18.02.p4.223231.

[30] B. Wang, X. Zhang, M. Wang, and P. Zhao, "Saliency distinguishing and applications to semantics extraction and retrieval of natural image," *2010 Int. Conf. Mach. Learn. Cybern. ICMLC 2010*, vol. 2, no. July, pp. 802–807, 2010, doi: 10.1109/ICMLC.2010.5580581.

[31] C. S. Wang, G. Q. Han, Y. Wo, and L. M. Liu, "An approach of content-based image retrieval based on image salient region," *2010 6th Int. Conf. Wirel. Commun. Netw. Mob. Comput. WiCOM 2010*, vol. 2, no. 1, pp. 6–10, 2010, doi: 10.1109/WICOM.2010.5601025.

[32] G. Cao, "Salient feature extraction for image retrieval," 2010.

[33] F. Alaei, A. Alaei, U. Pal, and M. Blumenstein, "Document Image Retrieval Based on Visual Saliency Maps," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Sep. 2019, pp. 7–12, doi: 10.1109/ICDARW.2019.30057.

[34] W. Wei, H. Chen, Y. Wen, and Y. Wang, "Saliency Object Detection Based on Regions Merging and Its Application in Image Retrieval," *J. Phys. Conf. Ser.*, vol. 1302, no. 2, p. 022005, Aug. 2019, doi: 10.1088/1742-6596/1302/2/022005.

[35] Y. H. Jacky Lam and S. Yildirim Yayilgan, "Saliency-Based Image Object Indexing and Retrieval," vol. 1, no. June, Springer International Publishing, 2018, pp. 269–277.

[36] A. G. Bors and A. Papushoy, "Image Retrieval Based on Query by Saliency Content," in *Visual Content Indexing and Retrieval with Psycho-Visual Models*, Cham: Springer International Publishing, 2017, pp. 171–209.

[37] Q. Zheng, S. Wei, J. Li, F. Yang, and Y. Zhao, "Uncovering the effect of visual saliency on image retrieval," *Commun. Comput. Inf. Sci.*, vol. 772, pp. 170–179, 2017, doi: 10.1007/978-981-10-7302-1_15.

[38] Y. Gao, M. Shi, D. Tao, and C. Xu, "Database saliency for fast image retrieval," *IEEE Trans. Multimed.*, vol. 17, no. 3, pp. 359–369, 2015, doi: 10.1109/TMM.2015.2389616.

[39] A. Papushoy and A. G. Bors, "Image retrieval based on query by saliency content," *Digit. Signal Process. A Rev. J.*, vol. 36, no. C, pp. 156–173, 2015, doi: 10.1016/j.dsp.2014.09.005.

[40] J. Z. Wang, "James Z. Wang Group." [Online]. Available: http://wang.ist.psu.edu/docs/home.shtml. [Accessed: 12-Oct-2019].

[41] M. Huiskes, B. Thomee, and M. Lew, "The MIRFLICKR Retrieval Evaluation." [Online]. Available: https://press.liacs.nl/mirflickr/. [Accessed: 01-May-2019].

[42] F. Larsson, "Linköping University, Computer vision lab," 2019. [Online]. Available: https://www.cvl.isy.liu.se/en/research/datasets/index.html. [Accessed: 01-Oct-2019].

[43] A. Krizhevsky, V. Nair, and G. Hinton, "CIFAR Dataset," *Toronto University - CS*, 2013. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html. [Accessed: 01-Jun-2019].

[44] "MSRA10K Salient Object Database," 2014. [Online]. Available: https://mmcheng.net/msra10k/. [Accessed: 01-Oct-2019].