Image retrieval using a scale-invariant feature transform bag-of-features model with salient object detection

Yih-Chearng Shiue
Department of Information Management,
National Central University,
No.300, Zhongda Rd., Zhongli Dist.,
Taoyuan City, Taiwan
Email: ycs@mgt.ncu.edu.tw

Sheng-Hung Lo*
Department of Information Management,
National Central University,
No.300, Zhongda Rd., Zhongli Dist.,
Taoyuan City, Taiwan
Email: shenghung.lo@gmail.com
*Corresponding author

Yi-Cheng Tian
Center for General Education,
Hsin Sheng College of Medical Care and Management,
No.418, Sec. Gaoping, Zhongfeng Rd., Longtan Dist.,
Taoyuan City, Taiwan
Email: tyc@hsc.edu.tw

Cheng-Wei Lin
Department of Information Management,
National Central University,
No.300, Zhongda Rd., Zhongli Dist.,
Taoyuan City, Taiwan
Email: domo.lin@gmail.com

Abstract: How to effectively retrieve digital images is a focus of image retrieval research. Developed in the 1990s, content-based image retrieval (CBIR) systems are used to extract low-level visual features. However, semantic gaps exist between these features and high-level semantic concepts. This study proposes an image retrieval solution based on a bag-of-features (BoF) model integrated with scale-invariant feature transform (SIFT) and salient object detection. An image search system based on this image retrieval solution, which used object images as the query image, was subsequently constructed. Overall, the results verify the feasibility of the object-based image retrieval solution. Finally, the enhanced image search method and precision
enabled constructing an image search system. The system is expected to improve through the search pattern, as well as improve the accuracy of images search, images search system to make a real attempt to solve the huge amount of data and images search difficult problems arising.

**Keywords:** image retrieval; content-based image retrieval; scale-invariant feature transform; bag-of-features model; k-means clustering.


**Biographical notes:** Yih-Chearng Shiue is a Professor in the Department of Information Management at National Central University, Taiwan (R.O.C). He holds a PhD in Management from the Tamkang University. His research interests include knowledge management and strategy management. His research has appeared in *Computers in Human Behaviour, European Journal of Operation Research, Journal of Educational Media & Library Sciences, Social Behaviour and Personality, Journal of Testing and Evaluation, Energy and Buildings*, and others.

Sheng-Hung Lo is a Lecturer in the General Education Centre and Director for Information Services Centre at the Hsin Sheng College of Medical Care and Management, Taiwan (R.O.C). He is also a doctoral candidate in the Department of Information Management at the National Central University, Taiwan. His research interests include ERP system, factory automation, strategic management and graphical analysis.

Yi-Cheng Tian is a Lecturer in the General Education Centre at the Hsin Sheng College of Medical Care and Management, Taiwan (R.O.C). He holds a PhD in the Department of Applied Mathematics at Chung Yuan Christian University, Chung-Li, Taiwan. His research interests include cluster analysis, fuzzy clustering and data analysis.

Cheng-Wei Lin is currently a Master student in the Department of Information Management at the National Central University, Taiwan. His research interests include cluster analysis and graphical analysis.

This paper is a revised and expanded version of a paper entitled ‘The study of salient object and BoF with SIFT for image retrieval’, presented at 5th International Conference on Frontier Computing in Tokyo, Japan, 13–15 July 2016.

---

1 Introduction

1.1 Background

Digital cameras have surpassed film cameras as the mainstream photographic instrument. Mobile phones with photographic functionality are also widely used to record images and enable delivering even higher image quality than some digital cameras do. Moreover, increasingly more digital images are uploaded online with the progressively expanded
capacity of storage media and the growing ubiquity of the internet. Thus, how to effectively retrieve digital images has become a major topic in image retrieval research.

Digital images are not only stored online; most of them are also copied to multimedia storage devices such as hard drives, USB flash drives, and compact discs. Images stored on such devices can be retrieved according to their corresponding tags or keywords, which are established by users prior to retrieval. Digital images are typically named with a serial number that is generated by digital photographic instruments, and users seldom rename them according to their content. In addition, in contrast to images stored online, which typically contain some annotation feature that enables keyword searches, images stored in multimedia storage devices require categorisation or tagging to ensure that they are properly arranged. However, accessing such images through the built-in search functionality of a computer is not always successful. Thus, this study proposes a search system that facilitates image retrieval.

1.2 Motivation

Typical online image search engines use keywords (e.g., file name, title, and content description) as queries. Different types of image-related textual information stored in multimedia databases can also be used as keywords in retrieving relevant images. However, images that contain complex concepts comprising ambiguous or subjective textual descriptions provide no accurate, objective, or concrete information about the overall conception or context, resulting in either inconsistency between image retrieval and user expectation or poor search performance (Veltkamp et al., 2001).

Introduced in the 1990s, content-based image retrieval (CBIR) differs from keyword-based image search solutions in that CBIR bases its search on the features that can be extracted from an image (e.g., colour, texture, and shape). Such features are categorised as low-level features that can be transformed into numerical vectors. The feature vectors can subsequently be used to evaluate the similarity between images to retrieve the most visually similar images. At present, image searches are performed largely through such online search engines as Google Image, Microsoft Bing, TinEye, and GazoPa.

Most CBIR systems perform retrieval on the basis of an input similar image/photos or user-drawn sketch as the query image. Users can input similar images/photos as the query image to conduct an online image search, through which similar or precise images are retrieved using search algorithms. This query-by-example method uses an entire image for the query, which differs from the object-based search solution proposed in this study. Querying with user-drawn sketches requires users to have sufficient drawing capability; however, this is often ineffective because not every user is skilled at drawing. Accordingly, this study proposes an image search system to address the limitations of CBIR systems.

The proposed system performed image searches using only object images as the query image. This enables users to specify any object in an image as the query image without the inconvenience of using entire images or user-drawn sketches for the query.

1.3 Purpose

This study was aimed at developing an image search system that facilitates retrieval among disarrayed images stored on personal computers. Rather than using an entire
image, the proposed system detects and then uses the salient object(s) of an image for the query.

1 Search method

Conventional image search systems use the ‘search-by-image’ solution, basing their searches on sketches or similar images provided by the user. The proposed image search system provides various object images for multi object query, enabling the user to select one or multiple object images as the query image to retrieve the target image.

2 Image retrieval precision

To improve its image retrieval precision, the bag-of-features (BoF) model based on the scale-invariant feature transform (SIFT) algorithm was integrated with salient object detection.

3 Construction of the image search system

Rectangular salient images were derived through salient object detection. The SIFT algorithm was implemented to extract features from these images and the BoF model was adopted to establish BoF vectors of the images.

The system adopts Euclidean distance to estimate the similarity between the object images and target images stored in a database; the target images and their corresponding names are then presented in descending order according to their similarity to the object images.

2 Literature review

2.1 Content-based image retrieval

CBIR is an image search solution based on visual feature extraction and searches for images stored in a multimedia database that correspond with the contents of an image specified by the user. Figure 1 presents the image-matching process of a CBIR system.

Figure 1  Image-matching process of CBIR systems (see online version for colours)
In a CBIR system, image contents are described by common low-level visual features such as colour, texture, and shape (Ojala et al., 2002). These features are extracted from images to estimate the similarity between the query image and images stored in a database, through which the most similar images are retrieved. However, the features do not accurately describe the semantic concepts of images, as indicated by the summary of strengths and weaknesses of the features in Table 1. In addition to low-level visual features, mid-level features have received increasing attention in recent years. Such features can be detected using SIFT, which is a computerised algorithm that can be implemented to detect and describe local features in images. In addition, it is translation-, scale-, and rotation-invariant (Lowe, 1999). The SIFT algorithm has been experimentally verified to be the most robust local feature descriptor in the presence of geometric transformation, as compared with other descriptors of the same type (Mikolajczyk and Schmid, 2005).

Table 1 Strengths and weaknesses of low-level visual features

<table>
<thead>
<tr>
<th></th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colour</td>
<td>High computational efficiency; invariant to image scaling and rotation</td>
<td>No description of image content or spatial distribution of colours</td>
</tr>
<tr>
<td>Texture</td>
<td>Describing spatial variations in pixel intensity and the surface characteristics of an object</td>
<td>Lack of texture segmentation methods that meet human perception</td>
</tr>
<tr>
<td>Shape</td>
<td>Relatively consistent with the intuitive feeling</td>
<td>Lack of robust mathematical foundations to address the deformation of the target image</td>
</tr>
</tbody>
</table>

Source: Yuan et al. (2011)

2.1.1 Bag-of-features model

The BoF model is based on the bag-of-words (BoW) model, a text-mining method that assumes that all words in a document are independent of each other, thereby disregarding the order of words, which are instead grouped into vocabularies, and lexical vectors are used as the representation of the document. Similarly, the BoF model treats each visual representation as an orderless collection of local features, and its effectiveness in image processing has been evidenced by previous studies (Zhang et al., 2011). This model represents images comprises two concepts: local features and codebooks (Yuan et al., 2011).

1 Local features

In the BoF model, feature extraction is performed on the basis of local feature descriptors such as SIFT, which has been extensively studied because of its reliability, efficiency, and simplicity in feature extraction (Lowe, 2004). It is also one of the most common descriptors used in the BoF model.

2 Codebooks

A codebook is the manner in which an image is represented as a collection of local features (Zhang et al., 2011). To generate a cookbook in the BoF model, a clustering algorithm (mostly k-means clustering) is used to cluster extracted local features, with
the centroid of the resulting cluster representing a visual word in the model. The occurrence counts of all visual words of an image are recorded in the codebook and quantified as feature vectors (Yuan et al., 2011).

2.1.2 Scale-invariant feature transform

The SIFT algorithm detects and describes the local features of an image. To identify scale-space extrema and extract extremum features that are invariant to location, scaling, and rotation, the key-points of the image are calculated through the following four stages:

1. scale-space extremum detection,
2. key-point localisation,
3. orientation assignment, and
4. key-point descriptor generation.

2.1.2.1 Scale-space extrema detection

- Key-point detection

The input image is convoluted with Gaussian filters at different scales. The difference between adjacent Gaussian images is estimated to effectively detect stable key-points in the scale space (Lowe, 1999), as expressed in equation (1) and equation (2):

\[ L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \]  
\[ G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2 + y^2)}{2\sigma^2}} \]

where \((x, y)\) is the pixel location of the image; \(\sigma\) is the standard deviation of a normal distribution, or the scale-space factor; \(G(x, y, \sigma)\) is the scale-variant Gaussian filter; \(*\) is a convolution operation; \(I(x, y)\) is the input image; and the convolution of \(G(x, y, \sigma)\) and \(I(x, y)\) yields \(L(x, y, \sigma)\), which is a blurred image (also referred to as a Gaussian image). The generation of the difference-of-Gaussian (DoG) images is illustrated by Figure 2, in which the original image at different scales is blurred with Gaussians to generate a set of scale-space images on the left, and the adjacent Gaussian images of similar scale spaces are subtracted to generate the DoG images on the right (Lowe, 2004).

- Local extrema detection

The process demonstrated in Figure 2 is continuously repeated to generate the DoG images, the amount of which decreases in multiples of 2 until potential key-points are identified. To determine whether the identified key-points are extrema, all feature points near the key-points are tested. The process of local extrema detection is illustrated in Figure 3, where the ‘X’ symbol denotes a potential key-point. If X is determined to be a maxima or minima relative to its 26 neighbours in the \(3 \times 3\) regions at the current and adjacent scales, then it is considered an extremum.
Figure 2  Generation of the dog images (see online version for colours)


Figure 3  Local extrema detection (see online version for colours)

2.1.2.2 Key-point localisation

Once all the extrema are identified, key-point selection is performed to filter inappropriate extrema. For example, feature values with low contrast are sensitive to noise and are therefore rejected using Taylor’s series formula (Brown and Lowe, 2002), which is expressed in equation (3) and equation (4):

\[ D(x) = D + \frac{\partial D}{\partial x} x + \frac{1}{2} x^2 \frac{\partial^2 D}{\partial x^2} x \]  

where \( D(x) \) is Taylor’s series formula; \( D \) is the DoG image; and \( x = x, y, \sigma \) is the offset from the key-point. The location of the extremum, \( x \), is determined by taking the derivative of \( D(x) \) relative to \( x \), with the function set to 0, as expressed in equation (4):

\[ \hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2} \frac{\partial D}{\partial x} \]  

2.1.2.3 Orientation assignment

For the identified key-points to achieve rotation-invariance, the location, scale, and orientation of the image feature points in a region around the key-point locations require estimation. The gradients of the other feature points around the key-point locations are computed to determine the gradient magnitude and orientation of these points, as expressed in equation (5) and equation (6):

\[ m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \]  

\[ \theta(x, y) = \tan^{-1}\left(\frac{(L(x, y+1) - L(x, y-1))}{(L(x+1, y) - L(x-1, y))}\right) \]  

where \( L(x, y) \) is the Gaussian image at each scale; \( m(x, y) \) is the gradient magnitude around the key-point location; and \( \theta(x, y) \) is the gradient orientation around the key-point location. The gradient orientations of the feature points in the region around the key-point are estimated using an orientation histogram. Peaks in the orientation histogram denote the dominant directions of the local gradients (Lowe, 2004).

2.1.2.4 Key-point descriptor generation

A key-point descriptor is generated by estimating the gradient magnitude and orientation at the image feature points in the region around the key-point location, as shown in Figure 4. These points are weighted by a Gaussian window, indicated by the overlaid circle on the left of the figure. They are then organised into a grid comprising 4 \times 4 subregions. Each subregion contains eight directions, producing 4 \times 4 \times 8 = 128 feature vectors, or 128-dimensional SIFT vectors.

Figure 4 displays a 2 \times 2 descriptor array estimated from an 8 \times 8 image array. However, this study adopted a 4 \times 4 descriptor array computed from a 16 \times 16 image array because the descriptor array yields the optimal results, as evidenced by Lowe (2004).
2.2 Visual attention

The human brain and visual system devote a preponderance of attention to some parts of images (Liu et al., 2011). Visual attention has been extensively studied in the fields of physiology, psychology, neuroscience, and computer vision. Applications for visual attention include image/video compression, automatic image cropping, advertisement design, and image collection browsing. Several studies (Navalpakam and Itti, 2006; Rutishauser et al., 2004) have demonstrated the effectiveness of visual attention in object recognition, tracking, and detection. Visual attention comprises two processes, namely bottom-up and top-down attention.

2.2.1 Salient object detection

Immediately before the human eye engages in image recognition, it identifies the salient regions of an image with remarkable speed and precision. Thus, approximating human object recognition capability is a major topic in machine vision and has received extensive attention in this field. This study adopted the method for salient object detection proposed by Liu et al. (2011).

Liu et al. (2011) revealed a supervised approach to salient object detection, proposing a set of novel features to describe salient objects, and optimally combining three features through conditional random field learning. These features are multiscale contrast (MSC), centre-surround histogram (CSH), and colour spatial distribution (CSD). After the feature map is normalised, these three features are optimally combined through conditional random field learning to generate a binary label map that separates a salient object from the background. To identify regions of interest, the smallest rectangle containing at least 95% salient pixels should be identified in the binary label map.

2.2.1.1 Multiscale contrast

Studies on visual attention have concluded that contrast is the most common parameter in detecting local features because contrast operators model the human visual receptive fields (Itti et al., 1998; Ma and Zhang, 2003; Pirnog et al., 2009). Thus, when the size and location of a salient object are unknown, contrast is typically estimated at multiple scales.
In this study, the MSC feature $f, x, I$ is conveniently defined as a linear combination of contrasts in the Gaussian image pyramid, as expressed in equation (7):

$$f(x, I) = \sum_{l=1}^{L} \sum_{x \in N(x)} \left\| I^l(x) - I^l(x') \right\|^2$$

where $I^l$ is the $l$th-level image in the pyramid; $L$ is the total number of pyramid levels ($L = 6$); and $N(x)$ is a $9 \times 9$ window. The feature map $f, I$ is normalised to a fixed range $[0,1]$. The results are shown in Figure 5.

**Figure 5** Multi-scale contrast; (a) original image (b) various scales feature map comparison (c) multi-scale comparison feature map (see online version for colours)

### 2.2.1.2 Centre-surround histogram

As Figure 5 shows, a salient object can be distinguished from its environs. Thus, this study proposes a regional salient feature that assumes that the salient object is surrounded by a rectangular area $R$, and a surrounding contour $R_s$ is constructed with the same area of $R$ (Figure 6). To distinguish the salient object in the rectangular area from its surroundings, several visual cues, such as intensity, colour, and texture, are subsequently used to measure the distance between $R$ and $R_s$. The chi-square distance between the histograms of RGB colour is adopted, as expressed by equation (8):

$$\chi^2 (R, R_s) = \frac{1}{2} \sum \left( \frac{R^i - R_s^i}{R^i + R_s^i} \right)^2$$

Figure 6 illustrates that the salient object (the girl) is most distinct, which is determined using the chi-square distance. To handle the varying aspect ratios of the object, five templates with different aspect ratios [0.5, 0.75, 1.0, 1.5, 2.0] are used. The most distinct rectangular area $R'(x)$ centred at each pixel $x$ is identified by varying the size and aspect ratio, as expressed in equation (9):
The size range of the rectangular area $R(x)$ is specified at $[0.1, 0.7] \times \min(w, h)$, where $w$ and $h$ are image width and height, respectively. The CSH feature $f_h(x, I)$ is defined as the sum of the spatially weighted distances, as expressed in equation (10):

$$f_h(x, I) = \sum_{[x' \in (x')]} w(x') \chi^2 \left( R^*(x'), R^*(x') \right)$$

where $R^*x'$ is the rectangular area centred at $x'$, which contains pixel $x$. The feature $f_h, I$ is normalised to the range of $[0, 1]$. Figure 6(b) shows three CSH feature maps.

**Figure 6** CSH (a) Chi-square distances with three different locations and sizes; (b) input images (top row) and CSH feature maps (bottom row) (see online version for colours)

Source:  Liu et al. (2011)
2.2.1.3 Colour spatial distribution

A CSH is a regional feature. In addition to local and regional features, a global feature may be related to salient object detection. As observed from Figure 5, the wider a colour is distributed in the image is, the less likely it is that the salient object contains this colour. Thus, the saliency of the object can be described according to the global spatial distribution of a specific colour.

The simplest approach to describing spatial distribution of a specific colour is estimating the spatial variance of that colour. First, all the colours in an image are represented by the Gaussian mixture model \( \{ w_c, \mu_c, \Sigma_c \}_{c=1}^C \), where \( \{ w_c, \mu_c, \Sigma_c \} \) respectively refers to the weight, mean colour, and the covariance matrix of the \( c \)th component. Equation (11) yields the probability of each pixel being assigned to colour component \( c \).

Next, the horizontal variance \( V_h(c) \) of the spatial position for each colour component \( c \) is determined using equation (12):

\[
p(c|I_x) = \frac{w_c}{\sum_c w_c} \frac{N(I_x|\mu_c, \Sigma_c)}{\sum_c w_c \frac{N(I_x|\mu_c, \Sigma_c)}}
\]

\[
V_h(c) = \frac{1}{|X_c|} \sum_{x \in X_c} p(c|I_x) \left( x_h - M_h(c) \right)^2
\]

\[
M_h(c) = \frac{1}{|X_c|} \sum_{x \in X_c} p(c|I_x) \cdot x_h
\]

where \( x_h \) is the x-coordinate of the pixel \( x \), and \( |X_c| = \sum_x p(c|I_x) \). The spatial variance of component \( c \) is expressed in equation (14). The term \( \{V(c)\}_c \) is normalised to the range of [0, 1]. The CSD feature is expressed in equation (15):

\[
V(c) = V_h(c) + V_v(c)
\]

\[
f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c))
\]

where \( f_s(\cdot, I) \) is normalised to the range of [0, 1].

3 System development

3.1 Research framework

The framework of the proposed image search system integrates the BoF model based on SIFT with salient object detection. Based on this framework (Figure 7), the image search system performs searches in three stages:

1. salient object detection,
2. BoF-based image transformation, and
3. similarity measurement.
3.1.1 Salient object detection

At this stage, the features of MSC, CSH, and CSD are defined as salient objects in images. During the experiment, all the images stored in the image database were resized to a minimum of 200 pixels in length or width (those smaller than 200 × 200 pixels retained their original size). However, the resized images contained distortions and slightly affected the precision of image search. After the salient objects were detected, the image backgrounds were filtered to retain the salient object images. The results of performing salient object detection on three images are shown by Figure 8.

Figure 8 Original vs. cropped salient object images; (a) input images (b) original salient object images (c) rectangular salient object images (cropped) (see online version for colours)
Figure 8  Original vs. cropped salient object images; (a) input images (b) original salient object images (c) rectangular salient object images (cropped) (continued) (see online version for colours)

3.1.2 Image transformation based on bag of features

The BoF model transforms the query image into a set of visual words. The transformation comprises three stages: feature extraction and description, visual vocabulary creation, and feature quantisation (Lv et al., 2013).

3.1.2.1 Feature extraction and description

CBIR systems categorised visual features in an image into global and local features. Global features are extracted from an entire image (e.g., colour histograms, texture distribution, and shape features). However, such features are sensitive to changes in the locations and scales of image objects (Niblack et al., 1993). Most studies related to image retrieval have explored local features (Yuan et al., 2011; Yang and Kurita, 2013) such as SIFT.

The SIFT features can be simplified into key-point detectors and key-point descriptors. Key-point detectors identify all the rotationally invariant features in an image, whereas key-point descriptors describe the appearance of regions around the key-points. Key-point detectors and descriptors facilitate image recognition by matching the key-points between the query image and database images. Thus, greater similarity between these images indicates higher numbers of matched key-points. SIFT is extensively used as a local feature descriptor in applications such as image classification, image retrieval, and image stitching (Yang and Kurita, 2013). In addition, the SIFT features have a high tolerance of beam, noise, and slight changes in the angle of view, and are therefore impervious to image size variations. The SIFT algorithm is implemented to compute the SIFT features in a salient object image; the image generates an unspecified number of key-points, each of which is represented by a 128-dimensional feature vector (Lowe, 2004).
3.1.2.2 Visual vocabulary creation

Most image retrieval systems based on local features use the BoW model to perform expanded image searches. The BoW model trains local features in order to build visual vocabulary and uses the vocabulary to quantify the features. After quantification, similar local features are clustered; the centroid of each cluster denotes one visual word. Therefore, a set of visual words can be used to represent an image, enabling image retrieval through the term frequency-inverse document frequency (TF-IDF) model (Lv et al., 2013).

Codebook creation is crucial to the BoW model because the visual feature vectors of an image can be generated according to codebooks. In addition, several studies have derived visual words through k-means clustering algorithm and high-dimensional data streams by mining correlated-clusters (Fehr et al., 2009; Fan et al., 2010; Yuan et al., 2011; Lv et al., 2013; Yang and Kurita, 2013) which has been extensively applied in image processing and is one of the simplest unsupervised algorithms (Wan and Qin, 2010).

3.1.2.3 Feature quantisation

Visual words in the BoW model can be referred to as a codebook. The frequency of all visual words in an image are recorded in the codebook and presented in the form of a BoF frequency histogram. Figure 9 illustrates the feature quantisation process.

Figure 9 Feature quantisation (see online version for colours)

The quantification of the recorded local features involves frequency statistics and weighting schemes (namely, TF, IDF, and TF-IDF) (Sivic and Zisserman, 2003). In this study, the TF-IDF weighting scheme was adopted to weight the vectors of the quantified features, as expressed in equation (16):

\[
\text{tf}\cdot\text{idf}_{i,j} = \begin{cases} 
\frac{f_{i,j}}{\sum_{i=1}^{k} f_{i,j}} \times \log \frac{N}{n_j} & \text{if } f_{i,j} > 0, \\
0 & \text{otherwise.}
\end{cases}
\]
where \( j \) is the image at location \( j \) in the image database; \( i \) is the visual word at location \( i \) in the codebook; \( k \) is the number of visual words in the codebook, which is equal to the number of clusters; \( N \) is the total number of images in the database; \( n_j \) is the total number of images containing visual word \( i \); \( f_{ij} \) is the frequency of visual word \( i \) in image \( j \); and \( tfidf_{ij} \) is the weighted value of the initial frequency \( f_{ij} \).

### 3.1.3 Similarity measurement

The similarity between the query image and database images is typically measured by estimating the distance between their feature vectors: the smaller the distance is, the greater the similarity is. Common distance functions for estimating this distance are Euclidean distance, cosine, Manhattan distance, Chebyshev distance, Minkowski distance, and Mahalanobis distance. This study used the Euclidean distance, a widely applied distance function, to determine the similarity between the query image and database images. Euclidean distance is expressed in equation (17):

\[
d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

where \( x \) and \( y \) denote two points at multidimensional spatial coordinates.

### 3.2 System construction and functionality

#### 3.2.1 Establish a model

1. image path: displays the link of a selected image data set folder
2. select folder: selects an image data set folder for image retrieval
3. indexing: performs salient object detection and BoF transformation for the selected image data set.

#### 3.2.2 Object image


#### 3.2.3 Query image

When an object picture listed on the object image panel is selected, the picture appears as the query image. A total of three object pictures can be selected as the query images. After the query images are determined, the user clicks the ‘search’ button to begin searching for target images.

#### 3.2.4 Search results

By estimating the similarity between the query images and database images, the system lists the top 25 images according to similarity in the Search Results panel. If no target
image is found, any picture from the search results that is deemed most similar to the target image can be selected as the query image to conduct further researches.

**Figure 10** Image search results (see online version for colours)

---

4 Experimental design and results

4.1 Experimental environment

All experiments were conducted using Matlab R2012a on Windows 8. The hardware specifications are as follows: Intel® Core™ i5-4200U CPU@1.60GHz 2.30GHz with 4G RAM.

4.2 Experimental design

This section introduces the experiment in this study, which were conducted to answer the following questions:

1. What are the effects of different cropped formats of salient object images on the image retrieval precision?

2. Given the effects of the codebook size on the image search results, what is the optimal cluster size for image retrieval?

3. How does the image search system perform when using the BoF model based on SIFT and salient object detection as compared with the BoF model based on SIFT only?
Are image searches viable when performed using one or multiple object images as the query image?

4.2.1 Retrieval performance assessment

The system’s image retrieval performance was estimated in terms of average precision (AP), as expressed in equation 18 and equation 19 (Yuan et al., 2011), and the recall function, as expressed in equation 20:

\[ p(i) = \frac{1}{M} \sum_{j=1}^{M} \gamma(i, j) \]  
\[ \gamma(i, j) = \begin{cases} 1 & \text{id}(i) = \text{id}(j) \\ 0 & \text{otherwise} \end{cases} \]  
\[ \text{Recall}(R) = \frac{\text{number of relevant images retrieved}}{\text{total number of relevant images}} \]

where \( p(i) \) is the precision of query image \( i \); \( \text{id}(i) \) and \( \text{id}(j) \) are the category of query images \( i \) and \( j \), respectively; \( \text{id}(i) \) is in the range of ten object images; and \( M \) is the total number of images belonging to category \( \text{id}(j) \) of image \( j \). Each object image type contains 100 pictures; thus, the maximum value of \( M \) is 100. In other words, \( p(i) \) denotes the percentage of images belonging to the category of query image \( i \) in the first \( M \) retrieved images. For example, if 17 of the first 25 retrieved images are similar to picture 4 of the query image of butterflies, then the retrieval precision is estimated as \( p(i) = \frac{17}{25} = 0.68 \).

4.3 Analysis of experimental results

4.3.1 Cropped formats of salient object images

After undergoing salient object detection, all the input images were cropped in two formats to obtain the salient object images. The percentage of images whose categories were completely identified was then estimated using both cropped formats of the salient object images (Table 2). For example, 70 salient object images of dogs were identified; thus, 70% of the salient object images of dogs were identifiable in terms of category (each salient object image comprised 100 pictures of the same category).

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Percentages (%) of objects identified in the original and rectangular salient object images</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bird</td>
</tr>
<tr>
<td>Original salient object images</td>
<td>99</td>
</tr>
<tr>
<td>Rectangular salient object images (cropped)</td>
<td>100</td>
</tr>
</tbody>
</table>
As shown in Table 2, objects in some of the original salient image types (‘ship’, ‘cat’, ‘horse’, and ‘human’) were less identifiable than in the corresponding rectangular ones, whereas those in other original salient images (‘bird’, ‘butterfly’, ‘flower’, and ‘house’) were almost equally identifiable. Overall, objects in the rectangular salient object images were more identifiable than those in original ones.

Figure 11  Mean retrieval rates at different codebook sizes (see online version for colours)

4.3.2 Codebook size

Based on the mean retrieval precision rates of the five-picture set of each object image, the optimal image search results were achieved with the codebook size of 200. These results are depicted in Figure 11, which presents the mean image retrieval precision rates for the ten object image types stratified by codebook size. In addition, this figure suggests that the mean retrieval rates for the rectangular salient images were noticeably higher than those for the original salient images. Thus, the cluster size was set to 200 in the subsequent experiments.

4.3.3 Image retrieval performance: SIFT-based BoF models with and without salient object detection

This experiment revealed that, compared with the conventional SIFT-based BoF model, the image retrieval precision of the SIFT-based BoF model with salient object detection improved. The mean retrieval precision rates improved for most of the original salient images that underwent salient object detection. Nevertheless, few of the original salient
images, namely cars, cats, and horses, attained decreased precision rates (Figure 12), likely because the contours of the salient objects of these images were partially removed during cropping.

**Figure 12** Image retrieval performance the sift-based BoF model with and without salient object detection (see online version for colours)

![Image Retrieval Performance Graph](image)

**Figure 13** Multi-object images (see only version for colours)

<table>
<thead>
<tr>
<th>Object Types</th>
<th>Boat, house</th>
<th>Tree, house</th>
<th>Dog, billboard, fence</th>
<th>Cat, tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images</td>
<td><img src="image" alt="Boat, house" /></td>
<td><img src="image" alt="Tree, house" /></td>
<td><img src="image" alt="Dog, billboard, fence" /></td>
<td><img src="image" alt="Cat, tree" /></td>
</tr>
<tr>
<td>Object Types</td>
<td>Horse, house, tree</td>
<td>Car, house, tree</td>
<td>Person, crown</td>
<td>Bird, meadow</td>
</tr>
<tr>
<td>Images</td>
<td><img src="image" alt="Horse, house, tree" /></td>
<td><img src="image" alt="Car, house, tree" /></td>
<td><img src="image" alt="Person, crown" /></td>
<td><img src="image" alt="Bird, meadow" /></td>
</tr>
</tbody>
</table>

### 4.3.4 Image retrieval precision based on single- and multi object images

This experiment was conducted to examine the feasibility of object images for image search. Image searches were performed using one, two, and three object images as the query image (Figure 13). Regarding the multiple-object search approach, two or three object images were integrated into one query image to conduct the search.
4.4 Optimal image search results

In the image retrieval experiments, using different cropped object images and codebook sizes improved the retrieval precision to varying extents, suggesting that the integrity of the rectangular salient objects in the images was relatively well-retained after the images were cropped through the salient object detection process, and that using such images improved the recognition of salient objects contained in them.

Dissimilar to the size of text vocabulary in information retrieval, the size of visual vocabulary, or codebook size, is determined by the number of key-points obtained through SIFT. However, a small visual vocabulary may lack discriminative capacity and cause two dissimilar key-points to be assigned to the same cluster, whereas a large vocabulary is susceptible to noise and computational complexity (Jiang et al., 2007). In this study, experiments were conducted with five codebook sizes (50, 100, 150, 200, and 250); the optimal image search results were obtained with a codebook size of 200.

Finally, the image retrieval precision analysis revealed that the search results varied according to the object image used because of differences between salient object images in the same categories. Moreover, some images were not retrieved at all when the existing portfolio of object images was used; increasing the number of object images in the search system could improve its capability of retrieving many images featuring considerably more complexity.

5 Conclusions and future studies

On the basis of the aforementioned experimental results, this chapter provides the conclusion, limitations, and research directions regarding the present study. The first section summarises the challenges and findings of the image retrieval experiments. The second section discusses the limitations of this study and proposes future research directions.

5.1 Conclusion

Current image retrieval technologies use either similar images/photos or sketches as the query images; however, both retrieval solutions have their limitations. Image querying based on a similar image/photo uses the entire image/photo, whereas image querying with sketches requires a sufficient impression of the target image and drawing capability of the user.

To address the limitations of conventional image retrieval solutions, this study proposed using object images as query images. This object-based image retrieval solution enables users to conduct searches with any object(s) within a target image. After a conventional SIFT-based BoF model was integrated with salient object detection to improve image retrieval precision, an image search system was constructed on the basis of the proposed research framework. The findings are as follows:

5.1.1 Modified image retrieval solution

An image retrieval precision analysis confirmed the feasibility of the proposed object-based image retrieval solution, which is more convenient and easier to use than
conventional retrieval solutions based on images or sketches that resemble the target image.

5.1.2 Image retrieval precision

After salient object detection was performed to identify salient objects in images, which filtered the image backgrounds and yielded rectangular salient object images, querying with the object images was less susceptible to interference from the background and other complex image features. In addition, the codebook size experiment revealed that the highest retrieval precision was attained with an image cluster size of 200. With the codebook size set to 200, the experiment on single-object image retrieval precision attained an average retrieval precision rate of 0.234 for all the object images (estimated by summing and averaging the mean retrieval precision rates of all these images). By contrast, the average retrieval precision rate for all the object images in the input images (without salient object detection) was 0.177, verifying that the retrieval precision of the SIFT-based BoF model can be improved through integrating the single-object querying approach with salient object detection.

The experiment on the double-object image retrieval yielded lower retrieval precision but a higher recall rate because the number of images containing two objects was limited. The experiment on triple-object image retrieval also yielded lower retrieval precision but a higher recall rate since relatively few objects contained three objects. The low retrieval precision in both experiments indicated that, because salient object detection retained only the salient objects in the images and filtered the image features, some images could not be retrieved on the basis of their content.

5.1.3 Operation of the image search system

The operation of the image search system was divided into ‘input’, ‘processing’, and ‘output’.

1 Input
   • Performing salient object detection and BoF transformation on all the images stored in a path specified by the user to establish the BoF vectors for the images.
   • Providing ten object image types with a combined 50 pictures and applying eight image deformation types yielded 450 object images as the query images.
   • Performing image searches based on one, two, or three object images as the query image.

2 Processing
   • Obtaining rectangular salient objects in the images through salient object detection.
   • Extracting features from the rectangular salient objects through SIFT.
   • Performing k-means clustering to cluster each SIFT feature, estimating the occurrence of the SIFT features in each cluster, and establishing the BoF vectors for the images.
   • Using the TF-IDF scheme to attenuate the effects of unimportant SIFT features on image retrieval.
3 Output

Euclidean distance was adopted to estimate the similarity between the query images and database images, which was used for ranking the retrieved images. The image search system presented the top 25 images in the search results panel, as well as the subsequently ranked images in the next page of this panel. In addition, the user could use any of the retrieved images as the query image to conduct further searches.

5.2 Limitations and research directions

1 Image data set

A total of 1,000 images contained in the MSRA-A image data set were used for the experiment. These experimental image data differ from the images users take in real-world contexts. Thus, future studies can investigate the application of real images in image retrieval.

2 Salient object detection

The salient object detection method adopted in this study has two limitations:

- Although the method rectangular crops salient objects from images, the resulting images contain part of the background. Thus, the contour of the salient object images alone cannot be retained.
- Problems with the hierarchical detection of the method remain to be solved.

Modifications can be made to the SIFT-based BoF model with salient object detection to determine whether the modified model improves image retrieval precision when object images are used as the query image.

3 Use different feature extraction methods

Feature extraction is currently based on SIFT. However, local feature descriptors for the BoF model also include more commonly applied descriptors such as speeded-up robust features (SURF) and local binary pattern (LBP). Both SURF and LBP can be integrated with the BoF model to improve the image retrieval precision of the model.

4 Global image search

Although this study used salient object detection to identify salient objects in images and then use the identified objects as the query image, the object-based image retrieval solution might not succeed in identifying salient objects in landscape images, which do not contain such objects.

5 Distortions in the resized images

The experimental image data set and object pictures therein were resized to less than 200 pixels in width or length, but they all contained distortions. An image-resizing solution that reduces distortions should be developed before image retrieval experiments are conducted.
References


