Cognitive Feature Fusion for Effective Pattern Recognition in Multi-modal Images and Videos

Yijun Yan

In the fulfilment of the requirement for the degree of

Doctor of Philosophy

Centre for Excellence in Signal and Image Processing

Department of Electronic and Electrical Engineering

University of Strathclyde

Supervised by

Doctor Jinchang Ren

Professor John Soraghan

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Yijun Yan

ACKNOWLEDGEMENTS

After a long period of four years, I finally finish my PhD study. It is an impressive period for me, and I get many self-improvements, not only in the research area, but also on a personal level. I would like to thank many people who have supported and helped me so much throughout this period.

Firstly, I would like to express my most sincere gratitude to my supervisor, Doctor Jinchang Ren for his long-lasting guidance and all the opportunities I was given to conduct my research. I am also grateful to my co-supervisor, Professor John Soraghan, for his encouragement and guidance.

I am also appreciative of all co-authors in my publications, external partners in collaboration with my research group, the journal reviewers involved in the peer-review process of my publications, and my external and internal examiners, respectively. I am grateful as well to all my colleagues during this period for sharing their time and expertise, and building positive office atmosphere.

This research would have been impossible without the funding from the University of Strathclyde and my supervisors. I really appreciate their economic support and I am confident that my research outcomes have returned this effort in the best possible way. I must thank as well the university staff for their guidance and treat.

Finally, I would like to thank my parents for their moral and emotional support in my life. And I am also grateful to all my friends for their encouragement and understanding along the way.

ABSTRACT

Image retrieval and object detection have been always popular topics in computer vision, wherein feature extraction and analysis plays an important role. Effective feature descriptors can represent the characteristics of the images and videos, however, for various images and videos, single feature can no longer meet the needs due to its limitations. Therefore, fusion of multiple feature descriptors is desired to extract the comprehensive information from the images, where statistical learning techniques can also be combined to improve the decision making for object detection and matching. In this thesis, three different topics are focused which include logo image retrieval, image saliency detection, and small object detection from videos.

Trademark/logo image retrieval (TLIR) as a branch of content-based image retrieval (CBIR) has drawn wide attention for many years. However, most TLIR methods are derived from CBIR methods which are not designed for trademark and logo images, simply because trademark/logo images do not have rich colour and texture information as ordinary images. In the proposed TLIR method, the characteristic of the logo images is extracted by taking advantage of the color and spatial features. Furthermore, a novel adaptive fusion strategy is proposed for feature matching and image retrieval. The experimental results have shown the promising results of the proposed approach, which outperforms three benchmarking methods.

Image saliency detection is to simulate the human visual attention (i.e. bottom-up and top-down mechanisms) and to extract the region of attention in images, which has been widely applied in a number of applications such as image segmentation, object detection, classification, etc. However, image saliency detection under complex natural environment is always very challenging. Although different techniques have been proposed and produced good results in various cases, there is some lacking in modeling them in a more generic way under human perception mechanisms. Inspired by Gestalt laws, a novel unsupervised saliency detection

framework is proposed, where both top-down and bottom-up perception mechanisms are used along with low level color and spatial features. By the guidance of several Gestalt laws, the proposed method can successfully suppress the backgroundness and highlight the region of interests. Comprehensive experiments on many popular large datasets have validated the superior performance of the proposed methodology in benchmarking with 8 unsupervised approaches.

Pedestrian detection is always an important task in urban surveillance, which can be further applied for pedestrian tracking and recognition. In general, visible and thermal imagery are two popularly used data sources, though either of them has pros and cons. A novel approach is proposed to fuse the two data sources for effective pedestrian detection and tracking in videos. For the purpose of pedestrian detection, background subtraction is used, where an adaptive Gaussian mixture model (GMM) is employed to measure the distribution of color and intensity in multi-modality images (RGB images and thermal images). These are integrated to determine the background model where biologically knowledge is used to help refine the background subtraction results. In addition, a constrained mean-shift algorithm is proposed to detect individual persons from groups. Experiments have fully demonstrated the efficacy of the proposed approach in detecting the pedestrians and separating them from groups for successfully tracking in videos.

Contents

List of Figures

List of Table

CHAPTER 1

1 INTRODUCTION

1.1 Motivation

Aiming to use computers to simulate and progressively replace the function of human vision, computer vision has been a research hotspot for several decades [30]. Typical applications can be found in image retrieval [31], surveillance[32], saliency detection [33], object tracking [34], 3D reconstruction [35], and robotics [36] et al. Among these applications, feature extraction is the key as it can extract representative information to characterise images and videos, which consequently helps to enhance the efficiency and efficacy of the algorithm [30].

With the rapid development of computer science and imaging sensor technologies, the type of images/videos has become more and more diverse (e.g. video image, logo image, thermal image, spectral image etc), which naturally facilitates a number of new applications e.g. trademark/logo image retrieval, object tracking in multi-modal video[37]. To tackle the challenges caused by these emerging technologies and applications, conventional feature extraction techniques are found to be lacking in processing the new data especially under complex situations, i.e. multimodal data sources and unusual colour/texture components [38]. As a result, more generic models are desirable for effective feature extraction, especially in addressing these challenges.

Motivated by human's visual perception mechanisms, cognitive model based feature fusion is emphasised in this thesis, which has been successfully applied to three typical applications of computer vision, i.e. trademark/logo image retrieval (TLIR), image saliency detection, and pedestrian detection and tracking in multi-modal videos. For each of the three applications, models and algorithms are successfully derived for feature extraction, fusion and image/video analysis. Comprehensive experiments show that the proposed methods for each topic yield improved results than their benchmarking peers.

Trademark/logo image retrieval (TLIR) as a branch of CBIR has been drawn more and more attention in recent years because of its huge commercial value e.g. copyright protection by detection of misuse of existing trademarks/logos [39]. More broadly, it can be used for anti-fake detection [39], brand related statistics on social media [40], vehicle logo for intelligent traffic-control systems [41], etc. However, the traditional CBIR technologies are mostly designed for the nature images, which are not very suitable for trademark/logo images. Nature images mainly contain colour and texture characteristics, whereas the trademark/logo images contain graphic shape and texts at different locations hence spatial information becomes important [42]. Moreover, most images in the real world have embedded noise caused by image acquisition, transmission and compression et al. Therefore, the accuracy, reliability and robustness of the methods to noise is also a challenging problem which will affect the retrieval performance. To this end, a novel noise-robust trademark/logo retrieval method is focused as the first challenging topic in this thesis.

Image saliency detection is to extract the region of interest within an image, i.e. predicting where the vision attention focuses [33]. Although it has been developed for 20 years, there are still many new concepts and methods proposed every year in order to meet more diverse needs [3]. There are two main challenging problems on this topic. First, most existing methods do not have strong theoretical support to fully explore the human visual mechanisms, although they may produce good results of saliency detection. Second, new models are needed for more difficult cases i.e. images with high diversity such as irregular shape, complexed background, and multiple regions of interest, et al [29]. As a result, effective models for unsupervised image saliency detection is focused as the second topic in this thesis, guided by human vision perception and cognitive models, which is also compared with the state-of-the-art unsupervised and supervised approaches including deep learning models.

In recent years, there is a trend in visual surveillance for object detection and tracking on multi-modal images such as RGBT (colour and thermal imagery) and RGBD (colour and depth imagery)[43]. Although RGB images can provide many detail information of colour, texture, and shape, they suffer from the illumination change which increases the difficulty of object detection. On the contrary, thermal imagery shows good robustness to illumination changes by sensing objects' surface temperature. However, its spatial resolution tends to be low and shows lack of detail as RGB image does. To this end, the combination of RGB and thermal image becomes a new trend in visual surveillance in the recent 1-2 decades[44, 45], which is also selected as the third topic in this thesis. By fusion of RGB and thermal images, effective models and approaches are developed for pedestrian detection and tracking. This also shows the potential of the proposed approach in dealing with videos.

1.2 Thesis Structure

The rest of this thesis is organised as follows:

Chapter 2 provides an overview of the background introduction and related work for trademark/logo image retrieval, saliency detection in the coloured image and pedestrian segmentation and tracking in video.

Chapter 3 introduce the proposed TLIR method wherein the colour feature extraction, spatial feature extraction and dynamic feature fusion and matching are detailed. The proposed method is evaluated on thousands of logo images with several categories and most of them are degraded by different level of noise. From the results, the proposed method is proved to be robust to the noise and has good retrieval performance.

Chapter 4 presents a saliency detection method guided by Gestalt law namely GLGOV. Inspired by Gestalt law, each model of proposed saliency detection method has strong theory background. Through doing comprehensive experiments, the proposed method is benchmarked with 10 state-of-the-art methods on five widely used database. And the experimental results show that the proposed method can produce reliable performance according to objective and subjective assessment.

Chapter 5 proposed a multi-model background subtraction method where a cognition guided two stage background subtraction method is applied to detect the foreground/ROIs by fusing the background subtraction results of visible and thermal images. In addition, an improved mean-shift method is also proposed to track the individual pedestrian in the foreground. The proposed background subtraction method is benchmarked with several state-of-the-art techniques on an open database. From the analysis, the proposed method yields the best performance in terms of precision, recall and F-measure.

Chapter 6 briefly summarize the contributions of this thesis and discuss some further improvement of the proposed methods in the future.

1.3 Thesis Contributions

In this thesis, three new methods for three different applications in pattern recognition field are proposed and evaluated. Generally, these methods aim to extract more effective features for better performance in terms of retrieval, saliency detection, and segmentation. A detailed summary of these contributions is highlighted in the following:

1. The methodology proposed for colour and spatial feature extraction aims to extract a more effective feature of trademark/logo images, and adaptive feature matching strategy aims to produce a good performance of retrieval based on different feature effectiveness in different images.
2. A Gestalt-law guided saliency detection framework is proposed, which has been evaluated on many popular databases and benchmarked with many state-of-the-art algorithms. With the proposed method, more effective feature extraction is achieved, at the same time, the regions of interest are also highlighted.
3. A fusion model of the thermal and visible image is proposed to detect and track the pedestrians on the street in the videos. Some cognitive knowledge is used to guide each key step in the model.

CHAPTER 2

2 Background and Literature Survey

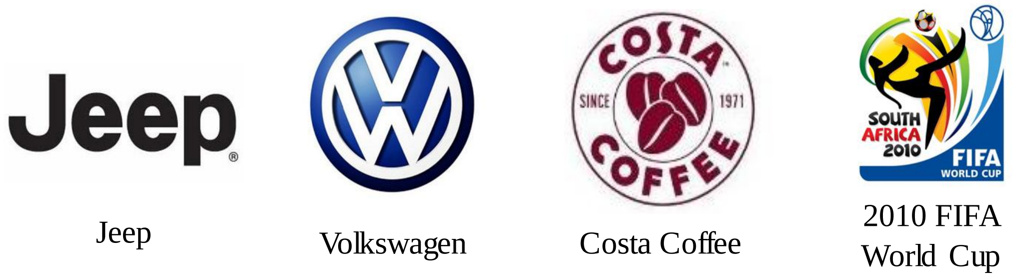
2.1 Introduction

According to the motivation of the present thesis, background and related work in feature extraction and Statistical Learning is introduced in this chapter. Section 2.2 describes the background of trademark/logo image retrieval, wherein the colour and spatial feature extraction methods are highlighted. Section 2.3 surveys the feature extraction methods in saliency detection and discusses the existing works comprehensively. Section 2.4 describes the development of background subtraction for object detection and tracking. Finally, a brief summary is given in Section 2.5.

2.2 Trademark/logo Image Retrieval

Content-based image retrieval (CBIR), as seen today, has drawn massive research attention in the last decade and been applied on many applications[31, 46, 47]. It is a technique to search user required image from large image database according to users' demands in the form of a query image. Instead of manually labeling the images by keywords or metadata, CBIR indexes images using visual contents such as colours, shapes, textures and other information that can be derived from the image itself. Therefore, it is considered as one of the most economical and effective solutions for analysis of image-based big data[48].

Some useful image retrieval methods have been proposed in recent years. Anuar et al, proposed a shape descriptor based retrieval system where Zernike moment and edge gradient co-occurrence matrix are used to extract the global and local feature of the logos. In [46], authors use a deep learning model i.e. AlexNet model to extract the rich feature of images for the purposed of precise retrieval. In [49], authors use speeded-up robust features (SURF) and maximally stable extremal region (MSER) algorithm to extract the visual and texture feature of the image, and the retrieval results was very satisfied. Mistry er al [50], multiple features such as colour auto-



(a)Word only mark (b)Device mark (c)Composite mark (d)Complex mark

Fig. 2.1 Examples of four typical types of trademarks

correlogram, colour moments, HSV histogram and Gabor wavelet, etc, are used to extract the hybrid features of the images, and various distance metrics such as Euclidean distance, City block distance, Minkowski distance and Mahalanobis distance, etc are used for retrieval.

With the rapid development of e-business, logos are found to have great commercial and social value. For example, the customer can search for a product by using an uploaded logo image. Companies can also use logo retrieval technologies for logo detection, trademark infringement detection and brand protection [51]. To this end, trademark and logo image retrieval (TLIR), as one subset of CBIR, is of great practical significance.

As shown in Fig.2.1, trademarks can be classified into four different types, i.e. the word-only mark, the device mark, the composite-mark and the complex mark [52]. For a word-only mark, it surely includes the text words or phrases. However, a device mark contains only icons or some graphic shapes. The composite mark is a

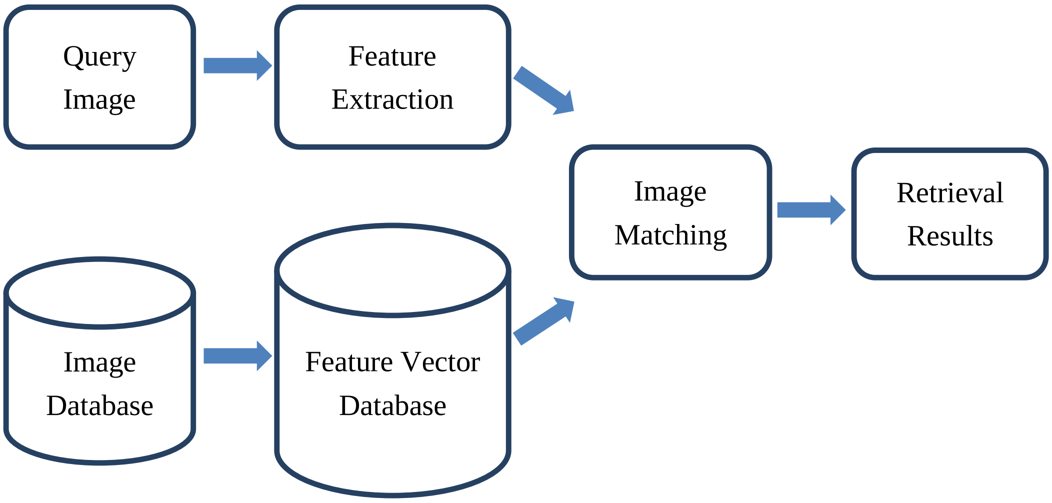


Fig. 2.2 Basic CBIR system.

combination of the previous two. Finally, the complex mark is a composite mark which may contain various added visual effects. In this work, we focus on using CBIR methods to improve the matching and retrieval of all these kinds of trademarks.

Generally, the most widely used feature in CBIR is colour, shape texture etc. The feature vector is generated by these features which are used to represent the main content of each image in the image database and the retrieval of the image is based on the similarities in their contents[53]. Fig.2.2 shows the workflow of the basic CBIR system, where the visual contents of the images in the database are generated by feature extraction methods and form as a feature vector database. The retrieval process in CBIR is that the user inputs the query image into the system first, and the main visual content of the query image will be converted into a set of feature vectors by the same feature extraction methods which were used to generate the feature vector database. In image matching module, the similarity is calculated between the feature vector of the query image and target images in the database. Eventually, the retrieval results will show the top rank of target images based on the similarity scores.

2.2.1 Colour-Based feature

Colour, as one of the low-level visual features, is widely used in CBIR [54-56]. It is one of the most important attributes for image matching and retrieval due to its invariance to image size and orientation. Several kinds of colour descriptors have been proposed:

1) Colour Histogram [57, 58]

   In CBIR, the colour feature is one of the basic features of the image, and it is also easily observed and extracted. The colour histogram is one of the colour descriptors. It can measure the probability of intensities of the three colour channels[59] by counting the pixels' occurrence in the image. The advantage of colour histogram includes low computation cost and not sensitive to the rotation and translation of the image [60, 61], and its disadvantage is lack spatial information.

2) Colour Moment [62, 63]

Colour moment provides an assumption that the distribution of colour in an image can be interpreted as a probability distribution[63]. Therefore, if the colour in an image has a certain probability distribution, the moments of that distribution can be used to represent that image. Three central moments are widely used i.e. mean, standard deviation and skewness of colour distribution in the image[62].

3) Colour Coherence vector[64]

   Colour coherence vector classifies the pixels by given colour bins as either coherent or incoherent. If the pixel is part of a large group of pixels of the same colour, it is a coherent pixel and vice versa. A pixel group is composed of connected pixels where any two pixels are adjacent to each other. Finally, the vector contains the information of colour group and the number of its coherent pixels and incoherent pixels.

4) Colour Correlogram [65, 66]

   Colour correlogram can describe the spatial correlation of colour in an image. It quantizes an image into a number of colours and calculate the spatial correlation of pairs of colours changes with the distance. It mainly has four advantages: (i) low computation cost, (ii) both colour and spatial information are included in the feature, (iii) small feature size and (iv) providing both global colour distribution and local colour correlation.

5) Dominant Colour descriptor (DCD) [67-70]

   DCD can describe a compact and effective colour representation in an image and provide the salient colour distribution in a ROI or an image. Due to its efficiency, it has been widely used for retrieving similar images from database and browsing of image database based on single or various colour values.

For most chromatic logo images, the most salient characteristic is a dominant colour which has two main components, i.e. the representative colours and the percentage of each colour. As a result, dominant colour descriptors (DCD) are considered to be one of the most suitable features and have been widely used in colour based CBIR applications [67-70]. Actually, DCD is one of the descriptors in

MPEG-7, as it can describe the representative colour distributions in an image or a region of interest in an effective and compact form.

In [71], an early concept of dominant colour extraction is proposed for human visual perception. The representative dominant colours of a quantization image are extracted by the available codebook. However, the colour codebook design for a natural image database is a time-consuming and challenging problem, as a fixed colour codebook makes a histogram comparison sensitive to quantization boundaries. Another widely used k-means based clustering approach is the generalized Lloyd algorithm (GLA) [72], however, it is unsuitable for the special case of data clustering in colour quantization [69]. In addition to its high computational cost, one main reason is that the GLA needs to predetermine the number of clusters and initial seeds, where these parameters vary under different experimental settings when the dominant colours are extracted. In a GLA, the salient colours are closely linked to colour distributions. Therefore, the most representative colours are located in the higher colour distribution range with smaller colour distance. However, this may not meet human visual perception as human eyes cannot distinguish colour with close distance. To solve this problem, Yang et al [69], presented a colour quantization method for dominant colour extraction, namely the linear block algorithm (LBA). It has been shown that the LBA significantly improves the efficiency of dominant colour extraction. Lu et al [54], also proposed a method to extract the global characteristic of an image based on the colour distribution and bitmap. As the local feature, the image bitmap presents the local characteristics of the image with the aim of improving retrieval accuracy. H. Shao et al, [67] proposed an image retrieval system framework based on MPEG-7 DCD in the HSV colour space. It extracts representative colour information by using a non-interval quantization algorithm, and a histogram intersection is applied to measure similarity.

2.2.2 Shape and spatial features

Usually, the shape descriptors carry semantic information, and shape extraction techniques can be categorized into contour-based and region-based approaches [73].

The former extracts feature based on boundary information of the region, for instance, the pixels along the object's boundary. In latter techniques, all the pixels within a shape region are taken into account for shape representation, rather than only considering the boundary information. Ling et al [74], proposed an advanced shape descriptor, which proved very good for context-sensitive shape retrieval. However, it suffers from computational complexity which means it can hardly be applied in online retrieval.

Many methods are established based on contour information such as the Fourier descriptor (FD)[75], curvature scale space (CSS)[76], edge direction histogram[77], wavelet descriptor[78], etc.

Global region-based approaches are also widely used in shape-based image retrieval studies [78-81]. Among these, the Zernike moment (ZM) has shown good performance in many related studies [79, 80, 82]. It outperforms many other shape descriptors in terms of compactness, robustness, accuracy, computation complexity and hierarchical representation. However, it can only depict global shape properties. In this case, most trademark retrieval studies use the integration of global and local descriptors to represent shape [77, 79, 83]. In these studies, the researchers use ZM as a global feature and contour-based shape descriptor to extract local features. It is proved that their proposed method gave good results in their experiment. However, these methods are not robust to different noises. Section V will give a detailed analysis on this.

In addition, spatial location is also useful to extract the regional information in an image. Region centroid and its minimum bounding rectangle are the most conventional features to provide the spatial location information[84, 85]. Although it is not sufficient to represent the semantic content of images, it can be very supportive for other feature extraction methods such as colour or shape-based descriptors.

2.2.3 Similarity measure

In CBIR systems, a similarity measurement is used to calculate the similarity between the query image and target images in the database in terms of the feature vectors. Then the retrieval results will be ranked based on similarity value, and the most similar target images will be at the top rank. To calculate the similarity value, most researchers employ the Minkowski-type metric to define the distance[86]. Assume we have two feature vectors of two images $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_n)$. The Minkowski metric is defined as

$$D(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^r)^{1/r}$$

When r equals to 1, it becomes to the Manhattan distance. When r equals 2, it is the Euclidean distance.

Generally, there are two ways to measure the similarity of two images.

(1) Image-Image match: This means the similarity of two images is calculated by their extracted global feature vectors. Then the overall similarity is defined as the weighted sum of the similarity of different types of feature descriptors between two images. Many works [69, 70, 77]use this way since it has high effectiveness and low computation cost.

(2) Region-Region match: Every image is split into several regions link with several feature vectors. Each region of the query image is only associated with the most similar region in the target image [87], then the overall similarity is weighted sum of the similarity between each region in the query image and its most similar region in the target image, and the value of weight is bound up with region size.

2.2.4 Feature Matching Strategy

To better extract the feature of the images in a retrieval system, instead of using single feature descriptor, multiple feature descriptors are used for effectiveness. Therefore, the weight of each descriptor becomes the key parameter in feature matching and affects the overall similarity measurement. The most two popular feature matching techniques are the weight-based solution (WBS) proposed by Jain and Vailaya [88] and the two-component solution (TCS) proposed by Wei et al [79]. In WBS, different weights are assigned to each descriptor, and the final dissimilarity value of each feature descriptor is calculated by Euclidean distance. In TCS, for each descriptor, the Euclidean distance is also used to calculate the final dissimilarity value and if the corresponding value is larger than the threshold value, a penalty value 1 is added to its current dissimilarity value. The total dissimilarity is summarized by the dissimilarity of each descriptor. However, the crucial problem for these two matching techniques is that the weight value in WBS and the threshold value in TCS have to be pre-defined. In previous work, the appropriate value for the database was determined, and so there was a good performance from the retrieval result. But if the database is changed, the result might be not ideal, because it is hard to obtain these values empirically [83]. In order to make some improvements to the matching strategy, Anuar et al [77] proposed a novel retrieval technique which split the matching process into two stages. In the first stage, the dissimilarity value D_g of the global descriptors is computed and an average global dissimilarity value is set as the threshold value. If the dissimilarity value is higher than the threshold value, the corresponding images are not further considered in the second stage. In the second matching stage, the dissimilarity value D_l of the local descriptor is computed and the total dissimilarity value is summarized by $D_g * \omega_g$ and $D_l * \omega_l$. In that study, the weight value ω_g and ω_l is set as 0.2 and 0.8, respectively. Nevertheless, the weighting value is still based on empirical evidence, just as with the two matching techniques mentioned above.

To solve the problem mentioned above, we propose a novel TLIR technique that integrates existing colour and shape based spatial descriptors together with an adaptive feature matching strategy. A dominant colour descriptor (DCD) is selected as the global feature detector and a shape-based spatial descriptor is selected as the local feature detector. In addition, K-means clustering is applied to optimize colour quantization. For low-quality logo images, the integration of DCD and K-means clustering gives better quantization results and results in clearly enhanced colour regions for precise extraction using the spatial descriptor. In addition, an enhanced feature matching strategy is also proposed. Unlike other retrieval systems, an adaptively determined weight is used to balance the significance between colour and shape features, where a fuzzy-based histogram analysis technique is proposed to calculate the adaptive weight. The proposed approach has been evaluated using coloured logo images collected online, and the experimental results shown in Section 3.3 have validated its superiority when benchmarked with other state-of-the-art methods.

2.3 Image Saliency Detection for Segmentation and Extraction of Objects

2.3.1 Background

Whenever we watch a video or look at a photo, we always pay attention to the regions or objects that we are interested in. This is a kind of attention mechanism of our human beings in neuroscience. In order to simulate this attention mechanism in computer vision, people start to develop smart algorithms to get the region of interests (ROIs) of an image. In a recent benchmark survey [33], quite a few unsupervised saliency detection methods are summarized and assessed, where the two main objectives of saliency detection are fixation prediction [6, 11] and salient object detection [20, 23, 27]. In fixation prediction, it aims to predict eye's gaze or motion through detecting sparse blob-like salient regions [89], whilst salient object detection is to detect the salient objects/regions in the scene [90]. According to the survey [33], much more salient object detection methods are proposed than those using eye fixation prediction, possibly due to their contributions to a wide range of applications including content-based image retrieval [91-93], image/video compression [94-96], image quality assessment [97-99], region of interest segmentation [24, 100, 101], and object detection [102-104], etc.

For human beings, our visual attention system is mainly made up by both bottom-up and top-down attention mechanisms that enable us to allocate to the most salient stimuli, location, or feature that evokes the stronger neural activation than others in the natural scenes [105-107]. Bottom-up attention helps us gather information from separated feature maps e.g. colour or spatial measurements, which is then incorporated to a global contrast map representing the most salient objects/regions that pop out from their surroundings [108]. Top-down attention modulates the bottom-up attentional signals and helps us voluntarily focus on specific targets/objects i.e. face and cars [109]. However, due to the high level of subjectivity and lack of formal mathematical representation, it is still very challenging for computers to imitate the characteristics of our visual attention mechanisms.

To extract features at the bottom level, colour plays an important role since it is a central component of the human visual system, which also facilitates our capability for scene segmentation and visual memory [110]. Colour is particularly useful for object identification as it is invariant under different viewpoints. We can move or even rotate an object, yet the colour we see seems unchanged due to the light reflected from the object into the retina remains the same. As a result, the salient regions/objects can be easily recognised intuitively for their high contrast to the surrounding background.

In addition to colour features, our visual perception system is also sensitive to spatial signals, as the retinal ganglion cells can transmit the spatial information within natural images to the brain [111]. As a result, our human beings pay more attention to the objects and regions not only with dominant colours but also with close and compact spatial distributions. Therefore, the main objective of saliency detection is to computationally group the perceptual objects on the base of the way how our human visual perception system works.

For modeling top-down attention, Al-Aidroos et al [112] proposed a theory named ‘background connectivity’ to describe the stimulus-evoked response of our visual

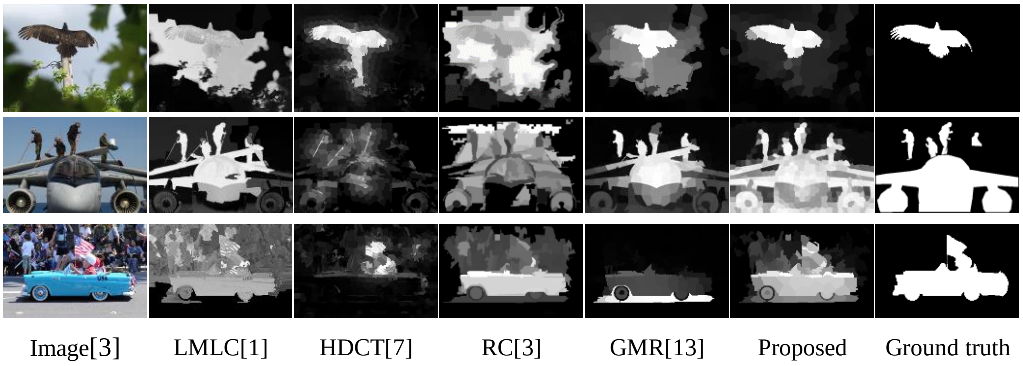


Fig. 2.3. Three examples of salient objects.

cortex. It is found that focus on the scenes rather than objects may increase the background connectivity. Inspired by this theory, we employed a robust background detection model to represent the background connectivity of top-down attention in the images as post-processing to further refine the saliency maps detected using gestalt-laws guided processing. This will be detailed in Section 4.3.4.

Fig.2.3 shows several examples in which the salient objects contain poor colour and/or spatial contrasts. As such, conventional approaches either fails to detect the object as a whole or results in massive false alarms. Within the proposed cognitive framework, salient objects can be successfully detected whilst the false alarms are significantly suppressed. Descriptions of the proposed salient model and its implementation are detailed in Sections 4.3.

2.3.2 Related Work

In the past decades, a number of salient object detection methods have been developed to identify salient regions in terms of the saliency map and capture as much as possible human perceptual attention. Inspired by a biologically plausible architecture [106] and the feature integration theory [113], Itti et al [6] proposed an epic saliency detection model in 1998. With multiple image features extracted including luminance, colour, and edge orientation, the saliency map is generated by using center-surround difference across these features. In the following two decades, quite a few landmark saliency models are developed, which are briefly reviewed below.

In general, saliency detection methods can be categorized into two classes, i.e. supervised and unsupervised approaches. Most unsupervised methods are driven by manually designed models which are mainly based on computing feature-based contrast followed by various mathematical principles and the assumption that salient regions or objects in the visual field would outstand from their surroundings. Meanwhile, some other mechanisms such as geodesic distance [114], minimum barrier distance transform[4], reconstruction error [18], have been adopting to detect salient objects in still images. Most supervised methods, especially deep learning based, have been introduced to image saliency detection recently. These methods [115-117] typically use CNNs to learn the deep feature of the image from which the saliency objects are selected. Comparing with unsupervised methods, deep learning based supervised methods are able to obtain better saliency maps, however high-performance computers even with particular graphics process units (GPU) are needed to cope with the lengthy training time. In addition, supervised methods may also suffer from lack of generality, especially when the training samples are limited and/or insufficiently representative. With deep learning, this drawback seems can be somehow overcome [118], yet at a cost of a large amount of data requested for training to learn prior knowledge. On the contrary, it seems our human vision can easily detect and recognise objects under complex scenes without supervision [119]. To this end, the proposed saliency method in Section 4 focuses mainly on unsupervised saliency detection.

Depending on whether a salient object is detected from pixels or regions, saliency detection techniques can be further categorized into two groups, i.e. pixel-based and region based. Herein the main difference between the two groups is whether the image is segmented into regions for saliency detection, using either colour quantization or pixel clustering. In [9], a contrast-based saliency map is proposed, where the colour difference between the pixel and its neighbours is determined to extract the attended areas using the fuzzy theory [120]. In [11] a biologically plausible bottom-up visual saliency model is presented based on the Markovian approach and mass concentration algorithm. More recently, an efficient method is introduced in Achanta et al. [17] to build high quality saliency maps using low-level features such as luminance and colour in the L*a*b* colour space. In [20], a salient region is detected by using the colour difference between the pixels and their average value in the image, again in the L*a*b* colour space. In [23], instead of treating the whole image as the common surround for any given pixel, the saliency map is defined by using colour difference between the given pixel and a local symmetric surround region. In Cheng et al [27], a global contrast-based method is proposed to determine the saliency value. The colour is quantized into a number of bins in the

L*a*b* colour space with the global colour contrast measured between colour bins. Furthermore, a colour space smoothing process is also introduced to reduce quantization artefacts before assigning similar saliency value to similar colour bins. In [121], the authors adopt a famous CNN model, fully convolutional network [122] (FCN), to detect the salient objects in the video frames. The network takes single frame image as input and transforms the entire frame to multidimensional feature representation through multilayer convolution networks. Then it up-samples the extracted feature by applying a stack of deconvolution networks. Finally, a fully connected layer and sigmoid activity function are used to generate the probability map with the same size of the input frame, in which larger values mean higher saliency values. In [29], the authors employ a stacked denoising autoencoder (SDAE) to learn powerful representation from the raw image data. It first generates multiscale inputs by using five scales of the original image size. The advantage of this progress is that small objects can be detected at a large scale while the inner regions of large objects can be highlighted at a small scale. According to the assumption where four boundaries of the image are background, SDAE based deep learning structure is used to model the image background and construct the residual maps for each scale. After that, these residual maps are integrated together for the final map. Finally, a post-processing stage with image refinement and region smoothing is applied for further improvement.

Although the aforementioned approaches are found to produce relatively good results on saliency detection, their robustness is limited when extending to large datasets due to increasing complexity of the scenes, especially the variations in terms of spatial size and layout between the salient objects and the image background. The reason here is, redundant and tedious information in the pixel makes the extremely high computational cost for pixel-level saliency computation. To this end, region-based contrast and saliency detection has become increasingly popular in recent years, especially using the superpixel based approach. In [27], a region-level saliency map is proposed based on both colour and spatial difference across the regions, where the spatial prior is used to highlight the salient regions. In [123], a contrast-based saliency estimation is proposed, where a given image is segmented into a number of homogeneous regions by using superpixel. The contrast and spatial distribution of these regions are measured and smoothed by using high-dimensional Gaussian filters for saliency detection. In [15], a superpixel based saliency detection method is proposed, where colour and spatial contrast across the superpixels are used for efficient saliency detection. In Kim et al [7], superpixel technique is employed to cluster the image into a number of regions, and high-dimensional colour transform is applied to extract rich feature from the image. Then K-nearest neighbour algorithm and random forest regression are used to estimate the saliency values. In Li's work [28], he proposed a deep contrast network with two streams. In the first stream, the feature of raw image is directly extract by DNN. In the second stream, unlike the first stream, he over-segment the image into a number of regions, and extract the deep regional features by DNN. Then DNN is fine-tuned and well trained. Finally, the saliency maps from two streams are integrated and refined.

In general, the whole process of bottom-up saliency detection can be divided into at least three stages, i.e. pre-processing, feature-based salient map generation, and post-processing. Apparently feature based salient map generation is the key in saliency models, where various colour and spatial features are extracted and measured in determining the saliency maps. The pre-processing is often for spatial and illumination normalization, image enhancement and image segmentation (only for region-based approaches). The post-processing, on the contrary, serves mainly for normalization and/or fusion of saliency maps, where object prior is widely used in region-based approaches. One optional stage is to extract the binary template of the salient object via thresholding the salient map, where histogram based adaptive thresholding such as Otsu's approach [124] is commonly used.

In Table 2.1, some typical unsupervised and supervised saliency detection approaches are summarized for comparison in terms of the features used and any adopted pre-processing and post-processing stages. In unsupervised pixel-level saliency detection methods, colour and spatial information is the most widely used features due to their importance in visual psychology [110, 111]. However, spatial

Table 2.1. Overview of some popular unsupervised saliency detection models

	Method	Pre-processing	Features for initial saliency map generation	Post-processing to refine the saliency map
Pixel-level unsupervised saliency detection	IT[6]	No	Colour, Intensity, Orientation	Normalization of saliency map
	MZ[9]	Image resizing, Colour quantization	Colour	Refinement via Fuzzy growing
	GB[11]	No	Colour, Orientation, Intensity	Normalization of saliency map
	SR[14]	Log spectrum	Intensity	No
	AC[17]	No	Colour, Luminance	Fusion of multi-saliency maps
	FT[20]	Gaussian filter	Colour, Luminance	No
	MSS[23]	No	Colour	No
	SEG[24]	No	Colour	No
	HC[27]	Colour quantization	Colour, Luminance	No
	MBD[4]	Minimum barrier distance transform	Colour, Intensity	Morphology, normalization and refinement.
Region-level unsupervised saliency detection	RC[3]	Colour quantization, Graph-based image segmentation	Colour, Luminance, Spatial	Object prior with colour refinement and hard constraints
	SP[15]	Colour quantization, Superpixel clustering	Colour, Spatial	Colour and spatial refinement
	LMLC[1]	Superpixel clustering	Colour, Spatial, Intensity	Colour refinement
	GMR[13]	Superpixel clustering	Colour	Manifold ranking
	DSR[18]	Superpixel clustering	Colour, spatial	Dense and sparse reconstruction
	RBD[22]	Superpixel clustering	Colour, spatial	Cost function
Supervised saliency detection	HDCT[7]	Superpixel clustering	Location, Colour, Texture, Shape	Object prior via Spatial refinement
	MDF[26]	Non-overlapping segmentation	Convolutional neural network	Multi-scale fusion
	DCL[28]	Superpixel clustering	Convolutional neural network	Fully connected conditional random field
	DRR[29]	Multi-scale transform	Deep reconstruction residual, stacked denosing autoencoder	Multi-scale fusion, region smoothing

features are excluded in some pixel-based approaches. In unsupervised region-based approaches, colour quantization and graph or superpixel based clustering is normally used. Combination of colour and spatial features are then employed for saliency map determination and refinement. In supervised approaches, region clustering, and multi-scale transform are generally used as pre-processing. After that, unlike unsupervised method, deep features are extracted by deep learning model such as CNN and SDAE. Last but not least, multi-scale image fusion or some optimization model help to further improve the final saliency map.

2.4 Multi-modal Sensor Fusion for Pedestrian Detection and Tracking

In the past decades, detection and tracking of video objects have drawn massive research attention and has always been a major task in the computer vision field [34, 125, 126]. As one subset of video object tracking, pedestrian detection and tracking have been applied to many applications such as visual surveillance [32, 127], driver-assistance systems [128, 129], human activity recognition[130, 131], etc. For pedestrian detection and tracking, visible camera and thermal imagery are two popularly used sources of image modalities, though not necessarily in a combined solution [132-134]. However, either visible image or thermal image has their advantages and disadvantages. Visible image can show detailed colour information, however it really suffers from lighting variations, cluttered backgrounds, artificial appearances i.e. shadows and etc. Since the object is detected by its temperature and radiated heat, thermal image can eliminate the influence of colour and illumination changes on the objects' appearance [135] in any weather conditions and at both day and night time. However, in some cases, e.g., occlusions, the thermal camera may fail to detect the object properly. In Fig.2.4, there are three pedestrian templates, for the one with a yellow rectangle; both visible and thermal image can detect it very well since it has high contrast to the background in the visible domain and human temperature in the thermal domain. For the template in the red rectangle, it has a compact shape in the thermal image. However, in the visible image, we can just identify it coarsely due to the similar appearance in colour of the background and the person's cloth. The one in green rectangle can be seen in the visible image but hardly observed in the corresponding thermal image. This is because thermography is only able to directly detect surface temperatures, and it cannot work well when the object is (partially) occluded. Moreover, it will detect any objects (e.g. windows and car in Fig.2.4) with surface temperature.

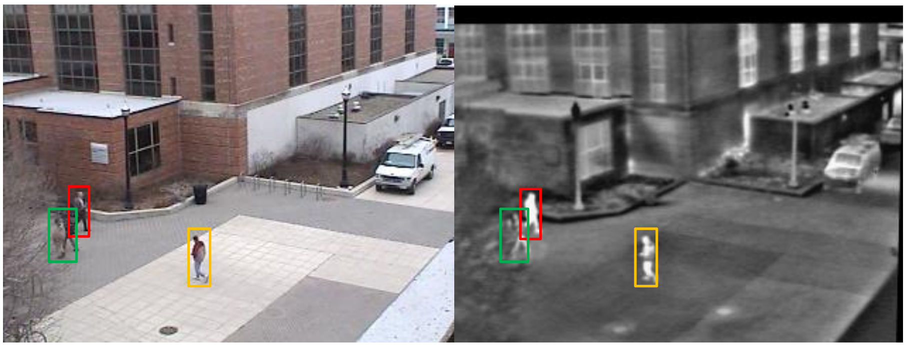


Fig. 2.4 Visible image of a scene (left) and thermal image of the same scene (right) [189].

2.4.1 *Background subtraction*

For the purpose of pedestrian detection, the basic operation is to separate the moving objects called “foreground” from the static scene called “background” in the video sequences. To achieve this, background subtraction plays an important role in it. The simplest way to build the background is to get a background image without any moving objects, and then find the difference between each frame and background image to detect the foreground. However, it is not easy to achieve in some cases due to such challenges as bad weather, illumination changes, shadow, dynamic background, etc[136]. Therefore, the model of background representation must be more robust and adaptive.

In 1997, Wren et al. [137] first proposed a background model based on a Gaussian function. However, single Gaussian distribution can hardly tackle with dynamic background in real scene due to low frame rate. To this end, Grimson and Stauffer[138] proposed the famous Gaussian mixture model (GMM) in where each pixel location has a mixture of *K* Gaussian distributions. After that, they initialized the parameter in the model by an online EM-based method[139] which improved the performance of GMM but increase the computation cost. Then several works had been proposed to improve the accuracy and reduce the computational time. The author in [140] proposed an adaptive GMM to make the parameters of GMM dynamically. In Lee’s work [141], adaptive learning rate is used to increase the convergence speed without changing the performance of GMM. Shimada [142]controlled the Gaussian mixture model by dynamic Gaussian component.

Except GMM methods, a large number of other mechanism-based background subtraction models have been proposed and applied on many different applications in recent year. Andrew et al[2] proposed a single-camera statistical segmentation algorithm where a combination of statistical background image estimation and Bayesian-based segmentation are used to achieve foreground detection. Domenico and Luca[5] proposed a fast background subtraction method based on a clustering algorithm with a condition based mechanism. Zhao et al[16] proposed a background modelling method for motion detection in dynamic scenes based on type-2 fuzzy Gaussian mixture model[143] and Markov random field (MRF) [144]. In[10], authors introduced a background subtraction framework based on texture feature. Furthermore, colour cues are clustered by the codebook scheme in order to refine the texture- based detection. Pierre-Luc points out in [8] that most background subtraction methods do not pay attention to the spatial or spatiotemporal relationship of each analysed pixel, and also suffer in complexity, computation cost, and versatility. Therefore, he proposed a spatiotemporal based background subtraction algorithm which has been proved low-cost and highly efficient. In addition, he also proposed another one using spatiotemporal feature descriptors in [12] in order to build an adaptive and flexible model rather than tuning parameters in different scenarios for optimal performance. In [145], authors proposed a background subtraction model based on optical flows to detect the multiple moving objects under complex outdoor scenes. [146] proposed a novel model named CLASS (collaborative low-rank and sparse separation) for moving object detection as well. Wang et al [147] presented a PTZ(pan-tilt-zoom) control model to detect the pedestrians. In [148], authors proposed a robust principal component analysis method to separate the foreground and background. It is robust to camera movement and dynamic backgrounds. In addition, it also proposed an efficient initialization method for low-rank and sparse matrics, which improve the separation of the moving objects. In [149], authors proposed a low rank and sparse representation based method where a new formulation of foreground detection in the low-rank representation is proposed to estimate the sparse outliers(foreground).

2.4.2 Materials

Many surveys and comparative studies have been published to evaluate these background subtraction methods [136, 150-152]. Both [152] and [150] have done a comprehensive evaluation on 29 methods implemented in the BGS Library [153](background subtraction library). And many background subtraction datasets have been published and are fully available online, such as ChangeDetection.net(CDnet2014)[154], BMC (Background Models Challenge)[155] and OCTBVS[1], etc. CDnet2014 database contains 11 video categories such as bad weather, thermal, shadow, object motion, etc. Each category has 4 to 6 video sequences, and most frames in the videos were annotated for obtaining the ground truth foreground and background. BMC database is composed of both synthetic and real videos. 20 urban video sequences rendered with the SiVIC simulator[156] is contained in the data, where two scenes: street and rotary, and five event types: cloudy w/o noise, sunny with noise, foggy with noise and wind with noise, are included. OCTBVS contains videos and images recorded in and beyond the visible spectrum (e.g., infrared). It currently has 11 sub-datasets, some of them are composed of infrared images, some of them contains thermal images, and some of them contains both thermal and visible images.

2.5 Summary

This chapter comprises the background and related work on three main works where feature extraction methods play important role in. In the first work, the concept of logo/trademark image retrieval is described where both colour and spatial features are the key features and retrieval results can be generated by using feature matching strategy.

In the second work, the development of saliency detection has been introduced, presenting a classic saliency detection chain with 3 stages: pre-processing, feature extraction and post-processing. Pre-processing is normally addressed by existing colour quantization and region clustering techniques. Therefore, most researchers

[1] http://vcipl-okstate.org/pbvs/bench/

mainly focus on the feature extraction stage, where many unsupervised and supervised methods have been fully explored. In post-processing stage, many useful refinements or optimized methods have been developed and help improve the final saliency map.

In third work, a review of object detection and tracking in the video is carried out, where two core techniques, background subtraction, and object tracking, are introduced. These have motivated the work as presented in this thesis and reported in detail in the next three chapters.

CHAPTER 3

3 Adaptive fusion of colour and spatial descriptors for trademark/logo image retrieval

3.1 Introduction

Due to their uniqueness and high value commercially, logos/trademarks play a key role in e-business based global marketing. However, existing trademark/logo retrieval techniques and content-based image retrieval methods are mostly designed for generic images, which cannot provide effective retrieval of trademarks/logos. Although colour and spatial features have been intensively investigated for logo image retrieval, in most cases they were applied separately. When these are combined in a fused manner, a fixed weighting is normally used between them which cannot reflect the significance of these features in the images. When the image quality is degraded by various reasons such as noise, the reliability of colour and spatial features may change in different ways, such that the weights between them should be adapted to such changes. In this chapter, an adaptive fusion of colour and spatial descriptors is proposed for coloured logo/trademark image retrieval. First, colour quantization and k-Means are combined for effective dominant colour extraction. For each extracted dominant colour, a component-based spatial descriptor is derived for local features. By analyzing the image histogram, an adaptive fusion of these two features is achieved for more effective logo abstraction and more accurate image retrieval. The proposed approach has been tested on a database containing over 2300 logo/trademark images. Experimental results have shown that the proposed methodology yields improved retrieval precision and outperform three state-of-the-art techniques even with added Gaussian, salt and pepper, and speckle noise.

The rest of the chapter is organised as follows: Section 3.2 details the proposed method in terms of colour feature extraction, spatial feature extraction, and feature matching strategy. Section 3.3 presents the comprehensive experimental results and discussed deeply. Finally, some concluding remarks and future work are summarized in Section 3.4.

3.2 Proposed Methods

The performance of image retrieval relies on the quality of original images, which can be affected by some disturbance, compression, and noise especially when images are transmitted using mobile devices. For a logo image degraded by noise or distortion, the customer may recognise it, but not the computer algorithm. Therefore, it is necessary to develop a Trademark/logo image retrieval(TLIR) system which is robust enough to such noise.

In our proposed approach, as shown in Fig.3.1 below, the workflow includes feature extraction, feature matching, adaptive fusion and content-based retrieval, where the feature extraction stage in the proposed algorithm involves image pre-processing and feature representation. In image pre-processing part, each image is quantized into a maximum of eight salient colours in order to extract the spatial layout information in each colour component. The feature representation part is used to extract a set of feature vectors of the images from their corresponding colour metrics, spatial descriptors, and adaptive coefficient. The descriptions of the relevant parts in detail are discussed in the following.

In the adaptive fusion stage, the adaptive coefficient (detailed in Section 3.2.3) for each image is calculated by fuzzy-based histogram analysis. The coefficient is used

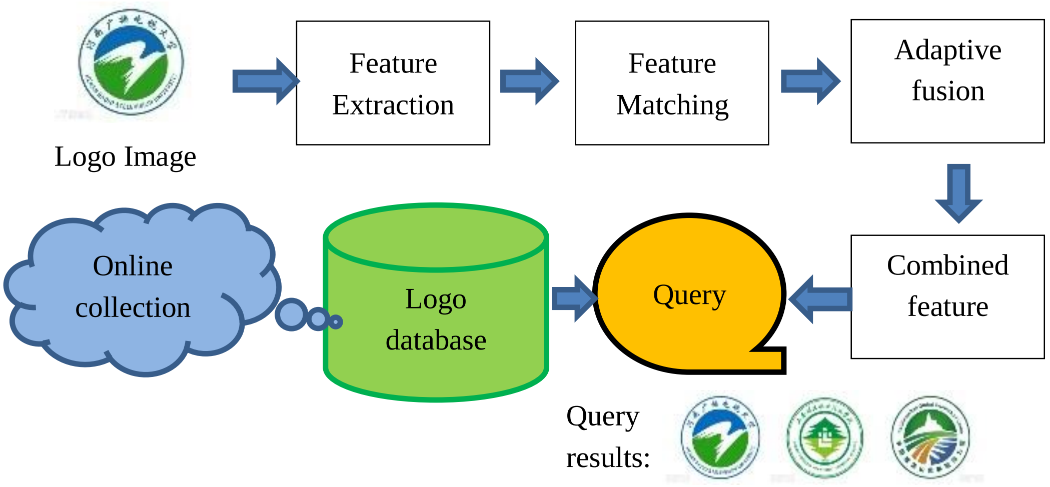


Fig. 3.1 The workflow of the proposed trademark image retrieval system

to balance the weight between the two features mentioned previously in order to optimize the matching and retrieval performance.

The rest of this Section is organised as follows. Section 3.2.1 introduces the method for colour feature extraction. Section 3.2.2 presents the local spatial layout feature and Section 3.2.3 elaborates the adaptive matching strategy to balance between colour and shape features.

3.2.1 Colour Feature Extraction

In this Section, a detailed description of the feature extraction methods is presented to determine the dominant colours from the noise-degraded images.

3.2.1.1 Global Colour descriptor

In general, a colour logo image has several salient colours. In order to extract these colours rapidly, the dominant colour descriptor is considered as the most efficient method. Linear block algorithm (LBA) [69] is a very efficient colour quantization method and its computation cost is also very low. Given an image I, the RGB colour space is uniformly divided into *i* coarse partitions with each separate colour component divided into two parts.

For each partition, the quantized colour is given by $C_i = (\bar{x}_i^R, \bar{x}_i^G, \bar{x}_i^B), i \in [1, 8]$, where $\bar{x}_i$ is the average value of distribution of the three colour components *red*, *green* and *blue*, for each partition center. Afterwards, the mutual distance of two adjacent C_i is calculated and similar “colour bins” are merged together using a weighted average agglomerative procedure as follows:

$$\begin{cases} x^R = x_1^R \times \left(\dfrac{p_{R,1}}{p_{R,1} + p_{R,2}}\right) + x_2^R \times \left(\dfrac{p_{R,2}}{p_{R,1} + p_{R,2}}\right) \\ x^G = x_1^G \times \left(\dfrac{p_{G,1}}{p_{G,1} + p_{G,2}}\right) + x_2^G \times \left(\dfrac{p_{G,2}}{p_{G,1} + p_{G,2}}\right) \\ x^B = x_1^B \times \left(\dfrac{p_{B,1}}{p_{B,1} + p_{B,2}}\right) + x_2^B \times \left(\dfrac{p_{B,2}}{p_{B,1} + p_{B,2}}\right) \end{cases} \tag{3.1}$$

In Eq. (3.1), p_R, p_G, p_B represent respectively the percentage of R, G, B components in the RGB image. The merge processes iterate until the minimum Euclidean distance between the adjacent “colour bins” center are larger than the

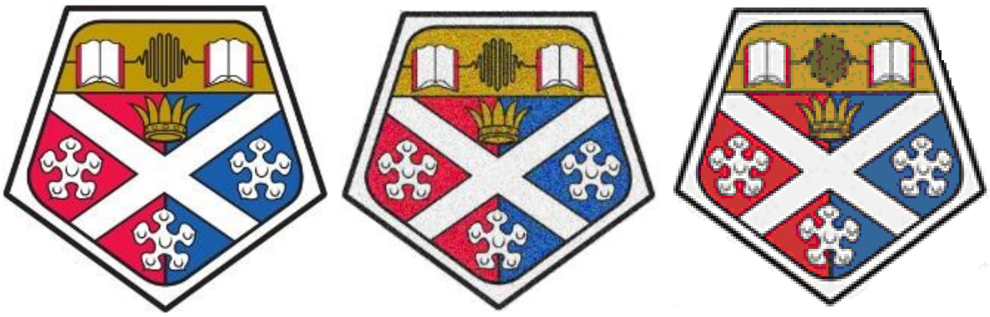


Fig. 3.2 One original logo image (left), its degraded version (middle) and result of colour quantisation (right)

threshold T_d. As the dominant colours should be significant enough, if the percentage of surviving colour is less than a threshold T_m, it will be merged in to nearest colour. In this study, as suggested in[69], we choose $T_d = 15$ and $T_m = 6\%$.

Finally, the LBA can quickly return the maximum eight dominant colour descriptors (DCDs) and DCD in MPEG-7 is defined as $F_C = (C_i, P_i), i \in [1, M]$, where C_i is the 3D dominant colour vectors, P_i is the percentage for each dominant colour and the sum of P_i equals 1, M is the total number of dominant colour for an image. Fig. 3.2 shows the logo image of the University of Strathclyde. In this image, there are five colours and the quantization result generated by LBA exactly meets that of human vision (Table 3.1).

However, if the image has poor resolution or noise, the LBA cannot give a proper quantization result as the noise affects both 3D dominant colour vectors and the percentage of each DC. In order to improve the robustness of image quality for colour quantization, the integration of LBA and K-means clustering algorithm is proposed. Since the Manhattan distance (L1) is known to be robust to noise [157-159], the L1 distance is used to compute the distance between the centroid and the points in the clusters.

Although the K-means algorithm is inarguably one of the most widely used methods for data clustering [160], its main disadvantages are that both K and the initial clustering centroids are hard to decide, whereas, the LBA can resolve these negatives, using its output including the number of dominant colours and 3D Dominant colour vectors. In addition, the K-means algorithm iteratively assigns a sample to the nearest center and each center is recalculated as the mean of all

Table 3.1 Dominant colour feature of Fig.3.2

DCs	Red	Green	Blue	Percentage (%)
BLACK	46	38	35	17.23
RED	221	26	71	8.1
BLUE	29	93	169	6.96
YELLOW	188	152	15	8.11
WHITE	251	251	251	59.61

samples assigned to it. Therefore, it compensates for the lack of colour quantization by the LBA. From Fig. 3.3, it can be seen that the quantization results of the integration of LBA and the K-means on a low-quality image gives better quantization performance than human vision. The peak signal to noise ratio (PSNR), and the mean square error (MSE) are also used to evaluate the quantization performance. In general, a higher PSNR value results in a lower MSE, i.e. a better quality of the quantized image.

Table 3.2 MSE and PSNR comparison of two quantization methods

Method	LBA		LBA_Kmeans	
Gaussian variance	PSNR (dB)	MSE	PSNR (dB)	MSE
0.01	20.83	672.06	21.51	608.96
0.02	20.85	673.63	21.49	615.78
0.03	20.80	683.04	21.41	628.25
0.04	20.71	695.99	21.30	643.37
0.05	20.58	713.98	21.14	664.71
0.06	20.40	738.77	20.94	691.42
0.07	20.21	768.19	20.73	720.47
0.08	20.00	798.05	20.51	749.31
0.09	19.76	835.15	20.26	784.89
0.1	19.50	880.48	19.98	830.10
0.2	17.01	1497.19	17.50	1399.36
0.3	14.80	2513.05	15.32	2290.77
0.4	12.96	3966.53	13.36	3664.60

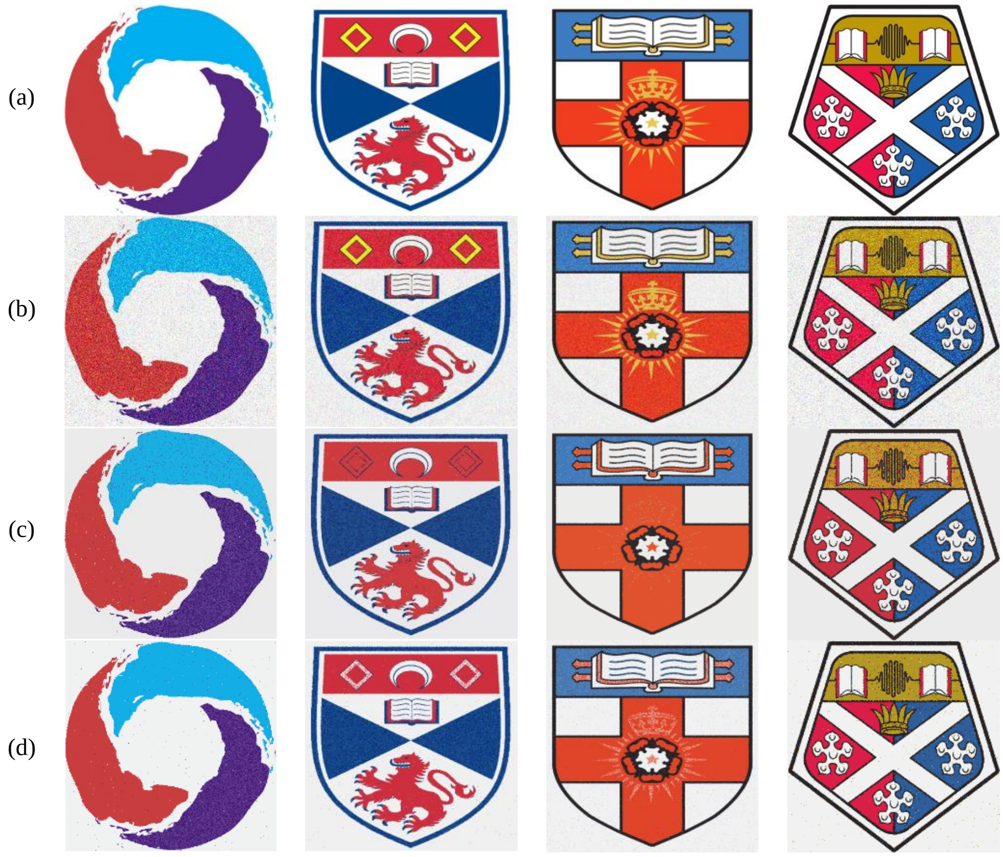


Fig. 3.3. Original example logo image(a). Logo image with Gaussian noise(b). Quantization results from using LBA method(c). Quantization results using combination of LBA and the k-means(d).

The MSE and PSNR between two images are given by:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2 \tag{3.2}$$

$$PSNR = 10\log_{10}(\frac{L^2}{MSE}) \tag{3.3}$$

where *N* is the total number of pixels in the image and *L* is the maximum dynamic range. For 8 bpp images, *L*=255. X_i is quantized image and Y_i is the reference image.

Table 3.2 shows the average PSNR and MSE value with several Gaussian variance levels evaluated on our logo image database. It shows that KLBA (the combination of LBA and K-means) gives better quantization results.

3.2.1.2 Colour dissimilarity

The DCD descriptor matching strategy is based on Yang [69]. The distance between images I_1 and I_2 is defined as:

$$Dis\_C(I_1, I_2) = 1 - SIM(I_1, I_2) \tag{3.4}$$

where $SIM(I_1, I_2) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} a_{i,j} S_{i,j}$, N_1 and N_2 are the number of dominant colours in I_1 and I_2, respectively. The similarity score between two representative colours is given by:

$$S_{i,j} = \left[1 - \left|p_q(i) - p_t(j)\right|\right] \times \min(p_q(i), p_t(j)) \tag{3.5}$$

Here, $p_q(i)$ and $p_t(j)$ are the percentage of the ith and jth dominant colour in query image and the target image, respectively. The term in the bracket, $1 - \left|p_q(i) - p_t(j)\right|$, is used to measure the difference of two colours in their percentage. The $\min(p_q(i), p_t(j))$ is the intersection of $p_q(i)$ and $p_t(j)$, which represents the similarity between two colours in percentage. If $p_q(i)$ equals to $p_t(j)$, then their percentage is same and the colour similarity is determined by $\min(p_q(i), p_t(j))$, otherwise, a large difference between $p_q(i)$ and $p_t(j)$ will decrease the similarity measure.

The similarity coefficient is given by:

$$a_{i,j} = \begin{cases} 1 - d_{i,j}/d_{max} & (d_{i,j} \leq T_d) \\ 0 & otherwise \end{cases} \tag{3.6}$$

$$d_{i,j} = \left\| c_i - b_j \right\| \tag{3.7}$$

where $d_{i,j}$ is the Euclidean distance between two colour clusters c_i and b_j. T_d is the threshold to judge if two colour features are similar, and $d_{max} = \alpha T_d$. The

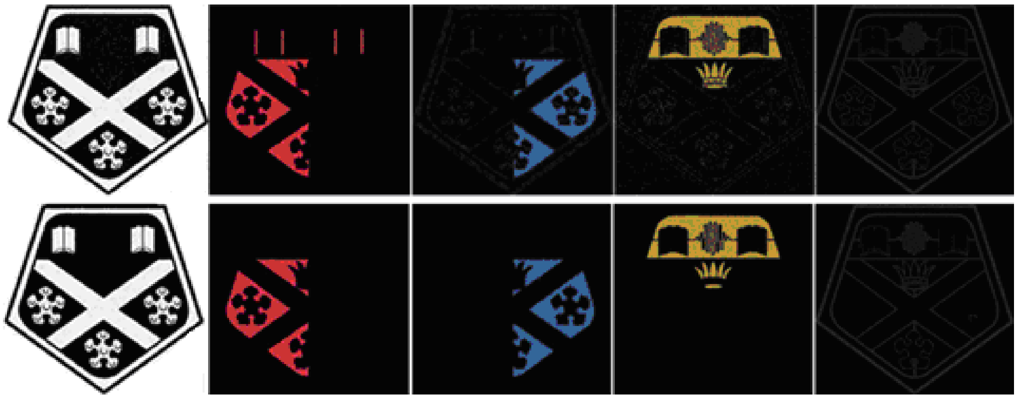


Fig. 3.4 Detailed comparison of layered components for the quantized image shown in Fig. 3.3, where the top and bottom rows are results from image segmentation and population and density filter, respectively

parameters α and T_d are set to 2 and 25 [69], respectively.

3.2.2 Spatial Feature Extraction

Due to the noise and degradation of the images, it is found that colour descriptor fails to precisely represent the dominant colours. As a result, spatial layout information is needed.

3.2.2.1 Local spatial layout descriptor

The spatial location in an image represents the spatial distribution information of the colours in the image. Kankanhalli [161] proposed a useful spatial descriptor, and the steps of the spatial information extraction are listed as follows:

i. Extracting dominant colours to split the image into several layers where each layer corresponds to one dominant colour (in Fig. 3.4).

ii. In each colour layer, every colour block is treated as spatial clusters. For example, there are three spatial clusters in red colour layer in Fig.3.4.

iii. For each cluster, let P is the total number of pixels of the image, define a lower threshold $T_l = P \cdot 0.03125\%$ and an upper threshold $T_u = P \cdot 0.625\%$. Also, define a density $\rho = \frac{N}{{l_{max}}^2}$, where N is the population of the cluster, and , l_{max} is the maximum length of the bounding box of the cluster [161].

iv. Discard clusters whose population is less than the lower threshold. For clusters whose population is less than the upper threshold but more than the lower threshold, as suggested in [161], eliminate them if their density ρ is less than 10%.

v. Find the centroid location from survived spatial clusters.

Then the colour-spatial feature vector is defined as:

$$V = \{C_i, P_i, x_{ij,}y_{ij}\}, (i\epsilon[1, I], j\epsilon[1, J]) \tag{3.8}$$

where C_i and P_i are the dominant colour and their corresponding percentage respectively; (x_{ij}, y_{ij}) is the centroid of the jth spatial cluster in the ith colour layer. I is the number of DC, and J is the number of the spatial cluster in a certain colour layer.

3.2.2.2 Spatial dissimilarity

Assume that the image I_q has i spatial clusters in the m^{th} dominant colour layer and the target image I_t has j spatial clusters in the n^{th} dominant colour layers. If the m^{th} dominant colour in I_q is similar to the n^{th} dominant colour in I_t, then the assignment function F maps every spatial cluster s in I_q to the closest colour cluster F_s in I_t .This is a 1-to-1 mapping and F is determined as follows.

i. Form the distance matrix, D_{ij}

$$D_{ij} = \left[d_{ij}\right]_{i\times j} \tag{3.9}$$

ii. Find the minimum entry (a,b) in D_{ij} with F(a)=b,

iii. Cut off the row a and column b of the matrix D_{ij}. If the matrix D_{ij} is non-degenerate, go to step 2 else stop. A matrix is considered degenerate when the number of rows or columns which are not yet struck off is zero.

For every matched spatial cluster, we use the spatial centroid for distance measure. The spatial distance *d* of the similar colour layer *m* is defined as:

$$d(m) = \frac{\sum_{k=1}^{\min(i_m,j_m)} \sqrt{\left(x_{k,m}^{Q} - x_{F_S(k),m}^{T}\right)^2 + \left(y_{k,m}^{Q} - y_{F_S(k),m}^{T}\right)^2}}{\min(i_m, j_m)} \tag{3.10}$$

and the final spatial distance of two images is averagely summed by $d(m)$.

$$Dis\_S = \frac{1}{M}\sum_{m=1}^{M} d(m) \tag{3.11}$$

where *M* is the number of similar colours between two images.

3.2.3 Adaptive Fusion for Feature Matching

In the proposed retrieval system, the first decision is whether two images have similar colour content. If not, this pair of images is considered as different. Otherwise, the system measures spatial feature difference for every similar colour layer. The overall dissimilarity is defined by

$$D = (1 - \gamma) \times Dis\_S + \gamma \times Dis\_C \tag{3.12}$$

where γ is used to weighting colour and spatial based features, Dis_S is the spatial dissimilarity and Dis_C is the colour dissimilarity.

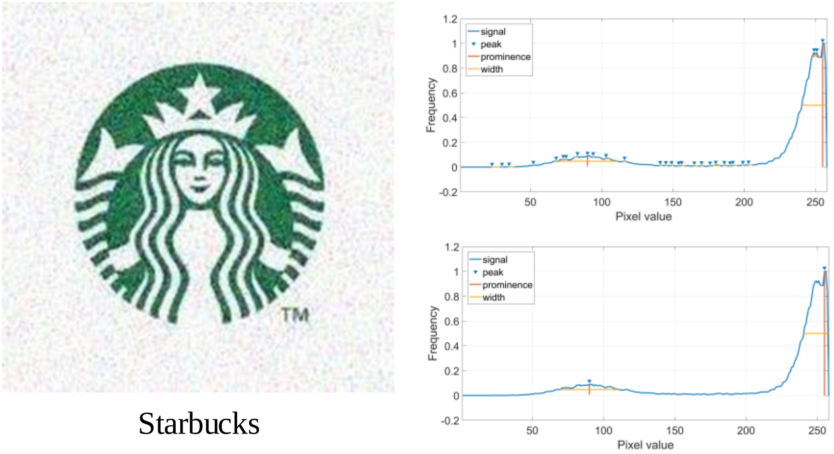

Starbucks

Fig. 3.5 The logo image with Gaussian noise (left), the peak point illustration without any constraint (upper right, with over 20 candidates), the peak point illustration with constraints (low right, two main peaks)

It is worth noting that the parameter γ is adaptively determined based on the reliability of the colour features in the query image. For a logo image, its dominant colours tend to produce peaks in the corresponding histogram. When the logo image is severely degraded, the colour features become less reliable and result in insignificant peaks. As a result, the reliability of colour features can be approximately determined by identifying the significance of the peaks in the corresponding image histogram as follows.

Convert the coloured logo image into grayscale and calculate its histogram with the frequency normalised into [0 1].

Apply window-based filtering on the 1D histogram vectors to remove noise and enhance the correct peak locations. The filter window is set as 3 for simplicity. The distance between each pair of adjacent peak points is required to be no less than T_d and the prominence of each peak point is no less than T_m , where $T_d = 15$ and $T_m = 6\%$ as defined in Section III-C. With these constraints, most noise-caused peak points are successfully removed, and the correct peak points are extracted (Fig.3.5).

For the i^{th} peak point, let P_i be its frequency and $h_{i,j}$ denote the frequency of its j^{th} neighbour point(s), where $i \in [1, m]$ and $j \in [1, n]$, then its reliability r_i is defined below,

$$r_i = 1 - \left(\frac{\sum_{j=1}^{n} \omega_{ij} h_{ij}}{P_i}\right) \tag{3.13}$$

$$\omega_{ij} = \frac{h_{ij}}{S_i}, \quad S_i = \sum_{j=1}^{n} h_{ij} \tag{3.14}$$

Finally, the parameter γ is determined as the average of all r_i as follows, where λ can be regarded as the upper limit of the weight of colour features which is usually set to 0.5 as suggested in [88]. This is useful to avoid overweighting of either colour or shape feature in the combined approach.

$$\gamma = \frac{\lambda}{m} \sum_{i=1}^{m} r_i \tag{3.15}$$

3.3 Results and Discussion

3.3.1 Data Sets

For performance evaluation of an image retrieval system, the database is one of the significant issues. Although MPEG-7 based datasets were popularly used in some trademark retrieval systems [77, 83], the images are not deliberately designed for performance evaluation of colour logo image retrieval systems. Therefore, in this chapter, a new colour logo image database, which contains over 2300 original images and a number of distorted versions, was created. The original trademark/logo images were collected from online sources, including a large number of categories such as air company brand, car brand, coffee brand, university logos, event logos, company brand and so on, where some of them are very similar to each other. The whole database will be made publicly available in the near future.

In the real world, some captured images will contain noise due to the error of the

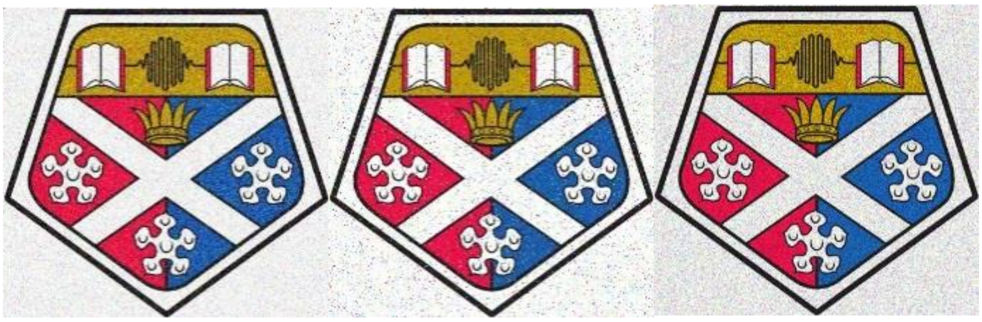


Fig. 3.6. Degraded logo images of the one shown in Fig.3.2. From *left to right*, the images are added with Gaussian noise (variance 0.02, PSNR=22.2 dB), salt and pepper noise (density=0.05, PSNR=16.57 dB) and speckle noise (variance 0.04, PSNR=18.46 dB)

circuitry of scanner or digital camera, analogue-to digital converter, illumination, and overheating problem, etc. To simulate the effect of such affecting factors, for each of the 2300 logo/trademark images, the noise was added to generate degraded versions for performance assessment. The distortion types include 13 levels of Gaussian noise (0.01:0.01:0.1 and 0.1:0.1:0.4), 4 levels of salt and pepper noise (0.05:0.05:0.2), and 4 levels of speckle noise (0.01:0.01:0.04). The logo images with added noise are considered to be the query image in the TLIR system. For the original logo image in Fig.3.2, its degraded versions using the three kinds of noise are shown in Fig.3.6 for comparison, where visual distortions can be clearly observed in these degraded images. In addition, 10 experiments at each level of noise have been performed and the final results are determined by the average values.

3.3.2 Evaluation criteria

To assess the robustness and accuracy of the new approach and the three benchmarking algorithms, a single value measurement, average precision (P_{ave}), is used to estimate the global performance of the retrieval systems. For a query image, if the system can retrieve its ground truth image in a number of retrieval results, P is 1, otherwise 0.

$$P_{ave}(R) = \frac{1}{Nq}\sum_{i=1}^{Nq} P_i(R) \tag{3.16}$$

where Nq is the number of query images, R is the number of top retrieved images which is set to 1 and 5. P(1) means the first retrieval result matches the query image, and P(5) means one of the top five results matches the query image.

The following three approaches are used for performance comparison and benchmarking with the proposed approach:

CBIR system proposed by Yang et al [69].

TLIR system proposed by Wei et al[79].

TLIR system proposed by Anuar et al[77].

In [69], MPEG-7 based dominant colour extraction was used for image retrieval, which will be compared to the proposed colour quantization approach. For [79] and

[77], global and local shape descriptors were used, and they will be compared with the proposed shape features.

And three kinds of noise, Gaussian, salt and pepper, and speckle noise, were respectively added to all the images of the new database. The retrieval results using the degraded images from these three kinds of noise are presented in separate sub-Sections as follows.

3.3.3 Retrieval results with added two different filters

Two simple filters are added to smooth the histogram in order to extract the real peak frequency, including the Median filter (MMF) and average filter (AMF). Here, the window size is set to 3. For the two filters, the corresponding retrieval results are quite comparable, yet the global performance of the median filter based adaptive fusion method is slightly better (Fig. 3.7). As a result, MMF is used in these experiments.

3.3.4 Retrieval results with added Gaussian noise

Gaussian variance varies in 13 levels within a wide range between 0.01 and 0.4, i.e. 10 levels from 0.01 to 0.1 and 3 levels from 0.1 to 0.4. The average retrieval precision in terms of $P_{ave}(1)$ and $P_{ave}(5)$ under various Gaussian variances is plotted in Fig. 3.8 for comparison. From these results, some findings can be summarized as follows.

First, when the Gaussian variance is no more than 0.08, the new approach outperforms all three other benchmarking algorithms. However, the performance degrades when the Gaussian variance exceeds 0.08, mainly due to inaccurate dominant colour extraction from images with severe noise.

Second, although it produces inferior retrieval precision when the variance of the Gaussian noise is low, Anuar’s approach [77] shows slightly better performance when the Gaussian noise variance is high. This is mainly due to the fact that the Zernike moment feature used is robust to Gaussian noise, yet poor performance from other kinds of noise as shown in the next two subsections. Third, the approaches

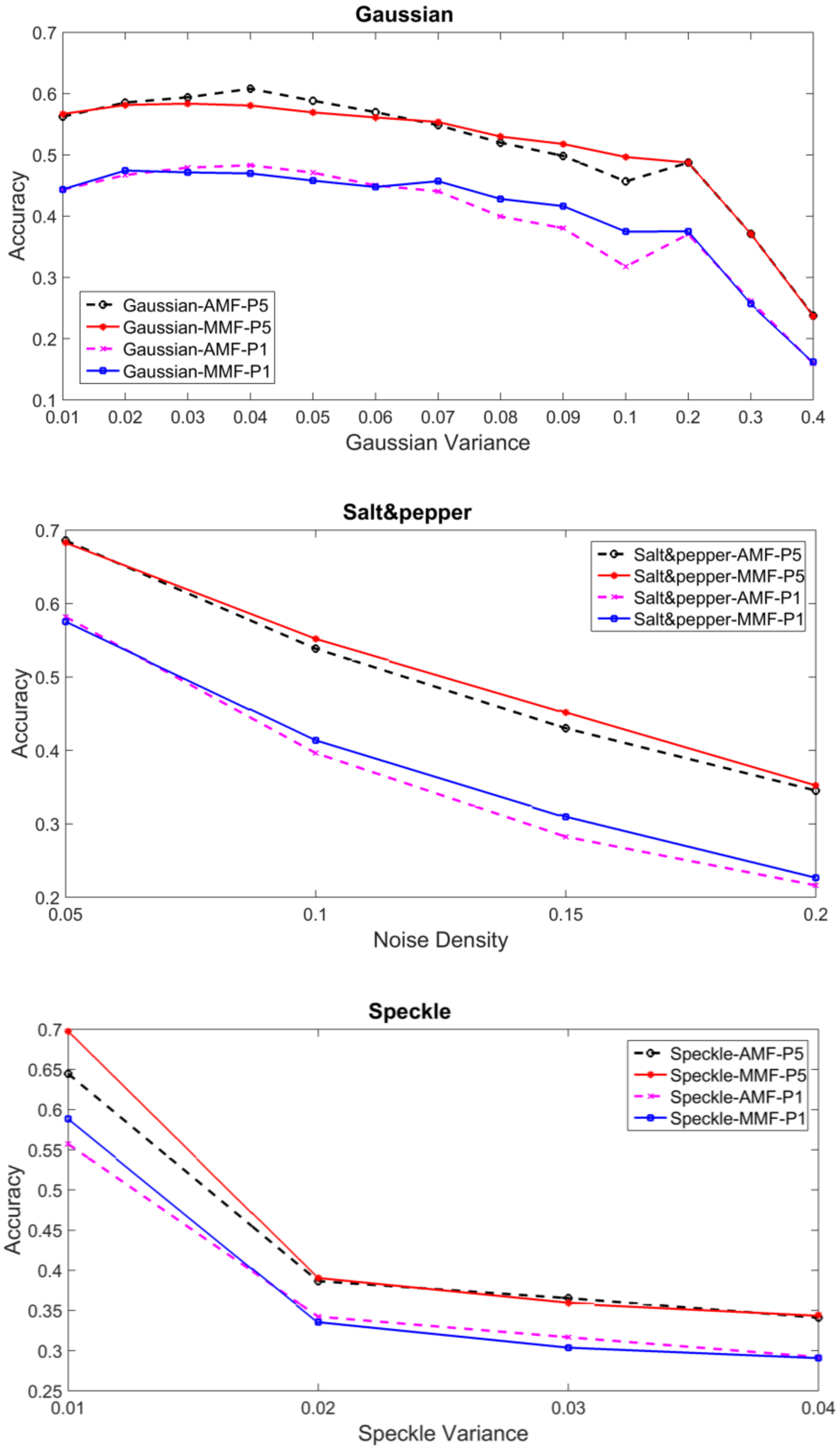


Fig. 3.7. The performance of our matching solution based on different filtering methods (R=1 & R=5). Gaussian noise dataset (top); salt and pepper noise dataset (middle); speckle noise dataset (bottom).

from Yang et al [69] and Wei et al[79] show much worse results than those from Anuar et al[77] and the new method. In addition, it seems that the performance of Wei’s approach is least affected by the added noise, as the average precision appears to nearly constant, though much lower when the Gaussian variance is small. The reason for that is the approach in [79] used Zernike moment as global feature descriptor and curvature as local descriptor. As mentioned before, Zernike moment is robust to noise, however, curvature calculation relies on the precise boundary extraction which is hardly possible in noisy conditions. As a result, the final retrieval performance is limited by the curvature feature extraction.

3.3.5 Retrieval results with added salt and pepper noise

For salt and pepper noise, four density levels from 0.05 to 0.2 were used for performance assessment, with the achieved $P_{ave}(1)$ and $P_{ave}(5)$ compared in Fig. 3.9, where some key findings are summarized as follows.

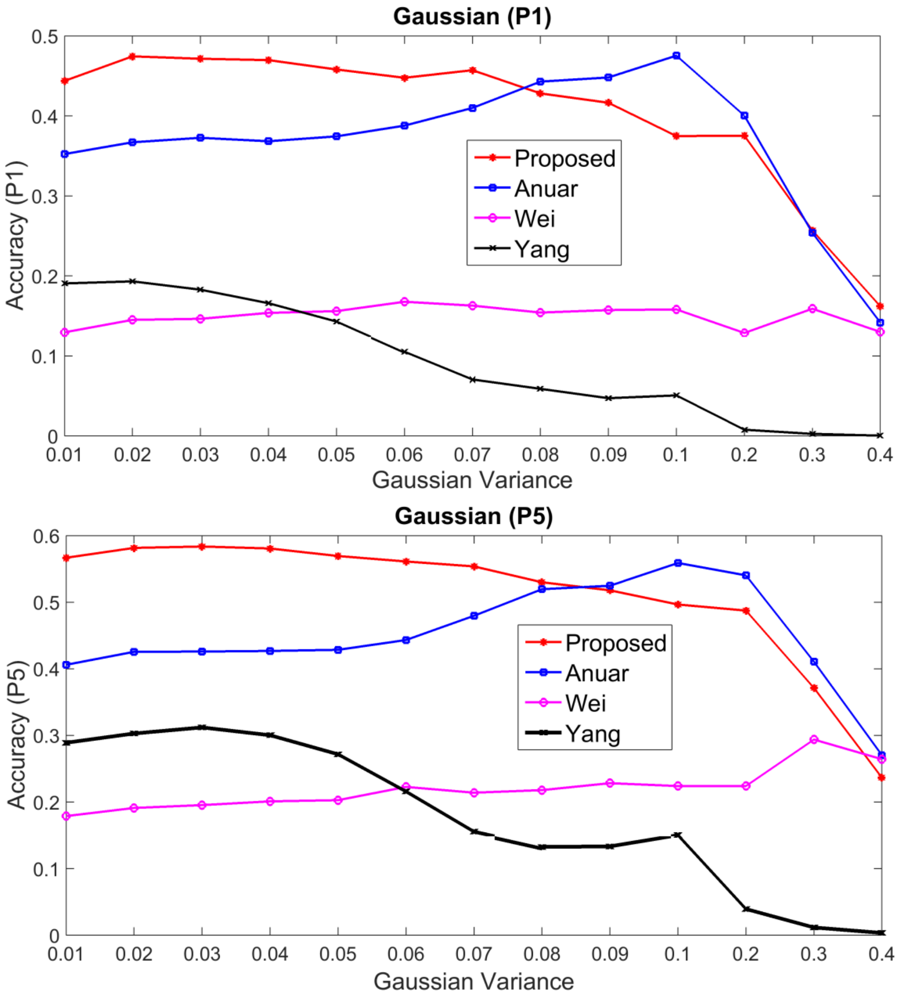


Fig. 3.8. Average retrieval precision from different retrieval approaches with added Gaussian noise when R is 1 (a) and 5 (b)

In this group of results, the new proposed approach consistently and significantly outperforms all others, which is mainly due to the fact that salt and pepper noise tends to have less effect on the extracted dominant colours. This also explains why Yang's approach [69] yields the second best results, as it is also based on dominant colour for image retrieval. However, thanks to the proposed feature fusion matching scheme, the new approach is far better than Yang [69]. Possibly due to affected gradients by salt and pepper noise, Anuar's approach [77] and Wei's approach [79], which rely on moments and edge gradients, generate much worse results here. Overall, the results from Wei's approach seem to be the worst in this group of experiments.

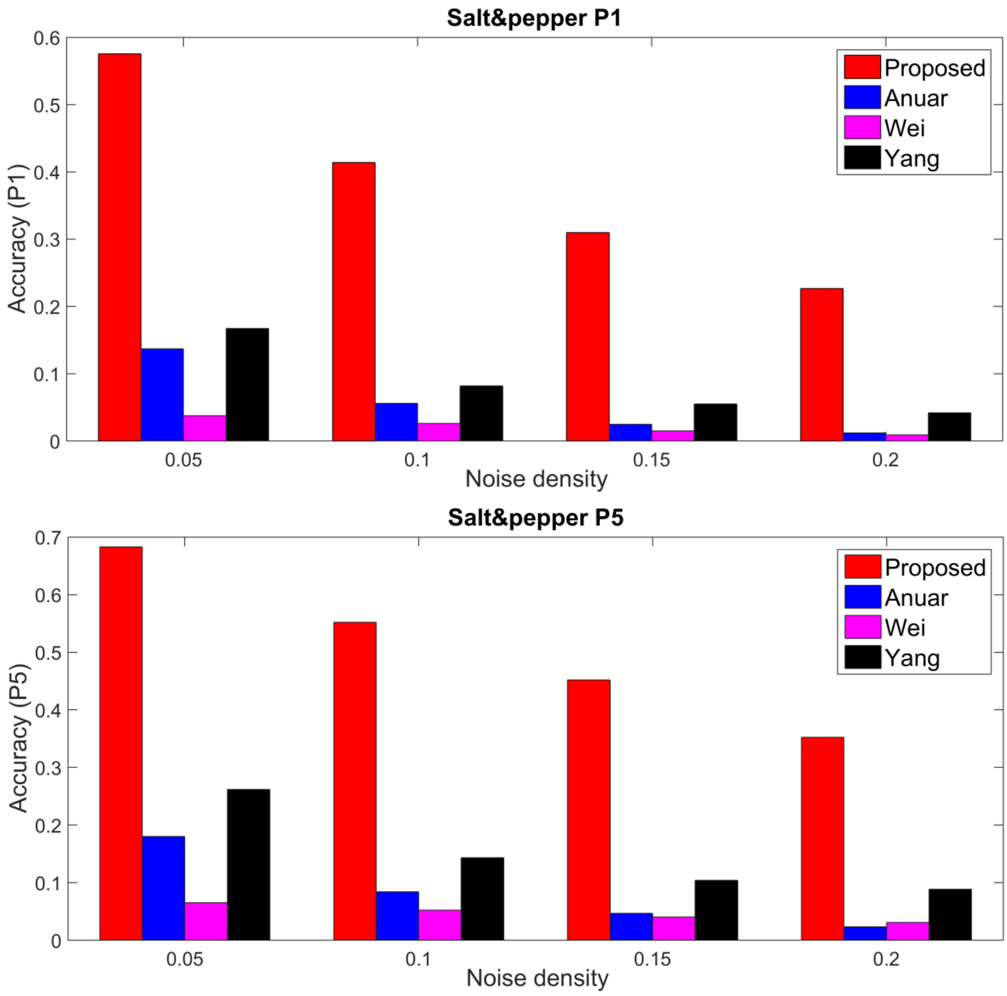


Fig. 3.9. Average retrieval precision from different retrieval approaches with added salt and pepper noise when R is 1 (a) and 5 (b)

3.3.6 Retrieval results with added speckle noise

For speckle noise, four variance levels from 0.01 to 0.04 were tested. The achieved average precision, $P_{ave}(1)$ and $P_{ave}(5)$, are plotted in Fig. 3.10 for comparison, from which some findings are summarized as follows.

First, the performance of the new approach presented in this work is among the best. Second, though Yang's approach [69] produces better results than Wei's one [79], without the proposed fusion scheme, its average precision is only nearly half of that of the new approach. In addition, the overall retrieval precision seems worse than those from the previous two groups of experiments, which means this type of noise heavily degrades the quality of image and leads to less effective results produced from the TLIR methods.

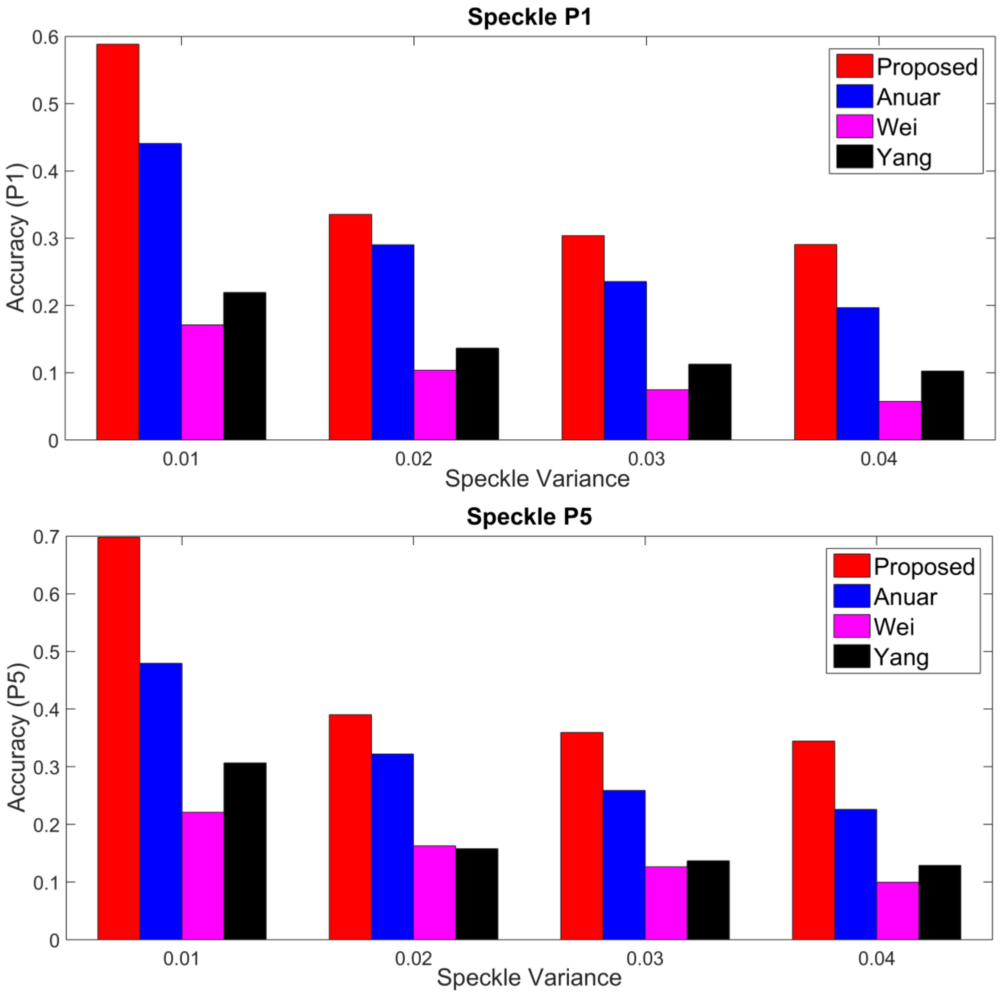


Fig. 3.10 Average retrieval precision from different retrieval approaches with added speckle noise when R is 1 (**a**) and 5 (**b**)

3.3.7 Retrieval results with the different matching strategy

The new adaptive matching scheme proposed in this work is also compared with other two popular matching strategies, i.e. the Weighted-based solution (WBS) [88] and the two-component solution (TCS) [79]. From Fig. 3.11, it can be seen that, based on each sample set, the performance of the new solution is better than that of the WBS and TCS. Since both matching strategies use fixed value for weight, which are very sensitive to the database. Therefore, they cannot always yield satisfied results on every database. However, our proposed strategy calculates the weights according to colour quality in the image. As a result, our feature matching strategy has more flexibility and usefulness.

3.3.8 Summary of main findings

With three kinds of typical noise added to the query images, the performance measured using average precision as shown in Fig. 3.8 to Fig. 3.10 has demonstrated that in general the proposed approach generates the best results and outperforms the three benchmarking algorithms. Though Yang [69] used a similar colour quantization approach, its performance is among the worst in the group, mainly due to inaccurate dominant colour extraction, especially in noise-degraded images, and lack of fusion of spatial descriptors. Wei’s approach [79], using edge curvature from a Canny edge detector as local features and the Zernike moment (ZM) as global features, seems to produce the worst results even the noise level is low. This is mainly because the local feature descriptor is easily affected by noise-degraded images. Anuar’s approach [77], due to the robustness of the edge gradient co-occurrence matrix (EGCM) feature, is found to be robust to high-variance Gaussian noise, though its performance under lower variance of Gaussian noise and salt and pepper noise appears to be much worse than the proposed approach. Meanwhile, the proposed adaptive fusion and matching solution are also more efficient than two other widely used matching solutions, WBS and TCS (Fig.3.11). In summary, by adaptive fusion of improved dominant colour extraction and component-based spatial descriptors, the proposed method becomes robust to noise degraded images and generates higher TLIR precision even with low-quality trademark/logo images.

3.4 Conclusion

In this chapter, a novel content-based trademark retrieval system is proposed for effective matching and retrieval of coloured logo/trademark images. By combining colour quantization and k-Means clustering, more accurate extraction of dominant colour is achieved as global features. Based on the determined dominant colours, component-based spatial descriptors can be successfully extracted as local features. More importantly, this chapter also proposes a fuzzy-based histogram analysis technique as a feature matching strategy for more accurate and robust retrieval of coloured logo/trademark images. In general, the proposed approach significantly outperforms three benchmarking retrieval algorithms. Also, the proposed adaptive fusion matching solution is better than WBS and TCS. These have been validated by using images with added Gaussian, salt and pepper and speckle noise. Future work will focus on detection and rotation invariant feature descriptors for more effective TLIR, especially in dealing with trademarks and logos appearing as part of larger images, where new object detection approaches can be used [162-165].

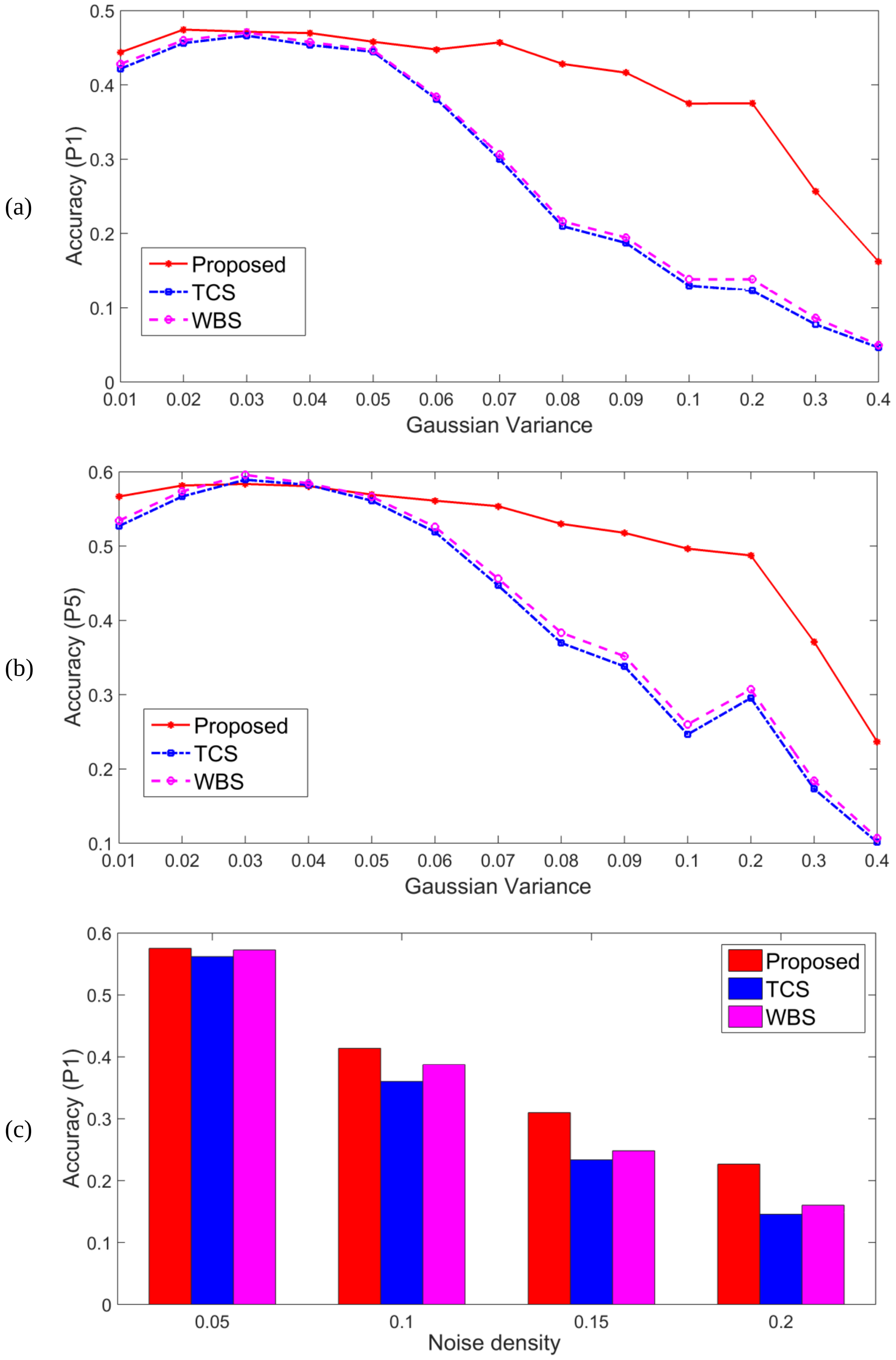


Fig. 3.11 The performance of our matching solution based on each data set, compared with the WBS and TCS. (a) Gaussian noise dataset (R=1); (b) Gaussian noise dataset (R=5); (c) salt and pepper noise dataset (R=1); (d) salt and pepper noise dataset (R=5); (e) speckle noise dataset (R=1); (f) speckle noise dataset (R=5).

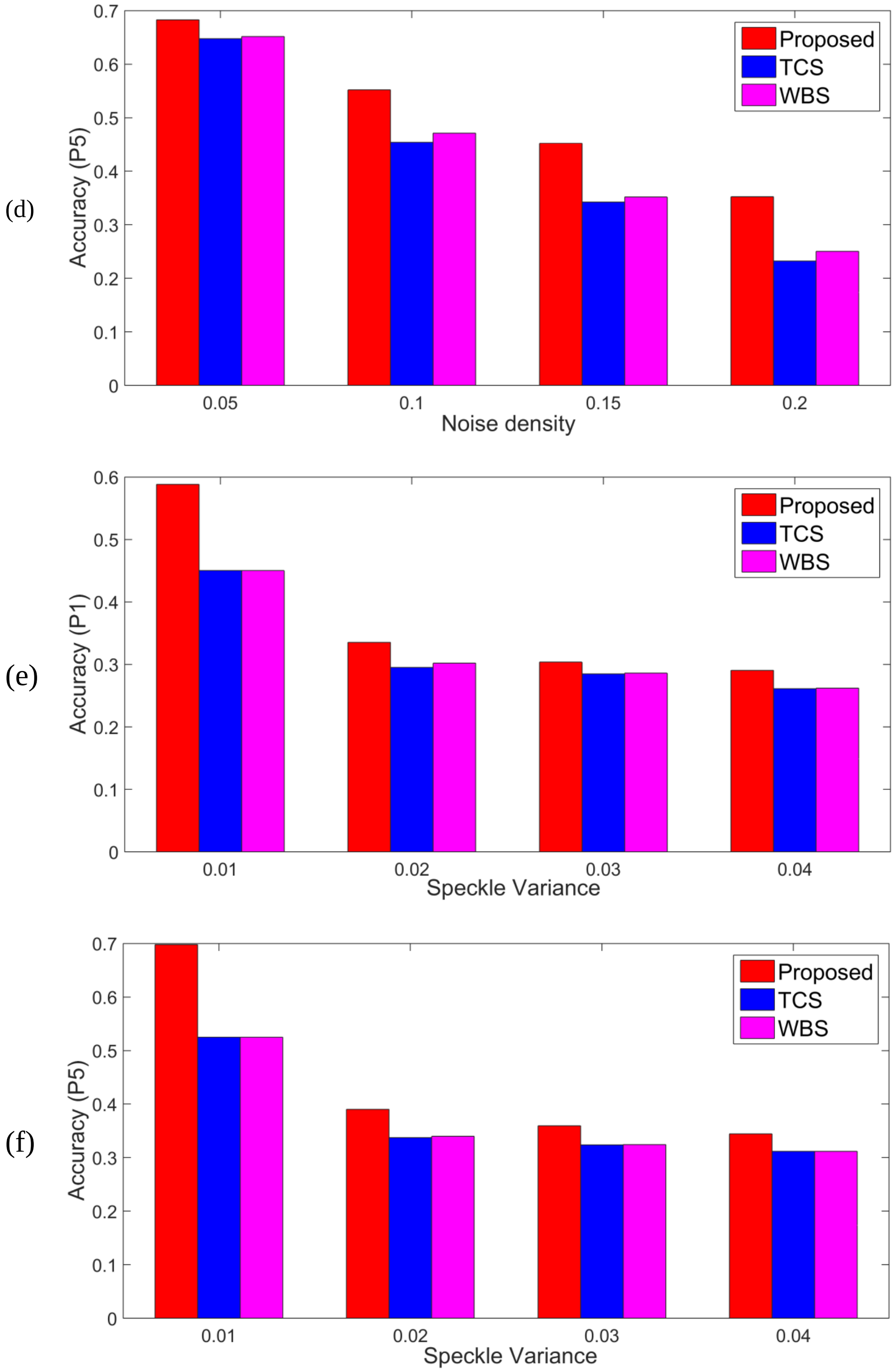


Fig.3.11 continued.

CHAPTER 4

4 Unsupervised Image Saliency Detection with Gestalt-laws Guided Optimization and Visual Attention Based Refinement

4.1 Introduction

Visual attention is a kind of fundamental cognitive capability that allows human beings to focus on the region of interests (ROIs) under complex natural environments. What kind of ROIs that we pay attention to mainly depends on two distinct types of attentional mechanisms. The bottom-up mechanism can guide our detection of the salient objects and regions by externally driven factors, i.e. colour and location, whilst the top-down mechanism controls our biasing attention based on prior knowledge and cognitive strategies being provided by visual cortex. In [108], it is found that the two attentional functions have distinct neural mechanisms but constantly influence each other to attention. However, how to practically use and fuse both attentional mechanisms for salient object detection has not been sufficiently explored. To this end, we aim to build a cognitive framework where separated model for each attentional mechanism is integrated together to determine the visual attention, refer to the salient object detection. Within our framework, the model of a bottom-up mechanism is guided by the gestalt-laws of perception. We interpreted gestalt-laws of homogeneity, similarity, proximity and figure and ground in link with colour, spatial contrast at the level of regions and objects to produce feature contrast map. The model of top-down mechanism aims to use a formal computational model to describe the background connectivity of the attention and produce the priority map. Integrating both mechanisms and applying to salient object detection, our results have demonstrated that the proposed method consistently outperforms a number of existing state-of-the-art approaches on five challenging and complicated datasets in terms of higher precision and recall rates, AP (average precision) and AUC (area under curve) values.

The main contributions of this work can be highlighted as follows:

1) We propose gestalt laws guided optimization and visual attention-based refinement framework (GLGOV) for unsupervised salient object detection, where bottom-up and top-down mechanisms are combined to fully characterize HVS for effective forming of objects in a whole;
2) We introduce a new background suppression model guided by the Gestalt law of figure and ground, where superpixel-level colour quantization and adaptive thresholding are applied to determine object-level foreground and background for the calculation of the background correlation term and the spatial compactness term to further suppress the background and highlight the saliency objects;
3) We have carried out comprehensive experiments on five challenging and complex datasets and benchmarked with eight state-of-the-art saliency detection models, where useful discussions and conclusions are achieved.

The rest of this chapter is organised as follows. The proposed framework by combining bottom-up and top-down HVS mechanisms for saliency detection is presented in Section 4.2, where the implementation detail is discussed in Section 4.3. Section 4.4 presents the experimental results and performance analysis. Finally, some concluding remarks are drawn in Section 4.5.

4.2 The Proposed Framework for Saliency Detection via GLGO

Although colour and spatial features have been widely used for salient object detection, the efficacy can still be fragile, especially in dealing with large objects and/or complicated background in the scenes [26]. The salient object often cannot be extracted as a whole (see examples in Fig. 4.1), though it is still relatively easy for our HVS to identify the full range of the salient objects. This shows a gap between existing approaches to an ideal one that can better exploit the potential of our HVS for more accurate salient object detection. To this end, we propose a Gestalt-law guided cognitive approach to calculate bottom-up attention. As gestalt-laws can characterize the capabilities of HVS to yield whole forms of objects from a group of simple and even unrelated visual elements [166], e.g. edges and regions, we aim to employ these laws to guide/improve the process of salient object detection.

As shown in Table 2.1, the principle of gestalt laws has been implicitly reflected in many existing approaches. This includes not only the colour and spatial features in representing the HVS but also their applications in other key stages. Examples herein can be found in homogeneity based colour quantization and pixel clustering in the pre-processing, similarity-based feature measurement in saliency map generation and object-prior based grouping in post-processing. In fact, the concept of gestalt laws was first introduced as a series of mechanisms to explain the human visual perception, dated back to 1920s in [167, 168], where Gestalt law based detection is proved to fit the human perception [169]. Starting from low-level features of the salient stimuli that influence perception at the bottom level way up to high level cognition, gestalt can be naturally used to define bottom-up objects [170, 171]. Although the gestalt theories are gradually used in saliency detection models explicitly [172-174] or implicitly [123, 175], there is no specified structure to regulate the relationship between Gestalt law and saliency models. How these laws can be systemically applied for salient detection remains unexplored. As a result, we aim to further explore various aspects of gestalt laws when applying in a bottom-up model and also combined with the top-down model for feature contrast map generation, which is detailed in Section 4.3.

A new saliency detection framework inspired by the Gestalt laws of HVS is proposed. The proposed framework contains six main modules, i.e. homogeneity, similarity and proximity, figure and ground, background connectivity, two-stage refinement and performance evaluation. The overall diagram of our saliency detection framework is illustrated in Fig.4.1 where corresponding gestalt laws and visual psychology used in different modules are specified and also detailed below.

The homogeneity module aims to group or cluster pixels into regions on the basis of human visual perception [168, 176]. According to Gestalt law of homogeneity and the gestalt perceptual organization theory [168], if pixels have similar intensity, colour, orientation or other features, they should be treated as homogeneity regions.

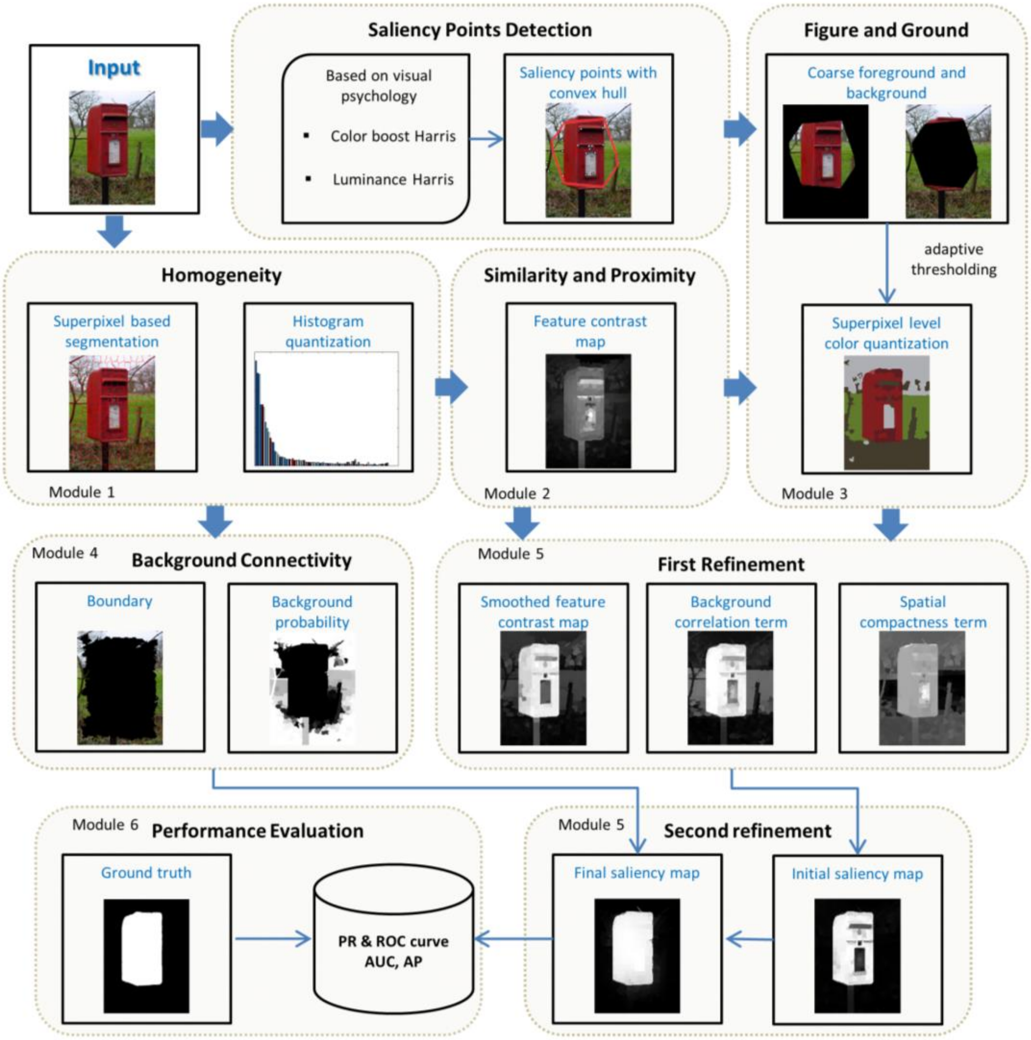


Fig. 4.1 The proposed framework in six modules guided by various gestalt laws (specified in brackets) and visual psychology.

To this end, there can be a number of homogeneity regions extracted in one image. Specifically, the simple linear iterative cluster (SLIC) method [177] is used to separate the image into regions namely superpixels, where the colour quantization is applied to reduce the number of colours based on colour homogeneity.

In similarity and proximity module, colour contrast and spatial contrast across superpixels are measured for extraction of the feature contrast map and smoothing process. Based on Gestalt laws of similarity and proximity, elements/objects tend to be perceived as a whole in cognition if they are close enough to each other and/or share similar appearance of visual features. In other words, if a superpixel with a low saliency value is surrounded by those with high saliency values, this superpixel should be assigned with a high saliency value as determined by its neighbouring superpixels provided that their colours are similar to each other. As such, a smoothing process should be introduced to refine the extracted saliency map, which will be defined in two-stage refinement.

The saliency point detection module is to coarsely split the whole image into foreground and background based on a convex hull formed by the detected saliency points. Inspired by human visual psychology [110, 111], the greyscale image and colour image are treated differently. Herein, we choose a luminance-based operator for greyscale images and Colour Boosted Harris (CBH) operator for colour images for detection of saliency points [178]. As this module is relative quite straightforward, it is not detailed in this chapter.

In figure and ground module, Gestalt law of figure and ground is applied to suppress the background and highlight the foreground objects. According to HVS, human's perception is comprised of objects and their surrounding background under observation, where salient objects can be automatically highlighted from the suppressed background. To this end, superpixel level colour quantization, adaptive thresholding, and a saliency points detection method [178] are well combined together for background suppression.

The background connectivity module is to calculate the background probability of each super-pixels. Inspired by background connectivity theory [179], attention to the scene or background of the image will increase the background connectivity in our visual cortex. Herein, we employ a component from the background detection method [22] to represent the background connectivity mathematically.

In the two-stage refinement module, the first refinement helps to smooth Gestalt law guided saliency maps by considering all superpixels from the foreground or background as well as fusion with the feature contrast map and background suppression map. The result from the first refinement will be fused with the background probability map derived from the background connectivity module and further smoothed in the second refinement process. Finally, the final saliency map is determined.

In the performance evaluation module, comprehensive experiments are carried out against a number of state-of-the-art methods on several widely used publicly available databases. Some widely used evaluation criteria such as PR curve, ROC curve, AUC and AP value are used for quantitative assessment. Detailed results are reported in Section 4.4.

4.3 Gestalt-laws Guided Optimization for Image Saliency Detection

In this Section, the implementation of the proposed saliency detection framework is detailed in five stages, i.e. homogeneity, similarity and proximity, figure and ground, background connectivity and two-stage refinement below.

4.3.1 Homogeneity clustering

For a colour image in three channels, it usually contains thousands of pixels with up to 256^3 possible values, where pixel-based saliency detection suffers from high computational cost. Therefore, we use two operations to reduce the dimension of the image and simplify the problem. First, based on Gestalt law of homogeneity, similar pixels spatially close to each other in the image can be considered as a whole. To this end, we use SLIC [177] to segment the image into a number of regions called superpixels which generally have inner colour consistency and a compact shape with decent boundary adherence. As suggested by [7], we set the number of superpixels n to 500. Secondly, using the same scheme as suggested in [15], we firstly transfer the RGB colour space to L*a*b* colour space, and uniformly quantize each colour channel into q bins in order to reduce the number of colours to generate a new image histogram H with q^3 bins. Herein we define q=12 instead of 16, as this change has little effect on the experimental results yet the number of colours and associated computation cost has been significantly reduced, i.e. proportionally to the reduction from 16^3 to 12^3.

From the quantized image, bins of dominant colours are selected to cover more than 95% of the image pixels. The remaining colour bins that cover less than 5% pixels are merged into the selected dominant colour bins based on minimum Euclidean distance criterion. An example of the colour quantization is shown in

Module 3, Fig. 4.1. After superpixel-based image segmentation and colour quantization, each superpixel $S_i(i = 1,2, \dots n)$ has its histogram H_i normalised by $\sum_{m=1}^{k} H_i(m) = 1$, where *k* is the index of the colour bins.

4.3.2 Similarity and Proximity

In human visual perception, usually, we are aware of the image regions that have high contrast to their surroundings. According to the Gestalt law of proximity and similarity, HVS tends to perceive elements that are close or similar to each other in a group. This actually explains that, at the superpixel-level, those image regions are normally grouped by spatial closeness and colour similarity of superpixels. In addition, those superpixels should have a similar level of low inner contrast and highly different from the surrounding superpixels. In this Section, we presented in detail the proposed saliency object detection method based on colour and spatial contrast at superpixel level.

First, for two superpixels S_i and S_j, their colour distance $C_d(S_i, S_j)$ is defined as the Euclidean distance between the mean colours of the two superpixels and scaled within the range of [0,1]. Here the colour distance C_d measures the dissimilarity of the colour appearance of the two superpixels.

In addition to the colour based dissimilarity, the Gestalt law of proximity for the two superpixels S_i and S_j is denoted as a spatial proximity distance $P_d(S_i, S_j)$, which is defined as the Euclidean distance between the centroids of the two superpixels. The larger the $P_d(S_i, S_j)$ is, the higher proximity between S_i and S_j is the spatial proximity distance can be further normalised within [0, 1] as follows:

$$\mathrm{P_{d_{norm}}(S_i, S_j) = \frac{P_d(S_i, S_j) - P_{d\_max}}{P_{d\_min} - P_{d\_max}}} \tag{4.1}$$

where P_{d_min} and P_{d_max} refer respectively to the minimum and the maximum of all possible values of P_d as determined from superpixels of the given image.

Based on the defined colour distance C_d and the normalised spatial proximity distance P_{d_norm}, the global colour contrast C_G for a superpixel S_i is defined by

$$\mathrm{C_G(S_i) = \sum_{j=1}^{n} A_j \cdot P_{d\_norm}(S_i, S_j) \cdot C_d(S_i, S_j)} \tag{4.2}$$

where n is the number of superpixels, and A_j is the normalised area of the superpixel S_j where a larger one contributes more to the global colour contrast value of S_i, where the sum of A_j is 1.

Similarly, we define the global spatial contrast value for a superpixel S_i below:

$$\mathrm{S_G(S_i)} = \frac{\sum_{j=1}^{n} \mathrm{P_{d\_norm}(S_i, S_j) \cdot C_I(S_i, S_j) \cdot L_j}}{\sum_{j=1}^{n} \mathrm{L_j} \sum_{j=1}^{n} \mathrm{P_{d\_norm}(S_i, S_j) \cdot C_I(S_i, S_j)}} \tag{4.3}$$

where L_j is the weight of the spatial layout, which is defined as the minimum distance from the centroid of superpixel S_j to any of the four image borders; C_I is the inter superpixel colour similarity defined below, where $\mathrm{H_i}$ and $\mathrm{H_j}$ refer respectively to the colour histogram of the two superpixels S_i and S_j with m as the index for the k colour bins.

$$\mathrm{C_I(S_i, S_j)} = \sum_{m=1}^{K} ((1 - |\mathrm{H_i(m) - H_j(m)}|) * \min(\mathrm{H_i(m), H_j(m)})) \tag{4.4}$$

Unlike conventional histogram intersection method, we introduce the term $(1 - |\mathrm{H_i(m) - H_j(m)}|)$ rather than $\min(\mathrm{H_i(m), H_j(m)})$ to measure the similarity of two histogram bins based on their frequencies. For example, given two pairs of histogram bins: $\mathrm{H_i(1) = 0.3, H_j(1) = 0.2}$, and $\mathrm{H_i(2) = 0.8, H_j(2) = 0.2}$, conventional histogram intersection method will produce the same result of 0.2, which fails to reflect the difference between the histogram bins. On the contrary, in our modified definition, the similarity terms for the two cases become 0.9 and 0.4, respectively, which improves the similarity measurement and makes it more consistent to human perceptions than conventional histogram intersection approach.

For simplicity, we define the initial color and spatial (ICS) contrast as a combined measurement of colour similarity and spatial proximity between two superpixels S_i and S_j:

$$\mathrm{ICS(S_i, S_j)} = \mathrm{P_{d\_norm}(S_i, S_j) \cdot C_I(S_i, S_j)} \tag{4.5}$$

Accordingly, Eq. (4.3) can be simplified as

$$\mathrm{S_G(S_i)} = \frac{\sum_{j=1}^{n} \mathrm{ICS(S_i, S_j) \cdot L_j}}{\sum_{j=1}^{n} \mathrm{ICS(S_i, S_j)}} \tag{4.6}$$

Based on both the global colour contrast C_G and the global spatial contrast S_G, the feature contrast value for superpixel S_i is defined as

$$\mathrm{FC}(\mathrm{S_i}) = \mathrm{C_G}(\mathrm{S_i}) \cdot \mathrm{S_G}(\mathrm{S_i}) \tag{4.7}$$

Although a salient object usually has a high contrast to the background, it may contain some low-contrast parts. As a result, the salient object cannot be detected as a whole as these low-contrast parts will be missed due to their small sizes and low saliency values. To tackle these problems, a smoothing process (detailed in 4.5) is applied to filter such superpixels.

4.3.3 Figure and Ground

Although the image background can be very complicated and even similar to the foreground, in most cases people can still recognise the salient objects without difficulty, and this can be explained by the Gestalt law of figure-ground that human’s perception is the fusion of observed objects and their surrounding background. To this end, we can enhance the contrast of salient superpixels for their easier detection by suppressing the saliency value of the superpixels in the background. Based on gestalt law of similarity and proximity, salient objects in superpixel level image are composed of several superpixels with similar colour appearance and close spatial distance. This actually indicates that salient objects are formed by several superpixels in a compact manner yet the distribution of background superpixels tends to be more dispersive. Based on this assumption, in this subsection, we introduce a novel background suppression model, the spatial distribution of superpixels is utilized to highlight the objects and suppress the background.

According to the dominant colour theory [180], images are usually quantized into up to 8 dominant colours. Similarly, for a given superpixel level image, the colours of all superpixels can also be divided into 8 coarse clusters in the L*a*b* colour space, where each of the three individual colour components is divided into two parts by a threshold $T^{(.)}$. For each of the eight colour clusters, the extracted dominant colour is denoted as $\mathrm{C_i} = [\bar{\mathrm{x}}_\mathrm{i}^\mathrm{L}, \bar{\mathrm{x}}_\mathrm{i}^\mathrm{A}, \bar{\mathrm{x}}_\mathrm{i}^\mathrm{B}]$, $\mathrm{i} \leq 8$, where $\bar{\mathrm{x}}_\mathrm{i}^{(.)}$ is the average value in one of the three colour components L*, a* and b*, and C_i will be used as the center for the corresponding colour cluster.

Herein the colour boosted Harris point (CBHP) operator [178] is employed to determine the coarse regions of foreground and background. For colour or greyscale images, the CBHP operator and luminance operator are respectively applied to detect the salient points, where a convex hull is computed to enclose all these salient points. All superpixels within the convex hull are defined as the foreground and the remaining superpixels as the background (see module 3 in Fig.4.1). Let $\mathrm{MP}_{\mathrm{G_f}}^{(.)}$ and $\mathrm{MP}_{\mathrm{G_b}}^{(.)}$ denote the mean colour (with three components) of the foreground G_f and the background G_b. To maximize the colour contrast, in each colour component the threshold T(.) is adaptively defined as $\mathrm{T}^{(.)} = \frac{1}{2}\left(\mathrm{MP}_{\mathrm{G_f}}^{(.)} + \mathrm{MP}_{\mathrm{G_b}}^{(.)}\right)$.

With the adaptively determined eight dominant colours above, we merge similar colours for simplicity below. We iteratively examine each pair of the extracted dominant colours and merge them if their Euclidean distance is less than a threshold $\mathrm{T_d}$. For each image, the threshold is adaptively decided as half of the Euclidean distance between $\mathrm{MP}_{\mathrm{G_f}}^{(.)}$ and $\mathrm{MP}_{\mathrm{G_b}}^{(.)}$, and the upper limit of $\mathrm{T_d}$ is set to 15.

Let C_i and C_j be two colour clusters, they can be merged by the weighted average agglomerative procedure [69]:

$$\begin{gathered} \mathrm{C}^{(\cdot)} = \rho \mathrm{C}_\mathrm{i}^{(\cdot)} + (1-\rho)\mathrm{C}_\mathrm{j}^{(\cdot)} \\ \rho = \frac{\mathrm{p_i}}{\mathrm{p_i} + \mathrm{p_j}} \end{gathered} \tag{4.8}$$

where p_i and p_j denote, respectively, the number of pixels in C_i and C_j . For each superpixel-level image, we can usually extract 3-8 dominant colours, located in background or foreground. In other words, we may have up to 16 colour clusters obtained, where half of them are from the foreground group and others from the background group. As the spatial information is excluded in colour quantization, within each colour cluster the superpixels are not necessarily spatially grouped together. This means that each colour cluster may contain several separated colour blocks composed of one or more superpixels. Here we define these blocks as objects $\mathrm{O_i}$. As a result, the superpixel-level image has now become an object-level image.

For each colour cluster C_k in either G_f or G_b, denote $\mathrm{M(C_k)}$ as its geographic center. The spatial compactness term of C_k is defined as the average distance between $\mathrm{M(C_k)}$ and each object $\mathrm{O_i}$ it contains:

$$\mathrm{Sd_f(C_k)} = \frac{1}{\mathrm{N_k}} \sum_{\forall \mathrm{O_i \subset C_k \subset G_f}} \|\mathrm{M(O_i)} - \mathrm{M(C_k)}\| \tag{4.9a}$$

$$\mathrm{Sd_b(C_k)} = \frac{1}{\mathrm{N_k}} \sum_{\forall \mathrm{O_i \subset C_k \subset G_b}} \|\mathrm{M(O_i)} - \mathrm{M(C_k)}\| \tag{4.9b}$$

where $\mathrm{k} \leq 8$, N_k denotes the number of objects in the cluster C_k and $M(O_i)$ refers to the geographic center of the object O_i.

Similar to $P_{d_{norm}}$ in Eq. (1), Sd can be scaled into [0, 1] by

$$\mathrm{N_{Sd}(C_k)} = \frac{\mathrm{Sd(C_k)} - \mathrm{Sd_{max}}}{\mathrm{Sd_{min}} - \mathrm{Sd_{max}}} \tag{4.10}$$

where $\mathrm{Sd(C_k)}$ is either $\mathrm{Sd_f(C_k)}$ or $\mathrm{Sd_b(C_k)}$; $\mathrm{Sd_{min}}$ and $\mathrm{Sd_{max}}$ refer respectively to the minimum and the maximum of all spatial compactness values of Sd.

By applying the spatial compactness, the salient objects can be highlighted. Herein we introduce a new background correlation term to further suppress the background and enhance the salient objects. This term is used to measure the connection degree between an object *O* and the image borders, where a large value indicates a high degree of overlapping between *O* and the image borders. As a result, the object *O* will more likely be classified into the background as we assume the salient object should be away from the image borders. The background correlation term is defined as

$$\mathrm{B_c(S_i)} = \frac{\mathrm{N_b}}{\mathrm{N_c}}, \quad \mathrm{S_i \subset O} \tag{4.11}$$

where N_c denotes the total number of superpixels in *O*, among which N_b is the number of superpixels in *O* that directly connects to the image borders. Using the maximum and the minimum of all possible values of $\mathrm{B_c}$, denoted as $\mathrm{B_{c\_max}}$ and $\mathrm{B_{c\_min}}$, we can obtain normalised $\mathrm{B_c}$ within [0,1] below.

$$\overline{\mathrm{B_c}}(\mathrm{S_i}) = \frac{\mathrm{B_c}(\mathrm{S_i}) - \mathrm{B_{c\_max}}}{\mathrm{B_{c\_min}} - \mathrm{B_{c\_max}}} \tag{4.12}$$

4.3.4 Background Connectivity

In this subsection, an effective method of background detection [22] is employed to extract the background connectivity in a numerical value by determining the probability of each superpixels being the background. For a given superpixel $\mathrm{S_i}$, the associated background probability can be determined by:

$$\mathrm{BP}(\mathrm{S_i}) = \frac{\mathrm{Len}(\mathrm{S_i})}{\sqrt{\mathrm{A}(\mathrm{S_i})}}, \tag{4.13}$$

$$\mathrm{A}(\mathrm{S_i}) = \sum_{\mathrm{j=1}}^{\mathrm{n}} \exp\left(-\frac{\mathrm{d^2_{geo}}(\mathrm{S_i}, \mathrm{S_j})}{2\sigma^2}\right), \tag{4.14}$$

$$\mathrm{Len}(\mathrm{S_i}) = \sum_{\mathrm{j=1}}^{\mathrm{n}} \exp(-\frac{\mathrm{d^2_{geo}}(\mathrm{S_i}, \mathrm{S_j})}{2\sigma^2}) \cdot \delta\,(\mathrm{S_j}). \tag{4.15}$$

where σ is set to 10, $\delta(\cdot)$ is 1 for superpixels on the image boundary and 0 otherwise as suggested in [65], and n is the number of superpixels. $\mathrm{d_{geo}}(\mathrm{S_i}, \mathrm{S_j})$ is the geodesic distance between any two superpixels, which is calculated by accumulating the weights $\mathrm{d_{app}}$, measured using the Euclidean distance between pixels along the shortest path formed by a sequence of adjacent superpixels between $\mathrm{S_i}$ and $\mathrm{S_j}$ on the graph, and l is the total number of superpixels in the determined path.

$$\mathrm{d_{geo}}(\mathrm{S_i}, \mathrm{S_j}) = \min_{\mathrm{S_1=S_i,S_2,\dots,S_l=S_j}} \sum_{\mathrm{i=1}}^{\mathrm{l-1}} \mathrm{d_{app}}(\mathrm{S_i}, \mathrm{S_{i+1}}) \tag{4.16}$$

To scale BP within [0,1], the parameter σ_{BP} is set to 1, and the normalization process is defined as:

$$\mathrm{NBP}(\mathrm{S_i}) = 1 - \exp(-\frac{\mathrm{BP^2}(\mathrm{S_i})}{2\sigma^2_{\mathrm{BP}}}) \tag{4.17}$$

4.3.5 Two-stage Refinement

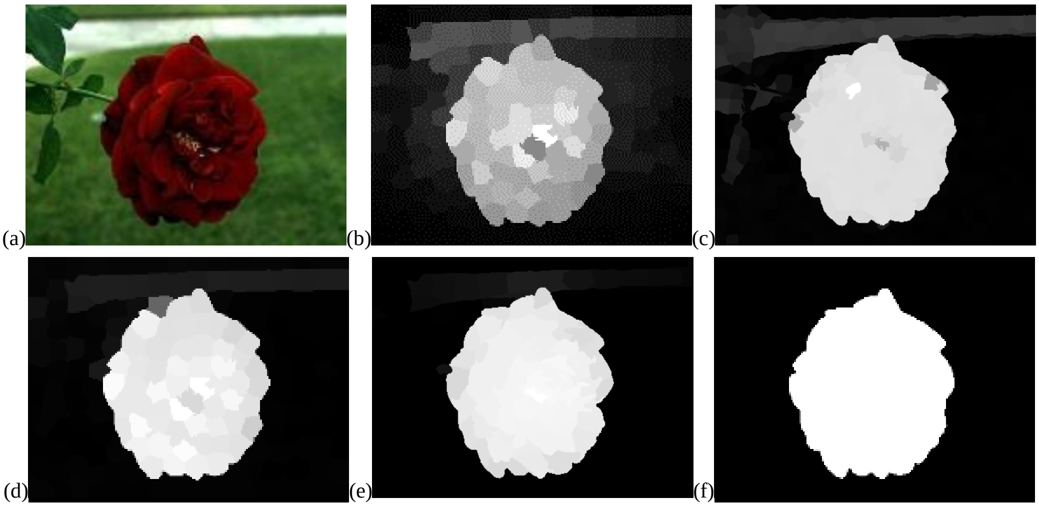


Fig. 4.2 An example of proposed saliency computation: (a) The original image, (b) Feature contrast map, (c) background suppression map, (d) first-stage refinement, (e) final saliency map after second-stage refinement, and (f) the ground truth

In the first stage refinement, the feature contrast map, spatial compactness term, and background correlation terms will be smoothed separately and fused together to generate initial saliency map.

For each superpixel, its saliency value is replaced using the weighted average of the saliency values of its surrounding superpixels. Similar to image filtering, the process applied to the superpixel level image can also smooth the saliency values for more consistent detection of salient objects. In [27], a linear varying smoothing operator is employed to smooth the colour contrast in the image. However, this smoothing procedure fails to address the spatial factor. In our approach, we improve this procedure with Gestalt law of similarity and proximity and apply it to superpixel level image filtering. For robustness, for a given superpixel $\mathrm{S_i}$, we choose $\mathrm{k} = \mathrm{n}/8$ nearest neighbours for weighted average to refine the feature contrast value of $\mathrm{S_i}$ locally as follows:

$$\mathrm{RFC}(\mathrm{S_i}) = \frac{1}{(\mathrm{k}-1)\mathrm{T}} \sum_{\mathrm{j}=1}^{\mathrm{k}} (\mathrm{T} - \mathrm{ICS}(\mathrm{S_i}, \mathrm{S_j})) \cdot \mathrm{FC}(\mathrm{S_i}) \cdot \mathrm{A_j} \tag{4.18}$$

where $\mathrm{T} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{ICS}(\mathrm{S_i}, \mathrm{S_j})$ is the sum of colour similarity and spatial proximity between $\mathrm{S_i}$ and its k nearest neighbours. A linearly-varying smoothing weight $\mathrm{T} - \mathrm{ICS}(\mathrm{S_i}, \mathrm{S_j})$ is used to assign larger weight to the superpixels which has similar colour and close to $\mathrm{S_i}$. $(\mathrm{k}-1)\mathrm{T}$ is the normalization term to scale the value into [0,1].

For the input image in Fig. 4.2(a), Fig. 4.2(b) shows the contrast-based saliency map obtained from Eq. (4.7). Due to the shadow-caused low contrast parts in the flower, the saliency map actually contains many low saliency valued superpixels. This will inevitably affect the successful detection of the salient object from the image. As a result, the smoothing process and background suppression model is applied with the results shown in Fig. 4.2(c). After smoothing, the saliency map has been improved in several ways. First, the low saliency values from the salient object have been enhanced. Second, the saliency values for all the superpixels become more consistent, which are actually raised. This shows that the proposed smoothing procedure can not only filtering low-contrast defects but also normalize the overall saliency values. Consequently, the quality of the adjusted saliency map is significantly improved.

For all the superpixels in C_k, they will be assigned the same spatial compactness term $N_{Sd}(C_k)$ as determined in Section 4.3. The lower this value is the higher likeness the salient object it has. Although $N_{Sd}(C_k)$ can measure the spatial compactness of all superpixels within C_k, it cannot differ between them even for incorrectly detected foreground superpixels. To overcome this drawback, using the similar process in Eq. (18), the Gestalt laws of similarity and proximity are applied to smooth the spatial compactness term for each superpixel S_i as follows:

$$R_{Sd}(S_i) = \frac{1}{(n-1)T}\sum_{j=1}^{n}\left(T - ICS\left(S_i, S_j\right)\right) \cdot N_{Sd}\left(S_j\right), S_j \subset C \qquad (4.19)$$

where again n is the number of superpixels, $T = \sum_{j=1}^{n} ICS(S_i, S_j)$ is the sum of colour similarity and spatial proximity between S_i and other superpixels. Note that, in global refinement, all superpixels rather than neighbouring superpixels are selected for weight average because the spatial compactness term has been extended from superpixel-level to object-level.

In addition, the background correlation term can also be globally refined by

$$R_{B_c}(S_i) = \frac{1}{(n-1)T}\sum_{j=1}^{n}\left(T - ICS\left(S_i, S_j\right)\right) \cdot \overline{B_c}(S_j), S_j \subset O \qquad (4.20)$$

Finally, the background suppression map is determined by using the conjunction of both the spatial compactness term $\mathrm{R_{Sd}(S_i)}$ and background correlation term $\mathrm{R_{B_c}(S_i)}$ below:

$$\mathrm{OP(S_i) = R_{Sd}(S_i) \cdot R_{B_c}(S_i)} \tag{4.21}$$

Based on Eq.(4.18) and Eq. (4.21), our initial saliency map is formed by

$$\mathrm{ISA(S_i) = RFC(S_i) \cdot\ OP(S_i)} \tag{4.22}$$

In the second refinement stage, the initial saliency map and the background probability map are fused by adopting a cost function [22] to optimize the whole procedure. Let $\mathrm{FSA_i}$ denote the final saliency value determined for the superpixel $\mathrm{S_i}$, the cost function is given by:

$$\begin{aligned}\mathrm{J(FSA_i)} = \min_{\mathrm{FSA_i}} \Bigg[& \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{NBP(S_i) \cdot FSA_i}^2 + \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{ISA(S_i) \cdot (FSA_i - 1)^2} \\ & + \sum_{\mathrm{i,j=1}}^{\mathrm{n}} \omega_{\mathrm{i,j}} \cdot \left(\mathrm{FSA_i - FSA_j}\right)^2 \Bigg],\end{aligned} \tag{4.23}$$

$$\omega_{\mathrm{i,j}} = \exp\Big(-\frac{\mathrm{d_{app}^2}\left(\mathrm{S_i, S_j}\right)}{2\sigma^2}\Big) + \mu \tag{4.24}$$

where three terms represent different cost. The first term $\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{NBP(S_i) \cdot FSA_i}^2$ is used to make the value of $\mathrm{S_i}$ close to zeros and enlarge background probability NBP. The second term $\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{ISA(S_i) \cdot (FSA_i - 1)^2}$ is used to make the value of $\mathrm{S_i}$ close to one and enlarge the saliency measurement. The last term is used for continuous saliency values. It is large in flat regions and small at region boundaries. The three terms are all squared errors and the optimal saliency map is computed by least-square. The parameter σ is set to 10 as defined in Eq. (4.14). As seen in Fig. 4.2 (d-e), by adding the two-stage refinement, the salient object can be significantly highlighted whilst the saliency value for the background is effectively suppressed.

The pseudocodes for the proposed framework are shown in Algorithm 1.

```
Algorithm 1: GLGOV
Input: image samples
Calculate the salient points by CBHP
Module 1
    1.1 Segment the image into a number of superpixels.
    1.2 Transfer RGB colour space to L*a*b* colour space.
    1.3 Quantize the image histogram into $12^3$ bins.
Module 2
    2.1 Calculate the color contrast map by Eqs 4.1 and 4.2.
    2.2 Calculate the spatial contrast map by Eqs 4.3 – 4.6.
    2.3 Calculate the initial feature contrast map by Eq 4.7.
Module 3
    3.1 Get the coarse background and foreground
    3.2 Calculate the threshold $\mathrm{T}^{(\cdot)} = \frac{1}{2}\left(\mathrm{MP}_{\mathrm{G_f}}^{(\cdot)} + \mathrm{MP}_{\mathrm{G_b}}^{(\cdot)}\right)$
    3.3 Divide L*a*b* colour space into maximum 8 clusters
Module4
    4.1 Calculate the spatial compactness term by Eqs 4.9 and 4.10.
    4.2 Calculate the background correlation term by Eqs 4.11 and 4.12.
Module 5
    5.1 Calculate the background connectivity by Eqs 4.13-4.17
    5.2 Refine feature contrast map by Eq 4.18
    5.3 Refine spatial compactness term and background correlation term by Eq 4.19
    and Eq 4.20, respectively.
    5.4 Form initial saliency map by Eqs 4.21 and 4.22
    5.5 Refine the initial saliency map by Eq 4.23
Output: Final saliency maps
Module 6
    Evaluate the performance in terms of precision, recall and F-measure.
```

4.4 Experimental Results

For performance evaluation of our proposed saliency detection method, in total 10 state-of-the-art algorithms are used for benchmarking, as listed below by the first letter of the name of methods. They are selected for two main reasons, i.e. high citation and wide acknowledgment in the community and/or newly presented in the recent years. Introduction to the datasets and criteria used for evaluation as well as relevant results and discussions are presented in detail in this Section.

- Bayesian saliency via low and mid-level cues (LMLC) [1]
- Dense and sparse reconstruction (DSR) [18]
- Graph-based manifold ranking (GMR) [13]
- Minimum barrier (MB+) [4]

- Region-based contrast (RC) [3]
- Salient region detection via high dimension colour transform (HDCT) [7]
- Superpixel based saliency (SP) [15]
- Robust background detection (RBD) [22]
- Multiscale Deep Features (MDF) [26]
- Deep Contrast Learning (DCL) [28]

4.4.1 Dataset description

In our experiments, five publicly available datasets including MSRA10K, DUTOMRON, THUR15K, ECSSD, and PASCAL-S are employed for performance assessment.

The MSRA10K dataset [3] contains 10000 images with pixel-level salient object labelling [33, 181]. This database has various image categories such as animals, flowers, humans and natural scenes, et al, where most images contain only one salient object. Given the large size and wide variety of contents, this dataset is very challenging for testing the efficacy of relevant saliency detection approaches. The THUR15K dataset [182] consists of 15000 images, which are divided into five categories: butterfly, coffee mug, dog jump, giraffe, and plane. Since it does not contain a salient region labelled for every image, we only use those labelled images (6232 in total) in our experiment for testing. The DUTOMRON dataset [13] is a very challenging database that contains 5166 images. Each image has one or more saliency objects and complex background. The ground truth is labelled by several experienced participants who are familiar with the goal of saliency detection.

The ECSSD dataset [183] and the PASCAL-S dataset [184] are two other challenging databases. ECSSD has 1000 semantically meaningful images with the complicated background, which is also widely used for saliency detection [33]. PASCAL-S has 850 natural images and contains people, animal, vehicles, and indoor objects. This dataset is widely used to recognise an object from a number of visual object classes in the real world [185].

4.4.2 Evaluation criteria

For quantitative performance assessment of the proposed saliency detection algorithm, several commonly used metrics are adopted in our experiments, which include the precision-recall curve (PR), average precision (AP), receiver operating characteristics (ROC) curve and area under the ROC curve (AUC). By varying a threshold from 1 to 255 and applying it to the determined saliency map, a series of binary images indicating the detected saliency objects can be produced, from which the PR, ROC curve, and AUC can be obtained for quantitative assessment.

The PR curve is formed by the true positive rate (TPR, also namely recall) versus positive predictive value (PPV, also namely precision) and the ROC curve is formed by the false positive rate (FPR) versus TPR. The three rates including TPR, PPV, and FPR are determined by $TPR = \frac{T_p}{T_p+F_n}, PPV = \frac{T_p}{T_p+F_p}, FPR = \frac{F_p}{T_n+F_p}$, where T_p, F_p, T_n and F_n respectively refer to the number of correctly detected foreground pixels of the salient object, incorrectly detected foreground pixels (false alarms), correctly detected background pixels (non-objects) and incorrectly detected background pixels (or missing pixels from the object). Specifically, these four numbers can be calculated by comparing the binary masks of the detected image and the ground truth.

In addition, the F-measure defined below is also used for comprehensive performance assessment:

$$F_{measure} = \frac{(1+\beta) \cdot Precision \cdot Recall}{\beta \cdot Precision + Recall} \tag{4.25}$$

where the parameter β is set to 0.3 to combine the precision and the recall rate as suggested in [20].

4.4.3 Assessment of the obtained saliency maps

To evaluate the performance of the proposed GLGOV method, we show comprehensive comparison results using the PR curve and AUC values on all the five datasets. For subjective assessment, several typical examples with either large objects or complicated backgrounds are shown in Fig. 4.3 for comparison. As can be seen, most of these benchmarking methods fail to highlight the objects as a whole or with a high contrast. However, our proposed method can successfully suppress the background regions and maintain the boundaries of the salient object due to the gestalt law guided cognitive framework. In addition, since the object can be well highlighted, this can further facilitate some potential applications e.g. classification of butterflies (rows 7 and 8) and flowers (row 6) and recognition of people (rows 3 and 4).

For quantitative assessment, the results are evaluated in terms of AP, AUC measurement on five datasets (Table 4.1), the running time for each test image and global performance for each method are listed in Table 4.2. In total there are 10 approaches benchmarked with ours in Table 4.1 and 4.2, where the first eight are unsupervised, and the last two are supervised ones using deep learning. All the

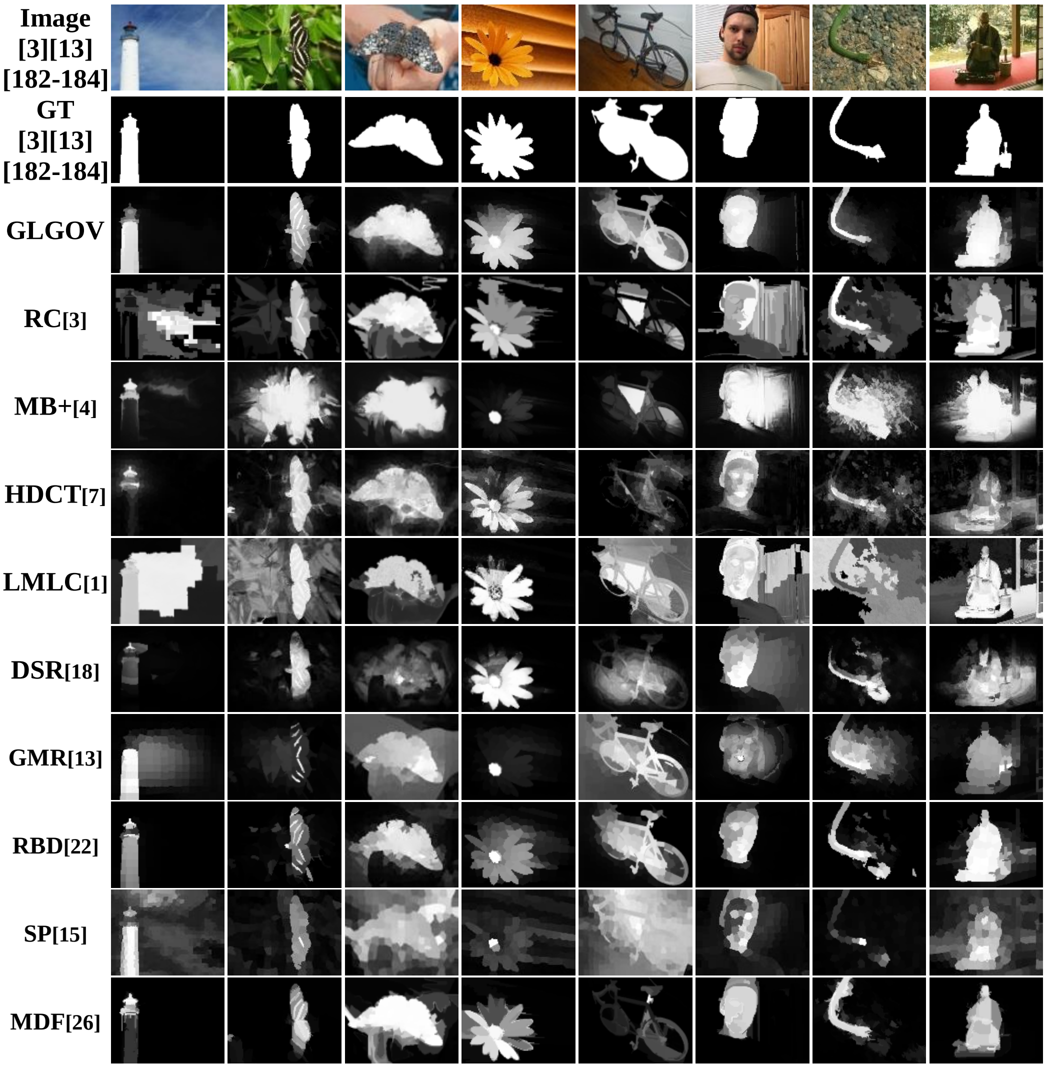


Fig. 4.3 Visual comparison. The ground truth (GT) in shown in the second column.

approaches are tested on a computer with Intel Dual Core i5-4210U 1.7 GHz CPU and 4GB RAM, where for consistency GPU is absent for deep learning-based approaches. Due to hardware configuration reasons, we cannot implement DCL hence we use their published saliency maps which are only available for two datasets, i.e. ECSSD and DUTOMRON.

For unsupervised approaches, the proposed approach yields the highest AUC in all the five datasets, followed by DSR, though the AP value from our approach is slightly less than those of DSR except the MSRA 10k dataset. However, our approach is 3.5 times faster than DSR. This has validated the efficacy of the proposed approach, especially the gestalt laws and background connectivity used in guiding the process of saliency detection. Although MB+ is the third or fourth place in this group of experiments, the extremely high efficiency makes it a good candidate for particular applications, i.e. online processing. It is worth noting that although RC, GMR and RBD have very low running time, the AUC and AP measurements they achieve on these datasets seem inferior. For RC, the reason for the degraded

Table 4.1. AUC and AP score (top two unsupervised methods are highlighted in red and green). *: deep learning based method

Method	**MSRA10K**		**PASCAL-S**		**ECSSD**		**DUTOMRON**		**THUR15K**	
	AUC	**AP**	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**
GLGOV	0.967	0.888	0.868	0.680	0.915	0.773	0.902	0.539	0.905	0.580
MB+ [4]	0.955	0.835	0.859	0.633	0.906	0.695	0.893	0.484	0.900	0.542
RC [3]	0.936	0.838	0.707	0.348	0.893	0.733	0.893	0.503	0.896	0.568
LMLC [1]	0.936	0.721	0.793	0.516	0.850	0.570	0.818	0.374	0.853	0.447
HDCT [7]	0.941	0.784	0.807	0.628	0.868	0.710	0.867	0.506	0.878	0.541
DSR [18]	0.959	0.878	0.866	0.699	0.915	0.788	0.900	0.578	0.902	0.612
GMR [13]	0.954	0.882	0.860	0.680	0.894	0.748	0.894	0.544	0.886	0.57
RBD [22]	0.944	0.876	0.822	0.663	0.890	0.763	0.854	0.526	0.856	0.579
SP [15]	0.923	0.810	0.780	0.576	0.848	0.682	0.837	0.475	0.843	0.518
MDF* [26]	0.973	0.871	0.908	0.762	0.941	0.829	0.917	0.649	0.929	0.639
DCL* [28]	-	-	-	-	0.971	0.897	0.934	0.675	-	-

Table 4.2. Overall and average AUC and AP score (top two unsupervised methods are highlighted in red and green). *: deep learning based method

Method	Time (s)	Overall(OA)		Average(AG)	
		AUC	AP	AUC	AP
GLGOV	1.68	0.930	0.715	0.911	0.692
MB+ [4]	0.05	0.921	0.665	0.903	0.638
RC [3]	0.25	0.905	0.669	0.865	0.598
LMLC [1]	140	0.878	0.557	0.850	0.526
HDCT [7]	4.02	0.899	0.648	0.872	0.634
DSR [18]	5.85	0.925	0.729	0.908	0.711
GMR [13]	0.5	0.893	0.706	0.873	0.681
RBD [22]	0.25	0.874	0.643	0.846	0.612
SP [15]	1.2	0.917	0.701	0.898	0.644
MDF* [26]	200	0.945	0.754	0.934	0.750
DCL* [28]	75	0.940	0.711	0.952	0.786

performance is mainly due to the hard constraints it used to reduce the saliency value near the image borders. These seem to work fine in the MSRA10K dataset. However, given a large number of images along with large variations in terms of image contents and complex background such as PASCAL-S, these constraints become less effective in refining the detected saliency maps. Similar to GMR, it works well on MSRA 10K and ECSSD, but fails to process complicated images such as THUR15K and DUTOMRON. For RBD, due to the lack of effective foreground detection, its precision is not good enough which also leads to its inferior segmentation performance (Table 4.4). For HDCT, the high dimensional colour transform used increases the success of foreground and background separation but increase the running time as well. Moreover, as aforementioned, this method does not totally fit the HVS and results in undesirable results when both the foreground and background contain the same colour elements. For LMLC and SP, regardless of the running time, their performance seems to be quite low due to the lack of effective post-processing to further refine the detected saliency maps. Thanks to our cognitive framework, the way our proposed method detects the salient object fit human visual attention very well. This has helped us suppress the background and highlight the main objects more effectively. Therefore, the results of several datasets show much less

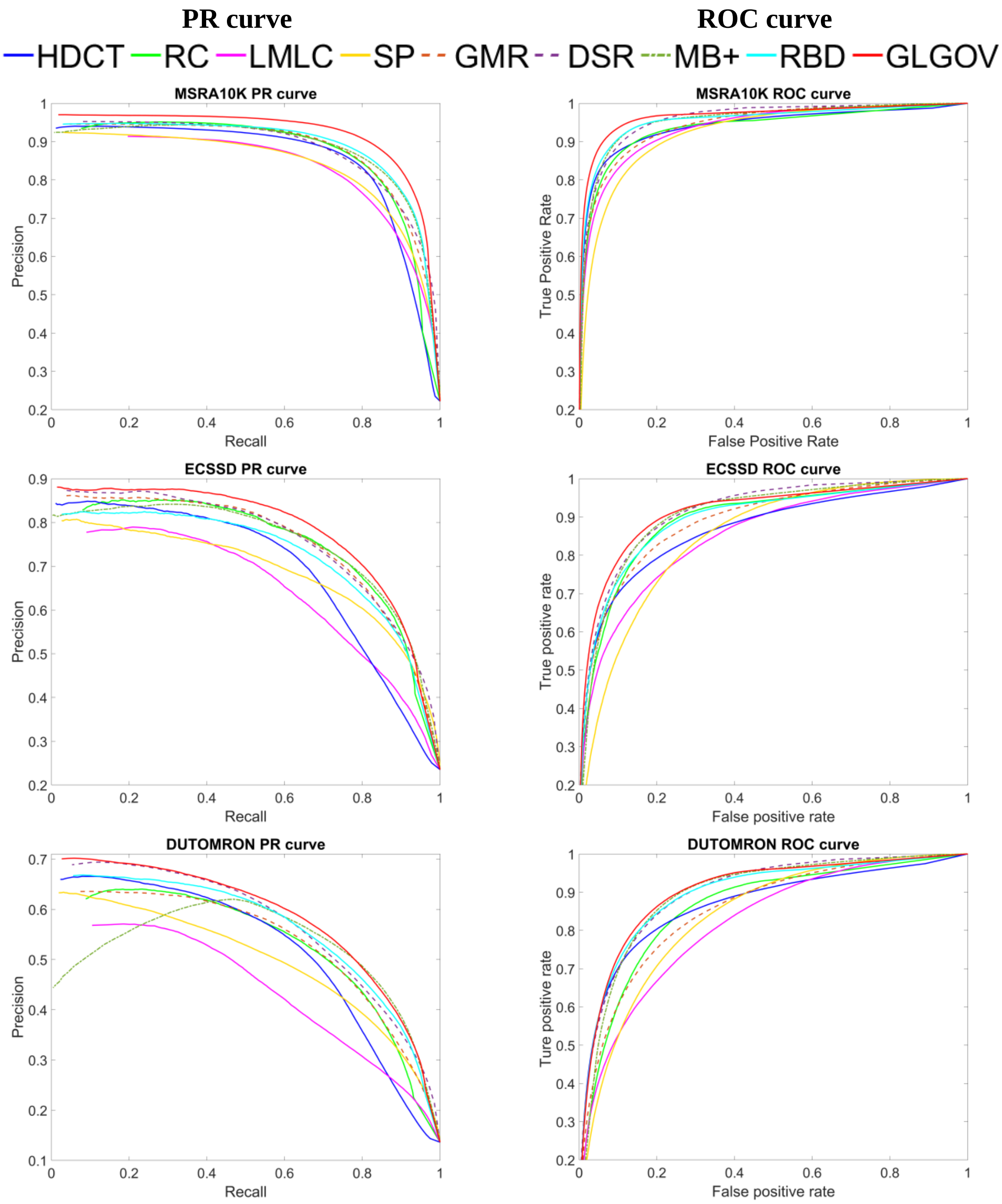


Fig. 4.4 PR and ROC curve

inconsistency. It is believed that our computation cost can be significantly improved by transplanting the code from MATLAB to C++ implementation.

For deep learning based supervised approaches including MDF and DCL, unsurprisingly they produce better results in terms of higher AUC and AP values than unsupervised ones, however, they suffer from lengthy training and testing time, also their results seem sensitive to the learning strategies used. This has constrained their applicability for specific tasks that need a nearly real-time response. For MDF, it gains 2% in AUC and 3.5% in AP than unsupervised approach, possibly due to the

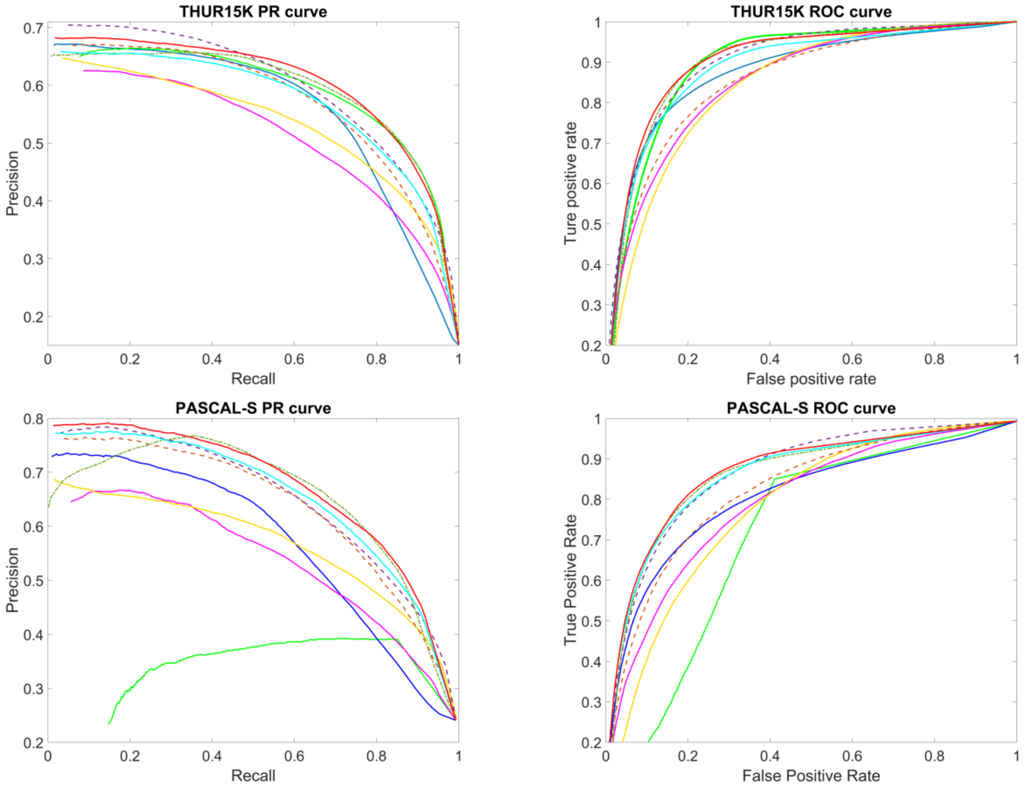


Fig. 4.5 PR and ROC curve (continue)

introduced multiscale CNN deep learning. Nevertheless, for the challenging MSRA10k dataset, MDF gains 0.6% in AUC but loses 1.6% in AP when comparing to our proposed unsupervised approach, despite the fact it uses 25% of the samples i.e. 2500 randomly selected images for training. This again shows the potential limitation of the supervised approach where unsupervised approach can supplement.

To further evaluate the performance of these approaches, we plot in Fig. 4.4 the PR curves and the ROC curves for the results obtained from the five datasets. For better visual effect, we only compare in Fig. 4.4 the results from the unsupervised approaches, as the advantages of deep learning based supervised methods, have been discussed according to the results in Table 4.1 and 4.2. As seen in Fig. 4.4, our approach almost outperforms all other unsupervised methods, especially on the MSRA10K and ECSSD datasets, yet the performance on the rest three datasets appears quite comparable to DSR. Although the curves from MB+ are close to those from our GLGOV approach, the much lower AP as shown in Table 4.1 indicates more false alarms in the detected results.

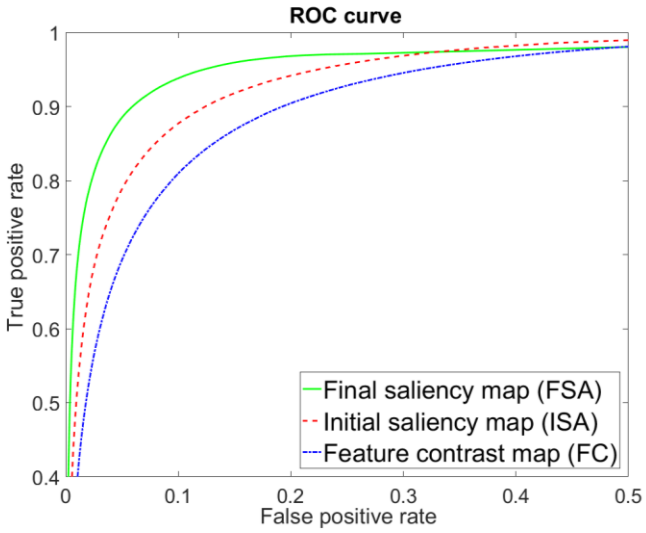


Fig. 4.6 Results of GLGOV with various settings on the MSRA10K dataset.

4.4.4 Key component analysis

In this subsection, we discuss the effect of several key components in the proposed method, where all the evaluations are carried out on the MSRA 10k dataset due to its popularity. As the proposed GLGOV framework is actually a multi-stage approach, in the following we assess the contributions of three major components of our algorithm, which include the feature contrast map FC, initial saliency map after the first refinement ISA, and the final saliency map with the second refinement FSA.

Fig. 4.5 shows the ROC curves obtained from these three settings, where the AUC measurements are given and compared in Table 4.3. For the feature contrast map FC, the result with an AUC at 93.3% seems undesirable. After applying the figure and ground in the first-stage refinement, the AUC reaches 95.58% with an increase of 2.28%. By further adding background connectivity model for the second-stage refinement, the AUC becomes 96.67%, i.e. an additional gain of 1.09%. Meanwhile, the running time has been increased from 1.16s to 1.47s and 1.68s after introducing

Table 4.3. AUC values and running time for our GLGOV approach under various settings on the MSRA 10K dataset.

Components	**FC**	**ISA**	**FSA**
AUC	0.9330	0.9558	0.9667
Time (s/image)	1.16	1.47	1.68

the first and second stage refinement, respectively. This has clearly demonstrated the contribution of the key components in our proposed GLGOV framework.

4.4.5 Validation of image segmentation

Based on the determined saliency maps, the salient objects can be extracted as binary masks, which can be further applied for performance assessment of the saliency detection approaches. Herein the OTSU approach [124] is used for adaptive thresholding to generate the binary masks of salient objects. For quantitative performance assessment, we calculate the average F-measure of all the test images over their ground truth maps and report the results in Table 4.4. As can be seen, our proposed GLGOV model consistently produces the highest F-measure on the MSRA 10k, PASCAL-S and the ECSSD datasets in comparison to other unsupervised peers, and also the second-best on the DUTOMRON and THUR15K datasets after DSR. If taking the five datasets as a whole, our proposed approach outperforms the second best, DSR, 1.5% and 1.1% in terms of the overall and average F-measure, respectively.

It is worth noting that the deep learning based supervised approaches are unsurprisingly high and surpass all unsupervised ones. Nevertheless, there is still some space for further improvement of their learning strategies. For example, the F-

Table 4.4. F-measure of segmented results (top two unsupervised methods are highlighted in red and green).

*: deep learning based method

Method	MSRA10K	PASCAL-S	ECSSD	DUTOMRON	THUR15K	Overall(OA)	Average(AG)
Proposed	0.8810	0.6625	0.7592	0.6041	0.6042	0.7320	0.7022
RC [3]	0.8395	0.4191	0.7277	0.5647	0.5947	0.6926	0.6291
HDCT [7]	0.8348	0.6074	0.7105	0.5947	0.5881	0.7017	0.6671
LMLC [1]	0.7501	0.5328	0.5880	0.4340	0.4816	0.5930	0.5573
SP [15]	0.8106	0.5689	0.6737	0.5367	0.5456	0.6640	0.6271
GMR [13]	0.8521	0.6467	0.7425	0.5959	0.5924	0.7133	0.6859
DSR [18]	0.8374	0.6476	0.7388	0.6205	0.6114	0.7174	0.6911
MB+ [4]	0.8484	0.6614	0.7226	0.5939	0.5977	0.7124	0.6848
RBD [22]	0.8610	0.654	0.7178	0.6011	0.5851	0.7156	0.6838
MDF* [26]	0.8794	0.7346	0.8173	0.6948	0.6628	0.7724	0.7578
DCL* [28]	-	-	0.8968	0.7375	-	0.7633	0.8172

measure from MDF is slightly less than our approach on the MSRA10K dataset, which shows that training on 2500 images seems insufficient to fully learn the characteristics of 10k images. This drawback is possibly overcome in DCL, as DCL seems more effective than MDF in the tested two datasets, ECSSD and DUTOMRON. Again, it shows that the performance of deep learning-based approaches can be very sensitive to the learning strategy used, regardless the extremely high computational resources and computational cost needed.

4.5 Conclusion

Inspired by both Gestalt laws optimization and background connectivity theory, in this chapter, we proposed GLGOV as a cognitive framework to combine bottom-up and top-down vision mechanisms for unsupervised saliency detection. Experimental results over five publicly available datasets have shown that our method helps to produce the best overall accuracy and average accuracy when benchmarking with a number of state-of-the-art unsupervised techniques. Additional assessments in terms of the PR curve, ROC curve, F-measure, AUC, and AP have also verified the efficacy of the proposed approach.

The most important finding in this work is the efficacy of bottom-up and top-down mechanisms for saliency detection, which are actually guided by necrologies such as Gestalt laws and background connectivity. On the one hand, the aim of saliency detection is to enable computers to recognise the salient object like a human. On the other hand, Gestalt laws are the main theories that describe the mechanism of HVS, whilst background connectivity can reflect our visual cortex reaction to stimuli. As such, these necrologies can be well introduced into the process of and support the modeling of saliency detection. Our outcomes showed that with the guidance of necrologies, the proposed unsupervised saliency methodology consistently produces good results on different datasets. Although there is still some gap to deep learning based supervised approaches, the unsupervised approach may supplement in cases there are no sufficient training samples and/or with limited computational resources.

For future work, we will focus on more in-depth guidance from Gestalt laws on saliency detection, where the laws of closure and continuity can be injected to further improve the performance. Texture feature and deep learning models will also be considered for saliency detection beyond colour contrast where semi-supervised or weakly supervised learning can be explored.

CHAPTER 5

5 Cognitive Fusion of Thermal and Visible Imagery for Effective Detection and Tracking of Pedestrians in Videos

5.1 Introduction

Pedestrian detection is an important task in urban surveillance, which can be further applied to pedestrian tracking and recognition. In general, visible and thermal imagery are two popularly used data sources, although not necessarily in a combined solution. However, either visible image or thermal image has their advantages and disadvantages. In general, coloured visible image in red-green-blue (RGB) has better distinguishability in human visual perception, yet it suffers from the effect of illumination noise and shadows. On the contrary, the thermal images don’t have colour and texture information but intensity instead. However, they are robust against illumination effects through its distinguishability varies according to environmental settings but very sensitive to surface temperature. For the purpose of pedestrian detection, background subtraction plays an important role. However, for either visible or thermal data, it may take high computational cost to obtain good results because of these disadvantages. To this end, we present an efficient framework to cognitively combine these two modalities of images for improved detection and tracking of salient objects from videos.

In this chapter, we proposed a two-stage background subtraction procedure based on human cognition knowledge on both visible and thermal images for fusion based pedestrian detection. In the first stage, we predict the background model by computing the median value of randomly selected frames in the videos and apply an adaptive threshold to detect binary foreground map along with knowledge-based morphological refinement. In the second stage, we employ an adaptive Gaussian mixture model to estimate the background model and generate the binary foreground map. Then, inspired by some previous works which put multi-modality image fusion concept into many application such as saliency detection and image registration

[186-188], our final foreground map can be obtained by the fusion of both visible and thermal binary maps. In addition, to deal with cases of occlusion or overlap, we also improved mean-shift tracking method to have a capability of scale change and identify the individual pedestrian template from a pedestrian group more efficiently. To evaluate the proposed method, a publicly available colour-thermal benchmark dataset OCTBVS[189] is employed here. For our foreground detection evaluation, objective and subjective analysis against several state-of-the-art methods have been done on our manually segmented ground truth. For our object tracking evaluation, comprehensive qualitative experiments have also been done on all video sequences. Promising results have shown that the proposed fusion based approach can successfully detect and track multiple human objects in most scenes regardless of any light change or occlusion problem.

This chapter makes the following three contributions:

- Based on the cognitive knowledge that colour and intensity information plays important role in our perception, we use adaptive Gaussian mixture model (GMM) to measure the distribution of colour and intensity in multi-modality images (RGB images and thermal images) and then integrate them together to build the background subtraction model.
- Inspired by biologically knowledge that human beings' shape is almost close to a rectangle shape, the morphological refinement with rectangle-shaped structure is integrated into fusion strategy to generate our final foreground map.
- Inspired by human being behavior knowledge that pedestrians usually have unique motion, constrained mean-shift is proposed to detect the single person from the group.

The rest of the chapter is organised as follows: Section 5.2 illustrates the framework of the proposed method. Section 5.3 describes the foreground detection approach. Section 5.4 elaborates the object tracking method. Experimental results are presented and discussed in Section 5.5. Finally, some concluding remarks and future work are summarized in Section 5.6.

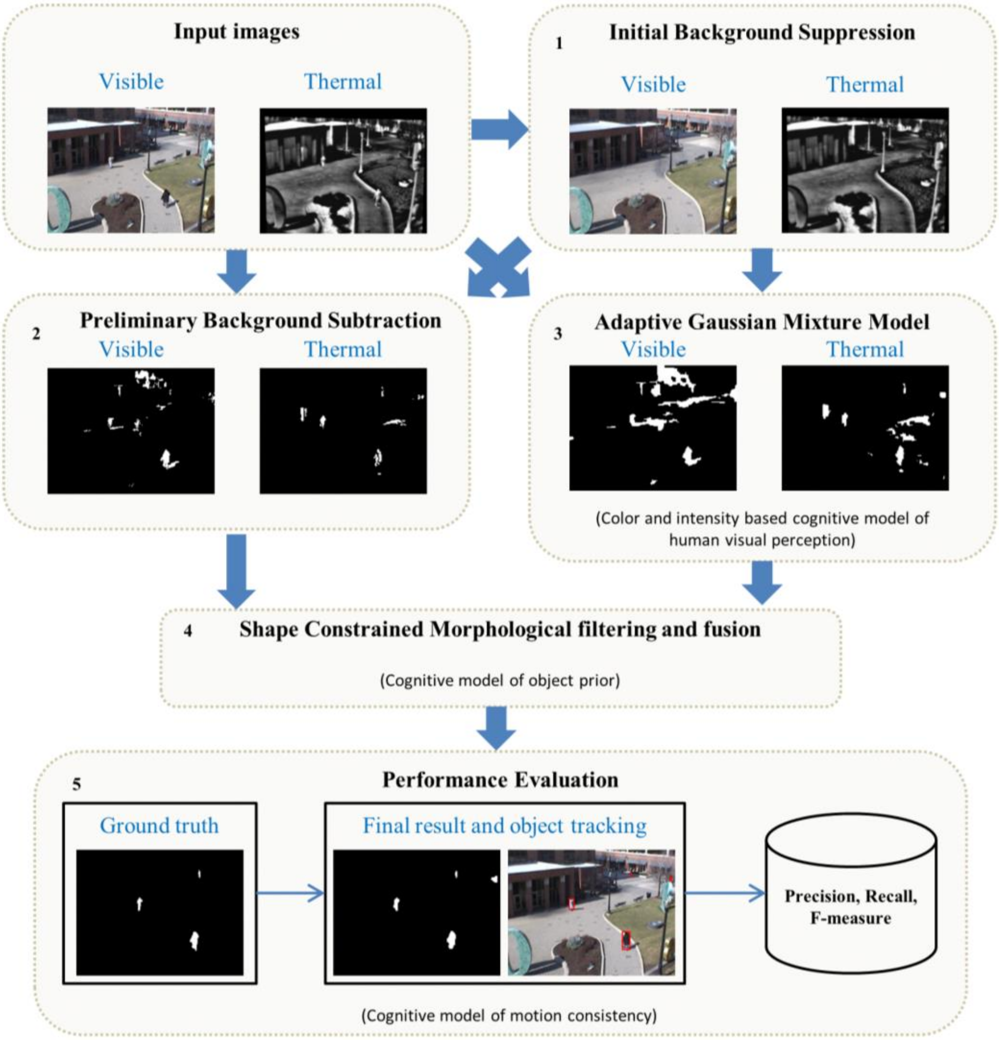

Input images
Visible
Thermal
1
Initial Background Suppression
Visible
Thermal
2
Preliminary Background Subtraction
Visible
Thermal
3
Adaptive Gaussian Mixture Model
Visible
Thermal
(Color and intensity based cognitive model of human visual perception)
4
Shape Constrained Morphological filtering and fusion
(Cognitive model of object prior)
5
Performance Evaluation
Ground truth
Final result and object tracking
Precision, Recall, F-measure
(Cognitive model of motion consistency)

Fig. 5.1. Proposed framework within five modules.

(where the selected images are from OCTBVS[189])

5.2 Overview of the proposed system

In this chapter, we proposed a two-stage background subtraction procedure based on human cognition knowledge on both visible and thermal images for fusion based pedestrian detection, and five modules are included in (Fig.5.1). In the first stage, we predict the background model by computing the median value of randomly selected frames in the videos (module 1) and apply an adaptive threshold to detect binary foreground map along with knowledge-based morphological refinement (module 2). In the second stage, we use the results from module 1 as prior frames and employ learning based adaptive Gaussian mixture model to estimate the background model and generate the binary foreground map (module 3). Then the initial and Gaussian-based foreground maps of both visible and thermal images will be refined by shape constrained morphological filtering and further fused together to get the final foreground map (module 4). In the performance evaluation (module 5), the proposed background subtraction method is compared against a number of state-of-the-art methods on a widely used publicly available video sequences. Some widely used evaluation criteria such as precision, recall, and F-measure are used for quantitative assessment. In addition, we also proposed constrained mean-shift tracking method to have a capability of scale change and identify the individual pedestrian template from a pedestrian group more efficiently (detailed in Section 5.3). Furthermore, the performance of object tracking is also evaluated by qualitative assessment. Detailed results are reported in Section 5.4.

5.3 Foreground detection

In this Section, a two-stage foreground detection method is applied for both visible and thermal images. Eventually, the desired foreground map is fused by the foreground detection results of two types of images with cognition based morphological process.

5.3.1 Random median background subtraction

To capture the initial region of pedestrians in the visible and thermal image, we first estimate the background model by computing a median map (Fig.5.1 Module 1) of N frames randomly selected from the video sequence. In other words, every pixel in the estimated background map is determined as the median value of all the pixels

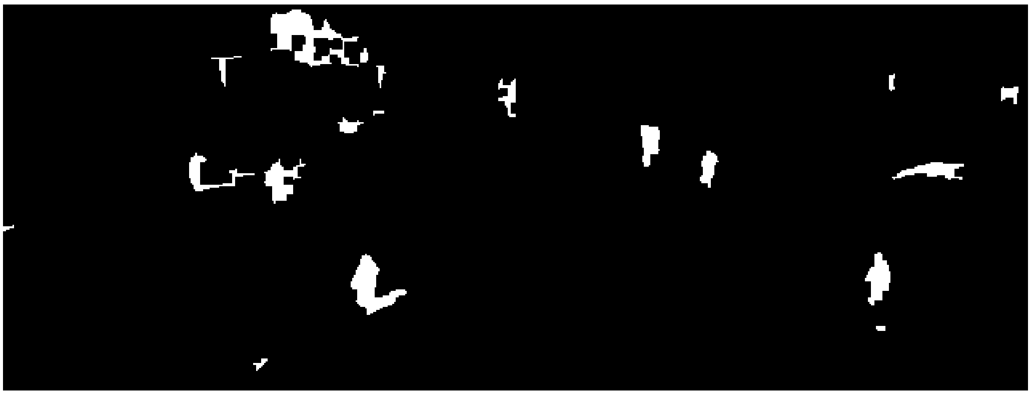


Fig. 5.2. The refined initial background subtraction results of visible (left) and thermal (right) images.

collected from N sequential frames at the same location. The initial background subtraction process for each visible or thermal frame is defined as:

$$\mathrm{BS}_{ini} = \left|\mathrm{I}_{frame} - \mathrm{I}_{med}\right| \tag{5.1}$$

where I_{frame} is the image of every frame, I_{med} is the median map of N randomly selected frames.

After that, we binarize the BS_{ini} with an adaptive threshold, i.e., OTSU [124] to get a binary image I_{bi} with coarse human body region (Fig.5.1 Module 2). However, I_{bi} contains many ambiguous contents and some objects that should be detected as a whole are fractured. Therefore, a cognitive-based morphology refinement is applied here to filter the insignificant region and integrate the potential objects. Since the object that we want to detect is pedestrian, and we can assume the shape of the pedestrians is an ellipse or a rectangle based on our cognition so that its major axis length is usually larger than minor axis length. Therefore, in our morphology refinement, we define a rectangle-shaped structuring element to connect separated regions together to be a whole object. The width and height of the rectangle is defined as $2n + 1$ and $2n + 3$ $(n \in Z_0^+)$, respectively. Here we set n as 1. Furthermore, as the size of a pedestrian in the video will not be small, we filter those noise regions by an empirical threshold (set as 0.7 in the chapter). From Fig.5.2, we can see in the refinement result I_r, the noise regions with the small area have been removed and every object has been integrated.

5.3.2 Adaptive mixture background subtraction

Although the random median background subtraction module can detect some potential objects, it still contains many false alarms due to lack of the analysis of scene changes, lighting changes and moving object, etc. Therefore, a learning-based background mixture model is employed here to estimate the foreground map under real scene. For a particular surface under particular lighting, a single Gaussian per pixel is sufficient to represent the pixel value. However, in practice, there are multiple surfaces due to the lighting conditions change. Thus, in order to fit the real-world situation and our human cognition, multiple adaptive Gaussians are necessary.

In this chapter, we model each pixel by a mixture of *K* Gaussian distributions and *K* is 5 as suggested in [138]. The probability of an observed pixel X_t at time t as background can be written as

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t; \mu_{i,t}, \Sigma_{i,t}) \tag{5.2}$$

where $\omega_{i,t}$ is the weight parameter of the i^{th} Gaussian in the mixture at time t, $\mu_{i,t}$ and $\Sigma_{i,t} = \sigma_i^2 I$ are the mean value and covariance matrix of the i^{th} Gaussian in the mixture at time t. $\eta(*)$ is the normal distribution of i^{th} Gaussian component.

$$\eta\left(X_t; \mu_{i,t}; \Sigma_{i,t}\right) = \frac{1}{(2\pi)^{\frac{D}{2}}\left|\Sigma_{i,t}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(X_t - \mu_{i,t}\right)^T \Sigma_{i,t}{}^{-1}\left(X_t - \mu_{i,t}\right)} \tag{5.3}$$

Then first B distributions are chosen as the background model

$$B = argmin_b(\sum_{i=1}^{b} \omega_i > T) \tag{5.4}$$

T is the minimum portion of the data that should be counted as background. For any new observed pixel value, X_t will be considered as foreground if it is more than 2.5 standard deviations away from existing *B* distributions.

The initial weights of *ith* distributions at time *t* is rewritten as:

$$\omega_{i,t} = (1-\alpha)\omega_{i,t-1} + \alpha\hat{p}(\omega_{i,t}|X_t) \tag{5.5}$$

where the parameters μ and Σ denote, respectively, the mean and covariance matrix of the first Gaussian component that matches the new observed pixel value, which will be updated by the following process:

$$\mu_{i,t} = (1-\alpha)\mu_{i,t-1} + \rho X_t \tag{5.6}$$

$$\Sigma_{i,t} = (1-\alpha)\Sigma_{i,t-1} + \rho(X_t - \mu_{i,t})\left(X_t - \mu_{i,t}\right)^T \tag{5.7}$$

$$\rho = \alpha\eta(X_t; \mu_{i,t}, \Sigma_{i,t}) \tag{5.8}$$

$$\hat{p}\left(\omega_{i,t}\middle|X_t\right) = \begin{cases} 1 & ; if\, \omega_{i,t}\ matches\ first\ Gaussian\ component \\ 0 & ; otherwise \end{cases} \tag{5.9}$$

where α is the learning rate. In this chapter, we choose 0.002 for α and 0.7 for T according to experimental results.

In addition, we use 10 random median background subtraction results to predict the initial value of the parameters $\omega_{i,t}, \mu_{i,t}\ and\ \Sigma_{i,t}$ for better performance. After the adaptive background mixture model is done, we can get the foreground map ${I_a}^{vis}$ and ${I_a}^{thm}$ of visible and thermal images (Fig.5.1 Module 3).

5.3.3 Fusion strategy

In order to generate the final foreground map and make the fusion result close to human perception, we add a morphological refinement to the results from the previous stage and integrate them together. For foreground map of visible and thermal image generated in Section 5.3.1 (i.e., ${I_r}^{vis}$and ${I_r}^{thm}$), and Section 5.3.2 (i.e., ${I_a}^{vis}$ and ${I_a}^{thm}$), we define a function $D(\cdot)$ that can dilate all the potential objects with a rectangle-shaped structuring element. We set *n*=0 because we want to smooth the edge for each object and connect the small gap between some object pieces. By doing so, the shape of the object will have continuity, which matches human perceptions. Then the foreground map of visible and thermal image can be built by fusion strategy as follows:

$$I_{vis} = ({I_a}^{vis} \cap D({I_r}^{vis})) \cup ({I_r}^{vis} \cap D({I_a}^{vis})) \tag{5.10}$$

$$I_{thm} = ({I_a}^{thm} \cap D({I_r}^{thm})) \cup ({I_r}^{thm} \cap D({I_a}^{thm})) \tag{5.11}$$

For visible images, the first term ${I_a}^{vis} \cap D({I_r}^{vis})$ in I_{vis} is used to improve ${I_a}^{vis}$ by integrating the morphological refinement of ${I_r}^{vis}$, and the second term is to improve the ${I_r}^{vis}$ by integrating the morphological refinement of ${I_a}^{vis}$. In this way, the noise in ${I_a}^{vis}$ and ${I_r}^{vis}$ can be removed and end up with smoothed shapes for the regions of interest. The two terms are fused together to generate the visible foreground map. Similar to the thermal image, two foreground maps ${I_a}^{thm}$ and ${I_r}^{thm}$ are refined and fused together to produce the thermal foreground map

Finally, the overall foreground map (Fig.5.1 Module 6) can be calculated by same fusion strategy as follows:

$$I_{final} = (I_{vis} \cap D(I_{thm})) \cup (I_{thm} \cap D(I_{vis})) \tag{5.12}$$

5.4 Object tracking

After background subtraction, a number of candidate regions can be derived. For any continuous frames, if the later frame has fewer candidate regions than the former frame, there will be only two situations. The first situation is one or more candidate regions in the former frame have been out of the later frame, and the other situation is some individual candidate regions in the later frame are detected as a whole in the foreground detection stage due to the inevitable overlap and occlusion problem. Fig. 5.3 (a,b) are two adjacent frames where there should be three pedestrian patterns detected in both frames. However, the background subtraction method considers the left two patterns in frame #2 as one candidate region, because of the small distance between them. Therefore, an improved mean-shift method is proposed in this section to track the individual objects in the second situation.

Conventional mean-shift method [190] has two main drawbacks. The first is that it tracks objects mostly based on the colour and texture features, where the spatial relationship among the objects is seldomly considered. Therefore, if an object's colour is similar to the surrounding background, it may be (partially) treated as background by the tracker in the following frame. For example, in Fig.5.3(c), the conventional mean-shift cannot locate and track two pedestrian patterns accurately. The second one is the similarity computation for probability density functions (PDFs) of candidate region and object region. In [190], it defines the distance between two PDFs as

$$\mathrm{d}(\mathrm{y}) = \sqrt{1 - \rho[\hat{\mathrm{p}}(\mathrm{y}), \hat{\mathrm{q}}]} \tag{5.13}$$

$$\rho[\hat{\mathrm{p}}(\mathrm{y}), \hat{\mathrm{q}}] = \sum_{\mathrm{u}=1}^{\mathrm{m}} \sqrt{\hat{\mathrm{p}}_{\mathrm{u}}(\mathrm{y})\hat{\mathrm{q}}_{\mathrm{u}}} \tag{5.14}$$

where $\rho[\cdot]$ is the Bhattacharyya coefficient, $\hat{\mathrm{q}} = \{\hat{\mathrm{q}}_{\mathrm{u}}\}_{\mathrm{u}=1\ldots\mathrm{m}}$ (with $\sum_{\mathrm{u}=1}^{\mathrm{m}} \hat{\mathrm{q}}_{\mathrm{u}} = 1$) is the discrete density from the *m*-bin histogram of the object region, $\hat{\mathrm{p}}(\mathrm{y}) = \{\hat{\mathrm{p}}_{\mathrm{u}}(\mathrm{y})\}_{\mathrm{u}=1\ldots\mathrm{m}}$ (with $\sum_{\mathrm{u}=1}^{\mathrm{m}} \hat{\mathrm{p}}_{\mathrm{u}}(\mathrm{y}) = 1$) is estimated as a given location y from the *m*-bin histogram of the candidate region. However, $\hat{\mathrm{q}}$ does not change with time which is not fit with human cognition because the surrounding of the object cannot be always same in the real scene. On the other hand, unchangeable $\hat{\mathrm{q}}$ will also increase the convergence cost because it will take more time to match object candidate and object model within difference background.

To overcome two problems mentioned above, we propose constrained mean-shift method where two improvements are introduced in the following. Firstly, the object model is updated in each frame in order to get real-time $\hat{q}$. Thus, the size of the $\hat{q}$ will change with the scale changing of the object. Meanwhile, the pedestrians usually move slowly which means their surrounding background in adjacent frames will not be changed too much. In this case, $\hat{p}(y)$ can be quickly matched with $\hat{q}$ in each frame. Secondly, we limit the shift range with the spatial information of the objects in adjacent frames. We define F_{n-1} and F_n as frame n-1 and frame n, R_{n-1}^{i} is the region i in F_{n-1}, and R_n^{j} is the region j in F_n, $X_{n-1}^{i,1}, X_{n-1}^{i,2}, Y_{n-1}^{i,1}, Y_{n-1}^{i,2}$ are the location elements of R_{n-1}^{i}, and $X_n^{j,1}, X_n^{j,2}, Y_n^{j,1}, Y_n^{j,2}$ are the location elements of R_n^{j}. After the location of R_{n-1}^{i} candidate in F_n (expressed as $X_n^{i,1}, X_n^{i,2}, Y_n^{i,1}, Y_n^{i,2}$) is determined by a conventional mean-shift algorithm in every iteration, we further refine this location by a displacement term represented as λ_x, λ_y.

Let $\lambda_x^{i,1} = X_n^{j,1} - X_n^{i,1}, \lambda_x^{i,2} = X_n^{j,2} - X_n^{i,2}, \lambda_y^{i,1} = Y_n^{j,1} - Y_n^{i,1}$ and $\lambda_y^{i,2} = Y_n^{j,2} - Y_n^{i,2}$ be the displacement terms, the new position of the object can be determined by using these displacement terms as follows.

$$\begin{cases} X_n^i = X_n^i + \lambda_x^{i,1}, & if\, \lambda_x^{i,1} > 0 \\ X_n^i = X_n^i + \lambda_x^{i,2}, & if\, \lambda_x^{i,2} < 0 \end{cases} \tag{5.15}$$

$$\begin{cases} Y_n^i = Y_n^i + \lambda_y^{i,1}, & if\, \lambda_y^{i,1} > 0 \\ Y_n^i = Y_n^i + \lambda_y^{i,2}, & if\, \lambda_y^{i,2} < 0 \end{cases} \tag{5.16}$$

As can be seen from the Fig. 5.3, define region 1 and region 2 in frame 1 are two individual object models, the corresponding object candidate should be limited in the region 3 in frame 2. In this case, the object group in the frame can be tracked separately in region 4 and 5 (shown in the right image in Fig. 5.3).

The Pseudocodes for our proposed pedestrian tracking system are shown in Algorithm 2.

```
Algorithm 2: RGBT video tracking
Input: visible and thermal video frames
    1. Calculate the initial background maps for both the visible and thermal videos
       by using the median filter over a number of randomly selected frames.
    2. Generate adaptive background subtraction results using Eqs 5.2-5.9.
    3. Fuse the foreground map of visible and thermal frames by Eqs 5.10-5.12.
Output: Final foreground map
    4. Separate the individual pedestrians from the groups by using constrained
       mean-shift method.
Output: Tracking results of every pedestrian.
```

5.5 Experimental results

5.5.1 *Dataset description and evaluation criteria*

To evaluate the performance of our foreground detection and object tracking methods, a publicly available database 03 OSU Colour-Thermal Database from OCTBVS [189] are employed here. Thermal sequences are captured by Raytheon PalmIR 250D thermal sensor and colour sequence are captured by Sony TRV87 Handycam colour sensor. All the frames in both sequences have a spatial resolution of 320*240 pixels. The number of frames in each video sequence is Sequence-1:2107, Sequence-2:1201, Sequence-3:3399, Sequence-4:3011, Sequence-5:4061, and Sequence-6:3303, respectively. Fig.5.3 shows some visible and thermal frames [189] and the results of our foreground detection method. For our foreground detection method, we do both qualitative (Fig.5.4) and quantitative (Table 5.1) analysis against

Table 5.1 Comparison of Precision, Recall and F-measure values

Methods	Precision	Recall	F-measure	Time(s)
GMG[2]	0.704	0.702	0.703	0.027
IMBS[5]	0.370	0.744	0.495	0.034
LOBSTER[8]	0.730	0.722	0.726	0.075
MultiCue[10]	0.260	0.888	0.403	0.033
SuBSENSE[12]	0.693	0.769	0.729	0.101
T2FMRF [16]	0.508	0.299	0.377	0.036
ViBe[19]	0.741	0.644	0.689	0.019
FA-SOM[21]	0.385	0.829	0.525	0.028
PBAS[25]	0.730	0.336	0.460	0.041
Proposed	**0.702**	**0.880**	**0.781**	**0.029**

Table 5.2 Key parameter α analysis

Learning Rate	Precision	Recall	F-measure
0.001	0.696	0.887	0.780
0.002	**0.702**	**0.880**	**0.781**
0.003	0.711	0.850	0.774
0.004	0.710	0.811	0.757

9 state-of-the-art methods i.e. GMG[2](derived from the first letter of three authors' family name: Godbehere, Matsukawa, and Goldberg), IMBS[5](Independent multimodal background subtraction), LOBSTER[8](local binary similarity patterns), MultiCue[10], SuBSENSE[12](self-balanced sensitivity segmenter), T2FMRF [16](type-2 fuzzy markov random field, ViBe[19](visual background extractor), FA-SOM[21](fuzzy adaptive SOM) and PBAS[25](pixel-based adaptive segmenter) on some manually segmented silhouettes. For our object tracking method, we do comprehensive qualitative experiments on all video sequences [189] (Fig.5.5).

For quantitative performance assessment of the proposed foreground detection algorithm, several commonly used metrics are adopted in our experiments, which include the precision, recall, F-measure. The precision value P and recall value R is determined by $P = \frac{T_p}{T_p + F_p}, R = \frac{T_p}{T_p + F_n}$, where T_p, F_p, and F_n respectively refer to the number of correctly detected foreground pixels of the pedestrians, incorrectly detected foreground pixels (false alarms), and incorrectly detected background pixels (or missing pixels from the object). Specifically, these three numbers can be calculated by comparing the binary masks of the detected image and the ground truth. Furthermore, since the database does not have ground truth, we obtain a manual segmentation of the pedestrian regions in 53 frames from Sequence 1. The F-measure is defined by: $F_{measure} = \frac{2 \cdot P \cdot R}{P + R}$.

5.5.2 Key parameter selection

We carefully choose the key parameter by investigating their changing on the performance. Initially, we choose the default parameter setting by theoretical and practical support. The parameter α and T in adaptive Gaussian mixture model are suggested to set as 0.002 (500 recent frames) [191] and 0.7 (it was 0.6 as suggested

Table 5.3 Key parameter T analysis

Threshold *T*	Precision	Recall	F-measure
0.4	0.688	0.892	0.777
0.5	0.688	0.892	0.777
0.6	0.691	0.891	0.778
0.7	**0.702**	**0.880**	**0.781**
0.8	0.723	0.825	0.771
0.9	0.763	0.593	0.667

in [191], but 0.7 is the best in this chapter). And Gaussian distribution number *K* is set as 5 as suggested in [138]. Table 5.2-5.4 measure the performance by changing the learning rate α, threshold of background portion and number of Gaussian distribution number, respectively. From Table 5.2, we can see that the precision will slightly increase with the raising of the learning rate and the recall shows the inverse trend against the precision. Since the learning rate decides how many recent frames are used to learn, and the larger the learning rate is, the less the recent frames are used. Generally, if we use less recent frames to predict the background, the local information will be more detailed. On the contrary, more recent frames will make the background have more global property. Therefore, the learning rate cannot be too large or too small and it is set as 0.002 in this chapter based on our practical measurement and [191]. From Table 5.3, we can find that the precision grows with the increasing of the T and recall still has the opposite tendency against the precision. The reason for that is when the portion of background is increased, some foreground regions and other noise might be considered as background. Although it somehow makes the precision increase, the recall will reduce sharply. However, if the portion of background is too small, many noises will be considered as foreground and the precision will be reduced due to the growth of the false alarm and the recall will not increase too much. In general, *T* is set as 0.6 as suggested in [191] which is good enough, but we choose 0.7 (70% of the time the background is present) here for the better performance. Table 5.4 shows that the number of Gaussian distribution does not make the performance too much difference as long as it is larger than 2. Therefore, we just follow the suggestion in [138] and set *K* as 5.

Table 5.4 Key parameter *K* analysis

Number of K	Precision	Recall	F-measure
2	0.688	0.892	0.777
3	0.696	0.889	0.780
4	0.701	0.880	0.781
5	**0.702**	**0.880**	**0.781**
6	0.702	0.880	0.781
7	0.702	0.880	0.781

5.5.3 Assessment of foreground detection method

To evaluate the quality of the extracted foreground map, we compare our proposed method with six state-of-art methods in terms of precision, recall and F-measure as the performance metrics with the results shown in Table 5.1. For fair comparison, instead of just comparing our fusion result with others' results on visible images, we do the same fusion strategy for each method where I_{vis} and I_{thm} are generated by those methods on visible and thermal images respectively. From the Table 5.1, we can see the precision of proposed foreground detection is not the highest, but the F-measure of our method outperforms other methods and our recall is the second best. IMBS, MultiCue, T2FMRF and FA-SOM yield bad performance due to their algorithms does not take too much account of the scene change. Although their methods work well in some indoor and outdoor data, those data do not have too much light change. However, in the 03 OSU Colour-Thermal Database from OCTBVS, the clouds make the big shadow on the ground and the light of the scene changes as time goes by. PBAS model the background by a history of recently frames. Although it updates the background over time to deal with gradual scene change, the update relies on the learning parameter. Therefore, in this database, it cannot yield as good results as reported in the original chapter. GMG, LOBSTER, SuBSENSE and Vibe almost have similar performance and very comparable with our proposed method.

However, these methods are mainly designed for the object detection in the small scene. And the objects in small scene usually have large size than the pedestrians in a surveillance system. Therefore, these methods can detect the pedestrians within close or middle range but not long range from the camera. In addition, affected by light change and weather condition, some details have been lost. As can be seen in the visible image in Fig.5.4, some pedestrians' shapes in GMG are not integrated; some pedestrians' shapes in GMG are fractured e.g. left person in the first image is split into two regions; some pedestrians that far away from the camera cannot be detected in SuBSENSE e.g. the 5th and 6th images. Hence, these methods have good quantitative results, but their qualitative results do not fit human's cognition. However, our foreground detection result generated by two-stage background subtraction procedure and fusion strategy where cognition-based knowledge is applied in to refine the procedure and guide the fusion strategy.

Although our proposed method yields the best performance in terms of F-measure, there are still rooms for further improvements. As seen, our proposed method has produced high recall value but relative low precision value just like other methods. There are two main reasons, i.e. missing detection and inaccurate ground truth mapping. For the cases of missing detection, this is mainly due to the failure in detecting objects dressing in the similar colour to the background whilst also behind obstacles. This can be possibly improved by introducing certain post-processing such as back-tracking. However, it can still be challenging in dealing with small objects which are frequently grouped together. This also explains the low accuracy of ground truth as in some cases the silhouettes of the pedestrians can be hardly defined accurately even in a manual way.

5.5.4 Assessment of object tracking method

To validate the performance of the proposed object tracking approach, all video sequences are used in our experiments. In Fig.5.5, detection and tracking results from these sequences are given to illustrate the extracted/tracked objects using their bounding boxes. As can be seen, the proposed method can give reliable pedestrian detection and tracking results under various conditions, including occlusion and light changes in terms of illumination and scale. When the pedestrians are independent, we can detect them very well with proper scale bounding box. We can also identify the people even they are overlapped such as the 1st and 3rd images in Fig. 5.5(d). In addition, when there is some occlusion like tree or wall such as the 2nd and 3rd images in Fig. 5.5(d), the 1st image in Fig. 5.5(e), 2nd and 3rd images in Fig. 5.5(f), the 6th image in Fig. 5.5(b) and 6th image in Fig. 5.5(c),etc. For the object is getting out of the screen such as the 3rd image in Fig. 5.5(e), 4th and 6th images in Fig. 5.5(f), we can still locate the objects and track their motion.

However, some failure cases such as 3rd and 4th images in Fig.5.5 (a), the 2nd image in Fig.5.5 (c) still exist in our tracking results. There are two main reasons, and the first one is that some pedestrians always walk together as a group from beginning to the end in the sequence, therefore, our tracking system always consider the pedestrian group as a single object. The second reason is that if one pedestrian leaves one pedestrian group and join in another pedestrian group, the tracking system cannot extract its own colour, texture and spatial feature, as a result, the mean-shift method cannot calculate its track.

5.6 Conclusion

In this chapter, we proposed a cognitive model by fusing visible and thermal images for pedestrian detection, along with an improved mean-shift method applied to track the pedestrians in videos. There are three key components in this model, i.e. foreground detection, fusion based object refinement and object tracking. By estimating the background model followed by two−stage background subtraction, foreground objects can be successfully detected. Shape constrained morphological filtering based fusion strategy helps to further refine the detected foreground objects. Finally, prediction based forward and backward tracking is found particularly useful to separate overlapped or occluded objects, and robust to the scale change. However, if some pedestrian in the pedestrian group cannot be detected individually from the beginning to the end, the tracking system will fail to estimate its own track and just estimate track of its group instead. In future work, deep learning model will be added to further enhance both foreground detection nd improve tracking procedure for precisely estimation of the objects' track even with some challenging situations.

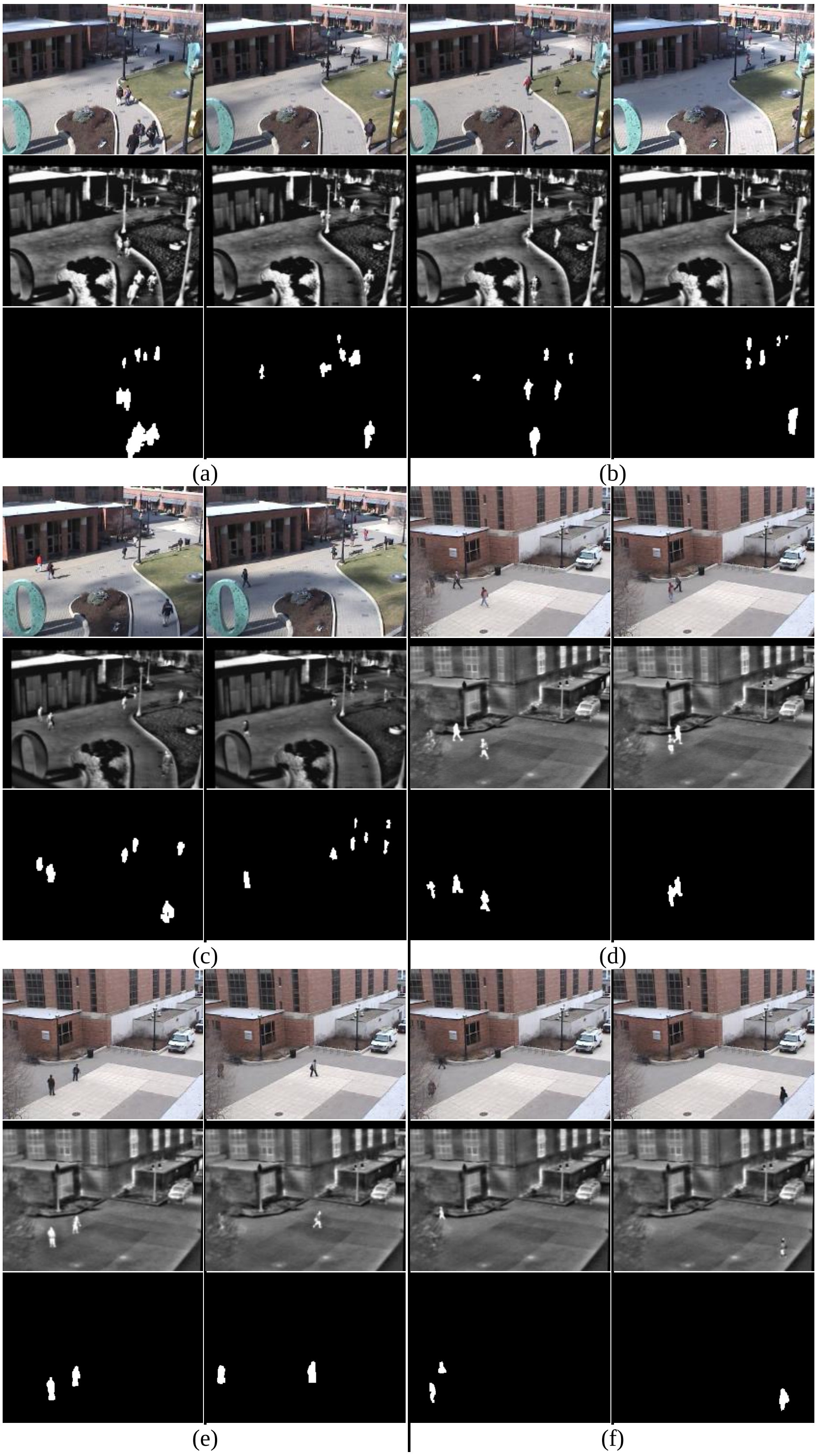


Fig. 5.3. Visual results of proposed foreground detection algorithm, (a) Sequence-1. (b) Sequence-2. (c) Sequence-3. (d) Sequence-4. (e) Sequence-5. (f) Sequence-6.

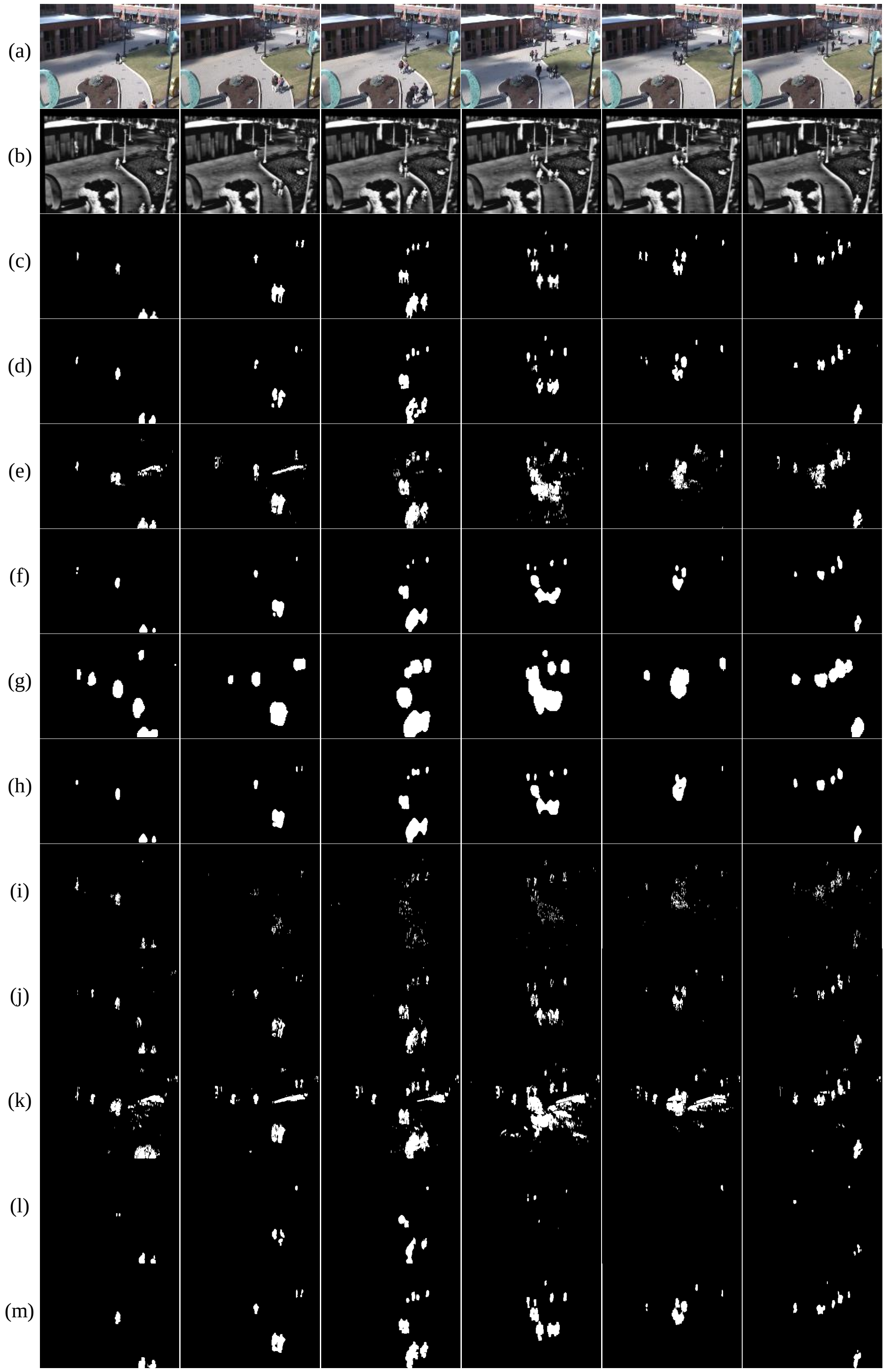


Fig. 5.4. Visual comparison, (a) RGB images[189], (b) thermal image[189], (c) ground truth[189], (d) GMG[2], (e) IMBS[5], (f) LOBSTER[8], (g) MultiCue[10], (h) SuBSENSE[12], (i) T2FMRF[16], (j) ViBe[19], (k) FA-SOM[21], (l)PBAS, (m) Proposed method.

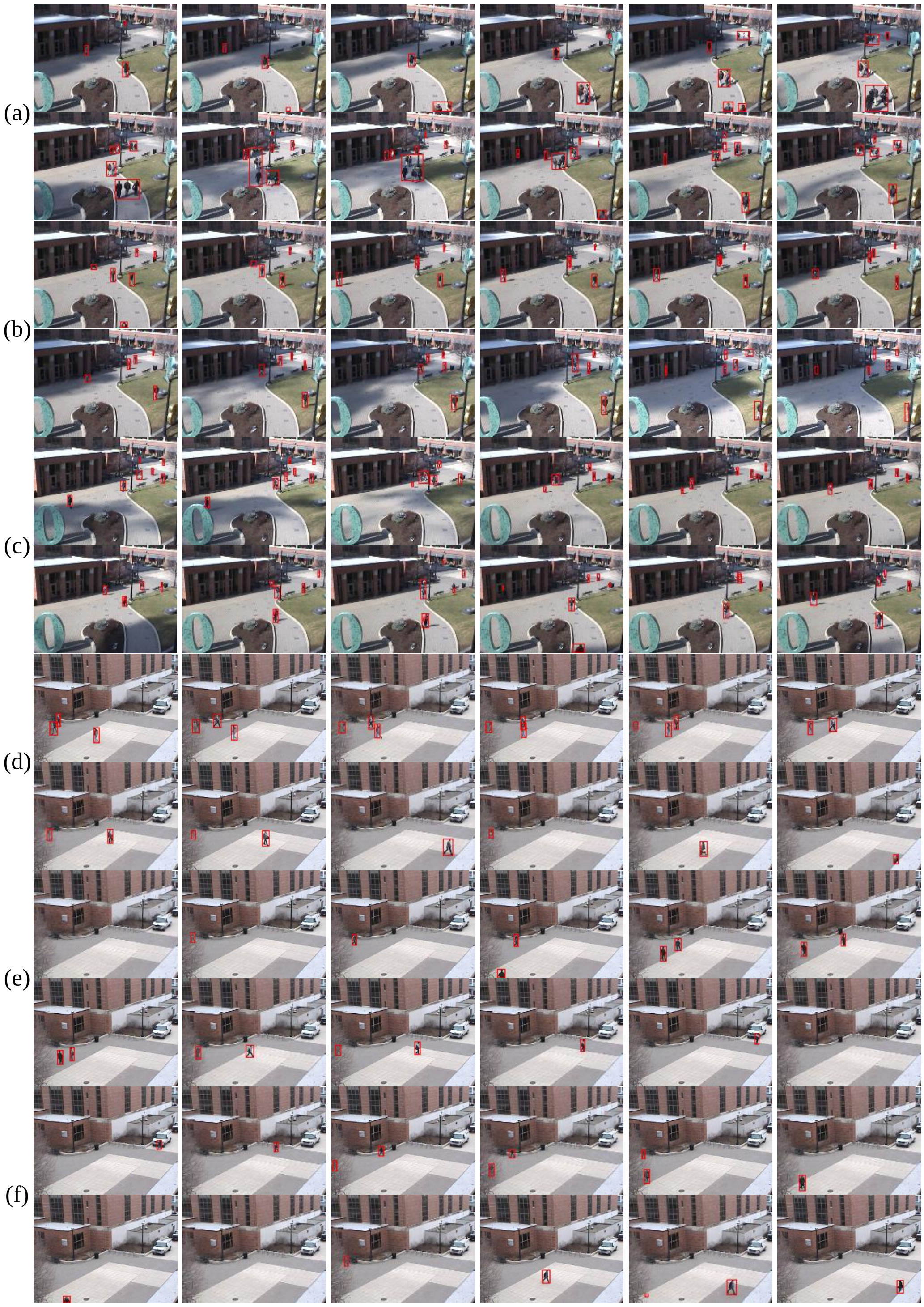


Fig. 5.5. Visual tracking results of the proposed approach across different images and scenarios. (a) -(f) Sequence1-6

CHAPTER 6

6 Conclusion and future work

6.1 Conclusion

The present thesis mainly focused on the methodologies for feature extraction, feature fusion, and statistical learning in different kinds of images. The main contributions cover three different area such as logo/ trademark image retrieval, saliency detection and the pedestrian detection in the videos.

In Chapter 3, a new logo/trademark retrieval system is proposed. For most logo images, colour and shape are their main characteristics. Therefore, by taking advantage of both colour and spatial layout descriptors, the feature of the logo images can be well extracted. Then an adaptive fusion strategy is used to balance the weight of both features based on the histogram distribution of the images and match the input query image with the target images in the database. All the data is acquired from Web and each image has three different noise types and totally 21 noise levels. The experimental results show that our method is robust to image noise and has good image matching performance.

In Chapter 4, a new unsupervised saliency detection model namely GLGOV is presented. According to neurology, our human visual attention system is made by top-down and bottom-up attention mechanism. The bottom-up mechanism makes our attention sensitive to the salient objects which have big contrast to their surrounding background in terms of colour and location. The top-down mechanism makes the visual cortex provide us prior knowledge and cognitive strategies to detect the salient objects. Inspired by these knowledge, the proposed method first detects the salient objects by a bottom-up model guided by Gestalt-law perceptions and then further refine the detection results by a top-down model guided by background connectivity principle. As such, these necrologies can be well introduced to support the saliency detection modeling. From the experimental results, it can be seen that our method can produce consistently good results on five datasets. However, as an unsupervised method, there is still some gap to those deep-learning based supervised approaches.

In Chapter 5, a cognitive model for pedestrian detection is proposed, where three key stages are included, i.e., foreground detection, fusion based object refinement, and object tracking. In foreground detection, a two-stage background subtraction on both visible and thermal images is introduced. Then the detected foreground objects are redefined by a fusion strategy guided by a cognitive knowledge. Finally, an improve mean-shift method, namely constrained mean-shift is proposed to separate overlapped or occluded objects. The proposed system is robust to the scale change and has better foreground detection performance than some state-of-the-art methods. However, in some cases, some pedestrian cannot be detected individually due to it is always a part of a group from the beginning to the end, and our system will just estimate the track of its group.

6.2 Future work

Although the contributions in the present thesis have achieved a certain level of success, there are still several challenges which can be translated to potential improvements and further investigation as summarized below:

1. For the logo/trademark retrieval, except the image with the different level of noise, some other challenging situations could be also considered, such as rotation of the logos, illumination changes, manually modification of the logos and occluded logos where some parts of the logos are hidden behind some other objects, etc. To achieve these, texture feature or deep feature descriptors can be used to better represent the image. In addition, the adaptive feature matching strategy can be also extended to fit multi-feature.
2. For saliency detection, it seems that the performance of unsupervised methods has already touched the up-limit and has a gap to those supervised methods, especially the deep learning based supervised methods. As a result, the deep learning model such as CNN will be the core of this work in the future and need to be further explored. Furthermore, more and more challenging

databases have been published in recent years, so that a more comprehensive experiment should be done to evaluate the effectiveness of the saliency model.

3. For pedestrian detection, the performance of background subtraction model is still under comprehensive investigation. Some other challenging conditions will be considered such as bad weather, low framerate, turbulence and night videos, etc. In addition, the combination of deep learning model and traditional methods in background subtraction progress is a clear line to follow in the future. Moreover, tracking progress will also be better optimized by other criteria, and benchmarked with some state-of-the-art methods on multiple databases.
4. Again, deep learning will be the focus of the future work, the improvement of some current deep learning models will be explored and applied in the different image processing related area.

REFERENCES

[1] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Transactions on Image Processing,* vol. 22, pp. 1689-1698, 2013.

[2] A. B. Godbehere, A. Matsukawa, and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *American Control Conference (ACC), 2012*, 2012, pp. 4305-4312.

[3] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 37, pp. 569-582, 2015.

[4] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1404-1412.

[5] D. Bloisi and L. Iocchi, "Independent multimodal background subtraction," in *CompIMAGE,* 2012, pp. 39-44.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis And Machine Intelligence,* vol. 20, pp. 1254-1259, Nov 1998.

[7] J. Kim, D. Han, Y. W. Tai, and J. Kim, "Salient Region Detection via High-Dimensional Color Transform," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[8] P.-L. St-Charles and G.-A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, 2014, pp. 509-515.

[9] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of the eleventh ACM international conference on Multimedia*, ed: ACM, 2003, pp. 374-381.

[10] S. Noh and M. Jeon, "A new framework for background subtraction using multiple cues," in *Asian Conference on Computer Vision*, 2012, pp. 493-506.

[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," presented at the Advances in neural information processing systems, 2006.

[12] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Flexible background subtraction with self-balanced local sensitivity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 408-413.

[13] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166-3173.

[14] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007.

[15] Z. Liu, L. Meur, and S. Luo, "Superpixel-based saliency detection," presented at the 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), 2013.

[16] Z. Zhao, T. Bouwmans, X. Zhang, and Y. Fang, "A fuzzy background modeling approach for motion detection in dynamic backgrounds," in *Multimedia and signal processing*, ed: Springer, 2012, pp. 177-185.

[17] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Computer Vision Systems*, ed: Springer, 2008, pp. 66-75.

[18] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2976-2983.

[19] O. Barnich and M. Van Droogenbroeck, "ViBe: a powerful random technique to estimate the background in video sequences," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 945-948.

[20] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," presented at the IEEE conference on Computer vision and pattern recognition (CVPR), 2009.

[21] L. Maddalena and A. Petrosino, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Computing and Applications*, vol. 19, pp. 179-186, 2010.

[22] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2814-2821.

[23] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," presented at the 17th IEEE International Conference on Image Processing (ICIP) 2010.

[24] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *European Conference on Computer Vision*, ed: Springer, 2010, pp. 366-379.

[25] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 2012, pp. 38-43.

[26] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Transactions on Image Processing,* vol. 25, pp. 5012-5024, 2016.

[27] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, "Global contrast based salient region detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[28] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478-487.

[29] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 25, pp. 1309-1321, 2015.

[30] M. S. Nixon and A. S. Aguado, *Feature extraction & image processing for computer vision*: Academic Press, 2012.

[31] L. Piras and G. Giacinto, "Information fusion in content based image retrieval: A comprehensive overview," *Information Fusion,* vol. 37, pp. 50-60, 2017.

[32] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 3457-3464.

[33] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing,* vol. 24, pp. 5706-5722, 2015.

[34] M. Li, C. Wei, Y. Yuan, and Z. Cai, "A Survey of Video Object Tracking," *International Journal of Control and Automation,* vol. 8, pp. 303-312, 2015.

[35] C. Haene, C. Zach, A. Cohen, and M. Pollefeys, "Dense semantic 3d reconstruction," *IEEE transactions on pattern analysis and machine intelligence,* vol. 39, pp. 1730-1743, 2017.

[36] P. Corke, *Robotics, Vision and Control: Fundamental Algorithms In MATLAB® Second, Completely Revised* vol. 118: Springer, 2017.

[37] X.-F. Zhang, A.-P. Zeng, S. Huang, M. Qing, and Y. Zhou, "An RGBD Tracker Based on KCF Adaptively Handling Long-Term Occlusion," in *International Conference on Intelligent Computing,* 2018, pp. 100-107.

[38] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE,* vol. 103, pp. 1449-1477, 2015.

[39] X. Jin, Y. Su, L. Zou, C. Zhang, P. Jing, and X. Song, "Video logo removal detection based on sparse representation," *Multimedia Tools and Applications,* pp. 1-20, 2018.

[40] Y. Gao, F. Wang, H. Luan, and T.-S. Chua, "Brand data gathering from live social media streams," in *Proceedings of International Conference on Multimedia Retrieval*, 2014, p. 169.

[41] A. P. Psyllos, C.-N. E. Anagnostopoulos, and E. Kayafas, "Vehicle logo recognition using a sift-based enhanced matching scheme," *IEEE transactions on intelligent transportation systems,* vol. 11, pp. 322-328, 2010.

[42] Y. Yan, J. Ren, Y. Li, J. F. Windmill, W. Ijomah, and K.-M. Chao, "Adaptive fusion of color and spatial features for noise-robust retrieval of colored logo and trademark images," *Multidimensional Systems and Signal Processing,* vol. 27, pp. 945-968, 2016.

[43] E. Kazakos, C. Nikou, and I. A. Kakadiaris, "On the Fusion of RGB and Depth Information for Hand Pose Estimation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 868-872.

[44] A. Leykin and R. Hammoud, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Machine Vision and Applications,* vol. 21, pp. 587-595, 2010.

[45] T. Alldieck, C. H. Bahnsen, and T. B. Moeslund, "Context-aware fusion of RGB and thermal imagery for traffic monitoring," *Sensors,* vol. 16, p. 1947, 2016.

[46] M. Tzelepi and A. Tefas, "Deep convolutional learning for content based image retrieval," *Neurocomputing,* vol. 275, pp. 2467-2478, 2018.

[47] D. C. G. Pedronette and R. d. S. Torres, "Unsupervised rank diffusion for content-based image retrieval," *Neurocomputing,* vol. 260, pp. 478-489, 2017.

[48] J. Yue, Z. Li, L. Liu, and Z. Fu, "Content-based image retrieval using color and texture fused features," *Mathematical and Computer Modelling,* vol. 54, pp. 1121-1127, 2011.

[49] S. Unar, X. Wang, and C. Zhang, "Visual and textual information fusion using Kernel method for content based image retrieval," *Information Fusion,* vol. 44, pp. 176-187, 2018.

[50] Y. Mistry, D. Ingole, and M. Ingole, "Content based image retrieval using hybrid features and various distance metric," *Journal of Electrical Systems and Information Technology,* 2017.

[51] X. Liu and B. Zhang, "Automatic Collecting Representative Logo Images from the Internet," *Tsinghua Science and Technology,* vol. 18, pp. 606-617, Dec 2013 2013.

[52] Y. S. Kim and W. Y. Kim, "Content-based trademark retrieval system using a visually salient feature," *Image and Vision Computing,* vol. 16, pp. 931-939, 1998.

[53] J. M. Patel and N. C. Gamit, "A review on feature extraction techniques in Content Based Image Retrieval," in *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*, 2016, pp. 2259-2263.

[54] T.-C. Lu and C.-C. Chang, "Color image retrieval technique based on color features and image bitmap," *Information Processing & Management,* vol. 43, pp. 461-472, 2007.

[55] C.-H. Lin, R.-T. Chen, and Y.-K. Chan, "A smart content-based image retrieval system based on color and texture feature," *Image and Vision Computing,* vol. 27, pp. 658-665, 2009.

[56] X.-Y. Wang, Y.-J. Yu, and H.-Y. Yang, "An effective image retrieval scheme using color, texture and shape features," *Computer Standards & Interfaces,* vol. 33, pp. 59-68, 2011.

[57] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognition,* vol. 46, pp. 188-198, 2013.

[58] S. Jeong, C. S. Won, and R. M. Gray, "Image retrieval using color histograms generated by Gauss mixture vector quantization," *Computer Vision and Image Understanding,* vol. 94, pp. 44-66, 2004.

[59] S. Murala, A. B. Gonde, and R. P. Maheshwari, "color and texture features for image indexing and retrieval," in *2009 IEEE International Advance Computing Conference, IACC 2009*, 2009, pp. 1411-1416.

[60] R. Brunelli and O. Mich, "Histograms analysis for image retrieval," *Pattern Recognition,* vol. 34, pp. 1625-1637, 2001.

[61] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom*, et al.*, "Query by Image and Video Content: The QBIC System," *Computer,* vol. 28, pp. 23-32, 1995.

[62] Z. Huang, "Contnt-based Image Retrieval Using Color Moment and Gabor Texture Feature," presented at the Machine Learning and Cybernetics, Qingdao, 2010.

[63] P. Maheshwary and N. Srivastav, "Retrieving Similar Image Using Color Moment Feature Detector and K-Means Clustering of Remote Sensing Images," 2008.

[64] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proceedings of the fourth ACM international conference on Multimedia*, 1997, pp. 65-73.

[65] J. Park, Y. An, I. Jeong, and J. Kang, "Image Indexing using Spatial Multi-Resolution Color Correlogram," presented at the IEEE International Workshop on Imaging Systems and Techniques, Krakow, 2007.

[66] L. Thurdsak, A. Kiyoaki, and K. Shozo, "A New Content-based Image Retrieval Using Color Correlogram and Inner Product Metric," presented at the Eighth International Workshop on Image Analysis for Multimedia Interactive Service, Santorini, 2007.

[67] H. Shao, Y. Wu, W. Cui, and J. Zhang, "Image Retrieval Based on MPEG-7 Dominant Color Descriptor," presented at the Young Computer Scientists, Hunan, 2008.

[68] R. Min and H. D. Cheng, "Effective image retrieval using dominant color descriptor and fuzzy support vector machine," *Pattern Recognition,* vol. 42, pp. 147-157, 2009.

[69] N.-C. Yang, W.-H. Chang, C.-M. Kuo, and T.-H. Li, "A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval," *Journal of Visual Communication and Image Representation,* vol. 19, pp. 92-105, 2008.

[70] A. Talib, M. Mahmuddin, H. Husni, and L. E. George, "A weighted dominant color descriptor for content-based image retrieval," *Journal of Visual Communication and Image Representation,* vol. 24, pp. 345-360, 2013.

[71] A. Mojsilovic, J. Hu, and E. Soljanin, "Extraction of perceptually important colors and similarity measurement for image matching retrieval and analysis," *IEEE TRANSACTIONS ON IMAGE PROCESSING,* vol. 11, November 2002.

[72] S. P.Lloyd, "Least Squares Quantization in PCM," *IEEE TRANSACTIONS ON IMAGE PROCESSING. Information Theory,* vol. 28, pp. 129-137, 1982.

[73] P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*: John Wiley & Sons, Inc., 2002.

[74] H. Ling, X. Yang, and L. J. Latecki, "Balancing deformability and discriminability for shape matching," presented at the Proceedings of the 11th European conference on computer vision conference on Computer vision: Part III, Heraklion, Crete, Greece, 2010.

[75] D. Zhang and G. Lu, "Shape-based image retrieval using generic Fourier descriptor," *Signal Processing: Image Communication,* vol. 17, pp. 825-848, 2002.

[76] D. Zhang and G. Lu, "A comparative study of curvature scale space and Fourier descriptors for shape-based image retrieval," *Journal of Visual Communication and Image Representation,* vol. 14, pp. 39-57, 2003.

[77] F. M. Anuar, R. Setchi, and Y.-k. Lai, "Trademark image retrieval using an integrated shape descriptor," *Expert Systems with Applications,* vol. 40, pp. 105-121, 2013.

[78] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition,* vol. 37, pp. 1-19, 2004.

[79] C.-H. Wei, Y. Li, W.-Y. Chau, and C.-T. Li, "Trademark image retrieval using synthetic features for describing global shape and interior structure," *Pattern Recognition,* vol. 42, pp. 386-394, 2009.

[80] D. Zhang and G. Lu, "A Comparative Study of Three Region Shape Descriptors," presented at the Digital Image Computing Techniques and Applications, Melbourne, Australia, 2002.

[81] M. E. ElAlami, "Unsupervised image retrieval framework based on rule base system," *Expert Systems with Applications,* vol. 38, pp. 3539-3549, 2011.

[82] D. Zhang and G. Lu, "Content-Based Shape Retrieval Using Different Shape Descriptors: A Comparative Study," presented at the Digital Image Computing Techniques and Applications, Melbourne, Australia, 2002.

[83] H. Qi, K. Li, Y. Shen, and W. Qu, "An effective solution for trademark image retrieval by combining shape description and feature matching," *Pattern Recognition,* vol. 43, pp. 2017-2027, 2010.

[84] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "An ontology approach to object-based image retrieval," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, 2003, pp. II-511.

[85] W.-Y. Ma and B. S. Manjunath, "Netra: A toolbox for navigating large image databases," *Multimedia systems,* vol. 7, pp. 184-198, 1999.

[86] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern recognition*, vol. 40, pp. 262-282, 2007.

[87] N. Suematsu, Y. Ishida, A. Hayashi, and T. Kanbara, "Region-based image retrieval using wavelet transform," in *Proc. 15th international conf. on vision interface*, 2002, pp. 9-16.

[88] A. K.Jain and A. Vailaya, "A Case Study with Trademark Image Database," *Pattern Recognition,* vol. 31, pp. 1369-1390, 1998.

[89] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280-287.

[90] A. Furnari, G. M. Farinella, and S. Battiato, "An Experimental Analysis of Saliency Detection with Respect to Three Saliency Levels," in *ECCV Workshops (3)*, 2014, pp. 806-821.

[91] H. Fu, Z. Chi, and D. Feng, "Attention-driven image interpretation with application to image retrieval," *Pattern Recognition,* vol. 39, pp. 1604-1621, 2006.

[92] S.-M. Hu, T. Chen, K. Xu, M.-M. Cheng, and R. R. Martin, "Internet visual media processing: a survey with graphics and vision applications," *The Visual Computer,* vol. 29, pp. 393-405, 2013.

[93] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Transactions on Image Processing,* vol. 22, pp. 363-376, 2013.

[94] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing,* vol. 19, pp. 185-198, 2010.

[95] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing,* vol. 29, pp. 1-14, 2011.

[96] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: an overview," *IEEE Transactions on Consumer Electronics,* vol. 46, pp. 1103-1127, 2000.

[97] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barbba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," presented at the International Conference on Image Processing (ICIP) 2007.

[98] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: based on eye-tracking data," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 21, pp. 971-982, 2011.

[99] Q. Ma and L. Zhang, "Saliency-based image quality assessment criterion," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, ed: Springer, 2008, pp. 1124-1133.

[100] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut," *IEEE Transactions on Multimedia,* vol. 14, pp. 1275-1289, 2012.

[101] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," presented at the IEEE 12th International Conference onComputer Vision, 2009.

[102] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 1915-1926, 2012.

[103] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE transactions on circuits and systems for video technology,* vol. 23, pp. 2009-2021, 2013.

[104] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 53, pp. 3325-3337, 2015.

[105] J. M. Wolfe, "Guided Search 2.0 A revised model of visual search," *Psychonomic Bulletin &amp; Review,* vol. 1, pp. 202-238, 1994.

[106] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*, ed: Springer, 1987, pp. 115-141.

[107] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience,* vol. 18, pp. 193-222, 1995.

[108] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: Different processes and overlapping neural systems," *The Neuroscientist,* vol. 20, pp. 509-521, 2014.

[109] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biology,* vol. 14, pp. R850-R852, 2004.

[110] K. R. Gegenfurtner, "Cortical mechanisms of colour vision," *Nat Rev Neurosci,* vol. 4, pp. 563-72, Jul 2003.

[111] E. Doi, J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner*, et al.*, "Efficient coding of spatial information in the primate retina," *The Journal of Neuroscience,* vol. 32, pp. 16256-16264, 2012.

[112] N. Al-Aidroos, C. P. Said, and N. B. Turk-Browne, "Top-down attention switches coupling between low-level and high-level areas of human visual cortex," *Proceedings of the National Academy of Sciences,* vol. 109, pp. 14675-14680, 2012.

[113] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology,* vol. 12, pp. 97-136, 1980.

[114] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *European conference on computer vision*, 2012, pp. 29-42.

[115] W. Zou and N. Komodakis, "HARF: Hierarchy-associated rich features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2016, pp. 406-414.

[116] Y. Tang and X. Wu, "Saliency detection via combining region-level and pixel-level predictions with CNNS," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 9912 LNCS, ed, 2016, pp. 809-825.

[117] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455-5463.

[118] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE transactions on neural networks and learning systems,* vol. 27, pp. 1135-1149, 2016.

[119] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron,* vol. 73, pp. 415-434, 2012.

[120] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic* vol. 4: Prentice hall New Jersey, 1995.

[121] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing,* vol. 27, pp. 38-49, 2018.

[122] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[123] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[124] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica,* vol. 11, pp. 23-27, 1975.

[125] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR),* vol. 38, p. 13, 2006.

[126] Y. Yan, J. Ren, H. Zhao, J. Zheng, E. M. Zaihidee, and J. Soraghan, "Fusion of Thermal and Visible Imagery for Effective Detection and Tracking of Salient Objects in Videos," in *Pacific Rim Conference on Multimedia*, 2016, pp. 697-704.

[127] O. Sidla, Y. Lypetskyy, N. Brandle, and S. Seer, "Pedestrian detection and tracking for counting applications in crowded situations," in *Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on*, 2006, pp. 70-70.

[128] J. Ge, Y. Luo, and G. Tei, "Real-time pedestrian detection and tracking at nighttime for driver-assistance systems," *Intelligent Transportation Systems, IEEE Transactions on,* vol. 10, pp. 283-298, 2009.

[129] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* pp. 1239-1258, 2009.

[130] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing,* vol. 28, pp. 976-990, 2010.

[131] R. Bodor, B. Jackson, and N. Papanikolopoulos, "Vision-based human tracking and activity recognition," in *Proc. of the 11th Mediterranean Conf. on Control and Automation*, 2003.

[132] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *null*, 2005, pp. 364-369.

[133] J. W. Davis and V. Sharma, "Robust background-subtraction for person detection in thermal imagery," *IEEE Int. Wkshp. on Object Tracking and Classification Beyond the Visible Spectrum,* 2004.

[134] J. W. Davis and V. Sharma, "Robust detection of people in thermal imagery," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 713-716.

[135] D.-E. Kim and D.-S. Kwon, "Pedestrian detection and tracking in thermal images using shape features," in *Ubiquitous Robots and Ambient Intelligence (URAI), 2015 12th International Conference on*, 2015, pp. 22-25.

[136] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review,* vol. 11, pp. 31-66, 2014.

[137] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, pp. 780-785, 1997.

[138] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, 1999, pp. 246-252.

[139] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological),* pp. 1-38, 1977.

[140] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 28-31.

[141] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE transactions on pattern analysis and machine intelligence,* vol. 27, pp. 827-832, 2005.

[142] A. Shimada, D. Arita, and R.-i. Taniguchi, "Dynamic control of adaptive mixture-of-Gaussians background model," in *Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on*, 2006, pp. 5-5.

[143] J. Zeng, L. Xie, and Z.-Q. Liu, "Type-2 fuzzy Gaussian mixture models," *Pattern Recognition,* vol. 41, pp. 3636-3643, 2008.

[144] S. Li, "Markov random field models in computer vision," *Computer Vision—ECCV'94,* pp. 361-370, 1994.

[145] Z. Tu, A. Zheng, E. Yang, B. Luo, and A. Hussain, "A Biologically Inspired Vision-Based Approach for Detecting Multiple Moving Objects in Complex Outdoor Scenes," *Cognitive Computation,* vol. 7, pp. 539-551, 2015.

[146] A. Zheng, M. Xu, B. Luo, Z. Zhou, and C. Li, "CLASS: Collaborative Low-Rank and Sparse Separation for Moving Object Detection," *Cognitive Computation,* vol. 9, pp. 180-193, 2017.

[147] Y. Wang, Q. Zhao, B. Wang, S. Wang, Y. Zhang, W. Guo*, et al.*, "A Real-Time Active Pedestrian Tracking System Inspired by the Human Visual System," *Cognitive Computation,* vol. 8, pp. 39-51, 2016.

[148] S. E. Ebadi, V. G. Ones, and E. Izquierdo, "Approximated robust principal component analysis for improved general scene background subtraction," *arXiv preprint arXiv:1603.05875,* 2016.

[149] X. Liu, G. Zhao, J. Yao, and C. Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Transactions on Image Processing,* vol. 24, pp. 2502-2514, 2015.

[150] Y. Xu, J. Dong, B. Zhang, and D. Xu, "Background modeling methods in video analysis: A review and comparative evaluation," *CAAI Transactions on Intelligence Technology,* vol. 1, pp. 43-60, 2016.

[151] S. Jeeva and M. Sivabalakrishnan, "Survey on background modeling and foreground detection for real time video surveillance," *Procedia Computer Science,* vol. 50, pp. 566-571, 2015.

[152] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding,* vol. 122, pp. 4-21, 2014.

[153] A. Sobral, "BGSLibrary: An opencv c++ background subtraction library," in *IX Workshop de Visao Computacional*, 2013, p. 7.

[154] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection. net: A new change detection benchmark dataset," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 2012, pp. 1-8.

[155] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequievre, "A benchmark dataset for foreground/background extraction," in *ACCV 2012, Workshop: Background Models Challenge*, 2012.

[156] Y. Dhome, N. Tronson, A. Vacavant, T. Chateau, C. Gabard, Y. Goyat*, et al.*, "A benchmark for background subtraction algorithms in monocular vision: a comparative study," in *Image Processing Theory Tools and Applications (IPTA), 2010 2nd International Conference on*, 2010, pp. 66-71.

[157] M. F. Calitz; and H. Rüther, "Least absolute deviation (LAD) image matching," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 51, pp. 223-229, 6 May 1996.

[158] M. Thomas and N. Kenneth, "Least squares and least absolute deviation procedures in approximately linear models," *Statistics & Probability Letters,* vol. 16, pp. 153-158, 1993.

[159] Y. Yan, J. Ren, Y. Li, J. Windmill, and W. Ijomah, "Fusion of Dominant Colour and Spatial Layout Features for Effective Image Retrieval of Coloured Logos and Trademarks," presented at the 1st IEEE International Conference on Multimedia Big Data, Beijing, 2015.

[160] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters,* vol. 31, pp. 651-666, 2010.

[161] Mohan S. Kankanhalli, Babu M. Mehtre, and H. Y. Huang, "Color and Spatial Feature for Content-based Image Retrieval," *Pattern Recognition Letters,* vol. 20, pp. 109-118, 1999.

[162] Y. Feng, J. Ren, and J. Jiang, "Object-based 2D-to-3D Video Conversion for Effective Stereoscopic Content Generation in 3D-TV Applications," *Broadcasting, IEEE Transactions on,* vol. 57, p. 500, June 2011.

[163] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An Object_Oriented Visual Saliency Detection Framework Based on Sparse Cosing Representations," *IEEE Trans. on Circuits and Systems on Video Technology,* vol. 23, pp. 2009-2021, December 2013.

[164] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Obvject Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning," *IEEE Trans. Geoscience and Remote Sensing,* vol. PP, 2014.

[165] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised Extraction of Visual Attention Objects in Color Images," *IEEE Trans. on Circuits and Systems on Video Technology,* vol. 16, pp. 141-145, January 2006.

[166] N. R. Carlson, H. Miller, J. W. Donahoe, and G. N. Martin, *Psychology: The Science of Behavior*. Ontario, CA: Pearson Education Canada, 2010.

[167] M. Wertheimer, "Untersuchungen zur Lehre von der Gestalt. II," *Psychological Research,* vol. 4, pp. 301-350, 1923.

[168] M. Wertheimer, "Laws of organization in perceptual forms," 1938.

[169] A. Desolneux, L. Moisan, and J.-M. Morel, "Computational gestalts and perception thresholds," *Journal of Physiology-Paris,* vol. 97, pp. 311-324, 2003.

[170] X. Li and G. D. Logan, "Object-based attention in Chinese readers of Chinese words: Beyond Gestalt principles," *Psychonomic bulletin & review,* vol. 15, pp. 945-949, 2008.

[171] G. Kootstra and D. Kragic, "Fast and bottom-up object detection, segmentation, and evaluation using Gestalt principles," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 3423-3428.

[172] Z. Wang and B. Li, "A two-stage approach to saliency detection in images," presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008.

[173] G. Kootstra, N. Bergström, and D. Kragic, "Gestalt principles for attention and segmentation in natural and artificial vision systems," presented at the ICRA 2011 Workshop on Semantic Perception, Mapping and Exploration (SPME), Shanghai, China, 2011.

[174] J. Wu and L. Zhang, "Gestalt saliency: Salient region detection based on gestalt principles," presented at the 20th IEEE International Conference on Image Processing (ICIP), 2013.

[175] Z. Ren, Y. Hu, L.-T. Chia, and D. Rajan, "Improved saliency detection based on superpixel clustering and saliency propagation," presented at the Proceedings of the 18th ACM international conference on Multimedia, 2010.

[176] S. E. Palmer, *Vision science: Photons to phenomenology*: MIT press, 1999.

[177] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 2274-2282, 2012.

[178] J. Van De Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 150-156, 2006.

[179] N. I. Córdova, A. Tompary, and N. B. Turk-Browne, "Attentional modulation of background connectivity between ventral visual cortex and the medial temporal lobe," *Neurobiology of learning and memory,* vol. 134, pp. 115-122, 2016.

[180] S. Jeannin, L. Cieplinski, J. R. Ohm, and M. Kim, "Mpeg-7 visual part of experimentation model version 9.0," *ISO/IEC JTC1/SC29/WG11 N,* vol. 3914, 2001.

[181] T. L. J. Sun, N.-N. Zheng, X. Tang, H.-Y. Shum, and P. Xi'an, "Learning to Detect A Salient Object," presented at the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[182] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "Salientshape: Group saliency in image collections," *The Visual Computer,* vol. 30, pp. 443-453, 2014.

[183] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155-1162.

[184] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280-287.

[185] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265-1274.

[186] J. Han, E. J. Pauwels, and P. De Zeeuw, "Visible and infrared image registration in man-made environments employing hybrid visual features," *Pattern Recognition Letters,* vol. 34, pp. 42-51, 2013.

[187] P. M. De Zeeuw, E. J. E. M. Pauwels, and J. Han, "Multimodality and multiresolution image fusion," in *VISAPP 2012 - Proceedings of the International Conference on Computer Vision Theory and Applications*, 2012, pp. 151-157.

[188] J. Han, E. J. Pauwels, and P. de Zeeuw, "Fast saliency-aware multi-modality image fusion," *Neurocomputing,* vol. 111, pp. 70-80, 2013.

[189] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer vision and image understanding,* vol. 106, pp. 162-182, 2007.

[190] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2000, pp. 142-149.

[191] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-based surveillance systems*, ed: Springer, 2002, pp. 135-144.

Appendix A: Publications by the Author

Journal Publications

1. Y. Yan, J. Ren, et al., ‘Adaptive Fusion of Colour and Spatial Features for Noise-Robust Retrieval of Coloured Logo and Trademark Images’, Multidimensional Systems and Signal Processing, vol. 27, no. 4, pp. 945-968, 2016.
2. Y. Yan, J. Ren et al., ‘Cognitive Fusion of Thermal and Visible Imagery for Effective Detection and Tracking of Pedestrians in Videos’, Cognitive Computation, pp. 1-11, 2017.
3. Y. Yan, J. Ren et al., ‘Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement’, Pattern Recognition 79, 65-78, 2018
4. J. Zheng, Y. Liu, J. Ren, T. Zhu, Y. Yan, et al., ‘Fusion of block and keypoints based approaches for effective copy-move image forgery detection’, Multidimensional Systems and Signal Processing, vol. 27, no. 4, pp. 989-1005, 2016.
5. Z. Yang, W. Chen, F. Cao, Y. Yan, et al., "Dimensionality Reduction Based on Determinantal Point Process and Singular Spectrum Analysis for Hyperspectral Images," *IET Image Processing,* 2018.

Conference Publications

1. Y. Yan, J. Ren, et al., ‘Fusion of Dominant Colour and Spatial Layout Features for Effective Image Retrieval of Coloured Logos and Trademarks’, in Proc. 1st IEEE Int. Conf. on Multimedia Big Data, Beijing, 2015.
2. Y. Yan, J. Ren, et al., ‘Fusion of Thermal and Visible Imagery for Effective Detection and Tracking of Salient Objects in Videos’, Advances in Multimedia Information Processing – PCM 2016, Part II, LNCS vol. 9917, pp. 697-704, Springer, 2016.
3. Y. Yan, J. Ren, et al., ‘Deep Background Subtraction of Thermal and Visible Imagery for Pedestrian Detection in Videos’, 9th International Conference on Brain Inspired Cognitive Systems (BICS 2018).
4. X. li, J. Ren, Y. Yan and J. Soraghan, ‘Knowledge based Fundamental and Harmonic Frequency Detection in Polyphonic Music Analysis’, Int. Conf. on Communications, Signal Processing and Systems, Harbin, 2017.
5. Z, Yang, W. Chen, Y. Yan, et al., ‘Unsupervised Hyperspectral Band Selection Based on maximum Information Entropy and Determinantal Point Process’, 9th International Conference on Brain Inspired Cognitive Systems (BICS 2018).
6. J. Zabalza, Z. Fei, C. Wong, Y. Yan, et al., ‘Making Industrial Robots Smarter with Adaptive Reasoning and Autonomous Thinking for Real-Time Tasks in Dynamic Environments: A Case Study’, 9th International Conference on Brain Inspired Cognitive Systems (BICS 2018).