Multi-script Text Extraction from Natural Scenes

Lluís Gómez and Dimosthenis Karatzas
Computer Vision Center
Universitat Autònoma de Barcelona
Email: {lgomez,dimos}@cvc.uab.es

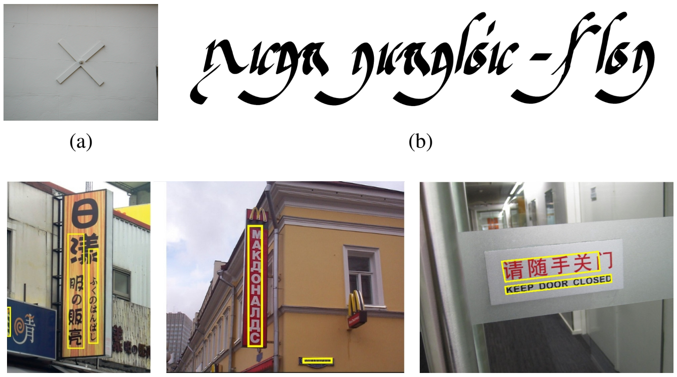
***Abstract*—Scene text extraction methodologies are usually based in classification of individual regions or patches, using a priori knowledge for a given script or language. Human perception of text, on the other hand, is based on perceptual organisation through which text emerges as a perceptually significant group of atomic objects. Therefore humans are able to detect text even in languages and scripts never seen before. In this paper, we argue that the text extraction problem could be posed as the detection of meaningful groups of regions. We present a method built around a perceptual organisation framework that exploits collaboration of proximity and similarity laws to create text-group hypotheses. Experiments demonstrate that our algorithm is competitive with state of the art approaches on a standard dataset covering text in variable orientations and two languages.**

I. Introduction

Text is ubiquitous in man-made environments and most of our daily activities imply reading and understanding written information in the world around us (shopping, finding places, viewing advertisements, etc). The automated localization, extraction and recognition of scene text in uncontrolled environments is still an open computer vision problem [1], [2], [3]. At the core of the problem lies the extensive variability of scene text in terms of its location, physical appearance and design.

A key characteristic of text is the fact that it emerges as a gestalt: a perceptually significant group of similar atomic objects. These atomic objects in the case of text are the character strokes giving rise to text-parts, be it well-separated characters, disjoint parts of characters, or merged groups of characters such as in cursive text. Such text-parts carry little semantic value when viewed separately (see Figure 1a), but become semantically relevant and easily identifiable when perceived as a group. Indeed, it can be shown that humans detect text without problems when perceptual organisation is evident irrespectively of the script or language - actually they are able to do so for non-languages as well (see Figure 1b).

In this sense, text detection is an interesting problem since it can be posed as the detection of meaningful groups of regions, as opposed to the analysis and classification of individual regions. Still, the latter is the approach typically adopted in state of the art methodologies. Some methods do include a post-processing stage where identified text regions are grouped into higher level entities: words, text lines or paragraphs. This grouping stage is not meant to facilitate or complement the detection of text parts, but to prepare the already detected text regions for evaluation, as the ground truth is specified at the word or text line level. This mismatch of semantic level between results and ground truth is a recognised problem that has given rise to specific evaluation methodologies that intend to tackle the problem [4].



(a) (b)

(c)

Fig. 1: (a) Should a single character be considered "text"? (b) An example of automatically created non-language text[1]. (c) Our method exploits perceptual organization laws always present in text, irrespective of scripts and languages.

We pose that the fact that a region can be related to other neighbouring ones is central to classifying the region as a text part, and is in-line with the essence of human perception of text, which is largely based on perceptual organisation. The research hypothesis of this work is thus that building an automatic text detection process around the above fact can help overcome numerous identified difficulties of state of the art text detection methods.

To test the above hypothesis a state of the art perceptual organisation computational model is employed to assess the meaningfulness of different candidate groups of regions. These groups emerge naturally through the activation of different visual similarity laws in collaboration with a proximity law.

Similarity between regions is not strictly defined in our framework. This is intentional, as due to design, scene layout, and environment effects different perceptual organisation laws might be active in each case. Since text is a strong gestalt, a subset of such laws are expected to be active in parallel (collaborating) at any given instance. As a result a flexible approach is proposed here where various similarity laws are taken into account and the groups emerging through the individual activation of each similarity law provide the evidence to decide on the final set of most meaningful groups. The resulting method does not depend on the script, language or orientation of the text to be detected.

[1] Daniel Uzquiano's random stroke generator: http://danieluzquiano.com/491

II. Related Work

The automatic understanding of textual information in natural scenes has gained increasing attention over the past decade, giving rise to various new computer vision challenges. Jung *et al.* [5] and Liang *et al.* [6] offer an exhaustive survey on camera-based analysis of text in real environments.

The large number of techniques proposed for text localization in natural scenes can be divided into patch-based, region-based, and hybrid approaches. Patch-based methods usually work by performing a sliding window search over the image and extracting certain texture features in order to classify each possible patch as text or non-text. Coates *et al.* [7], and in a different flavour Wang *et al.* [3] and Netzer *et al.* [8], propose the use of unsupervised feature learning to generate the features for text versus non-text classification. Wang *et al.* [2], extending their previous work [9], have built an end-to-end scene text recognition system based on a sliding window character classifier using Random Ferns, with features originating from a HOG descriptor. Mishra *et al.* [10] propose a closely related end-to-end method based on HOG features and a SVM classifier. Patch based methods yield good text localisation results, although they do not directly address the issue of text segmentation (separation of text from background) and thus require further preprocessing before recognition.

On the other hand, region-based methods are based on a typical bottom-up pipeline: first performing an image segmentation and subsequently classifying the resulting regions into text or non-text ones. Frequently these are complemented with a post-processing step where said regions assessed to be characters are grouped together into words or text lines. Yao *et al.* [11] extract regions in the Stroke Width Transform (SWT) domain, proposed earlier for text detection by Epshtein *et al.* [12]. Chen *et al.* [13] obtain state-of-the-art performance with a method that determines the stroke width using the Distance Transform of edge-enhanced Maximally Stable Extremal Regions (MSER). The effectiveness of MSER for character candidates detection is also exploited by Novikova *et al.* [14], while Neumann *et al.* [1] propose a region representation derived from MSER where character/non-character classification is done for each possible Extremal Region (ER). Other methods in this category make use of regions obtained from the image edge gradient [15], or by color clustering [16].

Pan *et al.* [17] obtain state-of-the-art accuracy following a hybrid method where a classifier using HOG features builds a text confidence map feeding a local binarization algorithm for region extraction.

There exist two main differences between current state-of-the-art approaches and the method proposed in this paper. On one side, methods relying on learning processes [2], [3], [7], [8], [9], [10] are usually constrained to detect the single script which they have been trained on. The feedback loop between localization and recognition they propose, although performing well on certain tasks (e.g. detecting English horizontal text), contradicts with the human ability to detect text structures even in scripts or languages never seen before. In comparison, the methodology proposed here requires no training, and is largely parameter free and independent to the text script.

On the other hand, there is an important distinction between the way the grouping is used by existing methods [11], [12], [13] and the way perceptual organisation is used here. In past work, grouping is used solely as a post-processing step, once the text parts have already been identified as such, the main reason being to address the results / ground truth semantic level mismatch mentioned before. As a matter of fact, we also use a post-processing step to enable us to evaluate on standard datasets. However, crucially, in our approach perceptual organisation provides the means to perform classification of regions, based on whether there exists an interpretation of the image that involves their participation to a perceptually relevant group.

III. Text Localization Method

We present a region based method where extracted regions are grouped together in a bottom-up manner guided by similarity evidence obtained over various modalities such as color, size, or stroke width among others, in order to obtain meaningful groups likely to be text gestalts, i.e. paragraphs, text lines, or words. Figure 2 shows the pipeline of our algorithm for text extraction where the process is divided in three main steps: region decomposition, perceptual organization based analysis, and line formation.

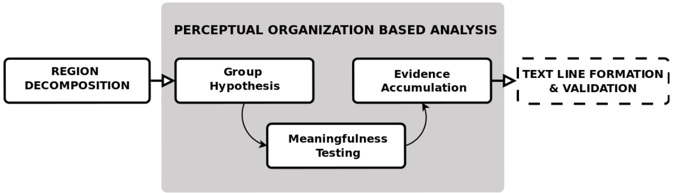


Fig. 2: Text Extraction algorithm pipeline.

A. Region Decomposition

The use of Maximally Stable Extremal Regions (MSER) [18] for detecting text character candidates in natural scene images is extensively used in recent state of the art methods for text detection [1], [13], [14]. The MSER algorithm builds a tree of regions with an extremal property of the intensity function over its outer boundary, and this property is normally present in all text parts as they are explicitly designed with high contrast in order to be easily read by humans.

The resulting MSER tree is pruned by filtering the regions that are not likely to be text parts by their size, aspect ratio, stroke width variance, and number of holes.

B. Perceptual Organization Clustering

The perceptual organization clustering is applied to the entire set of resulting MSERs in three stages. First we create a number of possible grouping hypotheses by examining different feature sub-spaces. Then, these groups of regions are analysed and we keep the most meaningful ones, thus providing an ensemble of clusterings. Finally, those meaningful clusterings are combined based on evidence accumulation [19].

1) Group Hypothesis Creation: We aim to use simple and low computational cost features describing similarity relations between characters of a word or text line. The list of features we use for this kind of similarity grouping are:

Geometrical features. Characters in the same word usually have similar geometric appearance. We make use of the bounding box area, number of pixels, and diameter of the bounding circle.

Intensity and color mean of the region. We calculate the mean intensity value and the mean color, in the L*a*b* colorspace, of the pixels that belong to the region.

Intensity and color mean of the outer boundary. Same as before but for the pixels in the immediate outer boundary of the region.

Stroke width. To determine the stroke width of a region we make here use of the Distance Transform as in [13].

Gradient magnitude mean on the border. We calculate the mean of the gradient magnitude on the border of the region.

Each of these similarity features is coupled with spatial information, i.e. x,y coordinates of the regions' centers, in order to capture the collaboration of the proximity and similarity laws. So, independently of the similarity feature we consider, we restrict the groups of regions that are of interest to those that comprise spatially close regions.

We build a dendrogram using Single Linkage Clustering analysis for each of the feature sub-spaces described above. Each node in the obtained dendrograms represents a group hypothesis whose perceptual meaningfulness will be evaluated in the next step of the pipeline.

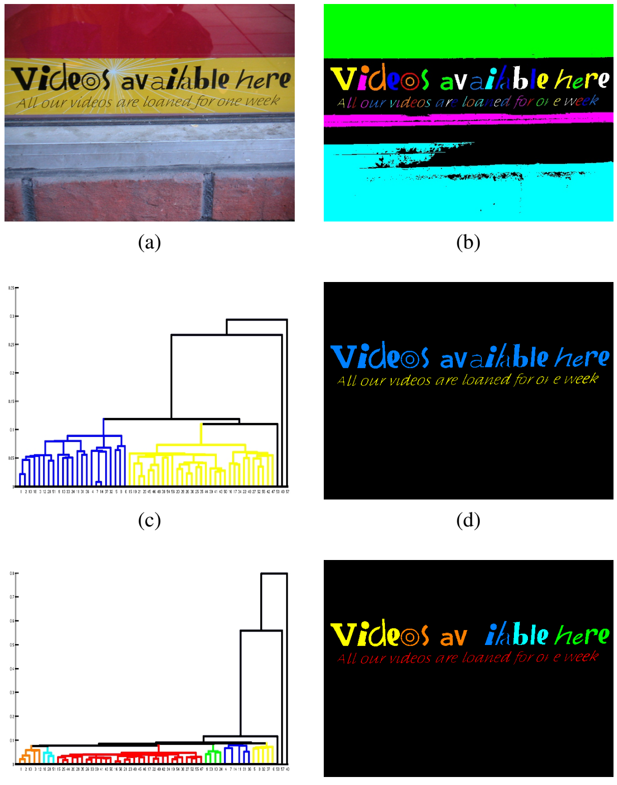
2) Meaningfulness Testing: In order to find meaningful groups of regions in each of the defined feature sub-spaces we make use of a probabilistic approach to Gestalt Theory as formalised by Desolneux *et al.* [20]. The cornerstone of this theoretical model of perceptual organization is the Helmholtz principle, which could be informally summarised as: "We do not perceive anything in a uniformly random image", or with its equivalent *"a contrario"* statement: "Whenever some large deviation from randomness occurs in an image some structure is perceived". This general perception law, also known as the principle of common cause or of non-accidentalness, has been stated several times in Computer Vision, with Lowe [21] being the first to pose it in probabilistic terms.

The Helmholtz principle provides the basis to derive a statistical approach to automatically detect deviations from randomness, corresponding to meaningful events. Consider that n atomic objects are present in the image and that a group G of k of them have a feature in common. We need to answer the question of whether this common feature is happening by chance or not (and thus is a significant property of the group). Assuming that the observed quality has been distributed randomly and uniformly across all objects, the probability that the observed distribution for G is a random realisation of this uniform process is given by the tail of the binomial distribution:

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

Where p is the probability of a single object having the aforementioned feature.

We make use of this metric in the dendrogram of each of the feature sub-spaces produced in the previous step separately to assess the meaningfulness of all produced grouping hypotheses. We calculate (1) for each node (merging step) of the dendrogram, using as p the ratio of the volume defined by the distribution of features of the samples forming the group with respect to the total volume of the feature sub-space. We then select as maximally meaningful a cluster A iif for every successor B and every ancestor C, it is $\mathcal{B}_B(k,n,p) > \mathcal{B}_A(k,n,p)$ and $\mathcal{B}_C(k,n,p) \geq \mathcal{B}_A(k,n,p)$. Notice that by using this maximality criteria no region is allowed to belong to more than one meaningful group at the same time. The clustering analysis is done without specifying any parameter or cut-off value and without making any assumption on the number of meaningful clusters, but just comparing the values of (1) at each node in the dendrogram.



(a) (b)

(c) (d)

(e) (f)

Fig. 3: (a) A scene image from the ICDAR2003 dataset, (b) its MSER decomposition. (c) Dendrogram of the feature subspace (x,y coordinates, intensity mean), and (d) the maximal meaningful clusters found; (e)(f) same for the feature subspace (x,y coordinates, stroke width).

Figure 3 shows the maximal meaningful clusters detected in a natural scene image for two of the feature sub-sets defined, in Figure 3d image regions are clustered in a three dimensional space based on proximity and intensity value, while in Figure 3f they are clustered based on proximity and stroke width. This behaviour, of arriving to different grouping results depending on the similarity modality examined is desirable, as it

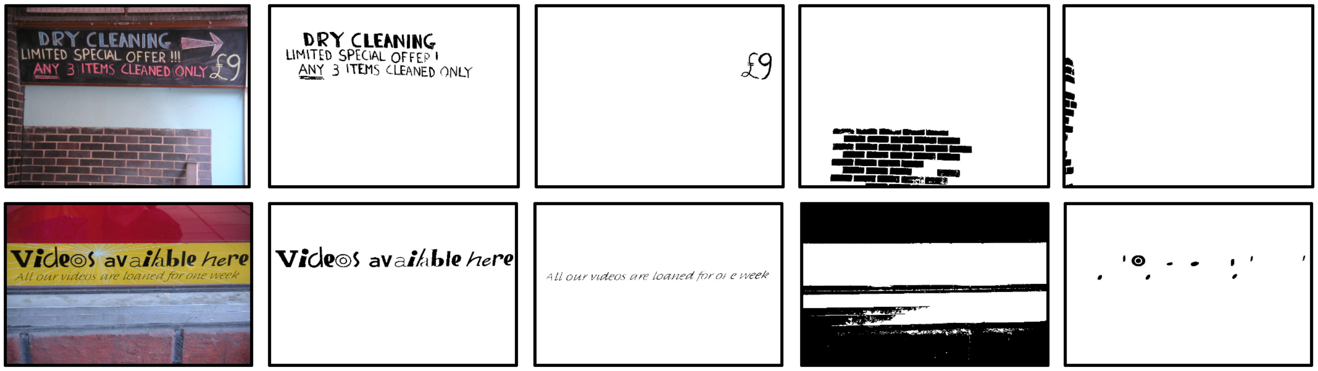

DRY CLEANING
LIMITED SPECIAL OFFER !!!
ANY 3 ITEMS CLEANED ONLY £9
Videos available here
All our videos are loaned for one week

Fig. 4: Maximally meaningful clusters validated through the Evidence Accumulation framework.

allows us to deal with variabilities in text design, illumination, perspective distortions, and so on.

3) Evidence Accumulation: Once we have detected the set of maximally meaningful clusters P^i in each feature sub-space $i \in N$, the clustering ensemble $\mathbb{P} = \{P^1, P^2, P^3, ..., P^N\}$ is used to calculate the evidence [19] for each pair of regions to belong to the same group, producing a co-occurrence matrix $\mathcal{D}$ defined as:

$$\mathcal{D}(i, j) = \frac{m_{ij}}{N} \quad (2)$$

Where m_{ij} is the number of times the regions i and j have been assigned to the same maximal meaningful cluster in $\mathbb{P}$.

The co-occurrence matrix $\mathcal{D}$ is used as a dissimilarity matrix in order to perform the final clustering analysis of the regions, by applying the same hierarchical clustering process described in section III-B2.

Figure 4 shows two examples of the results obtained where the method exhibits its ability to deal with a flexible definition of what a text group is: on the bottom row text characters do not have the same stroke width but they share the same color, while on the top, on the contrary, text characters do not have the same color but have similar stroke, size, etc.

As expected, not only text is detected as meaningful, but also any kind of region arrangement with a text like structure. It is important however to note at this stage that the algorithm produces relatively pure text-only clusters, with almost all text parts falling in clusters that contain virtually no non-text components. In order to filter non-text meaningful groups we use a combination of two classifiers. First, each region of the group is scored with the probability of being or not a character by a Real Adaboost classifier using features combining information of stroke width, area, perimeter, number and area of holes. Then, simple statistics of this scores are fed into a second Real Adaboost classifier for text/non-text group classification together with same features as before (but in this case for the whole group and not for independent regions) and a histogram of edge orientations of the Delaunay triangulation built with the group regions centers. Both classifiers are trained using the ICDAR2003 [22] and MSRA-TD500 [11] training sets as well as with synthetic text images, using different scripts, to ensure script-independence.

IV. Experiments and Results

The proposed method has been evaluated on two multi-script datasets for different tasks, in one hand for text segmentation on the KAIST dataset, and on the other for text localization in the MSRA-TD500 dataset. Despite we did also evaluation on the ICDAR 2003 Robust Reading Competition dataset, results are not reported here due to space, and because multi-script datasets fit better with the focus of this paper. The full list of results can be seen at http://dag.cvc.uab.es/text_localization, while the interested reader can try our method directly by submitting an image at the online demo available at the same url.

A. Text Segmentation

The KAIST dataset [16] comprises 3000 natural scene images, with a resolution of 640x480, categorized according to the language of the scene text captured: Korean, English, and Mixed (Korean + English). For our experiments we use only 800 images corresponding to the Mixed subset. For the pixel level evaluation precision p and recall r are defined as $p = |E \cap T|/|E|$ and $r = |E \cap T|/|T|$, where E is the set of pixels estimated as text and T is the set of pixels corresponding to text components in the ground truth. Table I show the obtained results on the KAIST dataset.

Method	p	r	f-score	time(s)
Our Method	0.66	**0.78**	**0.71**	**0.41**
Lee *et al.* [16]	**0.69**	0.60	0.64	n/a

TABLE I: KAIST dataset performance comparison.

B. Variable Orientation Text Detection

The MSRA-TD500 dataset [11] contains arbitrary oriented text in both English and Chinese and is proposed as an alternative to the ICDAR2003 [22] dataset where only horizontal English text appears. The dataset contains 500 images in total, with varying resolutions from 1296×864 to 1920×1280. The evaluation is done as proposed in [11] using minimum area rectangles. For an estimated minimum area rectangle D to be considered a true positive, it is required to find a ground truth rectangle G such that:

$$A(D' \cap G')/A(D' \cup G') > 0.5, abs(\alpha_D - \alpha_G) < \pi/8$$

where D' and G' are the axis oriented versions of D and G, $A(D' \cap G')$ and $A(D' \cup G')$ are respectively the area of their intersection and union, and α_D and α_G their rotation angles. The definitions of precision p and recall r are: $p = |TP|/|E|$, $r = |TP|/|T|$ where TP is the set of true positive detections while E and T are the sets of estimated rectangles and ground truth rectangles.

As the perceptually meaningful text groups detected by our method rarely correspond directly to the semantic level ground truth information is defined in (lines in the case of the MSRA-TD500 dataset), the proposed method is extended with a simple post-processing step in order to obtain text line level bounding boxes.

We consider a group of regions as a valid text line if the mean of the y-centres of its constituent regions lies in an interval of 40% around the y-centre of their bounding box and the variation coefficient of their distribution is lower than 0.2. Notice that, as we are considering text lines at any possible orientation, the orientation of the group (and consequently the definition of the y-axis) is always defined in relation to the axes of the circumscribed rectangle of minimum area for the given group. If the collinearity test fails, it may be the case that the group comprises more than one text line. Thus in such a case we perform a histogram projection analysis in order to identify the text lines orientation, and then split the inital group into possible lines by clustering regions on the identified direction. This process is iteratively repeated until all regions have either been assigned to a valid text line or rejected, using the collinearity test described above, or until no more partitions can be found.

Table II show a comparison of the obtained results with other state of the art methods on the MSRA-TD500 dataset.

Method	p	r	f-score	time(s)
TD-Mixture [11]	**0.63**	**0.63**	**0.60**	7.2
Our Method	0.58	0.54	0.56	**2.97**
TD-ICDAR [11]	0.53	0.52	0.50	7.2
Epshtein *et al.* [12]	0.25	0.25	0.25	6
Chen *et al.* [23]	0.05	0.05	0.05	n/a

TABLE II: MSRA-TD500 dataset performance comparison.

V. Conclusions

A new methodology for text extraction from scene images was presented, inspired by the human perception of textual content, largely based on perceptual organisation. The proposed method requires practically no training as the perceptual organisation based analysis is parameter free. It is totally independent of the language and script in which text appears, it can deal efficiently with any type of font and text size, while it makes no assumptions about the orientation of the text. Qualitative results demonstrate competitive performance and faster computation.

The approach presented opens up a number of possible paths for future research, including the higher integration of the region decomposition stage with the perceptual organisation analysis, and further investigation on the computational modelling of perceptual organisation aspects such as masking, conflict and collaboration.

VI. Acknowledgements

This project was supported by the Spanish projects TIN2011-24631 and 2010-CONE3-00029, the fellowship RYC-2009-05031, and the Catalan government scholarship 2013FI1126.

References

[1] L. Neumann and J. Matas, “Real-time scene text localization and recognition,” in *Proc. CVPR*, 2012. 1, 2

[2] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” in *Proc. ICCV*, 2011. 1, 2

[3] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, “End-to-end text recognition with convolutional neural networks,” in *Proc. ICPR*, 2012. 1, 2

[4] C. Wolf and J.-M. Jolion, “Object count/area graphs for the evaluation of object detection and segmentation algorithms,” *IJDAR*, 2006. 1

[5] K. Jung, K. I. Kim, and A. K. Jain, “Text information extraction in images and video: a survey,” *Pattern Recognition*, 2004. 2

[6] J. Liang, D. Doermann, and H. Li, “Camera-based analysis of text and documents: a survey,” *IJDAR*, 2005. 2

[7] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. Wu, and A. Ng, “Text detection and character recognition in scene images with unsupervised feature learning,” in *Proc. ICDAR*, 2011. 2

[8] 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. 2

[9] K. Wang and S. Belongie, “Word spotting in the wild,” in *Proc. ECCV*, 2010. 2

[10] A. Mishra, K. Alahari, and C. Jawahar, “Top-down and bottom-up cues for scene text recognition,” in *Proc. CVPR*, 2012. 2

[11] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, “Detecting texts of arbitrary orientations in natural images,” in *Proc. CVPR*, 2012. 2, 4, 5

[12] B. Epshtein, E. Ofek, and Y. Wexler, “Detecting text in natural scenes with stroke width transform,” in *Proc. CVPR*, 2010. 2, 5

[13] 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,” in *Proc. ICIP*, 2011. 2, 3

[14] T. Novikova, O. Barinova, P. Kohli, and V. Lempitsky, “Large-lexicon attribute-consistent text recognition in natural images,” in *Proc. ECCV*, 2012. 2

[15] J. Zhang and R. Kasturi, “Text detection using edge gradient and graph spectrum,” in *Proc. ICPR*, 2010. 2

[16] S. Lee, M. S. Cho, K. Jung, and J. H. Kim, “Scene text extraction with edge constraint and text collinearity,” in *Proc. ICPR*, 2010. 2, 4

[17] Y.-F. Pan, X. Hou, and C.-L. Liu, “Text localization in natural scene images based on conditional random field,” in *Proc. ICDAR*, 2009. 2

[18] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and Vision Computing*, 2004. 2

[19] A. Fred and A. Jain, “Combining multiple clusterings using evidence accumulation,” *IEEE Trans. PAMI*, 2005. 2, 4

[20] A. Desolneux, L. Moisan, and J.-M. Morel, “A grouping principle and four applications,” *IEEE Trans. PAMI*, 2003. 3

[21] D. G. Lowe, *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, 1985. 3

[22] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “ICDAR 2003 robust reading competitions: entries, results, and future directions,” *IJDAR*, 2005. 4

[23] X. Chen and A. Yuille, “Detecting and reading text in natural scenes,” in *Proc. CVPR*, 2004. 5