Demonstration of a Human Perception Inspired System for Text Extraction from Natural Scenes

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

This demonstration presents a prototype implementation of the method described in [1], able to detect textual content in scenes in real time. The text detection algorithm makes use of a high level representation of text as a perceptually significant group of atomic regions. This representation is independent from the particular script or language of the text, allowing the system to detect textual content in multi-script scenarios without any extra training.

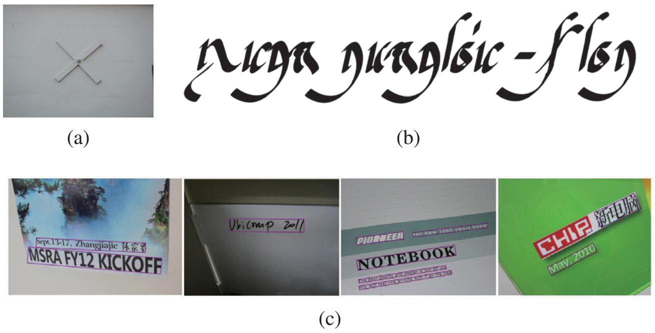
Mainstream state of the art methodologies are usually based on the classification of individual regions (connected components) or image patches. The approach followed here is distinctly different as text detection is not performed based on classifying individual components, but through searching for groups of components that share certain characteristics.

The inspiration comes from human perception. Humans make strong use of perceptual organisation to detect textual content, through which text emerges as a perceptually significant group of atomic objects (disjoint text parts such as characters or ideograms). Therefore humans are able to detect text even in languages and scripts never seen before (see Figure 1b), as long as such gestalts emerge by the way the text parts are arranged. As a result, the text extraction problem can be posed as the detection of such gestalts, as we show in [1].

From an implementation point of view (see Figure 2), the system creates group hypotheses, considering a number of different modalities in which atomic regions might be similar to each other (size, colour, background colour, stroke width, etc) and subsequently tests these hypotheses in terms of their meaningfulness. Meaningfulness is defined in a probabilistic way according to [2]. Given a group hypothesis comprising a set of regions which have a feature in common (they are similar in terms of that particular feature), meaningfulness measures the extent to which this common feature is happening by chance or not (thus it is a significant property of the group).

Similarity is explored in a number of modalities separately, but always in association with the spatial proximity of the atomic objects. The collaboration of the different similarity laws is taken into account at the end of the process through evidence accumulation [3], which provides a flexible way to identify maximal perceptual groups without any strict definition of the exact similarity laws invoked.

The system runs on a laptop computer using a Web camera to sense the environment. The current implementation offers real time performance at VGA resolution. The processing time of a frame is mainly dependent on the complexity of the scene and lesser on the frame resolution.


(a) (b)

(c)

Fig. 1: (a) Should a single character be considered "text"? (b) An example of automatically created non-language text[1]. (c) Example result images from the MSRA-TD500 dataset.

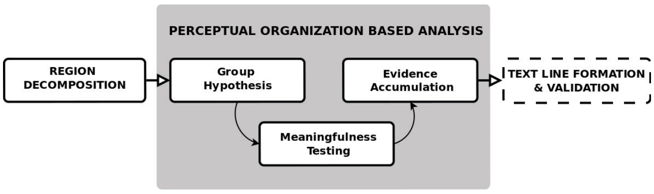


Fig. 2: Text Extraction algorithm pipeline.

Acknowledgements

The creation of this demo 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. Gomez and D. Karatzas, "Multi-script text extraction from natural scenes," in *Proceedings of the 12th Int. Conf. on Document Analysis and Recognition (ICDAR 2013)*, 2013. 1

[2] A. Desolneux, L. Moisan, and J.-M. Morel, "A grouping principle and four applications," *IEEE Trans. PAMI*, 2003. 1

[3] A. Fred and A. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. PAMI*, 2005. 1

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