Faculty of Engineering and Information Technology
University of Technology, Sydney

Detecting Text in Clutter Scene

A thesis submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy

by

Xia Cui

November 2014

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student:

Date:

Abstract

We often encounter cluttered visual scenes and need to identify objects correctly to navigate and interact with the world. As text takes the typical form of a human-designed informative visual object, retrieving texts in both indoor and outdoor environments is an important step towards providing contextual clues for a wide variety of vision tasks. Furthermore, it plays an invaluable role for multimedia retrieval and location based services.

Text detection from clutter background is nevertheless a challenging task because the text, being figures in image, can be presented in various ways with lots of room for uncertainty such as size, scale, font type, font texture and colour, unpredicted decorative elements put on the text, etc. The situation will be even more complicated if the text is presented in a clutter background where non-text objects possess similar low-level features to text. Further, all these objects are composed of distinct geometric shapes and they are similar with the essential composition elements of text objects. Pursuing a robust text feature descriptor is therefore always difficult because special feature descriptor is only a fragment of text existence. It needs the completely understanding of text.

Regarding the design, understanding, representation and calculating of text as one unitary process of text perceiving, we deal with the completely understanding and representation of text in image with many kinds of aspects in different levels. Without following the legend feature based solution, this research is motivated by perceptual image processing and the observation of painting masters. It will explore a brand new solution by investigating the spatial structure of text and the compositional complexity of the visual object (i.e. text) in image. The research will present the composition granularity indicator and expose novel discriminable attributes embedded inside text objects, which can successfully differentiate text regions and non-text regions on clutter backgrounds.

As figures in image with the clutter scene, it is merely the physical appearance of text which provides the perceptual content and plays a central role for text detection, i.e. location and coarse identification. During the view-construction of text, properties of individual character and textual organization of characters build up the physical appear-

ance. When observers see text appearance in clutter scene, they describe their feelings in terms of crowding effect and clutter. However, the appearance of text still has enough saliency to reveal an informative message. Accordingly, text not only has the characteristics of crowding effect and clutter but also follows the principles of saliency.

Significantly, the crowding effect of text is derived from the space regularity of inbuilt neighbouring letters which have commonalities beside their distinctiveness. In addition, low-level features of individual letters contribute to the commonalities and distinctiveness from the moment that the font is designed.

Therefore, the computational model of text appearance is built up to integrate the three-level properties, including features of individual characters (low-level features), properties for spatial regularity (i.e. neighbourhood, appearance similarity), and the crowding statistics property of space averaged over pooling regions.

In terms of image processing, if we consider the view construction of text, the features of individual characters in image processing are obtained on the basis of the properties of construction, including mean intensity, local RMS contrast, shape, pixel density, edge density, stroke width, straight line ratio, height to width ratio, stroke width to height ratio, etc.

For the purpose of calculating the properties of space regularity and the crowding space averaging property, the spatial elements and relations are quantified and these involve space granularity and composition rules.

If we examine the works of painters, especially impressionists, they use directional brushstroke or colour patches as space granularity to represent "formless" visual objects in space regularity instead of clear contour shape sketches. The space regularity of patches, i.e. repetitive patterns, can offer a compositional format to express an artist's feelings about an object rather than to simply describe it. Secondly, it is the harmonious proportions among component parts that bridle component space patches into objects. If we consider the painter's harmonious proportions, the component parts of an object can be said to react simultaneously so that they can be seen at one and the same time both together and separately.

Similarly, image is described by a set of grey space patches in multi-grey levels. In addition, each space patch groups pixels in position proximity and similarity, in just the same way as the colour patch is used by impressionists. The space organisation of them is also quantified as the measurement of space relations, especially in terms of the

neighbourhood and proportions among component parts. Moreover, the harmonious proportions among space patches are captured by the mathematical tool of geometric mean. Geometric mean (i.e., GM) is calculated over those space patches which possess the same grey level, and considered as the space granularity to form objects. Grey patches with the same GM are composed of GM regions, which are enlarged, extended kinds of pooling regions. Regions given by clusters which have resulted from similarity and neighbourhood are direct, compact pooling regions. Therefore, the statistical properties of space averaging are calculated over GM regions and image is represented as a set of GM regions over which text and other visual objects are analysed by GM indication.

Finally, the representation of an image and the three-level computational text model are put into practice to develop a new-brand algorithm on the public benchmark dataset and to design and implement an automatic processing system on the real big data of the bank cheque. The resulting performance of these tools/processes shows that they are highly competitive and effective.

Acknowledgement

First and foremost I want to thank my supervisor Professor Longbing Cao. It has been an honour to be his Ph.D. student. He has taught me, both consciously and unconsciously, how good big data is done. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example he has provided as a successful leader scientist and professor.

And I also want to thank my supervisor Dr. Qiang Wu. It has been a great gratitude to be his Ph.D student. He has taught me how good and joyful the thinking ability is done for research as a vigorous and rigorous scientist. I appreciate all his contributions of time and ideas to make my Ph.D full of creative thinking and enjoyment. The endless creative thinking and enthusiasm he has for his research encouraged me to make greater efforts. I am also thankful for the excellent example he has provided as a time manager.

The members of the AAI group have contributed immensely to my personal and professional time at UTS. The group has been a source of friendships as well as good advice and collaboration. I am especially grateful for the fun group of original AAI group members who stuck it out in grad school with me: Ziye Zuo, Wei Wei, Jiahang Chen, Yuming Ou, Zhigang Zheng, XinHua Zhu, Chao Luo, Junfu Yin, Xuhui Fan, and SongYin. I would like to acknowledge honorary group member QiaoYu Sun who was here a couple of years ago. We worked together along with Qiang Wu on the i-Cheque system development, and I am very much appreciated her enthusiasm, intensity, and amazing ability to manipulate text images. And I am also grateful for these honorary group members who have come through the lab: Hongxiu Zhu, Wei Li, Xiaodong Yue. Other past and present group members that I have had the pleasure to work with or alongside of are grad students Jingjiu Li, Mu Li, Can Wang, She Zhong, Chunming Liu, DongYu, JinSong, Pochung Zhang, RenJing, Renhua Song and the numerous visitors and rotation students who have come through the lab.

When I started my PH.D from the cheque document image processing, I have appreciated the camaraderie and local expertise of bank as well as the AAI group early on. In regards to the system design and development, I have appreciated David for his collaboration and the impressive skills.

In my attempted measurements of the crowding structure in an image with clutter scene, I thank the following people for helpful discussions with us: JinSong Xu, Song Yin, DongYu. I thank for their inspirational discussions with us. And I would also like to acknowledge the AAI Facility.

For this dissertation I would like to thank my reading committee members for their time, interest, and helpful comments. I would also like to thank members of my DA defense committee, Jian Zhang, Jinyan Li, Weichang Ye, for their time and insightful questions.

I am grateful to our group's past and present administrative assistant: Lei Zheng, YangBo, Collin Wise and Rayee who kept us organized and was always ready to help. And I am especially great grateful for and touched by the kindness and very generous support from Dr. Dan Luo during tough times.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. I was funded by the Australian IPRS Scholarship and APA for my first 3 years.

My time at UTS was made enjoyable in large part due to the many friends and groups that became a part of my life. I am grateful for time spent with roommates and friends, especially for elder friends as I made big decision for my life, and for many other people and memories. My time at UTS was also enriched by the AAI sport group, Omma’s Church and Campsie Congregation.

Lastly, I would like to thank my family for all their love and encouragement. I am very grateful for my parents who raised me with a love of science and supported me in all my pursuits, especially for the presence of my parents here for three of my years here to take care of my kids. For my kids who safeguard my heart and restore my courage. And most of all for my loving, supportive, encouraging, and patient husband whose faithful support during the final stages of this Ph.D. is so appreciated. Thank you.

Xia Cui

The University of Technology, Sydney

June 2014

Table of Contents

Chapter 1

Introduction

We often encounter cluttered visual scenes and need to identify objects correctly to navigate and interact with the world. As text is one typical human-designed informative visual object, retrieving texts in both indoor and outdoor environments provides contextual clues for a wide variety of vision tasks [14], and plays an invaluable role for multimedia retrieval and location based services, such as scene understanding, content-based image retrieval, object tracking [15], mobile robot navigation, assisting in navigation for visually impaired persons [16], and automatic geocoding. Thus, pursuing a robust and effective text detection method becomes one important visual search task for the purpose of applications.

1.1 Previous work

In the last few decades, there are many kinds of methods to be proposed to extract text. And the Optical Character Recognition (OCR) system is closely involved. Accordingly, there are OCR-based methods and non-OCR methods. However, most OCR systems are restricted to binary images of text or very simple background images, and non-OCR methods struggle to pursue reasonable flexibility and performance through one stable feature in wide variety of clutter scenes because most solutions are based on fragment conception of text properties, whereas they have to deal with the totality of text.

1.1.1 OCR-based method for text detection

When locations of text characters are approximately predictable and background interference does not resemble text characters (i.e. with simple background), many OCR-based algorithms have been developed. Most traditional OCR techniques are restricted to grey-level or binary images of text and many new commercial OCR systems [17] and algorithms based on them have been developed.

Liang *et al.*[18] have used texture flow analysis to perform geometric rectification of the planar and curved documents. Burns *et al.* [3] have performed a topic-based partition of the document image to distinguish text, white spaces and figures. Banerjee *et.al.*[19] have exploited the consistency of text characters in different sections to restore document images from severe degradation. This has been based on the model of Markov Random Field. Lu *et al.*[20] have proposed a word shape coding scheme through three topological features of characters for text recognition in the document image. Chen et al.[21] have used the AdaBoost method to learn image features which should be reliable indicators of text and have low entropy. Further, a commercial OCR system has been used to read the text or reject it as a non-text region. In detecting the sign, Chen et al. [22] have adapted a hierarchical framework embedding multi-resolution and multi-scale edge detection, adaptive searching, colour analysis, and affine rectification to normalize the intensity features for the OCR sign reading system. All the above algorithms share the same assumption that locations of text characters have a clean background. However, clutter scene usually has complex background, which arise difficulties for OCR reading.

1.1.2 Feature-based method

When there is a background interferences complex, off-the-shelf OCR software cannot handle these complicated interferences, since both the colour and edge of text are corrupted by strong spotlight, shadow and reflection meanwhile letter-like visual objects co-occurrence with text.

There are many different algorithms for text detection which can be roughly classified into two categories. The first category focuses on text region initialization and extension by directly using the distinct features of text, such as local extreme points (edge points), edge segments, stroke width, text-line, and the boundary box. And the other category, starting from exploiting the whole image compositional structure, partitions the whole image compositional structure to obtain the compositional elements of text, then based on these partitioned compositional structures the text region is initialized and extended.

In the first category, the distinct features of text characters consist of edge segments, stroke width, inner holes and a boundary box, text-line, and a wavelet coefficient representing text textural features.

Considering the text regions usually contain denser edges in the image, Shivakumara et al. [23] have applied different edge detectors to extract blocks full of the most apparent edges of text characters. However these have failed to remove the background noise resulted from pane, building and other objects that also possess high density of edge.

Phan et al. [24] performed a line-by-line scan in edge images to combine rows and columns with a high density of edge pixels into text regions. However, it divides the image spatially into blocks of equal size before grouping, and is very likely to break up text characters or text strings into fragments which fail to satisfy the texture constraints.

Pan et. al.[25] learned an over-complete dictionary from the edge segments of isolated character images by K-SVD and then used it to label the sparse from the given edge map of the image to get the text candidate. But the edge information is often corrupted by strong light, reflection and perspective distances. And as the K-SVD dictionary is designed for coding and de-noising, it can be confused by complex backgrounds with text-like areas.

Notably, recently, by all possible thresholds, local extremes (edge points in extended meaning) can be obtained in multi-threshold, and then Maximally Stable Extremal Regions (MSERs) [26] of data-dependent shape can be built by connecting the extremal points in a neighbourhood. Because of the excellent characteristics of MSERs，MSERs are detected and taken as candidate text regions in many recent studies. Based on those MSERs, many efficient pruning algorithms can be applied to locate the text. Neumann et al. [27-29] modified the original MSER algorithm to take region topology into consideration, leading to superior detection performance. Chen et al. [30] also proposed an extension to MSER, in which the boundaries of MSERs are enhanced via canny edge detection, to cope with image blur. Yin et al.[31, 32] , and Shi et al.[33] used a pruning algorithm to MSERs to get the text candidate to get high performance. But the MSERs measurements still need to explore.

To capture one important discriminative feature of text-constant stroke width，Liu et al. [34] designed a stroke filter to extract the stroke-like structures; B Epshtein et al [12] proposed a stroke width transform to find the value of stroke width for each image pixel, and then generated the text candidate by connected component analysis on the stroke map of an image. But the stroke width tends to slightly vary even within the same character. Parts of characters are often confused with similar looking background parts, such as the pane, bar and foliage.

To extract candidates of text regions, Kasar et al. [35] first assigned a bounding box to the boundary of each candidate character in the edge image and then detected text characters based on the boundary model (i.e. no more than 2 inner holes in each bounding box of alphabets and letters in English); Hasan et al.[36], Park et al.[37] and Uddin et.al[38] have designed robust morphological processing based on edge gratitude and edge directions respectively. But there are contrast, colour and size restrictions.

Kim [39] used SVM to analyse the raw pixels to find activated pixels which are highly related to text, and then a continuously adaptive mean shift algorithm (CAMSHIFT) is applied to these pixels with a high score about the text texture to obtain text regions. Tran et al.[40] modelled text string as multi-scale ridges representing its centre line and the skeletons of characters. By traversing the multi-oriented scene text lines, Shivakumara.P et al [41] proposed the boundary growing method works based on the concept of nearest neighbours to extract multi-oriented text. Kumar et al [42] used the wavelet coefficient to model text textual features and obtained characteristic wavelets of pure image or text classes, then located the text region. But this method is limited to the classification number of non-text background and that of text.

Instead of directly exploiting the distinct features of text, starting from exploiting the whole image compositional structure, the second category partitions these structures to get the structural elements of letter corps by which the text string is formed and probes where the structure form. It also serves to get the location of the letter corps as the text region.

Using colour similarity and colour variations analysis, Socotra et al.[43] combined them to generate the text region. However the unexpected background noises might share the same colours with text characters. Gao et al.[44] and Suen et al.[45] performed heuristic grouping and layout analysis to cluster edges of objects based on similar colour, position and size into text regions. Besides, in an image, many different kinds of colours are in tune with each other to form an intact picture. Yi et al [13] used the colour histogram and weighted K-means clustering to partition the original image into several colour layers. And on the main colour layers, combined with gradient-based partitions, text candidates could be extracted by a set of heuristic rules of properties of text such as text-line, stroke width, aspect ratio etc. However, since at least three letters can determine a certain text line, this method cannot handle cases where the text includes less than three letters.

In a word, these feature-based methods emphasize one stable specific feature to deal with the totality of text in wide variety of background based on intellect conception of text. Actually, it is in fragment, and a fragment, however cleverly put together, is still a small part of text existence whereas they have to face all kinds of challenges. And when we look at what is taking place in text in clutter scene, we begin to understand that it is one unitary process, we need the completely understanding and representation of text in image from the standpoint of calculating. If we get the completely understanding of text, especially, the physical appearance of text, we have a chance to solve the wide variety of problems in text detection, which arise from the interactions between text and its various complex backgrounds.

1.2 Our motivation and aim

For the purpose of pursuing the reasonably flexible algorithm of text detection, we have to deal with the completely understanding and representation of text in image through regarding the design, understanding, representation and calculating of text as one unitary process of text perceiving. It is a whole, and involves view construction, description and calculation of text with many kinds of aspects in different levels.

Thus, without following the legend feature based solution, our research is motivated by perceptual image processing and the observation of painting masters since text is one typical human-designed informative high perceptual visual object. It will explore a brand new solution by investigating the spatial regularity of text and the composition complexity of the visual object (i.e. text) in image. The research will present the composition granularity indicator of image, and expose novel discriminable attributes embedded inside text object which can successfully differentiate the text region and non-text region on a clutter background.

In an image, text is viewed as figures. Apparently, the physical appearance of text plays a central role in its detection, i.e., location and coarse identification. From the viewpoint of observation and vision perception, there are three-level constructions for text appearance in image, including features of individual character, letter-centred spatial organization, and object-centred display, shown in Figure 1.1, which work in harmony to make text congruent in terms of the ergonomic criteria of legibility, readability and conspicuity.

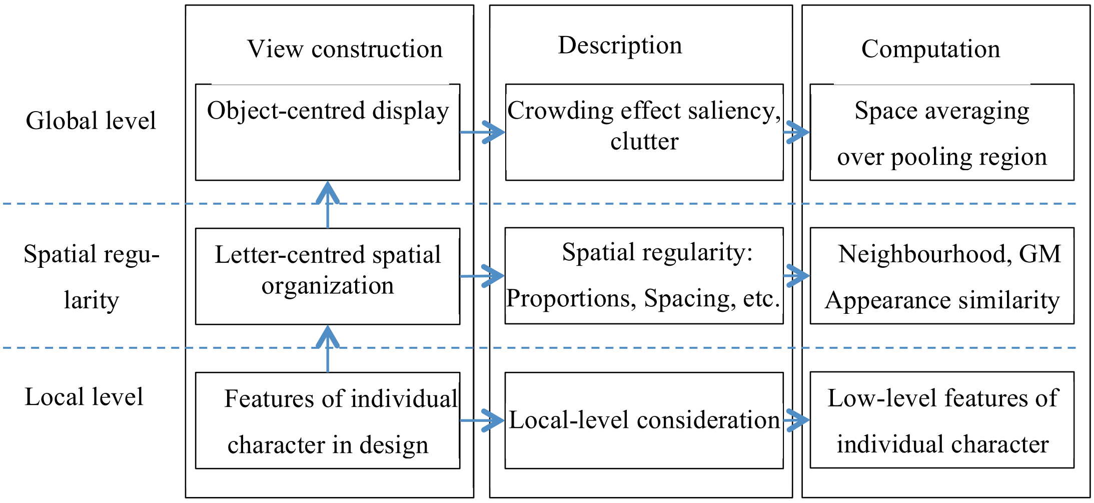

View construction
Description
Computation
Global level
Object-centred display
Crowding effect saliency, clutter
Space averaging over pooling region
Spatial regularity
Letter-centred spatial organization
Spatial regularity: Proportions, Spacing, etc.
Neighbourhood, GM Appearance similarity
Local level
Features of individual character in design
Local-level consideration
Low-level features of individual character

Figure 1.1 Three-level construction of text appearance and three-level computational model of text

In an object-centred display, there is the vision feeling of clutter and crowding in a clutter scene. That feeling is associated with the spatial organisation of the image and the text itself. For the image, there are too many items in limited space. For text, “it is as if there is a pressure on both sides of the word that tends to compress it. Then the stronger, i.e. the more salient or dominant letters, are preserved and they ‘squash’ the weaker, i.e. the less salient letters, between them.” This is the intrinsic crowding effect of text [46]. Meanwhile, text still has enough saliency to show an informative message in image with the clutter scene. Thus, text not only has the characteristic of a crowding effect but also follows the principles of saliency.

Significantly, crowding effect is derived from the space regularity of text, i.e., letter-centred spatial organization. The space regularity refers to the textural-like phenomenon showed by the in-built neighbouring letters which have commonalities beside their distinctiveness. In order to quantify and measure the space regularity, neighbourhood and appearance similarity among component parts are involved. The former is related to several kinds of spacing in typography design, including letter spacing, word spacing and line spacing. And the latter is highly related to the low-level features of individual letters.

Through the three-level construction, text is displayed and viewed in an image with a complex background. Correspondingly, the computational model of text is established to integrate the three-level properties or characteristics, including features of individual character (low-level features), properties for spatial regularity (i.e. neighbourhood, appearance similarity), and crowding statistics property of space averaged over pooling regions.

From the standpoint of calculating in image processing, we get the insight into these features from other fields, for instance, type design, crowding effect, clutter, saliency, and the way painters represent objects. From the type design, the functions of character features are understood completely, and these change the features for construction reasonably to be attributes in image processing. From the coexistence of crowding effect concurrences of clutter and saliency with text, the correlates among them inspire us to capture and represent text and image based on them. For the purpose of capturing the correlates and quantifying them in image processing, the ways in which painters understand and represent visual objects are mathematically applied, and tools are developed for the analysis of the composition of visual objects.

Gaining clues and inspiration from these fields, the three-level features or properties are calculated by an image-based method and integrated to discriminate the text region in non-text regions. Using this computation model of text, a brand-new solution is explored to break down the clutter background so as to detect text. Furthermore, this model is applied to process the big data of real bank cheques.

1.3 Methodology

As text represents the figures in an image, we mathematically dig the properties of text appearance in the view construction and description of text as illustrated in Figure 1.1. This enables us to reproduce the techniques that painters use to represent the spatial elements and proportions among component parts, and then we calculate the three-level features of text appearance in image processing.

In type design, the font stylish attributes and textual organisation contribute to the essential functions of text during the construction of appearance. They provide us with a comprehensive understanding of the roles of the features of individual character as many of them cannot be adequately caught through image processing alone. After investigating, we transfer them into image-based reasonable attributes of individual character and the measure of neighbourhood and appearance similarity of space regularity.

The coexistence of the crowding effect, clutter and saliency in text leads us to understand the correlates of them by investigating and summarising the theory, models, properties, and characteristics of them. It is the correlates that make us focus on the pooling region and region-level attributes or salient structure. Indeed, pooling region is related to its spatial elements and the relationships among them. Those elements across neighbour filters are tuned to similar orientations and integrated into an association receptive field which will become a pooling region or part of a pooling region.

For the purpose of capturing spatial elements and their relationships, we gain knowledge from painters. The first thing that we find from examining their work is that a boundary is merely a mathematical line. The impressionists, in particular, used directional brushstrokes or colour patches, which are small space patches with space regularity, to represent "formless" visual objects instead of clear contoured shapes. The second thing that we find is that proportions in all things bridle these space patches into different visual objects.

Like the painters, we can use a corps of multi-grey patches to represent an image. Grouping neighbouring similar pixels generates these. The proportions among the multi-grey patches bridle them into different visual objects. We then capture the proportions by the geometric mean (GM) among the grey patches and make it the indicator of spatial elements or the constitution of visual objects in an image. The spatial elements with the same GM form one kind of pooling region, and the statistical property of space averaged over the pooling regions is calculated over the GM regions.

After we figure out the calculation of the properties of the three-level computational model of text in the field of image processing, we use the model to develop an effective algorithm for text detection from the clutter scene on the public bench mark dataset, and also put it into practice on real big data i.e. the automatic processing of bank cheques.

1.4 The framework of our work

As text represents figures, for the purposes of text detection in image, its physical appearance plays a critical role. During the view construction of text, the properties of individual character and textual organisation of characters build up the physical appearance. When observers see text appearance, they speak of their feelings in terms of the crowding effect and saliency from the viewpoint of vision perception.

Thus, the properties of text appearance involve three-levels: object-centred level, space-regularity level, and individual character. The object-centred level properties refer to the properties of crowding effect and clutter and saliency. The properties of space-regularity level show the relationships of the concrete space organisation among spatial elements. The properties of individual character suggest those distinctive features of the form of characters. Our work focuses on the calculations of those properties in image processing so that we get a reasonable representation of image and three-level computational text models and then we put them into practice on public benchmark datasets and real big data i.e. bank cheques.

Firstly, the coexistence of crowding and saliency in text reveals two important aspects of the properties of text being the space averaging over pooling regions, and region-level salient structures in the object-centred level. This brings us to two subtasks: how do we capture the pooling regions, and how do we represent the composition of spatial elements in order to get salient structures at the region-level?

Noticeably, this perceptual feeling is derived from the characters' commonalities and spatial arrangement in regularity, which are inherent properties generated from the moment of typography design for text legibility, readability and conspicuity. The stylish attributes of fonts determines the characters' commonalities, including shape, stroke width and its related ratio, weight, line, size and its related proportions. In addition, several kinds of spacing in layout determine the spatial arrangement, such as letter spacing, word spacing and line spacing. While these inherent properties have provided important insights into the essential process of text appearance construction, the extent to which this understanding holds true in a clutter scene is less clear from the viewpoint of image processing. We therefore mathematically explore these properties (e.g. shape based on grey patches or edge points, pixel density, edge density, stroke width to height ratio, straight line ratio, and local RMS contrast etc) and transfer them into the viewpoint of image processing.

Additionally, focusing on the above subtasks presented by the correlate of crowding, clutter and saliency, we employ the ways in which painters represent spatial elements and relationships among them in image. Painters, especially impressionists, use directional brushstroke or colour patches as space granularity to represent visual objects as "formless" in space regularity instead of clear contoured shape sketches. This kind of space regularity of colour patches, i.e. repetitive patterns, can offer a compositional format to express an artist's feelings rather than to simply describe an object. Moreover, in painters' harmonious proportions, the component parts of an object react simultaneously so that they can be seen at one and the same time both together and separately. It is the harmonious proportions among component space patches which serve to bridle component space patches in objects.

Learning from the school of art, we describe image as a set of grey space patches in multi-grey levels. Each grey patch group neighbour pixels with similar grey tones just like the colour patches or directional brushstrokes used by impressionists. Further, among these space patches, a mathematical tool of geometric mean (GM) is used to implicitly catch the harmonious proportions of the component parts of a visual object.

Considering the statistical property of space tuning and grey tuning in crowding, GM is computed among the space patches containing the same grey level and it is considered as the space granularity to form objects. For grey tuning together with space tuning, greys which work together to form the same object are inclined to have a similar grain size so that they can be seen both together as a whole and separately as parts. Therefore, GM is regarded as an indicator of the space compositional granularity of an image and the image can be represented by several sets of grey patches at certain GM levels. In addition, the features of individual characters are extended into GM regions by space tuning, i.e. space averaging.

Clusters resulting from similarity and neighbourhood become the pooling region, thereby decreasing, modulating or breaking down the crowding effect. Accordingly, visual objects can be analysed at both the component patch level and the higher cluster level in certain GM levels.

Then, we put the representation of image and the three-level computational model of text into practice. In the clutter scene, a new solution is developed to detect text by GM analysis. For real big data—bank cheque image documents with various forms and styles, we discriminate the handwritten scripts from printed ones, and automatically read the legal amount and payee content.

Thus, the framework of our work is illustrated in Figure 1.2, which consists of three aspects. The first is about high perceptual related theory which describes the phenomenon of object-centred display, such as the crowding effect, clutter, saliency, and ergonomic guidelines. We explore the correlates among them and figure out the related subtasks in text detection, the subtasks are composed of feature of individual characters, the representation of spatial elements and the calculation of pooling region. The second is to provide solutions for those subtasks. Learning from the painters, and using the mathematical tool to capture the proportions in all things, we represent the image by a set of multi-grey level connected components, and define geometric mean (GM) on these components as the indicator of the space granularity of image. The pooling regions are regarded as GM regions. Further, features of individual characters, neighbourhood and appearance similarity are defined over regions given by the components and extended at the GM level. The third is to put the solutions into practice by developing an algorithm on the public benchmark dataset and a system over the bank cheque big data. All the experiments agree with our expectations and the performances are effective.

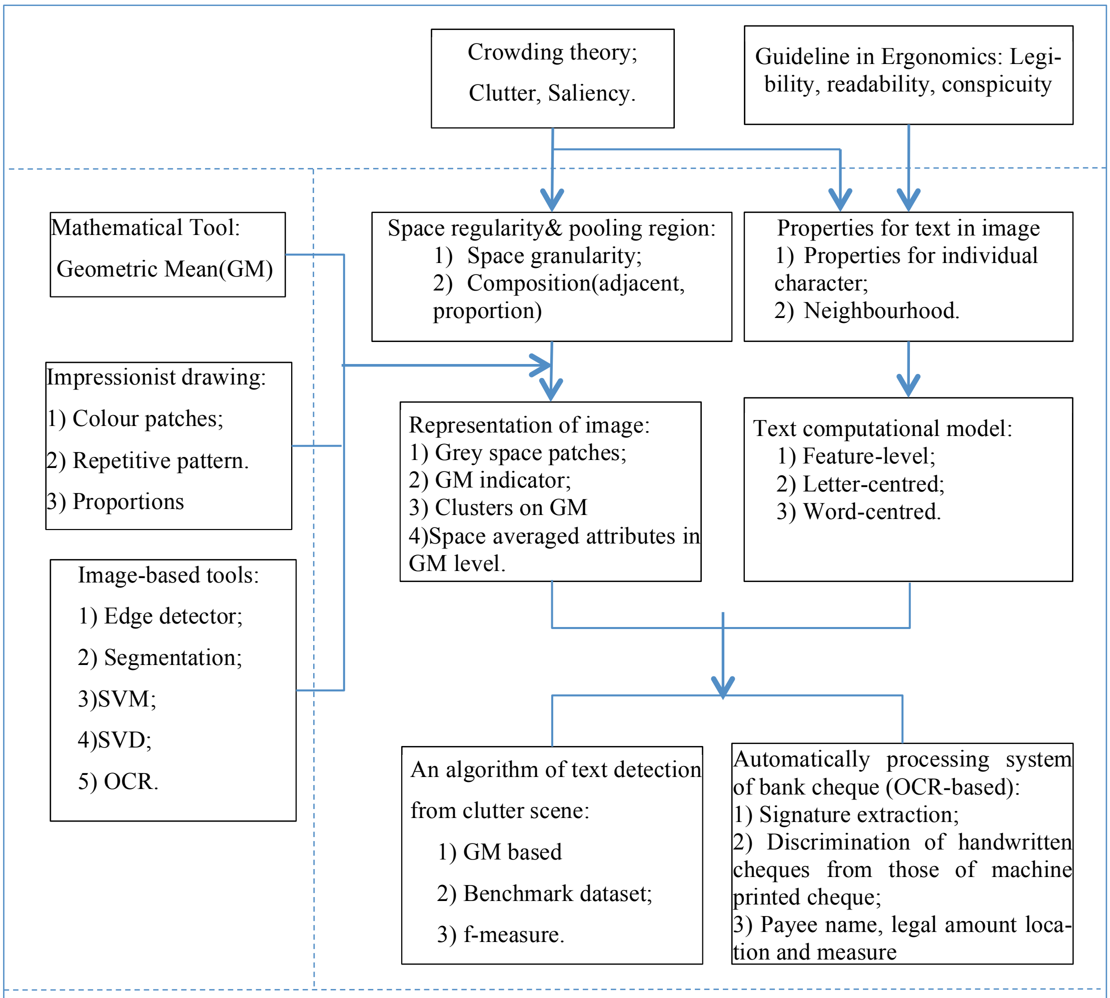

Crowding theory;
Clutter, Saliency.
Guideline in Ergonomics: Legibility, readability, conspicuity
Mathematical Tool:
Geometric Mean(GM)
Space regularity& pooling region:
1) Space granularity;
2) Composition(adjacent, proportion)
Properties for text in image
1) Properties for individual character;
2) Neighbourhood.
Impressionist drawing:
1) Colour patches;
2) Repetitive pattern.
3) Proportions
Representation of image:
1) Grey space patches;
2) GM indicator;
3) Clusters on GM
4)Space averaged attributes in GM level.
Text computational model:
1) Feature-level;
2) Letter-centred;
3) Word-centred.
Image-based tools:
1) Edge detector;
2) Segmentation;
3)SVM;
4)SVD;
5) OCR.
An algorithm of text detection from clutter scene:
1) GM based
2) Benchmark dataset;
3) f-measure.
Automatically processing system of bank cheque (OCR-based):
1) Signature extraction;
2) Discrimination of handwritten cheques from those of machine printed cheque;
3) Payee name, legal amount location and measure

Figure 1.2 the framework of the thesis work

1.5 Organization of our work

While text is ‘within-object conjunction’ [8, 47-50] or crowding effect [51-54] among built-in letters or words by nature, it nevertheless has salient features or structures to make it pop out in a clutter scene. The correlates among crowding, clutter saliency and the ergonomics criteria of text instigate our work by providing concrete subtasks of text detection. As illustrated in Figure 1.3, our thesis is organised as follows.

Chapter 1 deals with the problem of our work and introduces our motivation and aim. For the purpose of fulfilling our task, it makes clear the methodology and framework of text detection and provides the organisation of our work.

Chapter 2 deals with the concepts and theories related to text, including crowding effect, clutter, saliency and the ergonomics criteria of text. For the purpose of comprehensively understanding text, we investigate the different aspects of it. Starting from its origin of ergonomics, it transmits messages within the criterions of legibility, readability and conspicuity. From the standpoint of human perception, as an informative image figure, it has the perceptual content and semantic content to convey messages as clearly as possible. Thus, it has attention conspicuity and cognitive conspicuity, i.e., multi-level saliency from feature-level to object level. Meanwhile, it is a whole visual object as a corps of letters in space regularity (or coherence) and naturally leads to the crowding effect of the physical appearance of text. Especially, when the clutter scene is the background of text, text in space regularity is more difficult to discriminate from a similar background since clutter is tightly associated with space organisation. Therefore, the related concepts of crowding, saliency clutter and ergonomics criteria are introduced. In addition, edge point is the most informative point in image, which is not only important for features of individual character but also important for text in a string of letters. Therefore, another aim of this chapter is to introduce two kinds of edge operators, one with effective orientation, and the other with embedded confidence of location and both of them are used in the following chapters.

Chapter 3 deals with the object-centred properties of text appearance in a clutter scene, i.e., the globe properties in image. In terms of the informative figures in images, the coexistence of crowding effect, saliency, and clutter reveals the textual correlates. Indeed, the correlates enable us to figure out the concrete subtasks of text detection. As they coexist in text, all of their characteristics are shown in the text from the local level

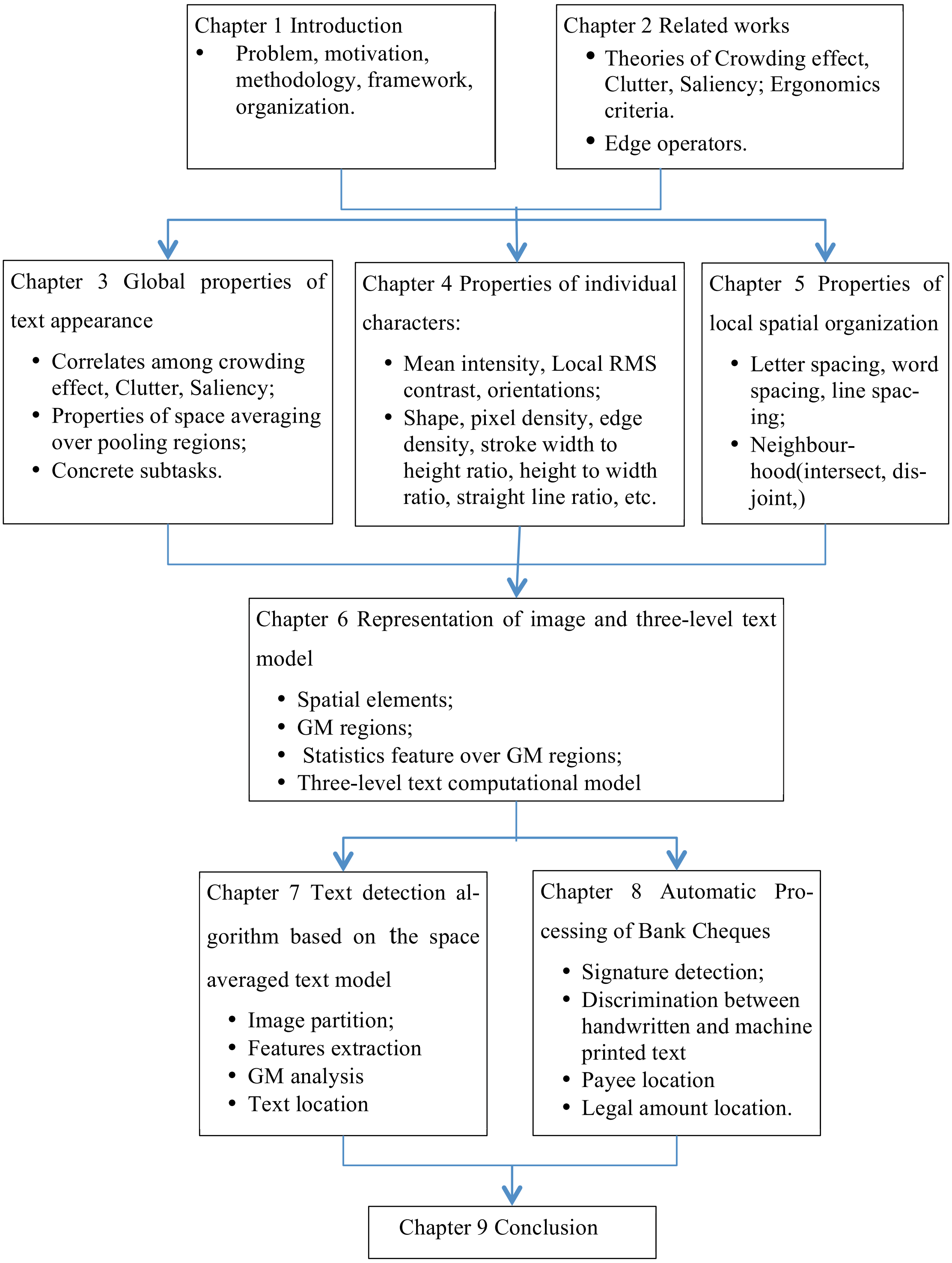

Chapter 1 Introduction
Problem, motivation, methodology, framework, organization.
Chapter 2 Related works
Theories of Crowding effect, Clutter, Saliency; Ergonomics criteria.
Edge operators.
Chapter 3 Global properties of text appearance
Correlates among crowding effect, Clutter, Saliency;
Properties of space averaging over pooling regions;
Concrete subtasks.
Chapter 4 Properties of individual characters:
Mean intensity, Local RMS contrast, orientations;
Shape, pixel density, edge density, stroke width to height ratio, height to width ratio, straight line ratio, etc.
Chapter 5 Properties of local spatial organization
Letter spacing, word spacing, line spacing;
Neighbourhood(intersect, disjoint,)
Chapter 6 Representation of image and three-level text model
Spatial elements;
GM regions;
Statistics feature over GM regions;
Three-level text computational model
Chapter 7 Text detection algorithm based on the space averaged text model
Image partition;
Features extraction
GM analysis
Text location
Chapter 8 Automatic Processing of Bank Cheques
Signature detection;
Discrimination between handwritten and machine printed text
Payee location
Legal amount location.
Chapter 9 Conclusion

Figure 1.3 The organisation of the thesis

to space organisation and the object-centred level. These include 1) statistical space averaging over the pooling region at the object-centred level which represents an important property of text appearance; 2) distinctive features of text at different levels making text possess saliency at the multi-level; and 3) space regularity which plays a critical role in relation to inner-letter space coherence. Consequently, we can delineate two clear subtasks as follows: 1) to dig the distinctive features from text construction and transplant them into image-based features in the area of image processing (i.e. observation). 2) To represent image by space organisation.

Chapter 4 deals with the distinctive characteristics of individual characters. From the viewpoint of construction, there are many properties for text legibility, readability and conspicuity, including luminance, viewing distance, shape, weight, line, stroke width, size, and the proportions among these measures. However, from the viewpoint of observation, some of them cannot be measured or are meaningless and some others need to be changed to be used in image processing or they should be kept as close as possible to the maximum level so that they can be used directly. Thus, the attributes of individual characters consist of mean intensity, local RMS contrast, shape, pixel density, edge density, stroke width, straight line ratio, height to width ratio, stroke width to height ratio, etc.

Chapter 5 deals with the local spatial organisation properties of text, i.e. the relationships among the letters which form text. The organisation properties include letter spacing, word spacing and line spacing. Letter spacing, word spacing and interline spacing are in harmony with each other and organised as a whole in order to keep crowding and readability in a good balance. All of them are highly related to the type size, and have a practical recommendation ratio respectively. From the viewpoint of image processing, all of them contribute to the space relationship of the neighbourhood, which is the most important space relation in multi-parts object---text detection.

Chapter 6 deals with deals with image representation by space organisation and text models in crowding. Image is formed by grey levels which are tune with each other. The question is how these grey levels form an image which leads to what we feel about an object.

If we consider painters, especially impressionists, they use directional brushstrokes or colour patches which are small space patches in space regularity to represent "formless" visual objects instead of clear contoured shape sketches. The space regularity of patches can offer a compositional format to express an artist's feelings about an object rather than to simply describe it. Secondly, the painter through his harmonious proportions makes the component parts of an object react simultaneously so that they can be seen at one and the same time both together and separately i.e. together, as a whole; and separately as component parts.

Similarly, image is described by many grey patches at the multi-grey level. Considering the crowding statistics property of space tuning, the granularity of the grey patches is defined as the average size of grey patches within each grey level. Since only the geometric mean can correctly capture these proportions among component parts in mathematics, the, geometric mean (GM) among grey patches is defined as the indicator of the granularity.

Grey tuning together with space tuning (i.e. greys which work together to form the same object) are inclined to have a similar grain sizes so that they can be seen at one and the same time both together (as a whole) and separately (as parts). Thus image can be mapped into several sets of grey patches in certain grain sizes and based on these sets of grey patches, visual objects can be analysed in terms of these grain sizes.

There is also another question as to how to model the text in image processing. As a figure, text is in harmony with painting theory and reveals similar phenomena.

According to the painting theory of Leonardo da Vinci [55], the object relies on these properties of object: volume, colour (or intensity), and shape. In image, the three properties are turned into space, intensity (or colour), and contour. In addition, when objects recede from the eye (or camera), the size of the object decreases. This means the space of the object in the image is reduced while the contour is lessened. If the distance is far away enough, the contours or boundary of the separated object disappear and the separated objects are merged into a whole.

The above image granularity provides a quantitative clue for text legibility. When the text is large enough, even separate letters becomes salient. Further, the text string turns out to be a salient object since there is no crowding for large stimuli. However, if the text is too small, the separate letters cannot be easily discerned and only the sheet resulting from the string of small letters can be distinguished. The granularity can roughly tell us whether the size of text is salient enough or not, or if the size is too small to be discerned as a sheet, or as ordinary sized text.

Moreover, the regularity of inter-letter spacing plays an important role in determining the strength of the crowding of text i.e. the space relations can quantitatively describe the regularity. Additionally, “straight line is ungodly” [56], text typically has distinct features beside its crowding effects and these are derived from the space regularity.

Combining the basic features, straight lines features and space regularity features, and then extending them on to different forms of granularity, the text model is built. This model represents both the string of text and a single salient letter.

Chapter 7 deals with the text extraction algorithm based text model in crowding, and a highly perceptive solution for text extraction is developed over the GM regions and the three-level text models which includes image partition, feature extraction, GM regions generation and GM analysis and text location. In this solution, Multi-Grey Connected Components (MGCC) are used to represent the intricate pattern of an image. Based on the GM indicator, we explore the composition theory among component parts, and the Geometric Mean (GM) is proposed as a new way to describe the compositional complexity of an object across meaningful MGCCs. Without following the legend framework based on supervised training, the proposed methods explore the input images on both pixel-levels through the MGCC and also at the semantic level through GM. In the end, the text regions are located and adjusted by the close regions generated from the edge points with embedded confidence. The proposed method sorts out several cases which failed when using the existing methods.

Chapter 8 deals with one practical application of our method. The representations of the image based on space regularity and the three-level computational modelling of text in the cluttered scene are put into practice through the automation of bank cheque processing. Based on the image partition, the signature can be extracted by the CSSD algorithm which also provides important features for discerning handwritten text from printed text. Meanwhile, based on the OCR, the payee name is found and measured, and the legal amount is also revealed through the means of path analysis or the VO string selection lexicon.

Finally, Chapter 9 provides the conclusion of our work.

1.6 Contribution

Thus, in summary, this thesis mainly contributes to the following aspects:

1. As text perceiving is a unitary process, which involves view construction, description and calculation of text, we explores the properties of text in local-level, spatial level and global level from the different aspects of text perception, and get the

completely understanding of text and formulate these properties into three-level computational model of text.

As text is figure in image, this computational model of text integrated features of individual characters (low-level features), properties of spatial regularity (i.e. neighbourhood, proportions among component parts, appearance similarity), and the crowding statistics property of space averaged over pooling regions.

2. Considering the view construction of text, the features of individual characters in image processing are obtained on the basis of the properties of construction, including mean intensity, local RMS contrast, shape, pixel density, edge density, stroke width, straight line ratio, height to width ratio, stroke width to height ratio, etc.
3. For the purpose of calculating the properties of space regularity and the crowding space averaging property in an image, learning from the works of painters, spatial elements and relations are quantified especially in terms of neighbourhood and proportions among component parts, and these involve space granularity and composition rules of an image. Further, Geometric Mean among space elements with the same grey level is proposed as an indicator of space granularity to capture the proportions among the component parts of an object. And those grey patches with the same GM image are named as GM regions, and image is composed of a set of GM regions.
4. Crowding pooling regions are built up as two kinds of regions: one is GM regions, named as the enlarged extended pooling regions, and the other is the compact pooling regions, which are given by clusters resulted from similarity and neighbourhoods. Therefore, the statistical properties of space averaging and those of spatial regularity are calculated over both pooling regions, and text and other visual objects are analysed over GM regions by GM indication.
5. A new brand algorithm of text detection has been developed based on the space crowding averaged model, that is, a highly perceptive solution of text extraction is developed over GM regions and the three-level text model. Experiment shows its effectiveness.
6. The computational model of text is used to build up a context-aware saliency computational model of document images with complex background, over which

the algorithms of handwriting scripts and signature extraction from bank cheques are developed.

7. A system of automatic processing of bank cheques has been developed by application of the computational model of text.

Chapter 2

Related Works

Being figures in image, text appearance in clutter scene give observers a feeling of crowding effect, clutter, but still it has salient enough to pop out the informative message. In addition, the inner-letters of text have commonalties and distinctiveness from the moment of font design, the appearance of text should be congruent with the Ergonomics criterions of legibility, readability and conspicuity. Therefore, one aim of this chapter is to introduce the related concepts of crowding, saliency, and clutter and Ergonomics criteria.

And in image processing, edge points represent the abundant information of figures, which are used to predict crowding and measure clutter, and also get desirable closed curves with increasing saliency. Thus, another aim of this chapter is to introduce two kinds of reliable edge operators.

2.1 Guideline in Ergonomics

Illustrated in Figure 2.1, text is one typical human-designed informative high perceptual visual object, including form and content. Form refers to the characters that have been arranged in a certain way, and typeface and textual organization are involved, which is the responsibility of the typographer. And content refers to the message that the author wants to communicate to the reader through its form, which is the responsibility of the author.

Before readers can ponder the ideas of the author, the message content needs to be reproduced in print. It involves author, type designer, type caster, typographer and compositor during the composition of letters into text with the aim of transmitting a message as clearly as possible. An important aspect of this process is to stimulate readers to search for the underlying structure of the message in order to aid comprehension and deeper processing of the information by enhancement of written language. According to three ergonomics criteria of *legibility*, *readability* and *conspicuity,* the well usability requirements of text is determined by three properties of text, including functional properties, semantic properties and textual organization.

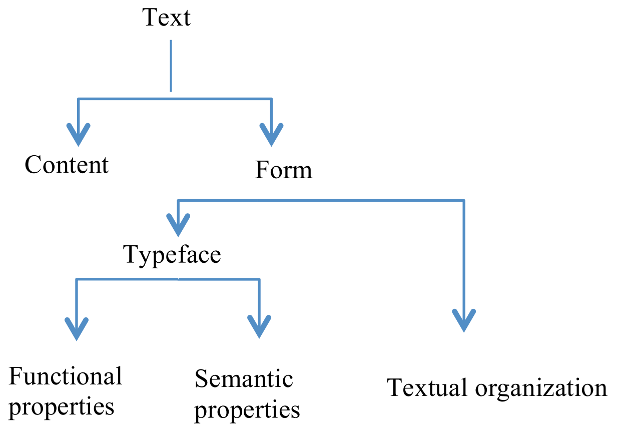


Figure 2.1 the general aspects of text

2.1.1 Legibility

"Legibility is the attribute of alphanumeric characters (letters and numbers) that makes it possible for each one to be identifiable from others. This depends on such features as stroke width, form of characters" and the amount of space between characters[57].

Notice, there is a distinction between character legibility and text legibility. Character legibility is the ease with which a person can identify an individual character as a particular letter or number. Legibility of text refers to the ease with which groups of characters are correctly identified as a word, with the result that the reader perceives meaningful sentences, which has high relation with composition of text.

High legibility is very important for reading-intensive print. And it is affected by level of illumination, background contrast and reader fatigue. If a text is not very legible, this will need more efforts for reader to identify the letters correctly.

2.1.2 Readability

"Readability is a quality that makes possible the recognition of the information content of material when it is represented by alphanumeric characters in meaningful groupings, such as words, sentences, or continuous text... [Readability] depends more on the spacing of characters and groups of characters, their combination into sentences or other forms, the spacing between lines, and margins than on the specific features of the individual characters" [57]. Thus, for a high level of readability, the composition of text needs to provide easy access to the information that is contained in the words, whose related science studies will be discussed in details from the different viewpoint in following sections. Apart from composition, readability is concerned with author's expressing precisely what one means in an unambiguous manner and will therefore not be discussed here.

There are differences between readability and legibility. Readability not only pertains to paragraphs of text, but also to tables, footnotes, and other special text formats. When a text is of low legibility, its readability is also low. When a text is not very readable, on the other hand, it is still possible that it is highly legible. For example, a brochure is printed in too small typeface and the characters have such indistinct shapes, that readers can hardly distinguish between the 'i' and the 'l' or the 'h' and the 'b'. In such a case, the text is of low legibility. Consequently, the text is not very readable either. Then, the brochure is reprinted in a more legible way, the same conditions of easy word distinction and correspondence between text and illustrations would make a more readable text. It is also possible, however, that the text has become highly legible, but that the illustrations are not numbered and are referred to in the text on a different page. In this case, readability would still be low.

2.1.3 Conspicuity

Conspicuity is the "quality of a character or symbol that makes it separately visible from its surroundings" [57]. Usually, the use purpose and requirement determine whether text is designed as a more conspicuous object than its surrounding.

When text is used in logotype or in sign for the purpose of navigation, such as traffic signs, hazard signs and billboards, highly conspicuous text receives more attention than visual objects, textual or other, which are less conspicuous. As a conspicuous object, according to Cole and Jenkins [58], is one that will, for any given background, be seen with certainty probability ($p>.9$) within a short observation time ($t<.25$ s) regardless of the location of the target with respect to the line of sight.

Here two kinds of conspicuity [59] get involved: attention conspicuity, which is the capacity of the target to attract attention when the observer’s attention is not directed to its likelihood of occurrence, and search, cognitive conspicuity, which is defined as the accessibility of the target when the observer is explicitly directed to look for the object.

Attention conspicuity depends upon the prominence of its physical properties compared with its background, which can be reduced by visual clutter[60] Some of variables affects attention conspicuity, such as font, letter size, spacing, and layout of characters, luminance and colour contrast with surround, distinctive shapes compared with other visual objects, display text content including information arising from the unusual or unexpected character of text. However, its computational work will be delta with in the following chapters.

2.2 Crowding effect

2.2.1 Definition

Letters are arranged to form text with textual properties. ‘‘It is as if there is a pressure on both sides of the word that tends to compress it. Then the stronger, i.e. the more salient or dominant letters, are preserved and they ‘squash’ the weaker, i.e. the less salient letters, between them.” This is Korte’s original, and often referred to, description of crowding [46]. And Levi, et al [8] suggests that the reader can experience this phenomenon by viewing Figure 2.2. Actually, the term ‘‘crowding” (‘‘Gedra¨nge”) has no counterpart in German perception research (including reading) and has been first used by Ehlers [61] in 1936.

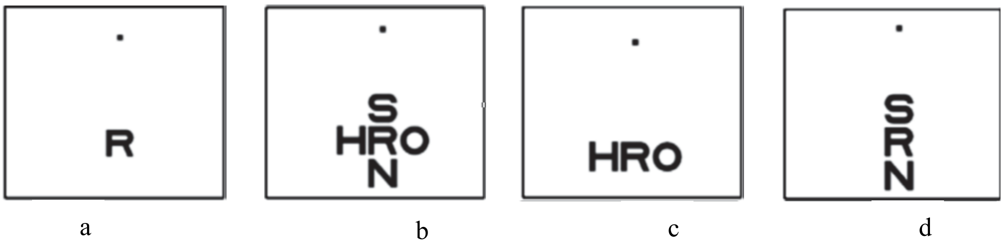


Figure 2.2 Crowding. After[8]. The reader can experience crowding by fixating the dot, and trying to identify one letter: in isolation (a), surrounded by 4 random flanking letters (b), surrounded by 2 horizontally placed random flanking letters (c), surrounded by 2 vertically placed random flanking letters (d).

Crowding, the deficit in discriminating target when other shapes are nearby, is a form of inhibitory interaction in spatial vision efficiency by limiting attention for recognition and attention against competition. Crowding impairs not only discrimination of object features and contours, but also the ability to recognize and respond appropriately to objects in clutter. Thus, studying crowding might lead to a better understanding of the processes involved in text detection from clutter scene.

2.2.2 Study objects

Traditionally, crowding has been studied with letters and numerals as stimuli [49, 62, 63]. And now there is great diversity in the stimuli used, ranging from letters, words, oriented bars, Gabor patches and shapes, faces, letter-like stimuli and natural scenes. From Bouma's highly influential report, it has stated that 'for complete visual isolation of a letter presented at an eccentricity of ϕ°, it follows that no other letters should be present (roughly) within $0.5\phi^\circ$ distance.'

Also, it has been demonstrated with discrimination of simple features like contrast, orientation, and spatial frequency [63, 64]. With simple detection, however, the effect seems to be much smaller. But recent study suggests that detection and coarse orientation discrimination are not immune to crowding[65]. And the crowding effects with these tasks depend on the number of flankers-performance, which is worse in the presence of more flankers.

Moreover, crowding effects have been reported to occur in a wide variety of tasks including: letter recognition [49] [66-68]; Vernier acuity [68, 69]; orientation discrimination [63, 70]; stereoacuity[71, 72] and face recognition [50, 54, 72, 73]. Crowding occurs for chromatic stimuli with equiluminant backgrounds, with similar extents to crowding in the luminance domain[74]. And crowding also occurs for moving stimuli [75].

Additionally, since natural environments are replete with structure and most of our visual field is peripheral, crowding represents the primary limit on vision in the real world, Wallis,et.al[76] has examined where crowding occurs in natural scene, revealed that target size, eccentricity, local Root-Mean-Squared contrast and edge density can be used to make reasonable predictions of the likelihood that an observer will experience crowding.

As a matter of fact, there is compound crowding[54]: in a given scene, crowding occurs selectively between features[8], object parts[50] and whole objects[72, 77]. Actually, object-cantered crowding effects demonstrate compound crowding within the same stimulus– crowding between the whole objects, and between the low-level features or parts that comprise each whole object –which suggests that crowding operates at multiple stages. These make an all-convergent crowding stage unlikely. Moreover, crowding is specific to the similarity between and the configuration of target and flanks. These evidence casts doubt on the idea that crowding is a unitary effect due to a single stage of processing, although this is implicitly assumed in most studies on crowding [8, 73], and suggest that there is multi-level crowding.

The emerging consensus from these studies is that crowding is a consequence of spatially pooling features within receptive fields of increasing size: information is averaged ([78-84] , or not resolved by attention ([85-87]and therefore some is lost. Moreover, there are a number of features or hallmarks of crowding and many models for crowding.

2.3 Saliency

Kim has proposed that image content can be divided into perceptual content and semantic content[88]. Perceptual content includes attributes such as colour, intensity, shape, texture, and their temporal changes, whereas semantic content means objects, events, and their relations.

Since text is one of the typical informative image figures, it has perceptual content and semantic content to transmit messages as clearly as possible from author to readers. Thus, it has both attention conspicuity and cognitive conspicuity even though it has crowding effect by nature. That is to say, it obviously has multi-level saliency from feature- level to object level.

2.3.1 General definition

Generally, saliency is defined as what captures human perceptual attention. Here two stages of visual processing are involved: first, the parallel, fast, but simple *pre-attentive* process; and then, the serial, slow, but complex *attention* process. Properties of pre-attentive processing have been discussed in literatures[89-91], the highly influential feature integration theory[89] and Koch's shifting attention selection[90] explains the visual search strategies.

2.3.2 Saliency map

To find the "proto objects", the attention process is often modelled using a saliency map: an internal map calculated by some preattentive mechanism[92] and representing the estimated priorities assigned to every location[93]. As a pioneer, Itti et al [94]has proposed a well-known saliency model, in which saliency is based on the centre-surround contrast of units modelling simple primary features such as colour, intensity and orientation. Thus visual input is first decomposed into a set of feature maps. Within each map, different spatial locations compete for saliency, such that only locations which locally stand out from their surroundings can persist. All feature maps feed, in a purely bottom-up manner, into a master "saliency map". Since then, the centre-surround scheme has been widely exploited in a variety of saliency models, due to its clear interpretation of the visual attention mechanism and its concise computational form. However, this saliency model focuses on identifying the fixation points that a human viewer would focus on first [94].

Generally, the bottom-up approaches consist of the following three steps [95]: s1) extraction: Multiple low-level visual features are extracted. s2) Saliency computation: using feature vectors, the saliency of each image pixel is computed and then normalized and linear/non s linear is combined to form a master map or a salient map. S3) A few key locations on the saliency map, just like human fixation locations, are identified by winner-take-all, or inhabitation-of return, or other non-linear operation.

A number of features are fed to the saliency map, including local contrasts of colour, orientation, texture and shape features, oriented sub-band decomposition based energy [96], ordinal signatures of edge and colour orientation histograms [97], Kullback-Leibler (KL) divergence between histograms of filter responses [98], local regression kernel based self-resemblance [99], and earth mover's distance (EMD) between the weighted histograms [100].

And the surrounding region of the centre pixel/region is selected as the maximum symmetric region [101], and the whole region of the blurred image in the frequency-tuned saliency model [102].

Besides, Mertsching et.al [103] used several region-based features to generate the region-level saliency map based on segmented regions. Note that, recently, to deal with complicated images, Liu et.al [104] have proposed a saliency tree for an image in a hierarchical saliency representation, in which each leaf node represents a primitive region with regional saliency generated by integrating global contrast, spatial sparsity, and object prior with regional similarities. Then by exploiting a regional centre-surround scheme based on node selection criterion, a systematic saliency tree analysis, a regional saliency map and pixel-wise saliency map are obtained.

2.3.3 Salient structure

Additionally, detecting salient structures is a basic task in perceptual organization. In a given image, image edges are more informative than other image parts, which are detected by the human visual system, and further filtered by the attention process, which is effectively able to discriminate the important image edges, known as the “figure,” from the less important edges known as the “background.”

To account for this phenomenon, using a computational theory, Shashua and Ullman [105, 106] have proposed a measure denoted saliency, which grows with the length and smoothness of the curve on which edge-points lie. And the properties of this saliency measure have been considered in [107]. Further, Berengolts A. et.al [108] have modified this saliency estimation mechanism that is based on probabilistically specified grouping cues and on curve length distributions.

2.3.4 Other computational schemes

Additionally, there are other formulations for measuring saliency based on different theories and principles such as information theory, frequency domain analysis, graph theory and supervised learning.

Within the framework of information theory, saliency refers to rarity represented by the self-information of local image features [109], the complexity measured by local entropy [110], and the average transferring information measured by entropy rate [111]. According to frequency domain analysis, the saliency map is generated by exploiting the spectral residual of the amplitude spectrum of Fourier transform [112], the phase spectrum of quaternion Fourier transform [113], and contrast sensitivity function in the frequency domain [114]. Using graph theory, a saliency map is generated at different levels by random walks on the weighted graph constructed at pixel level [115] [31] and block level [116], and the stochastic graph model is constructed on the basis of region segmentation [117]. In supervised learning, under the framework of conditional random field [118], a set of features including multi-scale contrast, centre-surround histogram and colour spatial distribution are integrated to generate the saliency map. Moreover, region feature vectors are mapped to saliency scores across multiple levels to generate the saliency map [119].

Recently, object-level saliency is generated research interest. W.Zhang et.al [120] exploit GMMs to explicitly construct a salient object/background model. A generic objectiveness measure [121] and object-level closed shape prior are effectively incorporated into the saliency models presented in [122] and [123] respectively. Under the framework of low-rank matrix recovery, region segmentation based object prior [124], centre prior, colour prior and learnt transform prior[125] are exploited for saliency detection .

2.4 Clutter scene

2.4.1 Definition

Clutter scenes are composed of numerous objects, textures and coloured regions, which are arranged in a variety of spatial layouts. And clutter is the state in which excess items, or their representation or organization, can cause crowding and lead to a degradation of performance at some task [126, 127]. This definition of clutter brings up two key points: the association between clutter and the representation or organization of space information, and the notion that clutter may depend upon the user's task.

Further, there are a number of conspicuous similarities between clutter and crowding[128]. First, both crowding and clutter increase with information density. Second, although they yet cannot be fully explained by acuity lost, both phenomena are most prominent in the periphery of the visual field. Third, both of them have closely relations to the spatial organization and degrade performance on visual tasks, which involves the excessive feature integration over inappropriately large area. Thus, crowding can model the clutter based on the analysis of spatial layout, and the measure of clutter can be used to capture and compute the crowding[129]. Inspired by these findings from these visual perception studies, clutter has its basis in visual crowding, and crowding is an important constituent of clutter. Thus, crowding and clutter may indeed be closely related concepts[128].

2.4.2 Clutter measurement

Starting from the practical applications of information visualization or display, there are strong attempts to quantify something like ‘clutter’, what they call information density.

In this context, several measures of information density have been proposed. For instance, the number of visual objects, and the number of vertices, the number of elements on a web page where an element consist of a word, graphic, or “interest area”, the number of entries in the source data matrix per unit area, the number of graphic tokens per unit area, the number of vectors needed to draw the visualization, the length of the program to generate the visualization, the amount of “ink” per unit area as a metric for simple black & white maps. Certainly, the amount of clutter has some dependency upon the number of objects, graphic tokens, or entries in the source data matrix in the display. However, counting the number of objects does not take into account the appearance or organization of the objects.

Since the clutter scene seems to be complex and congested in feature space, and has high relations to saliency and colour density, four kinds of measures are proposed to represent the different aspects of clutter: edge density, feature congestion, subband entropy, cluster density. Edge density is the space-averaged binary output of the edge detector, where higher values denote more ‘edge’ per unit of area. The feature congestion measure captures a bit of space organization as well as the sub-band entropy measure. Clutter density explicitly captures the space statistical properties, but still ignores the relationships among the regions.

Although all of them only capture a limited overview of the perceptual space organization of clutter, they can inspire us to explore and formulate an appropriate representation for our particular task. And the following sections will deal with the computational details of these measures.

2.4.2.1Edge density

A visual pattern is also seen to be complex if its parts are difficult to identify and separate from each other. Yet, paradoxically, when the parts are separated or conceptualized as a whole the valence of the complexity changes and the pattern becomes simpler [130]. This suggests that the perceived complexity of an image also depends on the amount of perceptual grouping, a characteristic independent of the quantity of parts that an observer perceives in the scene.

Exploring the space organization of clutter, Oliva et al[131] have attempted to determine what factors influence the human representation of 'complexity' which is clearly related to that of clutter. And they have suggested that complexity depends upon the quantity and variety of objects, detail and colour, as well as upon higher-level, more global concepts like the symmetry, organization, and 'openness' of the depicted space.

Since the presence of edges plays an important role in object discrimination in space, they have provided edge density[132] to predict subjective judgements of image complexity. The edge density measure attempts to capture the notion of clutter as number of objects by calculating the density of edges---the percentage of pixels that are edge pixels, as well as a likely correlation of clutter with high frequency content. And it implicitly captures the colour variability since colour variability co-varies to a large extent with edge density, when there is a change in object, there is an edge and there also tends to be a change in colour.

2.4.2.2 Feature congestion

From the viewpoint of the visual search in display, the search could be used to determine basic features of the visual system: search for a target defined by a basic feature ("feature search") would be paralleled, whereas search for a target defined by a combination of basic features ("conjunction search") would be serial. Intuitively, complex scenes should contain a larger variety of parts and surfaces styles, as well as more relationships between these regions, and there is less room in feature space to add new salient items. Rosenholtz et al [9, 126] refer to this condition as feature congestion and propose that feature congestion is one of the major causes of clutter, shown in Figure 2.3.

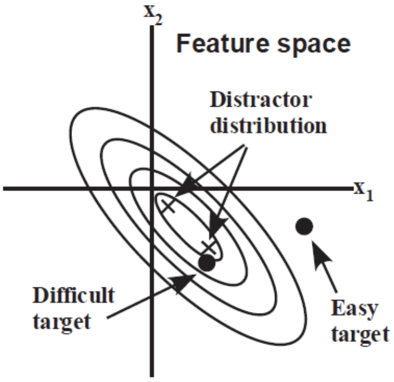


Figure 2.3 After[9] : Graphical depiction of the Statistical Saliency Model. Ellipses represent points of equal saliency. Outer ellipses correspond to greater saliency and easier searching.

Increased congestion leads to degraded performance, e.g. in visual searching. Thus they present the feature congestion measure of display clutter.

Moreover, the saliency of an item corresponds qualitatively to the ease of search for that item if it were the target, and correlates with the likelihood that a user makes an eye movement to that item. And the saliency measures are more versatile for evaluating displays. In looking for a qualitative measure of the clutter in a display, qualitative measures of search performance like saliency measures should suffice. There are roughly two categories of saliency: one based on biologically inspired mechanisms, and the other based on a functional rather than biological level.

In the first types, Itti's model[94] is perhaps the most popular model of bottom-up visual saliency, it starts with linear filters similar to the receptive fields found in the visual cortex, and then it is applied to a variety of nonlinear neural-like operations. Based on the centre-surround contrast of units modelling simply primary features such as colour, intensity and orientation, different spatial locations compete for saliency, such that only locations which locally stand out from their surroundings can persist.

In the second types, for example, Rosenholtz's model [133], these utilize the notion that the visual system is designed to characterize various statistical aspects of the visual display. First, it represents the features of each display element as a point p_i in an appropriate uniform feature space and these features are likely to include such things as contrast, colour, orientation, and motion. From the distribution of the features present in the display, we compute the mean and covariance of the distractor features, μ and $\sum$, respectively. The model then defines target saliency as the Mahalanobis distance Δ, between the target feature vector, T and the mean of the distractor distribution, where $\Delta^2 = (T-\mu)' \sum^{-1} (T-\mu)$. In this equation, T and μ are vectors, $\sum$ is a matrix, and the prime indicates a vector transpose. The model uses, as the measure of target saliency, the number of standard deviations between the target feature vector and the mean distractor feature vector.

Essentially, the statistical saliency model represents the local distribution of features by a set of covariance ellipsoids in the appropriate feature space, shown in Figure 2.3. The volume of the local covariance ellipsoid represented by Δ therefore gives a measure of the local clutter in a display, i.e. of the difficulty of adding a new, salient item to a local area of a display. Locally measuring the ellipsoid size, and pooling over the relevant display area, gives a measure of clutter for the whole display. A target with a feature vector on the $n\sigma$ ellipsoid will have saliency Δ=n. The farther out the target feature vector lies on these nested ellipsoids, the easier the predicted search.

According to the features in Itti's model[94] and the statistical computation method, the implementation of the feature congestion clutter measure involves four stages: 1) compute local feature (co)variance at multiple scales, and compute the volume of the local co-variance ellipsoid; 2) combine clutter across scale; 3) combine clutter across feature types; and 4) pool over space to get a single measure of clutter for each input image.

However, only a bit of perceptual space organization of clutter is captured implicitly by the feature congestion measure through looking at feature covariance; it essentially captures to some extent the measure of the grouping by similarity + proximity in the display.

2.4.2.3 Sub-band entropy measure

To the extent that an image contains redundancy, it can be represented with an efficient code while maintaining perceptual image quality. Currently, by making use of the same sorts of redundancies as human visual system, sub band image coding methods such as JPEG 2000 are efficient and highly successful.

Based on the notion that clutter is related to the number of bits required for sub band (wavelet) image coding, sub-band entropy measure is presented. A wavelet coder first decomposes the image into a set of sub-bands with different orientations and spatial frequencies. Then, the Shannon entropy within each sub-band is computed as follows:

$$H = \sum_i - p_i \log(p_i) \tag{2.1}$$

Here, p_i is the probability distribution of coefficients in each sub band, and this is estimated by binning the sub band coefficients into bins indexed by i , and computing a histogram. And the number of bins is equal to the square root of the number of coefficients. Finally, the clutter measure is computed as a sum of these sub band entropies,

Sub band entropy also implicitly deals with certain aspects of perceptual organization such as grouping by a combination of proximity and similarity as feature congestion measure does.

2.4.2.4 Cluster density with varying levels of clutter

In naval navigation display, Lohrenz et al [129, 134-136] have suggested that clutter is a function of "colour density" and "saliency". Saliency refers to how clearly one colour or feature "pops out" from the surrounding features in an image, which is estimated by a weighted average of colour gradients between adjacent features. Colour density refers to how closely-packed are similarly-coloured pixels within the image. And it is computed by clustering all the image's pixels in proximity of location and similarity of colour, such that adjacent pixels of similar colours cluster together, and calculating the density of pixels in each cluster as the number of clustered features divided by the area of the polygon for bounding cluster.

After clustering all pixels in the image into bounded polygons for a given "seed colour" s , a cluster density Dp is calculated for each cluster polygon p :

$$Dp = \sum (WcNc) / Ap \tag{2.2}$$

Where: WC = Weighting factor for colour c $= 1 - Ec/M$

Ec = Euclidean distance between colours c and s in the chosen colour space;

e.g., for CIE $L*a*b$:

$$= SQRT\left\lfloor (Lc - Ls)^2 + (a_c - a_s)^2 + (b_c - b_s)^2 \right\rfloor$$

M = Maximum distance between colours in chosen colour space

Nc = Number of pixels of colour c in the cluster polygon

Ap = Area of cluster polygon p .

The colour of each pixel in the cluster will be within a colour distance of z from all immediately surrounding pixels in the cluster, starting with pixels of colour s. In other words, the cluster will "chain" pixels together to form the cluster, starting with each pixel of colour s and subsequently including all other pixels within a geospatial distance of x, y and a colour distance of z . If $z = 0$, then $Dp = Ns / Ap$.

Note that there is an inverse relationship between clutter and "density": higher density tends to predict lower clutter, since density describes how closely-packed like-pixels are in the image.

Furthermore, based on cluster density, both local density D_s and global density D_I are defined as the local and global clutter metrics respectively.

Local density estimates how much an individual seed colour (s) contributes to the overall clutter of the image. D_s is computed as the weighted average of the densities for all clusters centred on colour s

$$D_s = \sum (Dp * Ap) / A_s \tag{2.3}$$

Where: Dp =Density of cluster p mentioned in above paragraph.

As = Sum of areas of all clusters for colour s.

Global density (D_I) estimates the clutter for the entire image and it is computed as the weighted average of the local clutter densities for all colours in the image:

$$D_I = \sum (Ds * A_s) \Big/ A_I \tag{2.4}$$

Where: Ds = Weighted average of clutter densities for all clusters centred on colour s

A_I = Sum of all A_s 's for image i .

Further, local clutter metric is defined as $(1 - Ds)$ for colours, and global clutter metric is defined as $(1 - D_I)$ for image i . However, the local clutter metric does not account for saliency, which might explain why certain features, for instance, red colour, are listed with higher clutter values than expected.

Then, local saliency of a given colour or features is estimated by a weighted average of the colour differences between each colour or feature of interest and the immediately adjacent colours or features. And global saliency is estimated as the weighted average of the local saliencies for all colours or features in the image. Greater colour distances result in greater saliency.

Finally, based on colour density and saliency, the clutter model is proposed as follows:

$$clutter = 15 * (1 - colour \quad density) * \exp\left[-6.3 * \exp\left(-{}^{saliency}\!/_{10}\right)\right] - 0.0002 \qquad (2.5)$$

Here colour density and saliency are the same as the above mentioned, and use local or global values corresponding to the local clutter and global clutter [135] respectively. For very low saliencies, clutter remains very low, regardless of colour density. When saliency is high, clutter becomes a function of colour density only.

2.5 Basic edge operators

Since edge density and orientation are very important to measure or predict crowding, clutter, and also edge points are significant to get desirable closed and infinite curves with increasing saliency. Thus, reliable edge operators are required. In our work, the Kirsch operator will be used to capture edge density and orientation, and the edge detector with embedded confidence will be used to obtain the reliable locations of edge points.

2.5.1 Kirsch operator

Edge information for a particular pixel is obtained by exploring the brightness of pixels in the neighbourhood of that pixel. If all of the pixels in the neighbourhood have almost the same brightness, then there is probably no edge at that point. However, if some of the neighbours are much brighter than the others, then there is a probably an edge at that point.

Measuring the relative brightness of pixels in a neighbourhood is mathematically analogous to calculating the derivative of brightness. Brightness values are discrete, not continuous, so we approximate the derivative function. A Kirsch edge detector algorithm is used to detect edges in 8-bit grey scale images.

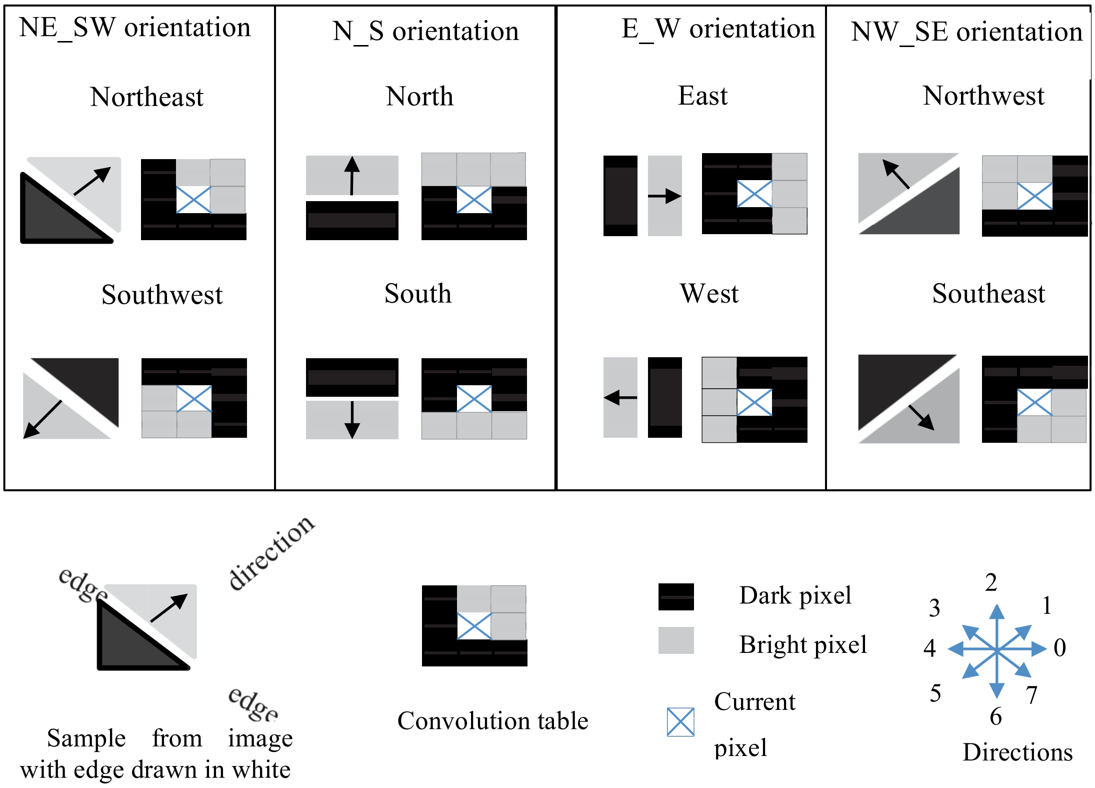


Figure 2.4 Four orientations and eight directions

The Kirsch operator is a non-linear edge detector that finds the maximum edge strength in a few predetermined directions. It identifies both the presence of an edge and the direction of the edge, illustrated in Figure 2.4. There are eight possible directions: north, northeast, east, southeast, south, southwest, west, and northwest.

The operator takes a single kernel mask and rotates it in 45 degree increments through all eight compass directions: N, NW, W, SW, S, SE, E and NE. For each direction, Figure 2.4 shows an example of edge, a convolution table, and the encoding of the direction. In the image sample, the edge is drawn in white and direction is shown with a black arrow. Notice that the direction is perpendicular to the edge. The trick to remember the edge direction is that the direction points to the brighter side of the edge. The eight directions are grouped into four orientations: NE_SW, N_S, E_W, and NW_SE.

The edge magnitude of the Kirsch operator is calculated as the maximum magnitude across all directions:

$$h_{m,n} = \max_{z=0,1,2,3,4,5,6,7} \sum_{i=-1}^{1} \sum_{j=-1}^{1} g_{ij}^{(z)} \cdot I_{m+i,n+j} \qquad (2.6)$$

Where z enumerates the compass direction kernels, and $g^{(z)}$ as shows

$$g^{(0)} = \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix}, g^{(1)} = \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix}, g^{(2)} = \begin{bmatrix} +5 & +5 & +5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix},$$

$$g^{(3)} = \begin{bmatrix} +5 & +5 & -3 \\ +5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}, g^{(4)} = \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix}, g^{(5)} = \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix},$$

$$g^{(6)} = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix}, g^{(7)} = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}.$$

For a convolution table $g^{(z)}$, calculating the presence and direction of an edge is done:

$$DerivNE = g^{(1)} * I; \quad DerivSW = g^{(5)} * I; DerivN = g^{(2)} * I; \quad DerivS = g^{(6)} * I;$$

$$DerivE = g^{(0)} * I; \quad DerivW = g^{(4)} * I; DerivNW = g^{(3)} * I; \; DerivSE = g^{(7)} * I.$$

Then, find the value and direction of the maximum derivative,

$$EdgeMax = \text{Maximum of eight derivatives}; \quad DirMax = \text{ Direction of } EdgeMax.$$

Notably, the following priority order determines which direction gets picked if more than one derivative has the same magnitude. (a) $DerivW$; (b) $DerivNW$; (c) $DerivN$; (d) $DerivNE$; (e) $DerivE$; (f) $DerivSE$; (g) $DerivS$; (h) $DerivSW$. This means that if, for instance, $DerivN$ and $DerivE$ are equal, $DerivN$ must be picked. Shown in Figure 2.5, the edge map is obtained by Kirsch operators with the presence of an edge and the direction of the edge.

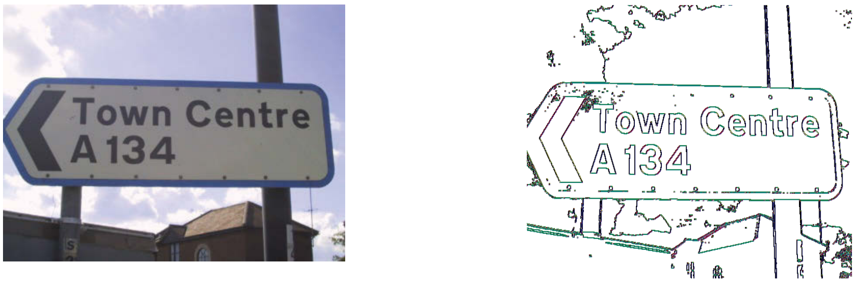


Figure 2.5 Input image and its Kirsch's edge map

2.5.2 Edge detection with embedded confidence

Usually, the optimality of an edge detector can only be assessed in the context of a well-defined task. That is, the quality of the edge map is directly related to the amount of supportive information it carries into the subsequent processing stages. Since this information is extracted after the edge map is generated, a measure of confidence should be associated with the bottom-up information stream.

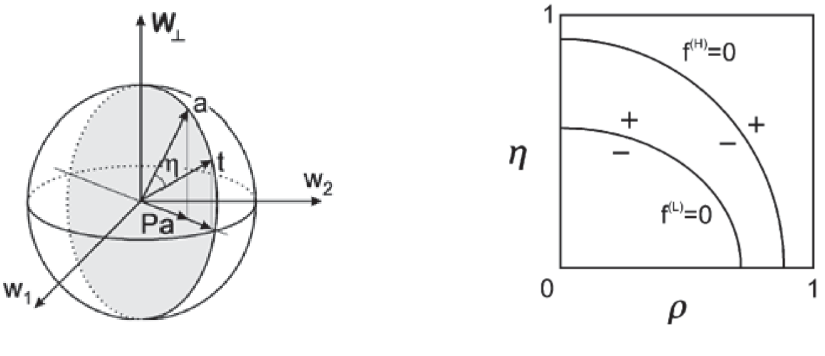
The paper [2] defined a confidence measure by using information inherently existing in the regular sampling lattice, which was not employed in the computation of the gradient magnitude, and proposed an edge detection approach with it.

2.5.2.1 Confidence measure

An often performed operation in computer vision and image processing is computing the weighted average of the data in a $(2m+1)\times(2m+1)$ window sliding over the image. The data $\{a_{ij}\}$and the weights $\{w_{ij}\}, i, j = -m, \ldots, 0, \ldots, m$ are combined to obtain

$$output = \sum_{i=-m}^{m}\sum_{j=-m}^{m} w_{ij} a_{ij} \tag{2.7}$$

And the output is associated with the centre of the window, i.e., the location on the sampling lattice corresponding to the window coordinates $i = j = 0$.



a) b)

Figure 2.6 a) Window operators as elements in a vector space, and b) $\rho-\eta-$diagram . After[2]

Using a_{ij} or w_{ij} as the element on the ith row and jth column, the $(2m+1)\times(2m+1)$data $\mathbf{A}$ and weight $\mathbf{W}$ matrices can be defined. The latter is the mask applied by the window operator. Written as a matrix inner product (2.7) becomes

$$output = \mathrm{trace}[\mathbf{W}^{\mathbf{T}}\mathbf{A}] = \mathrm{trace}[\mathbf{W}\mathbf{A}^{\mathbf{T}}] \tag{2.8}$$

Where the invariance properties of the trace are used, See Appendix A for a short compendium on matrices. The output of the window operator can be also written as a vector inner product, where the vectors $a = \mathrm{vec}[\mathbf{A}]$ and $w = \mathrm{vec}[\mathbf{W}]$ are obtained by stacking up the columns of the corresponding matrices

$$output = \mathbf{w}^{\mathbf{T}}\mathbf{a} = \mathbf{a}^{\mathbf{T}}\mathbf{w} \tag{2.9}$$

In $\mathbf{R}^{(\mathbf{2m+1})^2}$ the vector $\mathbf{w}$ defines a one-dimensional subspace and let $\mathbf{w}_\perp$ be its $\left[(2m+1)^2-1\right]$-dimensional orthogonal complement. Since for any $b \in \mathbf{w}_\perp$ the output of the window operator is 0, such data is "invisible" to the window operator. As a direct consequence we have

$$output = \mathbf{w}^{\mathbf{T}}(\mathbf{a}+\mathbf{b}) = \mathbf{w}^{\mathbf{T}}\mathbf{a} \tag{2.10}$$

Showing that a very large number of data vectors (image neighbourhoods) yield the same response. This fact is well-known in the vision literature. For example, it is often observed that the gradient operator can give a large spurious response in an apparently unstructured neighbourhood. Actually, by approaching the window operation in $R^{(2m+1)^2}$ it is possible to predict such behaviour. In practice a low-level computer vision task requires combining the output of several window operators. For example, the gradient is estimated using two differentiation masks. The two differentiation masks define a hyper-plane in $R^{(2m+1)^2}$. Let $\mathbf{w}_\perp$ be the $\left[(2m+1)^2-2\right]$-dimensional orthogonal complement of this plane, illustrated in Figure 2.6a. $\mathbf{w}_\perp$ is the null space of the gradient operator. The projector onto the subspace (plane) of the gradient operator is the $(2m+1)^2 \times (2m+1)^2$ matrix

$$\mathbf{p} = \frac{\mathbf{w}_1\mathbf{w}_1^{\mathbf{T}}}{\mathbf{w}_1^{\mathbf{T}}\mathbf{w}_1} + \frac{\mathbf{w}_2\mathbf{w}_2^{\mathbf{T}}}{\mathbf{w}_2^{\mathbf{T}}\mathbf{w}_2} \tag{2.11}$$

Without loss of generality it can be assumed that the data is normalized to a unit vector, $\|\mathbf{a}\| = 1$. Its projection onto the plane of the gradient operator is the vector $\mathbf{Pa}$. The definition of $\mathbf{w}_1$ and $\mathbf{w}_2$ implies that the orientation of $\mathbf{Pa}$ in the plane is the estimated orientation of the gradient $\hat{\theta}$. An ideal edge template t, with the same estimated gradient orientation $\hat{\theta}$ can now be defined. Thus, the unit vector t is always located in the plane $\langle \mathbf{a}, \mathbf{Pa} \rangle$ somewhere outside of the subspace of the gradient operator (Figure 2.6a)). Since only the estimated gradient orientation was used to define t, only the pattern of the data was taken into account. Inspecting Figure 2.6a, the definition of a simple measure of confidence for the presence of an edge in the data processed by the gradient operator is as thus:

$$\eta = \left| \mathbf{t}^{\mathrm{T}} \mathbf{a} \right| \tag{2.12}$$

Both t and a being unit vectors, η is the absolute value of the cosine of their angle in $R^{(2m+1)^2}$. Interpreted in the image domain, η is the absolute value of the correlation coefficient between the normalized data and the template. Moreover, the template is defined using only $\mathbf{Pa}$, i.e., the information contained in the subspace of the gradient, while η is computed based on t and a which are vectors in $R^{(2m+1)^2}$. The confidence measure incorporates information from both the data and the template which is not in the gradient subspace and, thus, is not used to determine $\hat{\theta}$. Therefore, η provides an independent estimate for the presence of the assumed edge model in the processing window.

2.5.2.2 Gradient estimation in the discrete domain

In the discrete domain, only the samples $f(i,j)$ are available and the two partial derivatives have to be computed by numerical differentiation. A possible approach is to approximate the local structure of $f(x,y)$ by a polynomial surface which takes the value $f(i,j)$ at the sampling points. The polynomial coefficients are then estimated by least-squares and the partial derivatives are analytical expressions in these coefficients. If orthogonal polynomials defined over a discrete interval are employed, all the computational steps can be replaced by an a priori computed differentiation mask.

A large family of differentiation masks are separable, the weights being obtained from the outer product of two one-dimensional sequences $s(i)$ and $d(j)$, $i,j=-m,\ldots,0,\ldots,m.$ these masks can be written as

$$\mathbf{W}=\mathbf{sd}^{\mathrm{T}} \tag{2.13}$$

The following properties are always satisfied for $i,j=-m,\ldots,0,\ldots,m$

$$\mathbf{s}(i)=\mathbf{s}(-i) \quad \mathbf{s}(0)\geq\mathbf{s}(i) \quad \sum_{i=-m}^{m}\mathbf{s}(i)=1$$

$$d(j)=d(-j) \quad d(0)=0 \quad \sum_{j=-m}^{m}d(j)=0 \tag{2.14}$$

The two sequences are orthogonal since

$$\mathbf{s}^{\mathrm{T}}\mathbf{d}=\sum_{i=-m}^{m}s(i)d(i)=\sum_{i=-m}^{-1}s(i)d(i)+\sum_{i=1}^{m}s(i)d(i)=0 \tag{2.15}$$

Their symmetry properties yield a four-fold symmetry/anti-symmetry for the mask $\mathbf{W}$ defined in (2.13)

$$w(i,j)=w(-i,j)=-w(-i,-j)=-w(i,-j) \quad w(i,0)=0 \quad i,j=-m,\ldots,0,\ldots,m. \tag{2.16}$$

The mask $\mathbf{W}$ performs numerical differentiation along the rows of the data followed by smoothing of the results.

Indeed,

$$output=\mathrm{trace}\left[\mathbf{W}^{\mathrm{T}}\mathbf{A}\right]=\mathrm{trace}\left[\mathbf{ds}^{\mathrm{T}}\mathbf{A}\right]=\mathbf{s}^{\mathrm{T}}\mathbf{Ad}=\mathbf{s}^{\mathrm{T}}\begin{bmatrix}a_{-m}^{T}d\\ \vdots\\ a_{m}^{T}d\end{bmatrix}=\sum_{i=-m}^{m}s_i\left(d^{T}a_i\right) \tag{2.17}$$

where a_i^T are the rows of the data matrix $\mathbf{A}$, thus $\mathbf{W}$ implements $\partial/\partial x$. Differentiation along the columns followed by smoothing, implementing $\partial/\partial y$, is obtained with the mask $\mathbf{W}^{\mathrm{T}}=\mathbf{ds}^{\mathrm{T}}$. This definition corresponds to the usual window coordinates, i.e., the positive $x-$axis points toward the right and the positive $y-$axis downward. Note that the relations between the two differentiation masks and their corresponding vectors is

$$\mathbf{w_1} = \mathrm{vec}(\mathbf{W}) \quad \mathbf{w_2} = \mathrm{vec}(\mathbf{W}^{\mathrm{T}}) \tag{2.18}$$

The Frobenius norm of **W**

$$\|\mathbf{W}\|_F = \left(\mathrm{trace}\left[\mathbf{W}^{\mathrm{T}}\mathbf{W}\right]\right)^{1/2} = \left(\mathrm{trace}\left[\mathbf{d}\mathbf{s}^{\mathrm{T}}\mathbf{s}\mathbf{d}^{\mathrm{T}}\right]\right)^{1/2} = \|\mathbf{s}\| \cdot \|\mathbf{d}\| \tag{2.19}$$

is the product of the vector norms of smoothing and differentiation sequences. The matrix $\mathbf{W}$ having rank one, its Frobenius norm is also equal to the sole nonzero singular value (A.7). Both masks are nilpotent since

$$\mathbf{W}\mathbf{W} = \mathbf{s}\mathbf{d}^{\mathrm{T}}\mathbf{s}\mathbf{d}^{\mathrm{T}} = \left(\mathbf{d}^{\mathrm{T}}\mathbf{s}\right)\mathbf{s}\mathbf{d}^{\mathrm{T}} = \mathbf{O} \tag{2.20}$$

based on (2.13). As expected, the mean value of the data matrix $\mathbf{A}$

$$\bar{a} = \frac{1}{(2m+1)^2} \sum_{i=-m}^{m} \sum_{j=-m}^{m} a_{ij}, \tag{2.21}$$

is discarded when the differentiation masks are applied. The estimated gradient magnitude is

$$\hat{g} = \left(\mathrm{trace}^2\left[\mathbf{W}^{\mathrm{T}}\mathbf{A}\right] + \mathrm{trace}^2\left[\mathbf{W}\mathbf{A}\right]\right)^{1/2} \tag{2.22}$$

and the estimated gradient orientation is

$$\hat{\theta} = \tan^{-1}\left(\frac{\mathrm{trace}\left[\mathbf{W}\mathbf{A}\right]}{\mathrm{trace}\left[\mathbf{W}^{\mathrm{T}}\mathbf{A}\right]}\right). \tag{2.23}$$

After gradient estimation, with the equation (2.22) and (2.23), every pixel in the image is associated with an edge (gradient) magnitude $\hat{g}$ and an edge orientation $\hat{\theta}_e$.

2.5.2.3 Nonmaxima suppression and hysteresis thresholding

Instead of the magnitudes it is more convenient to use their empirical cumulative distribution function. Let $\hat{g}_{[1]} \leq \hat{g}_{[2]} \leq \ldots \leq \hat{g}_{[k]}$ be the ordered set of distinct magnitudes values. Then, for a pixel its edge magnitude $\hat{g}_{[k]}$ is replaced with the probability

$$\rho_k = \mathrm{Prob}\lfloor \hat{g} < \hat{g}_{[k]} \rfloor \tag{2.24}$$

Note that ρ_k is the percentile of the cumulative gradient magnitude distribution. Every pixel is now associated with two values between 0 and 1, ρ and η. The former characterizes the estimated gradient magnitude, the latter the confidence in the presence of an edge pattern oriented according to the estimated gradient orientation. These two numbers define a point in the $\rho\eta-$diagram, illustrated in Figure 2.6 b). And in the context of $\rho\eta-$diagram, nonmaxima suppression and hysteresis thresholding can be defined.

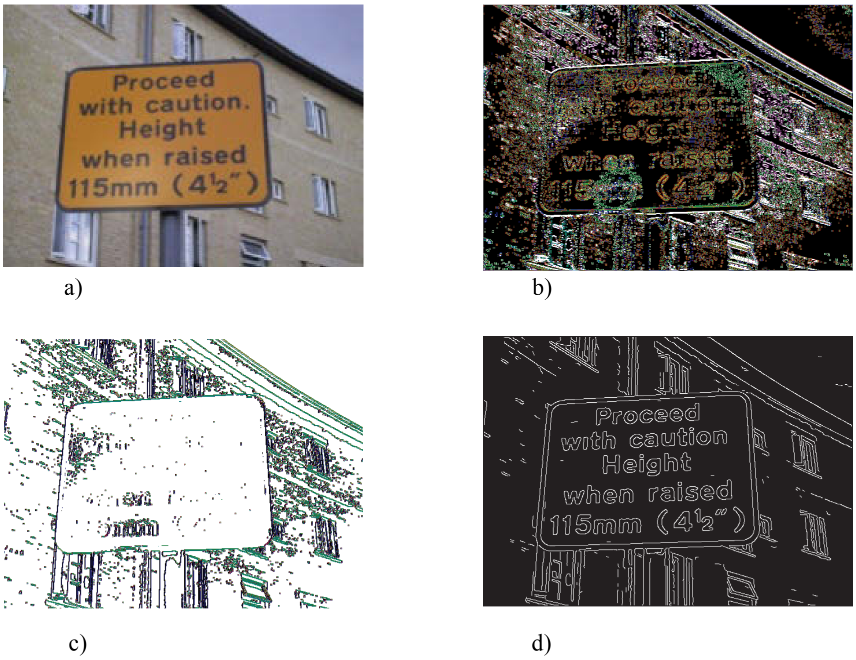


Figure 2.7 a) Original image. b) Traditional (Sobel) edge map. c) Kirsch edge detector; d) Edge map with embedded confidence

Let $f(\rho,\eta) = 0$be the implicit equation of a curve in the $\rho - \eta$plane. With a "point in polygon" algorithm from computational geometry, both nonmaxima suppression and hysteresis thresholding are used to determine if the point is inside or outside of the polygon defined by $f(\rho,\eta)$ and they calculate the coordinates of edge points with the embedded confidence η.

Since this edge detector embedded in confidence fills in most missed corners in the hysteresis thresholding step, for example, the edge map shown in Figure 2.7d), it can help us to obtain the contour or shape to the object and have access to the global information.

2.6 Summary

In term of function, text is congruent with Ergonomics criterion of legibility, readability, and conspicuity to transmit informative message to observers from the moment of construction of its appearance. And in object-centred display, its physical appearance makes observer feel crowding and clutter, meanwhile it's still salient enough to be perceived. Thus, these related concepts show different dimensions of text. Thus, the characteristics of all these aspects coexist in text and provide insights or complete understanding for text detection.

Additionally, edge points are very important to predict crowding or measure clutter or get salient structure, and also very significant to get desirable closed and infinite curves with increasing saliency. Two kinds of reliable edge operators are introduced, including Kirsch operator, and edge detector with embedded confidence. The former provides the edge orientation and edge point direction, and the latter provides the confident locations of edge points. Both of them will be applied to our task to get features in region-levels.

Chapter 3

Global Properties of Text Appearance

Text is corps of letters with in-built spatial coherence which makes text show a natural crowding effect. Meanwhile saliency is maintained by the salient structures originating from its design and construction. The coexistence of crowding, clutter and saliency reveals the different dimensions of text appearance in more global manner, and the characteristics in these dimensions describe the global properties of text figures in image.

For the purpose of breaking down, or at least decreasing the crowding effect to make the target "pop out" in the clutter scene, this chapter deals with the theories, characteristics and computational models of crowding, which also reveals the properties of text appearance in more global manner. And also, the correlates of crowding, saliency and clutter are discussed to help us understand the text figure completely, and figure out the concrete subtasks of text detection from image with clutter scene in computational way.

3.1 Characteristics of Crowding

Levi has summarized a number of features or hallmarks of crowding that are widely considered to be ‘‘true”. Some of these represent the ‘‘facts” that have to be explained by any viable theory of crowding [8, 54].

These "facts" tells us several key points: 1) Critical spacing is proportional to the signal size, there is an association between spacing and target size, just like letter spacing to letter size in type design; 2) Anisotropy means there is a different crowding effect on different orientations, and a stronger effect in horizontal orientations; 3) Different positions have different effects, this means that relative position is another significant factor. 4) Similarity between objects determines the crowding effect. The more similarity there is, the stronger crowding effect there is. 5) Averaging of signals. This connotes the statistical average of features of component stimuli.

3.1.1 Crowding, eccentricity and space density of objects

Crowding depends on the eccentricity of a target object and how densely spaced the surrounding objects are. At a given eccentricity, identification of a crowded target improves as the distance between the target and flankers increases.

In 1970, Bouma[49] has stated 'for complete visual isolation of a letter presented at an eccentricity ofϕ° , it follows that no other letters should be present (roughly) within $0.5\phi^\circ$distance.' This gives rise to the notion of a critical spacing that is proportional to eccentricity. While Bouma's proportionality is constant, b varies across studies, depending on how it is both measured and computed, but it is widely reported to be approximately 0.4–0.5. Also several recent studies have shown that the extent of peripheral crowding is more or less invariant to target size. So if the entire stimulus (including fixation) is scaled, i.e., the critical spacing of crowding scales with eccentricity, performance is unchanged. Outside the fovea, the critical spacing is surprisingly large, typically equal to about 1/3 the eccentricity, but at the fovea, estimates of critical spacing are quite small, 1/10th of a degree or less [67, 137].

The strength and extent of peripheral crowding are much greater than the strength and extent of masking [63, 138], so that in peripheral vision, the suppressive spatial interactions due to nearby flanks are not likely to be a consequence of simple contrast masking. Moreover, the threshold versus contrast function for crowding is quite unlike that for ordinary masking [139].

However, near the limit of resolution [140] in the normal fovea, the extent of ''crowding" is proportional to stimulus size and cannot easily be distinguished from ordinary masking [141]. For instance, with letters, crowding and masking may get confused in the fovea because of the effects of blur, so that what looks like crowding is actually partly masking.

With Gaussian or Gabor targets, the extent of ''crowding" is also proportional to stimulus size [141, 142]and is over a more than 50-fold range of target sizes. Over this large range, foveal ''crowding" is scale-invariant. Moreover, Polat et.al [143] have reported that threshold elevation for orientation discrimination is very similar to the detection of a Gabor patch among Gabor flankers, which has been ascribed to lateral masking.

These studies demonstrate that the critical spacing is proportional to the signal size, keeping the signal at the same eccentricity (zero), and both the strength and extent of foveal ‘‘crowding” can be predicted directly by the strength and extent of masking [141].

Note that, with Vernier targets, which are small and impervious to blur, crowding is qualitatively similar in foveal and peripheral vision [144].

3.1.2 Anisotropy

Crowding in peripheral vision is not isotropic. There is a very substantial radial-tangential anisotropy.

On average, crowding extends from about $0.1\times$ the target eccentricity in the tangential direction to $\approx 0.5\times$the target eccentricity in the radial direction [67]. And there is also a horizontal vertical asymmetry in crowding. In all four quadrants of visual space, crowding is significantly stronger when the target and distractors are horizontally rather than vertically arranged[145].

The extent of crowding is also reported to be field dependent. It is stronger when the distractors and target are within the same visual field than in separate visual fields, despite equal retinal distance [146]. Since flankers have a stronger effect on orientation discrimination (i.e., they reduce percentage correct responses more [85], and the ‘‘resolution of attention” (the minimum spacing at which observers can select individual items) [147] which is coarser in the upper visual field than in the lower field, crowding is stronger in the upper field than in the lower field.

3.1.3 Asymmetry

Peripheral crowding is asymmetric. Bouma has noted that two flankers (one on each side of the target letter) are much more potent than one, and that crowding is stronger with a single flanker at an eccentric locus greater than the target compared to a single flanker at an eccentric locus nearer to the fovea (at the same angular separation from the target) [49]. This inner–outer asymmetry occurs for recognition of letters [148], the identification Gabor patch orientation [149] and face recognition [77].

Motter and Simoni [150] have provided a very simple explanation for this asymmetry in terms of cortical geometry: ‘although the angular separations for near and far flankers are the same in visual space, the far flanker is actually closer to the target than the near flanker after mapping to cortical space”. But we still know of no similar explanation for the large radial–tangential anisotropy.

3.1.4 Crowding depends strongly on target/flanker similarity

Targets and flankers, as stimuli, are usually defined by changes in some properties, and called first-order stimuli and second-orders stimuli respectively. Here, stimuli which are defined by changes in luminance are first-order stimuli, and second-order stimuli are defined via contrast, texture, colour and motion. When targets and flankers are similar, they are likely to be grouped, and when they are dissimilar, they are ungrouped and the target ‘pops out’, i.e., salience. Thus, crowding will be stronger and more extensive when the target and flankers are similar in a number of dimensions. These dimensions include shape and size [151, 152], orientation [63, 142, 153], polarity [151, 154], spatial frequency [155] , depth [151] , colour [151, 156-158], synesthetic colour to some degree [159], motion [160] and order (first- vs. second- order [161].

Temporal grouping and spacing regularity also modulate crowding [162]. Crowding is maximal when targets and flankers are presented nearly simultaneously; presentation of targets before or after the flankers (by≈ 150ms) is sufficient to break crowding [163, 164].

And the strength of crowding depends monotonically on the target: flank contrast ratio [155]. Identical target and flank contrasts result in the strongest crowding from the simple grouping by contrast hypothesis. Importantly, at any target-to-flank spacing, the threshold and saturation contrasts of the flanks affecting the signal are the same [139].

Similarly, when a target and flanker seem to have a regular texture, it is difficult to make judgments about the target and crowding is strong. The regularity of inter-element spacing plays an important role in determining the strength of crowding: regular spacing leads to the perception of a single, coherent, texture-like stimulus, making judgments about the individual elements difficult [52]. For instance, the remarkably similar word crowding effect irrespective of the flanker configurations suggest that word crowding arises as a consequence of the interaction between low-level letter features [47].

Tuning along many of those dimensions would be expected based on low-level considerations and also on the basis of grouping. A set of seemingly “high-level” object-centred crowding effects can arise from “low-level” interactions between the features of letter-like elements[79]. The strong contrast polarity tuning [151, 165] has provided an important piece of evidence used to support both low-level and high-level explanations for crowding. Under certain conditions, increasing the size or number of flanking rings results in a paradoxical decrease in the magnitude of crowding—i.e., the bigger or more numerous the flanks, the smaller the crowding [162].

3.1.5 Statistical properties-average

The widely held notion is that crowded signals undergo a form of compulsory pooling or an averaging of signals.

The earliest reliable report by Parkes et al. [81] suggests the average ensemble orientation. It has elegantly demonstrated that although observers are unable to correctly report the orientation of an individual patch under conditions of crowding, they can reliably report the average ensemble orientation, which suggests that the local orientation signals are combined rather than lost.

This finding has been demonstrated throughout the crowding studies under a variety of different conditions and forms the basis of the faulty integration theory.

3.2 Theories of Crowding

There are many kinds of theories for crowding that range from the low-level receptive field to high-level attention from the different viewpoints comprised of optical physical proposals, neuronal proposals, attention proposals, and computational proposals [8].

These theories suggest several key points: 1) The association between crowding and shift select, and no crowding in large stimuli; 2) The distance over which spatial interaction occurs is related to the size of the receptive fields that are most sensitive to the target. This means that the target size, distance among spatial items have close relations to the size of the receptive field, and the crowding pooling region is related to the receptive field. Also, 3) Crowding is specific to the attentional selection region.

Thus the computational proposals are given, such as averaging position, coherent texture derived from spatial averaging, grouping.

3.2.1 Optical proposals

In foveal vision, the “crowding effect” has a strong relationship with both the eye’s point spread function and stimulus’ physical properties, such as the size and distance from the eyes.

Crowding is a consequence of the ‘‘physics” of the stimulus [166]. When the letters are small and closely spaced, the fovea “crowding effect” appears to be the omission of an interior letter and the merging of two neighbouring letters; neither spatial uncertainty nor split attention can explain this conduct.

However, the foveal “crowding” effect has, at least in part, been ascribed to the effect of the eye’s point spread function [167, 168]. It has also been argued that in foveal vision nearby flanks displace the ‘‘critical spatial frequency band” used to detect the orientation of the gap (horizontal vs. vertical) in a Landolt C, to higher spatial frequencies, thereby reducing the visibility of the cue [165].

Near the limit of visual acuity, the optical explanation suggests that crowding only occurs for small targets and does not occur for large blurred stimuli. Chung and Tjan [169] have found a shift in peak spatial frequency for all letter sizes, but only at the smallest letter separation. Although the shift is tiny, this provides a piece of important evidence that the human visual system shifts its sensitivity toward a higher (object) spatial-frequency channel when identifying letters in the presence of nearby letters.

3.2.2 Neuronal proposals

From the standpoint of the neuron, there are several kinds of explanations of crowding comprised of large receptive fields (spatial –scale shift), perceptive hyper-columns, long-range horizontal connections and contrast masking.

3.2.2.1Large receptive fields

When the target and flank overlap within the same neural unit, for instance when both fall within a single receptive field, crowding occurs. This means that crowding will occur over a range of target sizes, rather than just at the acuity limit, and that the flanking distance will be proportional to the target size.

In 1963, Flom et al [66] suggested that the distance over which spatial interaction occurs is related to the size of the receptive fields that are most sensitive to the target. Since peripheral vision is characterized by reduced visual acuity, larger receptive fields will be engaged, and this ‘‘scale shift” will result in proportionally larger crowding distances. The scale shift hypothesis can predict the fovea crowding effect. Actually, the extent of ‘‘crowding” does indeed depend on target size over a 50-fold range of target sizes [141]. In peripheral vision, with broadband stimuli (e.g. letters), the spatial extent of crowding will scale with the uncrowded acuity. Indeed, for Vernier acuity [68, 144] the spatial extent of crowding appears to scale with the unflanked Vernier acuity in both amblyopic and peripheral vision.

However, scale-shift hypothesis is not reasonable in peripheral vision with stimuli composed of narrow-band features. Actually, with narrow-band stimuli, crowding is largely independent of stimulus size in the periphery, depending only on eccentricity, and peripheral crowding extends over a greater distance even when tested with the same size (and spatial frequency) stimuli as the fovea [138].

3.2.2.2 Perceptive Hyper-columns

Directly starting from the eccentricity dependence (ϕ°) of crowding, in peripheral vision, the extent of crowding for letters is $\approx 0.5 \times \phi^\circ$in the radial direction as mentioned above. And for Vernier acuity, crowding extends approximately $0.1\phi^\circ$ at all eccentricities, about the size of a hyper-column in the primate visual cortex, leading to the suggestion that the extent of crowding corresponds to a fixed spacing on the cortex, and that crowding occurs when competing stimuli fall within the same (or an adjacent) ‘‘perceptive hyper-column” as the target [68].

This notion can predict both the eccentricity dependence and the relative target size independence [74]. Moreover, it does not treat the fovea as a ‘‘special case”. A foveal perceptive hypercolumn is $\approx 4 arc \min$, about the size of a just recognizable letter in the fovea and the distance over which flanks interfere with foveal Vernier acuity [68, 69].

3.2.2.3 Long-range Horizontal Connections

A recent study suggests that the switch from assimilation (crowding) to repulsion (salience, or making different stimuli pop out) [170] depends on cortical distances [171]. The cortical distances refer to the distance (up to 1-2mm) with which long-range horizontal connections between neurons extend in primate area, which translate to approximately $0.1\text{–}0.2\,\phi^\circ$ in peripheral vision [172].

Thus long-range horizontal inhibitory connections have approximately the requisite length to account for the extent of $\approx 0.1\text{–}0.2\,\phi^\circ$ of peripheral crowding (at least for Vernier acuity). However, the fixed cortical distance of long-range connections predicts interactions over a fixed retinal distance, rather than interactions that are related to the target size in the fovea.

3.2.3 Attention proposals

In the visual search task, even when items are easy to resolve *visually,* there are additional spatial constraints that may limit our ability to *select and scrutinize* individual items. Our attention will have the finest scale to operate on the spatial details of an individual item. Intriligator and Cavanagh[147] refer to this as ‘‘attentional resolution”.

Objects spaced more finely than this limit are beyond the limit of attentional resolution and thus cannot be selected individually for further processing based only on their location. He et al.[85] have argued that peripheral crowding results from limitations set by attentional resolution. In Cavanagh & Holcombe’s experiment, in the fixed location condition, both target and flankers appear to flicker in place, and a flickering test letter is flanked and crowded by flickering distractors.

However in the moving attention condition, the target appears to move, while the flankers don’t, therefore the test letter has no distractors along the target’s radial arm. Attention to a single location is revealed. Thus, Cavanagh et al [173] have suggested that crowding is specific to the attentional selection region and does not occur outside it. When attention moves with the guide, crowding is greatly diminished. Accordingly, crowding is specific to the arrangement of distractors within a moving attentional focus—and not set by the arrangement of distractors in retinotopic coordinates.

Notably, crowding is reduced when target and flankers differ in colour [151, 152] or when the target and flankers are the same colour, but the target appears on a different coloured background blob [174]. Po˜der has explained this on the basis that ‘‘exogenously controlled attention is attracted to the location of a salient colour singleton (either a target itself or a coloured blob), and [this] facilitates visual processing in that location”. He argues further that the coloured blob experiment rules out a non-spatial colour based selection.

3.2.4 Computational proposals

3.2.4.1 Abnormal integration at a stage beyond feature detection

Most crowding tasks require that the observer not only detects the features, but also isolates and localises them. One of core tasks of the visual system is to bind the features into a single percept (of object); however, feature binding can fail resulting in the experience of illusory conjunctions of physically disjunct features [175, 176].

Pelli et al.[139]suggest that both illusory conjunctions and crowding may be symptoms of excessive feature integration because small integration fields are absent from the periphery, leading to inappropriate feature integration by large peripheral integration fields. The process of inappropriate feature integration must also somehow suppress the detection of valid features [86], as might occur if the process of feature integration is a competitive one like the association field model [177] which integrates information across neighbouring filters tuned to similar orientations.

In Nandy and Tjan’experiments, the classification images added to the Gaussian noise fields contain sufficient information to reveal the second-order correlation structures of sub-template features, enabling us to infer the shape of the putative features used by the human observer and to compare them to features used by an ideal-observer model. The comparison provides a metric for feature validity. During crowding, they found a decrease in feature validity, consistent with the prediction of inappropriate feature integration and they also found a decrease in the number of valid features, which is not predicted by spurious feature integration.

And in letter identification, the crowding effects results from the inappropriate pooling of target and flanker features and that this integration is more likely to match a response template at a subsequent decision stage with similar rather than dissimilar flankers [51].

3.2.4.2 Loss of position information

In crowded displays, observers frequently mistakenly report a flanker rather than the target. This substitution-like effect of crowding reflects positional uncertainty. However, when required to report all the letters without correct order, the proportion of correct target responses is much higher. Clearly, some information about the target is preserved, but the location information is lost. A number of studies suggest this crowding effect.

Greenwood et al ‘s computational modelling [82] reveals that the perceived position in the presence of flankers follows a weighted average of noisy target- and flanker-line positions, rather than a substitution of flanker-features into the target. Together, the experimental results suggest that crowding is a pre-attentive process that uses averaging to regularise the noisy representation of position in the periphery.

In the presence of nearby distracting clutter-works for complex letter-like stimuli, a set of seemingly “high-level” object-centred crowding effects can arise from “low-level” interactions between the features of letter-like elements [79]. And a model based on the probabilistic weighted averaging of the feature positions within contours accounts for these experimental results.

And early in 2005, Strasburger [87] and Popple [178] have respectively found that the proportion of confusions between neighbouring positions is higher than that predicted by chance performance. Thus, ‘in a large part of the region where crowding occurs, the recognition of a character, irrespective of where in a string it is, is nearly as good as that of a singular character.” And the errors might be attributed directly to a noisy stimulus representation, with both object and positional uncertainty under crowded conditions.

3.2.4.3 Texture perception

Crowding occurrences explain the texture we see when object recognition fails [73]. Some reveal when the features of one or more objects encroach on the receptive field in which the object of interest (e.g. a letter) falls. The object then is difficult or impossible to see, and becomes part of a coherent texture. The implication is that the information kept consists of summary statistics, comprising information about distributions of feature values rather than localized feature maps.

Parkes et al. [81] have found that in peripheral vision, crowding in an orientation discrimination task is distinct from masking. Importantly, the orientation signals are pooled rather than being lost through masking. They have concluded that crowding reflects compulsory averaging of signals (e.g. image features), i.e. they form a texture.

For the natural image, J. Freeman and Simoncelli (2011) have developed a crowding model based on a texture synthesis algorithm so that spatial structure is synthesized within regions whose size scales with eccentricity. To test the model, this scale factor is varied to produce a set of naturalistic stimuli that are progressively ''texturized'' with eccentricity. Wallis et al [76] have examined where crowding occurs in natural images, requiring observers to identify which of four locations contains a patch of "dead leaves'' (synthetic, naturalistic contour structure) which are embedded into natural images. The results suggest that crowding models are based on the spatial averaging of features in the early stages of the visual system.

3.2.4.4 Configural grouping

In peripheral vision, there is a predilection to perceptually group elements and features into a Gestalt. Banks et al. [179] accounts for the inner/outer asymmetry on the basis that multi-element flankers group together separately from the target, reducing the cause of the crowding and thus the asymmetry. Livne and Sagi [180] provide strong evidence for configural effects in peripheral crowding. They arrange Eight Gabor patches to surround a central one in a way that creates several global configurations and allows for the Orientation discrimination and contrast detection of the central Gabor to be measured. These measurements reveal differences in the magnitude of crowding produced by the different configurations. They find that the crowding effect is stronger when random configurations are used and is reduced considerably when a smooth one is used. Further, there are configural effects in the object-level. Specifically, recognition of an upright target face is more strongly impaired when surrounded by a crowd of nearby upright faces, than by a crowd of inverted faces [72].

3.3 Models of crowding

Although many aspects of crowding have been studied, few are computational or make specific quantitative predictions. The large number of different models can be roughly classified into three basic types [54] : (i) masking, (ii) pooling (either of low-level features or by attention). Because of pooling, features of the target and the flankers are integrated, and, thus, feature identification is lost. (iii) substitution, in substitution models. Because of positional uncertainty or limited attentional resolution, features of the target and flankers are mislocalized or not "accessible" by attention. These models are largely descriptive and have been reviewed [8]. There are a few quantitative models of crowding as per the following.

Wilkinson et al. [64] have proposed a model incorporating spatial summation by complex cells and reciprocal inhibition between simple and complex cells. In this model, isolated visual contours are processed by simple cells, which suppress weak complex cell responses. However, in the presence of nearby similarly oriented flanking contours in a small area, complex cells respond vigorously because of spatial pooling, and they then suppress simple cell activity within their receptive field area. Notably, the pooling parameter of this texture model is based on simulations with the best fit to the data rather than on physiology or some other principled approach. Based on the principles of population coding, van den Berg et al [181] present a quantitative and physiologically plausible model for spatial integration of orientation signals.

Dayan and Solomon[182] take a very different approach, in which spatial selection of a target among flankers emerges through a process of Bayesian inference in a computational form and they build upon Parkes et al.'s proposal that ensemble properties are encoded in peripheral vision. Expanding upon the set of statistical features under consideration, Benjamin Balas et al [78] represent peripheral stimuli by the joint statistics of responses of cells sensitive to different positions, phases, orientation, and scale. This "textural" representation by summary-statistics predicts the subjective "jumble" of features often associated with crowding. And this representation can be widened to encompass a much wider range of inputs.

3.4 Breaking crowding

According to many aspects of crowding (as summarised above), under certain circumstances, crowding might be reduced or released completely [54].

- When targets and flankers are similar, they are likely to be grouped, and when they are dissimilar they are ungrouped and the target “pops out”.
- In multi-element flankers, when the flankers are grouped separately from the target, crowding can be reduced. Meanwhile when the target seems to be distinct from the flankers, crowding is weak or absent.
- When flankers are suppressed from visual awareness, crowding is released.
- Since Object-centred crowding effects adhere to all of the diagnostic criteria for crowding and are not due to masking, object-level crowding is released when there are similarity effects or groupings of low-level features.

3.5 Correlates among crowding, clutter and saliency

3.5.1 Saliency – “pop-up” in crowding

In pre-attentive stage, simply primary features such as colour, intensity and orientation, and certain low-level features such as, edges, or salient structure[105, 106] can “pop up” automatically. When an element becomes conspicuous by having a simple distinguishing property, the local saliency of the element occurs.

However, sometime the local elements are not salient in isolation as in simple properties case, instead the arrangement of the elements is what makes the corps of these elements unique and salient, and this saliency is defined as structural saliency, which occurs when the structure is perceived in a more global manner[105].

Moreover, in many case, this saliency is a property of the structure as a whole, i.e. parts of the structure are not salient in isolation. Salient structures appear to play a useful role in segmentation and recognition, since they allow us to immediately concentrate on objects of interest in the image.

Indeed, attention can be tied to objects, object parts, or groups of objects [183, 184]. Rensink has provides evidence for the view that rapid visual search cannot access the primitive elements formed at the earliest stages of visual processing; rather, it can access only higher level, more ecologically relevant structures [185-187]. Account for apparent blindness of observers, Rensink has introduced the notion of proto-objects, which are volatile units of visual information that can be bound into a coherent and stable object when accessed by focused attention.

Notably, Felisbert et al[170] suggest that salience has, at best, modest effects on crowding. When targets and flankers are dissimilar in crowding, they are ungrouped and the target "pop out", i.e., which increase the salience of target.

3.5.2 Crowding in clutter

Clutter focuses on capturing the space organization — grouping by similarity + proximity — in three kinds of viewpoints, including saliency, space averaged attributes and entropy. By saliency, feature congestion is proposed and measured, and by space averaged attributes, edge density and colour density are considered as measurement factors of clutter. The feature congestion measure captures a bit of space organization as well as the sub-band entropy measure.

Although the amount of clutter has some dependency on the edge density, the edge density measurement does not take into account the appearance or organization of the objects. Clutter density explicitly captures the space statistical properties, but still ignores the relationships among the regions.

However, they still can inspire us to explore and formulate an appropriate representation for our particular task, especially the region-level features.

In the cluttered image, spatial distortions could be detected, but sensitivity decreased as edge density is increased in real and random phase images [188]. This demonstrates that spatial discrimination is impaired in the cluttered image and the presence of edges plays an important role in spatial discriminations. Since crowding depends on visual features such as edges, this suggests that crowding also plays a critical role in the peripheral vision in the cluttered image.

As it is well known, crowding undergoes compulsory pooling or averaging, averaging models are able to account for crowding effects across a range of stimuli, from oriented gratings [81] and simple objects[78, 79, 82, 84, 181] to real cluttered images. And crowding can be broken by grouping. Similarly, image grouping processes can minimize the effects of crowding in cluttered images, for instance, crowding can be attenuated when flanks can be grouped together and/or segmented from a central target [180, 189-191]. Moreover, identification of objects containing internal structure is relatively less affected by crowding than for object silhouettes or for letters [192], for example, facial expression can be recognized even though facial features are crowded [193].

Given real image, several researchers [78, 80] have proposed that crowding may be an emergent property of statistical averaging among image features. Balas et al [78] have proposed that the visual system locally represents peripheral stimuli by the joint statistics of responses of cells sensitive to different position, phase, orientation, and scale. This "textural" representation by summary statistics predicts the subjective "jumble" of features often associated with crowding. Within a single pooling region, the difficulty of performing an identification task is correlated with peripheral identification performance under conditions of crowding. And within regions whose size scales with eccentricity, spatial structure can be synthesized by the crowding model based on a texture synthesis algorithm [80, 194].

Additionally, there are good estimates of the statistical distribution of luminance, contrast [195, 196] and edges [188, 197] in cluttered images. Specially, in real cluttered images, since those images are composed of a broad range of spatial and temporal structures, the standard contrast sensitivity function is a poor indicator of sensitivity to structure in cluttered scenes. And the sensitivity to spatial structure depends on the distribution of local edges as well as the local amplitude spectrum. Furthermore, for the purpose of examining where crowding occurs in arbitrary images, reverse correlation is used to analyse and determine local image statistics that correlate with task performance [76].

In this analysis, seven image statistics are concentrated, including luminance, RMS contrast, edge density, orientation, orientation variance, local amplitude spectrum slope, and the maximum moment of phase congruence [198].

Luminance corresponds to pixel intensity; RMS contrast is the variation in pixel intensity over space. Edge density is the space-averaged binary output of the edge detector, where higher values denote more “edge” pixels per unit of area. Orientation variance is the variability in orientation over space, bounded [0-1]. Local amplitude spectrum slope is the log-log local slope of the Fourier amplitude spectrum at every point in the image, where more negative slopes correspond to greater power at low spatial frequencies than higher, indicating that the image is more blurred.

And the experiment result reveals that target size, eccentricity, and local RMS (root mean square) contrast and edge density can be used to make reasonable predictions of the likelihood that an observer will experience crowding.

The above investigation of crowding in the cluttered image shows that the averaging property of crowding is still reasonable, and can be broken or minimized by grouping. Further, local image statistics such as target size, eccentricity, local RMS contrast and edge density, can reasonably predict where crowding occurs.

3.5.3 Four basic psychological principals

As it is well known, considering the context awareness, saliency follows four basic psychological principles [10]. In term of perception, considering the spatial organization or space regularity of in-built spatial elements, crowding and clutter also follows the four basic psychological principles of human visual attention as thus:

i) Local low-level considerations, including factors such as local contrast, orientation, and colour.

ii) Global considerations, which suppress frequently occurring features, while maintaining features that deviate from the norm. Or it will pop out proto-objects resulted from salient structures.

iii) Visual organization rules, which state that visual forms may possess one or several centres of gravity about which the form is organized. This is associated with space regularity in spatial elements.

iv) High-level factors, such as human faces.

Being the coexistence of crowding, clutter and saliency in text, these psychological principals are reasonable for text understanding and detection.

In global consideration, for text in a complex background, crowding and saliency are just like two sides of one coin. For crowding, it reveals that there is too much similarity among target and non-targets, statistic property of space averaging makes target be difficult to pop up; for saliency, it emphasis on salient structure or proto-objects derived.

For clutter, there are too many items in limited space in term of information density. However, all of them have close relations with space organization— grouping by similarity + proximity in objected-centred level, and statistic property of averaging is still effective over the groups. And saliency is required to pop out the target or to measure clutter.

Van's study[128] suggests that crowding places a limit on visual search performance in cluttered environments. But salience has, at best, modest effects on crowding[170], when targets and flankers are dissimilar in crowding, they are ungrouped and the target "pops out", i.e., which increases the salience of the target. And accordingly, clutter can be measured based on saliency, such as feature congestion.

In low local-level considerations, such as local contrast, orientation, they usually reflect the informative locations. Moreover, in the clutter scene, local image statistics such as target size, eccentricity, local RMS contrast and edge density, can reasonably predict where crowding occurs.

Similarly, the saliency map need to be fed a number features, including local contrasts of colour, orientation, texture and shape features, oriented sub-band decomposition based energy, ordinal signatures of edge, colour orientation histograms, Kullback-Leibler (KL)divergence between histograms of filter responses, local regression kernel based self-resemblance, and earth mover's distance (EMD) between the weighted histograms, salient structure, and region-based features.

In visual organization, reasonable regions and features over them are significant for calculation of saliency, crowding effect and clutter. Through visual or space organization, salient structures are obtained, space regularity--the cause of crowding effect---is generated, and visual clusters are formed. Consequently, attention pooling regions are revealed. Clutter associates the region with "interested area", and crowding is specific to the attentional selection region which is indeed "interested area" too. Moreover, the pooling region of crowding has tight relations to the size of the receptive field, critical spacing and target size. Accordingly, space organization— grouping by similarity + proximity—plays a critical role in these sorts of region's formation and generation.

In high level, distinctive features of text contribute to saliency through low local-level properties and local textual organization. Crowding depends on similarity among feature vector of target and that of distractors; clutter is measured by the distance between feature vector of target and that of distractor, the farther, the more salient and easier to search and less clutter. Thus, features and the similarity or dissimilarity over them are significant for crowding, saliency and clutter. The high-level factors or properties of individual characters and textual organization over them should be calculated.

As text is a corps of letters in space regularity, in which crowding and saliency coexists. The correlates among crowding clutter and saliency inspires us to computationally dig the distinctive features of text which contribute to saliency through low local-level properties and local textual organization, and also quantify the space organization to obtain the pooling region over which region-level features about clutter and crowding are captured.

3.6 Summary

The investigation of crowding theory and characteristics brings up several key points to the global property of text appearance: 1) The association between crowding and shift select, and no crowding in large stimuli; 2) The distance over which spatial interaction occurs is related to the size of the receptive fields that are most sensitive to the target. This notes that, target size, distance among spatial items have close relations to the size of receptive field, and the crowding pooling region is related to the receptive field. And also, 3) Crowding is specific to the attentional selection region.

Meanwhile, the characteristics of crowding tell us: 1) critical spacing is proportional to the signal size, there is an association between spacing and target size, just like letter spacing to letter size in text in type design; 2) Anisotropy means a different effect on different orientation discrimination; 3) different positions have different effects, this notes that relative position is another significant factor. 4) Similarity between objects. The more similar, the stronger the crowding effect is. 5) Averaging of signals.

Additionally, both crowding and clutter increase with information density, crowding can model the clutter based on the analysis of spatial layout, and the measure of clutter can be used to compute the crowding [129]. In the clutter scene, local image statistics such as target size, eccentricity, local RMS contrast and edge density, can reasonably predict where crowding occurs. And also, increasing saliency by curve formed by edge points is required to break down the crowding or clutter. Thus, for text detection in the clutter scene, both the local image statistics and text discriminative features need to be calculated to make text salient or pop out.

The correlation among crowding, clutter and saliency, brings up to two necessary subtasks: 1) for the purpose of the calculation of low-level properties, computationally track the discriminative features for text legibility, readability and conspicuity; 2) Guided by the characteristics and theories of crowding, interested regions or pooling regions need to be generated or formed to break down, or at least decrease crowding to make our target pop out, i.e., to represent image based on the spatial layout analysis of the space organization.

Chapter 4

Properties of Individual Character

The correlates among crowding, clutter and saliency, bring us one subtask to computationally track the discriminative features of individual character for text legibility, readability and conspicuity.

Firstly, individual characters must be distinctive, yet related, in their form and construction. Within a font, type designers constrain the shape of individual letters so that they are related in terms of the stylistic attributes of letters. And letters of the same font use a similar systematic reference frame type to create a family of objects for identification. The similar reference connotes that the letters have shared properties (commonalities) in addition to distinctiveness.

Those shared properties are significant not only for an independent individual character to be identified, but also for calculations of the statistical features of words or texts with word superiority effect. In addition, it also affects the strength of crowding.

Accordingly, this chapter deals with the properties of independent individual characters for text legibility, and introduces the relations between the legibility and attributes of letter form, and the view distance and luminance contrast [4, 199, 200] from the standpoint of the design and construction of text.

Most investigations of text legibility from the literature have used simplified stimuli and these have been presented on otherwise featureless backgrounds, a situation quite unlike the typical real world. While these studies have provided important insights into the essential process of crowding, the extent to which this understanding holds true in natural vision is less clear. Given that natural images are cluttered, this implies that most of the time, we are unable to restore the viewing distance or luminance to that of the viewpoint of design. Thus, in relation to the public benchmark dataset ICDAR 2003 for text detection in a natural scene, another aim of this chapter is to computationally explore character properties and their relations to text legibility.

4.1 Aspects of text in typography

With regard to legibility, readability and conspicuity, the aspects of typography determine the usability of text, including functional properties, semantic properties and textual organization.

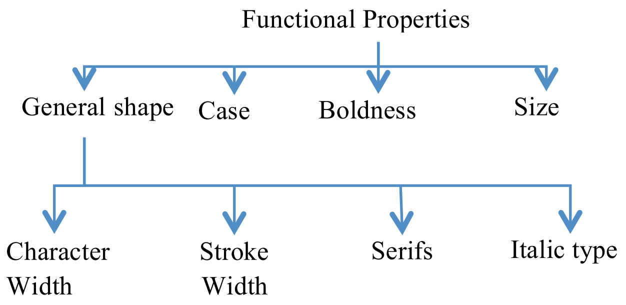


Figure 4.1 Functional Properties of text

With regard to the shape of a type, two kinds of properties should be distinguished [201]: functional and semantic properties. The functional properties of type allow the characters to be identified as a letter, illustrated in Figure 4.1, including general shape, case, boldness, and size. These terms will each be defined and discussed separately in the following sections. Of particular importance, the general shape involves character width, stroke width, serifs and Italic type, and these attributes attract significant research interest in terms of the computation work about text.

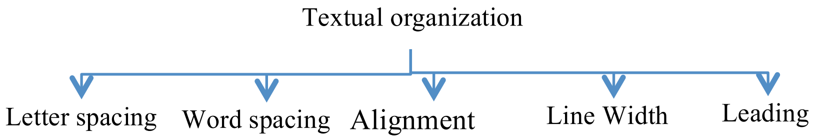


Figure 4.2 Textual organization of type of text

Textual organization indicates the spatial arrangement of words, text lines, and graphic illustrations (such as photographs) on the printing surface, which means the use of space and the composition of letters, in the way that letters are arranged to form text, illustrated in Figure 4.2.

All these spatial features of text consist of letter spacing, word spacing, alignment, line width, and leading. All of them contribute to the crowding effect which exists in letters [49] and this has significant implications for our work. Accordingly, we will deal with them in chapter 5.

The semantic properties of a letter or word trigger a cognitive or emotional response in the reader. These properties include the meanings attributed by aesthetics and the meaning attributed by association. For instance, the shape reveals, or rather suggests some sort of meaning because it is associated with a certain intellectual or emotional value. This is an important factor in the design of signs and advertisements. However, it is more of a craft than an objective selection process and therefore will not be discussed in this thesis.

4.2 The anatomy of type of text

The actual form of a letter depends on the type face, or font. The anatomy of type is demonstrated clearly in Figure 4.3. All the properties are named in the figure and contribute to the character's typeface, which are highly related to both legibility and readability. They are composed of the ascender, descender, bowl, counter, capital height, serif, stem, and X-height. The capital height and x-height usually are used to label type or text size, and the stem contributes to the line model of letter perception since it is a straight vertical stroke or the main straight diagonal stroke. However, from the viewpoint of computation, the truly distinctive features of type form are important to look for in order to make the character 'a'–sound when seeing the 'a'-form.

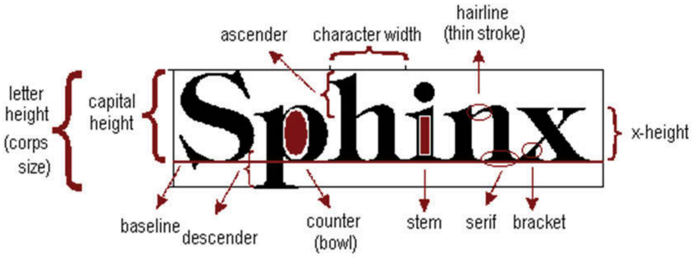


Figure 4.3 The Anatomy of type: after[202-204]

While the functional properties refer to the effect of these properties on legibility, some properties also relate to conspicuity. Owing to this, their effect on this criterion will also be discussed here. Not all of the separate features appearing in **Figure 4.3** have been investigated, only those properties which have been studied in isolation will be discussed.

An investigation will also be conducted in terms of the conventions and assumptions that exist in practical design and computation in science. Where possible, a comparison will be made of these views and the results of research. This overview will proceed from aspects of general shape, through line, stroke, case, weight and size. As well as this investigation, we will also explore them in images with the clutter scene in the public text detection benchmark dataset.

4.3 General shape

"The simplest forms (shapes) . . . preserve the characteristic structure, distinctiveness, and proportions of each individual letter" [205]. Letter form is therefore critical for optimal legibility and recognition.

4.3.1 Simple shapes

Firstly, the problem of shapes for letter-identity has received a large degree of attention from typographers, ergonomist and computer scientists.

When designing a set of characters, two issues must be addressed. First, the design of a character must match the reader's expectation as to how that character should appear. Second, the characters should be designed so that they can be easily discriminated from one another. Further, it is evident that a letter needs to be recognizable to convey meaning to a reader. An 'a' should therefore be identifiable otherwise it does not truly belong to the 'a'-category. It should not only be clear that it is a letter, but it also should be readily identifiable from other letter forms based on its feature map and letter map.

Following the lines of model of object recognition, based on crucial data from brain-damaged subjects, three levels of representation on prior to lexical access have been proposed [206], as illustrated in Figure 4.4: the first level of representation consists of a retinocentric, feature map; the second level consists of a stimulus-centred, letter-shape map; and the third level consists of a word- or object-centred, grapheme description.

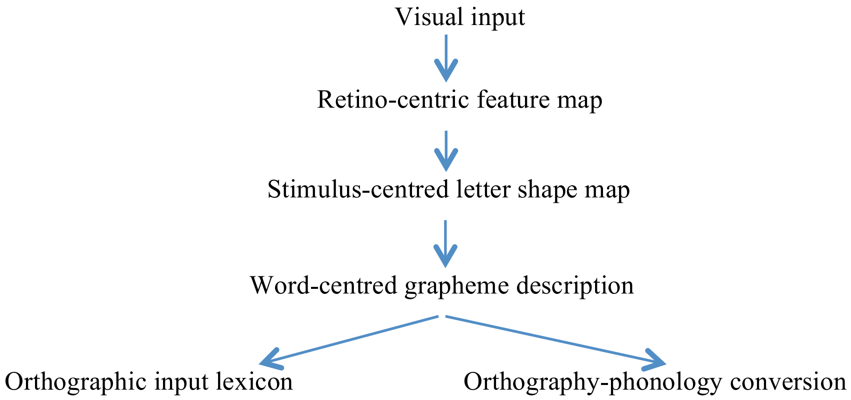

Visual input
Retino-centric feature map
Stimulus-centred letter shape map
Word-centred grapheme description
Orthographic input lexicon
Orthography-phonology conversion

Figure 4.4 Schematic representation of a model of early stages of the word-recognition process: after [206]

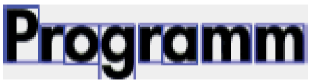


Figure 4.5 "Bouma" shape

For object identification, shape effects occur earlier than function effects[207].On the basis of colour and shape rather than luminance, infants more readily identify objects [208]. In addition, when focusing on identification of words or a string of letters, the word shape in Figure 4.5, i.e. "bouma shape" is an important variable [209]. However, identification of words written in Chinese characters doesn't rely on word-shape cues since Chinese characters are different from smoothing scripts (Alphabetic letters) [210].

Besides its importance in terms of the shape in individual letter identification, its feature map [211, 212] usually plays a determinative role and contains properties which share all the other letters in the same family. For the feature map, we need to dig computationally into the different dimensions of the functional properties of character in the following sections.

4.3.2 Representation of shape

As is well known, a representation of a shape consists of four independent components. A representation of a shape consists of four independent components [213, 214]: a set of primitives, a reference frame, a vocabulary of relations for within the reference frame, and the binding of elements to one another and to their locations and/or relations.

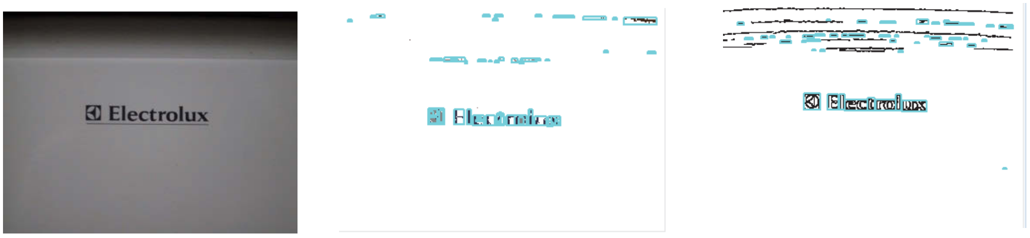
A set of primitives refers to a vocabulary of basic elements that can be put together to describe a shape, for instance, pixels in the case of a raw image, simple image features such as lines and vertices [215, 216], more complex features such as volumetric parts[217] , and approximations of volumetric parts [218].

A reference frame refers to a coordinate system that serves as the basis for specifying the arrangement of an object's features or parts, which means location and orientation.

A vocabulary of relations specifies how an object's features or parts are arranged within the reference frame. The most direct approach is simply coordinate-based coding the distance to representing the relations among computational models/parts of object recognition. An alternative to direct coordinate-based coding is to represent an object's features or parts in terms of their relations to one another. The resulting representation is referred to as a "structural description." Representing relations explicitly affords tremendous flexibility in the vocabulary and form of the relations expressed. In addition to expressing relative location, elements can be represented in terms of their relative size, orientation, etc [219].

The binding of elements to one another and to their locations and/or relations [218], is closely related to the issue of relations, but they are importantly quite different: The latter refers to the vocabulary of relations used to express the configuration of an object's features or parts; the former refers to the manner in which elements or properties are conjoined with one another and with their locations and/or relations.

For shape of text, we have two methods to capture it. One is by connected component pixels in the raw image, which forms the “inked area” of a letter, shown in Figure4.6 b). And the other is by contour in the confident edge map, shown in Figure 4.6 c), which connects the component edge point of contour. Based on them, locations are bounded and the structure or space relations can be quantified. The detail will be dealt with in the following chapters.



a) b) c)

Figure 4.6 a) input image; b) Shape is formed by "inked area"; c) Shape is obtained by contour.

4.4 Width and relative dimension of character

The term *character width* refers to the distance between the most leftward part and the most rightward part of the letter, excluding any attached space, shown in Figure 4.7. Usually, character width is expressed in its relation to character height since character height indicates the type size of a letter.

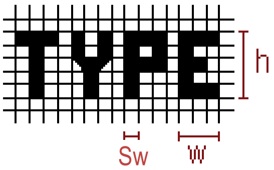


Figure 4.7 Width and relative dimension of character

Expressing width of a character in points is just for the purposes of determining the reader's perception of how large a letter is. Consequently, the most useful measurement is the height and height-to width *ratio*.

4.4.1 Height

Character height has the greatest impact on the distance at which a sign can be read and is the most obvious characteristic to be changed to improve large format legibility [220]. The best way to achieve visible displays for older drivers may be to increase letter height rather than increase luminance levels [221], although instrument panel size constraints may then become a factor.

Table 4.1 letter height (in centimeters) for various stroke widths-to height ratios at various distances

Height Dis Sw:H	Distance				
	70cm	3m	6m	30.5m	305m
1:6	0.25 cm	1.06 cm	2.12 cm	10.60 cm	106.0 cm
1 : 8	0.33 cm	1.41 cm	2.83 cm	14.15 cm	141.5 cm
1 : 10	0.41 cm	1.77 cm	3.54 cm	17.68 cm	176.8

When space limitations are a consideration, letters should be made as large as possible up to the point of very nearly filling the available space (margin less than the stroke width of the letters), in order to permit discrimination at a maximum distance. In this case, unlike any other variable, increased character height improves legibility at a distance; character height is limited only by the size of the sign. But still, there is the recommended letter height at various distances in Table 4.1.

Since different fonts have different character heights, fonts become another factor in relation to character height. In a laboratory setting, Bank Gothic Light and Dutch Regular are the most legible and readable fonts in large format display [222]. For instance, if contrast and lighting are equal, Commercial Script Regular is only legible when it is 4 times the size of Bank Gothic Light and Dutch Regular, therefore requiring a far larger sign so that it may be read at the same distance.

Based on experimental work on letter height for the legibility of text on display, there are several practical calculations in the literature for application design which consider the various factors, such as viewing distance [223-226], height to stroke width ratio [225], luminance [226, 227] and contrast [227].

However, in an image with the clutter scene, most of the time the luminance and view distance are unable to restore and we don't know the original design. Further, they must be locally matched to the image. Thus, local luminance and contrast will be important factors. Also in terms of image, intensity can be used to represent luminance, and local RMS contrast can serve to describe the local contrast. We will explore the relations among height, width to height ratio, intensity and local RMS.

Although the height depends on the size of sign in design, the height Pareto curve as a visual object in image suggests that there might be some limits set on it, as illustrated in Figure 4.8.

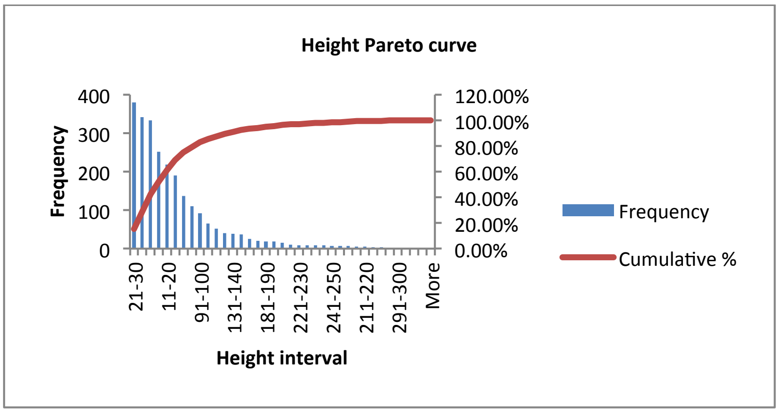


Figure 4.8 Height Pareto

Inspecting the figure, nearly 80% (79.27%) of the characters have their height in the range of [11, 90] pixels, and about 13% of the characters have their height in the range of [91,150] pixels, and 5% of the characters have their height in the range of [151,300] pixels, and about 3% of the characters have their height in the range of [301, 1000] pixels, and only about 0.2% of the characters have their height at less than 10 pixels.

Thus, paralleling the recommended height in Table 4.1, the height of the letter in digital images in the clutter scene still has some distribution, and can be divided into five ranges: [1, 10] pixels, [11, 90] pixels, [91, 150] pixels, [151,300] pixels and [301, 1000] pixels. All of these correspond to small fonts, common size fonts, medium size big fonts, big fonts and large fonts respectively.

4.4.2 Height and mean intensity, local contrast

4.4.2.1 The effects of mean intensity and local contrast

In the clutter scene, local contrast has an effect on the legibility of text embedded within the scene. We need to explore the relations among letter height, mean intensity and local contrast.

Without loss of generality, a region $Q_{M_Q \times N_Q}$ given by the tight rectangular boundary box of a letter, i.e. "Bouma" shape, over which the local RMS contrast can be computed by mean intensity $\bar{I}_Q$ of the local region and its standard differences σ_Q as follows:

$$RMS_Q = \frac{\sigma_Q}{\bar{I}_Q}, \quad \bar{I}_Q = \frac{1}{M_Q N_Q} \sum_{i=0}^{M_Q - 1} \sum_{j=0}^{N_{Q-1}} I_{i,j}, \quad \sigma_Q = \sqrt{\frac{1}{M_Q N_Q} \sum_{i=0}^{M_Q - 1} \sum_{j=0}^{N_Q - 1} \left(I_{i,j} - \bar{I}_Q\right)^2} \tag{4.1}$$

The relation between height and mean intensity is demonstrated in Figure 4.9. Inspecting especially Figure 4.9 c), the trend line is polynomial curve at the small slope with a linear intercept of 118.01; this suggests that mean intensity seems to have a very slight interaction with height. For the large letters, which are big enough to be individual salient visual objects, the relationship between their heights and mean intensity, as shown in Figure 9 a), suggests that mean intensity seems to have a bit more interaction with height. But when the heights become smaller, i.e. for big characters, the interaction is minimal. Further, for the various height characters, the mean intensity seems to have very little effect on them, oscillating as it does around 118. This finding leads us to explore the mean intensity distribution in the following section.

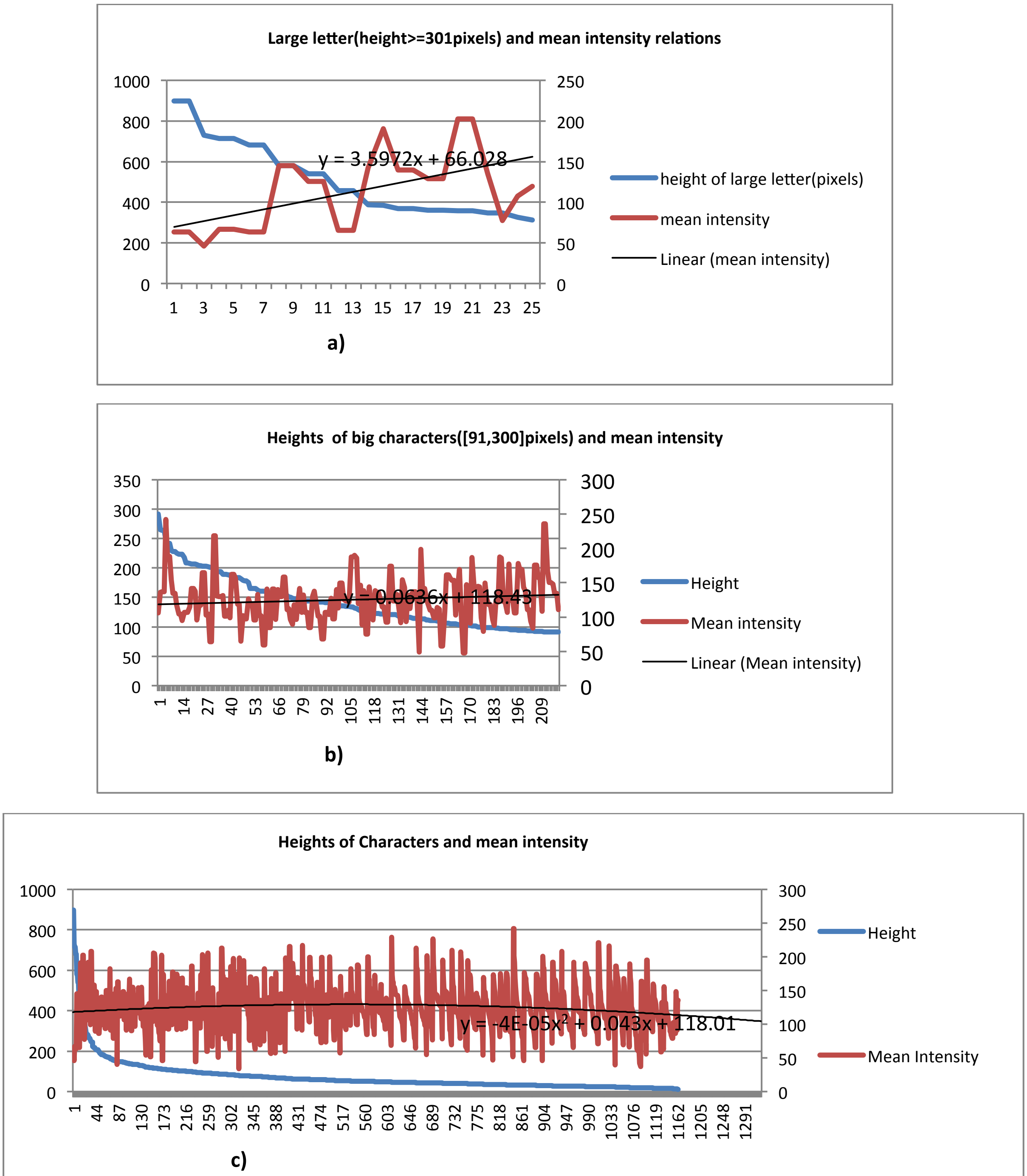


Figure 4.9 Relations between height and mean intensity

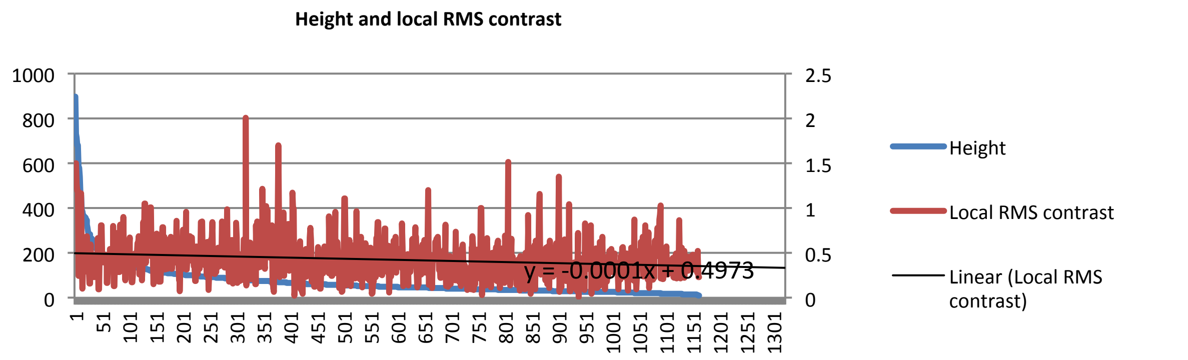


Figure 4.10 Relations between height and local RMS contrast

The relations between height and local RMS contrast are illustrated in Figure 4.10. Inspecting it, these are similar to those of mean intensity e.g. the trend line of local RMS contrast is a linear line on a very small slope with the linear intercept of 0.4973 as the height becomes smaller. This suggests that the local contrast does not seem to have an interaction with the height.

Even though both mean intensity and local RMS contrast do not seem to have an interaction with height, they have a similar shape in terms of trends lines as the height becomes small. We must explore the relations between mean intensity and local RMS contrast.

4.4.2.2 Mean intensity and local RMS contrast

Firstly, mean intensity has its own distribution, which is revealed in its histogram and Pareto curve, as shown in Figure 4.11. Inspecting it, about 80% of the characters have their mean intensity in the range of [80,170], and 10% of the characters have their intensity in the range of [61-80] and [171,180], and another 10% of the characters have their mean intensity in the range of [181,255] and [31,60]. Thus, the mean intensity can be divided into three intervals, such as [1, 80], [81,170], [171,255] which corresponds to nearly 10%, 80% and 10% of the characters.

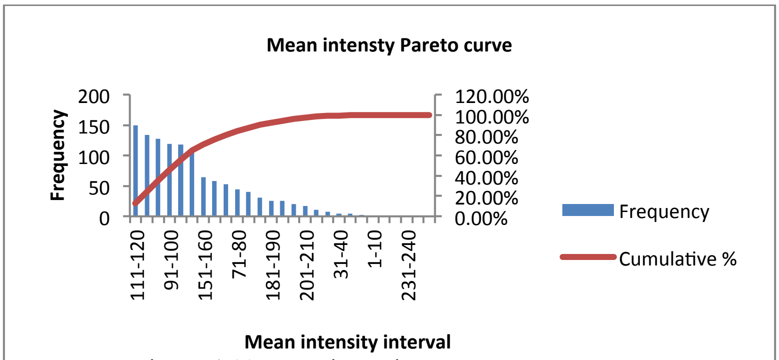
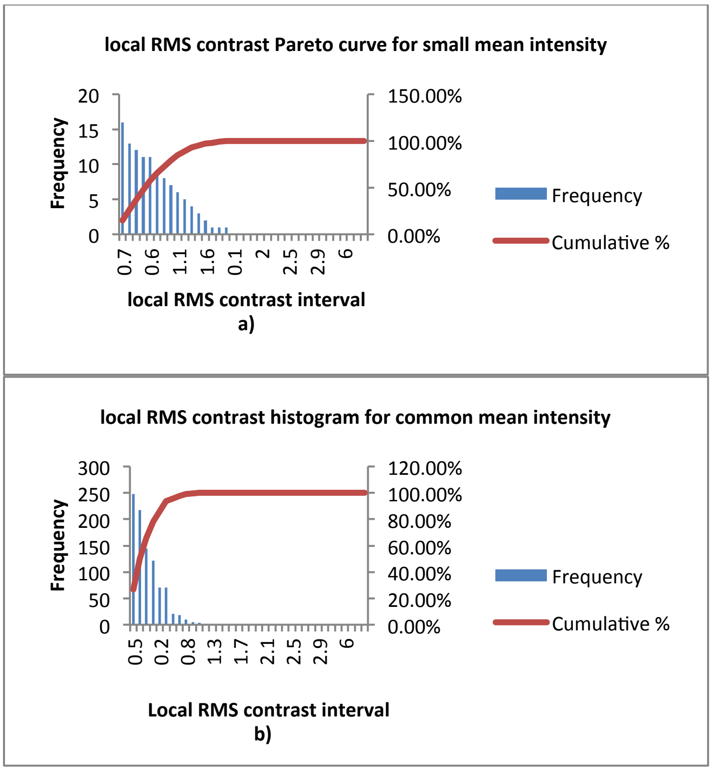
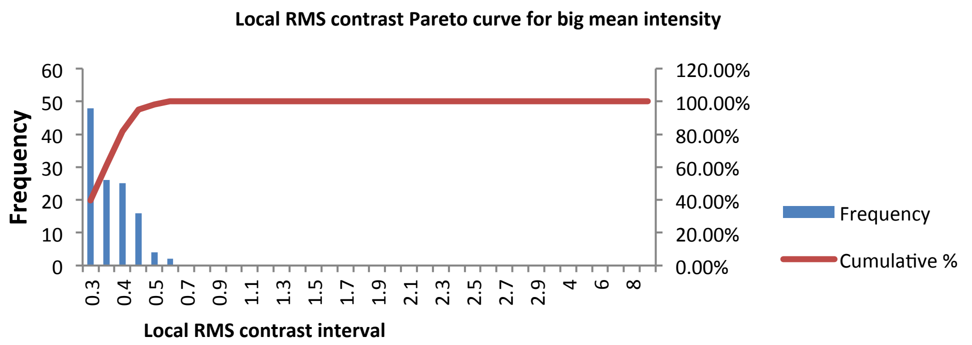


Figure 4.11 Mean intensity Pareto curve

When mean intensity is in the range of [1, 80] as illustrated in Figure 4.12 a), 85% of the local RMS contrast are in the range of (0.3, 1.2), 90% of them are in the range of (0.2, 1.2], nearly 10% of them are in the range of (1.2, 1.9], and nearly 1% are in the range of (0.1, 0.2].

When mean intensity is in the range of (80,170] as is manifest in Figure 4.12 b) nearly 80% of local RMS contrast are in the range of (0.2, 0.6], and about 10% of local contrast are in the range of (0.6, 0.9], that is, about 90% of local RMS contrast are in the range of (0.2, 0.9]. Further, nearly 10% of local RMS contrast is less than 0.2, and only about 1% of the local RMS contrast is bigger than 0.9.and less than 1.3.





c)

Figure 4.12 Local RMS contrast distribution at different mean intensity

When mean intensity is in the range of (170,255] as is displayed in Figure 4.12 c) all of the RMS contrast are less than 0.6, about 80% of them are in the range of (0.1, 0.4], about 13% of them are less than 0.1, and about 5% of them are larger than 0.4. Considering this analysis, it suggests that the bigger the mean intensity, the smaller the local contrast.

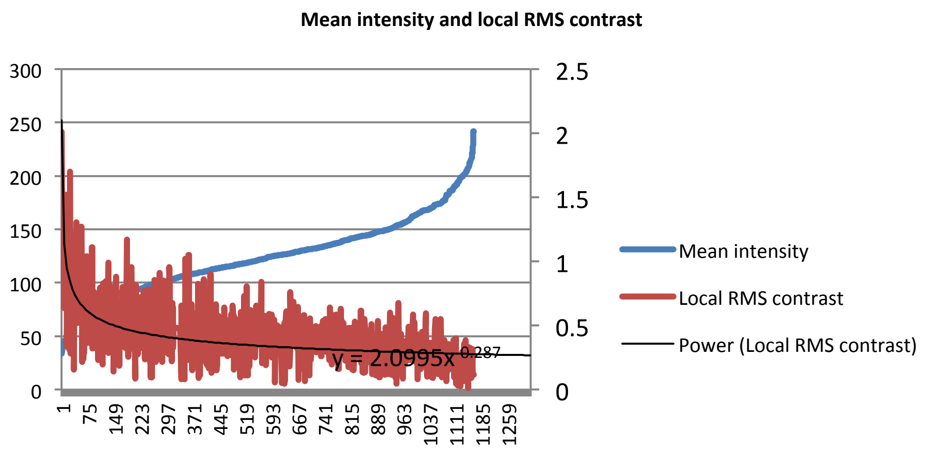


Figure 4.13 the inverse relations between mean intensity and local RMS contrast

Further, there is an inverse relationship between the mean intensity and the local RMS for various heights of letters, as is illustrated in Figure 4.13 and Figure 4.14. Inspecting Figure 4.13, when mean intensity is in the relatively small range, the trends line of the local RMS contrast decreases quickly, and then goes down slowly as the mean intensity goes up. In addition, the trends line of local RMS can be roughly modelled as one power function line when the mean intensity goes up. According to the trends lines of them, when mean intensity is small, there is an intersection between the local RMS contrast and mean intensity, and the intersection occurs when the mean intensity is around less than 80. Moreover, we can make the inverse relationship clear by the large characters, as shown in Figure 4.14.

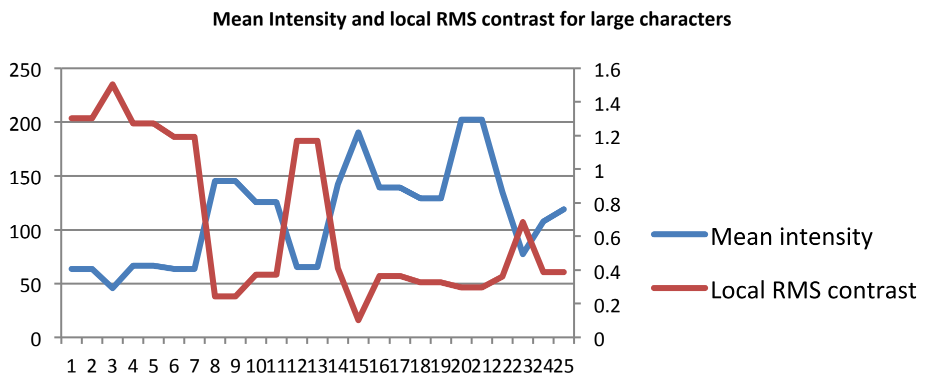


Figure 4.14 the inverse relations between mean intensity and local RMS contrast for large characters

4.4.3 Height-to-width ratio

Height-to-width ratio has mainly been studied for applications other than books. After all, the commonly used ratio in many book-typefaces seems to meet the criterion of legibility quite well.

For most applications, the width to height ratio is recommended at 3:5, that is, five is the maximum number of elements in the height of letters and three is the maximum number of elements in width. However, width to height ratio needs to be decided as per the particular situation. In aircraft cockpit displays, for all luminance levels, l a $H:W$ ratio of $1:1$ in general seems to be optimum [3], and it can improve the legibility distance [220], as is illustrated in Figure 4.15. For instance, translucent letters require a width to height ratio of 1:1[228] and for uniform stroke width capital letters, a marked loss in legibility is found when the letter width is narrower than $2/3$ of the height.

Additionally, Wourms et al. [229] have recommended that the width-to-height ratio ranges from 3:5 to 4:5. According to the U.S. ADAAG (Americans with Disabilities Act Accessibility Guidelines for Buildings and Facilities), the width-to-height ratio is recommended to be from 5:7 to 1:1.

Beyond that, with regard to dot matrix characters, 5×7 dot matrix characters are identified as being “acceptable”, 7×9 are recommended, and 9×13 are considered to offer improved performance [230]. For easy reading, the character matrix should be at least 7×9 and with 9×11 preferred [231]. And for VMS, 5×7 character matrix fonts are recommended to be employed [220], since a width-to-height ratio 5:7 matrix is slightly more legible than a 4:7 matrix [232]. Additionally, optimal legibility can be attained at the height of 9 pixels [233].

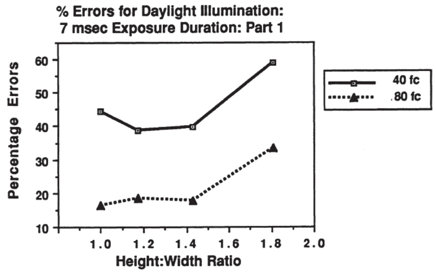


Figure 4.15 After Brown 1953[3]

Table 4.2 Descriptive statistics

Mean	1.448393009
Standard Error	0.01842031
Median	1.259095
Mode	1
Standard Deviation	0.876466129
Confidence Level(95.0%)	0.036122463

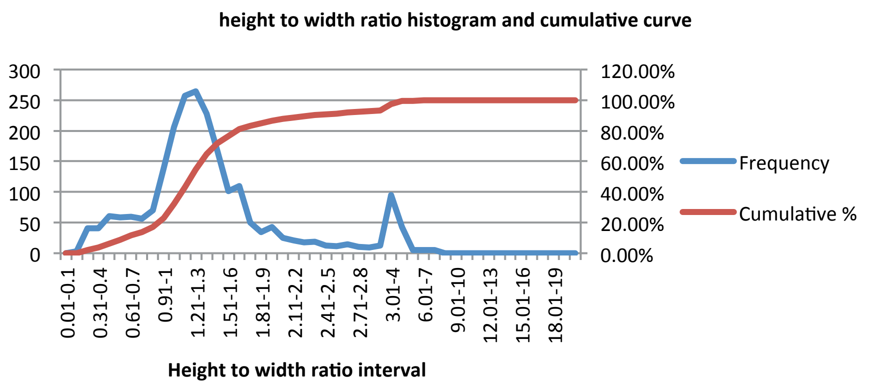


Figure 4.16 Height to width ratio histogram

For text in the document image with the clutter scene, the height-to-width distribution is shown in the histogram in Figure 4.16 and the Pareto curve in Figure 4.17. Further, its descriptive table is shown in Table 4.2. Inspecting them, the mean is 1.4483, the median is 1.225, and the mode is 1 which means that the height to width ratio of 1:1 occurs the most frequently, and this finding agrees with the recommended optimal ratio for legibility in the original design. Inspecting Figure 17, 80% of the height to width ratios are in the range of (0.4, 1.8], with 10% of them being in (1.8, 3], around 7% of them being in (3, 8], and approximately another 3% being under the 0.4.

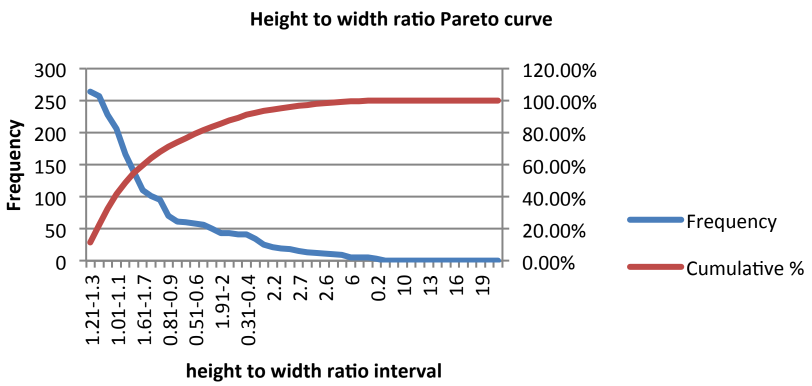


Figure 4.17 Height to width ratio Histogram and Pareto curve

4.5 Stroke width and contrast

Stroke width is the thickness of the stroke of a letter. It is usually expressed in terms of its relation to character height; the smaller the stroke width-to-height ratio, the skinnier the letters appear. Early in 1941, Uhlaner et al examined the stroke width of three-inch block letters (height equals width) as a factor in the legibility of highway signs. This study indicates that the optimal stroke width is 18% of the letter height.

Since the stroke width to height ratio captures the intrinsic construction of the letters, we can assume that it will achieve the maximum possibility when taking pictures, thus we can directly apply them as prior knowledge to roughly filter the non-text region in our work.

4.5.1 Stroke width-to-height ratio

As described in Buckler’s study, the stroke width-to-height ratio can vary from 1:6 to 1:10 with no significant loss in legibility and the stroke width-to-height range is best achieved from 1:6 to 1:8 [229].Typically, the stroke width-to-height ratio of 1:6 can be found in many display situations.

For the phenomenon of radiation, or sparkle in white characters, different luminance has a different effect on the white-on-black text and black-on-white text [234]. The optimal legibility for black characters is obtained for height-to-stroke width ratios at greater than 6:1, and the optimal legibility for white characters is obtained for height-to-stroke width ratios at greater than 12:1. After averaging many legibility factors, it is found that legibility is best for white-on-black characters except under conditions of high luminance [199]. Considering typeface, setting text in Times or Univers is more likely to result in consistent legibility rather than setting it in Baskerville, Rockwell [235].

In simple and complex systems, for black letters on a white background, under good illumination, the optimal stroke width-to-height ratio is from 1:6 to 1:8 (0.167 to 0.125); and for white letters on a black background, it is from 1:8 to 1:10(0.125 to 0.1). With reduced illumination, a lower ratio (higher proportion) is required to maintain the same level of legibility; bold type with a ratio of 1:5 (0.2) is suggested for low levels of illumination [228]. When the letters are transilluminated, the ratio can be set at 1:12 to 1:20 (0.083 to 0.05) [57].

With the use of a computer screen, Tahoma has a ratio of stroke width to height in the range from 1:5 to 1:8, and letter width to height near the recommended 3:5, that makes it highly legible [57, 226, 236] in the computer screen, especially for Power Point design [237].

4.5.2 Background colour, luminance contrast

When discussing the optimal stroke width-to-height ratio, the effect of background contrast and illumination should be considered. Background contrast is created by variation of colour, luminance and background material.

At background luminance levels greater than 3.77 $\mathrm{cd/m^2}$(1.1fL), red and green legibility data compare well. In terms of legibility, blue and green backgrounds also perform almost equally as well [238, 239]. In addition, white, yellow, and orange backgrounds produce similar legibility results, and have maximum legibility for luminance in the range of 3.4 to 34 $\mathrm{cd/m^2}$ (1 to 9.9fL). It appears that colour luminance contrast effects become inoperative at levels between 0.3 and 0.33 $\mathrm{cd/m^2}$.

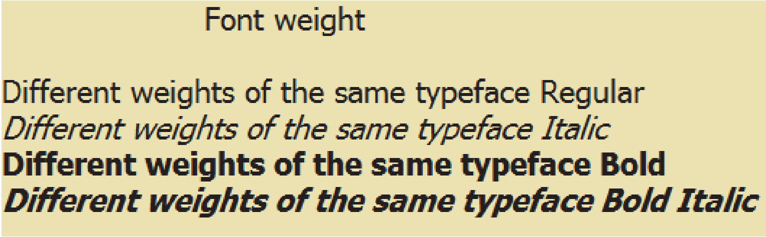
For signs with light (white, orange or yellow) backgrounds and black legends, the recommended optimal figure-to-background contrast is 12:1, the minimum luminance is 2.4 $\mathrm{cd/m^2}$ (0.7fL). At night, legend luminance contrast is the most important variable in sign legibility, and the maximum legibility is achieved at a contrast of 30 to 60:1. And the recommended minimum luminance applied to white legends with dark (green, blue, red, or brown) backgrounds is 0.4 cd/m^2 (0.12fL). And 30 cd/m2 is suggested for night time luminance and 1000 cd/m2 for bright daytime viewing [220]. Further, as visual acuity decreases, more light is needed to achieve equivalent performance, and improved daytime legibility for VMS will level off between 8% and 20% contrast [240].

For static traffic signs, the recommended luminance contrast ratio is 12:1 [241]. Under low ambient illumination, increasing the contrast ratio above 4:1 will have a minimal effect on detection performance especially for self-paced tasks. For the random-scan CRT[242], it can be superior for both threshold and comfort cases in both the 1.2:1 and 1.5:1 contrast ratio conditions. According to the guidelines, fully reflectorized signs should have a figure/background contrast ratio of 12:1[241].

For the dynamic variable message sign in traffic control [232], the recommended luminance contrast ratio between 8:1 and 12:1 should be used for light emitting technologies and 40% daytime and 50% night time contrast for light reflecting technologies for VMS. In addition, night time luminance should range from 30 to 230 cd/m2..

4.6 Weight

Weight of a style of type refers to the volume of white space its letters replace with ink within a contained area.



Font weight

Different weights of the same typeface Regular
Different weights of the same typeface Italic
Different weights of the same typeface Bold
Different weights of the same typeface Bold Italic

Figure 4.18 font weights

From the viewpoint of design, the weight of the lines in a type style may vary from “light” to “medium” to “ultra bold”, as shown in Figure 4.18, which is a family of different weights of the same typeface. Inspecting it, we can see the subtle differences within the same typeface of characters.

Thus, from the viewpoint of image processing, weight can also reveal the subtle intra-class differences. Indeed, it will be an important factor to determine the different types, such as handwritten script and machine-print characters. It also has close relations to different type fonts in the design field, for instance, italic type, serifs, boldness and case.

4.6.1 Intra-class difference: different type font

4.6.1.1 Italic type

Italic type: Compared to regular upright characters, italics are narrower and spaced closer together. They are preferred over both bold type and upper case type to stress pronouns since their slanted form slows down the speed of recognition and has perceived elegance.

It is widely acknowledged that italics are very useful for emphasizing individual words within a text to get conspicuity since they are attributed a *contrastive* role in text. The content of the italicized term is often opposed to that of other words in the text and draws attention to the sentence as a whole and not just to the term itself [243].

4.6.1.2 Serifs and non-serifs

Serifs and non-serifs: Serif refers to the little extra stroke added as a stop to the beginning and end of the main strokes of a character. Fonts with serif are called serifs and fonts without serifs are called sans serif. In general, it is recommended that serifs should be chosen for text type and sans serifs should be used for display type in order to provide more contrast.

We encounter printed material in a variety of forms, and these materials may use any of thousands of different type fonts: serif, sans serif, script and those that do not fit into the other three categories. Most of what we read uses a serif or sans serif font, and these kinds of fonts are usually appropriate in many conditions, although some will be more legible than others. Serif fonts, which have little embellishment, typically are used for the text. It may be easier to segregate words with serif fonts, and different letters may be easier to identify. However, there is no difference in reading speed for serif and sans serif fonts. When we consider fonts type for use on CRT and LCD computer monitors, we also need to consider point size, screen resolution, illumination conditions, background colours, viewing distance, and monitor size.

When reading from a distance, Gothics are found to be more legible than Romans. For highway use, a font called Clearview has been specially developed to improve the legibility and readability of road signs [244] because it has three properties: thinner stroke widths; lowercase letters with increased loop heights; and more open letter spacing for the lowercase Clearview [245]. Consequently, people recognize words at a 16% greater distance with the Clearview font than with traditional highway fonts, which at 55 mph translates into an additional 2s to read the sign.

4.6.1.3 Boldness

Boldness: The boldness of characters, also called *weight*, is a physical property of a letter that can be varied while keeping other properties unchanged. Bolder versions of a basic typeface are produced by increasing stroke width. Therefore, the bold typeface of a letter is heavier than that of the version of the basic typeface of the letter.

Like italics and case, bold words tend to draw too much attention [243]. However, reading speed doesn't benefit form bold text in the normal fovea and periphery [246]. Many signs, especially highway signs, are customarily read at the greatest possible distance, i.e. as soon as possible during approach. Good highway signs maximize sight distance by using low complexity-bold-lettering.

4.6.1.4 Case

Case refers to how characters are capitalized within a word or phrase. There are two types of cases for alphabets: Uppercase and lower case.

Like italics, upper case lettering is used for emphasis, which can be used to bring attention to a specific word or phrase.

Within the field of typography and cognitive science, there is a popular belief that text set in mixed upper- and lower-case is more legible than in all upper-case because text in all upper-case reduces the shape contrast for each word. Since lower-case characters vary in both height and average position, making words constructed with them more distinctive, lower-case characters are much easier to read than all capitals [247, 248] and 90% of readers prefer lower case text as compared to 10% for all upper-case. In addition, there is a common belief that the legibility of lower-case words should be greater than that of upper case words [249].

However, from the viewpoint of optical vision, enlarging any small object makes it more visible to achieve better visibility, while upper-case text perceived at a greater distance has a retarding effect on reading speed. Furthermore, the uppercase letters are classified more rapidly as letters (vs. non-letters) when they are preceded by a briefly exposed [250]. Moreover, since letter size determines legibility for low vision readers and for those viewing visually small text, uppercase text is more legible in terms of reading speed for readers with reduced acuity due to visual impairment, and in normally-sighted readers when text is visually small [251]. When point size is fixed, uppercase text is simply more legible and familiar acronyms are processed more quickly in the familiar uppercase than in lowercase [252, 253].

4.6.2 Weight and stroke width

From the viewpoint of image processing, weight refers to the “inked area”. It can be exactly calculated by the total amount of the connected component “inked” pixels filling the inked area, or be roughly estimated by the contour and the stroke width. Imaging opens the letter’s compositional stroke and strings them together, just like in on-line handwriting which strings each stroke. Accordingly, it will be considered as a rectangular stripe with the width of stroke width and the height of the length of contour. The length of contour can be estimated by the total amount of the edge points in the contour, thus

$$
\begin{aligned}
\text{"inked" area} &= \text{the total number of the connected component "inked" pixels} \\
&\approx \text{the stroke width} \times \text{the total number of the edge points in the contour}
\end{aligned}
$$

Considering the region given by the tight rectangular boundary box of the letter, both edge density and inked pixel density get involved as:

$$\text{"inked" pixel density} \approx \text{edge density} \times \text{the stroke width.} \tag{4.2}$$

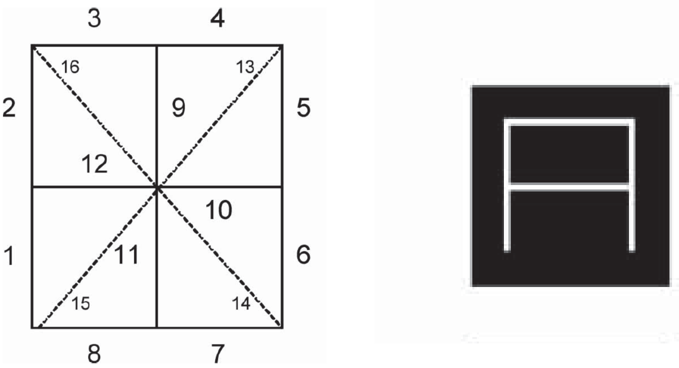
And this can be regarded as one of the necessary conditions of being a text object. Moreover, since weight describes many kinds of subtle differences among types in terms of design, it will play an important role with respect to our tasks e.g. region-level features and distinguishing the machine-printed text and handwritten text.

4.7 Straight line

“The straight line is godless [56]”, as text is one typical man-created visual object, the straight line is the basic element that gives form to a letter and determines the style of the type. And line terminations (the ends of letter parts) and horizontals appear to be the two most important features for letter identification [254, 255]. However, since there is ‘‘word superiority effect’’, little attention has been devoted to computational modelling of letter perception as well as the line property of character.

In 1974, Rumelhart and Siple[256] proposed the feature matrix for the letter computational model, see Figure 4.19 for a representation of the feature matrix, which consists of a set of 16 independent features that allow us to characterize any of 26 letters of the alphabet. For example, letter A is composed of eight features from this matrix: 1, 2, 3, 4, 5, 6, 10, and 12.

Based on the model, Grainger, Rey, and Dufau have claimed that the features of a letter mainly consists of lines of different orientation and curvature [257], and these have been tested in the interactive-activation computational model of letter perception.



“A” derived from
feature matrix:
1,2,3,4,5,6,10,12

Figure 4.19 Left side: the feature matrix (adapted from Rumelhart &Siple 1974), Right side: letter “A” derived from this matrix

The interactive-activation model of McClelland and Rumelhart [258] suggests a hierarchical organization of two levels: a feature and a letter level. The lines of different orientations are used as the representation at feature level and in feature-based identification visual processes at the letter-level. These representations are interconnected by feed-forward, feedback, and lateral connections, each being characterized by a fixed parameter that determines its weight. By systematically varying these parameter values, the predictions of the different computational instantiations can be tested. And the results are in favour of the computational model of letter perception.

From the viewpoint of human letter perception, characters are modelled as a set of lines of different orientation. However, for image processing, as figures in the background, proportions bridle each visual object in the same 2D space, and proportions are kept as the maximum possibility in the 3D space. Thus, we regard the line proportion of the length of the line segment in terms of the size of the letter as feature and call it the ratio of the straight line, denoted as RSL, as defined below:

$$RSL_H = \frac{\text{the length of Horizontal line segment}}{\text{Width of letter}};$$

$$RSL_V = \frac{\text{the length of Vertical line segment}}{\text{Height of letter}}. \qquad (4.3)$$

The line is also calculated based on the Kirsch edge map and the details are discussed in the three-level text computational model appearing in chapter 6.

4.8 Size and its related proportion

The size of type is described as the "depth of space required by one line of type", assuming a minimum distance between one line and the next. In other words, height is used to indicate size, without regard for the width of a character. And alternative measures of size are x-height, capital height, and total letter height.

In type design, there is a tension between considerations of distinctiveness and uniformity that are essential to the design process. Type designers use a systematic reference frame to create a family of objects for identification. The frame system constrains size proportions within the font, for example, the ratio of x-height to cap height and the length of ascenders and descenders are characteristics of a particular font [259]. The proportions vary somewhat among fonts, but within a restricted range thereby making fonts of the same point size appear larger or smaller. These regularity effects within a font lead to characters being related by weight, contrast, stress, or the axis of the letter. Meanwhile, individual characters are distinct in terms of the form and construction of the letter.

In the letter perception, the size of the letter weakly influences the efficiency for letter identification, but the efficiency is inversely proportional to perimetric complexity (perimeter squared over "ink" area, i.e. the contour of the letter form) [260].

For the human observer, since the size of the image projected on the retina is a function of both the size of the letterform and the viewing distance, the sizes of type are often characterized in terms of visual angle. Under optimal conditions, visual angles of between 0.35 and 0.40 degrees (equivalent to 8 to 10 point type) are capable of presenting a legible image to the viewer possessing perfect vision. And the fluent range extends over a factor of 10 in angular print size (x-height) from approximately 0.2° to 2°. Assuming a standard reading distance of 40 cm (16 inches), the corresponding physical x-heights are 1.4 mm (4 points) and 14 mm (40 points)[261].

For close up reading at a distance of 35 cm, such as in books, sizes between 9 and 11 points are predominantly used. And 10-point (9 pixels) type is best used for the text body, which is read faster than the type of other sizes, and has optimal legibility in electronic display [236]. However, 8-point type is most suitable for so called consultation text.

At a reading distance of 75 cm, for instance, when reading posters or signs, a character height of a little less than 4 mm could result in optimal reading performance (reading speed). For the commonly used typeface Times New Roman, this character height corresponds with a type size of 15 points. And in PowerPoint, it is best to use a font size of at least 22 points for bullets and 16 points for figure legends and axes, since these font sizes can project to screens at least 22 minutes of arc or 16 minutes of arc, as recommended for critical legibility or legibility, respectively [237].

In pixelated text reading, reading performance is impaired if as few as 6 x 6 binary pixels per character width are available, particularly with larger characters [262-265]. And a grid density of about 4 pixels per character width is needed to allow for accurate character definition [263, 265].

For text in an image within a clutter natural scene, illustrated in Figure 8, nearly 80% of the characters have their height in the range of [11, 90] pixels, and more than 90% of the characters have their height in the range of [11,150] pixels. Even though the height of characters depends on the size of sign in design, very large (>300 pixels height) or very small characters (<10pixels height) are small in terms of quantity, being about 3% and 0.2% respectively.

4.9 Summary

In type design, there is a tension between considerations of distinctiveness and uniformity that is essential to the design process. Individual characters must be distinct in their form and construction and the shapes of individual letters are constrained within a font by the similar systematic reference frame of design. Accordingly, they are related in terms of the shape, proportions, weight, contrast, and other stylistic attributes of letters.

From the standpoint of design, there are several important factors which contribute to the readability, legibility and conspicuity, including the size and its related proportions (height to width ratio), luminance, viewing distance, stroke width-to-height ratio, and weight.

However, most investigations of character properties for text legibility in the literature use simplified stimuli presented on otherwise featureless backgrounds, a situation quite unlike the typical natural world. Moreover, in relation to the image we cannot restore the above mentioned factors, such as real size, viewing distance and luminance. Thus, while these studies have provided important insights into the essential process of letters perceived, the extent to which this understanding holds true in natural vision is less clear.

In relation to image, luminance can be connected to the intensity, and contrast can be associated with the local RMS contrast. Moreover, proportions can be assumed to be kept at maximum possibility (like the real one) since they possess inbuilt, intrinsic properties for the form and construction of letters. On the public benchmark dataset for text detection within a clutter scene, we analyse the height, mean intensity, local contrast and the relations among them.

In an image with a cluster scene, nearly 80% (79.27%) of the characters have their height in the range of [11, 90] pixels, about 13% of the characters have their height in the range of [91,150] pixels, 5% of the characters have their height in the range of [151,300] pixels, about 3% of the characters have their height in the range of [301, 1000] pixels, and only 0.2% characters have their height at less than 10 pixels.

For mean intensity, about 80% of the characters have their mean intensity in the range of [80,170], 10% of the characters have their intensity in the range of [61-80] and [171,180], and another 10% of the characters have their mean intensity in the range of

[181,255] and [31, 60]. Thus, the mean intensity can be divided into three intervals, such as [1, 80], [81,170], [171,255] which corresponds to nearly 10%, 80% and 10% of the characters.

Moreover, both mean intensity and local contrast do not seem to have interactions with height. However, mean intensity and local contrast has inverse relations between their trend lines.

Additionally, besides the proportions of letters, another four particular properties are of interest, including line, weight, orientation and size. “Straight line is godless”, as text is a human created object while line represents the basic element in letters both in terms of letter perception and in image processing. Since line length varies broadly in comparison to height, proportions of line length to size will be regarded as significant features for individual characters. Weight is related to many kinds of type, that is, it tells the difference in terms of intra-class and gets involved in the features at the region-level. As a result, the weight related features in image processing will play a very important role in this project. Orientation is not only important to the text object but also to the clutter scene. Size is the basic element of physical appearance, thus it plays a significant role in our project.

In summary, from the standpoint of image processing, the properties of individual characters consist of mean intensity, local RMS contrast, stroke width to height ratio, height to width ratio, straight line ratio, and weight related attributes. Owing to these varying properties an image-processing based algorithm will be employed to extract them in the following chapters.

Chapter 5

Properties of Local Spatial Organization

Besides distinctiveness, through the commonalities in the shape, proportions, and other stylistic attributes of letters within a font, uniformity in letters is achieved. It is the uniformity that makes text contain verbal stimulus control features, including type, line, weight, orientation and size. Text can "say" something and be identified by readers. Consequently, text is "dressed up" [266] in the "costume" of typestyles which imbues meaning i.e. physical appearance. Also, the uniformity of letters contributes to text legibility in relation to the psychological concepts of perceivability, bias, similarity, and letter identification in computation [267].

Any pair of letters, beyond their distinctiveness, has a visual similarity that can be defined by, for example, the numbers of features (e.g line) the two letters have in common, which is referred to letter confusability. And for any letter, one can average its confusability with the 25 other letters to give a measure of that particular letter's overall letter confusability [268]. And for any letter string, one can measure its overall confusability.

In any letter string, there is a significant interaction between letter spacing and confusability on the flanked letters. Letter confusability together with the neighbouring spatial arrangement leads to a visual crowding effect since visual crowding refers to the detrimental effect that nearby or "flanking" objects have on the spatial processing of a test object by excessive feature integration.

As text is composed of letter strings or words, which have significant interactions between the organization of letters and their identification, and this perceptual organization can be modulated by the spacing between letters [269], changes in crowding can be modulated by the function of inter-letter spacing [48].

Although these studies are presented from the viewpoint of perception and design, they have provided important insights into the essential calculation of space organization in image processing. We will apply the rules as prior knowledge to quantify and measure the space organization in images.

Therefore, this chapter deals with those properties of the local spatial organization of text which directly relate to the arrangement of neighbouring letters in text, including letter spacing, word spacing, alignment, line width, and leading.

5.1 Textual organization and space

In many cases, the printed text is required to be conspicuous. There are two kinds of conspicuity that can be achieved. One is to emphasize single words or some paragraphs by distinctly different typefaces in order to receive more attention from the reader than other words or paragraphs. For instance, titles are often set in bold in papers or in magazines. And the other is to emphasize the entire area of text to get high figure background contrast by spatial arrangement, i.e. text organization in space, for the purposes of being conspicuous.

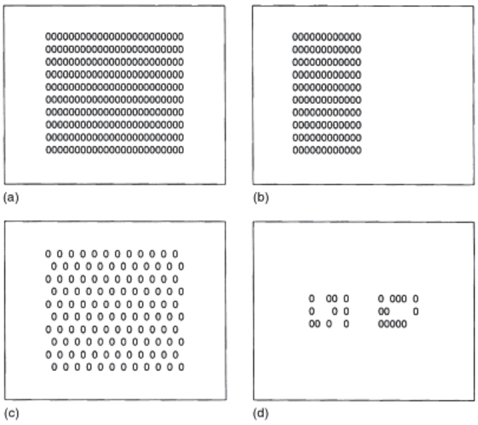


Figure 5.1 After [4, 5] examples of different display densities and grouping: a) overall density=100%, local density=81%; b) overall density=50%, local density=72%; c) overall density=50%, local density=39%; and d) grouping into two sets.

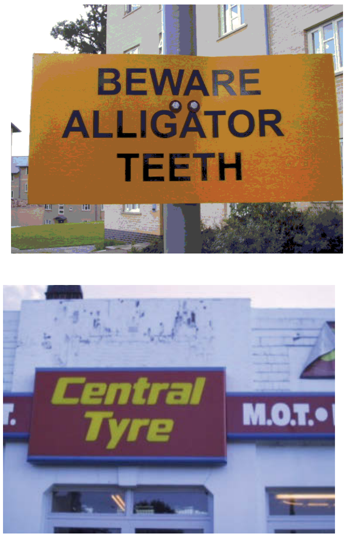


Figure 5.2 Samples of grouping into several sets

An important aim of textual organization is to optimize readability. Indeed, in the case of display, the characteristics of display must be related to the spatial array of characters on the display. Further, there are four basic characteristics of alphanumeric display formats that influence the ability of an observer to read or interpret the display [5]: overall density, local density, grouping and layout complexity, as illustrated in Figure 5.1.

Overall density is the number of characters shown over the total area of the display, often expressed as a percentage of the total character spaces available. Local density is the density in the region immediately surrounding a character, often manipulated by altering line spacing. For best readability, overall display density should be as low as possible, with local density at an intermediate level. This reduces lateral masking between display characters and increases the ease with which a reader can locate information in the display.

Grouping is related to the Gestalt organizational principles i.e. the extent to which items form well-defined perceptual groups. Layout complexity is the extent to which the arrangement of items on the frame follows a predictable visual scheme, which can be quantified by the formula adapted from information theory [270]. Grouping display elements will improve readability so long as the groups are appropriate, but there is a trade-off between grouping and layout complexity [5, 271]. More groups means higher complexity, and increased layout complexity can mean decreased readability. In addition, for text in the clutter scene, most of them exist in the grouping display, as shown in Figure 5.2.

Clearly, the dominant ingredient in the organization of text is space. Space is a particularly compelling tool for organizing text or a display because the visual system automatically attempts to group elements that are close together within the available space. In terms of typography, space consists of margin, letter spacing, word spacing, alignment, line width, and inter-line space. And we start from the most fundamental of spacing being the letter spacing [272] within a word.

5.2 Letter spacing and word spacing

In the clutter scene, forming text according to spatial arrangement is a function of position within a horizontal array. Therefore, the letter spacing plays a central role in calculating the relative distance between them and the neighbouring letters and this is determined by the height of the character. Further, word spacing plays an important role in calculating letters as a string of words.

We investigate the recommendation of them in the original design, and then take them into account to set the letter spacing for calculating the space organization in our task.

5.2.1 Letter spacing

Letter spacing, also called inter-letter spacing, or tracking, refers to the distance between the closest parts of two adjacent characters.

Inter-letter spacing generally applies to the overall spacing between all neighbouring characters in a set and affects the information density in a line or a block of text. High-density type allows the designer to fit more words on a line of text than when using type with regular spacing, whereas low-density type takes up more space per word.

In typographic designs, the type of font influences the letter spacing. There are two types of font, one of which yields only limited possibilities in letter spacing. A *monospace* font comprises an alphabet of which all letters have the same width. Such single width characters can be found on printers. A *proportional* font comprises an alphabet of letters which vary in the amount of space they take up, resulting in balanced letter spacing which occupies less space than single width characters.

Proportional spaced fonts use a different amount of horizontal space depending on the width of the letter. Thus, font size is a factor significantly affecting proportional spacing. For smaller font sizes, it results in tighter spacing while larger fonts lead to bigger spacing. Spacing with smaller fonts needs to be maintained at the default size or larger. For the regular font size of 10 to 12 points, spacing can be more condensed, about -5 pixels from the default, and it can still maintain good accuracy. For bigger font sizes the letter spacing can be decreased to -10 pixels or more without sacrificing text legibility[273].

For text legibility, the size of inter-letter spacing plays a central role since it manipulates the amount of lateral interference among neighbouring characters [274] or words [233, 275]. By increasing letter spacing, word legibility can be improved and gradually reach an asymptote close to single character legibility. Performance deteriorates non-linearly whenever letters are separated by at least 2 blank spaces, with the concomitant emergence of a word length effect, and the threshold of about 2 spaces is constant across variations in font size [276].The reason for this is that there is a dual effect [277]: increasing letter spacing improves individual letter identification but damages whole-word form (the unitization of words [278]) and/or parallel letter processing.

For application, there are solutions, for instance, the inter-letter spacing is recommended to range from 25 to 50 percentage of letter height for large format application in buildings [279]. In large format signage, inter-character spacing of 2/7th letter height and line spacing of 4/7th height are best.

Once the spacing within words has been determined, the word spacing can be determined.

5.2.2 Word spacing

The distance between the words of a line of text is called word spacing, or inter-word spacing [280]. It has an influence on word segmentation and facilitates word selection and identification [281]. Thus, it has an effect on reading performance[282].

In English, there is interplay between letter spacing and word spacing, which influences a font's readability. To enable a reader to easily distinguish between individual words, the distance between the last letter of one word and the first letter of the adjacent word needs to be significantly larger than the distance between adjacent letters within one word.

The optimal distance between two words is 25 percent of the type size [203]. For all practical purposes, based on practical experience in typography, the "i" rule is adopted as the conventional format for word spacing. This means that word spacing is about equal to the space occupied by the letter "i". In applications, inter word spacing is recommended to be in the range from 75 to 100 percent of letter height.

5.2.3 Setting in space organization calculation

Only when proportions among component parts are captured, the image or picture of the object is recognized as the same as the object itself in the real world. Therefore, in the clutter scene, we assume that the ratio of letter spacing or word spacing to the height of type size is kept at the maximum possibility when the image is taken.

Considering the recommendation of letter spacing and word spacing in relation to design, we have, after trial and error, set the letter spacing at less than 60 percent of letter height to calculate the letters into a string of words.

5.3 Text line

After inter-letter spacing and inter-word spacing are determined, the line of text is found as a whole. There are two attributes of the text line: alignment and line width.

5.3.1 Alignment and the importance of neighbourhood

Alignment is the way the lines of text are arranged in relation to each other within the margins that have been set. There are four options for alignment. We can choose to align the lines to either the left or the right, and justify or centre the lines of text.

Alignment on the left: It is also called ranged left or unjustified. When the text is aligned on the left only, the left margin is fixed whereas the right margin is not fixed. Therefore, all lines start at the same distance from the left margin, producing a ragged envelope where the lines end. For the purpose of increasing readability and reading comfort over the ragged envelope, rational spacing is made for the left range text breaking off the lines, which is useful for instructional text, and can convey the text information clearly as well as the text structure.

Justification: The other traditional way of aligning lines of text is justification. In order to avoid an uneven right margin, full justification is used to cut off the long words at the end of a line by hyphens, and vary both letter spacing and word spacing in such a way that all lines end at the same distance from the right margin. And with word processors in computers, the text can be easily justified in different ways: fill-justified, equal-justified, and micro-fill justified. Consequently, readers get used to the full justified arrangement of text. They do not grow accustomed to ragged right text. Further, it is more tedious to read unjustified papers than those that are justified. However, in terms of reading speed, all forms of justification read equally well, and there is no reading time superiority of ragged –right (unjustified) text [283] .

Uncommon arrangement: Neither right alignment nor centred text is commonly used in ordinary text. Both are most often used in brochures or volumes of poetry in order to draw attention to the informative content. And none of the special arrangements of text are significantly superior to the conventional arrangement.

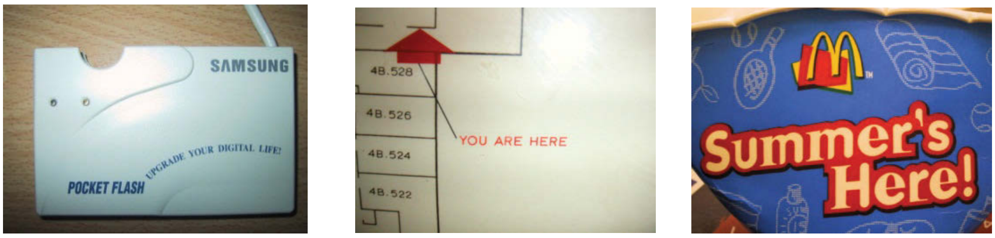


Figure 5.3 Samples of compliant alignment

For text in the clutter scene, most are used to draw attention to the informative content, thus uncommon arrangement or compliant arrangement are usually used, for instance in Figure 5.3, the principle of chunking is incorporated in several group sets of text in arbitrary places and these arrangement strongly emphasize each individual chunk.

Moreover, since the word recognition processes are very tolerant of text orientation [284], exhibiting a modest decline for orientations within +/-60 degree of the horizontal regardless of visual field meridians (right horizontal, upper-right diagonal, vertical, and upper-left diagonal), there is a compliant arrangement in which words are not aligned at all and are placed in harmony with the message content. Thus, especially in this case, only the neighbourhood of individual characters will provide significant clues for leading to the text chunks by grouping neighbouring letters in order to predict where the text is.

However, in the early stages of the detection task during which the visual object is not identified as letters, the neighbourhood needs to be estimated in the state of black box. In order to measure the neighbourhood without knowing any identification, we need to consider spacing, position, and physical appearance to calculate the proximity in position and the similarity in appearance. The computational detail will be discussed in Chapter 5.

5.3.2 Line width

Line width refers to the distance between the left and right margins of a text column, and is also called length of line. It affects both reading comfort and reading speed. If lines are too short, it will result in frequent hyphenation, and readers also very often have to shift their view to the next line thereby breaking read rhythm. On the other hand, if lines are too long, the reader will have a hard time focusing on the text. Furthermore, it can be difficult to continue onto the correct line in large blocks of text. The ideal line length depends on the design of typeface, type size, and line spacing.

For books, for the conventional type sizes (9 to 12 points) it has been concluded that, on the average, the optimal line width lies between 10 to 12 words per line. This amounts to a range of between 50 and 70 characters per line of text, also counting word spaces, to achieve the highest readability [285]. Thus, generally, the optimal line length for body text is considered to be 50 to 60 characters per line, including spaces [286], or 9 to 10 words, and anything from 45 to 75 characters is widely-regarded as a satisfactory length of line [203].

For signs, in limited space with high overload and information conflicts, the information density of the primary navigational message should be limited to a single glance. A simple message (i.e., one with few characters or elements), which can be made large to allow it to be seen further upstream, should be limited to a maximum of six words [287].

For text in the clutter scene, there is big uncertainty of line width since they are either from books or from signs. Accordingly, what we can do is to focus on its related factors: letter spacing and word spacing, that is, the neighbourhood measurement of letters.

5.3.3 Inter-line spacing

Inter-line spacing, also called leading, is the distance between the baseline of one line of text and the baseline of the next.

The amount of points of leading depends on the type size. A rule of thumb is that the leading that warrants optimal reading conditions is 125 percent of the type size for any size of type [203]. For instance, given a type size of 12 points, two lines of text would ideally be separated by 15 points in distance, i.e. 3-point leading.

Instead of the point measurement in print practice, computers offer the possibility to space lines at 1, 1.5, or 2 times the regular spacing. Regular spacing corresponds with the setting regularly chosen for the particular typeface in which the text is printed (not solid setting). 1.5 times spacing refers to a setting in which the width of the space between lines is twice as high as regular spacing.

In terms of readability, line spacing has significant main and interaction effects on both the proofreading time and detection rate [288]. Increasing interline blank spacing also speeds up the reading process overall, while also improving the identification of the words and the letters within words [278] presumably because it decreases the adverse effect of crowding between adjacent lines of text [289] . Thus, wider line spacing can lead to better accuracy and to faster reaction times [290]. Especially, when the contrast of the text is reduced, as may occur within intraocular light scatter or poor viewing conditions, spacing becomes particularly important [282].

Moreover, there are solutions for different applications. In screen-based proofreading, 1.5 line spacing is recommended for use. For large format application in buildings, the inter line spacing is recommended to range from 75 to 100 percent spacing [279]. Note that, word spacing should appear to be narrower than leading. If not, the evenly spaced look of the text is broken up by 'rivers' that appear to run through it vertically [204].

For text in the clutter scene, more blank space will increase the text conspicuity while maintaining the group set of the text. Thus, the inter-line spacing plays an important role in filtering out the non-text objects.

5.4 Keeping balance among spacing

Letter spacing, word spacing and inter-line spacing should keep in a balance so as to have a harmonious design on the whole. This means that letter spacing Ls should be smaller than word spacing Ws , which should be smaller than type size Ts , i.e., $Ls < Ws < Ts < Leading$. Leading should be the biggest number of points. Accordingly, there is a rule of thumb about the relationship of size and ratio [204], $Ws : Ts : Leading = 1:4:5$. But that ratio is not applicable when text is fully justified because both letter and word spacing are variable in that case.

Specifically, in signs in which the text is displayed within a background panel, spacing is essential to legibility. Just as white space gains attention in newspaper text, besides the five typical colours (red, blue, yellow, white and black) in standard traffic signs, signs that have blank space are more easily noticed. Blank space may be obtained by making signs larger or by removing secondary copy that has no navigational value, in order to increase the text conspicuity while maintaining the group set of text. More empty space should generally result in less secondary copy thereby leading to more conspicuous text. And the open space surrounding the copy area of a sign ideally should not be less than 60 percent of the sign or background area, such that a reasonable expectation of legibility will exist.

And in instances in which only letters comprise the total sign, such as channel letters on building walls, there is a useful rule of thumb which takes into account the letter spacing in a line as usually being 1/3 the width of an individual letter and this can give a surprisingly close determination of the actual length of the line of letters.

For text in the clutter scene, there are big uncertainties of position, arrangement and size, and only sorts of spacing, including letter spacing, word spacing and interline spacing, can be measured for the purpose of grouping neighbouring letters into words or text group sets. All of these spacing arrangements contribute to a neighbourhood of letters, words and text group sets. When we calculate the space relationships among visual objects, the first thing we need to do is measure one of the three types of spacing based on the size of characters.

5.5 Summary

Letter spacing, word spacing and interline spacing are in harmony with each other and organized as a whole text which keeps crowding and readability in good balance. All of them are highly related to the type size, and have a practical recommendation ratio to the type size respectively.

For text in the clutter scene, since there are big uncertainties of position, arrangement and size, only spacing contributes to the neighbourhood and can be measured for the purposes of grouping neighbouring letters into words or text group sets. Thus, they are integrated into the text computation model as attributes in the letter-centred level and word-centred description in chapter 6, and calculated in the image-based method for the purposes of quantifying space organization.

Chapter 6

Representation of Image and Text in Clutter Scene

Crowding is a consequence of spatially pooling features within receptive fields of increasing size: information is averaged or not resolved by attention. And clutter also starts from the strong attempts to quantify information density. Therefore, both crowding and clutter need a reasonable region in which to operate for the purposes of achieving calculated targets. Fundamentally, this region needs to represent image by quantifying the spatial layout or space organization and based on this representation, the computational model of text can be built up.

Inspired by a painter’s description of proportions among component parts in a picture, this chapter will mathematically deal with the region generation, quantification of the spatial representation of image, and the three-level text computational model in detail.

6.1 Space-averaged image representation

Space organization plays a significant role in clutter or crowding. We need to specify the region of image over which statistics are computed, and capture the organization explicitly—the grouping by similarity + proximity.

As described above, crowding is as a consequence of spatially pooling features within receptive fields of increasing size: information is averaged or not resolved by attention. And the pooling regions for computing statistics are smallest at the fovea and increase in size approximately linearly with increasing eccentricity.

Since the pooling region is tightly related to the receptive field size, we need to explore the conditions of increasing the receptive field size for the purpose of capturing the organization explicitly.

6.1.1 The size of pooling region

Early in 1952, Kufflor found that the receptive fields of light adapted cat retinal ganglion cells are approximately circular and have functionally distinct central and peripheral regions. And Hartline-Ratliff equations (1, 2), were used to model receptive fields. Then, for the response to moving bars, the concentric difference of Gaussians was proposed as a model for the receptive field of retinal ganglion cells. Later, arbitrary temporal phase differences between the centres and surroundings were included in the difference of Gaussians model [291].

Specially, for the response to drifting gratings, a Gaussian subunits solution[292] is presented as follows: (i) model receptive fields can be composed of any number of subunits, located anywhere in the $X - Y$ plane; (ii) the subunits are not required to be radially symmetric, i.e., any two-dimensional Gaussian function is allowed; and (iii) responses are predicted to gratings of any spatial frequency at any orientation. As a result of the constraint that the model of receptive fields must show linear spatial summation, the response can be calculated by a summation of the individual subunit responses.

Similarly, in vision contour integration research, Field and Hess discussed the linking between given elements in terms of the "association field" which integrates elements across neighbour filters tuned to similar orientations [177]. Moreover, Einevoll presented both the discrete and the continuous mathematical model for the spatial receptive-field organization [293].

These investigations of receptive fields provide us with the following principles. Given that the spatial element is anywhere in the 2D plane, those elements across neighbour filters are tuned to similar orientations and can be integrated into an association receptive field. Since crowding has the property of averaging within the receptive field, the pooling region can be obtained by integrating similar spatial elements in the neighbourhood.

6.1.2 Pooling region generating

6.1.2.1 Spatial elements

According to the theory of perspective, when objects recede from the eye or camera, the size of the object decreases; this means that the space of the object in image is reduced meanwhile the contour is lessened. If the distance is far enough away, the contours or boundary of separated objects disappear and those separated objects as parts are merged into a whole to show. This suggests that there are two spatial elements which need to be represented: the space occupied by the object and the contour or edge of the object.

Although we will explore the contour of the object based on the reasonable edge detector in the next chapter, contour will disappear when the space occupied is too small, and it can be distorted by crowding or clutter [188] [197]. Moreover, space occupied by the object exists and has discernible information. That is what impressionists do.

Significantly, impressionists realized that an object does not have its own colour but many individual patches of colour, so they used directional brushstrokes or colour patches, which are small space patches in space regularity, to represent "formless" visual objects instead of clear contoured shapes. Learning from this school of art, we have used an image-based method-connected component analysis to represent the space patches forming an object.

6.1.2.2 Geometric mean regions

Considering our task of text extraction, all the colour (RGB) images are converted to grey-scale. There is an advantage of grey scale objects over letters and silhouettes in a cluttered environment, in that the informative features of objects, defined by local variations in contrast, appear to mitigate the detrimental effects of crowding. Compared to letters or silhouettes, grey-scale objects (e.g. intact, aperture, and donut) require a much smaller increase of contrast in crowded conditions to restore accuracy to the uncrowded level. Thus, all of the following operations are on the grey-level image.

The space patches are defined solely by incorporating all the similar grey level pixels in the neighbourhood of the currently selected pixel, element or unit. And pooling regions in global image can be defined by GM regions. GM regions are defined by the property of the geometric mean function in space patches.

The concept can be explained informally. Imagine all possible grey patches in an image I , they are arranged in a consistent way to generate texture or form. This looks like the preconceptions about a “tree” or letters or “windows” or “sky”. If we are shown an object in an image, instead of rendering solid objects, many individual grey patches work together in harmony to make up a form, i.e., the form of an object is the harmonic combination of grey patches.

What and how is harmony created? ‘Do you not know that our soul is composed of harmony and that harmony is only produced when proportions of things are seen or heard simultaneously? [55]’ According to Leonardo Da Vinci, proportion is in all things, it is not only found in numbers and measurements but also in sounds, spaces and in whatsoever power there maybe. Harmony is composed of the union of its proportional parts reacted (e.g. seen or sounded) simultaneously. And a feeling of beauty in human being is born from these harmonious proportions. For visual objects in an image, their component grey patches are made to react simultaneously and can be seen at one and the same time both together and separately, their geometry and proportion are therefore keys in creating a structure and their proportion should ideally reflects the object in the real world.

How to capture these proportions among the component parts of a visual object? In mathematics, the geometric mean (GM) can capture the ratios to the reference value. The fundamental property of the geometric mean, which can be proven to be false for any other mean, is

$$GM\left(\frac{X_i}{Y_i}\right) = \frac{GM(X_i)}{GM(Y_i)} \tag{6.1}$$

This makes the geometric mean the only correct mean when averaging normalized results, and it is the results that are presented as ratios to reference values, i.e. geometric mean can capture the proportion. And a geometric mean is often used to compare different items– finding a single "figure of merit" for the items–when each item has multiple properties that have different numeric ranges. Thus using it among grey patches, the proportions of component parts to the whole image can be captured implicitly.

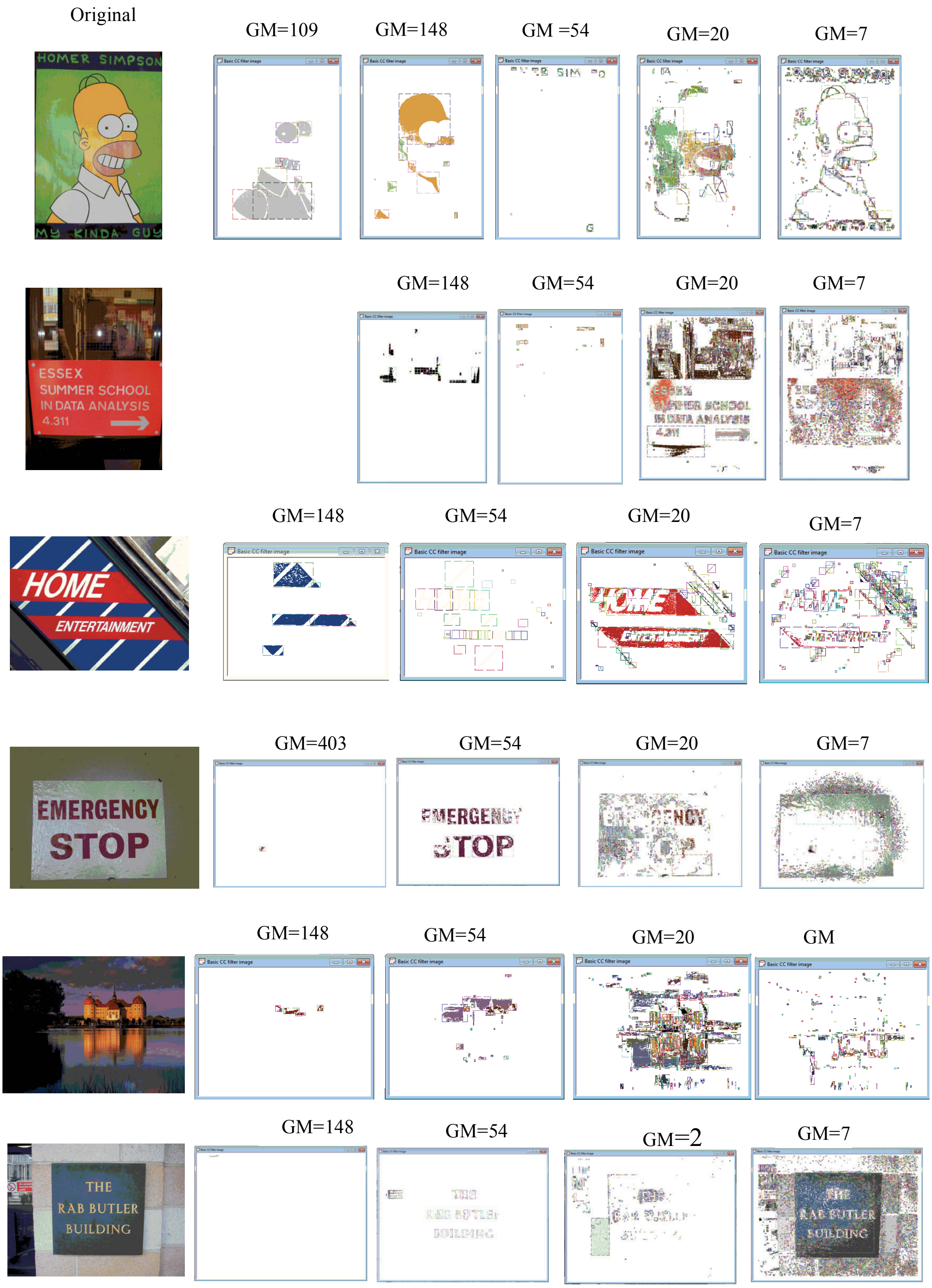


Figure 6.1 Images partitioned into GM regions

Image I is a mapping $I : D \subset R^2 \rightarrow S$. cc space patches/regions are well defined on images if:

1. S is totally ordered, i.e. a reflexive, antisymmetric and transitive binary relation $\le$ exists. In this paper, $S = \{0,1,\ldots,255\}$ is considered, and the regions can be defined on the images.
2. An adjacency (neighbourhood) relation $A \subset D \times D$ is defined. In this paper 8-neighbourhoods are used, i.e. $p,q \in D$ are similar and adjacent (pAq) iff $\sum_{i=1}^{d}|p_i - q_i| \le 1$, and p,q has the similar grey level s.

Region Q is a contiguous subset of D, i.e. for each $p,q \in Q$ there is a sequence $p,a_1,a_2,\ldots,a_n,q$ and pAa_1,a_iAa_{i+1},a_nAq. Let N_{s_i} be the number of the sequence points.

GM indicator Let $\Lambda_S = \{s \in S \mid Q_1,Q_2,\ldots,Q_{sg}\}$ be a set of regions in grey level s. The number of regions is sg. Among the set of regions Λ_s, the number of the sequence points is selected as the indicator of each region, and the geometric mean $GM_{\Lambda s}$ is computed as $GM_{\Lambda_s} = \left(\prod_{s\_i=1}^{sg} N_{s\_i}\right) = \sqrt[sg]{N_1 N_2 \cdots N_{sg}}$. And $GM_{\Lambda s}$ becomes an indicator of the Λ_s.

GM regions Let $\Psi_{GM} = \{\Lambda_{s1},\Lambda_{s2},\ldots,\Lambda_{sk}\}$ be a set of regions with a similar geometric mean, called GM regions, and Λ_{sj} with the number sg_j of space patches. Since there are similar geometric means, a set of values of geometric mean exist in the image, let $GM = \{GM_1,GM_2,GM_3,\ldots,GM_r\}$ denote it, and the number of geometric mean value is equal to r. Then, the image I can be represented as the set of GM regions $I = \bigcup_{i=1}^{r} \Psi_{GMi}$.

Table 6.1 Definitions used in following sections

In equation (5.1), if Y_i is the total amount of points in an image, and X_i is a variable of the number of points of grey patches in a given grey level, then $GM(X_i)$ can find a single “figure of merit” in the global image. If Y_i is a variable of the amount of points in whole objects, which can be formed in many kinds of grey levels, X_i is a variable of the number of points of component parts, and $GM(X_i)$ can capture a single “figure of merit” in an object as well as those in different objects.

Therefore, GM can be an indicator of a “figure of merit” in an image that has close relations to the proportion of component parts. Since GM can indicate the “figure of merit” among proportional component parts, GM regions get involved in object constitution. Shown in Figure 6.1, image is composed of the GM regions in different GM levels.

If several kinds of grey levels have similar proportions or rhythms in an image, they have the same GM value, and can form the same object, or in other words the same type of objects bring to human beings the similar feeling of beauty. So those grey patches with similar GM values are composed of GM regions. The formal definition of the GM regions concept and the necessary auxiliary definitions are given in Table 6.1. According to the Gaussian subunits solution and the associated field, the receptive field must show the linear spatial summation. And those GM regions, which are generated by neighbourhood proximity and similarity, can be considered as our pooling regions over which statistics are computed.

6.1.3 Statistics feature over GM regions

The investigation of crowding in the cluttered image shows that the averaging property of crowding is still reasonable, and local image statistics can reasonably predict where crowding occurs. Moreover, clutter can be measured by cluster density.

Bex et al shows that local image statistics, such as target size, eccentricity, local RMS contrast and edge density, can reasonably predict where crowding occurs. And those differing local positions, orientations, phases and spatial structures can be synthesized by summary statistic crowding modelling which is based on a textural representation [78]. However, the cluttered image is usually composed of a broad range of spatial and temporal structures, and the standard contrast sensitivity function is a poor indicator of sensitivity to structure in cluttered scenes. Contrast sensitivity does not increase monotonically with the image contrast at the target location but is similar for very high or very low local contrasts. But, contrast sensitivity does rise monotonically with the density of edge features at the test location. The sensitivity to spatial structure in natural scenes depends on the distribution of local edges as well as the local amplitude spectrum. Thus, for our task, over GM regions, statistical features are analysed, for instance pixel density, RMS contrast, edge density, orientation, direction density, and neighbourhood measurement. We need to analyse not only those general local statistical features, but also those discriminative features of text. The latter features of text will be discussed in section 6.2.

Neighbourhood measurement has tight relations to both the text object and non-text object, and we will define it over GM regions. Moreover, we'll deal with the general local statistical features. These include pixel density, orientation, direction density, RMS (root mean square) contrast and edge density. Pixel density corresponds to luminance; RMS contrast is the variation in pixel intensity over GM regions; direction density is the space-average direction output of edge detector, where has eight bins for eight directions respectively; Edge density is the space-averaged binary output of the edge detector, where higher values denote more edge pixels per unit area.

6.1.3.1 RMS contrast

Given an image $I_{M\times N}$, its global RMS contrast can be computed by mean intensity $\bar{I}$ and standard differences of intensity σ in the global image, shown in Equation (4.1) in Chapter 4 and rewritten as follows:

$$RMS = \frac{\sigma}{\bar{I}}\ ,\ \bar{I} = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} I_{i,j},\quad \sigma = \sqrt{\frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(I_{i,j} - \bar{I}\right)^2} \qquad (6.2)$$

And the local RMS contrast in a space patch $Q_{M_Q\times N_Q}$ can be computed by mean intensity $\bar{I}_Q$ of local region Q and its standard differences σ_Q as thus:

$$RMS_Q = \frac{\sigma_Q}{\bar{I}_Q},\ \ \bar{I}_Q = \frac{1}{M_Q N_Q}\sum_{i=0}^{M_Q-1}\sum_{j=0}^{N_{Q-1}} I_{i,j},\quad \sigma_Q = \sqrt{\frac{1}{M_Q N_Q}\sum_{i=0}^{M_Q-1}\sum_{j=0}^{N_Q-1}\left(I_{i,j} - \bar{I}_Q\right)^2} \qquad (6.3)$$

And over the GM Regions $\Psi_{GM\_\Lambda s} = \{\Lambda_{s1}, \Lambda_{s2}, \ldots, \Lambda_{sk}\}$, the average RMS is computed as

$$RMS_{\Psi_{GM\_\Lambda s}} = \frac{\sigma_{GM\_\Lambda s}}{\bar{I}_{GM\_\Lambda s}}$$

$$\bar{I}_{GM\_\Lambda s} = \frac{1}{\sum_{j=\Lambda s1}^{\Lambda sk} sg_j} \sum_{j=\Lambda s1}^{\Lambda sk} \sum_{s\_i=1}^{sg_j} \bar{I}_{j,s\_i},$$

$$\sigma_{GM\_\Lambda s} = \frac{1}{\sum_{j=\Lambda s1}^{\Lambda sk} sg_j} \sum_{j=\Lambda s1}^{\Lambda sk} \sum_{s\_i=1}^{sg_j} \sigma_{j,s\_i} \tag{6.4}$$

Where $\bar{I}_{j,s\_i}$ is the mean intensity of a space patch, and σ_{j,s_i} is standard intensity difference of the space patch, and both of them can be computed by equation (6.3).

6.1.3.2 Orientations, direction density and edge density

Based on the edge map obtained by Kirsch operators, edge density is computed over the space patches and GM regions. Given a space patch Q_i, Let $N_{s\_con\_i}$ denote the number of edge points over local region Q_i, and $N_{i\_dir\_z}$ denote the number of edge points in direction z, $z = \{0,1,2,3,4,5,6,7\}$ corresponding to the indicators of the above compass direction kernel. Over the GM Regions $\Psi_{GM\_\Lambda s} = \{\Lambda_{s1}, \Lambda_{s2}, \ldots, \Lambda_{sk}\}$, the averaged direction density $Density_{GM\_\Lambda s\_dir\_z}$ is computed as thus:

$$Density_{GM\_\Lambda s\_dir\_z} = \frac{1}{\sum_{j=\Lambda 1}^{\Lambda k} sg_j} \sum_{j=\Lambda 1}^{\Lambda k} \sum_{i=1}^{sg_j} N_{j,i\_dir\_z} \tag{6.5}$$

And the edge density $Density_{GM\_\Lambda s\_s\_con}$ is computed as thus:

$$Density_{GM\_\Lambda s\_s\_con} = \frac{1}{\sum_{j=\Lambda 1}^{\Lambda k} sg_j} \sum_{j=\Lambda 1}^{\Lambda k} \sum_{i=1}^{sg_j} N_{j,s\_con\_i} \tag{6.6}$$

6.1.3.3 Cluster and neighbourhood measurement

Image I, there are r levels of GM, let $GM = \{GM_1, GM_2, GM_3, \ldots, GM_r\}$ enumerate the values of GM. Therefore, image I consists of a set of GM regions $I = \bigcup_{k=1}^{r} \Psi_{GM_k}$, Ψ_{GM_k} denotes the GM region corresponding to GM_k. Over GM regions, adjacency (neighbourhood) relation $A_{cc} \subset \Psi_{GM_k} \times \Psi_{GM_k}$ is defined. For given regions $Q_p, Q_q \in \Psi_{GM_k}$, both the position and physical appearance of the region are considered. Space among the two regions corresponds to position change, which consists of the x coordinates change and y coordinates change. And the width and the height of the tight rectangular boundary of the region correspond to the physical appearance. And over the three dimensions, Gaussian functions are applied to measure the relations between Q_p and Q_q as follows:

$$\mathrm{Re}lation = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right), \qquad \Delta = \left|\xi_p - \xi_q\right| \tag{6.7}$$

Where ξ denotes four dimensions, including x-coordinates change, y-coordinates change, width, and height of the tight rectangular boundary, and σ is the Gaussian constant. Regions Q_p And Q_q are similar and adjacent $(Q_p A_{cc} Q_q)$ iff $\mathrm{Re}lations \in 1\sigma$ confidence interval.

Cluster Ω_l is a contiguous subset of GM regions Ψ_{GM_k}, i.e. for each $Q_p, Q_q \in \Psi_{GM_k}$, there is a sequence $Q_p, Q_{c1,} Q_{c2}, \ldots, Q_{cn}, Q_q$ and $Q_p A_{cc} Q_{c1}, Q_{ci} A_{cc} Q_{ci+1}, Q_{cn} A_{cc} Q_q$. Let $N_{gm\_c\_l}$ be the number of the sequence regions.

In GM regions Ψ_{GM_k}, there might be several clusters, whose locations and length of the sequence reflects and quantifies the composition of the cluttered scene.

6.2 Computational model of text

Type design is the beginning of letters. Considering the distinctiveness and uniformity, type designers constrain the shape of individual letters within a font by constraining size proportions, so that they are related in weight, contrast, and stress, or the axis, of the letter. The proportions vary somewhat among fonts, but they are within a restricted range which makes the fonts of the same point size appear larger or smaller. These regularity effects within a font imply the shared properties or commonalities in addition to the distinctiveness among letters. These commonalities in the shape, proportions and other stylistic attributes of letters within a font, lead to the uniformity. This suggests that letters forming text are related in weight, contrast, proportions, and they also share some attributes. Thinking of the space averaging property of the crowding effect, the commonalities among letters can be captured over pooling the regions explicitly to some extent.

Learning from the computational model in letter perception, a letter mainly consists of lines of different orientation and curvature, and two-level interactive-activation computational models are used to describe them. This suggests that the straight line plays a key role in letter perception. And we also pay attention to the straight line in letters in our task.

Additionally, in Ergonomics, text has two kinds of characteristics: properties for letter form, and properties of organization among letters. Since properties for letter form determines the form is a letter in perception, it definitely plays a significant role in the discrimination of a single letter. For properties of organization, it includes letter spacing, word spacing and line spacing. It especially corresponds to the space regularity of letters which leads to the crowding effect and summation of subunits in a receptive field. Therefore, these organizational properties should be quantified and computed in our task.

Summarising the properties in different aspects mentioned above, text can be described in three level computation models. Illustrated in Figure 6.2, the first level of representation consists of the features of letters, which correspond to the properties of the physical appearance of type but in an image-based representation. The second level consists of letter-centred attributes, that is, the relations among letters over both space locations and physical appearances in the image, such as letter spacing or letters neighbourhood. And the third level consists of a word-centred description, i.e. considering the space roles at play in the global image; we can compute it over the GM regions since text usually appears as one or several clusters in the GM regions.

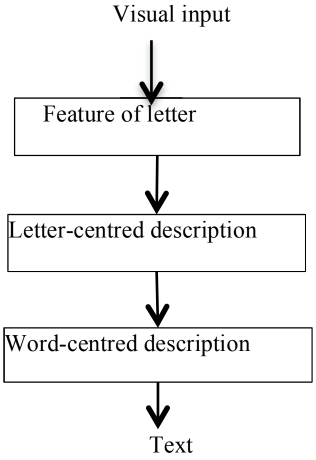


Figure 6.2 Schematic representation of a computational model of text

6.2.1 Feature of letters

As we know from chapter 3, the functional properties of a letter, which relate to its physical appearance include letter shape, line, size (width and height), weight, and proportion for the purposes of legibility. Given an image, these properties in Ergonomics can be translated in image-based attributes. The width and height and the related proportions of a letter can be used in a similar fashion to those in Ergonomics, thus we focus on those properties which we need to specially process in image.

6.2.1.1 Letter shape

Letter shape refers to the contour or form of a letter, or the inked space occupied by the letter. Therefore, it can be captured by grey space patch—region Q defined in 5.1.2.2, or formed by the edge point (i.e. boundary point).

Actually, the boundary of one thing with another is of the nature of a mathematical line, but not of a drawn line, because the end of one colour is the beginning of another colour—the boundary is a thing invisible. Image edge points depend on the detection algorithm. Indeed, edge detection is arguably the most important operation in low-level computer vision with its plethora of techniques.

However, the optimality of an edge detector can only be assessed in the context of a well-defined task. That is, the quality of an edge map is directly related to the amount of supportive information it carries into the subsequent processing stages. Since this information is extracted after the edge map is generated, a measure of confidence should be associated with the bottom-up information stream. Then, a task dependent top-down process can confirm (or discard) the hypotheses arising during the execution of the task and, thus, improve the overall performance. We will deal with one edge algorithm embedded in the confidence measure in the next chapter.

Based on the confident edge points, the letter form can be generated based on the neighbourhood measurement. And we will deal with it in the next chapter.

6.2.1.2 Pixel density and contour density

Weight can be obtained by the number of the "inked" pixels in a region$Q$. Considering the commonalities among letters, pixel density—the number of pixels per unit area, is defined to correspond with weight. Given a region$Q$with the number of pixelsN_{s_i}in grey level$s$, it has a tight rectangular boundary with the width $W$and height$H$, and then the pixel density$Density_{pixel}$ is defined as follows

$$Density_{pixel} = N_{s\_i} / (W \times H) \tag{6.8}$$

Additionally, if there is a number of an edge point $N_{con\_s\_i}$ in the region$Q$, contour point density $Density_{con_pixel}$is also defined as a variable related to weight, and is computed as

$$Density_{con\_pixel} = N_{con\_s\_i} / (W \times H) \tag{6.9}$$

Since font of the same typeface can have different weights, it means that there are somewhat intra-class differences, and we can tell the subtle differences of letters in the same category. And we will deal with them in an application in chapter 6.

6.2.1.3Line and line ratios

Line Although a letter mainly consists of lines of different orientation and curvature, text arbitrary sizes and complex background set big uncertainty on the length of lines, so a straight line ratio is defined to correspond to this property.

Based on the edge points with directions, edge in one orientation can be computed by connecting the edge points in corresponding directions. According to Kirsch, the horizontal line is composed of the edge points either in direction 2(north) or direction 6(south); the vertical line is formed by the edge points either in direction 0(east) or direction 4(west). A 45° straight line consists of the edge points either in direction 3(Northwest) or direction 7(southeast), and a 135° straight line is formed by the edge points either in direction 1(Northeast) or direction 5(Southwest). And we can connect the corresponding edge points to straight lines in eight directions respectively based on neighbourhood measurement as follows:

Let z enumerate the eight directions: $z=\{0,1,2,3,4,5,6,7\}$ which corresponds to 0^0, 45°, 90°, 135°, 180°, 225°, 270° and 315° respectively, edge map E is a mapping $E: Dir \subset R^2 \to z$. straight lines are well defined on edge map if:

1. z is ordered and $z=\{0,1,2,3,4,5,6,7\}$is considered, the straight lines can be defined on the edge map.
2. An adjacency (neighbourhood) relation $A_{dir} \subset Dir \times Dir$ is defined. In this paper 8 neighbourhoods are used, i.e., $p,q \in Dir$ are similar and adjacent (pAq)iff $\sum_{i=1}^{d} |p_i - q_i| \leq 1$, and p,q has the same direction.

Straight line L is a contiguous subset of Dir,i.e. for each $p,q \in L$there is a sequence $p,a_1,a_2,\ldots,a_n,q$and $pA_{dir}a_1, a_iA_{dir}a_{i+1}, a_nA_{dir}q$. Let N_{z_i}be the number of the sequence of edge points, i.e. the length of the line segment L.

Line ratios are defined over both horizontal orientation and vertical orientation in the region U given by the tightly rectangular boundary box of a letter since both horizontal line and vertical line are salient structure features of a letter in letter perception. The ratios are referred to the horizontal line length to the width of a letter RSL_H, and the vertical line length to the height of a letter RSL_V, and defined as follows:

$$RSL_H = \frac{\max_U(N_{2\_i}, N_{6\_i})}{Width \quad of \quad letter} \qquad RSL_V = \frac{\max_U(N_{0\_i}, N_{4\_i})}{Height \quad of \quad letter} \tag{6.10}$$

Note that the region U given by the tightly rectangular boundary box of a letter can be composed of the grey space patches Q, or be obtained by the letter shape.

6.2.2 Letter-centred features

For the purpose of readability, letter spacing, word spacing and interline spacing are in harmony with each other and organized as a whole text. And all of them are highly related to the type size. Here letter-centred features refer to the space relations among letters derived from similarity and proximity, i.e. adjacent. Considering the type size, i.e., the width and height of a letter, we deal with the relations among letters.

Given letters $letter1, letter2, \ldots, letterk$ and $letter2$, their tightly bounding boxes B_1, B_2, …, B_k are in width $W_1, W_2, \ldots, W_k$ and height $H_1, H_2, \ldots, H_k$ respectively. Relations among the letters are defined over the regions occupied by the bounding boxes as follows:

An adjacency (neighbourhood) relation A_L is defined over boundary boxes. In this paper, the size of the boundary boxes, and distances between the bounding boxes of letters are used. $(B_i A_L B_j)$ if their appearances are similar and their positions are near in space on these conditions:

i. $\Delta H_{ij} \le C_1 \max(H_i, H_j)$ and $\Delta Hij \le C_2 \min(H_i, H_j)$, $\Delta H_{i,j} = |H_i - H_j|$, C_1, C_2 are constants.

ii. Horizontal distance between the two letters $\Delta Space\_x < C_3 \max(H_i, H_j)$, here C_3 is a constant. Additionally, in the vertical direction, the y-coordinates of the two letters have common parts.

Letter strings *Str* is a contiguous subset of the boundary boxes occupied by letters, i.e. for each $B_i, B_k \in Str$ there is a sequence of $B_i, b1, b2, \ldots, b_n, B_k$ and $B_i A_L b_1, b_i \; A_L b_{i+1}, b_n A_L B_k$. Let N_{str_i} be the number of the contiguous bounding boxes of letters, i.e. the length of the string *Str*.

Note that, as it's well known, when nature needs a proportion to relate things and to provide order *on any scale*, it tends to use the golden ratio. And moreover, letter spacing is recommended to range from 25 to 40 percent of letter height, so we also use the golden ratio and practical recommendation to default set $C_1 = 0.618,\ C_2 = 0.382, C_3 = 0.4$.

6.2.3 Word-centred description

Corresponding with the word spacing and line spacing, word-centred description refers to the space regularity in global level, i.e. the cluster attributes in space. Since text usually is one or several clusters in GM regions, we define the word-centred attributes over GM regions. Over GM regions, there might be one or several clusters. Word-centred attributes include the statistical properties of each cluster and the relations of their space location.

Given a cluster $\Omega_l = \{Q_{p,}Q_{c1}, Q_{c2}, \ldots Q_n, Q_q\}$ in GM regions Ψ_{GM}, $N_{gm\_c\_l}$ denotes the number of the entire component space patches. And the statistical properties the cluster Ω_l are composed of average width $\overline{W}$, average height $\overline{H}$, average stroke width $\overline{W}_{stroke}$, average direction points $\overline{N}_{z\_i}(z = \{0,1,2,3,4,5,6,7\})$ and standard differences σ_W, σ_H, σ_{stroke}, σ_z, $(z = 0,1,2,3,4,5,6,7)$ respectively as thus:

$$\overline{W} = \frac{1}{N_{gm\_c\_l}} \sum_{i=1}^{N_{gm\_c\_l}} W_i, \qquad \sigma_W = \sqrt{\frac{1}{N_{gm\_c\_l}} \sum_{i=1}^{N_{gm\_c\_l}} \left(W_i - \overline{W}\right)^2}$$

$$\overline{H} = \frac{1}{N_{gm\_c\_l}} \sum_{i=1}^{N_{gm\_c\_l}} H_i, \qquad \sigma_H = \sqrt{\frac{1}{N_{gm\_c\_l}} \sum_{i=1}^{N_{gm\_c\_l}} \left(H_i - \overline{H}\right)^2}$$

$$\overline{W}_{stroke} = \frac{1}{N_{gm\_c\_l}} \sum_{i=1}^{N_{gm\_c\_l}} W_{stroke\_i}, \qquad \sigma_{stroke} = \sqrt{\frac{1}{N_{gm\_c\_l}} \sum_{i=1}^{N_{gm\_c\_l}} \left(W_{stroke\_i} - \overline{W}_{stroke}\right)^2}$$

$$\overline{N}_{z\_i} = \frac{1}{N_{gm\_c\_l}} \sum_{i=1}^{N_{gm\_c\_l}} N_{z\_i}, \qquad \sigma_z = \sqrt{\frac{1}{N_{gm\_c\_l}} \sum_{i=1}^{N_{gm\_c\_l}} \left(N_{z\_i} - \overline{N}_{z\_i}\right)^2} \qquad (6.11)$$

The relationships among clusters can be represented pictorially by the Venn diagram, in which sets are represented as the interiors of an overlapping circle (or other plane figures). Set combinations are represented by areas bounded by the circles, as shown in the following example for two clusters Ω_l, Ω_m, illustrated in Figure 6.3:

i. Intersect: $\Omega_l \cap \Omega_m = x$;

ii. Disjoint: $\Omega_l \cap \Omega_m = \phi, \quad \Omega_l \cup \Omega_m = n(\Omega_l) + n(\Omega_m)$;

iii. Subset: $\Omega_m \cap \Omega_l = \Omega_l, \Omega_l \cup \Omega_m = \Omega_m$.

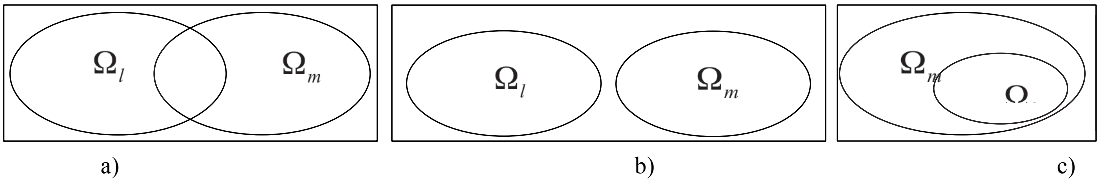


Figure 6.3 relationships among clusters over GM regions: a) intersect; b) Disjoint; c) Subset

6.3 Summary

Clutter is the state in which the organization of visual items can cause crowding and lead to a degradation of performance in some task. This definition of clutter brings up two key points: the association between clutter and the representation or organization of information, and the notion that clutter may depend upon the user's task. This presents the two corresponding questions. One is how the clutter scene images are represented by the spatial layout or space organization. And the other question is how to express the text in the clutter image, although there is another representation from the viewpoint of Ergonomics.

And the investigations of crowding in chapter1 provide important insight into the essential process of crowding as follows: 1) image features are pooled within receptive fields of increasing size, and 2) crowding can be broken by grouping or crowding of crowding. Clearly, what we need to specify, of course, is the region over which the statistical property of crowding is computed and un-crowding is realized by a grouping based on similarity.

Either with clutter or crowding, what we need to do is provide a reasonable representation of their spatial element and quantify their relations in space, that is, their neighbourhood together with similarity depends on our text detection work.

According to the theory of perspective, when objects recede from the eye or camera, the size of the object decreases. This means that the space of the object in the image is reduced while the contour is lessened. If the distance is far away enough, the contours or boundary of the separated object disappears and the separate objects are merged into a whole. This suggests that there are two types of spatial elements: space occupied by object and the contour or edge of the object.

Contour will disappear when the space occupied is too small, and it can be distorted by crowding or clutter, while the space occupied by the object exists and has discernible information. That is what impressionists do.

Especially, impressionists have realized an object does not have its own colour, but many individual patches of colour, so they use directional brushstroke or colour patches, which are small space patches in space regularity, to represent visual object "formless" instead of the clear contour shape sketch. Learning from the painters, we can use an image-based method--connected component analysis to represent the space patches forming an object. And spatial organization is explored among these connected component space patches.

This concept can be explained informally as follows. Imagine all possible gray patches in an image, they are arranged in a constitutional way to generate textures or form, which looks like the preconceptions about "tree" or letters or "windows" or "sky". If we are shown an object in an image, instead of rendering solid objects, many individual gray patches work together in harmony to make up a form, i.e., the form of the object is the harmonic combination of grey patches.

What is harmony? Harmony is only produced when proportions of things are seen or heard simultaneously [55]. According to Leonardo Da Vinci, proportion is in all things, it is not only found in numbers and measurements but also in sounds, spaces and in whatsoever power there maybe. Harmony is composed of the union of its proportional parts reacted (e.g. seen or sounded) simultaneously. And a feeling of beauty in the human being is born from these harmonious proportions. For visual objects in image, their component gray patches are also made to react simultaneously and can be seen at one and the same time both together and separately, their geometry and proportion are keys in creating a structure and their proportion should reflect the object in the real world.

In mathematics, the geometric mean (GM) can capture those ratios to the reference value. Thus using it among grey patches, the proportions of component parts to the whole image can be captured implicitly, and also GM can capture a single "figure of merit" in an object. Therefore, GM can be an indicator of "figure of merit" in an image which has close relations to the proportion of component parts. Since GM can indicate the “figure of merit” among proportional component parts, GM regions get involved in object constitution.

If several kinds of grey level have similar proportions or rhythm in an image, they have the same GM value, and they may form the same object, or the same type of objects which bring to the human being the feeling of beauty. So those grey patches with a similar GM value are composed of GM regions. The formal definition of the GM regions concept and the necessary auxiliary definitions are given in Table 1.

According to the Gaussian subunits solution and the associated field, the receptive field must show linear spatial summation. And those GM regions, which are generated by neighbourhood proximity and similarity, can be considered as our pooling regions over which statistics are computed.

Over GM regions, statistics features are defined and computed, including RMS contrast, orientation, direction density, and edge density. And neighbourhood is measured, together with similarity, adjacent relations are defined and clusters are obtained over GM regions. Thus, a space-averaged image representation is built up.

Moreover, to summarise the properties in different aspects investigated in type design, letter perception and computation, text can be described in terms of a three-level computation model. The first level of representation consists of the features of a letter, which correspond to the properties of the physical appearance of type but in image-based representation, such as letter shape, weight and pixel density, line and straight line ratios. The second level consists of letter-centred attributes, that is, relations among letters over both space locations and physical appearances in image– letter spacing or the letters neighbourhood. And the third level consists of a word-centred description, which considers the space roles of word playing in the global image. We also define the space relations over clusters since text is generally in one or several clusters in the GM regions.

The space averaged representations of image in the clutter scene and text model provide us with the insights of the computational features of text, and give us clear instructions for the algorithm of text extraction from the clutter scene.

Chapter 7

Text Detection Algorithm Based on the Space Averaged Crowding Model

The space averaged representations of image in the clutter scene and text model built in Chapter 4, provide us with insights into the computational features for text, and give us clear instructions for the algorithm of text extraction from the clutter scene.

Guided by these representations of image and text in the cluttered scene, instead of only focusing on the features of text, and considering the statistical properties of clutter and crowding, we propose a framework of text detection by integrating the properties of clutter/crowding into text features. Based on the quantification of the space organization of the clutter scene, a highly perceptive solution of text extraction is developed over GM regions and the three-level text model. This chapter will present details of the proposed algorithm, including feature extraction, component extraction, GM regions generation and GM regions analysis and text location.

7.1 Methodology

Based on the space averaged representation of image and the text three-level model, we propose a new framework to extract text strings with multiple sizes and colours from the cluttered scene. Illustrated in Figure7.1, the flowchart of our framework consists of three main parts: a) image partition to find text character candidates based on confident edge map and non-overlapped grey space patches. In this part, guided by the representation of image proposed in Chapter 4, image is represented by a set of GM regions with clutter features, meanwhile the image also is partitioned into a set of closed regions based on the gradient-based method, and post processing is then performed to remove the GM regions and closed regions. b) Attributes are computed over GM regions, including the general features of image and text features. Meanwhile, space adjacent relations among close regions are analysed in this part. c) In this part, the saliency of GM regions are analysed in each GM level. For our task, the GM level provides the granularity of the space element of the image.

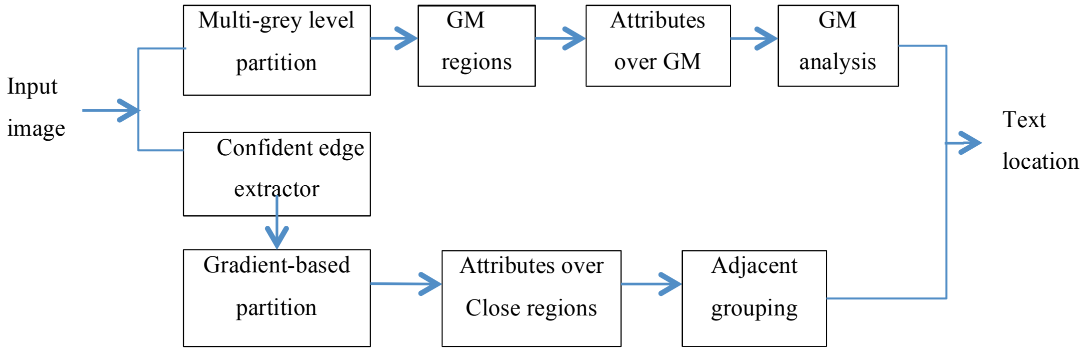


Figure 7.1 the flowchart of the proposed framework of text detection

Accordingly, for the large format text, even formed by only one character, it is still salient enough and detected by text structure features. For small text, according to the three aspects of perspective, when the text is too small, the boundary and colours of letters disappear, only the sheet resulting from the string of small letters can be discriminated. The investigation of text legibility and readability suggests to us that the measurement of small text is less than 5 pixels height. And the large format size is bigger than 72 points, being approximately 66 pixels height.

In each GM level, three-level features of text together with space-averaged image statistical features over the GM regions are used to GM analyse or learn by SVM to determine the text locations. And then with the cross-validation of close regions, the text regions are inferred.

The proposed framework is able to effectively detect text strings in arbitrary locations, sizes, orientations, colours and slight variations of illumination or the shape of the attachment surface. Compared with existing methods which focus on independent analysis of a single character, the text string structure is more robust in terms of distinguishing background interferences from text information. Experiments demonstrate that our framework outperforms the state-of the-art Robust Reading Dataset.

Overall, the algorithm introduced in this chapter offers the following primary contributions to robust detection of text with variations of scale, colour, and orientation from the clutter scene:

Starting from the high perception, instead of focusing on text features and considering the statistical properties of the clutter and the crowding based on the quantified space regularity, we explore the composition of visual objects on image and infer the patches of text from a perceptual view. Because the prosed method does not reply on heavy classifier training, it is not tangled with the constraints in legend solutions. Moreover, the proposed method, owing to its analysis of the composition of visual objects and the statistics of clutter instead of subtle local features, shows robustness in a clutter environment. This chapter contributes in the following aspects:

- Most existing work of text detection from clutter scene images only focuses on features of text, usually ignoring the features of the cluttered scene. We propose a new framework to robustly detect text strings with variations of orientation and scale from the cluttered scene by integrating three-level features of text based on clutter or crowding statistics properties, introduced in chapter 6.

- We formally quantify the space organization or the constitutional proportions of component parts of the visual objects, especially the composition of text. Therefore, a highly perceptive solution is proposed to analyse the visual objects in the image.

- We model text by combining the three-level features of text and the clutter statistic properties. Three-level features of text consist of features of letter, letter-centred space relations (the organization of letters), and word-centred attributes (i.e. attributes over GM regions). And the statistics properties of the clutter scene are computed over GM regions. Under this model, we develop an algorithm based on the GM partition and gradient-based partition to compute connected components of candidate characters. It is more robust and achieves better results than only using one of them.

After this, we deal with the detail in the following parts.

7.2 Image partition

According to the representation of image in the clutter scene, image partition is firstly performed to group together the adjacent pixels that belong to the same character, obtaining a map of space patches as candidates of text. Based on local gradient features and the uniform grey value of text characters, we design a gradient-based partition algorithm and a multi-grey level partition algorithm respectively.

7.2.1 Multi-grey connected component (MGCC)

The space patches are defined solely by incorporating all the similar grey level pixels in the neighbourhood of the currently selected pixel. And the whole image is represented as a set of grey patches in the multi-grey level. And among them, geometric mean is captured and GM regions are generated as pooling regions of crowding effect. The definition has been introduced in Chapter 4. Now the details are implemented by connected component analysis (CCA).

Image I is a mapping $I:\ D \subset R^2 \rightarrow S,\ S = \{0,1,\ldots,255\}$, Multi-grey level space patches/regions are well generated on images by the adjacent grouping. An adjacency (neighbourhood) relation $A \subset D \times D$ is defined. Here, 8-neighbourhoods are used, i.e. $p,q \in D$ are similar and adjacent (pAq) iff $\sum_{i=1}^{d} |p_i - q_i| \leq 1$, and p,q have the similar grey level s.

Region Q is a contiguous subset of D, i.e. for each $p,q \in Q$ there is a sequence p, a_1, a_2, $\ldots$, a_n, q and $pAa_1, a_iAa_{i+1}, a_nAq$. Let N_{s_i} be the number of the sequence points. Let $\Lambda_S = \{Q_1, Q_2, \ldots, Q_{sg}\}$, $s \in S$ be a set of regions in grey level s, and sg denote the number of regions in this grey level . Thus, for a 8-bit grey images (i.e. $0 \leq s \leq 255$), $I_{m\times n}$, its multi-grey Connected Components (MGCC) are described as below,

$$I_{m\times n} = \bigcup_{j=0}^{255} \Lambda s\_j = \bigcup_{j=0}^{255} \bigcup_{i=0}^{sg_j} Q_{j,i} \tag{7.1}$$

Based on the basic MGCC image segmentation, GM regions are obtained in the following way. Among the set of regions Λ_s, the number of the sequence points N_{s_i} is select as the flag of each region, and the geometric mean value $GM_{\Lambda s}$ is computed as $GM_{\Lambda_s} = \left(\prod_{s\_i=1}^{sg} N_{s\_i} \right) = \sqrt[sg]{N_1 N_2 \cdots N_{sg}}$. And $GM_{\Lambda s}$ becomes an indicator of the Λ_s.

Then, let $\Psi_{GM} = \{\Lambda_{s1}, \Lambda_{s2}, \ldots, \Lambda_{sk}\}$ be a set of regions with similar geometric mean, called GM regions, and Λ_{sj} with the number sg_j of space patches. Since there are similar geometric means, a set of values of geometric mean exist in the image, let

$GM = \{GM_1, GM_2, GM_3, \ldots, GM_r\}$denote it, and the number of geometric mean value is equal to r. Then, the image I can be represented as the set of GM regions:

$$I = \bigcup_{i=1}^{r} \Psi_{GMi} \tag{7.2}$$

7.2.2 Gradient-based partition

Generally, an edge map is associated with closed and infinite curves, and image can be partitioned into a set of regions given by these close curves. The increase in the saliency of closed curves is often considered desirable because these curves are usually more significant than their open counterparts with the same length. Indeed, closed curves are considered more salient by the HVS [8].

Similarly, for our task, un-crowding requires our target to be conspicuous through both salient structure and grouping according to space regularity. Thus, beside the straight line, the form or shape of character becomes another salient feature in feature level, and it will affect the letter-centred features. We need to select a confident edge detector. Then based on these confident edge points, the contour of the interested regions will be generated.

Usually, the close curves are considered as salient structure. Therefore, we call the regions given by contours as close regions (CRs). And image can be partitioned into a set of CRs.

7.2.2.1 Edge map with embedded confidence

Local contrast in a 2D image corresponds to discontinuities in depth or in surface orientations of 3D objects, changes in material properties, and variations in scene illumination. This can be captured by edge detection. Edge detection converts a 2D image into a set of curves. However, the optimality of an edge detector can only be assessed in the context of a well-defined task. That is, the quality of the edge map is directly related to the amount of supportive information it carries into the subsequent processing stages. Since this information is extracted after the edge map is generated, a measure of confidence should be associated with the bottom-up information stream. The paper

[Pami2001] defined a confidence measure by using information which inherently existed in the regular sampling lattice and was not employed in the computation of the gradient magnitude. The paper then proposed an edge detection approach.

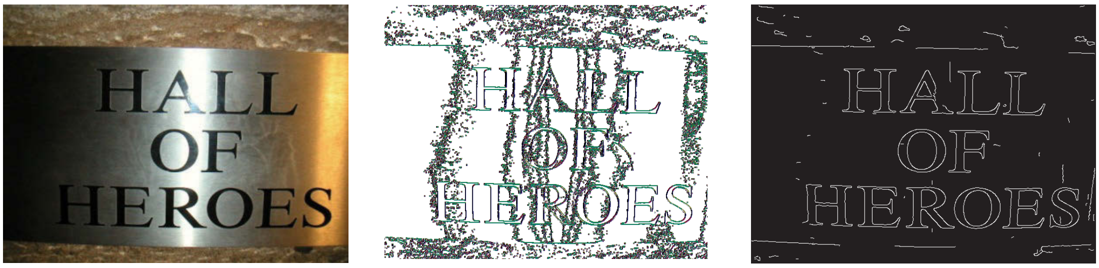
Based on a 5×5 gradient operator, for edge detection embedded confidence, the data is weighted with binomial weights and the simplest local structure model is assumed. The two sequences are

$$\begin{aligned} s(i) &= h_K(i;0,0) = [0.0625, 0.25, 0.375, 0.25, 0.0625]^{\mathrm{T}} \\ d(j) &= h_K(j;1,1) = [-0.125, -0.25, 0, 0.25, 0.125]^{\mathrm{T}} \end{aligned} \tag{7.3}$$

yielding the masks

$$\mathbf{W}_{dx} = \mathbf{W} = \begin{bmatrix} -0.0078 & -0.0156 & 0 & 0.0156 & 0.0078 \\ -0.0312 & -0.0625 & 0 & 0.0625 & 0.0312 \\ -0.0469 & -0.0938 & 0 & 0.0938 & 0.0469 \\ -0.0312 & -0.0625 & 0 & 0.0625 & 0.0312 \\ -0.0078 & -0.0156 & 0 & 0.0156 & 0.0078 \end{bmatrix} \tag{7.4}$$

$$\mathbf{W}_{dy} = \mathbf{W}^{\mathrm{T}}$$



b) b) c)

Figure 7.2 a) Original image. b) Traditional (Sobel) edge map.
c) Edge map with embedded confidence

After gradient estimation, every pixel in the image is associated with an edge (gradient) magnitude $\hat{g}$ and an edge orientation $\hat{\theta}_e$. Then, after performing hysteresis thresholding, the confident edge points are obtained.

Since this edge detector embedded in confidence fills in most missed corners in the hysteresis thresholding step, for example, the edge map shown in Figure 7.2c), it can help us to obtain the contour or shape to the object and give us access to the global information.

7.2.2.2 Shape generation based on edge points

A “point in polygon” algorithm with embedded confidence from computational geometry, is used to determine if the point is inside or outside of the polygon and the coordinate axes. Moreover, it fills in most missed corners. Thus, edge points determined by polygonal contours, which are modulated by both the magnitude and orientation of gradient, are reliable for generating the contours of an object.

Edge Image I_E is a mapping $I_E : D \subset R^2 \to S_E$. cc space patches/regions are well defined on images if:

3. S_E is binary value of edge points $S_E = \{0,1\}$, the regions can be defined on the edge point images.
4. An adjacency (neighborhood) relation $A_E \subset D \times D$ is defined. In this paper 8-neighbourhoods are used, i.e. $p,q \in D$ are similar and adjacent (pA_Eq) iff $\sum_{i=1}^{d}|p_i - q_i| \le 1$, and p,q has the similar value s.

Shape & Close Region Q is a contiguous subset of D, i.e. for each $p,q \in Q$ there is a sequence $p, a_1, a_2, \dots, a_n, q$ and $pA_Ea_1, a_iA_Ea_{i+1}, a_nA_Eq$. Let N_{se_i} be the number of the sequence points.

Let $\Lambda_e = \{Q_1, Q_2, \dots, Q_{M_e}\}$ be a set of regions. The number of regions is M_e. Then, the image I can be represented as the set of CR regions $I = \Lambda_e$.

Table 7.1 Contour generation algorithm and related definitions used in following sections

Based on these confident edge points, contours are generated by grouping together the edge points that belong to the same object. And the regions given by these contours are also interested by the vision task and defined as close regions since the majority of meaningful visual objects have close contours.

The contours are captured by the adjacent grouping together of the confident edge points, shown in Figure 7.3, e.g. meaningful contours are captured, and the isolated long horizontal lines are discarded. The close regions given by contours are interested and the image can be represented as a set of close regions. The contour generation algorithm is defined in Table 7.1 in detail.

Further, space relations (neighbourhood) among CRs are measured as those that are computed among grey space patches in Chapter 6. And based on the adjacent grouping algorithm which was defined in section 6.1.2.2, clusters are also formed. Similarly, the statistic properties of clutter or crowding are computed over them. Therefore, the features computed over CRs consist of the physical measures and the statistical properties of the tight rectangular boundary box, such as width, height, aspect ratio, mean intensity, local RMS contrast, and neighbourhood. Their computations are the same as the calculation over the regions given by MGCCs.

After image partitioning into a set of regions, the statistical properties of image in the cluttered scene and three-level features of text are computed over these regions.

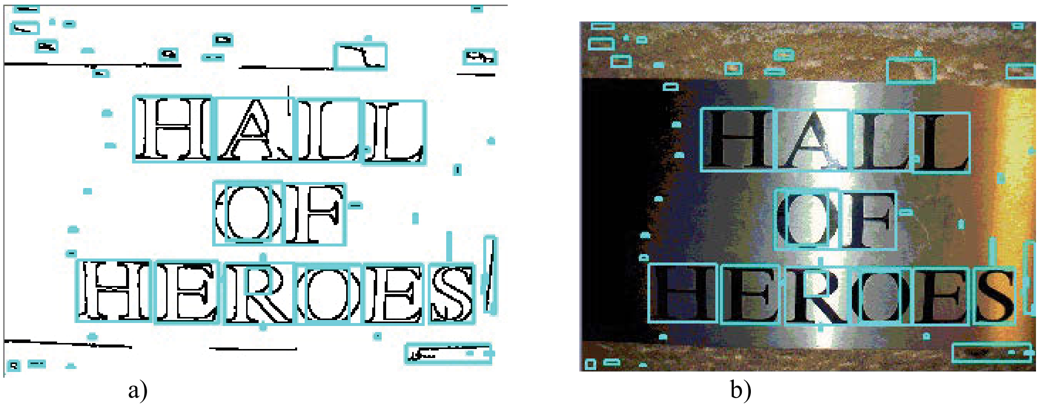


Figure 7.3 Close regions generated by adjacency grouping confident edge points. a) Edge point grouping. b) Regions given by contours

7.3 Features extraction

Both the statistical properties of image in the clutter scene and features of text are computed over regions given by MGCCs and CRs, and extended in the GM levels. We deal with them in the following sections.

7.3.1 Basic features of physical appearance

The basic features of the physical appearance of a region consist of the attributes of its tight boundary box (i.e., width, height, area, centroid, and aspect ratio), density of pixels, stroke width, orientation and density of directions, straight line ratio, mean intensity and local contrast.

Without the loss of generality, given a region Q_i formed by a MGCC, its tight rectangular boundary box is measured by width $W_{s,i}$ and height $H_{s,i}$. And let $N_{s\_con,i}$ denote the number of edge pixels in the region, and $N_{s,i}$ denote the connected pixels of it. Note that, Kirsch operator is adopted to extract the edge pixels in a region given by MGCC. To better adapt the variance of input images, the threshold inside Kirsch operator is well adjusted in order to assure enough edge points are identified inside Connected Components. The basic features of physical appearance are defined over them.

There may be a wide variety of shapes of the region Q_i. The size of it is determined by the width $W_{s,i}$ and the height $H_{s,i}$ of its rectangular tight boundary box. That is,

$$\begin{aligned} W_{s,i} &= x_{\max} - x_{\min} \\ H_{s,i} &= y_{\max} - y_{\min} \end{aligned} \tag{7.5}$$

where $x_{\max}, x_{\min}, y_{\max}, y_{\min}$ is the maximum and minimum coordinates in $x-$axis and $y-$axis respectively . Thus, the area of Q_i is defined as

$$area_{s,i} = W_{s,i} \times H_{s,i} \tag{7.6}$$

And the aspect ratio of the shape of Q_i is followed as

$$Aspect\_ratio_{s,i} = \frac{W_{s,i}}{H_{s,i}} \tag{7.7}$$

The centroid of the region Q_i is computed as

$$CC\_centroid_{s,i} = (x_c, y_c) = \left(\frac{1}{N_{s,i}} \sum_{j=1}^{N_{s,i}} x_j, \frac{1}{N_{s\_i}} \sum_{j=1}^{N_{s,i}} y_j \right) \tag{7.8}$$

Inside CC, the density of mass is considered as an important property to measure 2D visual patches. And edge density-the amount of edge point per area, is one important measure of clutter meanwhile the direction density is one key measure of the closure of a symmetric object. Thus, we compute these densities as:

$$D_{s,i}=\frac{N_{s,i}}{W_{s,i}\times H_{s,i}}\quad D_{s\_con,i}=\frac{N_{s\_con,i}}{W_{s,i}\times H_{s,i}}\quad D_{s\_dir,i}=\frac{N_{s\_dir,i}}{W_{s,i}\times H_{s,i}}\tag{7.9}$$

Where $N_{s_dir,i}$is the number of the edge points with the given direction *dir*. The direction of edge points is described statistically on a 8-bin Histogram of edge Direction (HoD). The number $N_{s_con,i}$of edge points is quantized onto eight directions $HoD(\varphi)_{s,i}$, let$HoD(\varphi)_{s,i}=N_{s_con,i}$,whose directions are closed to$\varphi\in\{0^{o},45^{o},90^{o},K\ ,315^{o}\}$. Simply, let $dir=\{0,1,2,3,4,5,6,7\}$ correspond to the directions respectively, and$N_{s_dir,i}$stands for the number of edge pixels in each direction.

As discussed in 5.2.1.3, since the straight line is a critical determinative features in letter perception, we have defined the straight ratio to capture it over regions. And the straight lines are extracted by the adjacent grouping algorithm. Here, an edge line inQ_i is considered as a line with at least two edge pixels consecutively linked together, and all these edge pixels are on the same edge direction dir. When$dir\in\{2,6\}$, it is horizontal line$lh_{s,i}$. When$dir\in\{0,4\}$, it is horizontal line$lv_{s,i}$. The ratio of the max length of straight line to the width or the height of the tight rectangular boundary box of Q_i, $RSL_{s,i}$ is defined as:

$$RSL_{s,i}=\{RSL_H,RSL_V\},\quad RSL_H=\frac{\max(lv_{s,i})}{H_{s,i}},\quad RSL_V=\frac{\max(lh_{s,i})}{W_{s,i}}\tag{7.10}$$

where $\max(\cdot)$stands for the maximum length of the line.

In addition, the stroke width is an important feature of character which can separate text from other elements of a scene [25]. In our work, the stroke width$sw_{s_i}$is also adopted along with other newly defined features in regionQ_i.

Then, consider that one of statistics properties of clutter is local RMS contrast, a wide variety of physical constituents contribute to the distribution of local luminance and contrast. Local luminance and contrast are measured in image patches formed by windowing with a circularly symmetric raised cosine weighting function. Given regionQ_i, the cosine weighting function is as thus:

$$w_j = 0.5\left(\cos\left(\frac{\pi}{r}\sqrt{\left(x_j - x_c\right)^2 + \left(y_j - y_c\right)^2}\right) + 1\right), \tag{7.11}$$

Where r is the patch radius, let $r = 0.5 * \min(W_{s,i}, H_{s,i})$. $\left(x_j, y_j\right)$Is the location of the jth pixel in the patch, and $\left(x_c, y_c\right)$is the location of the centre of the patch Q_i.

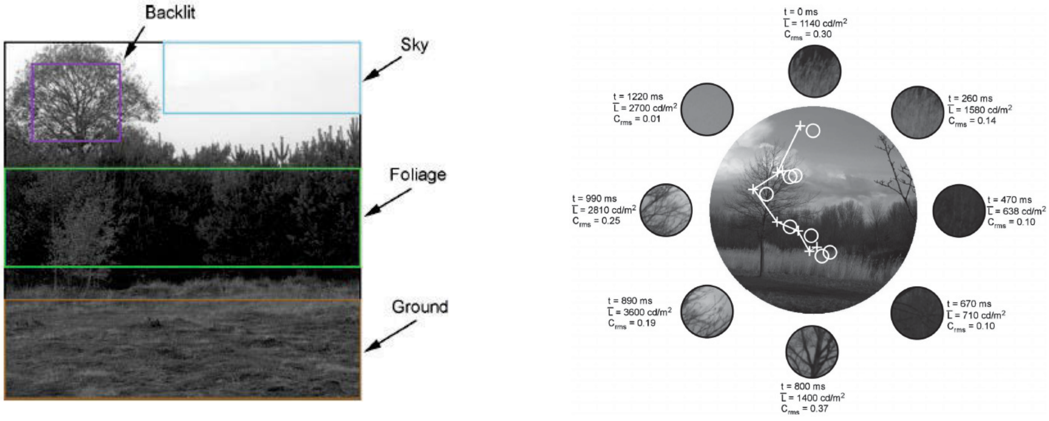
The local luminance and the root-mean-squared (RMS) contrast of each patch are measured through weight by the raised cosine window. The local luminance of a patch is defined by

$$\overline{L} = \frac{1}{\sum_{j=1}^{N} w_i} \sum_{j=1}^{N} w_j L_j, \tag{7.12}$$

Where N is the total number of pixels in the raised cosine window, L_j is the luminance of the jth pixel, and w_j is the weight of the raised cosine windowing function at the jth pixel. The RMS contrast of the patch is defined by

$$C_{rms} = \sqrt{\frac{1}{\sum_{j=1}^{N} w_j} \sum_{j=1}^{N} w_j \frac{\left(L_j - \overline{L}\right)^2}{\overline{L}^2}}. \tag{7.13}$$

According to [1], illustrated in Figure 7.4, the statistics result shows that the average RMS contrast in rural images is in the range 0.2–0.34 (depending on the analysis patch size), and the band-limited RMS contrast is in the range of 0.15–0.18 for rural images. The half-saturation contrast (c50) is expected to match the median contrast, which is in the range of 0.18–0.24, and it seems to be in reasonable agreement with the contrasts in natural scenes.


Backlit
Sky
Foliage
Ground
t = 0 ms
$\overline{L}$ = 1140 cd/m^2
C_{rms} = 0.30
t = 1220 ms
$\overline{L}$ = 2700 cd/m^2
C_{rms} = 0.01
t = 260 ms
$\overline{L}$ = 1580 cd/m^2
C_{rms} = 0.14
t = 990 ms
$\overline{L}$ = 2810 cd/m^2
C_{rms} = 0.25
t = 470 ms
$\overline{L}$ = 638 cd/m^2
C_{rms} = 0.10
t = 890 ms
$\overline{L}$ = 3600 cd/m^2
C_{rms} = 0.19
t = 670 ms
$\overline{L}$ = 710 cd/m^2
C_{rms} = 0.10
t = 800 ms
$\overline{L}$ = 1400 cd/m^2
C_{rms} = 0.37

a) b)

Figure 7.4 After[1] a) Example of hand segmentation of an image into regions containing "sky", "foliage", "ground", and "backlit foliage". b) Demonstration of the variation in contrast and luminance that might fall on a receptive field during a sequence of eye fixations. The circles show a receptive field at an arbitrary location relative to the fixation point.

7.3.2 The features of space relations among regions

The properties of space relations among regions correspond to the character organization and space regularity of crowding or clutter- grouping by similarity + proximity. Therefore, the proximity is defined over the distance between the two regions, and the similarity is still considered over the physical appearance features as follows.

Without the loss of generality, given any two regions Q_i and Q_j, their physical appearance are measured by the widths W_i, W_j and heights H_i, H_j of the tight rectangular boundary boxes respectively, and their positional change range along x-axis are $\left[x_{si}, x_{ei}\right]$ and $\left[x_{sj}, x_{ej}\right]$, meanwhile the positional change rang along y-axis are $\left[y_{si}, y_{ei}\right]$ and $\left[y_{sj}, y_{ej}\right]$. And their local mean intensities and local RMS contrast are $\overline{L}_i, RMS_i$ and $\overline{L}_j, RMS_j$.

Over the two dimensions of physical appearance and local RMS contrast, the similarity relation is measured by the Gaussian function

$$\mathrm{Re}\,lation = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right), \qquad \Delta = \left|\xi_i - \xi_j\right| \tag{7.14}$$

where ξ denotes the two dimensions of width, and height of the tight rectangular boundary box and one dimension of local RMS contrast, and σ is the Gaussian constant, it has three values corresponding to the three dimensions respectively. And regions Q_i and Q_j are similar in physical appearance iff $\mathrm{Re}lations \in 1\sigma$ confidence interval.

Over the position change, spaces among the two regions in horizontal direction $Space_H$ and vertical direction $Space_V$ can be computed. Consider that letter space is determined by letter height for readability, we measure the proximity in the reference to the height of the tight boundary box. Regions Q_i and Q_j are proximity iff $Space_H < \alpha \max(H_i, H_j)$ and $Space_V \leq 0$. Here α is a constant.

Thus, regions Q_i and Q_j are adjacent (neighbours) for each other iff Q_i and Q_j are similar and proximity. And the two regions can be connected or merged into one. Then with the adjacent measure, cluster is generated in GM regions.

7.3.3 Features over GM regions

Geometric mean can capture the proportion of the component parts of an object, and can find a single "figure of merit" for these parts. And GM regions get involved in object constitution, while image can be composed of the GM regions in different GM levels. As text has a highly salient structure, for the purpose of extracting the “figure of merit” or salient structure in the clutter scene, all features of region can be extended to GM-level.

Given GM regions $\Psi_{GM} = \{\Lambda_{s1}, \Lambda_{s2}, \ldots, \Lambda_{sk}\}$, $\Lambda_{si} = \{Q_1, Q_2, \ldots, Q_{M_{s,i}}\}$, and Q_j, $j = 1,2,\ldots,M_{s,i}$ stand for regions given by the corresponding MGCCs in the same grey value, features are computed over these regions as follows:

The average width and height of the regions as

$$W_\psi = \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{i=1}^{M_{s,i}} W_{s,i} \quad H_\psi = \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{i=1}^{M_{s,i}} W_{s,i}, \tag{7.15}$$

Where $W_{s,i}$, $H_{s,i}$ can be calculated according to Eq.(7.5).

The average aspect ratio of the shapes of regions in ψ_{GM}

$$Aspect\_ratio_{\psi} = \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{j=1}^{M_{s,i}} Aspect\_ratio_{s,j}, \tag{7.16}$$

where $Aspect\_ratio_{s,j}$ stands for the aspect ratio of the given region Q_j, and is calculated according to Eq. (7.7).

Inside ψ_{GM}, the average pixel density, average edge density and average edge directions will be used to measure the appearance of object and the clutter

$$\begin{aligned} D_{\psi_{s,i}} &= \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{j=1}^{M_{s,i}} D_{s\_j} \\ D_{\psi_{s\_con,i}} &= \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{j=1}^{M_{s,i}} D_{s\_con,j} \\ D_{\psi_{s\_dir,i}} &= \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{j=1}^{M_{s,i}} D_{s\_dir,j} \end{aligned} \tag{7.17}$$

Where $D_{s,j}, D_{s\_con,j}$ and $D_{s\_dir,j}$ stand for the pixel density, edge density and direction density of the given region Q_j with grey level s, which can be calculated according to Eq.(7.9).

The average straight line ratios are calculated as

$$RSL_{\psi} = \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{j=1}^{M_{s,i}} RSH_{s,j}, \tag{7.18}$$

Where $RSH_{s,j}$ denotes the straight line ratio of the region Q_j with grey level s, and can be obtained according to the Eq.(7.10)

And the average luminance and local RMS contrast are also calculated as

$$L_{\psi} = \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{j=1}^{M_{s,i}} \overline{L}_j, \qquad C_{\psi_{rms}} = \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{j=1}^{M_{s,i}} C_{rms,j} \tag{7.19}$$

where $\overline{L}_j$ denotes the local mean intensity in the region Q_j, and is calculated according to Eq.(7.12). $C_{rms,j}$ denotes the local RMS contrast in the region Q_j, which is computed according to Eq.(7.13).

Additionally, the average stroke width is as

$$sw_{\psi} = \frac{1}{\sum_{s=1}^{k} M_{s,i}} \sum_{s=1}^{k} \sum_{j=1}^{M_{s,i}} sw_j \tag{7.20}$$

7.4 Clusters analysis over GM and CRs

7.4.1 The composition of visual object over GM regions

In order to explore the semantic links across CCs, motivated by composition theory and the maximum entropy [294], a typical statistical model on the number of pixels in CCs is established to infer the composition of a visual object consisting of multiple CCs. Then, the CCs on the image obtained in Section 7.2.1 can be further organized. The CCs with similar composition complexity based on the newly proposed models will be grouped together. At the end of this section, it will be shown that such groupings will be close to human perception. The CCs belonging to the same object will be grouped together. In such a way, it will be easier to identify the object of interest on the image. It i not possible to have such semantic grouping if it is carried out based on the CCs in Section 7.2.1.

Given the grey level s, the number of pixels in the region Q_j is $N_{s,j}$. According to the Constitution Theory[294] on system complexity, Q_j is an element of a set $\Lambda_{si} = \{Q_1, Q_2, \ldots, Q_{M_{s,i}}\}$. $N_{s,j}$ will be considered as a random number following a typical statistical distribution. Its geometric mean $\hat{v}_s$ can be approximated as,

$$\hat{v}_s = \sqrt[M_{s,i}]{\prod_{j=1}^{M_{s,i}} N_{s,j}} \tag{7.21}$$

To simplify the computation, it usually calculates the logarithm of geometric mean

(GM) $\hat{v}_s' = \ln \hat{v}_s$. Thus,

$$\hat{v}_s' = C \sum_{j=1}^{M_{s,i}} N_{s,j}' \tag{7.22}$$

where $N_{s,j}' = \ln N_{s,j}$, and C is a constant.

In order to explore the statistics of random number $N_{s,j}$, it is presented in a more general term $z = N_{s,j}$. Its probability distribution function is noted as $\mathrm{f}(z)$. Accordingly, its logarithm of geometric mean v' is,

$$\begin{aligned} v' &= C \int_0^{\infty} \mathrm{f}(z) \ln z dz \\ \text{s.t.} &\int_0^{\infty} \mathrm{f}(z) dz = 1 \end{aligned} \tag{7.23}$$

In fact, the composition of CCs on a given grey level s can be seen as the composition of z on the same grey level. According to composition theory and maximum entropy [22], given a grey level s , the composition complexity of z can be presented by entropy:

$$H_s(z) = -\int_0^{\infty} \mathrm{f}(z) \ln \mathrm{f}(z) dz \tag{7.24}$$

To obtain (or maximize) $H_s(z)$, we represent Eq.(7.24) Using the Lagrange function conditioned on Eq.(7.23),

$$F_s(z, \lambda_1, \lambda_2) = -\int_0^{\infty} \mathrm{f}(z) \ln z dz + \lambda_1 \left(\int_0^{\infty} \mathrm{f}(z) dz - 1 \right) + \lambda_2 \left(\int_0^{\infty} \mathrm{f}(z) \ln z dz - \hat{v}_s' \right) \tag{7.25}$$

Further, λ_1, λ_2 can be regarded as constants given that the conditions of Eq. (7.23) are always satisfied. In order to maximize $\mathrm{f}(z)$, let $\frac{\partial F_s}{\partial z} = 0$,

$$\frac{\partial F_s}{\partial z} = -\mathrm{f}(z) \ln \mathrm{f}(z) + \lambda_1 \mathrm{f}(z) + \lambda_2 \mathrm{f}(z) \ln z = 0 \tag{7.26}$$

So we have

$$\ln \mathrm{f}(z) = \lambda_1 + \lambda_2 \ln z \tag{7.27}$$

then

$$f(z) = e^{\lambda_1} \cdot z^{\lambda_2} \tag{7.28}$$

Substitute Eq.(7.28) into Eq.(7.24), we have

$$H_s(z) = -\frac{\lambda_1 e^{\lambda_1} \cdot z^{\lambda_2 + 1}}{\lambda_2 + 1} - \lambda_2 \int_0^\infty \mathrm{f}(z) \ln(z) dz = -\frac{\lambda_1 e^{\lambda_1} \cdot z^{\lambda_2 + 1}}{\lambda_2 + 1} - v' = \phi(z) - v' \tag{7.29}$$

Obviously, there are two terms in Eq.(7.29). The first item $\phi(z)$ is a power distribution and the second item, in fact, corresponds to GM. According to the composition theory, the composition complexity of an object must be constant. Moreover, the composition complexities of two different objects are not the same. That is, this complexity will not change given various compositions of the object parts. Given an object of multiple grey levels, the composition complexity on various grey levels are the same i.e.

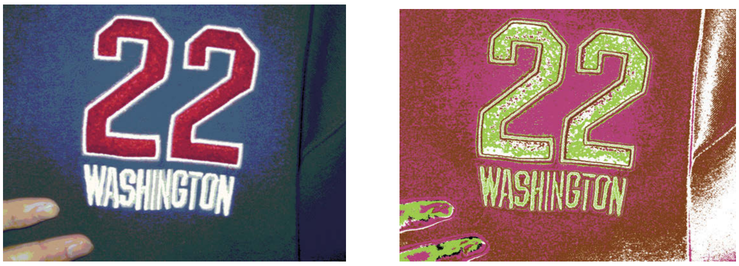
$$H_{s1}(z) \equiv H_{s2}(z), 0 \leq s1, s2 \leq 255 \tag{7.30}$$

Because $H_{s1}(z)$and$H_{s1}(z)$ are congruent, to simplify the analysis and maintain the congruence, we may reasonably enforce their corresponding items, power distribution and GM (see Eq. (7.22)), to ensure that they are congruent. Consequently, CCs on an image will be identified as parts of the same object if their $\phi(z)$ and geometric mean v' are the same.

Without estimating the parameters in Eq. (7.29), it cannot calculate $\phi(z)$. This paper proposes an approximation solution by first carrying out image segmentation according to GMs of all CCs. For a given image, it supposes that there are $\{s1, s2, \ldots sr\}$different GM values v'_s, on an image). Thus,

$$I_{m \times n} = \bigcup_{s=1}^{r} \Lambda_s \tag{7.31}$$

$\Lambda_s (1 \leq s \leq r)$ denotes all regions which are on one or several grey levels where the GM equals to v'_s. So the regions of the same GM are grouped together which satisfy the necessary condition of Eq. (7.30). That is, these regions are the candidate parts constructing the same object (their $\phi(z)$ may not be the same) (See an example in Figure 7.5 where the patches of same visual objects have the same GMs marked by the same



a) b)

Figure 7.5 Image partitioning based on proposed GM. a) original image; b) Image segmentation based on GM. The patches of the same GMs are marked by the same colour.

colour on the image). For the purpose of this paper on text detection, candidate text regions are now grouped together.

In Section 7.3.3, various features of MGCC defined in Section7.3.1 are extended to further define the features of a group of CCs of the same GM. A sophisticated analysis or SVM learning is then carried out on these new features to eliminate non-text objects on which $\phi(z)$is different from the text object although they have the same GM.

7.4.2 Analysis of GM regions

In order to detect the candidate text patches on an image, investigation on properties of characters in Chapter 3, 4 and 5, we define three conditions below,

$$1)\prod_{dir=0}^{7} N_{\psi_{s\_dir,i}} > 0,$$

$$2)D_{\Psi_{s\_con,i}} \times sw_{\psi} \ge D_{\Psi_{s,i}},$$

$$3)T_1 < Aspect\_ratio < T_2$$

T_1 and T_2 are two threshold values pre-defined.

For a given Λ_s, if its relevant features satisfy all three conditions above, Λ_s is regarded as the set of candidate text patches. Otherwise, it belongs to non-text objects.

So far, all candidate text patches have been allocated. In the next stage, three cases are investigated to firmly recognize the text patches. In the first case, the text information dominates the image, where all text patches stay on a single GM value v_s'. According to our work, this case can be interpreted by satisfying the following conditions,

$$\begin{cases} v_s' \ge v_j' \\ RSL_s \ge RSL_j \end{cases} \tag{7.32}$$

where $s \neq j, \forall j \in \{s1, s2, \ldots, sr\}$. v'_s is a dominated GM, and all patches in Λ_s correspond to text regions on image, which are also the most salient objects.

The second case is a bit more flexible than the first one. The text patches may stay on two different GMs. It is interpreted by satisfying the following conditions,

$$\begin{cases} v'_{s1} \geq v'_j \\ RSL_{s2} \geq RSL_j \end{cases} \tag{7.33}$$

where $s1 \neq s2, \forall j \in \{s1, s2, \ldots, sr\}, j \neq s1, j \neq s2$. Thus, all patches in $\Lambda_{s1} \cup \Lambda_{s2}$ belong to text regions on image.

If an image cannot match either case above, it will be viewed as the third case which is more complicated. Multiple conditions are defined below to progressively remove non-text patches. The rest of them will be regarded as text patches. For the text detection in a nature scene, the text region size would be constrained to a certain scale (this constraint may be very flexible and can depend on the data).

According our investigation in the section 4.41, about 3% of the characters have their height in the range of [301, 1000] pixels and they are called big character. Thus, given the size of image $WI \times HI$, if the ratio of MGCC size vs. image size is beyond the predefined threshold $\frac{1}{e}$, i.e., $\frac{W_{s,i}}{WI} \geq \frac{1}{e}$ and $\frac{H_{s,i}}{HI} \geq \frac{1}{e}, H_{s,i} \geq 300 pixels$, the corresponding regions Q_i are regarded as the large MGCCs. If any large CC satisfies one of conditions below, it will be regarded as non-text patches and removed from the image.

If $RSL_{s,i} < t_c$;

If $\frac{\max(D_{s\_dir,0}, D_{s\_dir,4})}{\min(D_{s\_dir,0}, D_{s\_dir,4})} > T_3$;

if $\frac{\max(D_{s\_dir,2}, D_{s\_dir,6})}{\min(D_{s\_dir,2}, D_{s\_dir,6})} > T_4$

if $D_{s,i} > T_5$ and $D_{s\_con,i} < T_6$;

Where t_c, T_3, T_4, T_5 and T_6 are defined empirically.

By exploring the three cases above, text patches will be firmly recognized. In order to better highlight the text regions on the image, the tightly bound boxes of CCs, instead of original CCs of various shapes, are used to pad text regions. The nearby patches of similar sizes are merged to each other.

7.4.3 Cross validation among GM regions and CRs

Clusters are generated by grouping proximity and similarity over GM regions according to 6.1.3.3. Meanwhile, based on the confident edge points, close regions (CRs) are obtained, and clusters over CRs are obtained by grouping by physical appearance similarity and position proximity.

For clusters, if clusters on GM regions are overlapped or embedded by clusters on close regions, the candidate GM clusters are considered as text. If they are exclusive, they will have average features of straight line RSL and local RMS contrast C_{rms} to filter them.

For isolated GM region, the region is distinguished by the determinative features, especially the straight line RSL and local RMS contrast C_{rms}.

7.5 Experiment

7.5.1 Data set

In order to provide a base line comparison, the proposed method is applied on the publicly available dataset ICDAR 2003[6]. This dataset has been widely used for text detection in natural scenes as a benchmark, and it contains 258 images in the training set and 251 images in the test set. The image regions containing text strings are labeled in XML files. In both the training set and test set, the images consist of varying spatial structures and varying texts.

7.5.2 Evaluation

The proposed algorithm is compared with benchmark methods with respect to $f-\text{measure}$ which it is a combination of two measures: precision p and recall r. According to [27], precision is the ratio of area of the successfully extracted text regions to the area of the whole detected region, and recall is the ratio of area of the successfully extracted text regions to the area of the ground truth regions. The area of a region is the

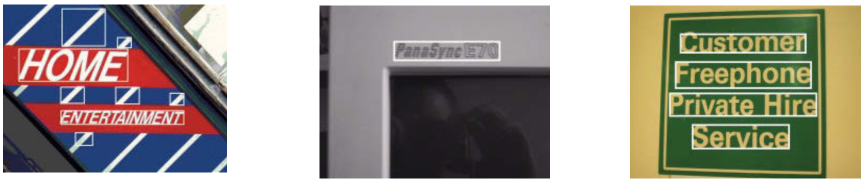


Figure 7.6 Sample results of text detection

number of pixels inside it. Low precision means over estimate while low recall means under-estimate. To combine p and r, f – measure is defined as below

$$f = \frac{1}{\frac{\alpha}{p} + \frac{(1-\partial)}{r}}, \tag{7.34}$$

where α represents the relative weight between these two metrics and $f \leq 1$. Larger f, better performance. In our evaluation, we set $\alpha = 0.5$.

7.5.3 Results and discussion

In the experiments, the thresholds defined in previous sections are determined empirically as $t_c = 0.3$, $T_3 = T_4 = 6$, $T_5 = 0.678, T_6 = 0.016$. Figure 7.6 shows the sample results of text detection.

There are several extremely difficult cases which failed in the existing[6, 7, 12, 13]. These cases include: 1) background of strong bright light; 2) blur image; 3) too small text; 4) very short text (e.g. less than 3 characters); 5) text behind mesh; 6) transparent text. Figure 7.7 demonstrates the encouraging performance of the proposed method on these cases. The proposed method not only extracts the subtle low-level text feature but also maintains the composition of visual object structure which is robust in relation to the variances of the image background.

Table 7.2 Performance comparison between the proposed method and benchmark methods presented in [6, 7]

	Precision	Recall	f-measure
Our method	**0.73**	**0.72**	**0.72**
GM	**0.67**	**0.64**	**0.625**
CC	**0.6**	**0.6**	**0.58**
Yao[11]	0.68	0.66	0.66
Epshtein[12]	0.73	0.60	0.66
Yi[13]	0.71	0.62	0.62
Becker	0.62	0.67	0.62
Chen	0.60	0.60	0.58
Ashida	0.55	0.46	0.50
David	0.44	0.46	0.45
Zhu	0.33	0.40	0.33
Wolf	0.30	0.44	0.35
J.Kim	0.22	0.28	0.22
Todoran	0.19	0.18	0.18
N.Ezaki	0.18	0.36	0.22

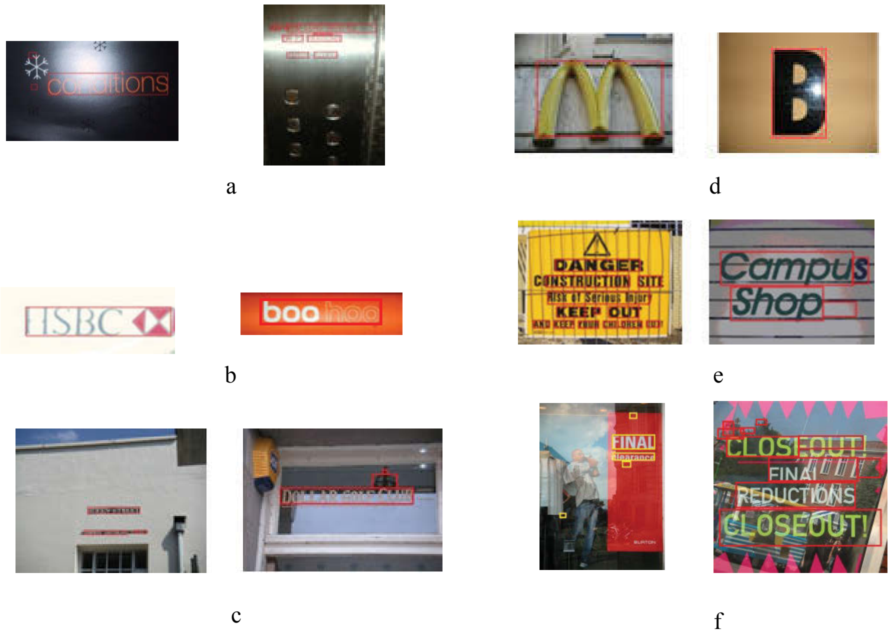
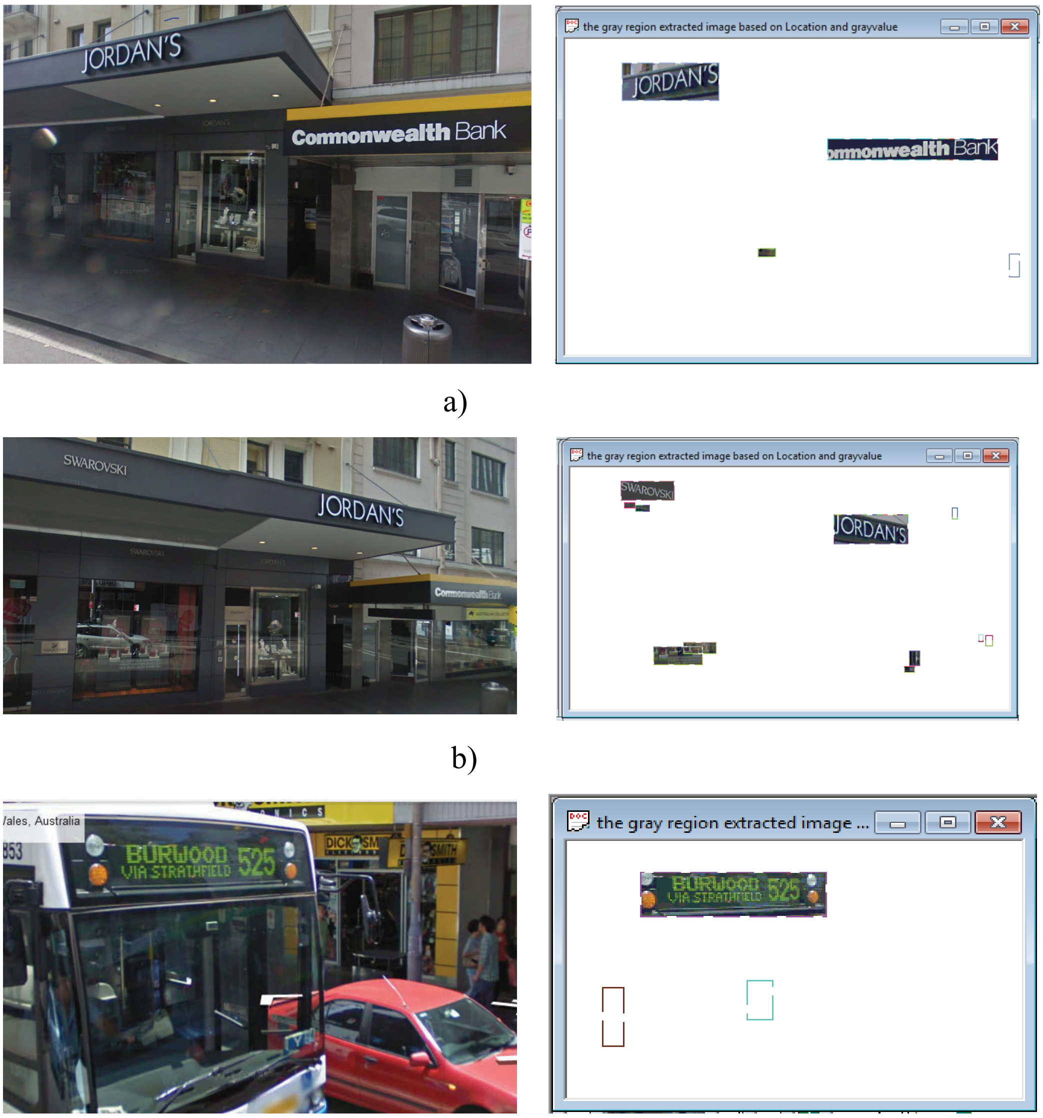


Figure 7.7 Results of text detection on extremely difficult cases which failed in the previous papers [42][293]. a) Strong highlights; b) Blur image; c) Too small text; d) Text with less than 3 characters; e) Text behind mesh; f) transparent text

The proposed method is compared with twelve benchmark methods which also carried out the experiments on the ICDAR2003 dataset. The reported performance of these benchmark methods and the proposed method is summarized in Table 7.2.

Given the impressive performance on the difficult cases shown in Figure 7.7, the proposed method demonstrates enough robustness overall when compared with previous methods. It is shown that the proposed method achieves an impressive f-measure on the top rank with the best recall rate against all these benchmark methods.

Further, the proposed method can be used to the real street view images, some experimental results of street view images are shown in Figure 7.8. Inspecting Figure 7.8 a), b) and c), the left column are original images and the right column shows the experimental results. These street view images with clutter scene full of windows, buildings and cars, however, the proposed method can detect the text effectively.


the gray region extracted image based on Location and grayvalue
JORDAN'S
Commonwealth Bank
SWAROVSKI
the gray region extracted image ...
BURWOOD 525
VIA STRATHFIELD
Wales, Australia
DICK SMITH

a)

b)

c)

Figure 7.8 Some experimental results on street view images

7.6 Summary

In above sections, a new solution is demonstrated, which describes the spatial structure of a natural scene image in a perceptual way for detecting the text regions on the image.

In the proposed solution, Multi-Grey Connected Components (MGCC) are used to represent the intricate pattern of an image. Based on the GM indicator, we explore the composition theory among component parts, and the Geometric Mean (GM) is proposed as a new way to describe the compositional complexity of an object across the meaningful CCs. Without following the legend framework based on supervised training, the proposed methods explore the input images on both pixel-levels through MGCC and also the semantic level through GM. In the end, the text regions are located on the image. The proposed method sorts out several cases which failed in the existing methods.

There are broad possible extensions to this work. The approach describing spatial structure is derived in a perceptual way which can be used for multi-object segmentation, semantic labelling and image quality assessment. It can also be used to discover the intrinsic compositional pattern for both visual objects and different shades of grey.

Chapter 8

Automatic Processing of Bank Cheques

Since text is still as a figure in document image, and also has the same properties of text appearance in local level, spatial level and global level, including individual character features, spatial relations (neighbourhood and appearance similarity), crowding effect and saliency. Therefore, the representations of image based on space regularity and three-level computational modelling of text in the cluttered scene can also have utility in terms of developing document processing systems which are capable of transferring the data present in documents like bank cheques, commercial forms, and government records into machine readable formats. As a large number of cheques in wide variety of layout have to be processed every day in a bank, an automatic reading system saves time and processing costs and offers better customer service. Thus, automatic cheque processing[295] is one of the most widely researched areas in document analysis and biometrics.

During the last decade, automatic cheque processing became an industrial problem. Some of the prominent vendors in the area of automatic bank cheque processing are ‘A2iA’, ‘Mitek’, ‘Parascript’ and ‘SoftPro’. The difficulties in developing an effective cheque reading system are the high degree of variability and uncertainty in the user-entered date information. People print or write the data zones in free style and there is no fixed format for cheques. There are differences not only in terms of background, but also in terms of the type and position of the machine printed and handwritten information. However, necessary data zones must be involved.

Generally, a cheque consists of the fields of legal amount, courtesy amount, date, and payee details which should be filled or printed by an account holder. The signature field is to be signed to ensure the authenticity of the cheque. Moreover, the two fields for filling the value of the cheque named the legal amount and courtesy amount are intended for redundancy. Legal amount contains the amount written in words, demonstrating the official value of the cheque. And the courtesy amount contains the amount written in numerals, which is supposed to be for courtesy purposes.

It is considered that a disagreement between the legal and courtesy amount shall be an indicator of amount alteration. Additionally, banks have the freedom to customize some parts of the cheques such as the background pattern, which is generally used to personalize the cheques. For example, they can use different fonts, special symbols, logos, lines, forms, different colours and imprinted textures.

For the purpose of the automatic reading and fraud detection of a cheque, we need to extract the information of payee details, legal amount and signature, and give the measure of the physical appearance of the characters. In this chapter, we will apply the representation of image and text in the clutter scene to automatically read and measure the scanned image of a bank check. This brings up three issues related to imaged-based technology: automated signature extraction; automatically reading the payee and legal amount; and the integrated system.

Practically, the user-entered information exists in two types of scripts: one is machine-printed text, and the other one is handwritten text. Since automatically processing the two different types of text involves adopting very different technology, how to differentiate them from each other becomes the first practical challenge, but most literatures have tended to ignore it. Additionally, signature is one typical form of handwritten text in a cheque.

Thus, we first start with signature extraction, which is entangled in the printed text background. We suggest that handwritten regions possess the typical information which follows the context-aware saliency principles. With such principles, the saliency is converted into a computation model, followed by a Context-aware Saliency Signature Detection (CSSD) algorithm. Experiments show that the proposed saliency approach can effectively detect the handwritten signature entangled in printed text. Further, SVM suggests that the two features, i.e. pixel density and contour density, play more important roles than other features in order to tell the intra-class differences between printed text and handwritten text. Based on the two important features averaged over the text string clusters, the type of cheque can be examined.

Then, we deal with automatically extracting the legal amount and payee content given the variety in font, size and background patterns. Accordingly, the whole document image is partitioned into space patches, and string clusters are generated through calculating the space relations – proximity and similarity. Then, three-level features of text are computed over the string clusters. Based on the three-level features, the type of image can be determined, that is to say, telling form or table type from non-form type. Furthermore, OCR is used to recognize the keywords, and the regions of payee and legal amount are identified and measured. These technologies are integrated into one whole cheque processing system, which provides automatic processing and human-intervention operations.

Experiment results on the real data (569,737 images) show that the system is effective. With the accuracy threshold of 70%, the successful detection rate of the payee name is found to be 98.7%. With the accuracy threshold of 80%, 90% and 95%, the successful detection rate of the payee name is stable at 97.93%. And for legal amount, the successful detection rate is found to be 98.13%, 97.27%, 94.60%, and 93.54% with the corresponding accuracy thresholds of 70%, 80%, 90%, and 95% respectively.

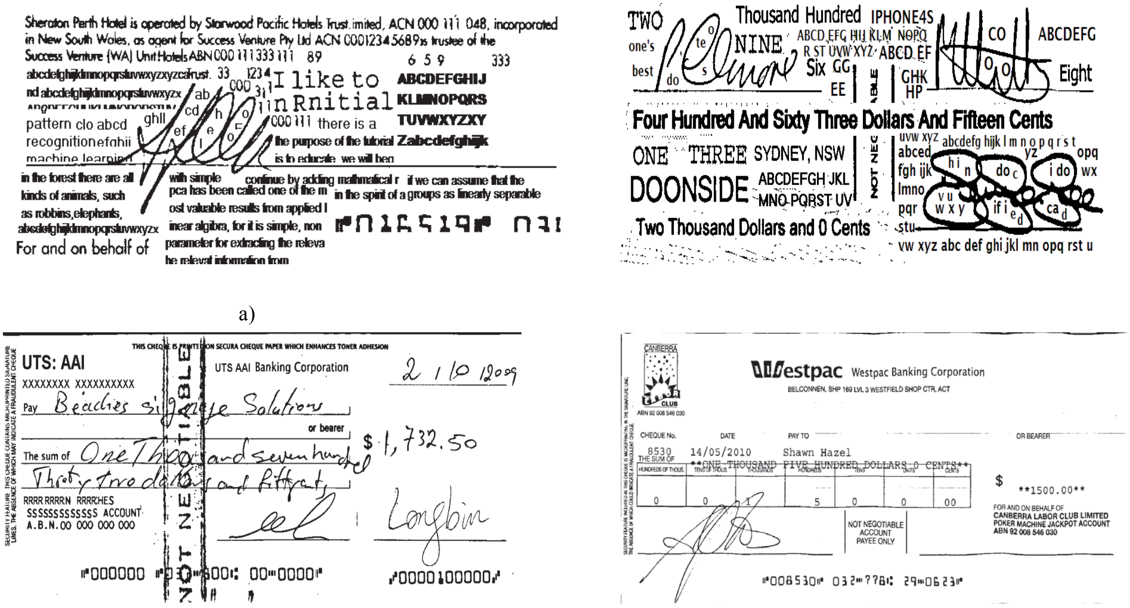
8.1 Signature extraction

Recently, many applications revive the study of handwritten scripts. Such applications include word spotting and text searching on professional documents and historical publications[296-302], and many applications, from reading of the legal amount or other field data on bank cheques [303-309] and of postal addresses [310, 311], revive the handwritten script study. In the word spotting task, if the given manuscript documents have clearly extracted text lines, they are further segmented into individual words and spotted or recognized by GHMM [296, 298], biologically inspired methods [297],W-TSV[301, 302]. However, all these applications are based on only processing either the printed text or handwritten text. They cannot detect the mixture of different types of texts as illustrated in Figure 1. In fact, this illustration discloses a very challenging problem, i.e. how the handwritten signatures can be extracted from the entangled printed text background.

In previous decades, many research outcomes have been reported on text detection and text recognition by using discriminative features.

For detection, it captures the difference between text/characters and non-text objects by using distinct features of character such as ridge points, edge points, SIFT points, stroke width, or the connected components (CC)-based features of printed text [299].

For recognition, it needs to capture the intra-character difference by using each individual character's discriminant features. For printed text, there are statistical features, geometrical and topological features [303]. And for handwritten text [304] or signature, the features proposed include Histogram of Orientation like features [296], biological inspired features [297], Weighted Topological Signature Vector (W-TSV) [302], and other features based on geometric centroids [312] and fixed points [313].



a)

c)

Figure 8.1 Entangled text with variety in layout, location, font and size.

However, the above approaches are not suitable for extracting handwritten text embedded with printed ones as shown in Figure 8.1. This is because the subtle difference between printed text and handwritten text is minor, which cannot be well described by such existing features. This section tackles this problem by exploring the conspicuity of handwritten text and demonstrates that such conspicuity can be used to detect handwritten text or signatures against printed text.

Letters are not a simple set of visual stimuli. They can differ from each other in basic features like the orientation of their line segments (e.g. A vs H) or line termination (X vs O). And letters that do not differ in basic features (e.g. L vs T) must still be discriminable or they would be of little use as members of an alphabet. They have not only a limited set of basic pre-attentive features, such as line, size, curvature, and several aspects of form (e.g line termination), but also space regularity which leads to informative text in perception. And the latter can makes the whole text pop out from the surroundings. If the document is full of text, the signature in the document can stand out from machine-printed text since its space organization is rare and different from the text. In this case, rather than saying that attention somehow identifies an object (conspicuity in pre-attentive), we would say that attention enables object recognition processes to work on a single item at a time. This notes that context information plays an important role in the consipicuity of the target object. And we can search the signature by the hierarchical structure of the component part of the text.

Our idea is to use context-aware object conspicuity, which has been previously studied in the research of saliency [10].However, the problem of the re-formulation of the saliency for the typical cases, shown in Figure 8.1, is not always straightforward. It requires integrating the detailed bottom-up object features and top-down domain knowledge for a concrete and feasible description of context-aware saliency. This section demonstrates such a reformulation and transforms it into a context-aware saliency signature detection algorithm.

8.1.1 The algorithm framework

We apply the representation of image and the three-level text model to describe the document image. Firstly, the document image is partitioned into multiple regions of interest through connected component analysis on a binary image. And then three-level text features are calculated, including basic features, letter-centred (space organization, i.e. letter string) features, and word-level (relations among string clusters) features. The three-level features agree with the four basic principles of human vision defined by the context-aware saliency [10] :

1. Local low-level considerations, including factors such as contrast and colour. In our case, such low-level factors include the edge point and other feature points that represent the contour or shape of an object on a binary image, but also it consists of those pixels that can be connected with their neighbours so that the space occupied by a visual object can be obtained.

2. Global considerations, which suppress frequently occurring features, and maintain features that deviate from the norm. In our case, an image is partitioned globally into multiple regions of interest through adjacent groupings by Connect Components Analy-

sis (CCA). And the features are computed over basic regions. Such global segmentation will later on help to highlight the conspicuity of the handwritten signature.

3. Visual organization rules, which state that visual forms may possess one or several centres of gravity through which the form is organized. In our case, this rule means the space regular organization in which letters form a text (i.e., letter string), which is measured by letter-centred features. Meanwhile, the features of regions given by CCA can be extended from CCA to regions given by letter strings. And such centres are the centres of letter strings formed with this rule.

4. High-level factors. In our case, the most important features are distinguished from the rest of the features through the process of SVD.

The framework of Context-aware Saliency Signature Detection (CSSD) is shown in Figure 8.2. We incorporate the principles (1), (2), (3) and (4) into a context-aware saliency computation model. Through this model, an image is partitioned into a set of space patches given by CCAs on the global level, and each space patch is considered as one visual object (i.e. VO). And each of them has attributes on multi-layers, including low-level, global-level, organization, and high-level factors. Through the SVD process, the relationship and sub-structure among VOs (essentially described by relevant connected components) can be captured, and the context-aware saliency is calculated. Based on the saliency, the salient signature VOs are figured out by the VO analysis.

Therefore, the algorithm framework contains three parts: the first is the image partition, the second is the context-aware saliency model, and the third is the VO analysis, as illustrated in Figure 8.2.

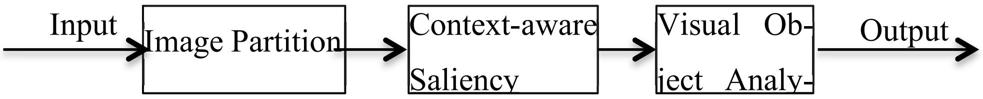


Figure 8.2 The framework of Context-aware Saliency Signature Detection

8.1.2 Image partition

8.1.2.1 VO regions

According to the theory of perspective, when objects recede from the eye or camera, the size of the object decreases, this means the space occupied by the object in image is reduced meanwhile the contour is lessened. If the distance is far away enough, the boundary of the separated object disappears and the separated objects are merged into a whole. What is more, the boundary of one thing with another is of the nature of a mathematical line (e.g. contour detected by edge detector). This suggests that space occupied by an object in an image plays a critical role in being visible and discernable.

Therefore, we regard those space patches occupied by the object as the building bricks of a visual object. Since the document image is full of characters, and each character usually occupies one space even if it consists of several line segments, we consider one connected component space patch as one basic Visual Object (VO) in bank cheque processing.

The formal definition of the VO concept and the necessary auxiliary definitions are given in Table 8.1.

Image I is a mapping $I : D \subset R^2 \rightarrow S,\ S = \{0,255\}$. Connected component space patches/regions are well defined on images if:

5. An adjacency (neighbourhood) relation $A \subset D \times D$ is defined. In this section 4-neighbourhoods are used, i.e. $p,q \in D$ are similar and adjacent (pAq) iff $\sum_{i=1}^{d} |p_i - q_i| \leq 1$, $d = 4$ stands for the number of neighbour, and p,q has the similar grey level $s = 0$.

VO region is a contiguous subset of D, i.e. for each $p,q \in VO$ there is a sequence $p, a_1, a_2, \ldots, a_n, q$ and $pAa_1, a_iAa_{i+1}, a_nAq$. Let N_i be the number of the sequence points. Let $\Lambda = \{VO_1, VO_2, \ldots, VO_M\}$ be a set of regions in grey level $s = 0$. Thus, for a binary image $I_{m \times n}$, it can be partitioned into a set of VOs as below,

$$I_{m \times n} = \Lambda = \bigcup_{i=1}^{M} VO_i.$$

Table 8.1. Definitions used in following sections.

8.1.2.2 The VO String and VO Block on visual organization

Letters together with harmonic space regularity form a whole text. Therefore, space regularity among letters corresponds to the visual organization of text. It consists of letter spacing, word spacing and interline spacing. All of them have tight relations to the type size of an individual letter.

With respect to the type size of individual letters, i.e., the width and height of a letter, the space organization is quantified and calculated. Here, we regard each VO as a character. Without the loss of generality, given VOs $VO_1, VO_2, \ldots, VO_k$, their regions of interest are those tightly rectangular bounding boxes $ROI_1, ROI_2, \ldots, ROI_k$ in width $W_1, W_2,$

…, W_k and height $H_1, H_2, \ldots, H_k$ respectively. Relations among these *VOs* are defined over those *ROIs* as thus:

An adjacency (neighbourhood) relation A_{VO} is defined over boundary boxes. In this section, the size of the boundary boxes, and the distance between the bounding boxes of letters are used. $(ROI_i A_{VO} ROI_j)$ iff their physical appearances are similar and their positions in space are near enough on these conditions:

i. $\Delta H_{ij} \le \xi_1 \max(H_i, H_j)$ and $\Delta Hij \le \xi_2 \min(H_i, H_j)$, $\Delta H_{i,j} = |H_i - H_j|$, ξ_1, ξ_2 are scale factors.

ii. Horizontal distance between the two letters $\Delta HSpace\_x < \xi_3 \max(H_i, H_j)$, here ξ_3 is a scale factor. Additionally, in the vertical direction, the y-coordinates of the two letters have common parts.

VO strings VO_Str is a contiguous subset of the boundary boxes occupied by letters, i.e. for each $ROI_i, ROI_k \in VO\_Str$ there is a sequence of $ROI_i, b_1, b_2, \ldots, b_n, ROI_k$ and $ROI_i A_{VO} b_1$, $b_i\ A_{VO} b_{i+1}$, $b_n A_{VO} ROI_k$. Let N_{str_i} be the number of the contiguous RIOs of VOs, i.e. the length of the string VO_Str, and ROI_{str_i} denote the regions given by the tightly rectangular boundary box of VO_Str in width W_{str_i} and height H_{str_i}.

Given two VO Strings VO_Str_i and VO_Str_j, an adjacency (neighbourhood) relation A_{VB} is defined over VO_Strs. Size of $ROI_{str}s$ and distances between them are used. $(VO\_Str_i A_{VO} VO\_Str_j)$ iff their physical appearances are similar and their positions in space are near enough on these conditions:

i. $\Delta H_{ij} \le \lambda_1 \max(H_i, H_j)$, $\Delta H_{i,j} = |H_i - H_j|$, λ_1 are scale factors.

ii. Vertical space between two VO strings follows the principle of line space in text: $\Delta VSpace < \lambda_2 \min(H_{str\_i}, H_{str\_j})$, and in horizontal direction, the distance of start point in $x -$ coordinate $\Delta x \le \lambda_3 \min(W_{str\_i}, H_{str\_i})$ Here, λ_2, λ_3 are scale factors.

VO Blocks *A* VO_Block is a contiguous subset of VO strings, i.e., for each VO_Str_1 and $VO\_Str_k \in VO\_Block$, there is a sequence of $VO\_Str_1, S_1, S_2, \ldots, S_n, VO\_Str_k$ and

$VO\_Str_1 A_{VB} S_1$, $S_i A_{VB} S_{i+1}, S_n A_{VB} VO\_Str_k$. Let N_{MB} denote the number of VO strings, and ROI_{Block_i} be the region of the tightly rectangular boundary box of the VO block.

8.1.3 Context-aware saliency computation model (CSCM)

In our case, each connected space patch given by adjacent grouping is regarded as one VO. In a global consideration, an input document image is defined as a set of co-occurring visual objects. A visual object (VO) is a visual information carrier that delivers the author's intention and catches part of the user's attention as a whole. A VO often represents a semantic object, for instance, character or symbol, words, a text sentence, signature, or non-text objects, such as lines, forms, logos, and pictures etc., as illustrated in Figure 8.3.

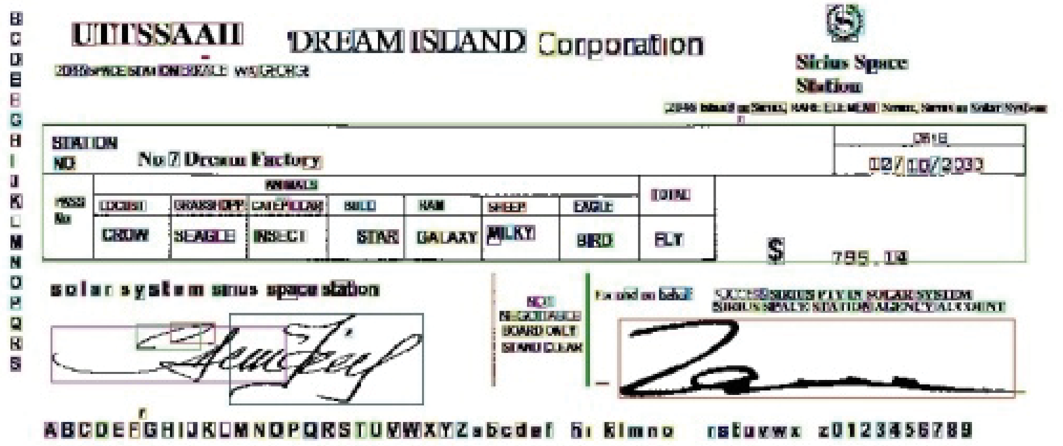


Figure 8.3 Visual Objects (VOs) marked by rectangle boxes

With regard to VOs, human processors are serial processors who handle one source of visual information at a time. While clearing bank cheques, they integrate various activities and maintain an appreciation of the dynamic time-consuming working situation by sampling information and shifting attention from one thing to another. They rely on judgment, experience, estimation, prediction and memory to fill in the gaps, and to shed less important information. Expectancy, motivation, and conspicuity all play a role in determining what a human processor will notice.

Of course, for human processors to find information useful, it must first be noticed. With respect to documents, characters, letter strings and words are all vivid informative figures, which all have legibility, readability and conspicuity. The issue of what attracts attention when processing depends on the degree to which the information content matches the informational needs of the processor. It brings up two key points, the association between the information content and conspicuity, and the notion that conspicuity may depend upon the user's task. The former suggests that there is conspicuity of the visual object compared to its surroundings; the latter tells us that it has distinguishable and discernable features which can be verified by processors.

For signature detection, the signature is observed impressively by its physical appearance, i.e. it stands out from its institutional surroundings. And obviously, it is our target of detection. Therefore, we consider the context-aware saliency of signature over the space patches given by adjacent groupings.

The four basic principles of the context-aware saliency of VO are discussed as follows. At the local level, the attributes of VO correspond to not only the local contrast features, i.e. edge points, but also those proper points which consitute the VO through adjacent groupings. In global consideration, the image is partitioned into a set of VOs which are the basic informative visual component parts and contribute to highlight the target.

And in visual organization, the space relations among VO are defined and calculated based on position proximity and physical similarity, and lead to a set of VO Strings and VO blocks, over which the space averaged statistical features are computed. SVD is used to decompose the attribute space of VO, and the conspicuous basic VOs and their interrelationships among CCs are exposed. Therefore two discriminative features are obtained to tell handwritten text from machine-printed text, including pixel density and edge density. Moreover, through the features of VO strings, signature handwritten cheques can be picked out from the printed ones.

The above mentioned factors will be dealt with in detail below.

8.1.3.1 Definition

We assign three attributes for each VO: region of interest (ROI), multi-layer attributes (MA) and occurring frequency (f).

Definition 1: Region-of-interest (ROI) is referred to a spatial region or segment within an image, i.e. VO region. As shown in Figure 3, a ROI can be denoted by its tight rectangular geometrical parameters, i.e. $\{Left, Top, Right, Bottom\}$ Inside a ROI, the pixels share some common properties (see the details below).

Definition 2: The context-aware saliency model for an image is defined over a set of VOs:

$$\{VO_i\} = \{ROI_i, MA_i, f_i\}, 1 \le i \le M, \tag{8.1}$$

Where, VO_i is the *i*th VO in an image, ROI_i is the ROI of VO_i, MA_i is the MA of VO_i, f_i is the occurring frequency f of discriminative features, M represents the total number of VOs in an image.

8.1.3.2 Multi-layer attributes

8.1.3.2.1 Attributes in local-level and global-level

Generally, edge points and their orientations are the predominant features of local contrast. Let N_{con_i} denote the number of edge pixels on a given VO_i. Thus, by combining its total number of sequence pixels N_i and N_{con_i}, one attribute of VO is formed and represented as $AL(N_i, N_{con_i})$

In high perception, colour, motion, orientation and size are the undoubted guiding attributes. Shape is the most vexed one because of its complications. We capture such shape related attributes inside the given VO_i, including the size and density of mass as below.

The size of VO_i is determined by the width W_i and the height H_i of its tightly rectangular boundary box,

$$W_i = x_{\max} - x_{\min}, \quad H_i = y_{\max} - y_{\min}, \tag{8.2}$$

where $(x, y) \in VO_i$. Thus, the area of ROI_i is defined by

$$Area_{ROI_i} = W_i \times H_i. \tag{8.3}$$

Inside VO_i, the density of mass is considered as an important property to measure 2D visual space patches. Accordingly, we define the pixel density D_p_i and edge density D_con_i of VO_i as follows:

$$D\_p_i = \frac{N_i}{W_i \times H_i}, \quad D\_con_i = \frac{N_{con_i}}{W_i \times H_i}. \tag{8.4}$$

Consequently, VO_i can be rewritten as a vector with two-layer attributes

$$VO_i = \left[N_i, N_{con_i}, W_i, H_i, D\_p_i, D\_con_i, Area_{ROI_i}\right], \tag{8.5}$$

Where the first two components are regarded as low-level attributes and the rest of the components are regarded as high-level components. What is more, when we consider the visual organization of text, we need to quantify the letter-centred attributes of space organization based on the three-level text model proposed in Chapter 4, and each VO is nested in its VO string and VO block.

8.1.3.2.2 The attributes over VO String and VO Block

The attributes of VO String and VO Block are statistical space averaged features. Without the loss of generality, given a VO string $VO\_Str_i = \{ROI_{1,} ROI_2, \ldots ROI_n\}$, its whole tight boundary box is measured by the width $Wstr$ and the height $Hstr$, let N_{str_i} denotes the number of the entire basic ROIs given by component space patch VOs. The statistical space averaged properties of the VO string are composed of average width $\overline{W}_{str}$ and average height $\overline{H}_{str}$ of its component ROI, average edge density $D\_str\_con\_i$, average pixel density $D\_str\_p\_i$, and those standard differences $\sigma_{W\_str}, \sigma_{H\_str}$. they are calculated respectively as thus:

$$\overline{W}_{str} = \frac{1}{N_{str\_i}} \sum_{i=1}^{N_{str\_i}} W_i, \quad \sigma_{W\_str} = \sqrt{\frac{1}{N_{str\_i}} \sum_{i=1}^{N_{str\_i}} \left(W_i - \overline{W}_{str}\right)^2}, \tag{8.6}$$

$$\overline{H}_{str} = \frac{1}{N_{str\_i}} \sum_{i=1}^{N_{str\_i}} H_i, \quad \sigma_{H\_str} = \sqrt{\frac{1}{N_{str\_i}} \sum_{i=1}^{N_{str\_i}} \left(H_i - \overline{H}_{str}\right)^2}, \tag{8.7}$$

$$D\_str\_p\_i = \frac{1}{Wstr \times Hstr}\sum_{i=1}^{N_{str\_i}} D\_p_i,$$

$$D\_str\_con\_i = \frac{1}{Wstr \times Hstr}\sum_{i=1}^{N_{str\_i}} D\_con_i. \tag{8.8}$$

Where, W_i, H_i, D_p_i, D_con_i are calculated according to the equation (8.2) and equation (8.4). Similarly, the space averaged statistical properties of the VO block are calculated over the VO block regions. Therefore, each VO is nested in its VO string and VO block with their respective statistical properties. Signature or handwritten text can be picked up by the hierarchical structure of its component VOs.

8.1.3.3 SVD decomposition of VOs' attributes space

8.1.3.3.1 Attributes space

In the attribute domain of VOs, the occurring frequency f of discriminative features will be captured through the process of SVD. We adopt basic attributes of VO_i to form an attribute matrix space Θ. Let τ_i denote the attribute vector of VO_i, thus

$$\Theta = \{\tau_0, \tau_1, \ldots, \tau_{M_k - 1}\}^{\mathrm{T}} \tag{8.9}$$

Here $\tau_i : \{D\_p_i, D\_con_i, N_i, N_{con_i}, Area_{ROI_i}, H_i, W_i\},\ i = 1,2,\ldots,M_k \xrightarrow{yields} \Theta : Span$ $\{D\_p_i, D\_con_i, N_i, N_{con_i}, Area_{ROI_i}, H_i, W_i\}.$

SVD is used to expose the conspicuous substructure and interrelationship among components, SVD decomposes Θ to

$$\Theta = \mathbf{U}\ \boldsymbol{\Delta}\ \mathbf{V}^{\mathrm{T}} \tag{8.10}$$

Instead of the analysis of the original attributes of VOs, the VO analysis (see section 4) is carried out on the sub-spaces: $\mathbf{U}, \boldsymbol{\Delta}$ and $\mathbf{V}$.

8.1.3.3.2 Visual object analysis on SVD subspace

From the view of transformation, SVD decomposition projects each row vector of matrix Θ , which represents each VO with multi-layer attributes, to the row vector $u_k = \{u_{k,1}, u_{k,2}, \ldots, u_{k,r}\}$, $r = 7$ of matrix $\mathbf{U}$ and projects each column vector of matrix Θ which represents the same attributes across all VOs to the column vector $v_j = \left[v_{j1}, v_{j2}, \ldots, v_{jr}\right]^{\mathrm{T}}$ of matrix $\mathbf{V}^{\mathrm{T}}$.

Diagonal matrix $\boldsymbol{\Delta}$ contains the sorted singular values in which the largest one is at the first position i.e. up-left corner of the matrix. That is, the first column vector of $\mathbf{U}$ (i.e. $\mathbf{u1}$) and the first row vector of $\mathbf{V}$ (i.e. $\mathbf{v1}$), which corresponds to the largest singular values, are used to analyse the relation between VOs and the relation between different attributes of VOs respectively.

The VOs represented by each row vector are an abstraction away from the noisy correlations found in the original attribute data space, and they best approximate the underlying structure of the dataset along each dimension independently. So we can index each individual VO_i from the row vector u_i of matrix $\mathbf{U}$ at the same row. Meanwhile, the representative of VOs that share substructure become more similar to each other, and VOs that were dissimilar to begin with may become more dissimilar as well, which results in an discriminable approximation of the data that contains substantially fewer dimensions than the original. In practical terms, we take the first column $\mathbf{U1}$ of $\mathbf{U}$ as the main discriminative vector of VOs to find out the class index of VOs in the matrix Θ by K-means clusters.

The majority of those VOs are similar, however, the conspicuous VOs are quite different from the majority ones, that is to say, the occurring frequency f of them is small. This results in two distinguished clusters, one consists of the majority components, and the other usually includes small numbers. The K-means algorithm can get it clearly. And the number of elements in each cluster can represent the occurring frequency f in the Equation 1. Moreover, K-means is operated on $\mathbf{U1}$ to calculate the saliency level of each VO, since saliency in the primate brain is represented at several levels.

In an extreme case, if there is of only one conspicuous VO represented by $\mathbf{u1}_i$ in $\mathbf{U1}$, we can write $Y_i = \mathbf{X}_i\mathbf{U1}$, $\mathbf{X_i} = \{\mathbf{x_i}\}$ and $\mathbf{x}_i = \mathbf{e}_j^{\mathrm{T}}$ is a vector form of the trivial basis, with all zero entries except one in the jth position. The index j is selected such that

$$\forall l \neq j, \left\| Y_i - \mathbf{e}_j^{\mathrm{T}}\mathbf{U1} \right\|_2^2 \leq \left\| Y_i - \mathbf{e}_l^{\mathrm{T}}\mathbf{U1} \right\|,$$

In general terms, each y_i is defined as $e_i^2 = \left\| y_i - \mathbf{e}_\mathbf{j}^{\mathrm{T}}\mathbf{U1} \right\|_2^2$ and the overall MSE is

$$E = \sum_{i=0}^{M-1} e_i^2 = \left\| \mathbf{Y_i} - \mathbf{X_i}\mathbf{U1} \right\|_F^2 \tag{8.11}$$

Now, the problem to index conspicuous VOs in $\mathbf{U1}$ can be considered to find some VOs that minimize the error E, subject to the limited structure of $\mathbf{X}$, whose columns must be taken from the trivial basis. $\min_{\mathbf{U1},\mathbf{X_i}} E, i = 0,1,\ldots,M$ subject to $\forall i,\ \mathbf{x_i} = \mathbf{e}_l^{\mathrm{T}}$ for some l.

And we minimize the expression in (11) by the following iterative process. Since saliency in the primate brain is represented at several levels, let s denote the saliency level, when $s = 1$, it corresponds to the most conspicuous VO. After repeatedly clustering into two clusters, the saliency levels of each individual VO can be labelled and for each saliency level, the cluster also is figured out. We shall call this algorithm "SVD-K" to parallel the name k-means. By sweeping through the first column $\mathbf{U1}$, it can index the *VO*s at the different saliency level as well as index those *VO*s in the whole *VO*s link list. The detail of this algorithm is introduced in Table 8.2.

```
Initialization: Get $\mathbf{U1}$ as the representation of $VO$s, and set the iterative number $J = 1$, $s = 1$ $\mathbf{u1}^s = \mathbf{U1}$;
  Repeat until convergence (stopping rule1):
    Step 1: The $s$ degree of salient VOs extraction.
            Repeat until convergence (stopping rule2):
      Step1.1 Partition the $\mathbf{u1}^s$ elements into 2 sets $\{R_{s,1}^{J-1}, R_{s,2}^{J-1}\}$ with the centres
        $\{\mu_{s,1}^{J-1}, \mu_{s,2}^{J-1}\}$,each holding the indices most similar to their centres,
              $R_{s,k}^{J-1} = \{\forall l \neq k, |\mathbf{u1}_l^s - \mu_{s,k}^{J-1}| < |\mathbf{u1}_l^s - \mu_{s,l\neq k}^{J-1}|\}$
      then, update each centres $\{\mu_{s,1}^{J-1}, \mu_{s,2}^{J-1}\}$ by
      $\mu_{s,k}^{J} = \frac{1}{|R_{s,k}|} \sum_{l \in R_{s,k}^{J-1}} \mathbf{u1}_l^s$
      Step 1.2 Set $J = J + 1$;
    Step 2. Update.
          Update the rows of the first column $\mathbf{u1}^s$, sweep the $\mathbf{U1}_l$ which corresponds
        to the rows in $\Theta$ selected by the nonzero entries in $\mathbf{X}_i$ with the all zero entries
        except a few positions, $\mathbf{u1}^s$ is reduced by the $\mathbf{U1}_l$ and updated into $\mathbf{U1}'$.
          Then, set $s = s + 1$, $\mathbf{u1}^s = \mathbf{U1}'$.
```

Table 8.2. K-means operated on SVD subspace to compute saliency level.

8.1.4 VO analysis

8.1.4.1 Salient visual object detection

Each element of the first column vector of $\mathbf{U}$ (i.e. $\mathbf{U}_1$) represents one VO respectively. The relationship (e.g. similarity) between VOs is noted by the clusters formed by SVD-Kmeans algorithm. Thus, the occurring frequency f of a VO in one cluster can be calculated by counting the total numbers of elements in this cluster, i.e. the occurring frequency of each element in $\mathbf{U1}$, which can be presented as a histogram of the element in $\mathbf{U1}$.

According to the theory of salience [12], the VOs identified as the salient ones usually have low occurring frequency in the image. Thus, the salient VOs can be detected by clusters since VOs are classified into two categories according to the histogram of element in $\mathbf{U}_1$: one with high occurring frequency and the other with low occurring frequency respectively. And each cluster has its corresponding saliency level. The first cluster with only a few elements is considered the most conspicuous cluster. And VOs in conspicuous clusters are considered to be the salient VOs, which are the candidates of signature.

8.1.4.2 Key features and signature detection.

The element values of the first row vector of $\mathbf{V}$ can indicate the importance of the different features of VOs. In this case, large values in $\mathbf{V}_1$ correspond to the important attributes of VOs. And the two most important attributes are selected: pixel density D_p_i and edge density D_con_i.

With respect to those salient VOs selected in 6.1.5.1, we can find their nested VO strings and the VO block. In the VO string, both pixel density and edge density are used to identify whether this VO string is to have a similar pixel density/edge density or not. The similarity is defined in terms of these conditions:

i. $Similarity(D\_str\_p\_i, D\_p_i) = \|D\_str\_p\_i - D\_p_i\| < \varepsilon$, $D\_str\_p\_i < \varepsilon_2$ and $D\_p_i < \varepsilon_2$;

ii. $Similarity(D\_str\_con\_i, D\_con_i) = \|D\_str\_con\_i - D\_con_i\| < \zeta$, $D\_str\_con\_i < \zeta_2$ and $D\_con_i < \zeta_2$.

Where $\varepsilon, \varepsilon_2, \zeta, \zeta_2$ are the predefined threshold. If the VO strings are similar to the nested VO, then the VO string is the signature.

8.1.5 Signature experiment

Based on the BME2 off-line signature corpus [15], a synthetic dataset is generated for our study (available on request). Each synthetic image mixes one of handwritten signatures from BME2 with a print text background image randomly obtained from the Internet. It contains 33 images embedding 33 original signatures from 33 signers, including 4 types of images: 1) single signature not-entangled with printed text, 2) single signature entangled with printed text, 3) multiple signatures not-entangled with printed text, and 4) multiple signatures entangled with printed text.

Two different experimental results are demonstrated: 1) salient VO detection based on context-aware saliency analysis (see Figure 8.4); and 2) signature detection on the printed text background (Figure 8.5).

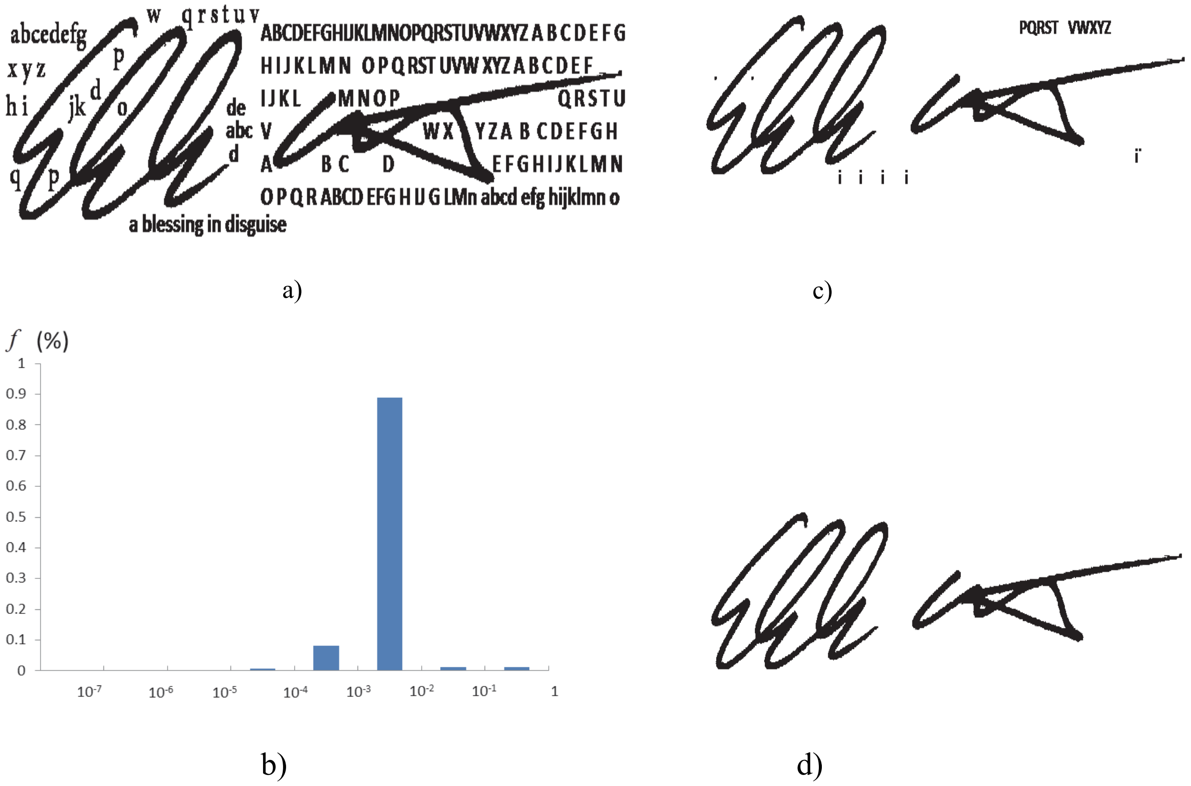


Figure 8.4 Salient VO detection: a) original image; b) histogram of elements; c) salient VOs (i.e. signature candidates); d) signature extracted finally

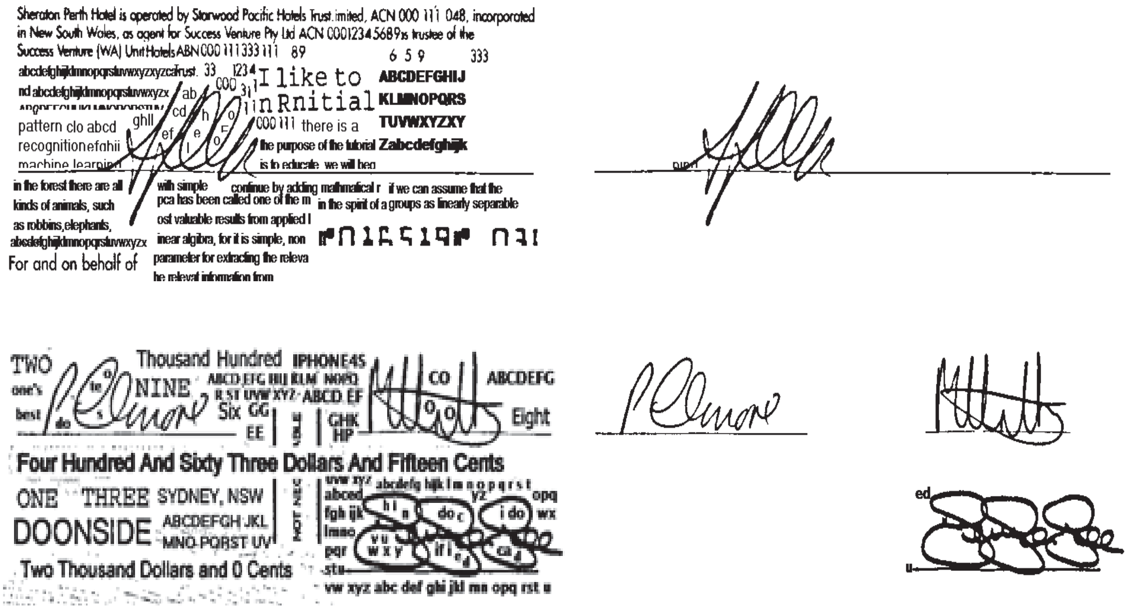


Figure 8.5 Results of signature detection. The left: input image; The right: signatures

All of the experiments are carried out with the same parameter setting. VOs are generated based on the adjacent grouping. The feature vectors of all VOs are spanned onto a feature space. Through SVD, the relationships among VOs and the importance of the individual attributes in the feature vector are exposed. The occurring frequency of each element of value in **U1** is normalized into $\left[0,1\right]$and quantified into 8 bins. Such frequency indicates the occurring frequency of the corresponding VOs in the image (see Figure 8.4 b). K-means is applied to this histogram of frequency to create two clusters. Salient VOs are selected as signature candidates which correspond to the lower frequency (see Figure 8.4 c). Then, we set the predefined threshold $\zeta_2 = \varepsilon_2 = 0.3$, and signatures are finally detected (see Figure 8.4 d).

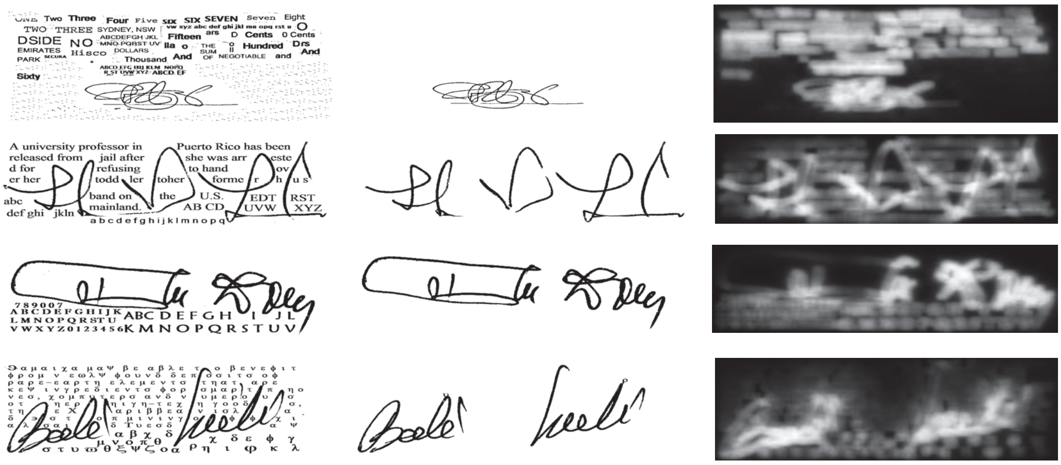


Figure 8.6 Signature detection on printed text back-ground Left: original images of four cases: single signature not-entangled with printed text, single signature entangled with printed text, multiple signatures not-entangled with printed text, multiple signatures entangled with printed text; Middle: Signature detection using the proposed method (CCSD). Right: context-aware saliency detection [10].

Figure 8.6 shows the results of signature detection based on the proposed method and common saliency detection. It demonstrates that the proposed method can achieve superior performance by extracting the clear signatures from the entangled print text background. Without the clear problem reformulation, common saliency detection cannot separate the handwritten signatures from the printed text (see the right column on Figure 8.6).

8.1.6 Extended application of handwritten cheque selection

By analysing the principles of saliency, we reformulate saliency detection in a computable manner for the case of handwritten signature detection from the entangled printed text. Such saliency is formed by visual objects with multiple attributes. Such redefined saliency is adopted and transformed into a context-aware saliency signature detection (CSSD) algorithm.

Using such a newly proposed saliency approach, the handwritten signatures are successfully extracted from the entangled printed text based on our preliminary dataset. Further, it can be extended into extracting handwritten text from printed ones.

According to SVD subspace analysis, there are two important features for discerning the handwritten text from the machine printed text: pixel density and edge density. Given a VO string, the histogram of pixel density and the edge density of the VO in the VO string is obtained. Then, the probability $P(D\_p_i < \varepsilon_1)$, $P(D_con_i < \varepsilon_2)$and the joint probability $P(D_p_i < \varepsilon_1, D_con_i < \varepsilon_2)$are computed over the histogram. The VO string is considered as handwritten text in a probability of$P(D_p_i < \varepsilon_1, D_con_i < \varepsilon_2)$. It is detected as handwritten text when it satisfies the following conditions:

If $P(D\_p_i < \varepsilon_1, D\_con_i < \varepsilon_2) > \eta$ and $D\_str\_con\_i < \varepsilon_2$ and $D\_p\_i < \varepsilon_1$, VO string is handwritten string.

Further, we apply the centre-surround principle to tell whether the information of the payee and the legal amount in a cheque is handwritten or not. Due to the fixation locations being usually biased towards the centre of the image and the designer's bias for keeping objects at the centre of the image, the information payee and legal amount lies toward the centre of the image. This is the most significant component part which needs to be filled. The others lying on the boundary of the document image belong to the background or the less important information. Thus, when all the VOs lying towards the centre are handwritten, the cheque can be considered as the handwritten cheque.

With the study of receptive field size, Enroth-Cugell measurements of the contrast sensitivity of X-cells to sinusoidal grating patterns of different spatial frequencies makes it possible to predict the radius at which the sensitivity of the central summating region falls to $1/e$(37%) of its maximum value. However, the diameter of the central region of a receptive field is equated either with the diameter of light which has the lowest incremental threshold or with the diameter of the boundary between the receptive field regions from which responses of opposite poliarity can be evoked by a small spot of light. In practical terms, the receptive field is considered to be an ellipse locating its centre at the centroid of the image. The semi-major axis and semi-minor axis of the ellipse are equal to $1/e$ (37%) of their maximum value (i.e. half the width in the horizontal direction, and half the height in the vertical direction) respectively. Let $Area_{hw\_str}$ denote the total area of the handwritten VO string in the receptive field, and $Area_{VOstr}$ denote the area of all the VO strings in the receptive field. The cheque is considered as a written cheque when the area $Area_{hw\_str}$ is considerably big compared to the area

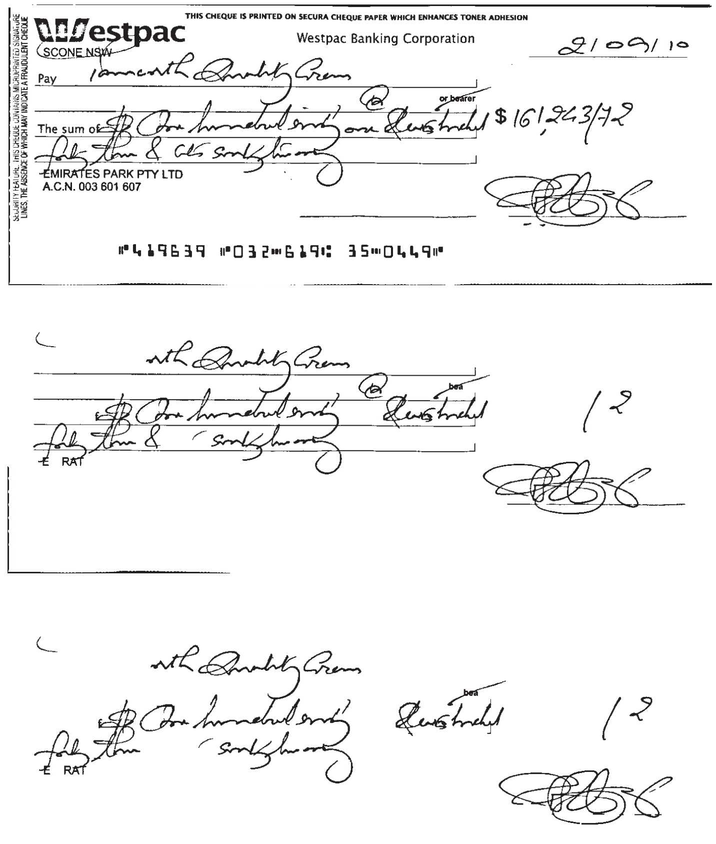

THIS CHEQUE IS PRINTED ON SECURA CHEQUE PAPER WHICH ENHANCES TONER ADHESION
Westpac
SCONE NSW
Westpac Banking Corporation
Pay
or bearer
The sum of
$
EMIRATES PARK PTY LTD
A.C.N. 003 601 607

Figure 8.7 Handwritten cheque selections. Top: input image; Medium: Candidates of handwritten VOs; Bottom: Handwritten VO strings and handwritten cheque la-

$Area_{VOstr}$ i.e., $\frac{Area_{hw\_str}}{Area_{VOstr}} > \lambda$, $\lambda > 0.5$.

As illustrated in Figure 8.7, the written strings are extracted based on features and this cheque is labelled as a handwritten cheque according to the centre-surround principle.

8.2 Automatic extraction of legal amount, payee name

8.2.1 The flow chart

Based on connected component space patches resulting from the image partition introduced in section 8.1.2, we determine which VO string and VO block represent the payee information or legal amount on the basis of lexicon and OCR recognition. The flowchart is illustrated in Figure 8.8, which includes three parts such as image partition, OCR over VO string, and VO Block selection based on lexicon and inference.

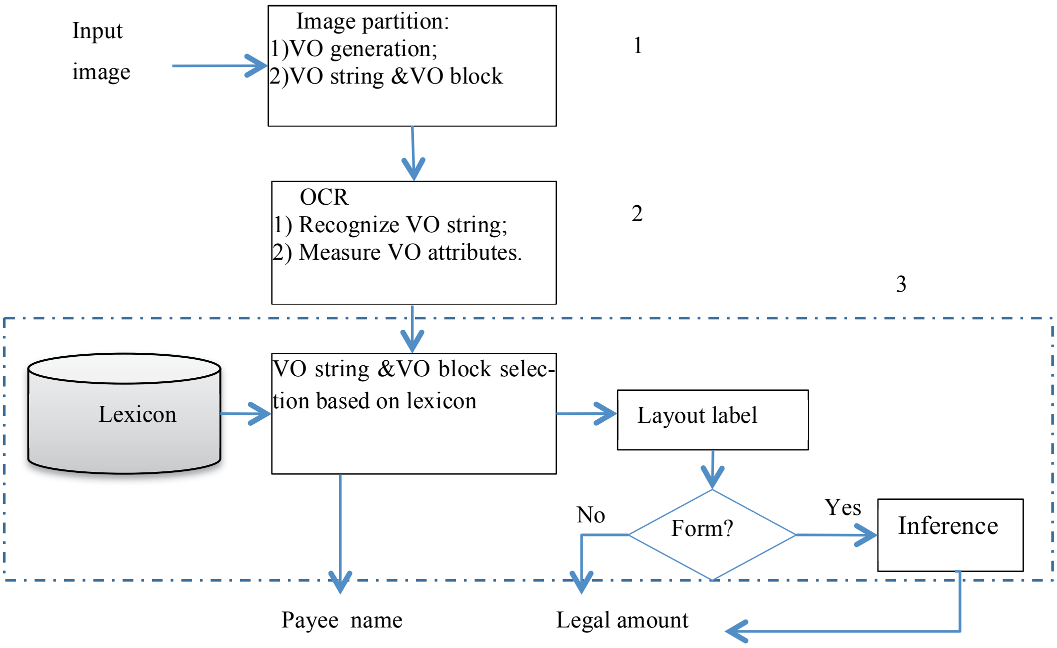


Figure 8.8 The flowchart of payee and legal amount extraction.

In the image partition, the image is split into a set of VO by the adjacent grouping through the connected component analysis, and then the VO strings and VO blocks are obtained according to proximity and similarity. Over VO strings and VO blocks, OCR recognizes them and verifies the keywords of the payee and legal amount, consequently, VO strings and VO blocks are selected for the compact filled payee name and legal amount, or inferred from those separate strings that are filled in various tables through path analysis.

8.2.2 Inferring

Besides one string of compact text, the legal amount may be filled as several separate strings of text in a wide variety of tables. Not all of them can be recognized correctly by the OCR engine since the accuracy limit of OCR and the deficiency of legibility and readability of the image overlap. For the pure text legal amount, the VO string can be selected as parts of the legal amount if at least one substring is recognized as a keyword in the lexicon. The VO string's nested VO block is then considered as the region of the legal amount. But for the legal amount filled in various tables, it needs reliable inference from the certain parts to uncertain ones based on prior knowledge of the financial form structure. This brings us two basic practical issues: how to know the legal amount has been filled in the table, and how to infer the table structure.

We apply the Bayesian inference theory to analyse the table. Suppose there are two linear functions, f and g, of two variables, s and r, of the form

$$\mathrm{f}(s) = as + d, \tag{8.12}$$

And

$$\mathrm{g}(s,r) = bs + cr + d', \tag{8.13}$$

Where a, b and c are the important slope parameters, and d and d' are constants that play no essential role in this theory. We introduce a third variable, y, into this system via the definition

$$y = g(s,r) \tag{8.14}$$

and we assume that r and s are related by the functional relationship

$$r = \mathrm{f}(s). \tag{8.15}$$

Such a system captures the idea that r is functionally dependent on s and y is functionally dependent on s and r. Since f and g are linear, changes in y and r are determined by the slope parameters, a, b and c. This system may be represented by the "path" diagram in Figure 8.9. The coefficients a, b and c are the "path" coefficients, or the "direct effects"; i.e., a is the direct effect of s on r, c is the direct effect of r on y, and b is the direct effect of s on y.

The “total effect” of s on y is found by substituting the equation for r into that for y. This yields

$$y = \mathrm{g}(s, \mathrm{f}(s)) = bs + c(as + d) + d' = (b + ca)s + (cd + d')$$

so that

$$y = (b + ac)s + d'' \tag{8.16}$$

Hence, the total effect of s on y is $b + ac$, which may also be calculated as the sum of the products of all the direct effects along all the paths connecting s and y in the path diagram in Figure 8.9 a); i.e., s to y yields b, and s to r to y yields ac, so the sum is $b + ac$.

r
a
c
b
s
y
R
a
c
b
S
Y

a)

b)

Figure 8.9 “Path” diagram

In our case, suppose there is a population U of “units” of the header row of a financial table, and for each unit u in U we can obtain measurements on three numerical variables, $S(u)$, $R(u)$, and $Y(u)$. In our application, the units are words of digits in the table; $S(u) = 1$ if u is encouraged to search, and $S(u) = 0$ if otherwise; $R(u)$ is the size of the VO string of legal amount in the digit u; and $Y(u)$ is u's the interest strings of the table cells.

As u varies over$U$, $(S(u), R(u), Y(u))$forms a trivariate distribution. This distribution can be used to define quantities such as the conditional expectation of R given$S$, $E(R|S = s)$. This conditional expectation is the average value of R for those units in U for which S(u) = s. The conditional expectation, $E(Y|R = r, S = s)$, has a similar definition in terms of averages over U. The expected value $E(Y|R = r, S = s)$ is the " true" regression function of Y on R and S in the sense that it is what one is trying to estimate by a least the squares regression fit of Y regressed on R and S.

And for our case, the table space structure can be considered as linear system. From Figure 8.9, there is a natural "causal order" to the variables $S, R,$and Y: S comes first, then R, and then Y. A path analysis uses a causal ordering to focus on certain regression functions; in the encouragement design, they are the two described above: $E(R|S = s)$ and $E(Y|S = s, R = r)$. Suppose, for simplicity, that they are both linear, i.e., that

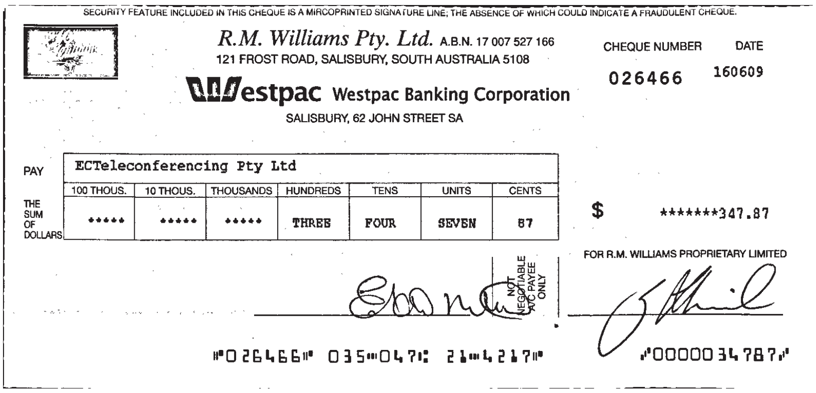
$$E(R|S = s) = \mathrm{f}(s) = as + d \tag{8.17}$$

$$E(Y|S = s, R = r) = g(s, r) = bs + cr + d'. \tag{8.18}$$

Since we are dealing with the measurements (OCR results) $S, R,$ and$Y$ rather than the abstract variable $s, r,$ and$y$, we relabel the nodes of the graph $S, R,$ and$Y$, as in Figure 8.9(b). The path coefficients in Figure 8.9 are just the (population) linear regression coefficients that may be estimated by a (linear) regression of R on$S$, and of Y on $S$and$R$. The same terminology is used as before for the direct effects: The regression coefficients are the direct effects. The "total effect" of S on$Y$, i.e.$b + ac$, can be interpreted as the coefficient of S in the regression of Y on$S$ alone:

$$\begin{aligned} E(Y|S) &= E(E(Y|S, R)|S) = E(bS + cR + d'|S) \\ &= bS + cE(R|S) + d' \\ &= bS + c(aS + d) + d' \\ &= (b + ca)S + dc + d' \end{aligned} \tag{8.19}$$

We use the phrase empirical path diagram to refer to any path diagram constructed from a causal ordering and the implied set of linear regression functions on units in the table structure. An empirical path diagram is, therefore, simply the result of computing certain regression coefficients and arranging them in the appropriate places in the diagram. Based on the relations we can estimate the effective VO strings which are filled in the various tables.



SECURITY FEATURE INCLUDED IN THIS CHEQUE IS A MIRCOPRINTED SIGNATURE LINE; THE ABSENCE OF WHICH COULD INDICATE A FRAUDULENT CHEQUE.

R.M. Williams Pty. Ltd. A.B.N. 17 007 527 166
121 FROST ROAD, SALISBURY, SOUTH AUSTRALIA 5108
Westpac Westpac Banking Corporation
SALISBURY, 62 JOHN STREET SA

CHEQUE NUMBER 026466
DATE 160609

PAY ECTeleconferencing Pty Ltd

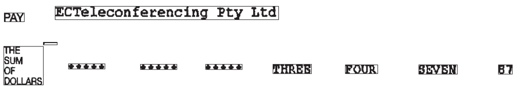
	100 THOUS.	10 THOUS.	THOUSANDS	HUNDREDS	TENS	UNITS	CENTS
THE SUM OF DOLLARS	*****	*****	*****	THREE	FOUR	SEVEN	87

$ *******347.87

FOR R.M. WILLIAMS PROPRIETARY LIMITED

NOT NEGOTIABLE A/C PAYEE ONLY

026466 035 047 21 4217 0000034787



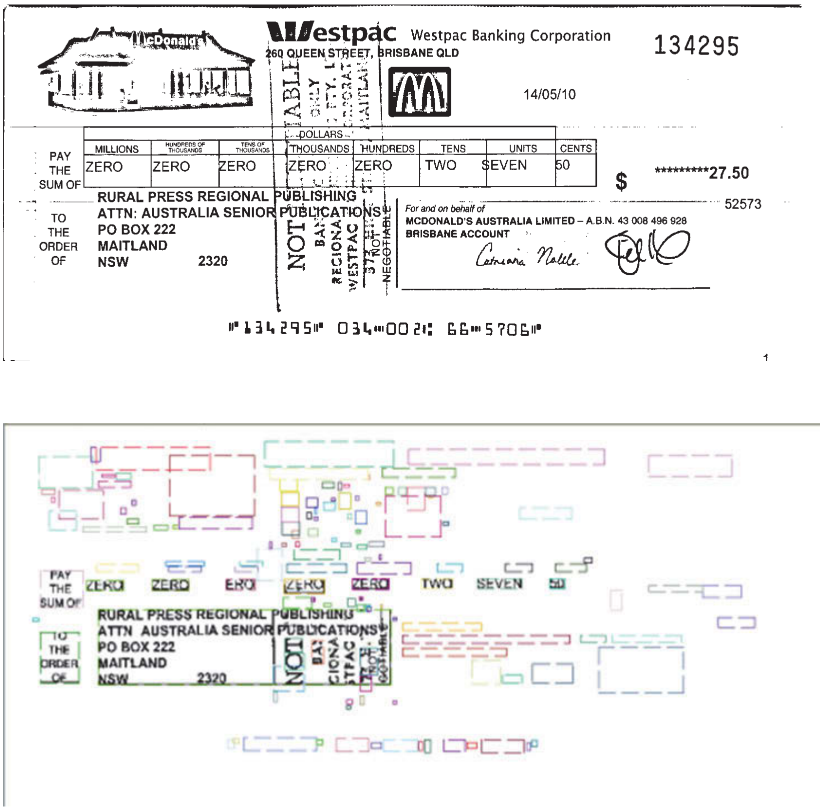
PAY ECTeleconferencing Pty Ltd

THE SUM OF DOLLARS ***** ***** ***** THREE FOUR SEVEN 87

Figure 8.10 Payee name and legal amount in table cheque are extracted

For example, illustrated in Figure 8.10, the stars filled in the table are obtained by path analysis. Here, “PAY” is a keyword for “Payee”, starting from it, the payee name of “ECTeleconferencing Pty Ltd” is figured out. And “The SUM OF DOLLARS” is a keyword for legal amount, both keywords are verified by OCR. And such units as “CENTS, UNITS, TENS, HUNDREDS, THOUSANDS, 10 THOUS, 100 THOUS”, are members of the population U of “units” of the header row of a financial table. For each unit u, we obtain measurements on three numerical variables $S(u)$, $R(u)$, and $Y(u)$. $S(u)=1$ if u exists to encouraged to search, and $S(u)=0$ if otherwise; $R(u)$ is the size of the VO string of legal amount in the digit u; and $Y(u)$ is u's the interest strings of the table cells. Through the path analysis, the table structure can be rebuilt, and the interested strings filled in table cells are found out, including “87, SEVEN, FOUR, THREE, *****,*****,*****.”

To some extent, this method can extract data with touching and overlapping in various fields of information, shown in Figure 8.11, there is a stamp overlaid on the cheque. By the image partition, effective VO strings are labelled in different colours, and then OCR verifies the keyword of payee “PAY” and legal amount “PAY THE SUM OF”. Starting from the payee keyword, the VO block of the payee name is found out, which includes only part of the stamp. Accordingly, by path analysis among units, table cells are picked out, which filters the stamp out.



Westpac Westpac Banking Corporation
260 QUEEN STREET, BRISBANE QLD

134295

14/05/10

PAY THE SUM OF

			DOLLARS				
MILLIONS	HUNDREDS OF THOUSANDS	TENS OF THOUSANDS	THOUSANDS	HUNDREDS	TENS	UNITS	CENTS
ZERO	ZERO	ZERO	ZERO	ZERO	TWO	SEVEN	50

$ *********27.50

52573

TO THE ORDER OF

RURAL PRESS REGIONAL PUBLISHING
ATTN: AUSTRALIA SENIOR PUBLICATIONS
PO BOX 222
MAITLAND
NSW 2320

NOT NEGOTIABLE

For and on behalf of
MCDONALD'S AUSTRALIA LIMITED – A.B.N. 43 008 496 928
BRISBANE ACCOUNT

⑈134295⑈ 034⑆002⑆ 66⑈5706⑈

1

Figure 8.11 Payee and legal amount extraction based on VO strings

8.3 System

8.3.1 System block diagram

All the above functions are integrated into a whole system. And the overall system block diagram is shown in Figure 8.12. For the purpose of fraud detection, both the measurement of the physical appearance and its semantic meaning are important. The system provides both the required measurement of the physical appearance of text and the semantic string content of them. And the task mainly consists of three parts: signature extraction, payee name extraction and legal amount extraction.

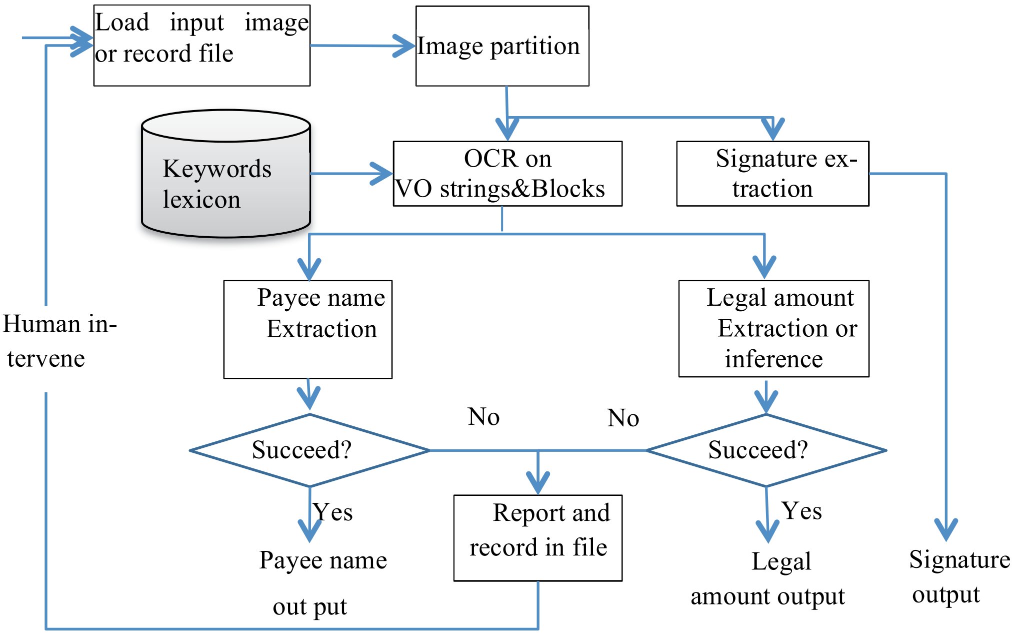


Figure 8.12 The overall system block diagram

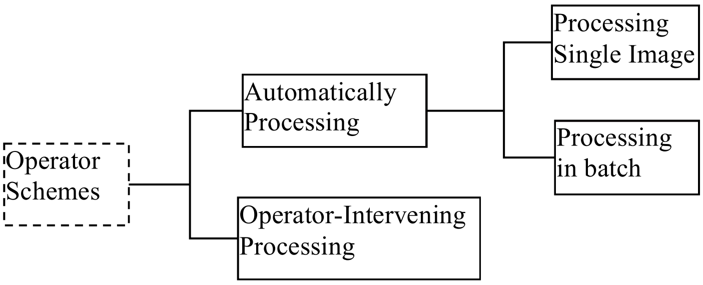


Figure 8.13 Operator schemes of system

Through context-aware saliency, signatures are extracted. And based on the keywords lexicon, the keywords for payee and legal amount are found out from the text strings resulting from OCR operations on VO strings. Meanwhile the OCR engine also measures the physical appearances of each character, such as height, width, stroke width etc.

Starting from the keyword, the payee name can be figured out from VO strings. Similarly, the whole text string of the legal amount is picked out according to its keywords. Besides this whole text string type, separate strings of the legal amount as filled out in various tables can also be searched by the path analysis of table units and table cells.

When one form of detections fails, the system will report the failure and record the file so that humans can intervene to fix this problem. Therefore, besides the completely automated processing in batches, this system also provides operator-intervening processes, shown in Figure 8.13. When the system automatically runs, it processes cheque images automatically in batches or in single cheque images, generates results and records the image file names of them without results and corresponding problems. And in the mode of operator-intervening running, the operator focuses on those images in the record file generated by the automatic running process. According to the record file, the operator loads the failed cheque image, and labels the failed area reported by recording, and then the system provides the measurement of each characters and interested string text in the labelled region.

8.3.2 System evaluation

8.3.2.1Data

Our system has effectively been tested on 569,737 real bank cheque document images with 49.8% machine-printed cheque images.

8.3.2.2 Key criterion

Since our system provides automatic running and human-intervention, the performance of the system depends on the automatic detection rate and operator intervention rate. They are defined with the frame of accuracy threshold (AT) (%), which is as high as possible, ideally 100%, also depends on the quality of cheque images.

Definition 1: Accuracy Threshold (AT) (%)

$$\text{AT\%} = \frac{\#\ \ \text{of Char. Detected in Key Area}}{\text{Real \# of Char. in Key Area}} \times 100\%, \tag{8.20}$$

Note that the real number of characters excludes the char. completely occluded by other information.

Definition 2: Automatic Detection Rate (ADR) (%)

$$\text{ADR \%} = \frac{\text{\# of Successful Cheques Processed Automatically *}}{\text{Total \# of Cheques to be Processed}} \times 100\%, \tag{8.21}$$

*Note that this is fully processed by the computer without operator intervention. We pursue ADR% as high as possible, ideally 100% if our knowledge consolidation is comprehensive. And for ADR%, it consists of the rate of successful cheques detected. The successful cheques processed automatically means the automatic detection rate of text of this system can come up with the readable rate that human eyes can do in the term of one accuracy threshold.

The rate of successful cheques detected automatically (ADSR%) and the rate of unsuccessful cheques detected automatically (ADUR %), and

$$\text{ADSR\%+ADUR\%=ADR\%} \tag{8.22}$$

Definition 3: Operator Intervention Rate (OIR) (%)

$$\text{OIR\%} = \frac{\text{\# of Successful Cheques Processed Manually *}}{\text{Total \# of Cheques to be Processed}} \times 100\%, \tag{8.23}$$

*Note that candidate cheques are given by pre-processing procedure. That is, if cheques cannot be processed by a computer, they will automatically ask for operator intervention. It is as low as possible, and it may be higher initially due to incomplete knowledge consolidation at the beginning. For OIR%, it includes the rate of successful cheques intervened by operators (OISR %) and the rate of unsuccessful cheques intervened by (OIUR %), and

$$\text{OISR\%+OIUR\%=OIR\%} \tag{8.24}$$

Table 8.3 The performance of system on real data (569,737 images)

	Accuracy Threshold (AT) (%)															
	70%				80%				90%				95%			
	ADR(%)		OIR(%)		ADR(%)		OIR(%)		ADR(%)		OIR(%)		ADR(%)		OIR(%)	
	ADSR%	ADUR%	OISR%	OIUR%	ADSR%	ADUR%	OISR%	OIUR%	ADSR%	ADUR%	OISR%	OIUR%	ADSR%	ADUR%	OISR%	OIUR%
2	84.54%		15.46%		84.54%		15.46%		84.54%		15.46%		84.54%		15.46%	
	83.24%	1.30%	15.46%	0%	82.47%	2.07%	15.46%	0%	82.47%	2.07%	15.46%	0%	82.47%	2.07%	15.46%	0%
	DR%=ADSR%+OISR%=98.70% **ER%=ADUR%+OIUR%=1.30%**				**DR%=ADSR%+OISR%=97.93%** **ER%=ADUR%+OIUR%=2.07%**				**DR%=ADSR%+OISR%=97.93%** **ER%=ADUR%+OIUR%=2.07%**				**DR%=ADSR%+OISR%=97.93%** **ER%=ADUR%+OIUR%=2.07%**			
3	98.10%		1.90%		98.10%		1.90%		98.10%		1.90%		98.10%		1.90%	
	96.23%	1.87%	1.90%	0%	95.37%	2.73%	1.90%	0%	92.70%	5.40%	1.90%	0%	91.64%	6.46%	1.90%	0%
	DR%=ADSR%+OISR%=98.13% **ER%=ADUR%+OIUR%=1.87%**				**DR%=ADSR%+OISR%=97.27%** **ER%=ADUR%+OIUR%=2.73%**				**DR%=ADSR%+OISR%=94.60%** **ER%=ADUR%+OIUR%=5.40%**				**DR%=ADSR%+OISR%=93.54%** **ER%=ADUR%+OIUR%=6.46%**			

Definition 4: Error Rate (ER) (%)

$$ER\% = \frac{\begin{array}{c}\#\text{ of Cheques Unsuccessfully Processed}\\ \text{Neither by computer nor by operator}\end{array}}{\text{Total }\#\text{ of Cheques to be Processed}} \times 100\%, \qquad (8.25)$$

And

$$ER\%=ADUR\%+OIUR\% \qquad (8.26)$$

It is as low as possible, ideally near to 0%. And they satisfy with the condition

$$ADR\% + OIR\% + ER\% = 100\% \qquad (8.27)$$

Our system has been tested on the real bank cheque document images, and the performance of our system is shown in Table8.3. Table 8.3 shows the system is stable and effective. With the accuracy threshold of 80%, 90% and 95%, the successful detection rate of the payee name is stable at 97.93%. And for legal amount, the successful detection rate is found to be 98.13%, 97.27%, 94.60%, and 93.54% with the corresponding accuracy thresholds of 70%, 80%, 90%, and 95% respectively.

8.4 Summary

Automation of bank check processing is an important and promising application of document recognition techniques. Cheque document image processing systems should be designed as important tools for reading automatically the wide variety of cheques provided by relevant units and individuals.

The difficulties in developing an effective cheque reading system are the high degree of variability and uncertainty in the user-entered date information. People print or write the data zones in free style and there is no fixed format for cheques. There are differences not only in background, but also in the type and position of the machine printed and handwritten information. The area of interest should be located first in those systems, which do not depend on specific cheque formats. And the three data zones are necessary to locate the payee name, legal amount and signature. All of them work to address the basic issue, how to segment or partition the image?

The representations of the image based on space regularity and three-level computational modelling of text in the cluttered scene can also have utility in relation to solving this basic issue. And based on the image partition, the signature can be extracted by the CSSD algorithm, which also provides the important feature of discerning handwritten text from printed text. Meanwhile, based on the OCR, the payee name is found out and measured, and the legal amount is also figured out by path analysis or the VO string selection lexicon.

Chapter 9

Conclusion

In relation to the figures in image in the clutter scene, it is the physical appearance of text that provides the perceptual content and plays a central role for text detection, i.e. location and coarse identification. When observers see text appearance in a clutter scene, they describe their feelings in perceptual terms e.g. crowding effect and clutter. However, the appearance of text still has enough saliency to reveal an informative message. Accordingly, text not only has the characteristics of crowding effect and clutter but also follows the principles of saliency. If we consider the textual coexistence of crowding, clutter and saliency, we come to understand the correlates among them, both at the beginning and end of the text detection process.

To understand the correlates, there are multilevel considerations, including local low-level, global level, visual organization and high-level considerations. In low local-level considerations, such as local contrast and orientation, they usually reflect the informative locations. At the global level, crowding and saliency can be thought of as two sides of the one coin. Crowding has the property of space averaging over the pooling region, while saliency tries to pop up the proto objects. Since text is composed of naturally inbuilt neighbouring letters, space averaging represents the inherent characteristics while the distinctiveness among letters helps to maintain enough saliency. It is necessary to figure out the distinctiveness (high-level factors) through inspecting the properties derived from font design, and this is also what we need to do in terms of high-level considerations. The consideration in high-level perception includes the features of individual characters and the similarity or dissimilarity of them, which crowding, clutter and saliency all depend on. In visual organization, the spatial element and its relations (i.e., reasonable regions and features) are significant. For saliency, the region is that given by salient structures or proto objects; for crowding the region is the pooling region, and for clutter, the region is the “interested region”. Once these fundamental steps are achieved, we can then go on to build up the computational three-level text model and the representation of an image by quantifying the space organization. Following this, the algorithm and the system of text detection in the clutter scene can be developed. This chapter provides the conclusions and recommendations that have resulted from this study.

Firstly, on the basis of the analysis and summarization of the theory and properties of crowding effect, saliency, and clutter, the correlation among them is obtained in three-levels, including the feature level, space organization, and the crowding space averaging level in a more global manner. This brings up to two basic subtasks: 1) computationally make tracks of the discriminative features for text legibility, readability and conspicuity; 2) interested regions or pooling regions need to be generated or formed to break down, or at least decrease the crowding to make our target pop out. In such a way, the image is represented on the basis of quantifying the space organization.

Secondly, since the font stylish attributes and textual organisation contribute to the essential functions of the physical appearance during the view construction of text in the local level type design, they provide us with a comprehensive understanding of the roles and the features of individual characters as many of them cannot be adequately caught through image processing alone. After investigating, we transfer them into image-based reasonable attributes of individual character, or the measure of the appearance of similarity in space regularity. In image processing, the attributes of individual characters consist of local RMS contrast, local mean intensity, edge density, pixel density, orientation and directions, height to width ratio, stroke with to height ratio, straight line to size ratios, and shape (e.g., grey patches occupied by character, contour, etc). Further, the measure of neighbourhood is defined by the distance between shapes according to the size of shape. Meanwhile, the measure of the appearance similarity is defined as the Gaussian function of the shape size. Moreover, to combine the neighbourhood and appearance similarity, the adjacent relation is defined among the characters.

Thirdly, for the purposes of quantifying the space organization in crowding and clutter, spatial elements and relations need to be calculated.

For spatial elements, according to features of individual characters, region-based spatial elements are preferred. According to the theory of perspective, when objects recede from the eye or camera, the size of the objects decreases; this means that the space of the object in the image is reduced while the contour is lessened. If the distance is far enough away, the contours or boundary of separated objects disappear and those separated objects as parts become manifest after they are merged into a whole. However, the space occupied by the object exists and has discernible information. This suggests that there are two spatial elements that need to be represented: the space occupied by the object and the contour or edge of the object. However, contour can be distorted in the clutter scene. Owing to this, a connected component analysis is applied to generate connected components to represent the space patches and image is represented as a corps of multi-grey level grey patches.

For relations among spatial elements, besides the common adjacent relation, there is another essential relation: proportions among component parts in image for proportions in all things. For adjacent relations defined by proximity and similarity among grey patches, clusters are obtained. For proportions, we receive inspiration from painters. If we examine the works of painters, especially the Impressionists, they use directional brushstroke or colour patches in repetitive patterns to represent a “formless” visual instead of clear shape sketches. These repetitive patterns can offer a compositional format to express an artist’s feelings about an object rather than to simply describe it. Furthermore, beside the adjacent relation in space, painters apply harmonious proportions among component parts to bridle them into visual objects. It is the painter’s harmonious proportions that make the component parts of an object react simultaneously so that they can be seen at one and the same time, both together and separately. In mathematics, the geometric mean (GM) can capture the ratios to the reference value. If we, therefore, apply it to the grey patches, the proportions of component parts in the whole image or an object can be captured implicitly. GM resembles the spatial granularity of an image. Since GM can find the “figure of merit”, it is defined as an indicator of spatial granularity and it gets involved in object constitution. Image can be composed of the GM regions in different GM levels. Based on GM, the constitution can be explored and we can recognize that the different grey levels forming the same visual object are inclined to have a similar GM value. Thus, the features of individual characters can be extended to the GM level and visual objects can be analysed at the GM level. Further, two kinds of pooling regions are clearly shown and these include regions given by clusters generated from adjacent relations and GM regions.

Based on these studies, the image is represented by a set of GM regions at several GM levels. In addition, the three-level computational model of text is built up and calculated over GM regions, including the feature-level, letter-centred level, and word-centred level. At the feature level, there are attributes of individual characters in image processing. At the letter-centred level, there are the attributes of the space relations among letters derived from the similarity of appearance and proximity of position, i.e. the adjacent relation. At the word-centred level, the attributes are defined over GM regions. Over GM regions, there might be one or several clusters. Word-centred attributes include the statistical properties of each cluster and the relations of their space location.

Finally, the computational model of text and image representation by the GM regions is put into practice for the purposes of developing a new algorithm of text detection in the clutter scene and a system for the automatic processing of a big data i.e. the real bank cheque. The performance of the algorithm is revealed to be comparable and the grey compositional structure of text becomes available for analysis. Notably too, the performance of the automatic processing system of the bank cheque is shown to be effective.

The system has been run on the real data (569,737 images). With the accuracy threshold of 70%, the successful detection rate of the payee name is found to be 98.7%. With the accuracy threshold of 80%, 90% and 95%, the successful detection rate of the payee name is stable at 97.93%. For the legal amount, the successful detection rate is found to be 98.13%, 97.27%, 94.60%, and 93.54% with corresponding accuracy thresholds of 70%, 80%, 90%, and 95% respectively.

The algorithm of the text detection based on the crowding model of text is submitted to Pattern Recognition, and the automatic processing system of the bank cheque is submitted to IJCAR.

In the future, the thesis works can be further studied in image processing, vision perception, and non iid study.

Firstly, since GM captures the proportions among component parts in an image and is regarded as the essential indicator of the spatial granularity in image, it can be used to analyse the intrinsic structure of the image from the standpoint of visual object composition. Therefore, GM together with features in its level can be applied to image segmentation, semantic labelling and image quality assessment. Also, it can be used to find the “figure of merit”, i.e. the most salient figures in an image and the GM method can be used to measure the crowding effect and clutter.

Secondly, our work provides a concrete computational formulation for crowding study which contributes to the important and growing research base examining the computation and application of the crowding effect.

Thirdly, since text computation model is built up on the basis of the unitary process of text perceiving, and other visual objects can be represented or understood in the same way from the standpoint of vision perception.

Additionally, spatial elements and relations among them are quantified, which suggest typical properties of non iid from the standpoint of composition of image, therefore this thesis work can be extended or become one typical case study in the growing field of non iid[314] research.

Appendix A

In this appendix, the matrix properties employed throughout the section 5.2.2 are reviewed. Let $\mathbf{A}$ be an $n \times p$ matrix having rank r. Without loss of generality will assume $r \le p \le n$. The trace of the matrix is

$$\text{trace}[\mathbf{A}] = \sum_{i=1}^{p} a_{ii} = \sum_{i=1}^{r} \lambda_i, \tag{A.1}$$

where λ_i are the eigenvalues of $\mathbf{A}$. The trace of a scalar is the scalar itself and the trace has the following invariance properties:

$$\begin{aligned} &\text{trace}[\mathbf{A}] = \text{trace}[\mathbf{A}^{\mathrm{T}}] \\ &\text{trace}[\mathbf{ABC}] = \text{trace}[\mathbf{CAB}] = \text{trace}[\mathbf{BCA}], \end{aligned} \tag{A.2}$$

Where $\mathbf{B}$ and $\mathbf{C}$ are matrices with corresponding dimensions. The invariance of trace to cyclic permutations is an important property which can often simplify matric manipulations.

The inner (scalar) product of two $n \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$

$$(\mathbf{AB}) = \text{trace}[\mathbf{A}^{\mathrm{T}}\mathbf{B}] = \text{trace}[\mathbf{B}^{\mathrm{T}}\mathbf{A}] \tag{A.3}$$

Satisfies all the well-known properties of inner product. The Frobenius norm of the matrix $\mathbf{A}$

$$\|\mathbf{A}\|_F^2 = \|\mathbf{A}^{\mathrm{T}}\|_F^2 = (\mathbf{A},\mathbf{A}) = \text{trace}[\mathbf{A}^{\mathrm{T}}\mathbf{A}] = \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}^2 \tag{A.4}$$

Is often used and the Cauchy-Schwartz inequality becomes

$$\left|trace[\mathbf{A}^{\mathrm{T}}\mathbf{B}]\right| \le \|\mathbf{A}\|_F \|\mathbf{B}\|_F \tag{A.5}$$

With equality iff $\mathbf{A} = \alpha\mathbf{B}$.

The singular value decomposition (svd) of $\mathbf{A}$ is defined as

$$\mathbf{A} = \mathbf{U\Sigma V}^T, \tag{A.6}$$

where $\mathbf{U}$ is an $n \times n$ and $\mathbf{V}$ a $p \times p$ orthonormal matrix. The $n \times p$ diagonal matrix Σ has r positive numbers arranged in descending order, the singular values σ_k of $\mathbf{A}$. The nonzero eigenvalues of $\mathbf{AA}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are σ_k^2. The Frobenius norm of $\mathbf{A}$ is then

$$\|\mathbf{A}\|_F = \left(\sum_{k=1}^{r} \sigma_k^2\right)^{1/2}. \tag{A.7}$$

The column vectors of $\mathbf{U}$ and $\mathbf{V}$ provide orthonormal bases for the different subspaces associated with matrix $\mathbf{A}$. The vector $\{u_1, u_2, \ldots, u_r\}$ span the rang $\mathrm{R}(\mathbf{A})$, and the vectors $\{v_{r+1}, v_{r+2}, \ldots, v_p\}$ span the null space $\mathrm{N}(\mathbf{A})$, while $\{v_1, v_2, \ldots, v_r\}$ span $\mathrm{R}(\mathbf{A}^{\mathrm{T}})$ and $\{u_{r+1}, u_{r+2}, \ldots, u_n\}$ span the null space $\mathrm{N}(\mathbf{A}^{\mathrm{T}})$. Thus, $\mathrm{R}[\mathbf{A}]$ and $\mathrm{N}[\mathbf{A}^{\mathrm{T}}]$ are orthonormal complements in R^n, while $\mathrm{R}(\mathbf{A}^{\mathrm{T}})$ and $\mathrm{N}[\mathbf{A}]$ are orthonormal complements in R^p.

Let the vectors $\{b_1, b_2, \ldots, b_q\}$ be an orthonormal basis for a $q \le n$ dimensional subspace $S \subseteq R^n$, the $n \times n$ projection matrix $\mathbf{P}$

$$\mathbf{p} = \sum_{k=1}^{q} b_k b_k^T \qquad \mathbf{P} = \mathbf{P}^{\mathbf{T}} \qquad \mathbf{P}^2 = \mathbf{P} \cdot \mathbf{P} = \mathbf{P} \tag{A.8}$$

has rank q, it is symmetric and idempotent, and projects orthogonally onto S. The rank $n - q$ matrix

$$\mathbf{Q} = \mathbf{I}_{\mathbf{n}} - \mathbf{P} \tag{A.9}$$

is the projection matrix onto the orthogonal complement of S in R^n.

The operator $\mathrm{vec}[\mathbf{A}]$ yields the vector a obtained by stacking up the columns of $\mathbf{A}$. It can be shown that

$$\mathrm{trace}[\mathbf{AB}] = \mathrm{vec}[\mathbf{A}^{\mathbf{T}}]^{\mathbf{T}} \mathrm{vec}[\mathbf{B}], \tag{A.10}$$

Let $f(\mathbf{A})$ be a scalar valued function of the matrix $\mathbf{A}$ and assume that

$$\mathbf{A} = \mathbf{A}_o + \delta\mathbf{A}, \tag{A.11}$$

Where $\mathbf{A}_o$ the uncorrupted is "true" value and $\delta\mathbf{A}$ is a zero-mean perturbation matrix with i.i.d. elements. Thus,

$\mathrm{vec}\,\delta\mathbf{A} = \delta a \sim G(0, \sigma^2 \mathbf{I}_{np})$. The variance of $f(\mathbf{A})$ can be approximated by error propagation. The linear approximation of $f(\mathbf{A})$ around $\mathbf{A}_o$ is obtained from the Taylor expansion

$$f(\mathbf{A}) = f(a) = f(a_0 + \delta a) \approx f(a_0) + \nabla f^T \delta a, \tag{A.12}$$

where ∇ is the gradient of f with respect to a computed in a_0. Assuming that the plug-in principle holds (the function of the mean can be used as substitute for the mean of the function) the variance becomes

$$\mathrm{var}[f(\mathbf{A})] \approx \sigma^2 \nabla f^T \nabla f = \sigma^2 \,\mathrm{trace}\left[\left(\frac{\partial f}{\partial \mathbf{A_0}}\right)^{\mathbf{T}} \frac{\partial f}{\partial \mathbf{A_0}}\right], \tag{A.13}$$

Where the derivation of a scalar function with respect to a matrix is the gradient matrix having as the ijth element $\frac{\partial f}{\partial a_{ij}}$. The gradient matrix is computed for the true value $\mathbf{A_0}$. The following gradient matrices:

$$\frac{\partial\, \mathrm{trace}[\mathbf{WA}]}{\partial \mathbf{A}} = \mathbf{W}^{\mathbf{T}} \qquad \frac{\partial\, \mathrm{trace}[\mathbf{W}^{\mathbf{T}}\mathbf{A}]}{\partial \mathbf{A}} = \mathbf{W} \tag{A.14}$$

BIBLIOGRAPHY

[1] R. A. Frazor and W. S. Geisler, "Local luminance and contrast in natural images," *Vision Research,* pp. 1585-1598, 2006.

[2] P. Meer and B. Georgescu, "Edge Detection with Embedded Confidence," *IEEE Trans on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 1351-1365, 2001.

[3] F. Brown, "A Study of the Requirements for Letters, Numbers, and Markings to be used on Trans-Illuminated Aircraft Control Panels(Part 4-Legibility of Uniform Stroke Capital Letters as Determined by Size and Height to Width Ratio and as Compared to Garamond Bold)," Philadelphia, PA: Naval Air Material Center1953.

[4] T. V. Z. Robert W. Proctor, *Human Factors in Simple and Complex Systems, Second Edition*: CRC Press, 2008.

[5] W. Thomas S. Tullis, "The Formatting of Alphanumeric Displays: A Review and Analysis," *Human Factors,* vol. 25, pp. 657-682, 1983.

[6] S.M.Lucas, A. Panaretos, L.Sosa, A. Tang, S.Wong, and R.Young, ""ICDAR2003"robust reading competitions," in *Proc. ICDAR*, 2003.

[7] S. M. Lucas, "ICDAR 2005 text locating competition results," in *Proceedings of the International. Conference on Document Analysis and Recognition*, 2005, pp. 80–84.

[8] D. M. Levi, "Crowding--an essential bottleneck for object recognition: a mini-review," *Vision Res,* vol. 48, pp. 635-54, Feb 2008.

[9] R. Rosenholtz, Y. Li, and L. Nakano, "Measuring visual clutter," *J Vis,* vol. 7, pp. 17 1-22, 2007.

[10] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans Pattern Anal Mach Intell,* vol. 34, pp. 1915-26, Oct 2012.

[11] X. B. Cong Yao, Wenyu Li, Yi Ma, Zhouwen Tu, " Detecting texts of arbitrary orientations in natural images," in *IEEE CVPR*, 2012, pp. 1083 - 1090

[12] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform," in *CVPR*, 2010, pp. 2963 – 2970.

[13] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans Image Process,* vol. 20, pp. 2594-605, Sep 2011.

[14] K. I. K. Keechul Jung, Anil K. Jain, "Text information extraction in images and video: a survey," *Pattern Recognition* vol. 37, pp. 977 – 997, 2004.

[15] D. D. Huiping Li, and Omid Kia, "Automatic Text Detection and Tracking in Digital Video," *IEEE Transactions on Image Processing,* vol. 9, pp. 147-156, 2000.

[16] M. B. Nobuo Ezaki, Lambert Schomaker, "Text Detection from Natural Scene Images:Towards a System for Visually Impaired Persons," in *ICPR* 2004, pp. 683 – 686.

[17] T. M. Breuel, "The OCRopus open source OCR system," in *IS&T/SPIE 20th Annual Symposium*, 2008.

[18] D. D. a. D. D. J. Liang, "Geometric Rectification of Camera-Captured Document Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, pp. 591-605, 2008 2008.

[19] A. M. N. a. C. V. J. J. Banerjee, "Contextual restoration of severely degraded document images," in *IEEE CVPR*, 2009, pp.,517-524.

[20] S. L. a. C. L. Tan, "Retrieval of machine-printed Latin documents through Word Shape Coding," *Pattern Recognition,* vol. 41, pp. 1799-1809, May 2008 2008.

[21] A. L. Y. X. Chen, "Detecting and reading text in natural scenes," in *IEEE CVPR*, 2004, pp. 366-373.

[22] J. Y. X. Chen, J. Zhang, and A. Waibel, "Automatic Detection and Recognition of Signs From Natural Scenes," *IEEE Transactions on Image Processing,* vol. 13, pp. 87-99, 2004 2004.

[23] W. H. a. C. L. T. P. Shivakumara, "An Efficient Edge based Technique for Text Detection in Video Frames," in *the Eighth IAPR Workshop on Document Analysis Systems*, 2008, pp. 307-314

[24] P. S. a. C. L. T. T. Phan, "A Laplacian Method for Video Text Detection," in *the 10th International Conference on Document Analysis and Recognition*, 2009, pp. 66-70.

[25] C. Y. S. Wumo Pan T. D. Bui, "Text detection from scene images using sparse representation," in *ICPR*, 2008, pp. 1-5.

[26] O. C. J. Matas, M.Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," presented at the BMVC, 2002.

[27] L. N. a. J. Matas, "Text Localization in Real-World Images Using Efficiently Pruned Exhaustive Search " presented at the ICDAR, 2011.

[28] L. M. Neumann, J., "Scene Text Localization and Recognition with Oriented Stroke Detection " presented at the Computer Vision (ICCV), 2013.

[29] J. M. L Neumann, "Real-time scene text localization and recognition " presented at the CVPR, 2012.

[30] S. T. H. Chen, G. Schroth, D. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," presented at the ICIP, 2011.

[31] X. Y. Xu-Cheng Yin, Kaizhu Huang, and Hong-Wei Hao, "Robust Text Detection in Natural Scene Images," *arXiv* 2013.

[32] X.-C. Y. X. Y. K. H. H.-W. Hao, "Robust Text Detection in Natural Scene Images " *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, pp. 970 - 983 2014.

[33] C. W. Cunzhao Shi, Baihua Xiao, Yang Zhang, Song Gao, "Scene text detection using graph model built upon maximally stable extremal regions," *Pattern Recognition Letters,* vol. 34, pp. 107-116, 2013.

[34] C. J. Q. Liu, and Y. Moon, "Text Segmentation based on Stroke Filter," in *International Conference on Multimedia*, 2006, pp. 129-132.

[35] J. K. a. A. G. R. T. Kasar, "Font and Background Color Independent Text Binarization " in *the Second International Workshop on Camera-Based Document Analysis and Recognition*, 2007, pp. .3-9.

[36] Y. M. Y. H. a. L. J. Karam, "Morphological Text Extraction from Images," *IEEE Transaction on Image Processing,* vol. 9, pp. 1978-1983, 2000.

[37] C.-J. P. K.-A. M. W.-G. O. H.-M. Choi, "An efficient extraction of character string positions using morphological operator " in *IEEE International Conference on Systems, Man, and Cybernetics*, 2000, pp. 1616-1620.

[38] M. S. S. Uddin, M. ; Rahman, T. ; Busra, U.S., "Extraction of texts from a scene Image using morphology based approach," in,*International Conference on Informatics, Electronics & Vision (ICIEV)*, 2012, pp. 876 - 880.

[39] K. J. Kwang In Kim, Jin Hyung Kim, "Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence archive,* vol. 25, pp. 1631-1639, December 2003 2003.

[40] A. L. H. Tran, T. H. L. Nguyen, A Boucher, "A novel approach for text detection in images using structural features," in *ICDAR*, 2005, pp. 627-635.

[41] P. Shivakumara, Sreedhar, R.P.,Trung Quy Phan , Shijian Lu ,Tan, C.L., "Multioriented Video Scene Text Detection Through Bayesian Classification and Boundary Growing " *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, pp. 1227 - 1235 2012.

[42] S. Kumar, Gupta, R. , Khanna, N. , Chaudhury, S., "Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model," *IEEE Transactions on Image Processing,* vol. 16, pp. 2117 – 2128,, 2007 2007.

[43] H. K. K. Sobottka, T. Perroud, and H. Bunke, "Identification of text on colored book and journal covers " in *the 5th International Conference on Document Analysis and Recognition*, Bangalore 1999, pp. 57 - 62

[44] J. G. a. J. Yang, "An Adaptive Algorithm for Text Detection from Natural Scenes," in *IEEE CVPR*, 2001, pp. 84-89.

[45] a. J. F. W. H. M. Suen, "Segmentation of Uniform Colored Text from Color Graphics Background," in *Vision, Image and Signal Processing, IEE Proceedings* 1997, pp. 317-332.

[46] W. Korte, "Uber die Gestaltauffassung im indirekten Sehen," *Zeitschrift fu¨ r Psychologie,* vol. 93, pp. 17-82, 1923.

[47] D. Yu, M. M. Akau, and S. T. Chung, "The mechanism of word crowding," *Vision Res,* vol. 52, pp. 61-9, Jan 1 2012.

[48] Y. Marzouki and J. Grainger, "Effects of stimulus duration and inter-letter spacing on letter-in-string identification," *Acta Psychol (Amst),* vol. 148, pp. 49-55, May 2014.

[49] H. Bouma, "Interaction effects in parafoveal letter recognition," *Nature,* vol. 226, pp. 177-8, Apr 11 1970.

[50] M. Martelli, N. J. Majaj, and D. G. Pelli, "Are faces processed like words? A diagnostic test for recognition by parts," *J Vis,* vol. 5, pp. 58-70, 2005.

[51] S. Zahabi and M. Arguin, "A crowdful of letters: disentangling the role of similarity, eccentricity and spatial frequencies in letter crowding," *Vision Res,* vol. 97, pp. 45-51, Apr 2014.

[52] T. P. Saarela, G. Westheimer, and M. H. Herzog, "The effect of spacing regularity on visual crowding," *J Vis,* vol. 10, p. 17, 2010.

[53] J. Grainger, I. Tydgat, and J. Issele, "Crowding affects letters and symbols differently," *J Exp Psychol Hum Percept Perform,* vol. 36, pp. 673-88, Jun 2010.

[54] D. Whitney and D. M. Levi, "Visual crowding: a fundamental limit on conscious perception and object recognition," *Trends Cogn Sci,* vol. 15, pp. 160-8, Apr 2011.

[55] T. W. IRMA A. RICHTER, MARTIN KEMP, *Leonardo da Vinci Notebooks*: Oxford University Press, 1980.

[56] Hundertwasser, "The straight line is godless," vol. http://www.hundertwasser.com/text/1.3.2.2/hl/62#titleanch, 1968.

[57] M. S. a. M. Sanders, E.J. , *Human Factors in Engineering and Design*. New York McGraw-Hill, 1993.

[58] a. S. E. J. Cole B.L., "Conspicuity of Traffic Control Devices," AIR 218-1, 1978.

[59] P. K. Hughes, and B.L. Cole, "What Attracts Attention When Driving?," *Ergonomics,* vol. 29, pp. 377-391, 1986.

[60] B. L. Cole, and P.K. Hughes, "A Field Trial of Attention and Search Conspicuity," *Human Factors,* vol. 26, pp. 299-313, 1984.

[61] H. Ehlers, "The movements of the eyes during reading," *Acta Ophthalmologica,* pp. 56-63, 1936.

[62] H. Strasburger, L. O. Harvey, Jr., and I. Rentschler, "Contrast thresholds for identification of numeric characters in direct and eccentric view," *Percept Psychophys,* vol. 49, pp. 495-508, Jun 1991.

[63] B. H. Andriessen J, "Eccentric vision: Adverse interactions between line segments," *Vision Research,* vol. 16, pp. 71–78, 1976.

[64] F. Wilkinson, H. R. Wilson, and D. Ellemberg, "Lateral interactions in peripherally viewed texture arrays," *J Opt Soc Am A Opt Image Sci Vis,* vol. 14, pp. 2057-68, Sep 1997.

[65] E. Poder, "Crowding with detection and coarse discrimination of simple visual features," *J Vis,* vol. 8, pp. 24 1-6, 2008.

[66] M. C. Flom, G. G. Heath, and E. Takahashi, "Contour Interaction and Visual Resolution: Contralateral Effects," *Science,* vol. 142, pp. 979-80, Nov 15 1963.

[67] A. Toet and D. M. Levi, "The two-dimensional shape of spatial interaction zones in the parafovea," *Vision Res,* vol. 32, pp. 1349-57, Jul 1992.

[68] D. M. Levi, S. A. Klein, and A. P. Aitsebaomo, "Vernier acuity, crowding and cortical magnification," *Vision Res,* vol. 25, pp. 963-77, 1985.

[69] G. Westheimer and G. Hauske, "Temporal and spatial interference with vernier acuity," *Vision Res,* vol. 15, pp. 1137-41, Oct 1975.

[70] G. Westheimer, K. Shimamura, and S. P. McKee, "Interference with line-orientation sensitivity," *J Opt Soc Am,* vol. 66, pp. 332-8, Apr 1976.

[71] T. W. Butler and G. Westheimer, "Interference with stereoscopic acuity: spatial, temporal, and disparity tuning," *Vision Res,* vol. 18, pp. 1387-92, 1978.

[72] E. G. Louie, D. W. Bressler, and D. Whitney, "Holistic crowding: selective interference between configural representations of faces in crowded scenes," *J Vis,* vol. 7, pp. 24 1-11, 2007.

[73] D. G. Pelli and K. A. Tillman, "The uncrowded window of object recognition," *Nat Neurosci,* vol. 11, pp. 1129-35, Oct 2008.

[74] S. P. Tripathy and P. Cavanagh, "The extent of crowding in peripheral vision does not scale with target size," *Vision Res,* vol. 42, pp. 2357-69, Sep 2002.

[75] P. J. Bex and S. C. Dakin, "Spatial interference among moving targets," *Vision Res,* vol. 45, pp. 1385-98, May 2005.

[76] T. S. Wallis and P. J. Bex, "Image correlates of crowding in natural scenes," *J Vis,* vol. 12, 2012.

[77] F. Farzin, S. M. Rivera, and D. Whitney, "Holistic crowding of Mooney faces," *J Vis,* vol. 9, pp. 18 1-15, 2009.

[78] B. Balas, L. Nakano, and R. Rosenholtz, "A summary-statistic representation in peripheral vision explains visual crowding," *J Vis,* vol. 9, pp. 13 1-18, 2009.

[79] S. C. Dakin, J. Cass, J. A. Greenwood, and P. J. Bex, "Probabilistic, positional averaging predicts object-level crowding effects with letter-like stimuli," *J Vis,* vol. 10, p. 14, 2010.

[80] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nat Neurosci,* vol. 14, pp. 1195-201, Sep 2011.

[81] L. Parkes, J. Lund, A. Angelucci, J. A. Solomon, and M. Morgan, "Compulsory averaging of crowded orientation signals in human vision," *Nat Neurosci,* vol. 4, pp. 739-44, Jul 2001.

[82] J. A. Greenwood, P. J. Bex, and S. C. Dakin, "Positional averaging explains crowding with letter-like stimuli," *Proc Natl Acad Sci U S A,* vol. 106, pp. 13130-5, Aug 4 2009.

[83] R. van den Berg, J. B. Roerdink, and F. W. Cornelissen, "On the generality of crowding: visual crowding in size, saturation, and hue compared to orientation," *J Vis,* vol. 7, pp. 14 1-11, 2007.

[84] J. Freeman, R. Chakravarthi, and D. G. Pelli, "Substitution and pooling in crowding," *Atten Percept Psychophys,* vol. 74, pp. 379-96, Feb 2012.

[85] S. He, P. Cavanagh, and J. Intriligator, "Attentional resolution and the locus of visual awareness," *Nature,* vol. 383, pp. 334-7, Sep 26 1996.

[86] A. S. Nandy and B. S. Tjan, "The nature of letter crowding as revealed by first- and second-order classification images," *J Vis,* vol. 7, pp. 5 1-26, 2007.

[87] H. Strasburger, "Unfocused spatial attention underlies the crowding effect in indirect form vision," *J Vis,* vol. 5, pp. 1024-37, 2005.

[88] H. K. Kim, "Efficient automatic text location method and content-based indexing and structuring of video database," *J. Visual Commun. Image Representation* vol. 7, pp. 336–344, 1996.

[89] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn Psychol,* vol. 12, pp. 97-136, Jan 1980.

[90] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum Neurobiol,* vol. 4, pp. 219-27, 1985.

[91] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Res,* vol. 40, pp. 1489-506, 2000.

[92] U. Neisser, *Cognitive Psychology*. New York: Appleton-Century-Crofts, 1967.

[93] S. C. J. Tsotsos, W. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling Visual Attention via Selective Tuning," *Artificial Intelligence,* vol. 78, pp. 507-545, 1995.

[94] C. K. Laurent Itti, and Ernst 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.

[95] F. A. W. Wolf Kienzle, Bernhard Scholkopf, and Matthias O. Franz, "A Nonparametric Approach to Bottom-Up Visual Saliency," *NIPS2006,* 2006.

[96] O. Le Meur and J. C. Chevet, "Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks," *IEEE Trans Image Process,* vol. 19, pp. 2801-13, Nov 2010.

[97] C. J. W. Kim, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 21, pp. 446–456, Apr. 2011 2011.

[98] V. M. D. Gao, and N. Vasconcelos, "The discriminant centersurround hypothesis for bottom-up saliency," in *NIPS*, 2007, pp. 497-504.

[99] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J Vis,* vol. 9, pp. 15 1-27, 2009.

[100] Y. T. Y. Lin, B. Fang, Z. Shang, Y. Huang, and S. Wang, "A visual-attention model using earth Mover's distance based saliency measurement and nonlinear feature combination," *IEEE Transaction Pattern Analysis and Machine Intellegence,* vol. 35, pp. 314–328, Feb. 2013 2013.

[101] R. A. a. S. Susstrunk, "Saliency detection using maximum symmetric surround salient region detection," in *IEEE ICIP*, 2010, pp. 2653–2656.

[102] S. H. R. Achanta, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," presented at the CVPR, 2009.

[103] M. Z. A. a. B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," *IEEE Trans. Image Process,* vol. 17, pp. 633–644, May 2008 2008.

[104] W. Z. Zhi Liu, and Olivier Le Meur, "Saliency Tree: A Novel Saliency Detection Framework," *IEEE TRANSACTIONS ON IMAGE PROCESSING,* vol. 23, pp. 1937-1952, MAY 2014 2014.

[105] S. U. a. A. Shaashua, "Structural Saliency: The Detection of Globally Salient Structures Using a Locally Connected Network," 1988.

[106] S. U. Amnon Sha'ashu, "Structural Saliency: The Detection of Globally Salient Structures Using a Locally Connected Network," in *Proc. Int'l Conf. Computer Vision (ICCV '88)*, 1988, pp. 321-327.

[107] T. A. a. R. Basri, "Extracting Salient Curves from Images: An Analysis of the Saliency Network," *Int'l J. Computer Vision,* vol. 27, pp. 51-69, 1998.

[108] A. Berengolts and M. Lindenbaum, "On the distribution of saliency," *IEEE Trans Pattern Anal Mach Intell,* vol. 28, pp. 1973-90, Dec 2006.

[109] N. D. B. B. a. J. K. Tsotsos, "Saliency Based on Information Maximization," presented at the NIPS, 2005.

[110] T. K. a. M. Brady, "Saliency, scale and image description," *J. Comput. Vis.,* vol. 45, pp. 83-105, Nov. 2001 2001.

[111] Y. W. W. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *IEEE CVPR*, 2010, pp. 2368–2375.

[112] X. H. a. L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE CVPR*, 2007, pp. 1-8.

[113] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans Image Process,* vol. 19, pp. 185-98, Jan 2010.

[114] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans Pattern Anal Mach Intell,* vol. 28, pp. 802-17, May 2006.

[115] C. K. J. Harel, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[116] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Random walks on graphs for salient object detection in images," *IEEE Trans Image Process,* vol. 19, pp. 3232-42, Dec 2010.

[117] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): meaningful attention using stochastic image modeling," *IEEE Trans Pattern Anal Mach Intell,* vol. 32, pp. 693-708, Apr 2010.

[118] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," *IEEE Trans Pattern Anal Mach Intell,* vol. 33, pp. 353-67, Feb 2011.

[119] J. W. H. Jiang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *IEEE CVPR*, 2013, pp. 2083–2090.

[120] Q. M. J. W. W. Zhang, G. Wang, and H. Yin, "An adaptive computational model for salient object detection," *IEEE Trans. Multimedia,* vol. 12, pp. 300–316, Jun. 2010 2010.

[121] T. D. B. Alexe, and V. Ferrari, "What is an object," in *IEEE CVPR*, 2010, pp. 73–80.

[122] T. L. L. K. Y. Chang, H. T. Chen, and S. H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *ICCV*, 2011, pp. 914–921.

[123] J. W. H. Jiang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *BMVC*, 2011, pp. 1–12.

[124] K. K. W. Zou, Z. Liu, and J. Ronsin, "Segmentation driven lowrank matrix recovery for saliency detection," in *BMVC*, 2013, pp. 1–13.

[125] X. S. a. Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *IEEE CVPR.*, 2012, pp. 853–860.

[126] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin, "Feature Congestion: A Measure of Display Clutter," in *Proceedings of CHI 2005*, Portland, OR, USA, 2005.

[127] J. A. Stuart and H. M. Burian, "A study of separation difficulty. Its relationship to visual acuity in normal and amblyopic eyes," *Am J Ophthalmol,* vol. 53, pp. 471-7, Mar 1962.

[128] R. van den Berg, F. W. Cornelissen, and J. B. Roerdink, "A crowding model of visual clutter," *J Vis,* vol. 9, pp. 24 1-11, 2009.

[129] M. R. Beck, M. C. Lohrenz, and J. G. Trafton, "Measuring search efficiency in complex visual search tasks: global and local clutter," *J Exp Psychol Appl,* vol. 16, pp. 238-50, Sep 2010.

[130] H. F., Ed., *The Growth of Structural and Functional Complexity during Evolution* (The Evolution of Complexity. 1999, p.^pp. Pages.

[131] A. Oliva, M. L. Mack, M. Shrestha, and A. Peeper, "Identifying the Perceptual Dimensions of Visual Complexity of Scenes," in *Proc. 26th Annual Meeting of the Cognitive Science Society*, 2004.

[132] M. L. Mack and A. Oliva, "Computational Estima-tion of Visual Complexity," presented at the Presented at the 12th Annual Object, Perception, Attention, and Memory Conference, 2004.

[133] R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," *Vision Res,* vol. 39, pp. 3157-63, Sep 1999.

[134] M. C. Lohrenz, G. J. Layne, S. S. Edwards, M. L. Gendron, and J. T. Bradley, "Feature Clustering to Measure Clutter in Electronic Displays," NAVAL RESEARCH LAB STENNIS SPACE CENTER MS2006.

[135] M. Lohrenz, M. R. Beck, J. G. Trafton, and M. Gendron, " Measurement and analysis of clutter in electronic displays," Naval Research Laboratory, Washington, DC.2009.

[136] M. C. Lohrenz, M. Gendron, and M. R. Beck, "Clearing up the clutter," *The Defense Management Journal,* pp. 167-169, 2008.

[137] G. Wolford and L. Chambers, "Lateral masking as a function of spacing," *Percept Psychophys,* vol. 33, pp. 129-38, Feb 1983.

[138] D. M. Levi, S. Hariharan, and S. A. Klein, "Suppressive and facilitatory spatial interactions in peripheral vision: peripheral crowding is neither size invariant nor simple contrast masking," *J Vis,* vol. 2, pp. 167-77, 2002.

[139] D. G. Pelli, M. Palomares, and N. J. Majaj, "Crowding is unlike ordinary masking: distinguishing feature integration from detection," *J Vis,* vol. 4, pp. 1136-69, Dec 30 2004.

[140] M. V. Danilova and V. M. Bondarko, "Foveal contour interactions and crowding effects at the resolution limit of the visual system," *J Vis,* vol. 7, pp. 25 1-18, 2007.

[141] D. M. Levi, S. A. Klein, and S. Hariharan, "Suppressive and facilitatory spatial interactions in foveal vision: foveal crowding is simple contrast masking," *J Vis,* vol. 2, pp. 140-66, 2002.

[142] S. Hariharan, D. M. Levi, and S. A. Klein, ""Crowding" in normal and amblyopic vision assessed with Gaussian and Gabor C's," *Vision Res,* vol. 45, pp. 617-33, Mar 2005.

[143] U. Polat and D. Sagi, "Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments," *Vision Res,* vol. 33, pp. 993-9, May 1993.

[144] D. M. Levi and S. A. Klein, "Vernier acuity, crowding and amblyopia," *Vision Res,* vol. 25, pp. 979-91, 1985.

[145] C. Feng, Y. Jiang, and S. He, "Horizontal and vertical asymmetry in visual spatial crowding effects," *J Vis,* vol. 7, pp. 13 1-10, 2007.

[146] T. Liu, Y. Jiang, X. Sun, and S. He, "Reduction of the crowding effect in spatially adjacent but cortically remote visual stimuli," *Curr Biol,* vol. 19, pp. 127-32, Jan 27 2009.

[147] J. Intriligator and P. Cavanagh, "The spatial resolution of visual attention," *Cogn Psychol,* vol. 43, pp. 171-216, Nov 2001.

[148] K. M. B. WILLIAM P. BANKS, and DOUGLAS W. LARSON, "The asymmetry of lateral interference in visual letter identification," *Perception & Psychophysics,* vol. 22, pp. 232-240, 1977.

[149] Y. Petrov, A. V. Popple, and S. P. McKee, "Crowding and surround suppression: not to be confused," *J Vis,* vol. 7, pp. 12 1-9, 2007.

[150] B. C. Motter and D. A. Simoni, "The roles of cortical image separation and size in active visual search performance," *J Vis,* vol. 7, pp. 6 1-15, 2007.

[151] F. L. Kooi, A. Toet, S. P. Tripathy, and D. M. Levi, "The effect of similarity and duration on spatial interaction in peripheral vision," *Spat Vis,* vol. 8, pp. 255-79, 1994.

[152] T. A. Nazir, "Effects of lateral masking and spatial precueing on gap-resolution in central and peripheral vision," *Vision Res,* vol. 32, pp. 771-7, Apr 1992.

[153] D. M. Levi, S. Hariharan, and S. A. Klein, "Suppressive and facilitatory spatial interactions in amblyopic vision," *Vision Res,* vol. 42, pp. 1379-94, May 2002.

[154] R. Chakravarthi and P. Cavanagh, "Temporal properties of the polarity advantage effect in crowding," *J Vis,* vol. 7, pp. 11 1-13, 2007.

[155] **S. T. Chung, D. M. Levi, and G. E. Legge, "Spatial-frequency and contrast properties of crowding," *Vision Res,* vol. 41, pp. 1833-50, Jun 2001.**

[156] **C. Gheri, M. J. Morgan, and J. A. Solomon, "The relationship between search efficiency and crowding," *Perception,* vol. 36, pp. 1779-87, 2007.**

[157] **G. J. Kennedy and D. Whitaker, "The chromatic selectivity of visual crowding," *J Vis,* vol. 10, p. 15, 2010.**

[158] **E. Poder and J. Wagemans, "Crowding with conjunctions of simple features," *J Vis,* vol. 7, pp. 23 1-12, 2007.**

[159] **E. M. e. a. Hubbard, "Individual differences among graphemecolor synesthetes: brain-behavior correlations," *Neuron,* pp. 975-985, 2005.**

[160] **T. Banton and D. M. Levi, "Spatial localization of motion-defined and luminance-defined contours," *Vision Res,* vol. 33, pp. 2225-37, Nov 1993.**

[161] **S. T. Chung, R. W. Li, and D. M. Levi, "Crowding between first- and second-order letter stimuli in normal foveal and peripheral vision," *J Vis,* vol. 7, pp. 10 1-13, 2007.**

[162] **D. M. Levi and T. Carney, "Crowding in peripheral vision: why bigger is better," *Current Biology,* vol. 19, pp. 1988-93, Dec 15 2009.**

[163] **A. Huckauf and D. Heller, "On the relations between crowding and visual masking," *Percept Psychophys,* vol. 66, pp. 584-95, May 2004.**

[164] **J. Ng and G. Westheimer, "Time course of masking in spatial resolution tasks," *Optom Vis Sci,* vol. 79, pp. 98-102, Feb 2002.**

[165] **R. F. Hess, S. C. Dakin, N. Kapoor, and M. Tewfik, "Contour interaction in fovea and periphery," *J Opt Soc Am A Opt Image Sci Vis,* vol. 17, pp. 1516-24, Sep 2000.**

[166] **R. F. Hess, S. C. Dakin, and N. Kapoor, "The foveal 'crowding' effect: physics or physiology?," *Vision Res,* vol. 40, pp. 365-70, 2000.**

[167] **L. Liu and A. Arditi, "Apparent string shortening concomitant with letter crowding," *Vision Res,* vol. 40, pp. 1059-67, 2000.**

[168] **L. Liu, "Can the amplitude difference spectrum peak frequency explain the foveal crowding effect?," *Vision Res,* vol. 41, pp. 3693-704, Dec 2001.**

[169] **S. T. Chung and B. S. Tjan, "Shift in spatial scale in identifying crowded letters," *Vision Res,* vol. 47, pp. 437-51, Feb 2007.**

[170] **F. M. Felisbert, J. A. Solomon, and M. J. Morgan, "The role of target salience in crowding," *Perception,* vol. 34, pp. 823-33, 2005.**

[171] **I. Mareschal, M. J. Morgan, and J. A. Solomon, "Cortical distance determines whether flankers cause crowding or the tilt illusion," *J Vis,* vol. 10, p. 13, 2010.**

[172] **B. L. Beard, D. M. Levi, and S. A. Klein, "Vernier acuity with non-simultaneous targets: the cortical magnification factor estimated by psychophysics," *Vision Res,* vol. 37, pp. 325-46, Feb 1997.**

[173] **A. O. H. Patrick Cavanagh "Non-retinotopic crowding," *Vision* vol. 7, pp. 338-, 2007.**

[174] **E. Poder, "Effect of colour pop-out on the recognition of letters in crowding conditions," *Psychol Res,* vol. 71, pp. 641-5, Nov 2007.**

[175] A. Treisman and H. Schmidt, "Illusory conjunctions in the perception of objects," *Cogn Psychol,* vol. 14, pp. 107-41, Jan 1982.

[176] J. M. Wolfe and K. R. Cave, "The psychophysical evidence for a binding problem in human vision," *Neuron,* vol. 24, pp. 11-7, 111-25, Sep 1999.

[177] D. J. Field, A. Hayes, and R. F. Hess, "Contour integration by the human visual system: evidence for a local "association field"," *Vision Res,* vol. 33, pp. 173-93, Jan 1993.

[178] A. V. Popple and D. M. Levi, "The perception of spatial order at a glance," *Vision Res,* vol. 45, pp. 1085-90, Apr 2005.

[179] W. P. Banks, D. W. Larson, and W. Prinzmetal, "Asymmetry of visual interference," *Percept Psychophys,* vol. 25, pp. 447-56, Jun 1979.

[180] T. Livne and D. Sagi, "Configuration influence on crowding," *J Vis,* vol. 7, pp. 4 1-12, 2007.

[181] R. van den Berg, J. B. Roerdink, and F. W. Cornelissen, "A neurophysiologically plausible population code model for feature integration explains visual crowding," *PLoS Comput Biol,* vol. 6, p. e1000646, Jan 2010.

[182] P. Dayan and J. A. Solomon, "Selective Bayes: attentional load and crowding," *Vision Res,* vol. 50, pp. 2248-60, Oct 28 2010.

[183] J. Duncan, "Selective attention and the organization of visual information," *J Exp Psychol Gen,* vol. 113, pp. 501-17, Dec 1984.

[184] P. R. Roelfsema, V. A. Lamme, and H. Spekreijse, "Object-based attention in the primary visual cortex of the macaque monkey," *Nature,* vol. 395, pp. 376-81, Sep 24 1998.

[185] R. A. Rensink and J. T. Enns, "Preemption effects in visual search: evidence for low-level grouping," *Psychol Rev,* vol. 102, pp. 101-30, Jan 1995.

[186] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vision Res,* vol. 40, pp. 1469-87, 2000.

[187] J. O. R. Rensink, and J. Clark, "To See or not to See: The Need for Attention to Perceive Changes in Scenes," *Psychological Science,* vol. 8, pp. 368–373, 1997.

[188] P. J. Bex, "(In) sensitivity to spatial distortion in natural scenes," *J Vis,* vol. 10, pp. 23 1-15, 2010.

[189] P. J. Bex, S. C. Dakin, and A. J. Simmers, "The shape and size of crowding for moving targets," *Vision Res,* vol. 43, pp. 2895-904, Dec 2003.

[190] T. Livne and D. Sagi, "How do flankers' relations affect crowding?," *J Vis,* vol. 10, pp. 1 1-14, 2010.

[191] T. P. Saarela, B. Sayim, G. Westheimer, and M. H. Herzog, "Global stimulus configuration modulates crowding," *J Vis,* vol. 9, pp. 5 1-11, 2009.

[192] J. M. Wallace and B. S. Tjan, "Object crowding," *J Vis,* vol. 11, 2011.

[193] J. Fischer and D. Whitney, "Object-level visual information gets through the bottleneck of crowding," *J Neurophysiol,* vol. 106, pp. 1389-98, Sep 2011.

[194] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," ***International Journal of Computer Vision,*** **vol. 40, pp. 49-70, 2000.**

[195] R. M. Balboa and N. M. Grzywacz, "Power spectra and distribution of contrasts of natural images from different habitats," ***Vision Res,*** **vol. 43, pp. 2527-37, Nov 2003.**

[196] D. G. Pelli and P. Bex, "Measuring contrast sensitivity," ***Vision Res,*** **vol. 90, pp. 10-4, Sep 20 2013.**

[197] P. J. Bex, S. G. Solomon, and S. C. Dakin, "Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure," ***J Vis,*** **vol. 9, pp. 1 1-19, 2009.**

[198] P. Kovesi, "Phase congruency detects corners and edges," presented at the In The Australian Pattern Recognition Society Conference, Sydney, Australia, 2003.

[199] S. G. Paul Green, Kristine Zeltner, and Susan Adams, "Leigibility of Text On Instrument Panels: A Literature Review," The University of Michigan Transportation Research Institute, Ann Arbor, Mighigan1988.

[200] W. L. Paul Green, Gretchen Paelke and Colleen Serafin, "Suggested Human Factors Design Guidelines for Driver Information Systems," 1993.

[201] Bartram, "The perception of semantic quality in type: Differences between designers and non-designers," ***Information Design Journal,*** **vol. 3, pp. 30-37, 1982.**

[202] Z. B, ***Studies in Legibility of Printed Text.*** **Stockholm: Almqvist & Wiksell, 1965.**

[203] H. J., ***Designing Instructional Text, Kogan Page.*** **London Nichols, 1978.**

[204] S. P., ***Type and Typography.*** **Copenhagen, Denmark: Danish Design Center, 1995.**

[205] E. Johnston, ***Writing & illuminating, & lettering*** **London: Pitman., 1945 21st reprint.**

[206] A. E. Hillis and A. Caramazza, "Deficit to stimulus-centered, letter shape representations in a case of "unilateral neglect"," ***Neuropsychologia,*** **vol. 29, pp. 1223-40, 1991.**

[207] E. Yee, S. Huffstetler, and S. L. Thompson-Schill, "Function follows form: activation of shape and function features during object identification," ***J Exp Psychol Gen,*** **vol. 140, pp. 348-63, Aug 2011.**

[208] Z. Kaldy and E. Blaser, "How to Compare Apples and Oranges: Infants' Object Identification Tested With Equally Salient Shape, Luminance and Color Changes," ***Infancy,*** **vol. 14, pp. 222-243, Mar 2009.**

[209] T. F. C. Alice F. Healy, "A developmental evaluation of the role of word shape in word recognition," ***Memory & Cognition,*** **vol. 20, pp. 141-150, 1992.**

[210] T. Tougakiuchi, "[Role of whole letter shape cues in Chinese character identification]," ***Shinrigaku Kenkyu,*** **vol. 69, pp. 33-8, Apr 1998.**

[211] S. Singh, "Optical Character Recognition Techniques: A Survey," ***Emerging Trends in Computing and Information Sciences,*** **vol. 4, pp. 545-550, 2013.**

[212] K. B. Amarjot Singh, and Akshay Bhasin, "A Survey of OCR Applications," *International Journal of Machine Learning and Computing,,* vol. 2, 2012.

[213] S. E. Palmer, "Fundamental aspects of cognitive representation," in *Cognition and Categorization*, E. R. a. B. B. Lloyd, Ed., ed Hillsdale NJ: Lawrence Erlbaum., 1978, pp. 259-303.

[214] J. E. Hummel, "The Complementary Properties of Holistic and Analytic Representations of Shape," in *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes* M. A. P. a. G. Rhodes, Ed., ed Oxford, New York: Oxford University Press, 2003, pp. 1-17.

[215] S. Edelman, "Representation is representation of similarities," *Behavioral & Brain Sciences,* pp. 449-498, 1998.

[216] M. Riesenhuber, & Poggio, T. , "Hierarchical models of object recognition in cortex," *Nature Neuroscience,* pp. 1019-1025, 1999.

[217] D. Marr, & Nishihara, H. K., "Representation and recognition of three dimensional shapes, Series B. 200, ," 1978, pp. 269-294.

[218] J. E. Hummel, & Biederman, I., "Dynamic binding in a neural network for shape recognition," *Psychological Review,* pp. 480-517, 1992.

[219] J. E. Hummel, & Stankiewicz, B. J.. "Two roles for attention in shape perception: A structural description model of visual scrutiny," *Visual Cognition,* pp. 49-79, 1998.

[220] P. M. a. M. Garvey, D. M. , "Changeable Message Sign Visibility," Washington, D.C. 1996

[221] R. a. L. Mourant, G. , "Luminance Specifications for Automobile Instrument Panels," *Human Factors,* vol. 18, pp. 71-84, 1976.

[222] P. M. Garvey, Zineddin, A. Z., and Pietrucha, M. T., " Letter Legibility for Signs and Other Large Format Applications," in *Human Factors and Ergonomics Society 45th Annual Meeting*, Minneapolis, Minnesota, 2001, pp. 1443-1447.

[223] G. A. a. A. Peters, B.B., "These 3 Criteria for Readable Panel Markings," vol. 30, pp. 55-57, 1959.

[224] S. Smith, "Letter Size and Legibility," *Human Factors,* vol. 21, pp. 661-670, 1979.

[225] J. a. K. Ducan, S., "Legibility of LED and Liquid Crystal Displays," in *Proceedings of the Society for Information Display*, Farnborough, Hants,UK, 1976, pp. 180-186.

[226] D. o. Defense, " MIL-STD-1472F," ed. Washington, DC: Government Printing Office, 1999, pp. 33,101.

[227] G. L. Howett, "Size of Letters Required for Visibility as a Function of Viewing Distance and Observer Visual Acuity," National Bureau of Standards, Washington, D.C1983.

[228] H. H., "NAVSHIPS Display Illumination Design Guide vol.2: Human Factors; ," Naval Electronics Laboratory Center, San Diego, California1973.

[229] D. F. Wourms, Cunningham, P. H., Self, D. A., and Johnson, S. J., "Bus Signage Guidelines for Persons with Visual Impairments: Electronic Signs," Federal Transit Administration, Washington D.C.2001.

[230] J. Laycock, "The Legibility of Passive Displays," in *Proceedings of the Society for Informaiton Display*, Farnborough, Hants, UK, 1985, pp. 89-93.

[231] H. L. a. M. Snyder, M.E., "Information Transfer from Computer-Generated Dot-Matrix Displays," Research Trigangle Park, NC: U.S. Army Research Office1978.

[232] C. L. Dudek, "Guidelines on the Use of Changeable Message Signs. Federal Highway Administration, ," Federal Highway Administration, Washington, D.C.1991.

[233] J. E. Sheedy, M. V. Subbaram, A. B. Zimmerman, and J. R. Hayes, "Text legibility and the letter superiority effect," *Hum Factors,* vol. 47, pp. 797-815, Winter 2005.

[234] P. R. Hind, Tritt, B.H., and Hoffmann, E.R., "Effects of Level of Illumination, Strokewidth, Visual Angle and Contrast on the Legibility of Numerals of Various Fonts.," in *Proceedings of the Australian Road Research Board, Eighth Conference*, Parkville, Victoria, Australia, 1976, pp. 46-55.

[235] R. L. Spencer H., and Coe B. , "Spatial and typographic coding with bibliographical entries," *Programmed Learning and Educational Technology* vol. 12, pp. 95-101, 1975.

[236] J. E. Sheedy, Subbaram, M. V., Zimmerman, A.B., and Hayes, J. R., "Text legibility and the letter superiority effect," *Human Factors,* pp. 797–815, 2005.

[237] V. L. P. Francis T. Durso, John S. Burnett, and Eric J. Stearman, "Evidence-Based Human Factors Guidelines for PowerPoint Presentations," *Ergonomics in Design,* 2011.

[238] P. L. a. B. Olson, A, "The Nighttime Legibility of Highway Signs as a Function of Their Luminance Characteristics," *Human Factors,* vol. 21, pp. 145-160, 1979.

[239] P. L. Olson, Sivak, M., and Egan, J.C., "Variables Influencing the Nighttime Legibility of Highway Signs," The University of Michigan Transportation Reseach Institute, Ann Arbor, Michigan1983.

[240] M. a. H. Colomb, R. , "Legibility and Contrast Requirements of Variable-Message Signs," *Transportation Research Record: Journal of the Transportation Research Board,* pp. 137-141, 1991.

[241] M. a. O. Sivak, P.L., "Optimal and Minimal Luminacne Characteristics for Retroreflective Highway Signs," *Transportation Research Record: Journal of the Transportation Research Board,* pp. 53-57, 1985.

[242] S. Rogers, Spiker, V., and Cicinelli, J., "Luminance and Luminace Contrast Requirements for Legibility of Self-Luminous Displays in Aircraft Cockpits," *Applied Ergonomics,* vol. 17, pp. 271-277, 1986.

[243] M. E., "Typeface emphasis and information focus in written language," *Applied Cognitive Psychology,* vol. 6, pp. 345-359, 1992.

[244] P. M. Garvey, Zineddin, A.Z. ,and Pietrucha, M.T. , "Development of a New Guide Sign Alphabet.," *Ergonomics in Design,* vol. 6, pp. 7-11, 1998.

[245] B. K. Jeffrey D. Miles, Sarah Hammond and Fan Ye, "Evaluation of guide sign fonts," T. A. M. T. Institute, Ed., ed. College Station, Texas 77843-3135, 2014.

[246] G. K. Jean-Baptiste Bernard, Jasmine Junge , Susana T.L. Chung "The effect of letter-stroke boldness on reading speed in central and peripheral vision," *Vision Research,* vol. 84, pp. 33-42, 24 May 2013 2013.

[247] M. A. Tinker, "The influence of form of type on the perception of words," *Journal of Applied Psychology,* vol. 16, pp. 167-174, 1932.

[248] M. A. Tinker, & Patterson, D. G., " Influence of type form on speed of reading," *Journal of Applied Psychology,* vol. 13, pp. 205-219, 1929.

[249] J. M. Cattell, "The time taken up by cerebral operations," *Mind,* vol. 11, pp. 524–538, 1886.

[250] A. M. Jacobs and J. Grainger, "Automatic letter priming in an alphabetic decision task," *Percept Psychophys,* vol. 49, pp. 43-52, Jan 1991.

[251] J. C. Aries Arditi, "Letter case and text legibility in normal and low vision," *Vision Research 47* pp. 2499–2505, 2007.

[252] D. Besner, Davelaar, E., Alcott, D., & Parry, P., "Wholistic reading of alphabetic print: Evidence from the FDM and the FBI," in *Orthographies and reading: Perspectives from cognitive psychology, neuropsychology and linguistics*, L. Henderson, Ed., ed Hillsdale, NJ: Erlbaum, 1984, pp. 121-135.

[253] P. H. Seymour and M. V. Jack, "Effects of visual familiarity on "same" and "different" decision processes," *Q J Exp Psychol,* vol. 30, pp. 455-69, Aug 1978.

[254] D. Fiset, C. Blais, C. Ethier-Majcher, M. Arguin, D. Bub, and F. Gosselin, "Features for identification of uppercase and lowercase letters," *Psychol Sci,* vol. 19, pp. 1161-8, Nov 2008.

[255] D. Fiset, C. Blais, M. Arguin, K. Tadros, C. Ethier-Majcher, D. Bub, and F. Gosselin, "The spatio-temporal dynamics of visual letter recognition," *Cogn Neuropsychol,* vol. 26, pp. 23-35, Feb 2009.

[256] D. E. Rumelhart and P. Siple, "Process of recognizing tachistoscopically presented words," *Psychol Rev,* vol. 81, pp. 99-118, Mar 1974.

[257] J. Grainger, "Cracking the orthographic code: An introduction," *Language and Cognitive Processes,* pp. 1-35, 2008.

[258] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception: Part 1. An account of basic findings," *Psychol Rev,* vol. 88, pp. 375-407, 1981.

[259] P. Baines and A. Haslam, *Type & typography (2nd ed.).* London: Laurence King., 2005.

[260] D. G. Pelli, C. W. Burns, B. Farell, and D. C. Moore-Page, "Feature detection and letter identification," *Vision Res,* vol. 46, pp. 4646-74, Dec 2006.

[261] G. E. Legge and C. A. Bigelow, "Does print size matter for reading? A review of findings from vision science and typography," *J Vis,* vol. 11, 2011.

[262] K. Cha, K. W. Horch, R. A. Normann, and D. K. Boman, "Reading speed with a pixelized vision system," *Journal of the Optical Society of America a-Optics Image Science and Vision,* vol. 9, pp. 673-677, May 1992.

[263] A. P. Fornos, J. Sommerhalder, B. Rappaz, A. B. Safran, and M. Pelizzone, "Simulation of artificial vision, III: do the spatial or temporal characteristics of stimulus pixelization really matter?," *Invest Ophthalmol Vis Sci,* vol. 46, pp. 3906-12, Oct 2005.

[264] L. Fu, S. Cai, H. Zhang, G. Hu, and X. Zhang, "Psychophysics of reading with a limited number of pixels: towards the rehabilitation of reading ability with visual prosthesis," *Vision Res,* vol. 46, pp. 1292-301, Apr 2006.

[265] J. Sommerhalder, E. Oueghlani, M. Bagnoud, U. Leonards, A. B. Safran, and M. Pelizzone, "Simulation of artificial vision: I. Eccentric reading of isolated words, and perceptual learning," *Vision Res,* vol. 43, pp. 269-83, Feb 2003.

[266] J. J. Terry L. Childers, "All Dressed Up With Something to Say: Effects of Typeface Semantic Associations on Brand Perceptions and Consumer Memory," *Journal of Consumer Psychology,* vol. 12, pp. 93-106, 2002.

[267] S. T. Mueller and C. T. Weidemann, "Alphabetic letter identification: effects of perceivability, similarity, and bias," *Acta Psychol (Amst),* vol. 139, pp. 19-37, Jan 2012.

[268] M. F. a. M. Coltheart, "Introduction Letter recognition: From perception to representation," *Cognitive Neuropsychology,* vol. 26, pp. 1-6, 2009.

[269] M. Zorzi, C. Barbiero, A. Facoetti, I. Lonciari, M. Carrozzi, M. Montico, L. Bravar, F. George, C. Pech-Georgel, and J. C. Ziegler, "Extra-large letter spacing improves reading in dyslexia," *Proc Natl Acad Sci U S A,* vol. 109, pp. 11455-9, Jul 10 2012.

[270] G. Bonsiepe, "A method of quantifying order in typographic design," *Journal of Typographic Research,* pp. 203-220, 1968.

[271] T. S. Tullis, " An evaluation of alphanumeric, graphic, and color information displays," *Human Factors,* pp. 541-550, 1981.

[272] S. P., *Type and Typography*. Copenhagen, Denmark: Danish Design Center, 1995.

[273] S.-n. Y. Yu-Chi Tai, and John Hayes, James Sheedy, "Effect of Character Spacing on Text Legibility," Vision PerformanceInstitute, Pacific University, Oregon, USA.

[274] S. N. Yang, Y. C. Tai, H. Laukkanen, and J. Sheedy, "Effects of ocular transverse chromatic aberration on near foveal letter recognition," *Vision Res,* vol. 49, pp. 2881-90, Nov 2009.

[275] M. Perea, C. Moret-Tatay, and P. Gomez, "The effects of interletter spacing in visual-word recognition," *Acta Psychol (Amst),* vol. 137, pp. 345-51, Jul 2011.

[276] F. Vinckier, E. Qiao, C. Pallier, S. Dehaene, and L. Cohen, "The impact of letter spacing on reading: a test of the bigram coding hypothesis," *J Vis,* vol. 11, pp. 1-21, 2011.

[277] S. J. Crutch and E. K. Warrington, "The relationship between visual crowding and letter confusability: towards an understanding of dyslexia in posterior cortical atrophy," *Cogn Neuropsychol,* vol. 26, pp. 471-98, 2009.

[278] J. P. Van Overschelde and A. F. Healy, "A blank look in reading: the effect of blank space on the identification of letters and words during reading," *Exp Psychol,* vol. 52, pp. 213-23, 2005.

[279] M. A. Mary-Jane Carroll. (2014, Text Legibility and Readability of Large Format Signs in Building and Sites. Available: http://www.udeworld.com/documents/designresources/pdfs/TextLegibilityandReadabilityofLargeFormatSignsinBuildingandSites.pdf

[280] T. J. Slattery and K. Rayner, "Effects of intraword and interword spacing on eye movements during reading: exploring the optimal use of space in a line of text," *Atten Percept Psychophys,* vol. 75, pp. 1275-92, Aug 2013.

[281] C. Zang, F. Liang, X. Bai, G. Yan, and S. P. Liversedge, "Interword spacing and landing position effects during Chinese reading in children and adults," *J Exp Psychol Hum Percept Perform,* vol. 39, pp. 720-34, Jun 2013.

[282] S. Blackmore-Wright, M. A. Georgeson, and S. J. Anderson, "Enhanced text spacing improves reading performance in individuals with macular disease," *PLoS One,* vol. 8, p. e80325, 2013.

[283] J. H. Coll, J. Fjermestad, and R. Coll, "An eight experiment sequence to determine reading equality," *Information & Management* pp. 231-242, 1998.

[284] A. Chaparro and C. Liao, "The effect of text orientation, visual meridian, and inter-character spacing on word identification in the retinal periphery," *Perception,* vol. 32, pp. 1339-50, 2003.

[285] M. D., "The graphic design of text," *Intercom,* February,1996 1996.

[286] E. Ruder, *Typographie: A Manual of Design* Verlag Niggli AG, March 1, 2001.

[287] B. T. Kuhn, P. M. Garvey, M. T. Pietrucha, A. R. P. o. the, and U. S. S. Council., "Sign Legibility: Impact of Color and Illumination on Typical On-Premise Sign Font Legibility.," A Research Project of the United States Sign Council1998.

[288] A. H. Chan, S. N. Tsang, and A. W. Ng, "Effects of line length, line spacing, and line number on proofreading performance and scrolling of Chinese text," *Hum Factors,* vol. 56, pp. 521-34, May 2014.

[289] S. T. Chung, "Reading speed benefits from increased vertical word spacing in normal peripheral vision," *Optom Vis Sci,* vol. 81, pp. 525-35, Jul 2004.

[290] J. Ling and P. v. Schaik, "The influence of line spacing and text alignment on visual search of web pages," *Displays,* vol. 28, pp. 60-67, 2007.

[291] C. Enroth-Cugell and J. G. Robson, "Functional characteristics and diversity of cat retinal ganglion cells. Basic characteristics and quantitative description," *Invest Ophthalmol Vis Sci,* vol. 25, pp. 250-67, Mar 1984.

[292] R. E. Soodak, "Two-dimensional modeling of visual receptive fields using Gaussian subunits," *Proc Natl Acad Sci U S A,* vol. 83, pp. 9259-63, Dec 1986.

[293] G. T. Einevoll and P. Heggelund, "Mathematical models for the spatial receptive-field organization of nonlagged X-cells in dorsal lateral geniculate nucleus of cat," *Vis Neurosci,* vol. 17, pp. 871-85, Nov-Dec 2000.

[294] Z. x. wen, *Constitution theory.* Beijing,China: China Science and Technology Press, 2003.

[295] R.Jayadevan, S. R. Kolhe, P.M.Patil, and U.Pal, "Automatic processing of handwritten bank cheque images: a survey," *IJDAR,* vol. 15, pp. 267-296, 2012.

[296] J. Edwards, Y. W. Teh, D. Forsyth, R. B. M. Maire, and G. Vesom, "Making latine manuscripts searchable using ghmm's," presented at the NIPS, 2004.

[297] T. v. d. Zant, L. Schomaker, and K. Haak, "Handwritten-word spotting using biologically inspired features," *IEEE TRANS. PAMI,* vol. 30, pp. 1945-1955, 2008.

[298] J. Chan, C. Ziftci, and D. Forsyth, "Searching Off-line Arabic Documents " presented at the CVPR, 2006.

[299] Rath, T. M., and R. Manmatha, "Word spotting for historical documents," *International Journal on Document Analysis and Recognition,* vol. 9, pp. 139-152, 2007.

[300] T. Adamek, N. E. Connor, and A. F. Smeaton, " Word matching using single closed contours for indexing handwritten historical documents," *Int. Journal on Document Analysis and Recognition,* vol. 9, pp. 153-165, 2007.

[301] R. A.-H. Mohamad, L. Likforman-Sulem, and C. Mokbel, "Combining slanted-frame classifiers for improved HMM-based arabic handwriting recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 31, pp. 1165–1177, 2009.

[302] Y. Chherawala and M. Cheriet, "W-TSV: Weighted topological signature vector for lexicon reduction in handwritten Arabic documents," *Pattern Recognition,* vol. 45, pp. 3277-3287, 2012.

[303] R. Palacios and A. Gupta, "A system for processing handwritten bank checks automatically," *Image and vision computing,* vol. 26, pp. 1297-1313, 2008.

[304] V. K. Madasu and B. C. Lovell, "Automatic Segmentation and Recognition of Bank Cheque Fields," 2005 2005.

[305] V. Madasu, Mohd, H. M. Yusof, M. Hanmandlu, and K. Kubik, "Automatic Extraction of Signatures from Bank Cheques and Other Documents," presented at the DICTA'03, 2003.

[306] G. Kaufmann and H. Bunke, "Automated reading of cheque amounts," *Pattern Analysis & Applications* pp. 132-141, 2000.

[307] K. K. Kim, J. H. Kim, Y. K. Chung, and C. Suen, "Legal amount recognition based on the segmentation hypotheses for bank check processing," in *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR '01)*, 2001, pp. 964–967.

[308] A. K. Talele and S. L. Nalbalwar, "Automatic Extraction of Legal and Courtesy amount, Payee Name and signature in Bank Cheque Processing

System " *International Journal of Engineering Science and Technology (IJEST),* vol. 3 pp. 4417-4425, 2011.

[309] M. Mehta, R. Sanchati, and A. Marchya, "Automatic Cheque Processing System," *International Journal of Computer and Electrical Engineering,* vol. 2, pp. 761-765, 2010.

[310] S. N. Srihari, "Recognition of handwritten and machine-printed text for postal address interpretation," *Pattern Recognition Letters,* pp. 291–302, 1993.

[311] C.-L. Liu, M. Koga, and H. Fujisawa, " Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1425–1437, 2002.

[312] H. N. Prakash and D. S. Guru, "Geometric Centroids and their Relative Distances for Off-line Signature Verfication," in *Proc. ICDAR*, 2009, pp. 121-125.

[313] M. A. Ferrer, J. B. Alonso, and C. M. Traviso, "Offline Geometric Parameters For Automatic Signature Verification Using Fixed-point Arithmetic," *IEEE Trans. PAMI,* vol. 27, pp. 993-997, 2005.

[314] R. G. Cinbis and J. V. a. C. Schmi, "Image categorization using Fisher kernels of non-iid image models," presented at the CVPR, United States, 2012.