Universitat
Autònoma
de Barcelona

Master in Computer Vision and Artificial Intelligence

Report of the Master Project

Option: Computer Vision

Perceptual Organization for Text Extraction in Natural Scenes

Author: **Lluís Gómez i Bigordà**
Advisor: **Dimosthenis Karatzas**

Acknowledgements

First of all I would like to thank my advisor, Dimosthenis Karatzas, for the countless ideas, conversations, and reviews that inspired the work done in this Thesis, but most of all I thank you for your confidence.

I have great thanks for the big family at the Computer Vision Center: doctorates, students, technical and administration staff, and especially to the members of my research group, the Document Analysis Group. You all make research activity more motivating, exciting and comfortable.

The same applies also for all the teaching and coordination staff of the Master in Computer Vision and Artificial Intelligence, from whom I have learned so much over the last year.

Finally, I would like to give a warm thank you to my family and friends, specially to Carme, Dolors, Silvia, and Antònia, for your energy, understanding, and support.

ABSTRACT

The automated understanding of textual information in natural scenes is an important problem to solve for the Computer Vision and Document Analysis community. In this Thesis we approach the problem of text detection and extraction from an anthropocentric point of view, arguing that the Gestalt grouping laws, as a primary process in the human vision system, is something inherent in the complex cognitive task of reading. Our research hypothesis is therefore that placing Perceptual Organization at the heart of the task of Text Extraction in Natural Scenes will help increasing the performance obtained by current state of the art methods.

We present a new method for text extraction in natural scenes inspired by the Perceptual Organization Theory. The method combines two different clustering techniques in a single parameter-free clustering procedure. This Perceptual Organization Clustering is totally independent of the language or script in which text appears, can deal with any kind of font types and sizes, and is not constrained to horizontally aligned text. The obtained results in several experiments on the ICDAR2003 dataset give rise interesting conclusions in the context of perceptual grouping of text components and achieve sufficient success to warrant further development.

Keywords: *Perceptual Organization, Gestalt grouping laws, clustering, a contrario clustering, Helmholtz Principle, Evidence Accumulation Clustering, text localization, text extraction, segmentation, natural scenes.*

Contents

Chapter 1

Introduction

Reading is the complex cognitive process of transforming forms (letters and words) into meanings [1] in order to understand and interpret written content. Learning to read is a long road for humans, requiring explicit instruction, practice and refinement. Although the majority of the reading activity is carried out over written words on paper, such as books or magazines, texts may also appear written on objects around us: text on street signs, a motto painted on a wall, and even text produced by arranging small stones on the sand. Environment text is an important source of information in our daily life.

The automated understanding of textual information is the main goal of Document Image Analysis (DIA). We live in the days of mobile and wearable computing revolution where billions[1] of pervasive, personal, and portable devices, equipped with integrated built-in digital cameras, flood our streets and have become indispensable in our daily life. Providing those already ubiquitous imaging devices with reading capabilities is a highly desirable goal for the Computer Vision and Document Analysis community. Reading text in natural scenes will enable new and exciting applications such as automatic translation, way-finding and navigation aid, support tools for elderly and visually impaired persons, or contextual image indexing and search among many others. It is therefore not surprising that the field has gained increasing attention of the researchers in the last decades, nevertheless there is still no general-purpose solution that can work in any scenario.

Text found in natural scenes is usually present with the explicit intention to be easily read by humans, as it is at the end a human communication tool. This important characteristic makes for a clear difference between text detection and any other object detection task in Computer Vision. This is, text in natural scenes is usually designed in a way that exploits human perception laws, and thus human perception inspired Computer Vision techniques may have the key to the problem we want to solve. For example, wayfinding pannels normally present text characters in white colour over a high

[1] `http://www.strategyanalytics.com/default.aspx?mod=reportabstractviewer&a0=6216`

contrasted constant colour background, making it visible from large distances and easy to read. In the same direction, all writing systems at large make use of some sort of rules regarding the distance, similarity and arrangement between characters, mainly because it makes easier the task of reading. It is not by chance that the well known Gestalt theory, established back in 1921 by Max Wertheimer [2], proposes a set of grouping laws for how human observers organize objects together using the principles of proximity, similarity and good continuation. This evident connection between Gestalt laws and writing systems is going to be the basis for the research drawn in this thesis.

1.1 Goals and research hypothesis

The main goal of this thesis is to explore the link between the task of Text Extraction in Natural Scenes and the Perceptual Organization theoretical framework, in order to build a successful text extraction method on its basis. Our research hypothesis is that placing Perceptual Organization at the heart of the task of Text Extraction in Natural Scenes will help increasing the obtained performance at the current moment. This idea arise from the observations that Gestalt laws of perception play an important role in the core of writing systems [3] being one of the main processes involved in reading [4] [5], and thus have to be taken into account in order to solve the problem of Text Extraction in Natural Scenes in a perceptually inspired bottom-up approach.

1.2 Contributions

The contribution of the present work is mainly the novelty of the proposed method, making use of existing Gestalt mathematical formulations together with a clustering ensemble framework, leading to a combined methodology that stems from different disciplines in a beautiful and coherent theoretical way. In the process we intend to contribute some clue in removing the barriers that usually separate Gestalt theory from Computer Vision analytical perspective.

1.3 Outline

The next chapter defines in detail the problem of Text Extraction in Natural Scenes, starting with an introduction to Camera Based Document Analysis systems and following a description of technical challenges, the range of possibile applications, and the main methods that have been used to approach the problem. The chapter includes a state-of-the-art review as well as an account of some standard datasets available.

In the third Chapter the main theory of perceptual organization is presented, from the origins in Gestalt Theory to the different mathematical formalizations existing in the literature and focusing into the probabilistic approach proposed by Desolneux et al. [6] [7]. The chapter ends with a short introduction to clustering ensembles and the sort of solutions that Evidence Accumulation Clustering can provide for our goals.

The fourth Chapter describes our proposed method for Text Extraction in Natural Scenes integrating some of the theories explained in the former chapters, including validation assessments, qualitative and quantitative results on the ICDAR2003 dataset.

Finally, the last Chapter provides conclusions, a general discussion on the proposed method, and future lines of development on the light of the obtained results.

Chapter 2

Reading Text in Natural Scene Images

Document Image Analysis(DIA) is a mature field in the crossroads between Pattern Recognition and Computer Vision, with its roots in the first Optical Character Recognition (OCR) systems developed around the 60's, and is considered today one of the more fruitful fields of Computer Vision. Nowadays OCR is considered a solved problem when a clean binarized and well formatted input image, with text in a standard font and language is provided. For instance, available OCR solutions, both commercial and open source, perform extremely well in documents generated with modern printers and standard font types. However, there are still some open research issues on OCR dealing with the recognition of rare font types or noisy copies and degraded documents, where one has to deal with difficult text segmentation, non easily separable adjacent characters, or broken strokes. Furthermore, handwritten documents, colour-image document analysis, complex layouts and text segmentation in born-digital images are examples of harder problems still unsolved and really hot topics for the Document Image Analysis community.

Reading Text in Natural Scene Images is therefore no more than an evolution of the Document Image Analysis research arising from one side as a consequence of the normal trend on incorporating new text containers and on the other by the widespread availability of digital cameras.

2.1 Camera Based Document Analysis

Compared with the long tradition of DIA, Camera Based Document Analysis is a relatively young field of research, mainly because digital camera devices appeared later into the scene but also because the quality and resolution of such devices was initially not enough to compete with scanner technology available at the time, thus making scanners the only realistic source of imaging for the DIA community. This however has changed drastically in the recent years and nowadays one can find a wide range of both professional and consumer-grade digital cameras. Moreover, those pocket devices are really portable, easy to use, cheap, and multi-task oriented (usually integrated in multi-purpose computing devices). All this makes digital cameras not only a real alternative to traditional scanners but also creates an exciting breed of new DIA applications that can be in the hands of milions of users that up to now have no plans on aquiring a scanner.

Jung et al. [8] and Linag et al. [9] made an exhaustive survey on the new challenges in the field as well as pointing to new emerging applications setting the basis for the research on camera-based text analysis. Linag et al. [9] define a common architecture for a Text Information Extraction System, involving several differentiated steps: detection, localization, tracking, extraction, enhancement, and recognition. The border between detection and localization tasks on one side and between extraction and enhancement on the other is not sharp and depending on the approach one can use only one of the denominations for a single step. Despite the schema has demonstrated to be a good starting point to the problem, more recently some authors have proposed end-to-end systems [10] [11] where the detection and recognition tasks work together in a feedback loop that allows the correction of errors in previous steps of the system.

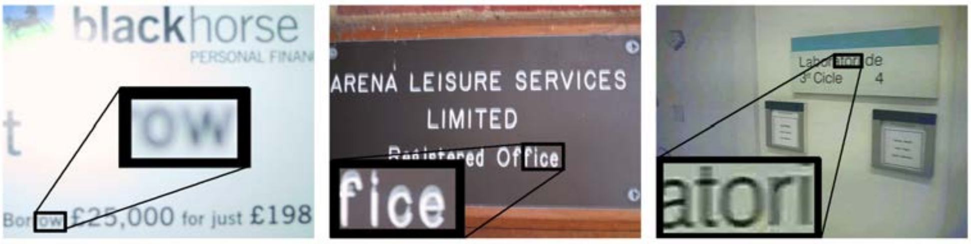
Robust reading tasks on Camera based Document Analysis(CBDA) are usually separated into caption text and scene text. The former is usually overlaid on images or video, and has a lot in common with other DIA fields like colour-image document analysis or reading born-digital text images. On the other hand, reading scene text, which exists naturally in the image, is definitely the most difficult task for CBDA.

2.1.1 Challenges

Most of the difficulty in Text Extraction from Natural Scenes comes from the fact that we have virtually no any restriction on how scene text can appear within an image. Thus, the first big challenge, is to deal in a robust way with high variability in font type, size, colour, orientation, and alignment. Other challenges to deal with stem from certain disadvantages inherent to the capturing technology of camera based systems, and of course from the intrinsic complexity of natural scenes including perspective distortions and cluttered scenarios.

In general, images with **low resolution** make difficult any task of image segmentation and at the same time are not well posed input for the final OCR step in the system, the same applies for **blur** and **uneven focus** that can appear in the image in situations where we are shooting **moving objects**, the **camera is not static**, or the **light is low**, if the sensor is not fast enough to maintain an optimal shutter speed. **Colour quantization** performed in low-profile cameras makes a high colouring difference when comparing a scanner obtained image and a digital camera one, **sensor noise** tends to also be a common problem in consumer-grade imaging devices. Finally, another important challenge in some digital cameras is that one must work with **compressed images** and has no access to the raw data.

Complex backgrounds and **occlusions** are the most difficult challenges to address in a completely

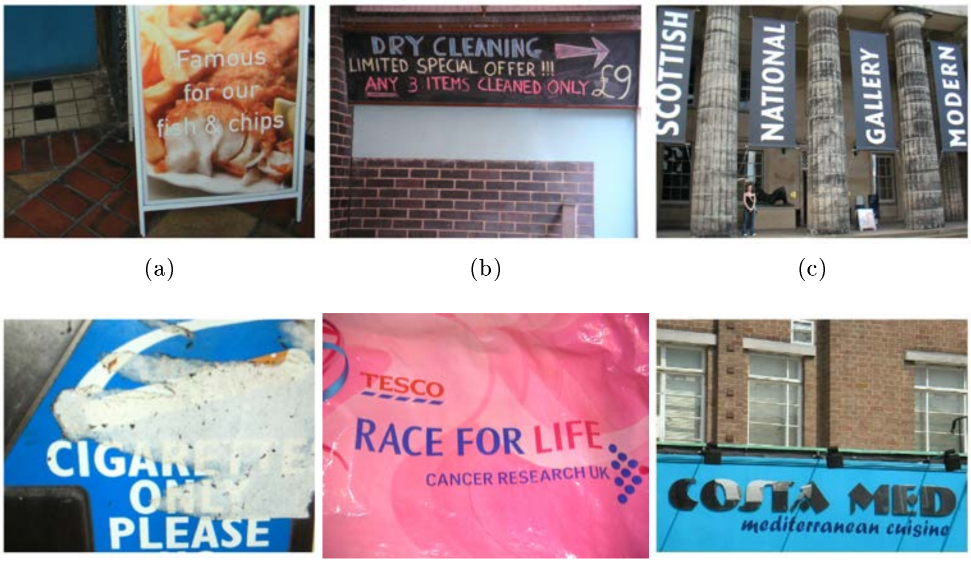


(a) (b) (c)

Figure 2.1: Examples of inherent challenges of Digital Camera imaging: (a) uneven focus, (b) blurring, (c) effects of image compession

uncontrolled environment, **low-contrast** and **uneven lighting**, that can appear in the form of reflections and shadows, makes things even worse. The projective nature of camera based imaging makes the planar text in natural scenes to appear with **perspective distortion**, other kind of distortion arises when text appears in **non-planar and non-rigid surfaces** or when using a **wide-angle lens**.

Besides, when one considers portable devices as the final target for automated text reading systems, another important challenge to have in mind is the necessity for fast algorithms, as **limited computing resources** are often available on the device.



(a) (b) (c)

(d) (e) (f)

Figure 2.2: Complexty of Natural Scenes: (a) cluttered background, (b) high font type variability, (c) vertical text, (d) occlusion, (e) non-plannar surface, (f) shadows and highlights.

2.1.2 Applications and socio-economic impact

Reading text in Natural Scenes can provide an important source of information for automatic annotation and indexing of images and video. Possible applications include digital image and video **archiving** both in **semantic labelling** and **keyword-based image search**, **video annotation** following MPEG-7 like standards, World Wide Web search engines, real-time information retrieval and targeted advertisement.

Assisted navigation and translation tools are another important kind of applications with recurrent appearance in the literature, on one side **elderly and visually impaired persons** and on the other **foreigner visitors** can take advantage of these technologies for reading printed information or text present in urban environments.

License plate reading, **address block location**, **cargo container and warehouse merchandise code reader** are some applications of camera based text reading suitable for diverse automated systems. In the same manner, text data could provide useful information to **mobile robots** and **intelligent vehicle systems**.

2.2 Text Localization and Extraction Approaches

The large number of techniques proposed for Text Localization and Extraction in natural scenes span bottom-up, top-down, and hybrid approaches. **Region based methods** make use of connected components typically in bottom-up approaches, grouping together regions likely to be text characters. On the other hand, **Texture-based methods** work in a top-down approach, usually performing a sliding window search over the image and extracting some texture features like e.g. HOG or HaarLike features in order to classify, using single classifiers or classifier ensembles, each possible window as text or non-text. Hybrid methods are being increasingly used for text localization and extraction tasks proposing many different combinations and flavours of bottom-up and top-down approaches.

2.3 State of the Art

For a recent region based method, one may refer to Epshtein et al. [12] where the authors propose a new operator, so-called Stroke Width Transform (SWT), assuming the fact that text has a constant stroke width, in order to group together by pairwise geometrical and colour similarities, the connected components (in the stroke domain) that are likely to be characters into text lines. Moreover, stroke width based features have been extensively used for the detection of text regions with promising results. Recent examples include work by Jung et al. [13] for character segmentation and Li et al. [14] who propose an approach based on multiple stroke integration and stroke filtering. Chen et al. have obtained state-of-the-art results in [15] with a method that determines the stroke width using a novel approach based on the Distance Transform in order to pairwise group connected components and form text lines; in the same work the authors propose an interesting technique to enhance the outline of Maximally Stable Extremal Regions (MSER) combining them with a Canny edge detector. A fast and simple region based method is proposed by Merino-Gracia et al. in [16], based on MSER as a candidate text region detector and using a cascade of text classifying filters on simple geometrical and texture features; the method provides, despite being far in performance of other state-of-the-art methods, a close to real-time implementation.

Regarding Texture based methods, Coates et al. [17], and in a different flavour Netzer et al. [18], propose a prior knowledge free method using an unsupervised feature learning algorithm to generate the features for classification and a linear SVM, trained for "text/non-text" discrimination using labelled windows from the ICDAR 2003 Dataset, that is finally exploited using a sliding window for text detection. Lee et. al. [19] propose the use of multi-scale sequential search approach with a Modest AdaBoost classifier trained with 6 types of feature sets including derivatives and edge information, Gabor Filter response, statistical texture measures and Connected Component Analysis. Another AdaBoost method, but using Haar-like features, is also proposed in [20] by Song et al. Wang et al. propose in [10], as an extension of their previous work in [21], an innovative end-to-end scene text recognition system based on a sliding window character detector using Random Ferns, where each character is a category of the classifier and the features consist of applying randomly chosen thresholds on randomly chosen entries in a HOG descriptor, trained with synthetically generated data and finally using the Pictorial Structures formulation to detect words in the image using a dictionary based grammar (lexicon).

Hybrid state-of-the-art approaches include Pan et al. [22] where a WaldBoost classifier is trained with HOG features to detect text-like windows in a top-down approach while in a second stage the authors propose a Conditional Random Field to label Connected Components as "text" or "non-text" before they finally create text line hypotheses in the scene again with a Conditional Random Field model. Neumann et al. [11] [23] propose an end-to-end method for localization and recognition allowing multiple text line hypotheses where an initial character recognition is done in a region representation derived from Maximally Stable Extremal Regions (MSER) and a SVM classifier is trained with geometrical features for "text/non-text" classification over each Extremal Region (ER); another SVM is then used for text line hypothesis formation and, after a perspective distortion rectification stage, character recognition is performed over all ER in each text line hypothesis, using a one-against-one SVM trained with 40 synthetic font types, and scoring each line hypothesis individually using a dictionary approach thus selecting the most probable hypothesis. Park et al. [24] perform an initial segmentation separating regions into chromatic and achromatic according to the colour distribution, followed by the identification of text objects by a SVM using moment features on the wavelet coefficients of the Wavelet Transformation of each component. Anthimopoulos et al. [25] propose a hybrid method where regions are detected based on the edge map of the image and then classified, in a refinement stage, with a SVM trained on features obtained by a new Local Binary Pattern based operator.

Kunishige et al. [26], in an innovative and clever proposal demonstrate the importance of environmental context, modelled by scene components such as sky or buildings, in regulating the probability of text existence at a specific region in an image. They first perform a Connected Component decomposition by trinarization to extract then a dual set of character and environmental context features that are then combined in a Random Forest based multi-stage classifier.

As conclusions from the state-of-the-art review one can highlight some trends in the task of Text Extraction from Natural Scenes. For instance, there is an increasing number of hibrid methods performing with top scores in the standard datasets. On another hand the recurrent use of stroke based features as well as the use of Maximally Stable Extremal Regions for character candidates detection is also remarkable. Finally, novel and original ideas arise, as the use of unsupervised learning or the charaterization of environmental context as a cue for text localization, in some way emphasising the increasing interest of researchers in the field. In parallel, more "traditional" techniques are still in use and achieving good results as well.

2.4 Datasets

The ICDAR 2003[1] and 2005[2] datasets used for the Robust Reading Competitions [27] [28] [29] have been made public and have become a common and standard benchmark tool for comparing the methods proposed by the researches, tracking the evolution and the performance of the current state of the art.

Other public datasets in use are the Street View Text[3] (SVT) dataset harvested from Google Street View and used in [21] [10], the Street View House Numbers[4] (SVHN) Dataset used in [18], the NEOCR[5] Natural Environment OCR Dataset [30], the KAIST[6] Scene Text Databaset [31] [32] and the dataset from Microsoft Research India (MSRI)[7] used in [12].

[1] http://www.iapr-tc11.org/mediawiki/index.php/ICDAR_2003_Robust_Reading_Competitions

[2] http://http://robustreading.opendfki.de/wiki/SceneText

[3] http://vision.ucsd.edu/~kai/svt/

[4] http://ufldl.stanford.edu/housenumbers/

[5] http://www.iapr-tc11.org/mediawiki/index.php/NEOCR:_Natural_Environment_OCR_Dataset

[6] http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database

[7] http://research.microsoft.com/en-us/um/people/eyalofek/text_detection_database.zip

Chapter 3

Perceptual Organization

Perceptual Organization relates to the processes by which humans are able to organize incoming sensations into meaningful information. From the point of view of the Gestalt psychologists the fundamental principle behind Perceptual Organization is the law of prägnanz, which states that we tend to order our sensorial experience in a regular, orderly, symmetric, and simple way. This elementary idea has been developed by Gestalt psychologists in what are called the Gestalt grouping laws of Perceptual Organization.

The first reference in the literature to the Gestalt grouping laws appeared in the seminal paper by Whertheimer [2] in which the author established in a descriptive way how objects with some common features get grouped together to form a gestalt, a larger visual object resulting from the sum of its parts. The theory was further developed by several authors [33] [34] [35]. A complete list of the elementary grouping principles found in more recent works [36] [37] [38] include:

- Proximity principle: leads us to group stimuli that are close together as part of a unique object.
- Similarity principle: leads us to group similar stimuli as part of a single gestalt.
- Good continuation principle: we tend to group together lines and forms that follow the same direction and not the ones that define an abrupt direction change.
- Closure principle: we tend to see complete figures or forms even if an object representation is incomplete or partially occluded by other objects.
- Symmetry principle: we tend to group symmetrical objects around their central point even when they are unconnected.
- Common fate principle: visual elements moving in the same direction at the same rate are perceived as part of the same gestalt.
- Constant width principle: we tend to group parallel lines together perceiving them as the boundaries of a constant width form.

- Figure-ground articulation: we tend to articulate our visual field into two different components: the ground and the figure, perceived as standing in front of the ground.
- Past experience principle: in some cases we tend to group stimuli together just because they were together often in our past experience.

Moreover, all the grouping laws described above can be applied in a recursive way, leading to the concept of global gestalts, formed by several partial gestalts.

The set of all Gestalt principles can be easily reproduced experimentally and still nowadays is widely mentioned and accepted by perceptual psychologists. None of them has been rejected, instead some new principles have been developed [39] [40]. However, the Gestalt laws of perception entail some unsolved concerns, such as the way they apply in cluttered images rather than the simple synthetic pictures normally used to illustrate them. A similar issue appears when two or more elementary grouping laws appear together not collaborating for a common organization but in conflict, sometimes one dominating the others, sometimes causing ambiguity between two or more possible interpretations. In this sense, a common critic point of view argue that Gestalt theory of Perceptual Organization provides us with a set of descriptive principles, but without a globally accepted computational model of perceptual processing [41] [42].

Another source of critics to the Gestalt grouping laws stem from its lack of the notion of Attention, a concept that plays a central role in today's perceptual psychology. However, some authors argue that the selectivity provided by Gestalt principles through a bottom-up saliency mechanism can be aligned with the notion of attention [43] and both processes complements each other with multifaceted interactive relations [44].

Besides, despite the fact that the Gestalt laws are in general considered to belong to a primary process of the visual system, there is not a single biological explanation on why they work. The gestalists originally presented the idea that the grouping laws are an inherent and fundamental part of the perception system shaping the way on how we perceive the world. On another hand, a more recent point of view suggest that the Gestalt grouping laws are related with the geometric statistics of the external world [45] [46], and somehow learned from our experience.

Up to here we have seen a general introduction to the Gestalt Theory of Perceptual Organization, emphasizing on one side both the critics and unsolved issues, that could lead to the idea that the theory is a kind of anachronism; and on the other, the fact that the theory is still alive, simply because it works. In the rest of this chapter we are going to concentrate in specific mathematical formulations of the Gestalt grouping laws that are going to help us to propose a method for text extraction in natural scenes inspired by Perceptual Organization.

3.1 Mathematical Formulation

Many mathematical formulations have been proposed by Computer Vision researchers in order to formalize the Gestalt theory principles. In [47] Zahn explore the natural relation between Gestalt grouping laws and graph theoretical methods for clustering. In a similar manner, Ahuja et al. [48] used Voronoi tessellations to extract perceptual structures in dot patterns. Zobrist et al. [49] have proposed a distance function for gestalt grouping, Santini et al. [50] and Shepard [51] have also published some theoretical discussion on Gestalt similarity functions. Ullman et al. [52], and Guy et al. [53] in a different approach, have proposed techniques to detect perceptually salient structures from local features. An interesting approach for embedding Gestalt laws in Markov Random Fields has been proposed by Zhu in [54]. More recently, Desolneux et al. [6] have proposed a Gestalt probabilistic approach based on the Helmholtz Principle; their method has been generalized for hierarchical clustering validity assessment in [55]. However, despite the efforts, it is still difficult to formulate in an explanatory way how Gestalt laws can be exploited in real world problems, especially when different partial gestalts conflict.

What follows in this chapter is an introduction to the Gestalt inspired clustering techniques which drives the research done in this Thesis and which proved useful in the development of the method presented in the next chapter.

3.1.1 Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis that fits pretty naturally with the Gestalt laws of Proximity and Similarity. The idea is to build a hierarchy of clusters from the input data in an agglomerative way: starting from each observation being a cluster itself and then iteratively merging the pair of clusters that are more close (or similar) following a linkage criteria; or in divisive way: starting from a single cluster holding all the observations together and recursively splitting each cluster. Obviously if the used metrics and linkage criteria are the same both agglomerative and divisive strategies result in the same hierarchy.

When going to search clusters in a set of observations one have to choose the appropriate distance metric according to the goals of the analysis. For instance, if we want to capture perceptual groups formed by the proximity principle and our input data are the 2D coordinates of a set of points in the plane, for sure we are going to use the Euclidean distance as distance metric. However, the decision is not going to be so obvious in the case of similarity measures depending on the features we use. Think for example in clusters formed by color similarity and the different available color-spaces that we can use to characterize our data. There are many different metrics that can be used to perform Hierarchical Clustering, and the decision is going to shape the results. So does with the linkage criteria which specify the dissimilarity of clusters as a function of pairwise distances of the observations in each set, being the most commonly used the Single Linkage Clustering (SLC), Complete Linkage Clustering (CLC) and the

Average Linkage Clustering. Appendix A presents a practical example of Hierarchical Clustering analysis for gestalt detection in a 2D point set where the differences between these linkage criterias is discussed.

As a final note on Hierarchical Clustering it is important for us to highlight here that once the Linkage Criteria is applied to the data, and a dendrogram is constructed, one have to choose heuristically the number of clusters in the resulting partition. Common strategies for this final step include the use of a fixed number of clusters and the selection of the partition with a maximum lifetime in the dendrogram. However, using a fixed number of clusters implies, like in the well known *k-means* clustering method, that we have, or can infer, prior knowledge about the underlying clusters in our data, and this is not the case in general when detecting gestalts as we want to do. On the other hand, selecting the partition with a maximum lifetime does not assure always a meaningful organization.

3.1.2 Graph Theoretical Methods

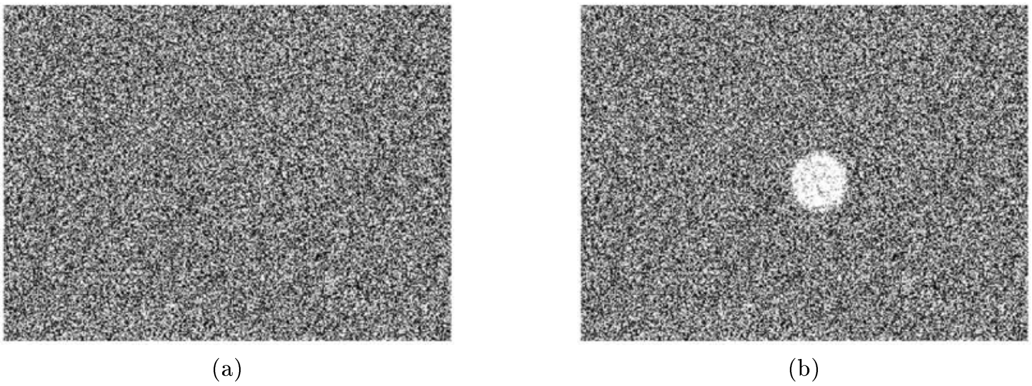
Hierarchical Clustering techniques are closely related with some Graph Theoretical Methods, in particular Single Linkage Clustering is in fact essentially the same as finding the Euclidean Minimum Spanning Tree (and sorting the edges) [56] [57] [58] [47]. In an early paper by Zahn [47] the author demonstrates how Graph Theoretical Methods can detect and describe Gestalt clusters in several grouping scenarios. Zahn [47] propose a method where the Minimal Spanning Tree of a given dataset is prunned removing those edges bridging separate clusters (see Appendix A for details). Thus, in the same manner as for Hierarchical Clustering, some heuristic threshold or rules should be defined in order to exploit Graph Theoretical Methods for Gestalt cluster detection.

3.1.3 A probabilistic Approach to Gestalt Theory

In this section we are offering an introduction to a probabilistic approach to the Gestalt laws of Perceptual Organization proposed by Desolneux at al. [6] [7] that has been developed in recent years and is gaining interest from researchers in both the Computer Vision and Cognitive Science communities. The theory has been applied to many different perceptual problems: detection of object alignments [6] [59], proximity and similarity clusterings [6], vanishing point detection [60], edge detection [61], good continuation [62] [63], shape recognition [64], and so on. Moreover the *"a contrario"* method proposed has been generalized for Hierarchical Clustering validity assessment [55]. The cornerstone of all this theoretical model is the Helmholtz Principle.

Helmholtz Principle

The Helmholtz Principle could be defined informally with the following sentence: "We don't perceive anything in a uniform random image", see Figure 3.1a. From this simple and quite logical sentence we can now define the same principle in an *"a contrario"* statement in this way: "Whenever in an image some large deviation from randomness occur some structure is perceived". This is, what we are going to search for in our data in order to detect gestalts is an arrangement of observations that has a very low probability to happen by chance. In Figure 3.1b for example, we have the same random distribution of samples as in Figure 3.1a but with a clearly noticeable deviation from randomness in the center of the image. This group of pixels constitute, under the Helmholtz Principle we have defined, a gestalt. The lack of perception in random images was first stated by Attneave [65], a gestalt psychologist, back in 1954; the same idea has been pointed by several Computer Vision researchers [66] [54], being Lowe [66] the first to pose it in probabilistic terms.


(a)

(b)

Figure 3.1: The Helmholtz Principle in action: (a)we don't perceive anything in a uniform random image, (b)whenever a large deviation from randomness occur a structure is perceived.

The Helmholtz Principle provides us the basis to derive a mathematical mechanism to automatically detect those deviations from randomness, or gestalts. For that, Deasolneux et al. propose the use of the binomial distribution:

$$\mathcal{B}(n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Under the independence assumption, the binomial distribution answers the following question: "What is the probability of k samples out of n to have a common property that for a single sample has probability p". As in fact we are interested in events with at least k samples, but not necessarily exactly k, having a common property, we are going to use the tail of the binomial function in this form:

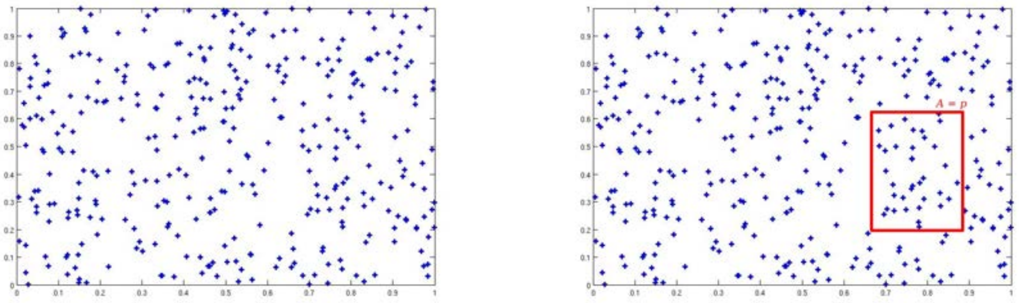
$$\mathcal{B}(k,n,p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

Under this mathematical formulation the task of searching gestalts in a given dataset should be approached, in a closely related way with the hypothesis testing statistical framework, calculating the probability of each possible configuration of samples to happen by chance. This is what Desolneux et al. have called the Number of False Alarms (NFA), which essentially entails the idea that meaningful organization of samples should have a small NFA and thus a very small probability to happen by chance, and have defined a ϵ-meaningful event as an event with:

$$NFA = N_{conf}\mathcal{B}(k,n,p) < \epsilon$$

Where N_{conf} is a constant value for the number of tests performed in a given dataset.

As a simple example, let's consider the uniform random distribution dataset of two-dimensional points in Figure 3.2. If we have a normalized feature space, i.e. x and y coordinates of the points in the 2D feature space are random numbers between 0 and 1, the area covered by the distribution is 1 and we can use the area of any region of the feature space as the probability of a sample to fall in this region. For example, the probability of one sample in our uniform random distribution dataset to fall in the right half of the feature space is obviously 0.5.



(a)

(b)

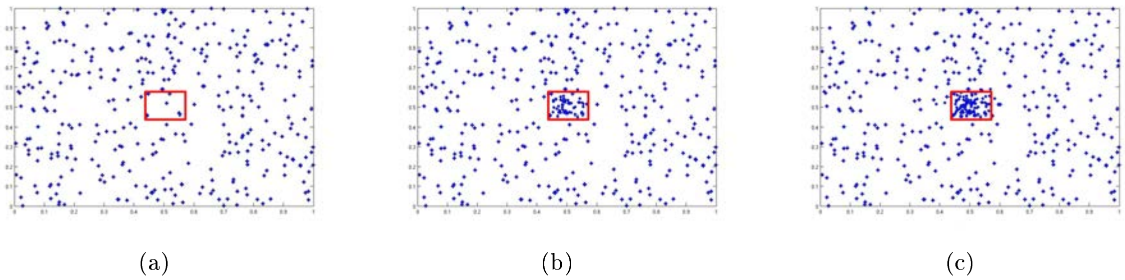
Figure 3.2: (a)A uniform random distribution two-dimensional dataset in a 1-normalized feature space, (b)the probability of one sample to fall in a given region is equal to the area of the region.

Thus, under the independence assumption, the probability of at least k samples out of n to fall in a rectangular region of the feature space with area p can be calculated with the tail of the binomial distribution:

$$\mathcal{B}(k,n,p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

Whatever the region chosen, if all samples in our dataset come truly from a uniform random distribution, as in Figure 3.2b, this probability will be close to 1, indicating that effectively this configuration has large probability to occur merely by chance. On the other side, when in some region of the feature space there is a deviation from randomness, see Figure 3.3, the NFA in this region will start decreasing indicating that a meaningful organization of samples exists.

The model shown here, using the area (or volume in higher dimension spaces) of the bounding hyper-rectangle containing a set of samples in a normalized feature space as a probability measure, is only one



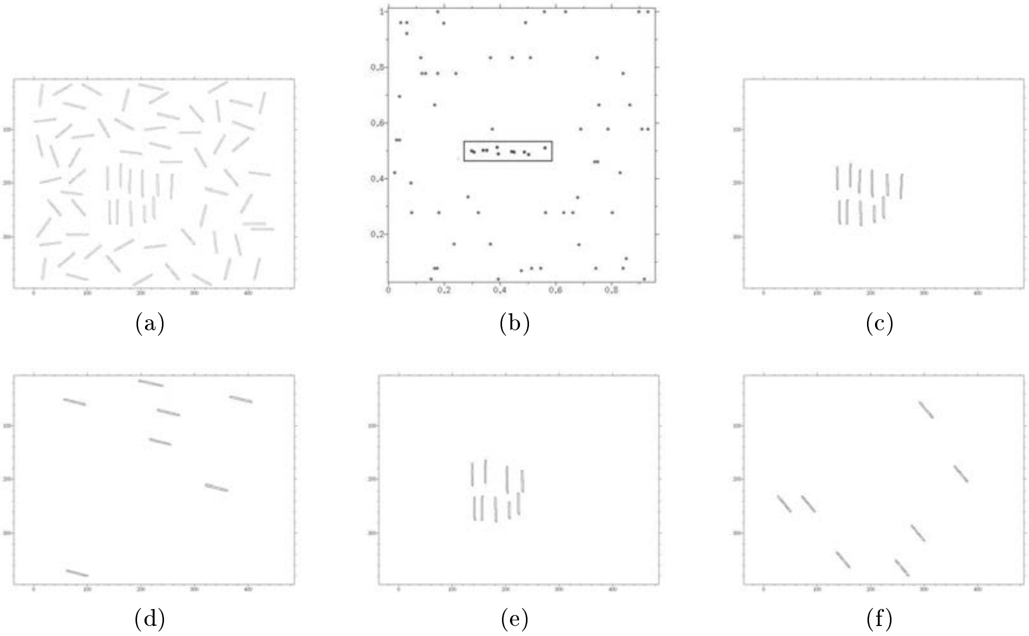
(a) (b) (c)

Figure 3.3: NFA of the subset in a given region of the feature space decreases as a deviation from randomness occur: NFA in (b) is smaller than NFA in (a), NFA in (c) is smaller than in (b).

of the possible ways to calculate the NFA of an event. In fact, Desolneux et al. have proposed different NFA formulations depending on the problem to solve. However, the model we have just introduced here is probably the most extensive, in the sense than can be used in any feature space independently on its dimensionality. Moreover, this method has been generalized for hierarchical clustering validity assessment by F.Cao et al. [55].

The idea is that using this method over the dendrogram of a Hierarchical Clustering, we can obtain a measure of the meaningfulness at each merge step by means of the NFA of this particular configuration. This procedure allows for comparison of meaningfulness in each branch of the tree and introduces the concept of maximal meaningful clusters. Detecting the maximal meaningful clusters in a given dataset is a parameter free process to identify gestalts in an image, omitting the necessity of heuristic thresholds which constitute one of the main drawbacks in the use of Hierarchical Clustering and Graph Theoretical Methods for detecting gestals. On another hand, the automatic detection of max. meaningful clusters does not make any assumption on the number of clusters we are looking for, and it's not constrained to full partitions of the data. This suits the problem we want to solve in the sense that, unlike some problems in data analysis where we expect to obtain a full partition of the input data, here we don't know if there are meaningful clusters at all, or even if the set of meaningful clusters will be or not a partition of the dataset.

As an example of maximal meaningful clusters detection in Figure 3.4 we reproduce an experiment done by F.Cao et al. in [55]: In Figure 3.4c a unique maximal meaningful cluster is found for the given dataset in the 2D feature space x-coordinate/orientation, while in Figures 3.4d, 3.4e, and 3.4f three maximal meaningful clusters are detected when analyzing only the orientation of the regions. This example, appart from serving as an ilustration of the described technique, serves to highlight the importance of the feature space selection in order to find meaningful information, in other words the technique described here will only be useful for Perceptual Organization if the feature space where our data is represented has the ability to express the kind of perceptual organization we are trying to detect.

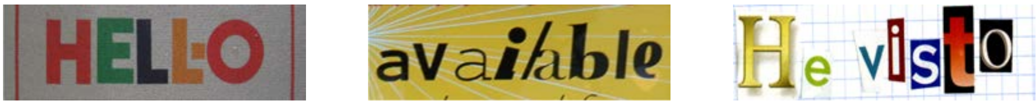

(a) (b) (c)

(d) (e) (f)

Figure 3.4: An example of maximal meaningful clusters detection reproduced from [55] (images are property of their authors). (a) A set of regions of the image to be analyzed, (b) 2D-representation of the objects in the space (x-coordinate,orientation) with the unique maximal meaningful group indicated by a rectangle, (c) the maximal meaningful cluster found in (b). On the other hand, three maximal meaningful clusters are found (d), (e) and (f) when using a single feature, the orientation of the regions, for clustering.

3.2 Clustering Ensembles

The Gestalt mathematical formulation by Desolneux et al. seen in the previous section is going to be an important part of the Text Extraction method presented in this Thesis. However, there is still an important decision to be made before we can use this technique for text detection: what features should we use in order to group regions of the image? Many methods in the literature of text detection in natural scenes assume that all characters in a word, or text line, have the same color and stroke width, and similar sizes, but this is not always true (See Figure 3.5). The different configurations in color, stroke or size of the characters in a single word could appear as a design factor but also as a consequence of natural scene challenges. For example, uneven illumination can produce strong colour differences between characters in the same word, the same applies for perspective distortion and the size of characters. Hence, we arrive to the conclusion that there is no single best feature (or set of features) in order to cluster character candidates for text detection.



(a) (b) (c)

Figure 3.5: There is no single best feature for character clustering: Characters in the same word may appear with different color(a), stroke width(b) and sizes(c).

A possible solution for solving the above problem is provided by the field of Clustering Ensembles [67] [68], an important and powerful methodology emerged as an elaboration of the classical clustering problem that has shown to be useful for improving the robustness and stability of existing unsupervised solutions. In a similar way as Classifier Ensembles do in Supervised Machine Learning, Clustering Ensembles combine a set of different "weak" clusterings in order to identify the true underlying clusters in a given dataset.

3.2.1 Evidence Accumulation

Evidence Accumulation is a simple but powerful Clustering Ensemble method proposed by Fred et al. [69] [70]. The main idea of the algorithm is to combine the information provided by an input set of different clusterings in a co-occurrence matrix as in a voting system. Such co-occurrence matrix can be used then as a similarity matrix in order to perform a final clustering over it. The input of the algorithm is a set of clusterings $\mathbb{P}$:

$$\mathbb{P} = \{P^1, P^2, P^3, ..., P^N\}$$

Where each P^i is the result of a different cluster analysis. $\mathbb{P}$ can be produced in several ways:

- (1) Using different data representations:
  - (a) Different pre-processing or feature extraction.
  - (b) Sub-sets of features.
  - (c) Bootstrapping (bagging).
- (2) Using different clustering methods:
  - (i) Different clustering algorithms.
  - (ii) Same clustering algorithm but different parameters.
  - (iii) Different inter-pattern similarity measures.

A voting scheme combines the outcomes of the initial clusterings using a co-occurrence matrix, $\mathcal{D}$, of object pairs defined as:

$$\mathcal{D}(i,j) = \frac{m_{ij}}{N}$$

where m_{ij} is the number of times the feature vectors i and j are assigned to the same cluster among the N initial clusterings. To decide on the final grouping, Hierarchical Clustering is applied on the co-occurrence matrix $\mathcal{D}$.

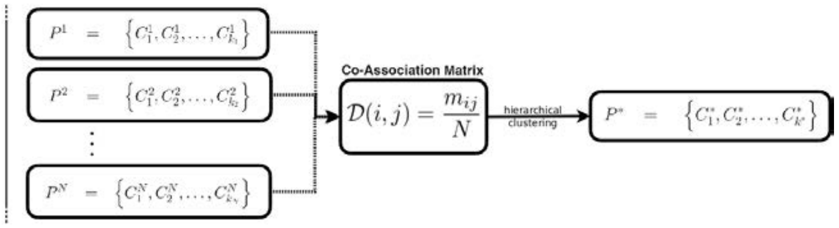
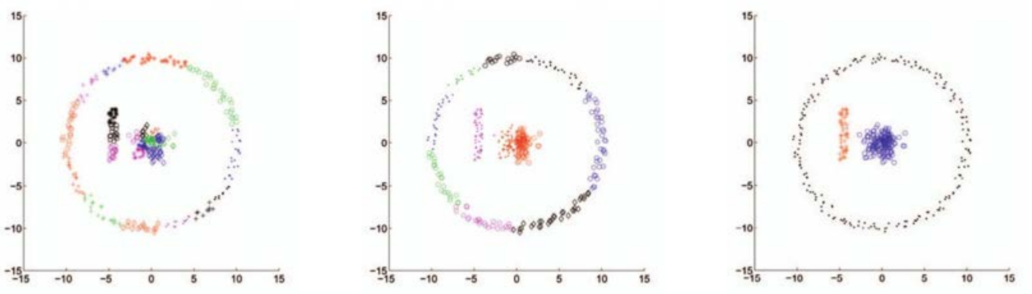

$P^1 = \{C_1^1, C_2^1, \ldots, C_{k_1}^1\}$
$P^2 = \{C_1^2, C_2^2, \ldots, C_{k_2}^2\}$
$P^N = \{C_1^N, C_2^N, \ldots, C_{k_N}^N\}$
Co-Association Matrix
$\mathcal{D}(i,j) = \frac{m_{ij}}{N}$
hierarchical clustering
$P^* = \{C_1^*, C_2^*, \ldots, C_{k^*}^*\}$

Figure 3.6: Flow diagram of the Evidence Acumulation Clustering algorithm.

Although Evidence Accumulation Clustering is not inspired by Perceptual Organization, the objective of detecting natural clusters directly links Perceptual Organization when the features describing the data have a perceptual meaning. As an example in figure 3.7 we can see a reproduction of an experiment realized by Fred et al. in [69] where the result obtained by EAC in a difficult dataset of 2D points match exactly with the partition one would expect from a "perceptual" point of view. This behavior makes the EAC a valid approach in the task of detecting visual gestalts in an image; a proposal in this direction can be found in the work by Alajlan [71] for retrieval of hand-sketched envelopes in logo images.

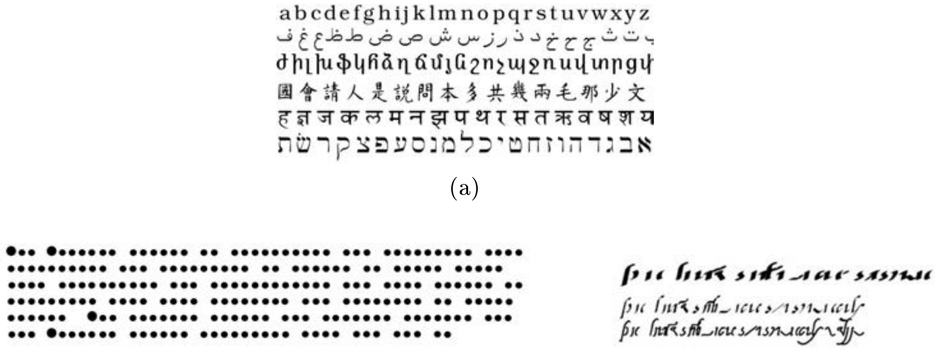


(a) (b) (c)

Figure 3.7: Evidence Accumulation Clustering results on a difficult dataset: (a) and (b) Two different *k-means* runs with, respectively, $k = 25$ and $k = 11$, (c) The resulting clustering appling EAC on 200 different *k-means* clusterings. This images are reproduced from [69] and owned by A. Fred and A. K. Jain.

3.3 The Role of Gestalt Laws in Text Detection

Many researchers in cognition and perception psychology have pointed to the close relationship between the Gestalt laws of perception and the core of the writting systems [3], arguing the gestalt laws as one of the main processes involved in the reading activity [4] [5]. This relationship transpires in an intuitive way when observing that writing systems at large share in common that Gestalt principles of proximity and similarity between glyphs are present. Not only alphabetic scripts but also ancient pictographic and ideographic writting systems seem to be designed in a way to exploit the power of Perceptual Organization rules (See Figure 3.8a): We use to write aligned and equally separated glyphs with high contrast to its background, with a constant stroke, and similar color and sizes; despite, as seen in Section 3.2, those similarity rules mentioned are not strict but allow a certain flexibility. Moreover, this rules apply independently of the style of the glyphs, so a kind of text structure can be perceived also when objects without meaning are organized in this way (See Figures 3.8b and 3.8c).



(a)

(b)

(c)

Figure 3.8: All writing systems seems to be designed in a way to exploit the power of Perceptual Organization laws (a). Text structure is independent on the glyphs style: (b) a text-like pattern using black dots, (c) random scrawl experiment[1](Figure owned by Daniel Uzquiano).

In a similar manner as when we talked about the biological explanation of Gestalt laws, one may be asked why and how this common geometric structure emerged in the evolution of writing systems. Either way, it is compelling that a strong relationship exists between writing systems and perceptual organization. Accordingly, many times in the literature one finds text elements used as examples on how Gestalt laws apply in different situations [2] [3] [36] [37]; moreover, gestalt principles are taught as important guidelines for graphic designers, especially those working in new font types and graphical user interfaces.

Regarding the task of automatic text detection, the kind of information provided by gestalt laws of perception is sometimes exploited in region based bottom-up methods in a post processing stage, where a rule based methodology is used for text candidates grouping and at the same time to filter out false detections [16] [15] [72] [12]. The methodology proposed in this thesis goes beyond the mere use of heuristic rules when it comes to exploiting the perceptual organization, placing it at the heart of the text extraction method for natural scenes presented in the next Chapter.

[1] `http://danieluzquiano.com/491/entre-el-garabato-y-la-letra`

Chapter 4

A Perceptual Organization Method for Text Detection in Natural Scenes

Building a Text Extraction Method inspired by Perceptual Organization necessarily involves a clustering procedure where atomic objects, in our case character candidates, are grouped together in larger gestalts, i.e. words and text lines. As we have seen in Chapter 2 similar grouping processes are a common step of bottom-up based methods for text detection. The novelty of the method proposed here is that this clustering procedure is done in an early stage of the work-flow (Fig.4.1) and stems directly from a Perceptual Organization model, while other methods usually do the grouping in a heuristic post-processing stage. One of the main contributions of our method is to show that looking separately into perceptually meaningful groups of regions, rather than to the whole connected component set of the image, the task of text detection is better posed. Thus, we suggest a region based method, making use of connected components in a bottom-up approach, where regions are grouped together under the evidence of simple features like colour, size, aspect ratio, or stroke width among others, in order to obtain meaningful groups likely to be text gestalts, i.e. paragraphs, text lines, or words.

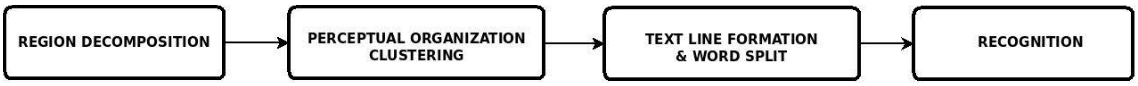

REGION DECOMPOSITION
PERCEPTUAL ORGANIZATION CLUSTERING
TEXT LINE FORMATION & WORD SPLIT
RECOGNITION

Figure 4.1: Pipeline of the proposed method for Text Extraction in Natural Scenes..

At this point we can state two initial assumptions we are making in order to build our method, one at the atomic level of characters, and the other at the grouping level of words and text lines.

At the atomic level we assume that characters are non overlapping connected components of the image, with a constant colour, and a noticeable contrast with their immediate background. This assumption follows naturally from the fact that text is designed to be read by human beings, therefore a high contrast is employed by design to make it stand out. We are going to see in the next Section how this assumption perfectly fits with the definition of Maximally Stable Extremal Regions(MSER), and so we are going to use the MSER technique for the initial region decomposition of the image.

At the group level we assume that Gestalt laws of proximity, similarity, and good continuation are always present in text. However, as we don't know exactly in which way this grouping principles collaborate or conflict, specially when talking about "similarity" as seen in Section 3.2, we are going to formalize a flexible definition of what "similarity" means for us; that is more permeables at the time of considering whether two character candidates should be grouped together or not. This second assumption has lead us to use a combination of the Meaningful Cluster concept, as defined by Desolneux et al., and the Evidence Accumulation method in a common step for Perceptual Organization Clustering.

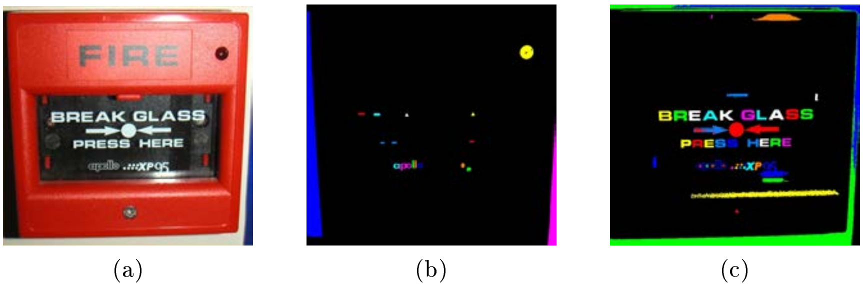
Although the detection of perceptually meaningful text groups is a desirable goal itself, the results obtained by our Perceptual Organization Clustering cannot be directly compared with other methods, because the lack of a reference benchmarking dataset at this grouping level. In fact, perceptually meaningful text groups might arise at different levels, that correspond or not to semantically defined ones (i.e. paragraphs, text lines and words). These groups, although perceptually meaningful, rarely correspond directly to the level ground truth information is defined. Hence the perceptual organization capabilities of our algorithm are difficult to evaluate in a direct manner. Thus, our method proposal is extended with simple post-processing in order to obtain word level bounding boxes and be able to evaluate on the ICDAR2003 Dataset. We are going to search for text lines candidates in the detected meaningful groups, and subsequently split those lines into single words. Finally, a simple template matching technique allows the detection and filtering of repetitive structures which tend to be confused with text, while a set of simple rules for noise removal are applied before sending candidate regions to the final OCR stage.

4.1 Maximally Stable Extremal Regions

As noted in Section 2.3 the use of Maximally Stable Extremal Regions (MSER) [73] algorithm for detecting text character candidates in natural scene images is a common trend in state of the art methods for text detection. MSER was originally proposed by Matas et al. [73] as a method for the detection of robust image features applied to the wide-baseline stereo problem, and has been recently used as part of systems for license plate detection [74] [75] and text detection in natural scenes [13] [16] [11] [23]. The main advantages of the use of MSER are its fast computation and its high illumination invariance, making the segmentation process robust over changes in lighting conditions and more suitable to work with different imaging sources than other approaches based for example on thresholding or color clustering. The whole MSER component tree can be build in linear time over the size of the input image [76].

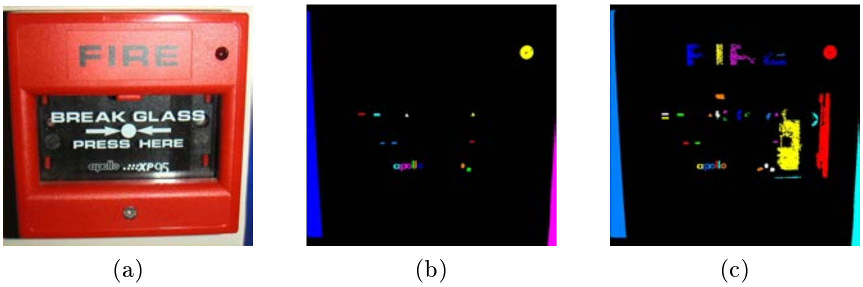
Moreover, the original grey-level MSER detector has been extended to colour images [77] and volumetric regions [78] [79].

MSER builds a tree of regions with an extremal property of the intensity function over its outer boundary, this is regions with a noticeable contrast to their immediate background, and this property is normally present in all text characters as they have been designed with high contrast to be easily read by humans (see Figure 4.2).



(a) (b) (c)

Figure 4.2: An image from the ICDAR2003 dataset (a) its MSER- (b) and MSER+ (c).



(a) (b) (c)

Figure 4.3: Effect of the Δ cut-off parameter: (a) an image from the ICDAR2003 dataset, (b) MSER- with $\Delta = 25$, (c) MSER- with $\Delta = 9$.

The MSER algorithm search then for regions with high shape stability in the sequences of nested regions, bounded by a cut-off parameter Δ, in the component tree. Notice that the larger Δ value the highest the contrast between detected MSER and its boundaries (see Figure 4.3). The MSER implementation used in our method, provided by the well known vlfeat[1] open source C++ library by Andrea Vedaldi and Brian Fulkerson, use four more parameters in order to discard regions wich are too small or too big (*Min_size* and *Max_size*), too unstable (*Max_variation*), or duplicated (*Min_diversity*). See in Appendix B for details on the MSER definition and implementation parameters.

[1] http://www.vlfeat.org/

We make use of the following set of fixed parameter values validated over the ICDAR2003 train set: $\Delta = 25, Min\_diversity = 1, Max\_variation = 0.7$. Aditionally, in the proposed method the resulting MSER tree is prunned by filtering the regions that are not likely to be characters by their aspect ratio, the variance of the stroke width, and the number of holes of the region.

4.2 Perceptual Organization Clustering

Our Perceptual Organization Clustering is going to be applied to the whole MSER tree in a two layer architecture (See Figure 4.4) where the clustering techniques described in Chapter 3 are combined together. In the first layer, meaningful clusters are detected using the probabilistic approach to Gestalt Theory proposed by Desolneux at al. [6] [7]. This meaningful cluster detection is done separately in several hierarchical representations of the MSER set, each one using a different similarity feature space, thus providing a meaningful clustering ensemble. In the second layer, all those initial clusterings are combined using the Evidence Accumulation method proposed by Fred at al. [69] [70], in a single clustering where groups of MSERs emerge from the evidence of having some sort of text structure, despite no closed definition is being provided from what a text group consists of.

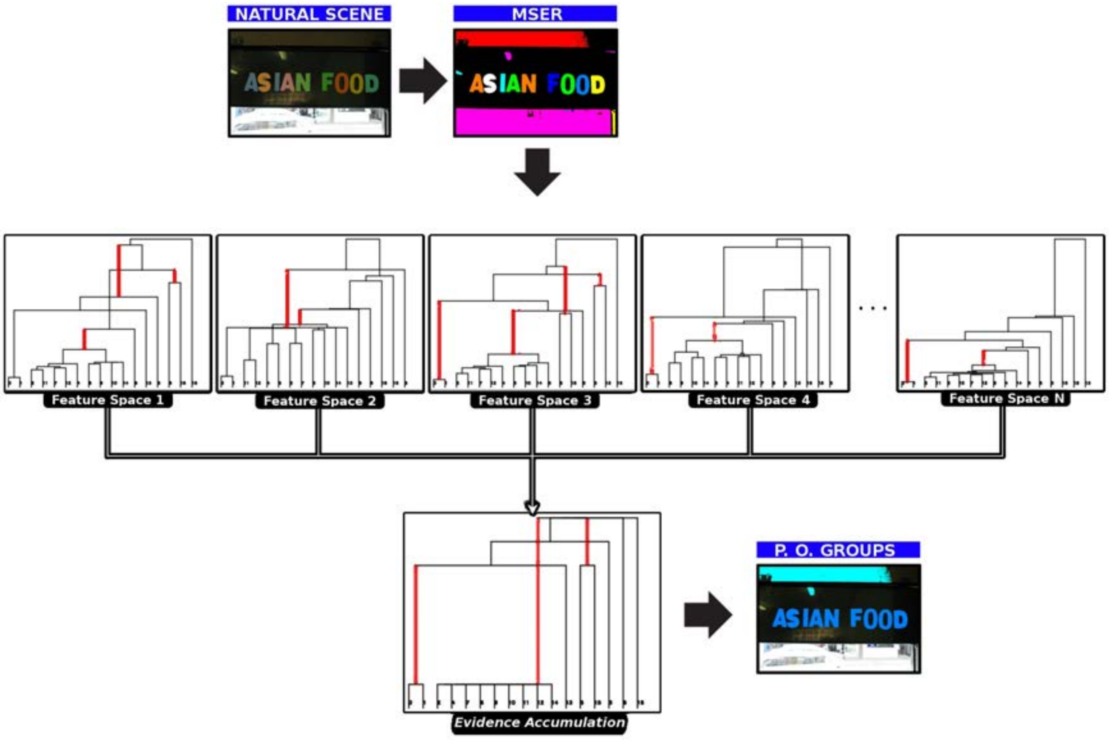


Figure 4.4: Detail of the work-flow in the Perceptual Organization Clustering stage of the proposed method.

We aim to use simple and low computational cost features that describe similarity relations between characters of a word or text line. So, in a sense, we are treating each MSER region as a text character candidate, and classifying it as character if and only if it can be grouped with neighbouring regions in a text like structure. A possible list of features one can use for this kind of similarity grouping are:

- Intensity value or color information of the inner region.
- Intensity value or color information in the outer boundary of the region.
- Simple geometrical features like area or diameter of the enclosing circle of the region.
- Stroke width of the region.

From the list above, the stroke width of a region is perhaps the only feature strictly related with text detection. Its efficiency has already been discussed in Section 2.3 where we saw that it is extensively used in recent text detection methods. To determine the stroke width of a region we make here use of the Distance Transform as proposed in [15] by Chen et al.

On another hand, colour information and simple geometrical features are also commonly used in text detection methods as pairwise similarity measures for grouping text characters together.

Our clustering ensemble $\mathbb{P}$ is a set of clusterings $\mathbb{P} = \{P^1, P^2, P^3, ..., P^N\}$ where each P^i is the result of a cluster analysis of the regions in the image using a different data representation provided by a sub-set of features. A simple option for constructing $\mathbb{P}$ is to analyze each feature separately in a unidimensional feature space. Instead, we propose the use of combined feature spaces where each one of the "similarity" measures is coupled with spatial information, i.e. x,y coordinates of the region center, in order to capture the collaboration of the Gestalt laws of proximity and similarity. Thus, the possible feature sub-sets to construct the clustering ensemble are:

#	Features	Dim.
1	Intensity mean of the inner region + x,y coordinates of the region center	3
2	L*a*b* color mean of the inner region + x,y coordinates of the region center	5
3	Intensity mean in the boundary + x,y coordinates of the region center	3
4	L*a*b* color mean in the boundary + x,y coordinates of the region center	5
5	Area of the region + x,y coordinates of the region center	3
6	Diameter of the enclosing circle + x,y coordinates of the region center	3
7	Stroke width of the region + x,y coordinates of the region center	3

Table 4.1: Feature sub-sets to construct the clustering ensemble.

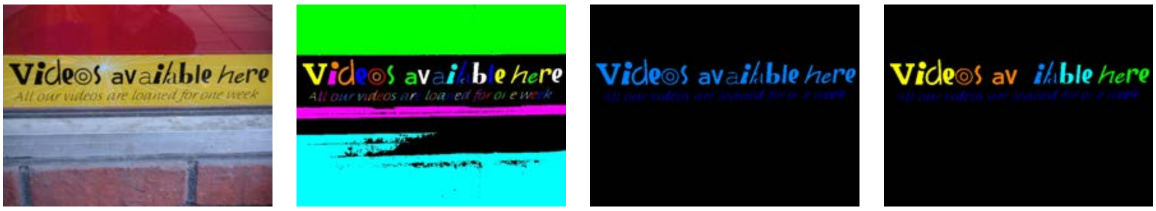
What we are trying to express with this coupled feature spaces is the fact that the proximity law of Perceptual Organization is something always present as far as text is concerned, in the sense that we can not talk about text characters in the same line if there is not a "reasonably" small distance between them. So, independently of the similarity measure we are going to evaluate, we restrict the groups of regions that are of interest to those that comprise spatially close regions.

The list of feature sub-sets in Table 4.1 can be extended with other similarity features, also some of them can be omitted as they encode highly correlated information, e.g. intensity and colour are probably going to result in similar clusterings. Moreover all these feature spaces can be combined in different ways creating different clustering ensembles. In the results section we provide the different f-scores obtained over the ICDAR2003 dataset by using different combinations of these feature sub-sets.

Whatever the number and nature of the different feature sub-sets we use, the way to combine them is as follows. Each one of the feature sub-sets defines a similarity hierarchy of the regions of the image based on which we can evaluate how meaningful is a given cluster of regions at each merge point of the hierarchy tree by means of its NFA:

$$NFA = N_{conf}\mathcal{B}(k, n, p)$$

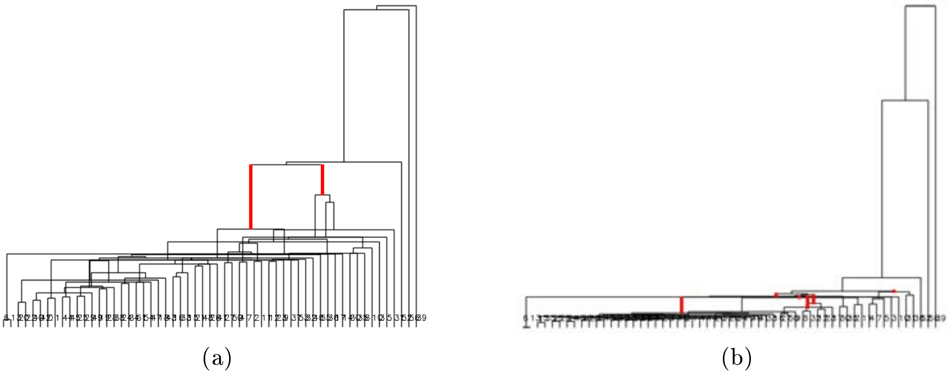
Being N_{conf} a constant value, and thus not relevant in order to compare the NFA of different clusters in the hierarchy, we aim to detect the maximal meaningful clusters by comparing the probability of at least k of our n samples falling in a region of the feature space with a relative area p, this is the tail of the binomial distribution $\mathcal{B}(k, n, p)$ at each merge step of the similarity tree. Assuming a uniform noise distribution as the background random process producing the input data, detecting the smallest values of the NFA in the hierarchical clustering of a given feature sub-set means detecting the clusters that have the largest deviations from randomness, corresponding to detecting the maximal meaningful clusters from a Perceptual Organization point of view.



(a) (b) (c) (d)

Figure 4.5: A natural scene image from the ICDAR2003 dataset (a) with its MSER decomposition (b), and the maximal meaningful clusters in feature spaces #1 and #7 of Table 4.1 (c) and (d).

Figure 4.5 shows the maximal meaningful clusters detected in a natural scene image using two different feature sub-sets, in Figure 4.5c regions are clustered by proximity and intensity value similarity, while in Figure 4.5d they are clustered by proximity and stroke width similarity. This situation is a good example of the kind of conflict between gestalt laws that our Perceptual Organization Clustering is able to solve by means of the Evidence Accumulation method. Figure 4.6 show the dendrograms of the hierarchies resulting from the clusterings in Figure 4.5, maximal meaningful clusters are indicated with a red line. Notice that the clustering analysis is done without specifying any parameter or cut-off value and without making any assumption in the number of clusters k, but just comparing NFA values of each branch merge in the dendrogram. Thus, we select as maximal meaningful a cluster A if for every successor B and every ancestor C, one has $NFA(B) > NFA(A)$ and $NFA(C) \geq NFA(A)$. Notice that using this maximallity criteria no region is allowed to belong to more than one meaningful group at the same time.


(a)

(b)

Figure 4.6: Dendrograms of the clusterings in Figure 4.5c and 4.5d, maximal meaningful clusters are indicated with a red line.

Once we have detected the set of maximal meaningful clusters P^i in each feature sub-set $i \in N$, the clustering ensemble $\mathbb{P} = \{P^1, P^2, P^3, ..., P^N\}$ is used to acumulate grouping evidence of each pair of regions in the co-occurrence matrix $\mathcal{D}$ defined as:

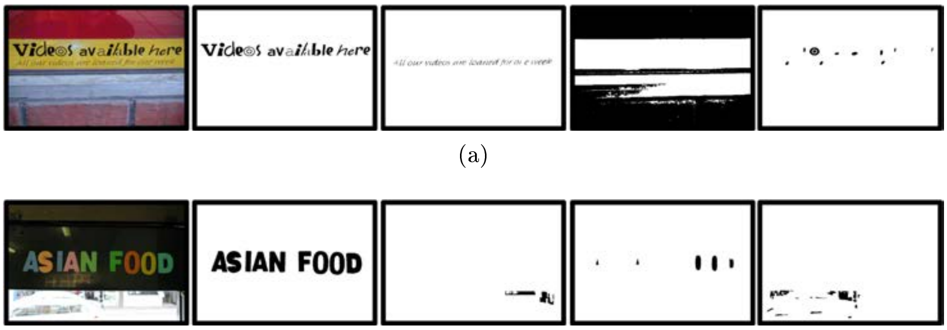
$$\mathcal{D}(i,j) = \frac{m_{ij}}{N}$$

Where m_{ij} is the number of times the regions i and j have been assigned to the same maximal meaningful cluster in $\mathbb{P}$. Hence, if two regions a and b have not been clustered together in any of the N initial clusterings in $\mathbb{P}$ the value $\mathcal{D}(a,b)$ is zero, while if a and b have been clustered N times in $\mathbb{P}$ the value of $\mathcal{D}(a,b)$ is one, indicating a strong evidence that a and b are part of a meaningful structure. The co-occurrence matrix $\mathcal{D}$ can now be used as a dissimilarity matrix in order to perform the final clustering of the regions.

While Fred at al. propose in [69] the use of the maximum lifetime criterion as an automated mechanism to decide on the number of clusters in the final Evidence Accumulation step, here we make use of the Maximum Meaningful Cluster detection in the co-occurrence matrix $\mathcal{D}$. There is however an important difference between the hierarchy in this final clustering analysis and the ones made when constructing the clustering ensemble, in this case we do not have a feature space where feature vectors representing our sample data can be spatially located, but instead we just have the distances between each pair of data samples. Thus, there is no direct way to calculate the area of the bounding volume of a set of samples in its feature space and we can not evaluate the NFA value as we did before. To overcome this drawback we propose the use of Complete Linkage Clustering over the matrix $\mathcal{D}$ and using the normalized distance where each merge is done as the probability value for the NFA formula. In Complete Linkage Clustering the criteria used to choose the two different clusters A and B that are going to be merged at each step of the clustering process is:

$$max\{d(a,b) : a \in A, b \in B\}$$

Thus, the distance at which A and B are merged in a single cluster can be used as an approximation of the volume of the cluster $\{A \cup B\}$ relative to the distance of the more distant pair of samples. Despite the approximative nature of this procedure, in the results Section its validity is demonstrated experimentally, providing better results than using the maximum lifetime criterion.



(a)

(b)

Figure 4.7: Natural Scene images and the clusters detected by Perceptual Organization Clustering: Text regions are grouped together despite (a) text characters not having the same stroke width but the same colour, and in (b), on the contrary, text characters not having the same colour but having the same stroke width.

Figure 4.7 shows two examples of the results obtained with the proposed Perceptual Organization Clustering where the method exhibits its ability to deal with a flexible definition of what a text group is: in Figure text characters do not have the same stroke width but the same colour, while in Figure , on the contrary, text characters do not have the same colour but have the same stroke width. In Figure 4.8 more examples of Perceptual Organization Clustering are shown (some small and non relevant groups are omitted), notice how text characters are always grouped together.

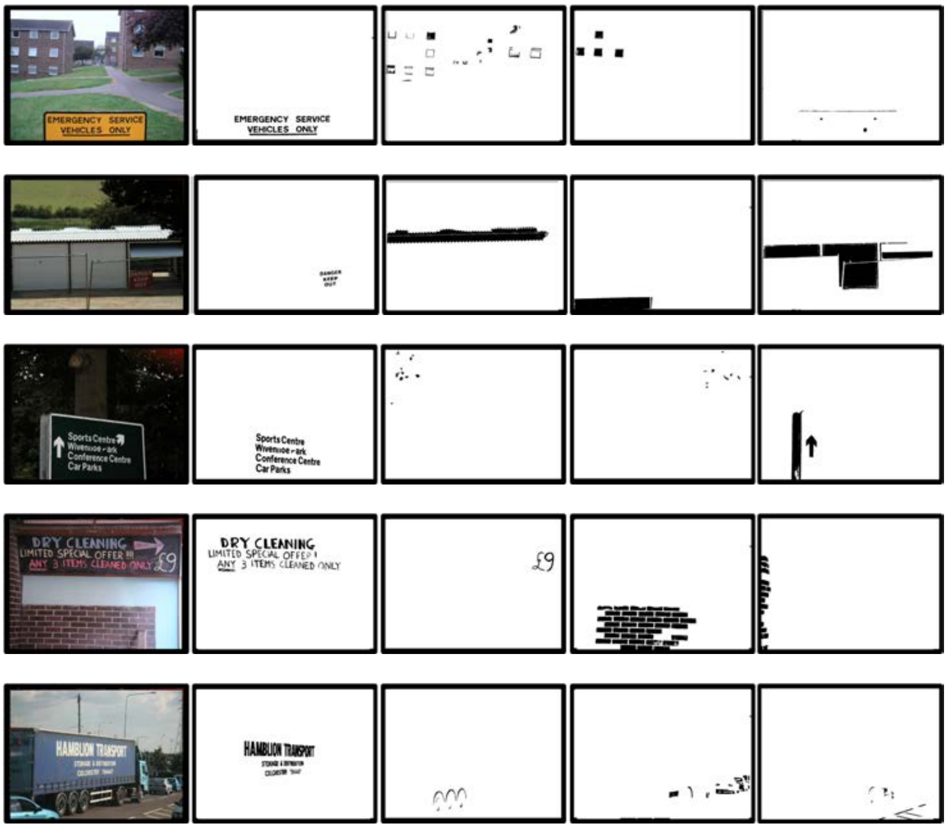


Figure 4.8: Resulting clusters of the Perceptual Organization Clustering method in natural scene images from the ICDAR2003 dataset.

As expected, not only text is detected as meaningful, but also any kind of region arrangement with a text like structure: the windows in a building, bricks, car wheels, or traffic lane marks are also detected as meaningful structures of the image because in some sense these organizations are perceptually relevant behind the criterion we have modelled. The next steps in the pipeline of the presented method are going to deal with this non-text groups in order to discriminate and omit them in the final text detection sets. Nonetheless, the algorithm producing relatively pure clusters, where almost all text characters fall in clusters that contain virtually no non-text components is very helpful in the task of filtering non-text meaningful groups.

4.3 Text Line Formation and Word Split

Once we have grouped together those regions that are likely to be part of a text structure in the image, we can analyze each of these groups in order to find possible text lines and split them in separated words if possible. At the same time this procedure is going to allow the detection of clusters which despite having a meaningful text structure are not possible to be text because they cannot be organized in valid text lines.

The definition of what is a valid text line is taken from the typographic model of a text line shown in figure 4.9; notice that here we assume collinear text characters. In order to evaluate whether a given set of regions forms a valid text line itself we can analyze the distribution of the y-coordinates of the regions centres with respect to the centre y-coordinate of the bounding box of the whole group. According to the typeface model in figure 4.9 there exists an interval around the y-center of the text bounding box where all the samples of this distribution should lie. We consider a set of regions as a valid text line if the mean of this distribution lies in an interval of 40% in the middle of the box height and the Coefficient of Variation of the distribution is lower than 0.21. This threshold values have been validated, using the training set of the ICDAR2003 dataset, in order to deal in the better way possible with a large font type variety and with independence of the possible combinations of small caps, capital letters, and characters with descendant or ascendant in the same line (see Figure 4.10). Notice that, as we are considering text lines at any possible orientation, the centres of the regions may be rotated, if needed, according to the angle of the circumscribed rectangle of minimal area for the given group (see Figure 4.11).

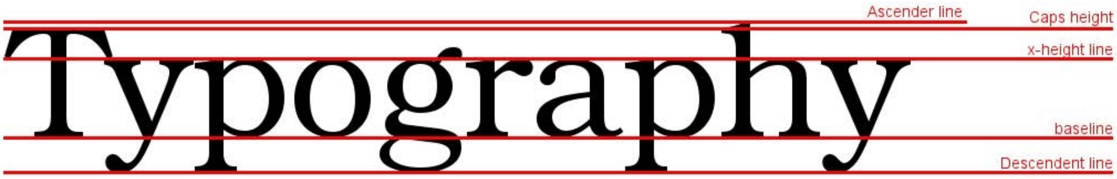
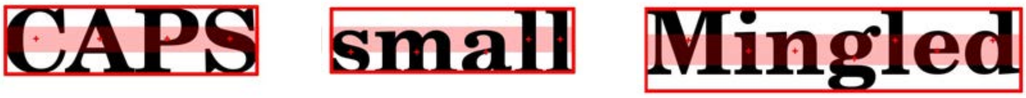


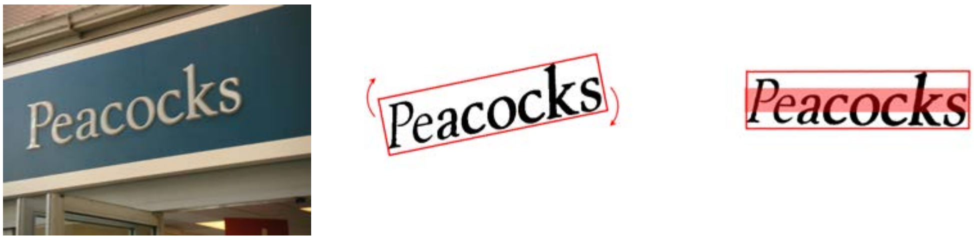
Figure 4.9: Typographic model of a text line.



(a) (b) (c)

Figure 4.10: Distribution of the character centres y-coordinates in a text line: a valid text line can be characterized by its mean and Coefficient of Variation.

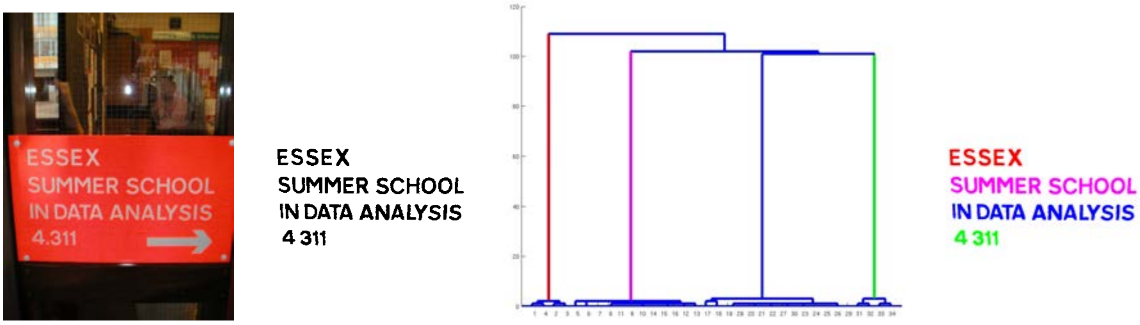
This simple test, together with non-overlaping and minimum height ratio constrains, suffices for us to identify groups of regions likely to form a text line, this is meaningful clusters of regions where the good alignment principle is present. Thus, given a Meaningful Cluster obtained from the previous stage in our method's pipeline we can first check if it can be considered a text line itself. Otherwise, if the collinearity test fails, it may be the case that the group analyzed is not a single text line but a group of lines, and thus we should try to separate them by a hierarchical clustering analysis of the distribution of regions centres y-coordinates. In this case we perform a histogram projection analysis in order to



(a) (b) (c)

Figure 4.11: (a) A natural scene image from the ICDAR2003 dataset, (b) a cluster of regions and its circumscribed rectangle of minimal area, (c) a rotated version of the regions in (b) where our definition of what is a valid text line make sense.

identify the text lines orientation, landscape or portrait with respect to its bounding box, and then we do hierarchical clustering with Single Linkage and cutting the dendrogram with the maximum lifetime criterion on the the distribution of the regions centres y-coordinates (see Figure 4.12). The process is iteratively repeated until all regions have been assigned to a valid text line, using the same collinearity test described above, or until no more partitions can be done.

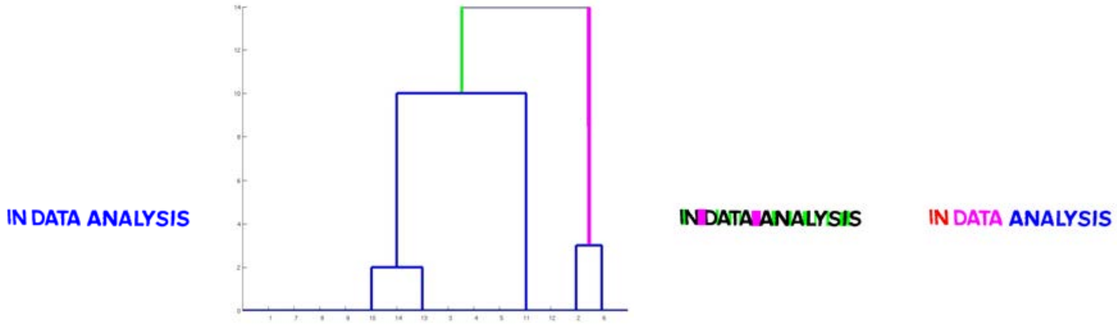


(a) (b) (c) (d)

Figure 4.12: Line separation by Hierarchical Clustering over the distribution of the y-coordinates of the regions centres: (a) a natural scene image from the ICDAR2003 dataset, (b) one of its perceptually meaningful clusters, (c) Hierarchical Clustering over the distribution of the regions centers y-coordinates, (d) text lines found by maximum lifetime in the clustering dendrogram.

After the text line formation is done, a similar procedure can be devised in order to split the text line candidates found into words, using in this case the inter-region distances distribution. For that, we sort horizontally the regions by its centre x-coordinate, filtering those that fall completely inside another region, and then calculate the distance between each pair of consecutive regions as the x-axis distance between its nearest pixels.

In a similar way as we did for validating text lines, a group of aligned regions is considered a valid text word if the coefficient of variation of the inter-character distances distribution is smaller than a threshold T_w, a value of $T_w = 0.65$ has been obtained again by validation over the ICDAR2003 training dataset. So, given an hypothetical text line the first thing to do is to check if it can be considered a word itself using the test just described. If the validity test fails, it may be the case that the text line has more than one word and then we proceed again with a Hierarchical Clustering analysis for the detection of two different types of inter-regions distances, namely the inter-character space and inter-word space distances (See Figure 4.13). In this case the dendrogram cut is done by a fixed number of clusters, $k = 2$, as we know the desired inter-region distance partition, and finally the words are split using the inter-word distance spaces found, which should be the cluster with a larger mean among the two clusters detected. Again the clustering procedure is performed iteratively until all the regions have been assigned into valid words or no more partitions are possible.



(a) (b) (c) (d)

Figure 4.13: Word separation by Hierarchical Clustering over the distribution of the inter-region distances: (a) one of the regions group identified as a valid text line in Figure 4.12, (b) Hierarchical Clustering analysis over the distribution of the inter-region distances, (c) inter-region partition for $k = 2$ in the clustering dendrogram, (d) the word separation found.

It is important to notice that the procedures proposed for text line formation and word separation have important limitations: mainly both methods are very sensible to noise in the incoming cluster of regions; besides, character kerning in some types of fonts could result in wrong word separation as well as it may happen if some characters in a word are missing in the region decomposition step. However, the proposed method is simple and effective enough for our purposes, especially because a good performance in the previous stages should assure the absence of noise or missing regions in the text clusters.

4.4 Recognition Feedback

Up to this point all the procedures described as part of our proposed method have a pure bottom-up approach, just trying to identify meaningful arrangements of regions which have a text-like structure in terms of proximity, similarity, and good continuation principles of the Perceptual Organization theory. However, the method is prone to produce a number of False Positives, mainly confusing for text regular human-made structures in the scene, like windows in a building, or other kind of region regularities in the scene. In fact, as already stated before in Section 3.3, a text-like structure does not guarantee the presence of text until it can be effectively read in some way. Thus, at this point a recognition step can be done in order to try transforming the forms in meaningful structures into meaningful concepts, providing a valuable feedback on what is really text and what is not.

Furthermore, although our method's objective is to exploit the laws of Perceptual Organization for text detection and not to recognise the text found, the feedback obtained by recognition with an optical character recognition (OCR) step is used to improve the performance of text detection. For this we have integrated the well known tesseract[2] open source OCR library into our method's pipeline. The recognition score, at character level, for each region is evaluated in order to detect regions with a very low confidence of being letters; moreover, if all regions in a word cluster are recognized as the same character (can happen in the case of repetitive non-text structures) we eliminate the word as probably being a False Positive.

It is known that conventional OCR systems are not the best option for reading text in natural scenes, mainly because in many cases it is extremely difficult to do a proper pixel-level segmentation of the characters, approaching the kind of challenge exploited by CAPTCHA tests. However, as our region based method works at pixel level, we are actually able to produce such a pixel-level segmentation. In this sense what we do is to build a binary image for each character candidate, additionally rotating the regions along the text line orientation if needed, and then send this binarized version of the region to the OCR for recognition. This simple process corresponds to the text extraction ability of our method. In other words we are not only detecting text but we can also perform an pixel-level segmentation of it. Even though in most of the cases this simple segmentation is not enough for obtaining a good recognition (for reasons external to our algorithm, e.g. low-resolution or blurring of the original image) it is a first step for further development in a more complex extraction process able to deal with degraded characters. Figure 4.14 shows two examples of the OCR outcomes, one where the text segmentation is easy and one where it is not; it can be seen that in the second case the results, even though they are not perfect, are intuitively much better than the ones obtained by sending the original image directly to the OCR system. Another possibility for future work is to build our own character classifier taking advantage of the pixel level performance of the proposed method in a similar way as in Neumann et al. do [11] [23].

4.5 Results

The proposed method has been extensively evaluated on the ICDAR2003 Robust Reading Dataset in different experiments. In this dataset, there are 258 images for training and 251 for testing, with

[2] `http://code.google.com/p/tesseract-ocr/`

varying resolutions from 640 × 480 to 1280 × 960. The evaluation schema [27] defines the precision and recall for a given image as:

$$p = \frac{\sum_{r_e \in E} m(r_e, T)}{|E|}$$

$$r = \frac{\sum_{r_t \in T} m(r_t, E)}{|T|}$$

where T is the ground-truth set, E are the estimated rectangles, and $m(r, R)$ is the best match for a rectangle r in a set of rectangles R, $m(r, T) = max(m_p(r, r')) \mid r' \in R$, where the match m_p between two rectangles is the area of intersection divided by the area of the minimum bounding box containing both rectangles.

Table 4.2 shows different results obtained by using different feature-sets for the construction of the clustering ensemble, being the best f-score obtained, $f = 0.55$, with the combination of all feature spaces proposed in Section 4.2.

In Table 4.3 we can compare the results obtained using the Evidence Accumulation Clustering with the best scored feature combinations and with the same features but performing clustering in a single feature space without using EAC to combine clusterings over different feature sub-spaces. The results obtained confirm the hypothesis that a clustering ensemble method is able to adapt better to the variability in similarity measures for text grouping.

Table 4.4 compares the results obtained on the ICDAR2003 dataset using two different criteria for deciding the final clusters in the Evidence Accumulation Clustering and demonstrates that detecting maximal meaningful clusters in the co-occurrence matrix performs better than using the maximal lifetime criterion.

Finally Table 4.5 shows a comparison of the obtained results with other methods on the ICDAR2003 dataset. We have evaluated three variants of our method using three different ground-truth sets: the first one, which is the standard ground-truth for the given dataset, aims for the task of text localization at word level, while the second and third ones do it at sentence level and block (paragrph) level respectively. Our intention with this triple evaluation is to be able to extract conclusions about the performance of our method at different stages of its pipeline. For instance, the performance of the method at the block level concern about how good is our Perceptual Organization Clustering step on grouping text characters which belongs to same paragraph together, without any post-processing. The results for this experiment achieve a 0.71 recall value, while performance is obviously near to zero because all meningful clusters found are treated as text blocks. On another hand results obtained at sentence level concern about the performance of the whole method but without the word split step, thus provinding important information on how this single step performs. For the sentence level evaluation we have joined

Feature sub-sets (see id. in Table 4.1)	Precision	Recall	f-score
#1	0.58	0.47	0.52
#2	0.53	0.43	0.47
#3	0.51	0.41	0.45
#4	0.47	0.39	0.43
#5	0.54	0.43	0.48
#6	0.54	0.45	0.49
#7	0.54	0.45	0.49
#1 + #2	0.57	0.47	0.52
#1 + #5	0.57	0.47	0.52
#1 + #6	0.56	0.46	0.51
#1 + #7	0.57	0.47	0.52
#2 + #3	0.52	0.42	0.46
#5 + #6	0.54	0.45	0.49
#1 + #2 + #7	0.55	0.46	0.50
#1 + #3 + #5	0.54	0.43	0.48
#1 + #5 + #6	0.58	0.47	0.52
#1 + #5 + #7	0.55	0.47	0.51
#1 + #5 + #3 + #7	0.57	0.47	0.52
#1 + #2 + #5 + #6	0.57	0.46	0.51
#1 + #3 + #4 + #5 + #6	0.57	0.47	0.52
#1 + #3 + #5 + #6 + #7	**0.60**	0.49	0.54
#1 + #2 + #3 + #5 + #6 + #7	**0.60**	**0.50**	**0.55**
#1 + #2 + #4 + #5 + #6 + #7	0.59	0.49	0.53

Table 4.2: ICDAR2003 dataset performance evaluation different feature-sets for the construction of the clustering ensemble.

the words bounding boxes of the complete ICDAR test dataset into sentences, while for the block level we evaluate with a small subset of 25 images.

The final results of the proposed method both at word and sentence levels are quite far from the best scores of state-of-the-art methods but still sufficient for taking the proposed methodology into account for further development. A plentiful collection of qualitative results for the proposed method at the task of word localization is provided in the Appendix C. In the next Chapter some conclusions are presented in the light of these results and possible lines of future work are proposed.

Feature sub-sets (see Table 4.1)	EAC	Precision	Recall	f-score
#1 + #2 + #3 + #5 + #6 + #7	Yes	**0.60**	**0.50**	**0.55**
#1 + #2 + #3 + #5 + #6 + #7	No	0.55	0.45	0.50

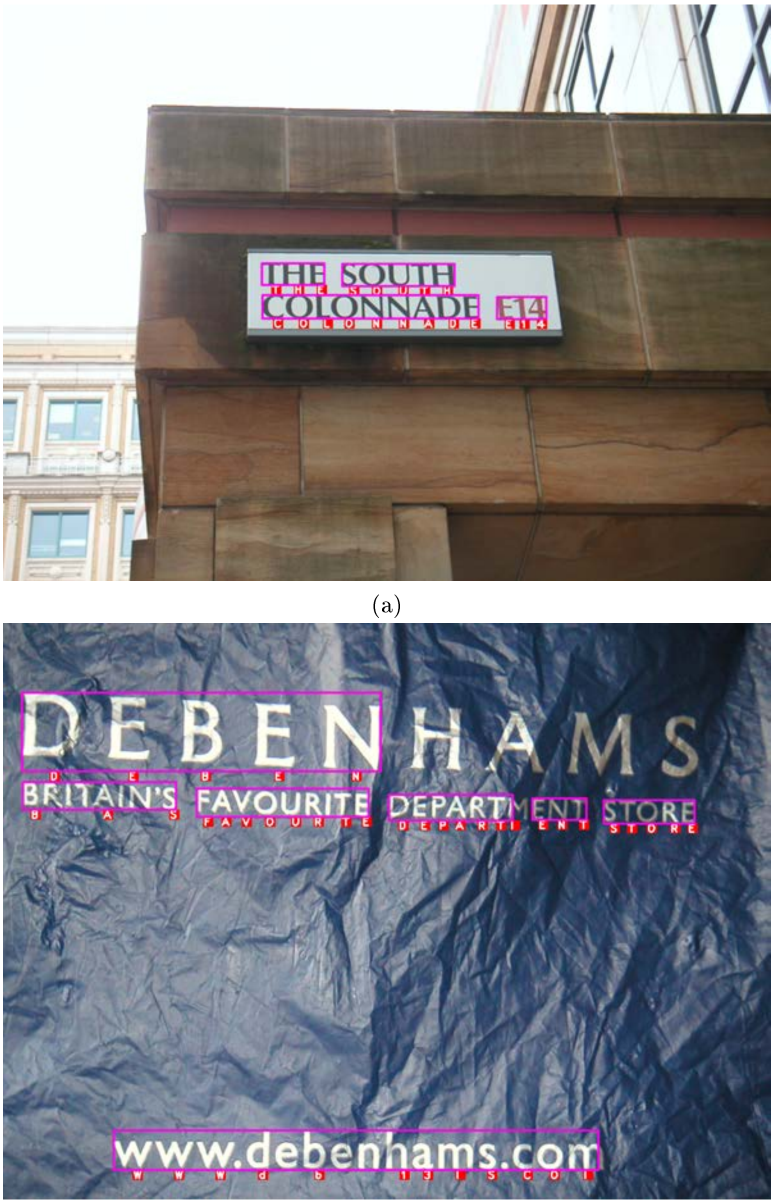
Table 4.3: ICDAR2003 dataset performance evaluation using Evidence Accumulation Clustering versus a single clustering analysis over a combined feature space.

Feature sub-sets (see Table 4.1)	EAC criterion	Precision	Recall	f-score
#1 + #2 + #3 + #5 + #6 + #7	max. meaningful clusters	**0.60**	**0.50**	**0.55**
#1 + #2 + #3 + #5 + #6 + #7	max. lifetime	0.56	0.46	0.51

Table 4.4: ICDAR2003 dataset performance evaluation using different criteria for deciding the final clusters in Evidence Accumulation Clustering.

Method	Precision	Recall	f-score	time(s)
Lee et al. [19]	0.66	0.75	0.70	n/a
Pan et al. [22] *	0.67	0.71	0.69	2.43
Chen et al. [15]	0.73	0.60	0.66	0.2
Epshtein et al. [12]	0.73	0.60	0.66	0.94
Becker (2005 winner)	0.62	0.67	0.64	14.4
Our Method *	**0.66**	**0.58**	**0.62**	**2.5**
Neumann and Matas [11]	0.59	0.55	0.57	n/a
Our Method	**0.60**	**0.50**	**0.55**	**2.61**
Merino et al. [16] *	0.51	0.67	0.55	0.2
Ashida (2003 winner) [27]	0.55	0.45	0.50	8.7
ICDAR 2003 average [27]	0.39	0.46	0.39	5.3
ICDAR 2005 average [28]	0.32	0.32	0.31	4.25
Full	0.10	0.06	0.08	

Table 4.5: ICDAR2003 dataset performance comparison. Methods marked with an asterisk evaluate at sentence level, while the others do it at word level.



(a)

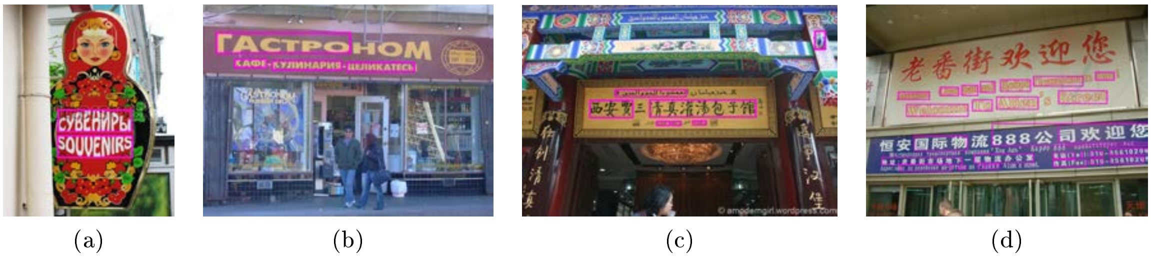
(b)

Figure 4.14: Examples of the OCR results for two images of the ICDAR2003 Dataset: (a) good recognition is achieved in images where the text segmentation is easy, (b) in images where text segmentation is more difficult the recognition results obtained using our extraction method, even though not perfect, are intuitively better than sending the original image to the OCR system.

Chapter 5

Conclusions and Future Work

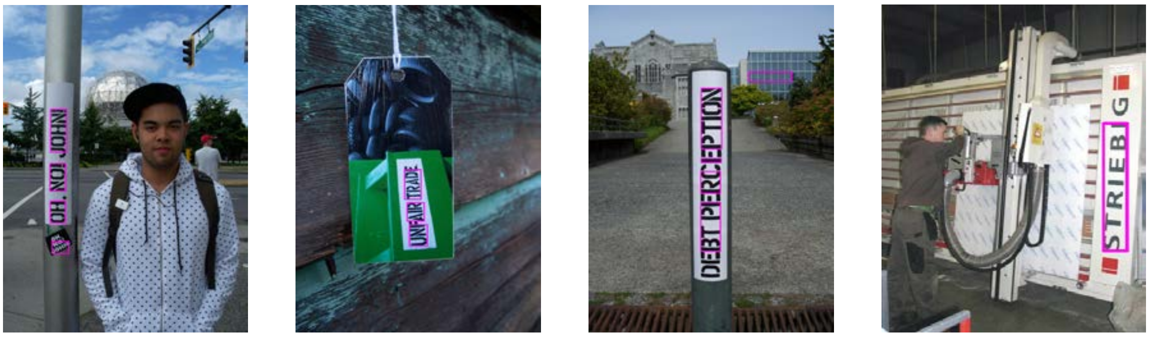
A new method for text extraction in natural scenes inspired by the Perceptual Organization Theory has been presented in this Thesis. The method combines two different clustering techniques, namely an *"a contrario"* clustering method for detecting meaningful clusters proposed by Desolneux et al. [6] [7] and the Evidence Accumulation Clustering ensemble method proposed by Fred et al. [69], in a single clustering step that we have called Perceptual Organization Clustering. We have placed this perceptual clustering procedure at the heart of the text detection task, being this the main contribution of this Thesis, motivated by the fact that text is always constructed in a way that exploits the Gestalt laws of perceptual organization.



(a) (b) (c) (d)

Figure 5.1: The proposed Perceptual Organization Clustering is able to detect text in any languages and scripts.

It is important to highlight that the Perceptual Organization Clustering is totally independent of the language or script in which text appears, can deal with any kind of font types and sizes, and is not constrained to horizontally aligned text, see Figures 5.1 and 5.2 for some examples. This interclass invariance itself is an advantage of the proposed method over others where the detection algorithm is trained for a given language or script. Another important property of the method is that the perceptual clustering stage is fully parameter-free, not using any kind of heuristic or parameter to detect meaningful text-like structures. Moreover, the proposed method works at pixel level and therefore allows text to be extracted in a direct way enabling the interfacing with an OCR module.



(a) (b) (c) (d)

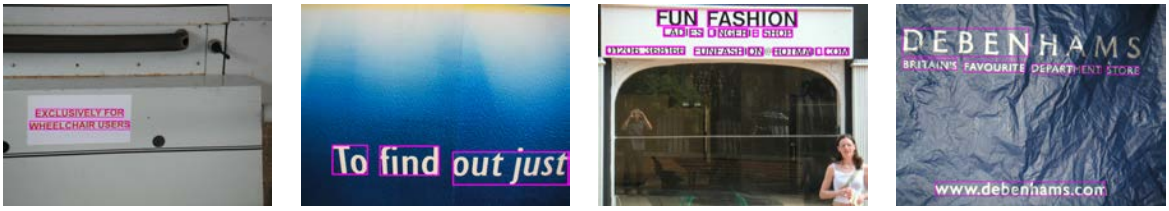
Figure 5.2: The proposed method is not restricted to horizontally aligned text.

The experiments performed show some interesting conclusions in the context of perceptual grouping of text components. For example, the results in Table 4.2 demonstrate that in this clustering approach there is no single best feature for grouping text characters and thus the combination of several similarity features increases the method's performance. On another hand, a similar conclusion arises from the results shown in Table 4.3, demonstrating that a clustering ensemble method combining the outcomes of several clustering analysis using different similarity features performs better than a single clustering analysis in a multidimensional feature space, thus better modelling this flexible definition of "similarity" between characters. The same idea can be qualitatively evaluated in the examples shown in Figure 4.7.

The proposed method has been evaluated on the ICDAR2003 dataset in order to compare its performance with other approaches and, despite the results obtained being still far from the top-scored state-of-the-art methods, it has shown sufficient success to warrant further development. An analysis of the qualitative results of the Perceptual Organization Clustering, e.g. in Figures 4.7 and 4.8, reveals high recall rates in the detection of text structures, although this recall rate has not been quantitatively evaluated mainly because of the absence of a reference benchmarking dataset at this abstract "text structures" grouping level. On the contrary, the seemingly low recall obtained with the ICDAR 2003 evaluation benchmark can be explained by a lack of precision in the text line formation and word separation post-processing steps.

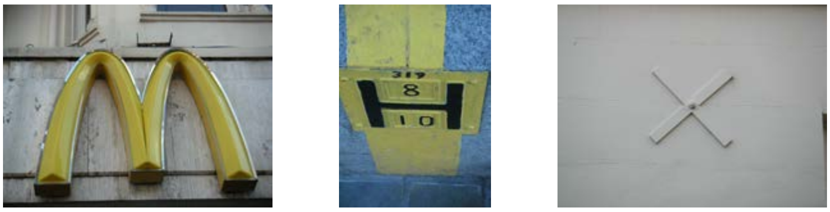
At this point it is important to note that a low performance score under the evaluation method used in the ICDAR2003 Dataset does not necessarily mean that our method is bad at finding text. The evaluation framework of ICDAR 2003 explicitly targets word detection. Hence it could be the case that the algorithm detects all the text in a scene but is not able to separate correctly the words in the text cluster as shown in the examples of Figure 5.3. The evaluation score for this behaviour is very low because the main objective in the ICDAR2003 competition is precisely to locate the individual words.

For this reason some authors [16] [22] opt to evaluate their methods at the sentence level, building an alternative ground truth for the ICDAR2003 Dataset, while many others have suggested the use of an evaluation scheme allowing for imprecise matches and dealing with the correspondence problem (one-to-many, etc.). Moreover, another difficulty of the ICDAR2003 evaluation scheme is the inability to perfectly match the estimated bounding boxes with the manually annotated ones in the ground truth, thus the obtained f-score can vary from 0.8 to 1 even when all the words are correctly detected. For example, in our case adding a small amount of space at the borders of the estimated bounding boxes increases the final f-score by 2%. All this together with the presence of very hard text to be located and single characters annotated as text that by definition are not detectable by a grouping algorithm (see Figure 5.4) makes the ICDAR2003 Dataset a somewhat inadequate and non-realistic testbed for our method.



(a) (b) (c) (d)

Figure 5.3: A bad word separation can cause poor performance under the ICDAR2003 evaluation scheme even when all the text in the scene is detected: (a) $f = 0.48$, (b) $f = 0.68$, (a) $f = 0.56$, (d) $f = 0.57$.



(a) (b) (c)

Figure 5.4: The ICDAR2003 dataset has some single characters annotated as text which can not be detected by the grouping based proposed method.

In any case, being the Perceptual Organization Clustering step the main contribution of this Thesis and in the light of its good qualitative results it seems clear that more elaborated pre-processing and post-processing steps will probably enhance the obtained results for the proposed method, thus driving two different lines for future developments:

In the pre-processing stage a better pruning method over the MSER tree, e.g. selecting the regions with higher border energy as proposed by Merino et al. [16], using edge enhanced regions as proposed by [15], or directly classifying regions as character or non-character in a pre-filtering stage as in [23], seems adequate in order to increase the recall of the whole method.

For the post-processing, using a texture based classifier at the cluster level, in order to filter out those clusters which despite being meaningful do not correspond to text groups, seems promising in order to increase the method's performance. Moreover, other approaches for the recognition stage better suited for natural scene images, like the approximate nearest neighours methodology proposed by Iwamura et al. [80] [81], could also help improving the results obtained by simply passing the extracted regions to a conventional OCR system. In the line of some end-to-end state-of-the-art methods [10] [23], better recognition would permit a more interactive feedback between the recognition and grouping processes, allowing the restoration of broken groups, e.g. searching for lost parts of a text line or word, and at the same time favoring better text line formation and word separation, making use of lexicon grammar models as in [21] [10] [23].

The Perceptual Organization Clustering itself can also be improved in several ways, being the most natural to increase the number of features used for the construction of the clustering ensemble, in order to enrich our flexible definition of similarity between characters. Another possible improvement is to use weights in the evidence accumulation contribution of the initial clusterings, trying to capture which ones of the features used are more significant in describing text.

Regarding computational time complexity, we have seen in Table 4.5 that the proposed method achieves an acceptable rate on the ICDAR2003 dataset. Besides, although the 2.6 seconds average obtained might suggest that the algorithm is not usable in a real time system, it is important to notice that in our case time complexity is not really a function on the size of the image but on the number of detected regions. We have made experiments with 640 × 480 video streams where our method is able to achieve between 7 and 10 frames per second when the scenario is not excessively cluttered. This behavior leads to new paths to explore in terms of optimization, for example with the use of approximate calculations for the meaningful clusters detection, or with a adaptable filtering module for the initial regions set.

Finally, another promising future line of development could emerge from the integration of the Perceptual Organization method proposed here with other fields of Human Perception inspired Computer Vision. For instance, an Active Vision system driven by the input of Eye-Tracker technology as proposed by Kobayashi at al. [82] and Gröger [83] will enable better and faster performance suitable for real-time video processing. On another hand, saliency maps [84], modeling the way humans select locations of interest in an image, could also enhance the proposed method in a similar way, but at the same time our Perceptual Organization Clustering method for text detection can serve as contextual input for saliency computational models [85] [86] [87] [88].

Appendices

Appendix A

Hierarchical Clustering and Graph Theoretical Methods for Gestalt Detection

A.1 Hierarchical Clustering

In order to see how Hierarchical Clustering can be used for gestalt getection, and at the same time discuss on how the the different possible linkage criteria affect the clustirng results, lets consider a simple example with a 2D points dataset:

$$
\begin{aligned}
X = \{&(2,2),(2,3),(2,4),(2,5.5),(2,6.5),(2,7.5),(4,2),(4,3),(4,4),(4,5.5),(4,6.5),(4,7.5),\\
&(6,2),(6,3),(6,4),(6,5.5),(6,6.5),(6,7.5),(8,2),(8,3),(8,4),(8,5.5),(8,6.5),(8,7.5)\}
\end{aligned}
$$

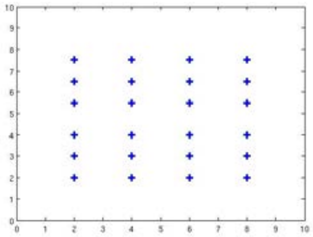
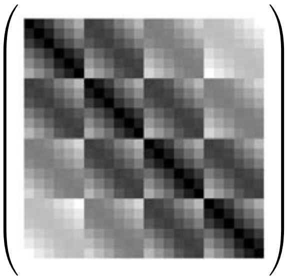


Figure A.1: A simple two dimensional point set to be analyzed with Hierarchical Clustering.

Our metric measure for clustering is the Euclidean distance, hence our distance matrix represented as a heat map, where white color represents the maximum distance in our dataset (8.1394), is:

$$D_{i,j} =$$

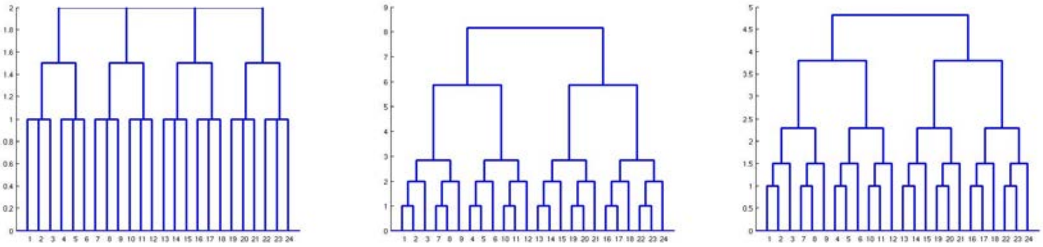


To apply hierarchical clustering on X in an bottom up approach, i.e. agglomerative clustering, each observation x_i starts in its own cluster and then pairs of clusters are merged iteratively until all observations are in the same cluster. The different linkage strategies differ in the criteria used to choose the two different clusters A and B that are going to be merged at each step of the process:

- Single Linkage Clustering
$$min\{d(a,b) : a \in A, b \in B\}$$
- Complete Linkage Clustering
$$max\{d(a,b) : a \in A, b \in B\}$$
- Average Linkage Clustering
$$\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$$

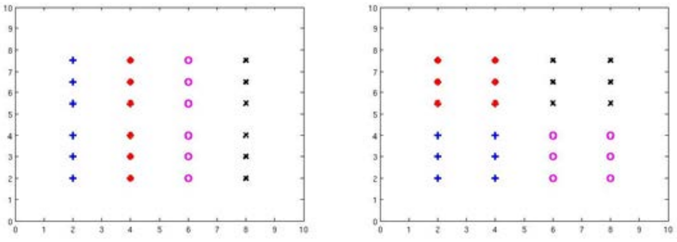
Other linkage criteria include the Minimum Energy Clustering and Ward's Minimum Variance Method, where at each step the criterion for choosing the pair of clusters to merge is based on minimizing the value of an objective function, e.g. the total within-cluster variance.

Applying SLC, CLC ans ALC to the two-dimensional point dataset in Figure A.1 results in the dendrograms shown in Figure A.2, where the main difference is that Single Linkage is emphasizing connectedness of patterns in clusters while Complete Linkage and Average Linkage emphasize compactness (See in Figure A.3 the resulting partitions for k=4). Thus, if we focus on the "factor of nearness" to make clusters of connected components then it is clear that what we want to use is Single Linkage Clustering, however if we want to focus on the "factor of similarity" the choice is again not so obvious. For example if we use color mean information as a "similarity" measure then it could make sense to use Complete or Average Linkage Clustering.



(a) (b) (c)

Figure A.2: Dendrogram representation of SLC, CLC, and ALC hierarchical clustering over the point set in Figure A.1: (a) Single Linkage Clustering, (b) Complete Linkage Clustering, (c) Average Linkage Clustering.



(a) (b)

Figure A.3: Partitions for *k=4* using: (a) Single Linkage Clustering, (b) Complete Linkage Clustering.

In Figure A.3 the resulting partition for k=4 is shown, but different k values could have been chosen as well. As pointed in Section 3.1.1 the final partition can also be obtained using the maximum lifetime criterion.

A.2 Graph Theoretical Methods

Minimum Spanning Tree

An edge-weighted graph $\mathcal{G}$ is a 3-tuple $\mathcal{G} = (V, E, \omega)$, where:

- V is the finite set of nodes
- $E \subseteq V \times V$ is the set of edges
- $\omega : E \to R$ is the edge weighting function.

A tree is a connected acyclic graph, and a spanning tree of a connected graph $\mathcal{G}$ is a tree in $\mathcal{G}$ containing all the nodes V in $\mathcal{G}$. I f we define the weight of a tree as the sum of weights of its edges, then the Minimal Spanning Tree (MST) in $\mathcal{G}$ is the spanning tree in $\mathcal{G}$ with the minimum weight among all the possible spanning trees in $\mathcal{G}$.

The solution of the MST is deterministic but there may be several solutions if the graph contain edges with the same weight. Two common MST algorithms are Prim's and Kruskal's, both with $\mathcal{O}(n^2)$ complexity. There are more complex algorithms to obtain the MST in quasi-linear time in some concrete situations, for example in 2D using Delaunay triangulation results in an algorithm with $\mathcal{O}(n \cdot log(n))$ complexity.

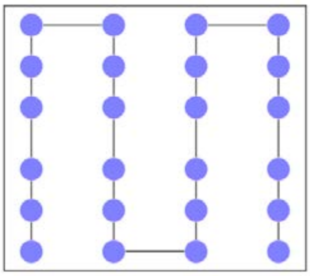
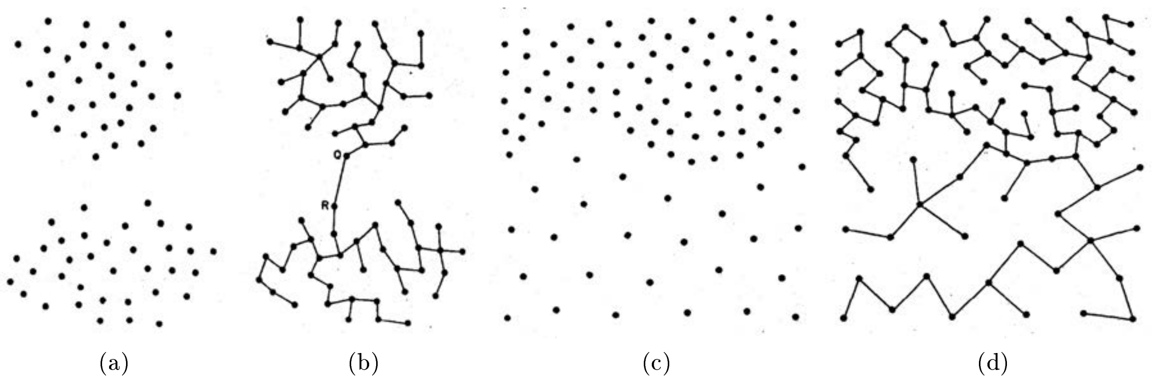


Figure A.4: A Minimum Spanning Tree of the two-dimensional point dataset in Figure A.1.

Once we have computed the MST of our data, the clustering procedure consists in to identify and remove those edges bridging separate clusters. Thus, one needs a criterion to decide which of the edges should be deleted. Zahn [47], for example, proposes a criterion for this type of two-dimensional perceptually observable clustering consisting in deleting those edges whose weight is significantly larger than the average of near edges. Different criteria should be defined depending on the problem to solve, for example the touching clusters problem or the density gradient problem (see Figure A.5) needs a more elaborated deletion criteria to be solved using an MST approach [47]. All this brings us again to the same point as in the previous section: some heuristic threshold or rules should be defined in order to exploit Graph Theoretical Methods or Hierarchical Clustering for Gestalt cluster detection.



(a) (b) (c) (d)

Figure A.5: The touching clusters problem (a), its MST (b), the density gradient problem (c), and its MST (d). This images are reproduced from [47].

Appendix B

MSER Definition

Let's consider a gray level image I as a mapping $I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow S$, where usually $S = \{1, 2, ..., 255\}$, and an adjacency relation $A : \mathcal{D} \times \mathcal{D}$ between pixels defined by:

$$pAq \iff \sum_{i=1}^{d} |p_i - q_i| \leq 1$$

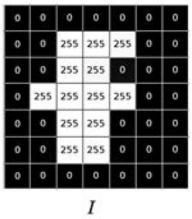
A **Region** $\mathcal{Q}$ is a contigous subset of $\mathcal{D}$:

$$\forall p, q \in \mathcal{Q}, \exists \{p, a_1, a_2, .., a_n, q\} : \{pAa_1, a_iAa_{i+1}, a_nAq\}$$

The **(Outer) Region Boundary** $\partial\mathcal{Q}$ of a region $\mathcal{Q}$ is the set of pixels adjacent to at least one pixel of $\mathcal{Q}$ but not belonging to $\mathcal{Q}$:

$$\partial\mathcal{Q} = \{q \in \mathcal{D} \setminus \mathcal{Q} : \exists p \in \mathcal{Q} : qAp\}$$

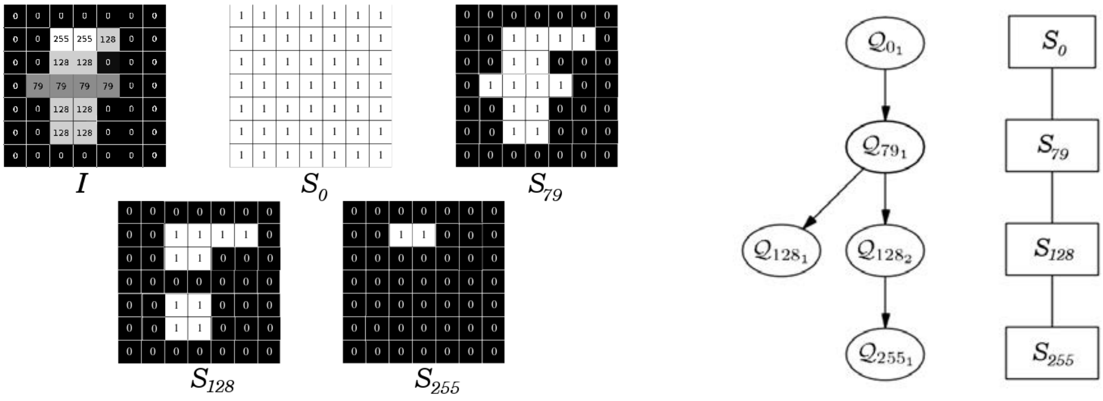
An **Extremal Region** $\mathcal{Q} \subset \mathcal{D}$ is a region such that either $\forall p \in \mathcal{Q}, q \in \partial\mathcal{Q} : I(p) > I(q)$ (maximum intensity region), or $\forall p \in \mathcal{Q}, q \in \partial\mathcal{Q} : I(p) < I(q)$ (minimum intensity region). This duality leads to the definition of MSER+, bright regions with darker boundary, and MSER-, dark regions with lighter boundaries. Note that MSER- can be obtained by detecting MSER+ in the intensity inverted image.



0	0	0	0	0	0	0
0	0	255	255	255	0	0
0	0	255	255	0	0	0
0	255	255	255	255	0	0
0	0	255	255	0	0	0
0	0	255	255	0	0	0
0	0	0	0	0	0	0

Figure B.1: A maximum intensity region: a contigous subset of pixels with a larger intensity value than the pixels in its outer boundary.

Thus, an **Extremal Region** is a connected component of the level set $S_l = \{p : I(p) \leq l\}, l \in S = \{1, 2, ..., 255\}$. At each intensity level l we have multiple disjoint **Extremal Regions** in the level set S_l, and we can construct the **Extremal Regions Tree** by connecting regions $\mathcal{Q}_l \in S_l$ and $\mathcal{Q}_{l+1} \in S_{l+1}$ if and only if $\mathcal{Q}_l \subset \mathcal{Q}_{l+1}$ (See Figure B.2).



(a) (b)

Figure B.2: Building the component tree of an image: (a) grayscale image I and its disjoint Extremal Regions at each possible level set S_l, (b) the component tree.

Finally, an **Extremal Region** is a **Maximally Stable Extremal Region** $\mathcal{Q}_{i^*}$ if it is a local minimum of the area grow function in a sequence of nested extremal regions $Q_1, .., Q_{i-1}, Q_i : (Q_i \subset Q_{i+1})$. Where the grow function $q(Q_i)$ stablishes the stability criterion:

$$q(Q_i) = \frac{|Q_{i+\Delta} \setminus Q_{i-\Delta}|}{|Q_i|}$$

Where $|\cdot|$ denotes cardinality of the region and $\Delta \in S$ is a cut-off parameter of the method. $q(Q_i)$ values are low for a given region if the regions along the sequence, bounded by the cut-off parameter, have similar area. So, selecting a local minimum of the area grow function is just the same as selecting regions with a high shape stability along a branch of the Extremal Region Tree. Notice that the larger Δ value the highest the contrast between detected MSER and its boundaries.

The MSER implementation used in our method, provided by the vlfeat[1] library by Andrea Vedaldi and Brian Fulkerson, refine the obtained MSER tree by running a set of filtering tests, walking from the bigger to the smaller regions:

- $a_- \leq |Q_l|/|Q_\infty| \leq a_+$: exclude MSERs too small or too big ($|Q_\infty|$ is the area of the image).

[1] `http://www.vlfeat.org/`

- $q(Q_l) < q_+$: exclude MSERs too unstable.
- For any MSER Q_l, find the parent MSER $Q_{l'}$ and check if $|Q_{l'} - Q_l|/|Q_{l'}| < d_+$: remove duplicated MSERs.

Appendix C

Qualitative Results

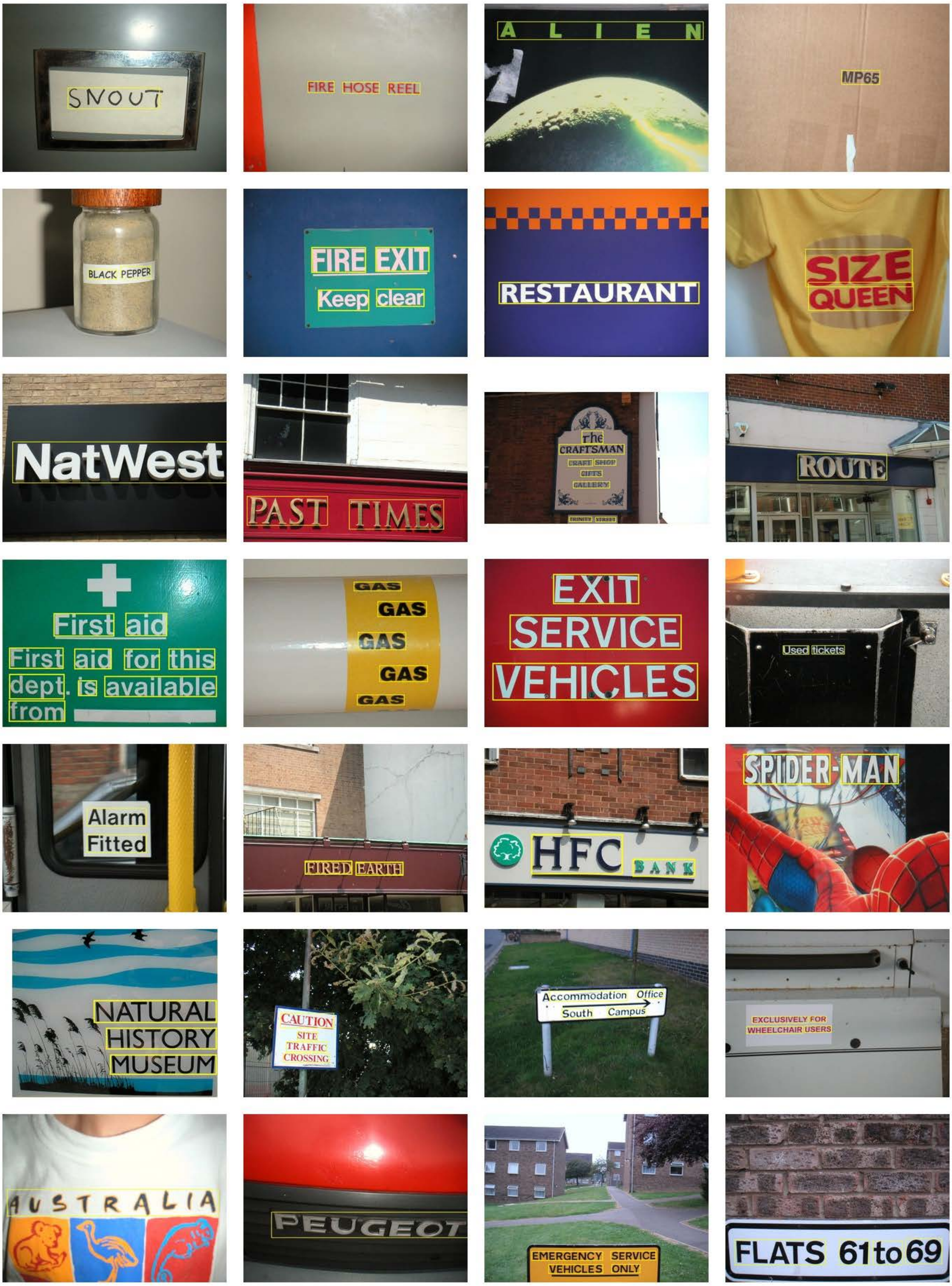


Figure C.1: Qualitative results with $f \geq 0.9$ on the ICDAR2003 dataset.

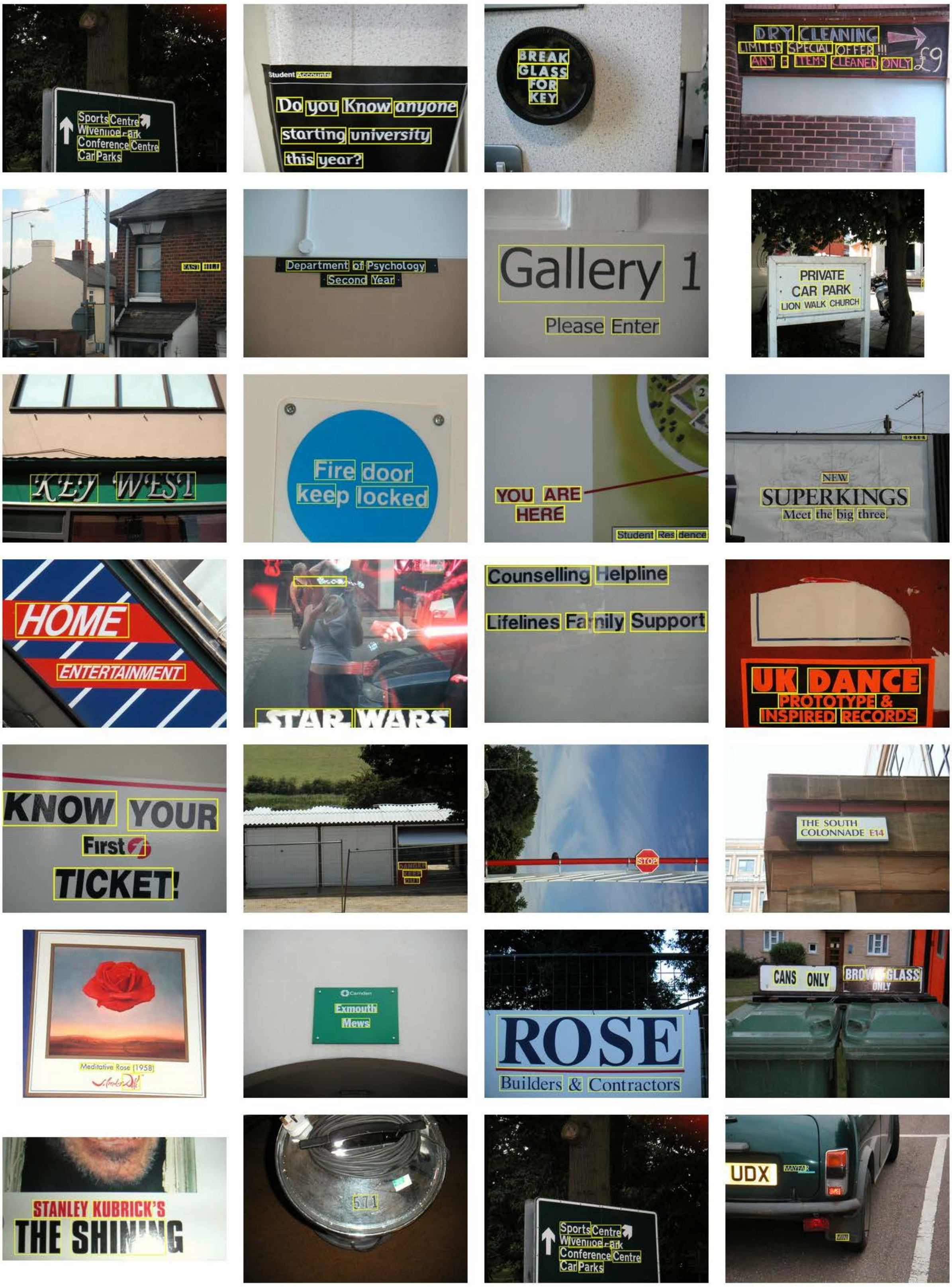


Figure C.2: Qualitative results with $0.9 > f \geq 0.75$ on the ICDAR2003 dataset.

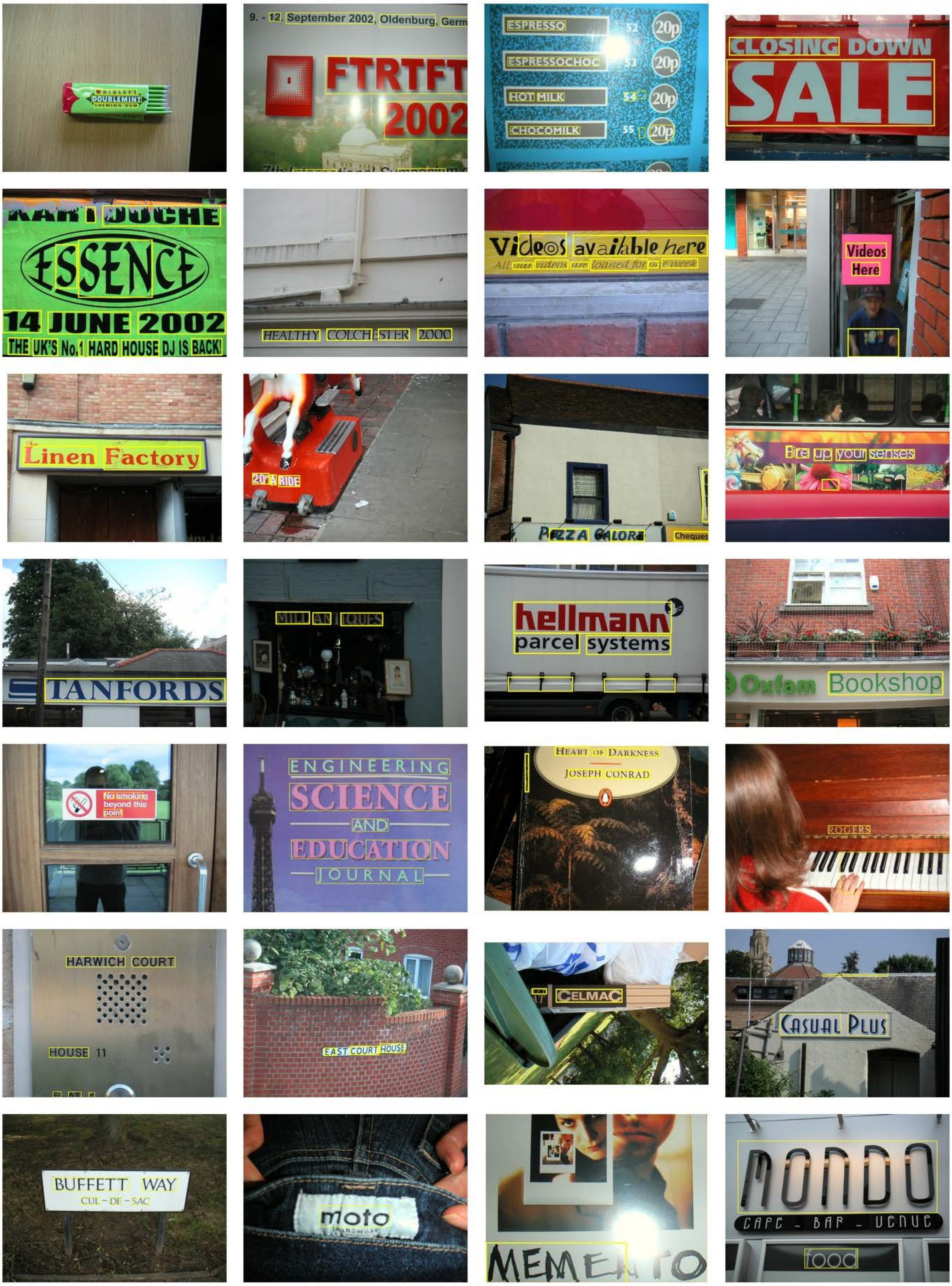


Figure C.3: Qualitative results with $0.75 > f \geq 0.5$ on the ICDAR2003 dataset.

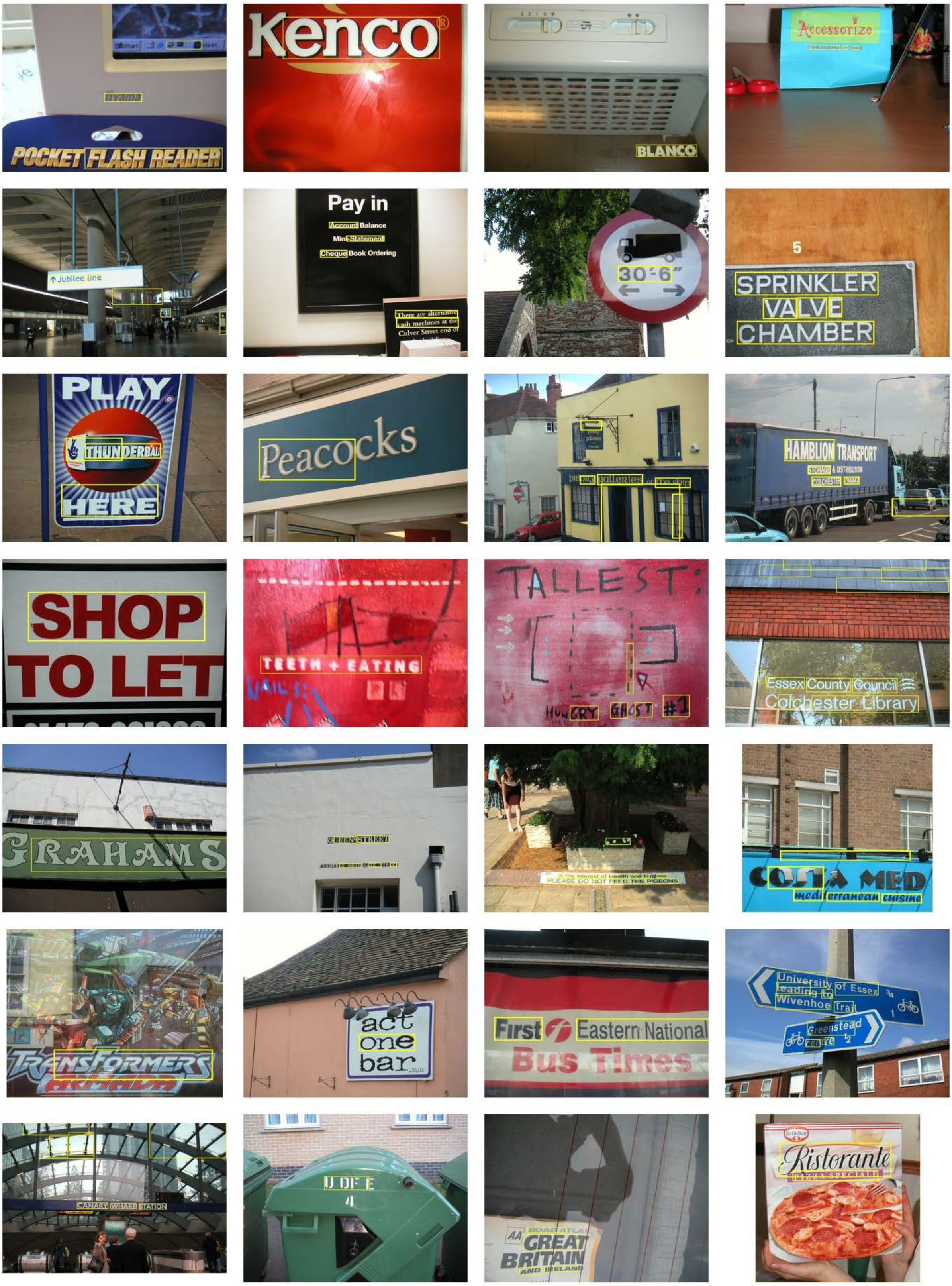


Figure C.4: Qualitative results with $0.5 > f \geq 0.25$ on the ICDAR2003 dataset.

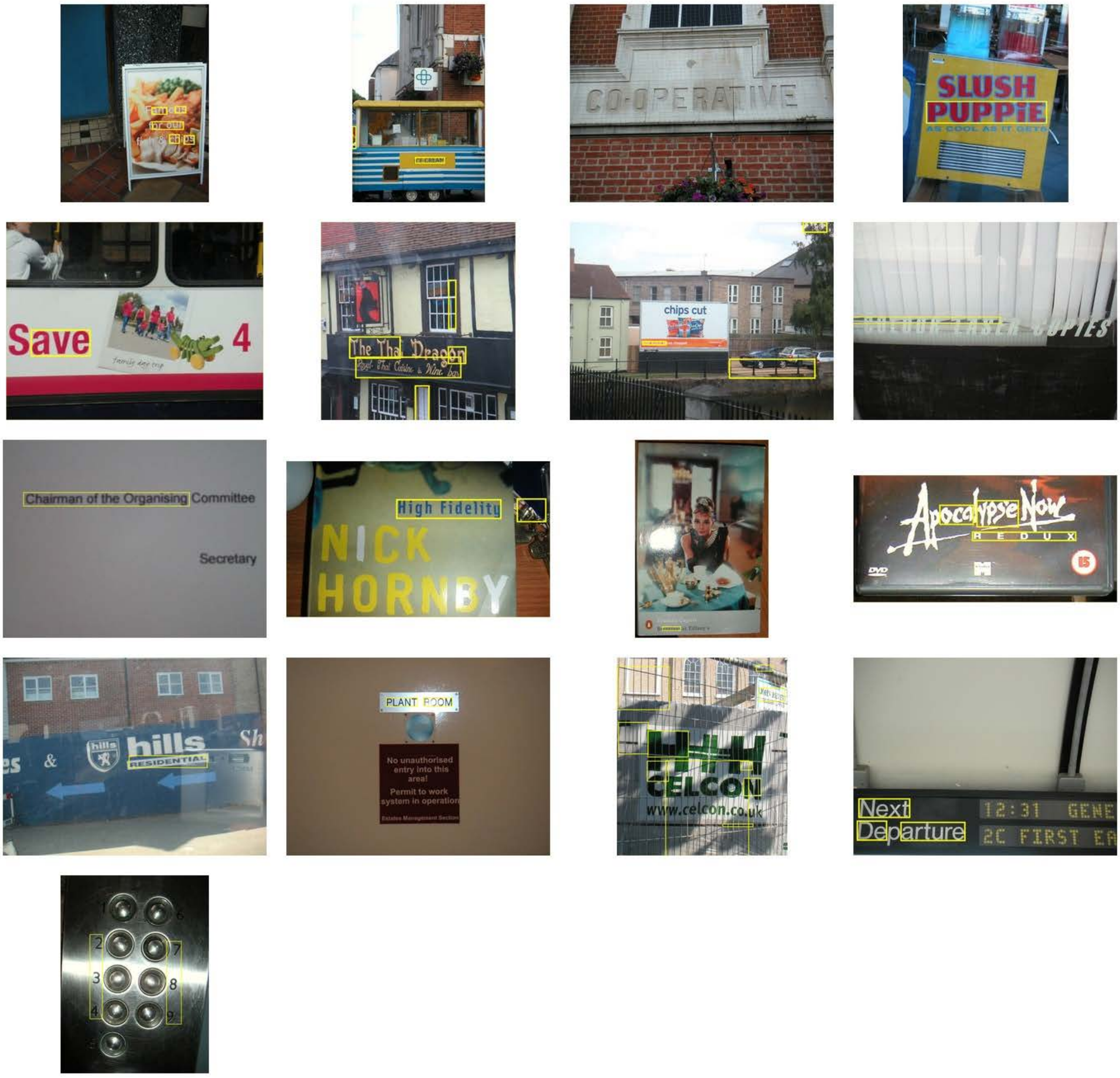


Figure C.5: Qualitative results with $0.25 > f \geq 0.1$ on the ICDAR2003 dataset.

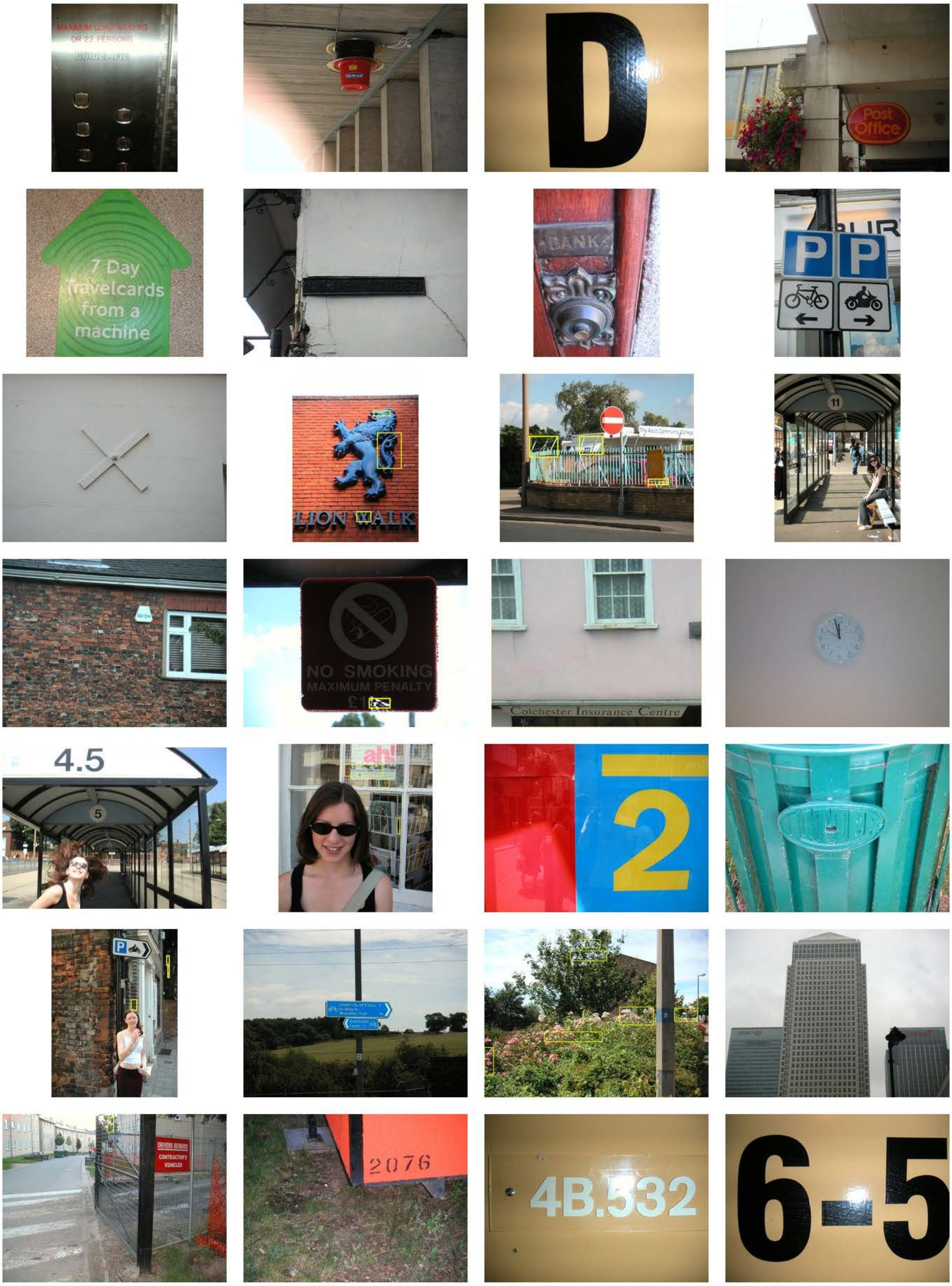


Figure C.6: Qualitative results with $f < 0.1$ on the ICDAR2003 dataset.

Bibliography

[1] K. Nation, “Form-meaning links in the development of visual word recognition,” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 364, no. 1536, p. 3665–74, December 2009.

[2] M. Wertheimer, “Untersuchungen zur lehre der gestalt, II,” *Psychologische Forschung*, vol. 4, pp. 301–350, 1923, translation published as "Laws of Organization in Perceptual Forms", in W.Ellis (ed.), "A Source Book of Gestalt Psychology", pp 71-88, 1938.

[3] D. G. Pelli, N. J. Majaj, N. Raizman, C. J. Christian, E. Kim, and M. C. Palomares, “Grouping in object recognition: The role of a gestalt law in letter identification,” *Cognitive Neuropsychology*, vol. 26, no. 1, pp. 36–49, 2009.

[4] Boder and S. Jarrico, *The Boder Test of Reading-Spelling Patterns: A Diagnostic Screening Test for Subtypes of Reading Disability*, N. Y. Grunne and Stratton, Eds., 1982.

[5] N. Bell, “Gestalt imagery: A critical factor in language comprehension,” *Annals of Dyslexia*, vol. 41, pp. 246–260, 1991.

[6] A. Desolneux, L. Moisan, and J.-M. Morel, “A grouping principle and four applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 508–513, 2003.

[7] A. Desolneux, L. Moisan, and J.-M. Morel, *From Gestalt Theory to Image Analysis: A Probabilistic Approach.* Springer-Verlag, collection "Interdisciplinary Applied Mathematics", 2008.

[8] K. Jung, K. I. Kim, and A. K. Jain, “Text information extraction in images and video: a survey,” *Pattern Recognition*, vol. 37, no. 5, pp. 977 – 997, 2004.

[9] J. Liang, D. Doermann, and H. Li, “Camera-based analysis of text and documents: a survey,” *International Journal on Document Analysis and Recognition*, vol. 7, no. 2, pp. 84–104–104, Jul. 2005.

[10] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” *Computer Vision, IEEE International Conference on*, vol. 0, pp. 1457–1464, 2011.

[11] L. Neumann and J. Matas, “A method for text localization and recognition in real-world images,” in *Computer Vision – ACCV 2010*, ser. Lecture Notes in Computer Science, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Springer Berlin / Heidelberg, 2011, vol. 6494, pp. 770–783.

[12] B. Epshtein, E. Ofek, and Y. Wexler, “Detecting text in natural scenes with stroke width transform,” *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 2963–2970, 2010.

[13] C. Jung, Q. Liu, and J. Kim, “A new approach for text segmentation using a stroke filter,” *Signal Processing*, vol. 88, no. 7, pp. 1907–1916, 2008.

[14] L.-J. Li, J. Li, and L. Wang, “An integration text extraction approach in video frame,” *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, vol. 4, pp. 2115–2120, 2010.

[15] H. Chen, S. Tsai, G. Schroth, D. Chen, R. Grzeszczuk, and B. Girod, “Robust text detection in natural images with edge-enhanced maximally stable extremal regions,” *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 2609 – 2612, September 2011.

[16] C. Merino-Gracia, K. Lenc, and M. Mirmehdi, “A head-mounted device for recognizing text in natural scenes,” *Proc. of Int. Workshop on Camera-based Document Analysis and Recognition*, pp. 27–32, September 2011.

[17] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng, “Text detection and character recognition in scene images with unsupervised feature learning,” *Document Analysis and Recognition, International Conference on*, vol. 0, pp. 440–445, 2011.

[18] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[19] J.-J. Lee, P.-H. Lee, S.-W. Lee, A. Yuille, and C. Koch, “Adaboost for text detection in natural scene,” *Document Analysis and Recognition, International Conference on*, vol. 0, pp. 429–434, 2011.

[20] Y. Song, Y. He, Q. Li, and M. Li, *Reading text in street views using Adaboost: Towards a system for searching target places*, 2009, pp. 227–232.

[21] K. Wang and S. Belongie, “Word spotting in the wild,” in *Computer Vision – ECCV 2010*, ser. Lecture Notes in Computer Science, 2010, vol. 6311, pp. 591–604.

[22] Y.-F. Pan, X. Hou, and C.-L. Liu, “Text localization in natural scene images based on conditional random field,” *Document Analysis and Recognition, International Conference on*, vol. 0, pp. 6–10, 2009.

[23] L. Neumann and J. Matas, “Real-time scene text localization and recognition,” *Computer Vision and Pattern Recognition (CVPR) 2012, 25th IEEE Conference on*, 2012.

[24] J. Park and G. Lee, "A robust algorithm for text region detection in natural scene images," *Electrical and Computer Engineering, Canadian Journal of*, vol. 33, pp. 215–222, 2008.

[25] M. Anthimopoulos, B. Gatos, and I. Pratikakis, "A hybrid system for text detection in video frames," *Document Analysis Systems, IAPR International Workshop on*, vol. 0, pp. 286–292, 2008.

[26] Y. Kunishige, F. Yaokai, and S. Uchida, "Scenery character detection with environmental context," *Document Analysis and Recognition, International Conference on*, vol. 0, pp. 1049–1053, 2011.

[27] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J.-M. Jolion, L. Todoran, M. Worring, and X. Lin, "ICDAR 2003 robust reading competitions: entries, results, and future directions," *International Journal on Document Analysis and Recognition*, vol. 7, pp. 105–122, 2005, 10.1007/s10032-004-0134-3.

[28] S. Lucas, "ICDAR 2005 text locating competition results," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, aug.-1 sept. 2005, pp. 80 – 84 Vol. 1.

[29] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," *Document Analysis and Recognition, International Conference on*, vol. 0, pp. 1491–1496, 2011.

[30] R. Nagy, A. Dicker, and K. Meyer-Wegener, "NEOCR: A configurable dataset for natural image text recognition," in *Camera-Based Document Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Iwamura and F. Shafait, Eds. Springer Berlin Heidelberg, 2012, vol. 7139, pp. 150–163.

[31] M. S. C. Jehyun Jung, SeongHun Lee and J. H. Kim, "Touch TT: Scene text extractor using touch screen interface," *ETRI Journal*, 2011.

[32] K. J. SeongHun Lee, Min Su Cho and J. H. Kim, "Scene text extraction with edge constraint and text collinearity link," *20th International Conference on Pattern Recognition (ICPR)*, August 2010, istanbul, Turkey.

[33] W. Köhler, *Gestalt Psychology*, N. Y. Liveright, Ed., 1929.

[34] K. Koffka, *Principles of Gestalt Psychology*, B. New York: Harcourt, Ed., 1935.

[35] W. Metzger, *Laws of Seeing*, 2006, cambridge, MA: MIT Press. (Original work published in German in 1936).

[36] G. Kanizsa, *Grammatica del Vedere*, 1980, il Mulino, Bologna/Editions Diderot, Arts et Sciences, 1980 / 1997.

[37] D. Todorovic, "Gestalt principles," *Scholarpedia*, vol. 3, no. 12, pp. 53–45, 2008.

[38] S. Palmer, *Vision Science. Photons to Phenomenology.*, M. M. P. Cambridge, Ed., 1999.

[39] S. Palmer, “Common region: a new principle of perceptual grouping.” *Cognitive Psychology*, vol. 24, pp. 436–447, 1992.

[40] S. Palmer and I. Rock, “Rethinking perceptual organization: The role of uniform connectedness.” vol. 1, pp. 29–35, 1994.

[41] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information.*, S. F. Freeman, Ed., 1982.

[42] V. Bruce, P. Green, and M. Georgeson, *Visual perception: Physiology, psychology and ecology (3rd ed.)*, LEA., Ed., 1996.

[43] C. van Leeuwen, D. Alexander, C. Nakatani, A. Nikolaev, G. Plomp, and A. Raffone, “Gestalt has no notion of attention. but does it need one?” *Humana Mente*, vol. 17, pp. 35–68, 1994.

[44] R. Kimchi, “Perceptual organization and visual attention,” in *Attention*, ser. Progress in Brain Research, N. Srinivasan, Ed. Elsevier, 2009, vol. 176, pp. 15 – 33.

[45] E. Brunswik and J. Kamiya, “Ecological cue-validity of proximity and of other gestalt factors,” *The American journal of psychology*, vol. 66, pp. 20–32, 1953.

[46] I. Rock, *An Introduction to Perception.* New York: Macmillan, 1975.

[47] C. T. Zahn, “Graph-theoretical methods for detecting and describing gestalt clusters,” *IEEE Transactions on Computers*, vol. 20, pp. 68–86, 1971.

[48] N. Ahuja and M. Tuceryan, “Extraction of early perceptual structure in dot patterns: integrating region, boundary, and component gestalt,” *Comput. Vision Graph. Image Process.*, vol. 48, no. 3, pp. 304–356, 1989.

[49] A. Zobrist and W. Thompson, “Building a distance function for gestalt grouping,” *IEEE Transactions on Computers*, vol. 24, pp. 718–728, 1975.

[50] S. Santini and R. Jain, “Similarity measures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 871–883, 1999.

[51] R. N. Shepard, “Multidimensional scaling, tree-fitting, and clustering,” *Science*, vol. 210(4468), pp. 390–398, October 1980.

[52] S. Ullman and A. Sha’ashua, “Structural saliency: The detection of globally salient structures using a locally connected network,” Cambridge, MA, USA, Tech. Rep., 1988.

[53] G. Guy and G. Medioni, “Inferring global pereeptual contours from local features,” *International Journal of Computer Vision*, vol. 20, pp. 113–133, 1996.

[54] S.-C. Zhu, “Embedding gestalt laws in markov random fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 1170–1187, 1999.

[55] F. Cao, J. Delon, A. Desolneux, P. Musé, and F. Sur, "An a contrario approach to hierarchical clustering validity assessment," 2004.

[56] A. G. Arkadev and E. M. Braverman, *Computers and Pattern Recognition.* Washington, D. C.: Thompson Book Co., 1967.

[57] S. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, pp. 241–254, 1967.

[58] J. C. Gower and G. J. S. Ross, "Minimum spanning trees and single linkage cluster analysis," *Appl. Statistics*, vol. 18, pp. 54–64, 1969.

[59] A. Desolneux, L. Moisan, and J.-M. Morel, "Meaningful Alignments," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 7–23, Oct. 2000.

[60] A. Almansa, A. Desolneux, and S. Vamech, "Vanishing point detection without any a priori information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 4, pp. 502 – 507, april 2003.

[61] A. Desolneux, L. Moisan, and J.-M. Morel, "Edge detection by helmholtz principle," *Journal of Mathematical Imaging and Vision*, vol. 14, no. 3, pp. 271–284, May 2001.

[62] F. Cao, "Application of the gestalt principles to the detection of good continuations and corners in image level lines," *Comput. Vis. Sci.*, vol. 7, no. 1, pp. 3–13, Jun. 2004.

[63] F. Cao, P. Musé, and F. Sur, "Extracting meaningful curves from images," *Journal of Mathematical Imaging and Vision*, vol. 22, pp. 159–181, 2005.

[64] F. Cao, J. Delon, A. Desolneux, P. Musé, and F. Sur, "A unified framework for detecting groups and application to shape recognition," *J. Math. Imaging Vis.*, vol. 27, no. 2, pp. 91–119, Feb. 2007.

[65] F. Attneave, "Some informational aspects of visual perception," *Psychological Review*, vol. 61, pp. 183–193, 1954.

[66] D. G. Lowe, *Perceptual Organization and Visual Recognition.* Norwell, MA, USA: Kluwer Academic Publishers, 1985.

[67] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2003.

[68] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1866–1881, Dec. 2005.

[69] A. Fred and A. Jain, "Combining multiple clusterings using evidence accumulation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 6, pp. 835 – 850, jun 2005.

[70] A. Fred and A. Jain, "Data clustering using evidence accumulation," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4, 2002, pp. 276 – 280 vol.4.

[71] N. Alajlan, “Retrieval of hand-sketched envelopes in logo images,” in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Kamel and A. Campilho, Eds. Springer Berlin / Heidelberg, 2007, vol. 4633, pp. 436–446.

[72] N. Ezaki, K. Kiyota, B. T. Minh, M. Bulacu, and L. Schomaker, “Improved text-detection methods for a camera-based text reading system for blind persons,” in *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, ser. ICDAR ’05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 257–261.

[73] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004, British Machine Vision Computing 2002.

[74] J. Matas and K. Zimmermann, “Unconstrained licence plate and text localization and recognition,” in *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, sept. 2005, pp. 225 – 230.

[75] M. Donoser, C. Arth, and H. Bischof, “Detecting, tracking and recognizing license plates,” in *Proceedings of the 8th Asian conference on Computer vision - Volume Part II*, ser. ACCV’07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 447–456.

[76] D. Nistér and H. Stewénius, “Linear time maximally stable extremal regions,” in *Proceedings of the 10th European Conference on Computer Vision: Part II*, ser. ECCV ’08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 183–196.

[77] P.-E. Forssen, “Maximally stable colour regions for recognition and matching,” in *Computer Vision and Pattern Recognition, 2007. CVPR ’07. IEEE Conference on*, june 2007, pp. 1 –8.

[78] M. Donoser and H. Bischof, “3d segmentation by maximally stable volumes (MSVs),” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, 0-0 2006, pp. 63 –66.

[79] A. Vedaldi, “An implementation of multi-dimensional maximally stable extremal regions,” University of California, LA (UCLA), Tech. Rep.

[80] M. Iwamura, T. Tsuji, and K. Kise, “Memory-based recognition of camera-captured characters,” in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS ’10. New York, NY, USA: ACM, 2010, pp. 89–96.

[81] M. Iwamura, T. Kobayashi, and K. Kise, “Recognition of multiple characters in a scene image using arrangement of local features,” in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, sept. 2011, pp. 1409 –1413.

[82] T. Kobayashi, T. Toyama, F. Shafait, A. Dengel, M. Iwamura, and K. Kise, “Recognizing words in scenes with a head-mounted eye-tracker,” in *IAPR International Workshop on Document Analysis Systems*. IEEE, 3 2012.

[83] D. Gröger, "Gaze as implicit feedback for text detection in real scenes," Master's thesis, Techinische Universität Kaiserlautern, 2012.

[84] C. Koch and S. Ullman, "Shifts in selective attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[85] M. Cerf, E. P. Frady, and C. Koch, "Using semantic content as cues for better scanpath prediction," in *Proceedings of the 2008 symposium on Eye tracking research &#38; applications*, ser. ETRA '08. New York, NY, USA: ACM, 2008, pp. 143–146.

[86] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of Vision*, vol. 9, no. 12, 2009.

[87] A. D. A. Shahab, F Shafait and S. Uchida, "How salient is scene text?" in *IAPR International Workshop on Document Analysis Systems.* IEEE, 3 2012, pp. 317–321.

[88] F. Konuskan, "Visual saliency and biological inspired text detection," Master's thesis, Technical University Munich & California Institute of Technology, 2008.