Knowledge-Based Systems 179 (2019) 8–20

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A decisive content based image retrieval approach for feature fusion in visual and textual images[☆]

Salahuddin Unar[a], Xingyuan Wang[a,b,*], Chunpeng Wang[c], Yu Wang[a]

[a] *Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China*
[b] *School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China*
[c] *Qilu University of Technology (Shandong Academy of Sciences), Shandong, China*

HIGHLIGHTS

- A decisive CBIR system is proposed that considers visual and textual contents.
- The system is innovative in dealing with visual as well as textual images.
- The system first classifies the query image into textual/non-textual and then extract its features.
- The system support three modes of retrieval: Image query, Keywords, and combination of both.
- The system is efficient and applicable for low-power hardware, and free from training efforts.

ARTICLE INFO

Article history:
Received 28 January 2019
Received in revised form 18 April 2019
Accepted 1 May 2019
Available online 14 May 2019

Keywords:
CBIR
Image retrieval
Information retrieval
Knowledge-based system
Feature fusion

ABSTRACT

Image content analysis plays a dynamic role in various computer vision applications. These contents can be either visual (i.e. color, shape, texture) or the textual (i.e. text appearing within images). Both the contents involve fundamental characteristics of an image and thus can be an enormous asset for any intelligent application. For content based image retrieval (CBIR) systems, most of the art methods are either annotated text based or the visual search based. Due to high demand of multitasking, there is a great need of a system that can combine visual as well as textual features. Consequently, this work proposes a decisive CBIR approach that combines visual and textual features to retrieve similar images. Firstly, the method classifies the query image as textual and non-textual. If any text appears within the image then the query image is classified as textual, and the text is detected and formed as Bag of Textual words. If the query image is classified as non-textual, the visual salient features are extracted and formed as Bag of Visual words. Next, the method fuses the visual and textual features, and top similar images are retrieved based on the fused feature vector. It supports three modes of retrieval: Image query, Keywords, and a combination of both. The experimental results on four datasets show the efficiency and accuracy of the proposed approach for visual and textual images.

1. Introduction

The explosive growth of massive digital devices has produced a huge amount of digital images in the last two decades. These devices are capable to capture, store, and share the images in an efficient and effective way. On the other hand, such a scenario has produced a huge repository of chaotic images with an improper organization in databases. To organize these images, content based image retrieval is working as backbone since the last few decades, and many scientists from multimedia and computer vision communities are working on this hot issue [1–3].

According to content based image retrieval (CBIR), the system retrieves a number of similar images when the user requests the system by providing a query image. The similar images are retrieved based on color, shape, and texture. In early years, image retrieval methods employed keywords based model (i.e. text-based image retrieval) to retrieve similar images. However, the limitation with TBIR (text-based image retrieval) is the necessity of meta-data (e.g. image tags and description) in advance. Such methods need a large amount of human labor to manually annotate each image that is more time consuming for millions of images. In later years, CBIR methods were introduced that were

[☆] No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.05.001.

[*] Corresponding author at: Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China.

E-mail addresses: imulticoder@gmail.com (S. Unar), wangxy@dlut.edu.cn (X. Wang).

https://doi.org/10.1016/j.knosys.2019.05.001
0950-7051/© 2019 Elsevier B.V. All rights reserved.

more user-friendly and may consider human perception more seriously.

A typical CBIR architecture has two core functionalities: data insertion and query processing. The data insertion function is performed independent of user interaction and is applied to all the data in the database. The main purpose of data insertion is to extract the visual salient features from all the images of the database. The query processing function is performed when the user requests the system by providing the query image, specific color, or desired sketch. The similarity is computed between the query image and the database images, and top similar images are retrieved based on chosen metrics. Another possibility for the user is to provide some random keywords and request the system to find similar textual images.

Day by day, as the number of images is increasing, the same way contents within images are being more complexed. Several new areas are being challenged for contents analysis with semantic learning [4–7]. The standard art methods have explored distinct techniques for visual search only. However, no standard method still exists to search textual images. The text appearing within images is certainly a precious clue to perceive the images and thus can be adopted to retrieve the images. Considering the textual contents along with visual contents may certainly enrich computer vision applications. For example, to help the visually impaired people to understand the image contents and to assist a robot to perform specific actions. Therefore, the feasibility of text appearing within the images might be useful for automatic indexing and retrieval.

Several studies proposed to extract visual features for retrieving similar images based on low-level visual features and descriptors [8]. Yang et al. [9] proposed a method using local visual attention features. SURF features are employed for detecting the salient features and visual attention feature is extracted around the strongest salient points. In [10], Liu et al. proposed image retrieval method that combines local binary pattern and color information features. Local binary pattern is capable to extract textural features while color information feature is capable to describe the color information of salient features. Tang et al. [11] proposed a novel visual feature based on discrete wavelet transform which helps to convert the input image into a normalized image. A color space model is employed that transforms the edge image of the normalized image to segment it into non-overlapping blocks. To achieve better accuracy, Zhang et al. [1] proposed a novel framework that considers visual salient objects within a complex background. Van et al. [12] introduced a method to create a binary signature on the basis of color and shape of interest of objects in an image. Sometimes, the rotation and scaling may affect the retrieval results. To cope with this limitation, multidimensional scaling can be beneficial to resist the image rotation [13,14].

Moreover, the embedded and scene text recognition has gained much attention recently [15–17]. However, a very rare work can be seen for detected text based image retrieval [18,19]. Neumann et al. [18] introduced a method that detects and recognize the scene text and employ the detected text to retrieve the similar images. Mishra et al. [19] proposed a method that detects the text in textual images and use that text as a query to search the similar textual images. Tang et al. [20] proposed a method that detects the text through edge cure and multiple features. The feature discrimination of textual regions is achieved by fusing low-level features and CNN-based deep features. These methods can only detect and localize the text from textual images. There is a great need to exploit the detected text in CBIR applications to retrieve similar textual images.

Furthermore, the different methods have been proposed that considers semantic learning of visual and textual contents. In [21], Cui et al. proposed a novel method that considers hybrid textual-visual relevance learning. The visual features are extracted and fused with social tags and annotations. In [22], Li et al. introduced a method that is capable to evaluate object and scene tag importance. The tags are measured from the human annotated description. Zhang et al. [23] proposed a method that learns feature paradigm from large-scale images and their tags. Their main purpose is to understand the semantic gap between the image and the social tags using deep learning. Until now, these methods still depend on the manually annotated description and tags that may not available most of the time. To overcome such limitation, we introduce automatic extraction of text and employ as the keywords/tags to retrieve the images.

In this paper, we propose a novel CBIR approach that combines visual and textual features to retrieve similar images. The approach first classifies the query image into the textual and non-textual category. For textual query image, the text within the image is detected and recognized. For non-textual query image, the salient visual features are extracted. Both the features (i.e. visual and textual) are fused together to form a final feature vector. Top rank similar images are retrieved by computing the similarity distance on the final feature vector. The proposed approach is applicable for visual as well as textual images. The main contributions of this work are as follow.

- A novel CBIR approach is proposed that considers visual and textual contents of the image to retrieve similar images.
- The proposed approach classifies the visual and textual images, and extracts its involved features.
- The approach is innovative in dealing with textual as well as visual images.
- The approach considers three modes of retrieval: Image query, Keywords, and a combination of both.
- The approach is efficient and effective for low-power hardware and independent of training efforts.

The rest of the paper is organized as follow: In Section 2, the proposed approach is described along with query classification, and the formation of bag of textual words and bag of visual words. In Section 3, the experiments are given with implementation detail and evaluation results. Section 4 concludes the proposed method.

2. Proposed method

In last few years, several methods have been proposed for content based image retrieval by many researchers and scientists. Most of them are based on low-level features and descriptors. The state-of-the-art methods are applicable for either visual search or the social tags based semantic learning, and no standard method has developed that consider retrieving the textual images. In this work, we introduce a novel decisive approach for visual as well as textual search. The proposed approach first classifies the query image into textual/non-textual. If the image is classified as a textual query, the text appearing within the image is localized and recognized. Bag of textual words model is employed that uses recognized words as keywords. If the image is classified as a non-textual query then visual salient features within the query image are extracted. Bag of visual words model is employed to store the extracted visual salient features. Next, the visual and textual features are fused together to form a fused feature vector. Different similarity measures are computed on the fused feature vector in order to ensure better accuracy. The top rank similar images are retrieved based on the fused feature vector. The proposed approach support three modes of retrieval: Image query, Keywords, and a combination of both. The schematic diagram of the proposed approach is given in Fig. 1.

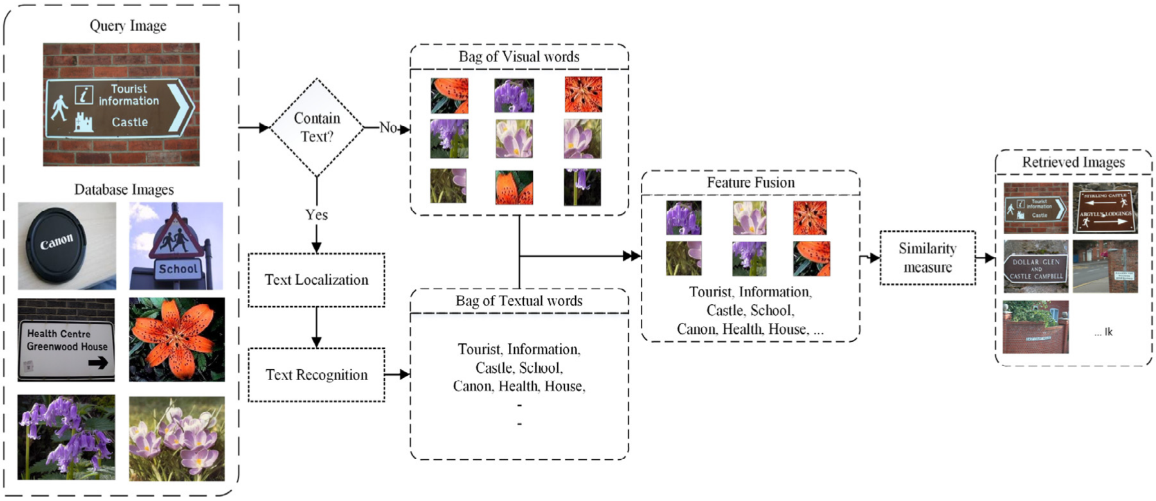


Fig. 1. Schematic diagram of proposed approach.

2.1. Text/non-text query classification

In order to distinguish text and the non-text query image, a decisive technique is used that classifies the query images. Since the proposed approach is efficient and applicable for textual as well as visual images (i.e. such images that does not contain text). The purpose of this step is to save the computation time of visual images without proceeding for the text detection step. This way, only the probable textual query image will proceed for text detection. Text/non-text query classification is useful in certain applications such as text detection, text tracking, live event detection, and others [24–27]. In our method, we divide each query image into 16 different blocks and each block is screened for text/non-text. If any of the blocks is noticeable with the text, the query image further proceeds for text localization and recognition steps. If all of the blocks marked with non-text then the query image proceeds for visual features extraction.

Generally, it is realized that the text comprises of sharp edges as compared to non-text blocks. The proposed method first finds the potential edges by using Sobel edge for each query image. Sobel edge filter is applied to each color frequency of the RGB image, as shown in Fig. 2. A single output image is formed by combining three edge images which consider the maximum edge pixel values.

Generally, the text does not appear within the whole image, despite it contain a limited space of an image. Hence, we divide the whole image into different blocks and block-wise text screening is performed. The method divides the query image of size 256 × 256 into 16 blocks of size 64 × 64 one by one. The text screening for textual images is stopped once any of the blocks is found to be a probable text block. If no text block is found then the visual features are extracted in the next phase. Inspired from the work [28], the proposed method achieves an average of three high-frequency sub-bands using wavelet decomposition in LH, HL, and HH. Next, the median and mean moments are computed to classify each block as text or non-text block by averaging the sub-bands. The median and mean moment for each probable block are given as

$$Me(I) = \begin{cases} SI(N^2/2), & N \text{ is odd} \\ \dfrac{SI((N^2-1)/2) + SI((N^2+1)/2)}{2}, & N \text{ is even} \end{cases} \tag{1}$$

$$\mu_2^{Me}(I) = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(I(i,j) - Me(I))^2 \tag{2}$$

$$\mu_3^{Me}(I) = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(I(i,j) - Me(I))^3 \tag{3}$$

$$M(I) = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}I(i,j) \tag{4}$$

$$\mu_2^{M}(I) = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(I(i,j) - M(I))^2 \tag{5}$$

$$\mu_3^{M}(I) = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(I(i,j) - M(I))^3 \tag{6}$$

where SI represents the pixel values of each block having size 64 × 64 , $N = 64$, and I represents the input matrix. From the above equations, a set of three features i.e. first, second, and third order moment is obtained. Next, these three features form a vector that represents all of 16 blocks and features are normalized to a binary value of [0, 1], given as

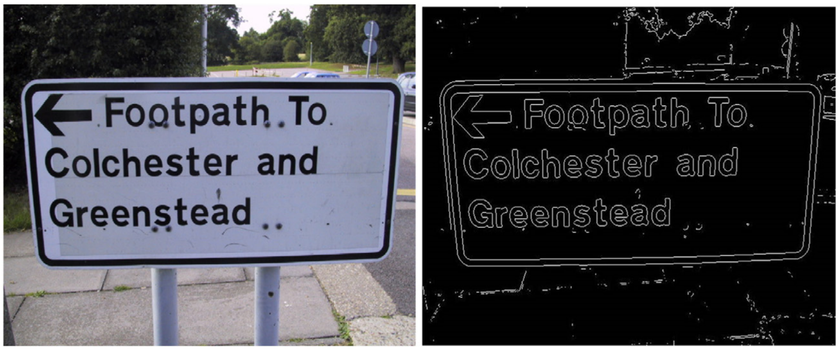
$$V = \{v_1, v_2, v_3, \ldots, v_{16}\} \tag{7}$$

To classify the vector into two classes i.e. textual and non-textual, a k-means clustering approach with $k = 2$ is applied. Those clusters having high mean value than the threshold are considered to be the text blocks. The reason to consider high mean value is the features showing the text pixels involve distinct sharp edges than the features showing non-text pixels. We apply k-means clustering on 16 feature vectors from Eq. (7) to find the potential text block. Consequently, the 16 blocks of the query image are classified as textual and non-textual blocks. Fig. 3(a) shows the 16 blocks of a query image and Fig. 3(b) shows the classified text and non-text blocks. The dark black blocks are classified as text blocks while light dark blocks are classified as non-text blocks. Once any of the blocks is found to be a potential text block, the query image is further processed for text localization step. If none of the blocks found to be the text block, the visual features are extracted without proceeding further for text localization step.

2.2. Bag of textual words

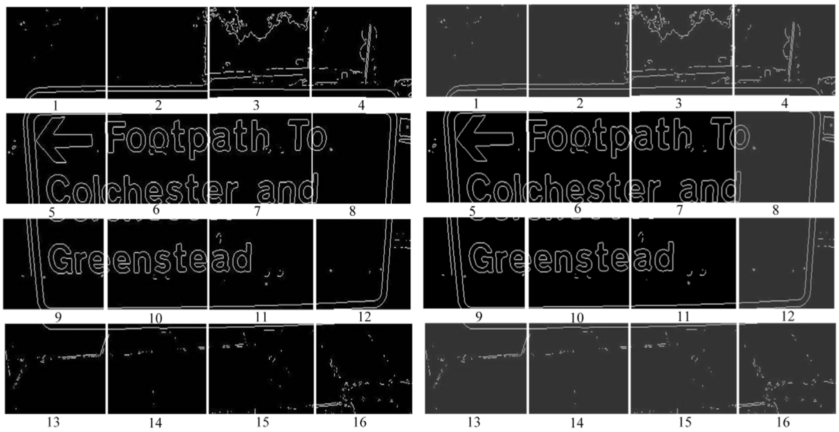
2.2.1. Text localization

Text in images conveys a valuable information and a clear clue about the scene within an image. To detect and recognize the text within images has equal and important value as visual



(a) (b)

Fig. 2. Image classification. (a) Original image (b) Sobel filter.



(a) (b)

Fig. 3. Classification of query image. (a) 16 blocks of query image (b) Selected text blocks (dark black) and non-text blocks (light black).

features in CBIR. In recent years, text detection and recognition have been considered a significant area for several computer vision applications including multilingual translation, automotive assistance, and image and video retrieval [29–33].

Text localization aims to extract the position of the text within the textual images. Textual images can be categorized into three main types: scene images, document images, and born-digital images. Scene images contain text that may be captured unintentionally e.g. text appearing within the banner, poster, building, shirt, door, wall, etc. and contain a complex background. Document images are structured in a sort of document that contains isolated text without any complex background. Born-digital images are the images in which the text is generated through post-processing techniques in the computer. Here we focus on scene images in which the text appears in a complex background with low resolution. Once the query image is classified as a textual image, the text appeared within the image is detected and recognized through certain step-filters. First, we apply Maximally stable extremal regions (MSER) algorithm to find the correspondence between the image components with different viewpoints [34]. MSER algorithm finds the potential connected components by adjusting a certain range of threshold value.

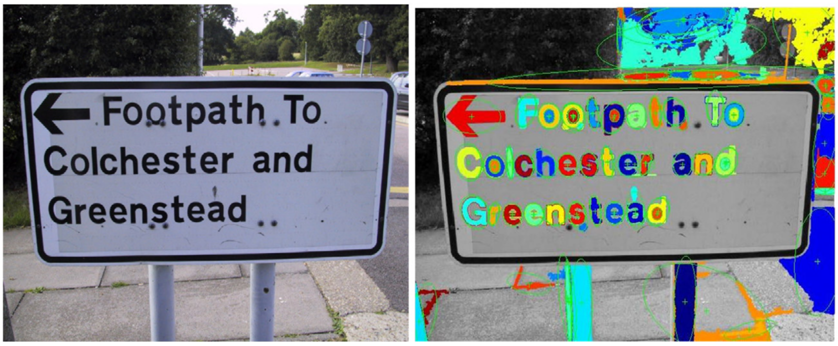
Given Q_m as the region areas segmented by the threshold m and Δ as the variation of m, we have

$$q(m) = \frac{Q_{m+\Delta} - Q_{m-\Delta}}{Q_m}, m \in [0, 255] \tag{8}$$

If $q(m)$ in Eq. (8) is a local minima, Q_m is considered as MSER which means the extracted region is potential to a variation of the threshold m. Smaller Δ results more MSERs. Most of the character candidates are detected after applying MSER as shown in Fig. 4(b).

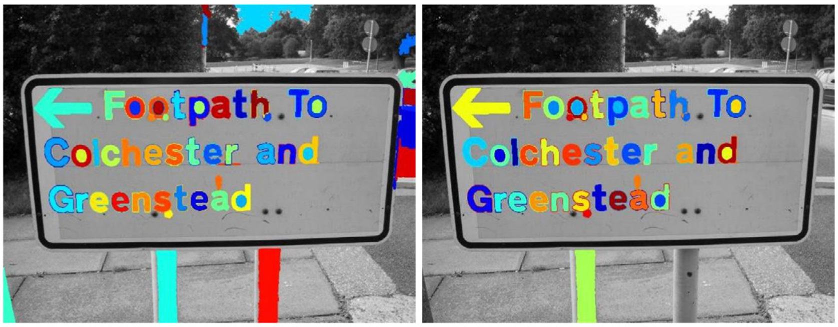
MSER extracted most of the stable character candidates but these are not the accurate text regions and many non-text objects may still exist. To reject false text regions, we apply two step-filters namely geometric filter and stroke feature transform that is the improved version of stroke width transform [35]. In geometric-filter, certain geometric constraints such as width, height, aspect ratio, etc. are used to sort out text and non-text character candidate regions. Initially, too wide and too narrow regions are discarded on its first place. Next, the connected components having too small and too large aspect ratio are also discarded. Such objects that contain a large number of holes are also discarded. Lastly, such objects that have the size of less than 5 pixels are also discarded. After applying the geometric filter, several non-text objects are rejected, as shown in Fig. 5(a).

After applying the geometric filter, certain non-text objects may still exist that can be further filtered out through the stroke feature transform filter. Generally, the stroke width in a single character remains the same within the same character, while a substantial change is found in the stroke width of non-text objects. Stroke pixels are detected by shooting a pixel spark from an edge pixel p_x to its conflicting edge pixel p_y along its gradient direction d_x. Each element of SWT is set to ∞ initially and the Canny edge detector is applied to find the potential edge pixels. Next, a pixel ray is passed along its gradient direction d_x for each edge pixel p_x to ensure the inadequate pixels along its way to an opponent pixel p_y. A ray is declared effective if current ray p_T satisfies the following two conditions:



(a) (b)

Fig. 4. Textual regions detection. (a) Original query image (b) Strongest detected textual regions.



(a) (b)

Fig. 5. Non-textual regions removal. (a) Geometric based (b) SFT based.

(i) Stroke width condition: p_T is a true pixel and its current gradient direction d_{cur} is conflicting to actual gradient direction d_x as:

$$||d_T - d_x| - \pi < \frac{\pi}{2} \tag{9}$$

(ii) Stroke color condition: the distance between the median ray color C_r and current pixel's color p_T(represented as C_T) satisfies $\|C_T - C_r\| > \lambda_c$. Here λ_c is calculated by linearly decreasing function regarding the numbers of pixels in p_T from 200 to 100. The current edge pixel is considered to be an edge pixel if the color gap is found and its direction is rechecked as in step (i) with the more lower threshold, given as

$$|| d_T - d_x| - \pi| < \frac{\pi}{6} \tag{10}$$

The rays that satisfies the above conditions are succeeded while remaining rays are rejected. The further filtration of invalid rays whose median colors are unlikely to be its neighbor pixels are also filtered out. Consequently, a stroke width mean value and median RGB color values are assigned to remaining valid pixels in order to create a stroke width map and a stroke color map. Fig. 5(b) shows the filtered out non-text objects after applying the stroke feature transform filter.

2.2.2. Textual words formation

After applying MSER algorithm and the step-filters, the majority of the non-text objects are removed. The next step is to group and merge the remaining pixels into meaningful words. We use some common heuristic constraints to merge the remaining text regions into lines. Generally, the text appears in linear form and characters involve the similar width, height, and stroke values. The two components will be merged into one if they satisfy any one of the following constraints:

(i) Merging two adjacent and overlapped connected components as:

$$\begin{aligned} & d(C_i, C_j) < t \\ & \wedge \min(h_i, h_j)/\max(h_i, h_j) > 0.5 \\ & \wedge \operatorname{abs}(R_i - L_j) \leq t_2 \\ & \wedge \operatorname{abs}(Cen_i - Cen_j) \leq t_1 \end{aligned} \tag{11}$$

where $d(C_i, C_j)$ represents the color difference between the two regions, h_i and h_j represents the height of two regions, R_i and L_j are maximum and minimum column index of two regions, respectively, and Cen_i, Cen_j represents the rows of center of regions. We set t to 25, and t_1 and t_2 to half of maximum width and height of two connected components, respectively.

(ii) If one region is found within another region, discard the smaller one.

Connected components satisfying the above constraints will be merged and the remaining components are supposed to be false and discarded, as shown in Fig. 6. The process ends when no components can be merged further.

The next step is to recognize the remaining textual components through optical character recognition (OCR). We employ Google's open source Tesseract OCR engine [36]. The recognized words are exploited as keywords for retrieving the images. The images which contain maximum words with high confidence score will be retrieved first. An auxiliary value of '1' is assumed if no textual word found. Unlike the method [37] that employed a character-level input model, we employ word-character recurrent language model that uses both word-level and character-level inputs [38]. Given the word w_t as an input at each time step t from the extracted words and long short-term memory (LSTM). The word-level input is changed into a high-dimension space by a

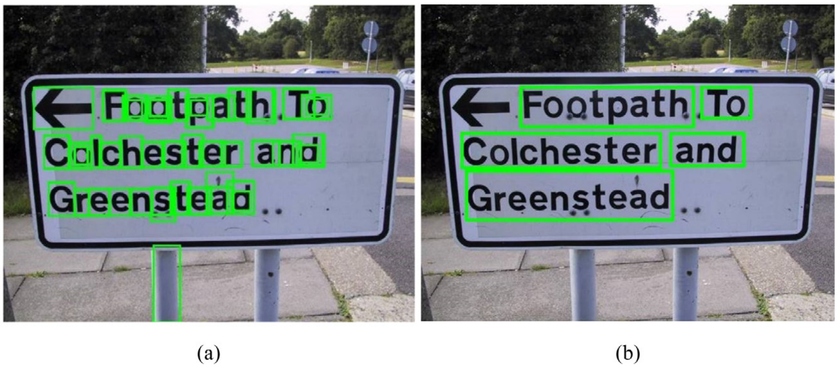


Fig. 6. Textual keywords formation. (a) Characters line merging (b) Obtained keywords.

word table $\mathrm{E} \in \mathbb{R}^{|V| \times d}$, where d is the dimension of a word vector and $|V|$ is the size of vocabulary given as

$$\mathrm{x}_{w_t}^{word} = \mathrm{E}^{\perp}\mathrm{w}_{w_t} \tag{12}$$

where $\mathrm{W}_{w_t} \in \mathbb{R}^{|V|}$ is a significant vector whose ith component is 1 and rest components are 0. A bidirectional LSTM is used to project character-level input into a word vector, and the last event of forward and reverse recurrent networks are combined linearly as

$$\mathrm{x}_{w_t}^{\mathrm{char}} = \mathrm{W}^f \mathrm{h}_{w_t}^f + \mathrm{W}^r \mathrm{h}_{w_t}^r + \mathrm{b} \tag{13}$$

where $\mathrm{h}_{w_t}^f$ and $\mathrm{h}_{w_t}^r$ belongs to $\mathbb{R}^d$ that are last events of forward and reverse LSTM, respectively. Whereas, $\mathrm{x}_{w_t}^{\mathrm{char}} \in \mathbb{R}^d$ represents a vector of word w_t using a character input, and $\mathrm{W}^f, \mathrm{W}^r \in \mathbb{R}^{d \times d}$ and $\mathrm{b} \in \mathbb{R}^d$ both are trainable elements. The resultant vectors $\mathrm{x}_{w_t}^{\mathrm{word}}$ and $\mathrm{x}_{w_t}^{\mathrm{char}}$ are fused together as

$$\mathrm{g}_{w_t} = \sigma(\mathrm{V}_{\mathrm{g}}^{\perp}\mathrm{x}_{w_t}^{\mathrm{word}} + b_{\mathrm{g}}) \tag{14}$$

$$\mathrm{x}_{w_t} = (1 - \mathrm{g}_{w_t})\mathrm{x}_{w_t}^{\mathrm{word}} + \mathrm{g}_{w_t}\mathrm{x}_{w_t}^{\mathrm{char}} \tag{15}$$

where $\sigma(.)$ is a sigmoid function, $\mathrm{v}_{\mathrm{g}} \in \mathbb{R}^d$ is a weight vector, and $b_{\mathrm{g}} \in \mathbb{R}$ is a bias scalar. The output vector x_{w_t} from Eq. (15) is used as an input to the LSTM model. The affine transformation of a hidden state h_t is given as

$$\Pr(w_{t+1} = k \,\big|\, w_{<_{t+1}}) = \frac{\exp(\mathrm{v}_k^{\perp}\mathrm{h}_t + b_k)}{\sum_{k'} \exp(\mathrm{v}_{k'}^{\perp}\mathrm{h}_t + b_{k'})} \tag{16}$$

where v_k is kth column of $\mathrm{V} \in \mathbb{R}^{d \times |V|}$ and b_k is kth component of $\mathrm{b} \in \mathbb{R}^d$.

2.3. Bag of Visual Words

In content based image retrieval, Bag-of-Visual-Words (BoVW) model is a promising approach to represent the images for image retrieval. In the early years, several methods have been proposed to retrieve the similar images based on the low-level features. The low-level features include color, shape, texture, and spatial position, and are known as global features. However, these features are variant to the rotation, translation, scaling, and affine transformation [39,40]. Moreover, these features are insufficient and perform poorly to recognize specific objects. To overcome such limitations, local descriptors were proposed in later years. Local descriptors finds the local structures in images and process the images more robustly. These local descriptors are invariant to rotation, translation, scaling, and affine transformation. Since our method is proposed for textual as well as visual images, a simple descriptor is insufficient as used in [37]. It is difficult for a single feature space to characterize the different varieties of images. The visual images may have different salient features than textual images, thus cannot be put into the same category under a single feature. We adopt a color descriptor along with local descriptor for a different subset of visual features. Firstly, the low-level features are extracted from each image using a color descriptor and a feature descriptor. Next, both the descriptors are quantized into a visual codebook using a clustering approach. Finally, a frequency histogram of visual words represents the image as bag-of-visual-words. The frequency histogram of bag-of-visual-words proceeds for the fusion step as a feature vector.

2.3.1. Color quantization

Color is the most significant visual characteristic for humans as well as machines' perception. Color quantization is the process to reduce the spatial color distribution with minimal visual distortion in an image. A color histogram is an effective color feature representation and widely used in computer vision tasks. There are several color models such as RGB color space, CIE $L^*a^*b^*$, CIE XYZ, and HSV extensively used for digital image processing [41]. Each color space has its own pros and cons according to the specific application. We use RGB color space model for color feature extraction due to high feasibility in digital image processing. RGB color space consists of low computational complexity and high robustness. In this model, the R, G, and B channels are uniformly quantized into several bins represented by Q_R, Q_G, and Q_B, respectively. The color feature of a single image that quantizes the entire color space into $Q_R * Q_G * Q_B$ colors varying from 0 to $Q_R * Q_G * Q_B - 1$, given as

$$H_{RGB} = Q_G Q_B R + Q_B G + B \tag{17}$$

Here, we consider $Q_R = Q_G = Q_B = 4$. The R, G, and B channels are quantized as

$$R = \begin{cases} 0, R \in [0, 64] \\ 1, R \in [65, 128] \\ 2, R \in [129, 192] \\ 3, R \in [193, 255] \end{cases}, G = \begin{cases} 0, G \in [0, 64] \\ 1, G \in [65, 128] \\ 2, G \in [129, 192] \\ 3, G \in [193, 255] \end{cases},$$

$$B = \begin{cases} 0, B \in [0, 64] \\ 1, B \in [65, 128] \\ 2, B \in [129, 192] \\ 3, B \in [193, 255] \end{cases} \tag{18}$$

The three-channel vector of RGB is constructed into a single dimension vector that quantizes the entire color space into 64 color values. The color values range from 0 to 63 and one dimension color histogram H_{RGB} is formed.

2.3.2. Features extraction

The extraction of low-level features detects and identifies a bunch of salient keypoints within an image. We consider well-known Binary Robust Invariant Scalable Keypoints (BRISK) descriptor for salient visual features extraction [42]. BRISK algorithm is greatly efficient and effective feature descriptor over state-of-the-art SIFT and SURF descriptors. It is simple yet effective, rotation, and multiscale invariant corner detector. It is a binary descriptor in which each keypoint is a bit which represents the binary output. The fast computation and high efficiency make it more reliable for low-complexity and low-power hardware.

Given the set $\mathcal{A}$ for all sample keypoint pairs as

$$\mathcal{A} = \{(\mathbf{p}_m^i, \mathbf{p}_m^j) \in \mathbb{R}^2 \times \mathbb{R}^2 \,|i < N \wedge j < i \wedge i, j \in \mathbb{N}\} \tag{19}$$

where $\mathbf{p}_m^i \in \mathbb{R}^2$, $i = 1, 2, 3, \dots, N$ denotes the location of a sampling keypoint in a coordinate system located at (x_m, y_m), scaled with σ_m, and rotated with angle θ_m. An intensity value $I(\mathbf{p}_m^i, \rho_i)$ is achieved by averaging the pixel values at locations around $\mathbf{p}_m^i$ and ρ_i depends on distance from center of sampling points. To compute the binary comparisons up to $N(N-1)/2$ for each pair in $\mathcal{A}$, we have

$$b = \begin{cases} 1, I(\mathbf{p}_m^i, \rho_j) > I(\mathbf{p}_m^i, \rho_i) \\ 0, \text{ otherwise} \end{cases} \tag{20}$$

The pattern is composed of two subsets i.e. short-distance pairs $\mathcal{S}$ and long-distance pairs $\mathcal{L}$ given as

$$\mathcal{S} = \left\{(\mathbf{p}_m^i, \mathbf{p}_m^j) \in \mathcal{A} \,\middle|\, \left\|\mathbf{p}_m^j - \mathbf{p}_m^i\right\| < \delta_{\max}\right\} \subseteq \mathcal{A} \tag{21}$$

$$\mathcal{L} = \left\{(\mathbf{p}_m^i, \mathbf{p}_m^j) \in \mathcal{A} \,\middle|\, \left\|\mathbf{p}_m^j - \mathbf{p}_m^i\right\| > \delta_{\min}\right\} \subseteq \mathcal{A} \tag{22}$$

The local gradient is computed over $\mathcal{L}$ pairs in the subset as

$$\mathbf{g}(\mathbf{p}_m^i, \mathbf{p}_m^j) = (\mathbf{p}_m^j - \mathbf{p}_m^i) \cdot \frac{I(\mathbf{p}_m^j, \sigma_j) - I(\mathbf{p}_m^i, \sigma_i)}{\left\|\mathbf{p}_m^j - \mathbf{p}_m^i\right\|^2} \tag{23}$$

The estimation of characteristic direction D_c at keypoint c is given as

$$D_c = \begin{pmatrix} g_x \\ g_y \end{pmatrix} = \frac{1}{\mathrm{L}} \sum_{(\mathbf{p}_m^i, \mathbf{p}_m^j) \in \mathrm{L}} \mathbf{g}(\mathbf{p}_m^i, \mathbf{p}_m^j) \tag{24}$$

where g_x represents the gradient in the x-axis and g_y represents the gradient in the y-axis. Hamming distance is used to match two binary strings and the bit-by-bit differences are counted using simple XOR function. The extracted and matched salient keypoints are shown in Fig. 7.

2.3.3. Codebook formation

In this section, the extracted and matched salient keypoints are quantized into the visual vocabulary. Quantization is the process to locate the nearest center in feature space. To improve recall and minimize quantization error, multiple vocabularies are generated and each descriptor is quantized into multiple visual words by multiple vocabularies. For a given set of salient keypoints $Y = (y_1, y_2, y_3, \dots, y_e)$, the vocabulary of K visual words is created on Y by k-means clustering approach. The vocabulary is achieved by optimizing quantization distortion given as

$$\min_{\{c_k\}} \sum_{k=1}^{K} \sum_{y \in C_k} d(y, c_k)$$
$$C_k = \{y\,|d(y, c_k) < d(y, c_{k'})\,, \forall k' \neq k\} \tag{25}$$

where C_k represents kth cluster, c_k represents the center of the cluster C_k and indicates a visual word, and $d(\cdot)$ shows squared Euclidean distance. To train L vocabularies in order to reduce the total quantization error, we have

$$\min_{\left\{c_k^l\right\}} \sum_{l=1}^{L} \left(\sum_{k=1}^{K} \sum_{y \in C_k^l} d(y, c_k^l) \right)$$
$$C_k^l = \{y\,\left|d(y, c_k^l) < d(y, c_{k'}^l)\right., \forall k' \neq k\} \tag{26}$$

where C_k^l represents kth cluster of lth vocabulary, c_k^l represents the center of the cluster C_k^l and indicates kth visual word of lth vocabulary, and $c_{k'}^l$ represents k'th visual word of lth vocabulary.

The salient keypoints and descriptors are used to generate the vocabularies. Each image is represented as a frequency histogram of salient features that are in the image. The frequency histogram shows the number of occurrence of a visual word in an image and proceeds further as a feature vector.

2.4. Feature fusion

Feature fusion is the process to combine two or more feature vectors into a single feature vector that is more discriminative than a single input feature vector. Feature fusion can be achieved at certain levels such as matching-level, feature-level, or decision-level. Since we have two data modalities (i.e. visual and textual) we consider feature-level fusion according to which the features extracted from input entities are initially combined and then further processed as a single unit to perform the final analysis. Canonical correlation analysis (CCA) is one of the well-known statistical technique to deal with mutual representation between two random feature vectors [43].

Consider two matrices $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{q\times n}$ containing n input feature vectors from two distinct modalities. Both the feature vectors with p and q dimensionality are extracted from each modality. Assume $S_{xx} \in \mathbb{R}^{p\times p}$ and $S_{yy} \in \mathbb{R}^{q\times q}$ represents the inside components covariance matrices of X and Y, and $S_{xy} \in \mathbb{R}^{p\times q}$ and $S_{yx} = S_{xy}^T$ represents the middle components covariance matrices. Given S for overall $(p+q) \times (p+q)$ covariance matrix containing information on the relationship between the following pair of features as

$$S = \begin{pmatrix} cov(x) & cov(x, y) \\ cov(y, x) & cov(y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \tag{27}$$

CCA generates the linear combinations $\overset{*}{X} = W_x^T X$ and $\overset{*}{Y} = W_y^T Y$ which optimizes the pair correlation between two feature vectors as

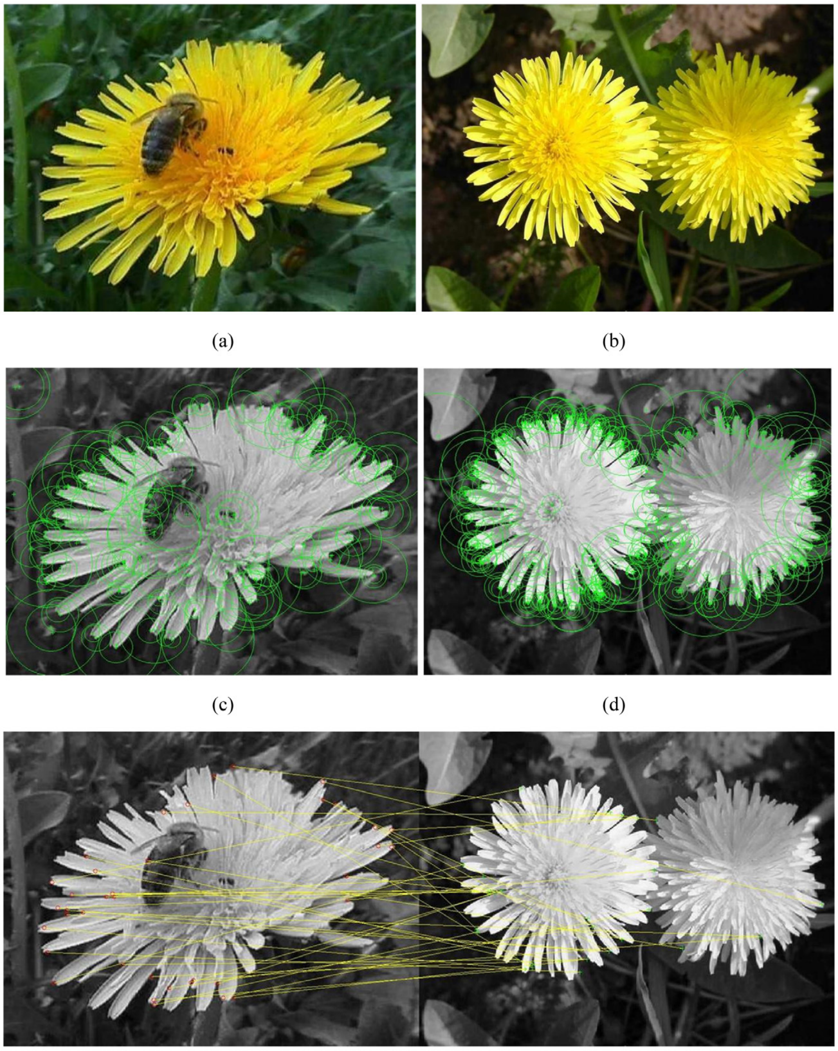
$$corr(\overset{*}{X}, \overset{*}{Y}) = \frac{cov(\overset{*}{X}, \overset{*}{Y})}{var(\overset{*}{X}) \cdot var(\overset{*}{Y})} \tag{28}$$

where $cov(\overset{*}{X}, \overset{*}{Y}) = W_x^T S_{xy} W_y$, $var(\overset{*}{X}) = W_x^T S_{xx} W_x$, and $var(\overset{*}{Y}) = W_y^T S_{yy} W_y$. The optimization is achieved using Lagrange multipliers by increasing the covariance between $\overset{*}{X}$ and $\overset{*}{Y}$by adjusting the variables $var(\overset{*}{X}) = var(\overset{*}{Y}) = 1$ [44]. The linear transformation matrices W_x and W_y are given as

$$S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \hat{W}_x = R^2 \hat{W}_x \tag{29}$$

$$S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} \hat{W}_y = R^2 \hat{W}_y \tag{30}$$

where $\hat{W}_x$ and $\hat{W}_y$ represents the eigenvectors, and R^2 represents the square of canonical correlations. The non-zero eigenvalues in both the equations is $d = rank(S_{xy}) \leq \min(n, p, q)$ that is arranged descending $r_1 \geq r_2 \geq \cdots \geq r_d$. The transformation matrices W_x and W_y containing the arranged eigenvectors corresponds to non-zero eigenvalues, and $\overset{*}{X}, \overset{*}{Y} \in \mathbb{R}^{d\times n}$ represents the recognized canonical variations. The covariance matrix for sorted data can be



(a) (b)

(c) (d)

(e)

Fig. 7. Visual salient features extraction. (a) Query image (b) database image (c)–(d) strongest salient keypoints (e) keypoints matching.

given as

$$\overset{*}{S} = \left(\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & r_1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & r_2 & \cdots & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & r_d \\ \hline r_1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 0 & 0 & \cdots & r_d & 0 & 0 & \cdots & 1 \end{array} \right) \tag{31}$$

The above Eq. (31) shows the canonical variations containing non-zero correlation on their corresponding indices. The identity matrix on top left and bottom right shows the canonical variations are uncorrelated in each feature vector. The final feature-level fusion is performed by summation or concatenation of transformed feature vectors given as

$$Z_1 = \overset{*}{X} + \overset{*}{Y} = W_x^T X + W_y^T Y = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{32}$$

$$Z_2 = \begin{pmatrix} \overset{*}{X} \\ \overset{*}{Y} \end{pmatrix} = \begin{pmatrix} W_x^T X \\ W_y^T Y \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{33}$$

where Z_1 and Z_2 are discriminant features of canonical correlation.

3. Experimental results and discussion

In this section, we perform an extensive number of experiments on different datasets to analyze the performance of proposed approach. The detailed discussion is given as follow:

3.1. Datasets and evaluation measures

3.1.1. Datasets

Since the proposed approach is efficient and effective for textual as well as visual images. We evaluated the proposed approach on four different datasets containing visual and textual images.

ICDAR 2013/2011 Dataset. In this dataset [45] all the images are textual images containing different text on different objects such as signboards, books, posters, banners, etc. The dataset contains 462 images including 229 training and 233 test images under varying dimension from 350 × 200 to 3888 × 2592 .

Street View Text (SVT) Dataset. The dataset [46] contains textual images that are captured with Google street view. The dataset contains 350 images having a different dimension from 1024 × 768 to 1920 × 906 . The text appearing within the images is low-quality font and very challenging to recognize.

Wang's Dataset. The dataset [47] contains 1000 visual images categorized into 10 different groups. The dimension of each image is either 256 × 384 or 384 × 256 .

Oxford Flowers Dataset: The dataset [48] contains 1360 visual images categorized into 17 different groups. Each group consists of 80 related images. The dimension of each image varies from 499 × 499 to 1057 × 500 .

3.1.2. Similarity measure

The distance between the extracted feature vectors is evaluated by a similarity measure. The proposed approach computes four distinct similarity measures that are Euclidean distance, Canberra distance, Manhattan distance, and Cosine similarity. Consider a feature vector $F_{DB_i} = \{w_1, w_2, \dots, w_N\}$ for each database image where N is the number of fused features in database image and a feature vector $F_q = \{w_1, w_2, \dots, w_N\}$ for each query image q, where N is the number of fused features in the query image. The similarity measures for Euclidean distance, Canberra distance, Manhattan distance, and Cosine similarity are given as follows, respectively:

$$D(F_{DB_i}, F_q) = \left(\sum_{i=1}^{N}(F_{DB_i} - F_q)^2\right)^{\frac{1}{2}} \tag{34}$$

$$D(F_{DB_i}, F_q) = \sum_{i=1}^{N} \frac{\left|F_{DB_i} - F_q\right|}{\left|F_{DB_i}\right| + \left|F_q\right|} \tag{35}$$

$$D(F_{DB_i}, F_q) = \sum_{i=1}^{N} \left|F_{DB_i} - F_q\right| \tag{36}$$

$$D(F_{DB_i}, F_q) = \frac{F_{DB_i}.F_q}{\left\|F_{DB_i}\right\| \left\|F_q\right\|} \tag{37}$$

where F_{DB_i} represents the feature vector of ith image in the database and F_q represents the feature vector of the query image q.

3.1.3. Textual evaluation measures

The evaluation of text detection is essential since the proposed method exploits the detected keywords for image retrieval. The text localization is evaluated on well-known measures of precision p, recall r, and f-measure given in [49] as

$$p' = \frac{\Sigma_{r_e \in E} m(r_e, T)}{|E|} \tag{38}$$

$$r' = \frac{\Sigma_{r_t \in T} m(r_t, E)}{|T|} \tag{39}$$

$$f = \frac{1}{\frac{\alpha}{p'} + \frac{(1-\alpha)}{r'}} \tag{40}$$

where E represents the number of estimated words, T represents the ground truth values, and f represents the combination of precision and recall values. The variable α is used to adjust the relative weights of p and r values.

3.1.4. Retrieval evaluation measures

The commonly used evaluation measure for image retrieval is *mean Average Precision (mAP)* that is the average of all query images. We adopt *mAP* to evaluate the proposed approach for final retrieval of images.

$$P(R_k) = \frac{\#(relevant\ Images \cap retrieved\ Images)}{\#(retrieved\ Images)} \tag{41}$$

where R_k represents the number of top retrieved images and we set $k = 10$ for retrieving the top ten similar images. Consider the set of relevant query images $q_i \in Q$ is $\{I_1, I_2, \dots, I_m\}$, where Q represents a set of all queries. To formulate the mean average precision (mAP), we have

$$mAP(Q) = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{m}\sum_{k=1}^{m} P(R_k) \tag{42}$$

Table 1
Classification rate of textual and non-textual query image on different datasets.

Dataset	Classification rate (%)	
	Textual	Non-textual
ICDAR	92.57	7.43
SVT	57.31	42.69
CBIR Flower	15.89	84.11
Wang	20.92	79.08

3.2. Implementation

This section describes the comprehensive implementation and discusses the output results. Firstly, the proposed approach classifies the query image as a textual and non-textual image. If the query image is found to be textual image then the text appearing within the image is localized, detected, and recognized. The recognized text is formed as keywords for text-based retrieval mode. Next, the same image is also processed through the visual features extraction step. If the query image is found to be non-textual image then the only visual features are computed. To ensure the accuracy and efficiency of the text detection step, we perform text detection experiments on ICDAR and SVT datasets which are the textual datasets. Next, the retrieval experiments are performed on all four datasets for similar images retrieval. The main purpose is to retrieve top rank similar images that have maximum similar visual features along with maximum similar textual words.

3.2.1. Text/non-text query classification

We achieve the text/non-text query classification on all the query images from the textual and visual datasets. We set a high classification ratio by adjusting the threshold since if any visual image is marked as a textual image it will be processed for text detection. There is no cost of sabotage if no text is found except a tiny cost of computation time. The obtained results of query classification for different datasets are given in Table 1. For ICDAR dataset, the results show that only a little percentage of images are marked as non-textual. For SVT dataset, the classification rate is not that good due to high complexity and excessive visual salient features. For CBIR flower and Wang datasets, a small number of images are falsely classified as a textual image.

3.2.2. End-to-end text detection

The query image classified as a textual image is processed for text detection and recognition. In this step, two experiments are conducted on ICDAR and SVT datasets.

Experiment I. Text detection. The text appearing within the textual query image is localized and detected in this experiment. The results are obtained using the textual evaluation measures defined in Section 3.1.3. The obtained results for ICDAR and SVT datasets are given in Tables 2 and 3. It is worth to mention that the proposed method achieves better precision and recall values than state-of-the-art methods for ICDAR. However, for SVT dataset, the obtained results are inferior due to query classification step. A good percentage of the images are classified as non-textual images during the decisive step as can be seen in Table 1.

Table 2
Text detection results on ICDAR 2013/2011.

Method	Precision	Recall	*f*
Kumar et al. [50]	0.75	0.70	0.72
Yin et al. [51]	0.83	0.65	0.73
Wang et al. [52]	0.82	0.68	0.74
Neumann et al. [53]	0.82	0.71	0.76
Unar et al. [37]	0.81	0.79	0.79
Our method	0.83	0.82	0.82

Table 3
Text detection results on SVT Dataset.

Method	Precision	Recall	*f*
Wei et al. [54]	0.18	0.41	0.25
Neumann et al. [18]	0.19	0.32	0.24
Yu et al. [55]	0.27	0.35	0.30
Wang et al. [52]	0.59	0.48	0.53
Unar et al. [37]	0.54	0.51	0.52
Our method	0.47	0.42	0.44

Table 4
End-to-End text recognition results on ICDAR 2013/2011.

Method	Precision	Recall	*f*
Weinman et al. [56]	0.36	0.41	0.33
Neumann et al. [53]	0.44	0.45	0.45
Unar et al. [37]	0.56	0.51	0.53
Our method	0.52	0.59	0.55

Table 5
End-to-End text recognition on SVT dataset.

Method	*f-score*
Wang et al. [57]	0.46
Alsharif et al. [58]	0.48
Neumann et al. [53]	0.68
Unar et al. [37]	0.71
Our method	0.64

Table 6
Performance evaluation (mAP) for Image Query search mode.

Method/Dataset	ICDAR 2013/2011	SVT	Wang	Oxford flowers
Yang et al. [59]	–	–	0.76	0.64
Elalami [60]	–	–	0.76	0.69
Walia et al. [61]	–	–	0.66	–
Lin et al. [62]	–	–	0.72	0.63
Liu et al. [63]	0.67	0.53	–	–
Unar et al. [37]	0.79	0.68	–	–
Our method	0.81	0.75	0.78	0.71

Table 7
Performance (mAP) evaluation for Keywords Query search mode.

Method/Dataset	ICDAR 2013/2011	SVT
Mishra et al. [19]	0.65	0.56
Neumann et al. [18]	–	0.23
Unar et al. [37]	0.71	0.63
Our method	0.74	0.59

Table 8
Performance evaluation (mAP) for Image+Keywords Query search mode.

Method/Dataset	ICDAR 2013/2011	SVT	Wang	Oxford flowers
Yang et al. [59]	–	–	0.76	0.64
Elalami [60]	–	–	0.76	0.69
Walia et al. [61]	–	–	0.66	–
Lin et al. [62]	–	–	0.72	0.63
Liu et al. [63]	0.67	0.53	–	–
Unar et al. [37]	0.74	0.67	–	–
Mishra et al. [19]	0.65	0.56	–	–
Neumann et al. [18]	–	0.23	–	–
Our method	0.79	0.74	0.77	0.72

Experiment II. Text recognition. In this experiment, the localized and detected text is recognized and formed as Bag of textual words for image retrieval. The results are obtained using precision and recall evaluation measures. Precision p is the ratio between the total keywords correctly recognized to the total keywords recognized by the system. Recall r is the ratio between the total keywords correctly recognized to the total keywords localized. The overlap ratio is set to be greater than 70% in order to achieve more keywords. Tables 4 and 5 show the obtained results for text recognition. The proposed approach outperforms state-of-the-art methods for ICDAR dataset. However, for SVT, the recognition rate is not that good since the limited textual regions were localized as can be seen in Table 3.

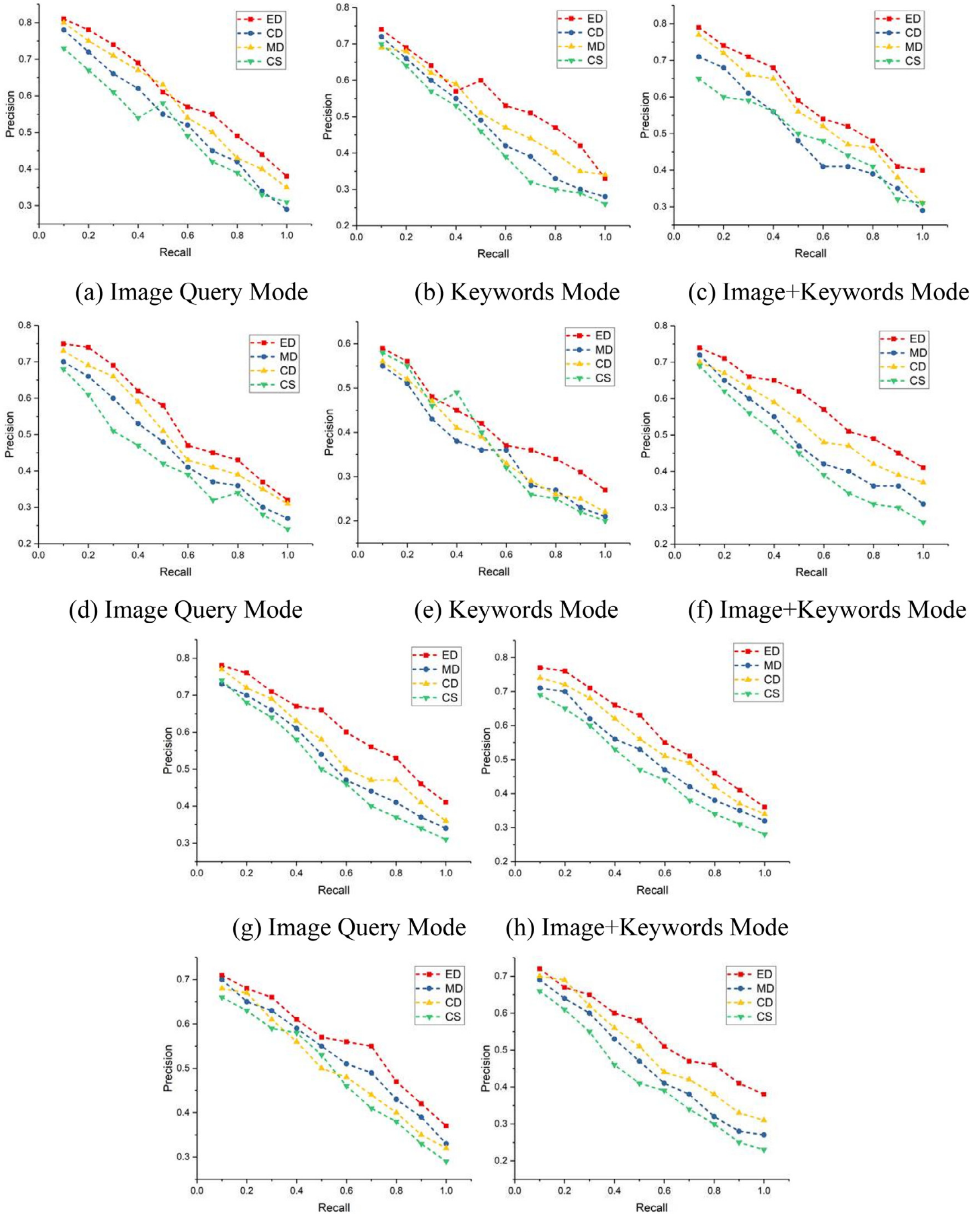
Although both the ICDAR and SVT datasets are textual datasets but the biggest difference between them is the former contains the focused text that is easy to be recognized. However, the later contains the unfocused text with several other objects such as banners, vehicles, trees, people, etc. Such objects may affect the query classification step but certainly helpful for visual search retrieval in the next step of extracting the visual features.

3.2.3. *Image retrieval evaluation*

In this section, the performance of the proposed method is evaluated for all four datasets in terms of retrieval measures defined in Section 3.1.4. The proposed method has three modes of retrieval: (i) Image Query (ii) Keywords (iii) Image Query + Keywords. The experiments are performed on all three modes of search. Considering textual images, 100 random images are selected as query images from ICDAR dataset, and 50 random images are selected as query images from SVT dataset. Considering visual images, 100 random images 10 from each group are selected as query images from Wang dataset, and 170 random images 10 from each group are selected as query images from Oxford Flowers dataset.

Experiment I. This experiment is based on the first and default mode of retrieval that is Image Query mode. This mode acts by default until the user specifies another mode. In this mode of search, the user provides a query image that is either a visual image or the textual image then the method classify it as textual or non-textual. If it contains any text within the image then the text is detected and formed as keywords. If no text is found then the visual salient features are computed in the next step. For Image Query mode, the method retrieves top k images that have maximum similar textual strings along with maximum similar salient features. The similarity measures defined in Section 3.1.2. are used to evaluate the similarity between each query image and the database image. The overall performance accuracy of the proposed approach is achieved in terms of mean Average Precision (mAP). The obtained results given in Table 6 show that the proposed method outperforms state-of-the-art methods. For SVT dataset, although fewer keywords were recognized but the maximum salient visual features are matched. Consequently, the proposed approach significantly outperforms the art methods.

Experiment II. This experiment is based on the second mode of retrieval that is Keywords-based search. This mode is efficient and capable of textual images search only since the textual images contain textual cues. In this mode, the user provides the random keywords and the method attempt to find similar textual images that contain provided keywords. For top k similar images, two matching functions are performed: Exact substring match and Approximate substring match. Exact substring match searches the images that have exact matching keywords with the user provided keywords and approximate substring match searches the images that have most probable matching keywords. The



(a) Image Query Mode (b) Keywords Mode (c) Image+Keywords Mode

(d) Image Query Mode (e) Keywords Mode (f) Image+Keywords Mode

(g) Image Query Mode (h) Image+Keywords Mode

(i) Image Query Mode (j) Image+Keywords Mode

Fig. 8. Average precision and recall at different distance measures for each mode. (a)–(c) For ICDAR (d)–(f) For SVT (g)–(h) For Wang (i)–(j) For Oxford Flowers dataset.

exact substring match has the highest priority over approximate substring match, thus retrieved first. The obtained results given in Table 7 show that the proposed approach outperforms state-of-the-art methods for ICDAR dataset. However, for SVT, the precision rate is not that good since the method considered salient features more than the only textual features during classification.

Experiment III. This experiment is based on the third mode of retrieval that is Image+Keywords search. In this mode of search, the user provides a query image along with random desired keywords. The method first classifies the query image and find its visual and textual features. The method locates the visual salient features with user-provided query image and matches the maximum similar keywords with user-provided keywords. A 50% ratio is set between the visual salient features and keywords. The exact substring match has higher priority over approximate substring as in Experiment II. An auxiliary value of 1 is added if no keywords are matched. The obtained results given in Table 8 show the proposed method outperforms state-of-the-art methods for the datasets.

Table 9
Average retrieval time (ms) performance for different datasets.

Dataset/Mode	Image Query	Keywords	Image+Keywords
ICDAR	5937	5813	6272
SVT	6551	6475	6780
Wang	3627	–	3806
Oxford Flowers	3194	–	3472

3.2.4. Average precision at different distance measures

The similarity measures defined in Section 3.1.2. are evaluated since each distance measure have different consequences. The distance measure computes the distance between the two feature vectors. The impact of different distance measures including Euclidean distance, Canberra distance, Manhattan distance, and Cosine similarity is computed to assure the retrieval accuracy. The effects achieved in terms of precision and recall values are shown in Fig. 8. The results show that no matter what distance measure is selected, the precision and recall values are opposing with increasing the number of searches.

3.2.5. Retrieval time comparison of different modes

In CBIR, time consumption for retrieving the similar images is an important parameter to be considered. Extracting maximum features may lead to more computation and retrieval time since the number of features and retrieval time are inversely proportional. The proposed approach finds a good balance between the retrieval time and retrieval accuracy. The obtained results are given in Table 9 for retrieval time of each mode. The proposed approach first classifies the query image and then retrieve the similar images accordingly. Hence, the classification step may slightly affect the retrieval time.

4. Conclusion

In this work, we proposed a novel CBIR approach that retrieves the top rank similar images by considering visual and textual features. The proposed approach first classifies the query image as textual or non-textual and extract its features. If any text appears within the query image, it is classified as a textual query and then the text is detected. If no text appears, it is classified as non-textual query and the salient visual features are extracted. We use Bag of textual words and Bag of visual words model to store the textual and visual features, respectively. Each model consists of certain step-filters to compute the visual and textual features and combined together to form a fused feature vector. The similarity measures are computed based on the fused feature vector and top rank images are retrieved. The proposed method supports three modes of search: Image query, Keywords, and a combination of both. In Image query mode, the user provides the query image and similar images are retrieved by considering salient features and keywords. In Keywords mode, the user provides some random keywords and the similar textual images are retrieved that have maximum similar strings. In the last mode, the user provides query image along with some random keywords and top rank images are retrieved based on the salient features as well as keywords. Experimental results on four datasets show the efficiency and accuracy of the proposed approach for visual and textual images over state-of-the-art methods. The current approach is implemented and validated on English language textual dataset. In future, we intend to implement and validate our approach for textual images which contain textual contents given in different languages (e.g. Arabic, Chinese). We also intend to decrease the computation time and introduce new tactics for learning the low-level features.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Nos: 61672124, 61370145 and 61173183), the Password Theory Project of the 13th Five-Year Plan National Cryptography Development Fund (No: MMJJ20170203).

References

[1] J. Zhang, S. Feng, D. Li, Y. Gao, Z. Chen, Y. Yuan, Image retrieval using the extended salient region, Inf. Sci. (NY) 399 (2017) 154–182, http://dx.doi.org/10.1016/j.ins.2017.03.005.

[2] C. Reta, I. Solis-Moreno, J.A. Cantoral-Ceballos, R. Alvarez-Vargas, P. Townend, Improving content-based image retrieval for heterogeneous datasets using histogram-based descriptors, Multimedia Tools Appl. (2018) http://dx.doi.org/10.1007/s11042-017-4708-8.

[3] M.K. Kundu, M. Chowdhury, S.R. Bulo, A graph-based relevance feedback mechanism in content-based image retrieval, Knowl.-Based Syst. 73 (2014) 254–264, http://dx.doi.org/10.1016/j.knosys.2014.10.009.

[4] H. Xu, C. Huang, D. Wang, Enhancing semantic image retrieval with limited labeled examples via deep learning, Knowl.-Based Syst. 163 (2019) 252–266, http://dx.doi.org/10.1016/J.KNOSYS.2018.08.032.

[5] E. Yildizer, A.M. Balci, T.N. Jarada, R. Alhajj, Integrating wavelets with clustering and indexing for effective content-based image retrieval, Knowl.-Based Syst. 31 (2012) 55–66, http://dx.doi.org/10.1016/J.KNOSYS.2012.01.013.

[6] S. Unar, X. Wang, C. Zhang, C. Wang, Detected text-based image retrieval approach for textual images, IET Image Process. 13 (2018) 515–521, http://dx.doi.org/10.1049/iet-ipr.2018.5277.

[7] E. Rashedi, H. Nezamabadi-pour, S. Saryazdi, A simultaneous feature adaptation and feature selection method for content-based image retrieval systems, Knowl.-Based Syst. 39 (2013) 85–94, http://dx.doi.org/10.1016/J.KNOSYS.2012.10.011.

[8] M.E. Elalami, A novel image retrieval model based on the most relevant features, Knowl.-Based Syst. 24 (2011) 23–32, http://dx.doi.org/10.1016/j.knosys.2010.06.001.

[9] H.-Y. Yang, Y.-W. Li, W.-Y. Li, X.-Y. Wang, F.-Y. Yang, Content-based image retrieval using local visual attention feature, J. Vis. Commun. Image Represent. 25 (2014) 1308–1323, http://dx.doi.org/10.1016/j.jvcir.2014.05.003.

[10] P. Liu, J.M. Guo, K. Chamnongthai, H. Prasetyo, Fusion of color histogram and LBP-based features for texture image retrieval and classification, Inf. Sci. (NY) 390 (2017) 95–111, http://dx.doi.org/10.1016/j.ins.2017.01.025.

[11] Z. Tang, Z. Huang, H. Yao, X. Zhang, L. Chen, C. Yu, Perceptual image hashing with weighted DWT features for reduced-reference image quality assessment, Comput. J. 61 (2018) 1695–1709, http://dx.doi.org/10.1093/comjnl/bxy047.

[12] T.T. Van, T.M. Le, Content-based image retrieval based on binary signatures cluster graph, Expert Syst. 35 (2018), http://dx.doi.org/10.1111/exsy.12220.

[13] Z. Tang, Z. Huang, X. Zhang, H. Lao, Robust image hashing with multidimensional scaling, Signal Process. 137 (2017) 240–250, http://dx.doi.org/10.1016/J.SIGPRO.2017.02.008.

[14] Z. Tang, X. Zhang, S. Zhang, Robust perceptual image hashing based on ring partition and NMF, IEEE Trans. Knowl. Data Eng. 26 (2014) 711–724, http://dx.doi.org/10.1109/TKDE.2013.45.

[15] H. Wu, B. Zou, Y. qian Zhao, J. Guo, Scene text detection using adaptive color reduction, adjacent character model and hybrid verification strategy, Vis. Comput. 33 (2017) 113–126, http://dx.doi.org/10.1007/s00371-015-1156-1.

[16] Y. Zheng, Q. Li, J. Liu, H. Liu, G. Li, S. Zhang, A cascaded method for text detection in natural scene images, Neurocomputing 238 (2017) 307–315, http://dx.doi.org/10.1016/j.neucom.2017.01.066.

[17] S. Unar, A.H. Jalbani, M. Shaikh, K.H. Memon, M.A. Ansari, Z. Memon, A study on text detection and localization techniques for natural scene images, Int. J. Comput. Sci. Netw. Secur. 18 (2018) 99–111.

[18] L. Neumann, J. Matas, Real-time scene text localization and recognition, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2012) 3538–3545, http://dx.doi.org/10.1109/CVPR.2012.6248097.

[19] A. Mishra, K. Alahari, C.V. Jawahar, Image retrieval using textual cues, in: Proc. IEEE Int. Conf. Comput. Vis. 2013, pp. 3040–3047, http://dx.doi.org/10.1109/ICCV.2013.378.

[20] Y. Tang, X. Wu, Scene text detection via edge cue and multi-features, Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR (2017) 156–161, http://dx.doi.org/10.1109/ICFHR.2016.0040.

[21] C. Cui, P. Lin, X. Nie, Y. Yin, Q. Zhu, Hybrid textual-visual relevance learning for content-based image retrieval, J. Vis. Commun. Image Represent. 48 (2017) 367–374, http://dx.doi.org/10.1016/j.jvcir.2017.03.011.

[22] S. Li, S. Purushotham, C. Chen, Y. Ren, C.-C.J. Kuo, Measuring and predicting tag importance for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 8828 (2017) 1–14, http://dx.doi.org/10.1109/TPAMI.2017.2651818.

[23] H. Zhang, X. Shang, H. Luan, M. Wang, Learning from collective intelligence: Feature learning using social images and tags, ACM Trans. Multimed. Comput. Commun. Appl. 13 (2016) http://dx.doi.org/10.1145/2978656.

[24] R.A. Sinoara, J. Camacho-Collados, R.G. Rossi, R. Navigli, S.O. Rezende, Knowledge-enhanced document embeddings for text classification, Knowl.-Based Syst. 163 (2019) 955–971, http://dx.doi.org/10.1016/J.KNOSYS.2018.10.026.

[25] H. Zhang, G. Zhong, Improving short text classification by learning vector representations of both words and hidden topics, Knowl.-Based Syst. 102 (2016) 76–86, http://dx.doi.org/10.1016/J.KNOSYS.2016.03.027.

[26] C. Liu, W. Wang, G. Tu, Y. Xiang, S. Wang, F. Lv, A new centroid-based classification model for text categorization, Knowl.-Based Syst. 136 (2017) 15–26, http://dx.doi.org/10.1016/J.KNOSYS.2017.08.020.

[27] J. Lu, Zhongxing. Ye, Yuru. Zou, Zhongxing ye yuru zou huber fractal image coding based on a fitting plane, IEEE Trans. Image Process. 22 (2013) 134–145, http://dx.doi.org/10.1109/TIP.2012.2215619.

[28] P. Shivakumara, A. Dutta, T. Quy Phan, C. Lim Tan, U. Pal, A novel mutual nearest neighbor based symmetry for text frame classification in video, Pattern Recognit. 44 (2011) 1671–1683, http://dx.doi.org/10.1016/j.patcog.2011.02.008.

[29] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, Image–text sentiment analysis via deep multimodal attentive fusion, Knowl.-Based Syst. (2019), http://dx.doi.org/10.1016/J.KNOSYS.2019.01.019.

[30] Z. Xia, X. Wang, L. Zhang, Z. Qin, X. Sun, K. Ren, A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing, IEEE Trans. Inf. Forensics Secur. 11 (2016) 2594–2608, http://dx.doi.org/10.1109/TIFS.2016.2590944.

[31] S. Unar, A.H. Jalbani, M.M. Jawaid, M. Shaikh, A.A. Chandio, Artificial urdu text detection and localization from individual video frames, Mehran Univ. Res. J. Eng. Technol. 37 (2018) 429–438, http://dx.doi.org/10.22581/muet1982.1802.18.

[32] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, Q. Dai, Effective uyghur language text detection in complex background images for traffic prompt identification, IEEE Trans. Intell. Transp. Syst. 19 (2018) 220–229, http://dx.doi.org/10.1109/TITS.2017.2749977.

[33] E.S. Tellez, D. Moctezuma, S. Miranda-Jiménez, M. Graff, An automated text categorization framework based on hyperparameter optimization, Knowl.-Based Syst. 149 (2018) 110–123, http://dx.doi.org/10.1016/J.KNOSYS.2018.03.003.

[34] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image Vis. Comput. 22 (2004) 761–767, http://dx.doi.org/10.1016/j.imavis.2004.02.006.

[35] W. Huang, Z. Lin, J. Yang, J. Wang, Text localization in natural images using stroke feature transform and text covariance descriptors, Proc. IEEE Int. Conf. Comput. Vis. (2013) 1241–1248, http://dx.doi.org/10.1109/ICCV.2013.157.

[36] R. Smith, D. Antonova, D.-S. Lee, Adapting the tesseract open source OCR engine for multilingual OCR, in: Proc. Int. Work. Multiling. OCR - MOCR '09, 2009, p. 1, http://dx.doi.org/10.1145/1577802.1577804.

[37] S. Unar, X. Wang, C. Zhang, Visual and textual information fusion using kernel method for content based image retrieval, Inf. Fusion 44 (2018) 176–187, http://dx.doi.org/10.1016/j.inffus.2018.03.006.

[38] Y. Miyamoto, K. Cho, Gated word-character recurrent language model, in: Proc. 2016 Conf. Empir. Methods Nat. Lang. Process., 2016, pp. 1992–1997, http://dx.doi.org/10.18653/v1/D16-1209.

[39] J. Lu, C. Xu, Y. Xu, Multiplicative noise removal in imaging: An exp-model and its fixed-point proximity algorithm, Appl. Comput. Harmon. Anal. 41 (2016) 518–539, http://dx.doi.org/10.1016/J.ACHA.2015.10.003.

[40] J. Lu, H. Yang, L. Shen, Y. Zou, Ultrasound image restoration based on a learned dictionary and a higher-order MRF, Comput. Math. Appl. 77 (2019) 991–1009, http://dx.doi.org/10.1016/J.CAMWA.2018.10.031.

[41] Z. Tang, X. Zhang, X. Li, S. Zhang, Robust image hashing with ring partition and invariant vector distance, IEEE Trans. Inf. Forensics Secur. 11 (2016) 200–214, http://dx.doi.org/10.1109/TIFS.2015.2485163.

[42] B. Brisk, O.R.B. Card, BRISK: Binary Robust Invariant Scalable Keypoints, 2011, pp. 1–8.

[43] Q. Sen Sun, S.G. Zeng, Y. Liu, P.A. Heng, D.S. Xia, A new method of feature fusion and its application in image recognition, Pattern Recognit. 38 (2005) 2437–2448, http://dx.doi.org/10.1016/j.patcog.2004.12.013.

[44] J. Lu, L. Shen, C. Xu, Y. Xu, Multiplicative noise removal with a sparsity-aware optimization model, Inverse Probl. Imaging 11 (2017) 949–974, http://dx.doi.org/10.3934/ipi.2017044.

[45] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L.G.I. Bigorda, S.R. Mestre, J. Mas, D.F. Mota, J.A. Almazan, L.P. De Las Heras, ICDAR 2013 robust reading competition, Proc. Int. Conf. Doc. Anal. Recognition ICDAR (2013) 1484–1493, http://dx.doi.org/10.1109/ICDAR.2013.221.

[46] W. Kai, B. Babenko, S. Belongie, End-to-end scene text recognition, in: Comput. Vis., ICCV, 2011 IEEE Int. Conf., 2011, pp. 1457–1464, http://dx.doi.org/10.1109/ICCV.2011.6126402.

[47] J.Z. Wang, J. Li, G. Wiederhold, SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 947–963, http://dx.doi.org/10.1109/34.955109.

[48] M.E. Nilsback, A. Zisserman, A visual vocabulary for flower classification, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2 (2006) 1447–1454, http://dx.doi.org/10.1109/CVPR.2006.42.

[49] S.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, ICDAR 2003 robust reading competitions, Proc. Int. Conf. Doc. Anal. Recognition, ICDAR 2003 (2003) 682–687, http://dx.doi.org/10.1109/ICDAR.2003.1227749.

[50] T.P. Kumar, M.V.P. Reddy, P.K. Bora, Multi-oriented text detection from video using sub-pixel mapping, Proc. Int. Conf. Comput. Vis. Image Process. 460 (2017) 531–541, http://dx.doi.org/10.1007/978-981-10-2107-7.

[51] X.C. Yin, W.Y. Pei, J. Zhang, H.W. Hao, Multi-orientation scene text detection with adaptive clustering, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 1930–1937, http://dx.doi.org/10.1109/TPAMI.2014.2388210.

[52] X. Wang, Y. Song, Y. Zhang, J. Xin, Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis, Pattern Recognit. Lett. 60–61 (2015) 41–47, http://dx.doi.org/10.1016/j.patrec.2015.04.005.

[53] L. Neumann, J. Matas, Real-time lexicon-free scene text localization and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 1872–1885, http://dx.doi.org/10.1109/TPAMI.2015.2496234.

[54] Y. Wei, Z. Zhang, W. Shen, D. Zeng, M. Fang, S. Zhou, Text detection in scene images based on exhaustive segmentation, Signal Process., Image Commun. 50 (2017) 1–8, http://dx.doi.org/10.1016/j.image.2016.10.003.

[55] C. Yu, Y. Song, Y. Zhang, Scene text localization using edge analysis and feature pool, Neurocomputing 175 (2016) 652–661, http://dx.doi.org/10.1016/j.neucom.2015.10.105.

[56] J.J. Weinman, Z. Butler, D. Knoll, J. Feild, Toward integrated scene text reading, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 375–387, http://dx.doi.org/10.1109/TPAMI.2013.126.

[57] T. Wang, D.J. Wu, A. Coates, A.Y. Ng, End-to-end text recognition with convolutional neural networks, in: 21st Int. Conf. Pattern Recognition, vol. 2012, 2012, pp. 3304–3308.

[58] O. Alsharif, J. Pineau, End-to-end text recognition with hybrid HMM maxout models, arxiv prepr. 2013 - arxiv.org, 2013, ArXiv1310.1811.

[59] J. Yang, B. Jiang, B. Li, K. Tian, Z. Lv, A fast image retrieval method designed for network big data, IEEE Trans. Ind. Informatics (2017) 2350–2359, http://dx.doi.org/10.1109/TII.2017.2657545.

[60] M.E. Elalami, A new matching strategy for content based image retrieval system, Appl. Soft Comput. J. 14 (2014) 407–418, http://dx.doi.org/10.1016/j.asoc.2013.10.003.

[61] E. Walia, A. Pal, Fusion framework for effective color image retrieval, J. Vis. Commun. Image Represent. 25 (2014) 1335–1348, http://dx.doi.org/10.1016/j.jvcir.2014.05.005.

[62] C.H. Lin, R.T. Chen, Y.K. Chan, A smart content-based image retrieval system based on color and texture feature, Image Vis. Comput. 27 (2009) 658–665, http://dx.doi.org/10.1016/j.imavis.2008.07.004.

[63] G.-H. Liu, J.-Y. Yang, Z. Li, Content-based image retrieval using computational visual attention model, Pattern Recognit. 48 (2015) 2554–2566, http://dx.doi.org/10.1016/j.patcog.2015.02.005.