Towards Modelling an Attention-based Text Localization Process

Antonio Clavelli[1], Dimosthenis Karatzas[1] Josep Lladós[1], Mario Ferraro[2], and Giuseppe Boccignone[3]

[1] Computer Vision Center, Universitat Autonoma de Barcelona
Edificio O, Campus UAB, 08193 Bellaterra (Cerdanyola) Barcelona, Spain
{aclavelli,dimos,josep}@cvc.uab.cat
[2] Dipartimento di Fisica, Universitá di Torino
via Pietro Giuria 1, 10125 Torino, Italy
ferraro@ph.unito.it
[3] Dipartimento di Informatica, Universitá di Milano
via Comelico 39/41, Italy
giuseppe.boccignone@unimi.it

Abstract. This note introduces a visual attention model of text localization in real-world scenes. The core of the model built upon the proto-object concept is discussed. It is shown how such dynamic mid-level representation of the scene can be derived in the framework of an action-perception loop engaging salience, text information value computation, and eye guidance mechanisms.
Preliminary results that compare model generated scanpaths with those eye-tracked from human subjects are presented.

Keywords: text localization, visual attention, eye guidance

1 Introduction

Text localization and recognition in real-world scenes has received in recent years significant attention [8], [21], [7], yet it is considered an open problem due to the complexity of the task.

Differently from mainstream research in this field, here we present some preliminary steps towards a model for actively detecting and locating text within unconstrained complex scenes relying on visual attention and eye guidance mechanisms. The model, in its perceptual component focuses on the concept of proto-objects, suitable to provide a unifying perspective on the integration of low-level salience and high-level text information value, in order to determine a probabilistic distribution of where to look next. In its eye-guidance component such distribution is used to stochastically sample the actual gaze shift in the vein of foraging-based methods that have been recently used to mimic inter- and intra-subject variability in generating visual scanpaths [1], [2].

Currently, most published methods for text localization and recognition in natural scenes are conceived as an extension to work done in printed document

processing. Namely, they are based on sequential pipeline processing consisting of three steps - text localization, text segmentation and processing by an OCR. In particular, for text detection (the determination of the presence of text in a given image or frame) and localization (determining the exact location of text in the image/frame and generating bounding boxes around it) a number of approaches have been proposed mainly relying on either region-based and/or texture-based methods (cfr. [14], for a recent survey). However, in contrast to the printed document setting: (1) real-world texts are often short snippets written in different fonts and languages, and embedded in a cluttered background; (2) their alignment does not follow the rules of printed documents; (3) many words are proper names that prevent an effective use of a dictionary [8].

But further and more fundamental for motivating the perspective taken here, automatic text detection in natural scenes, beyond posing such technical challenges, is likely to address applications that call for an active or animate vision paradigm [16], where visual attention plays a central role. Visual impairment assistance system, tourist assistance system, Unmanned Ground Vehicle navigation in urban environments, or Domestic Service Robots are some examples. Indeed, for such applications, where the movement of the body, head, and eyes of the observer (either natural, artificial or hybrid) determines the quality of what is perceived, the classic approaches to text localization, relying on a passive vision rationale are likely to be methodologically inadequate [16].

2 Background

The use of attention-based mechanisms and representation models for the task of text detection is relatively new. Some attempts have been made recently to integrate bottom-up saliency computation, a step shared by a variety of computational visual attention systems [3], into the text detection and recognition pipeline, e.g., [15], [12], [6]. In most cases, the intent is to use saliency as a tool to detect Regions of Interest (ROI) where text might be present. However, the presence of text may or may not correlate with saliency computed from the bottom-up [13]. On the other hand, saliency computation is not a sufficient condition (and it might not even be a necessary one [16]) to support an active approach based on visual attention that is deeply grounded in eye guidance mechanisms [16], [10], [17].

Indeed, when inspecting real-world scenes, human observers direct long shifts (saccades) to actively reposition the center of gaze on circumscribed regions of interest, the so called "focus of attention" (FOA); the succession of gaze shifts is usually referred to as a scanpath [10]. Significant information is generally provided by the objects placed in the context of a scene and observed under a given task [16] (e.g., cars and people in a urban scene for a walking pedestrian).

Evidence has been given that text is yet another entity that frequently captures humans gaze in natural scenes [4], [20]. Text detection in real world is crucial for people to perform efficiently in everyday life, for example, by drawing attention to traffic signs or displays showing directions to a hospital. Cerf et al.

[4] have shown that, in natural images, faces and text attract gaze independent of the task: they are difficult to ignore, even if there is a real cost associated with looking at them. Text attractiveness has been further investigated and confirmed by Wang et al. [20], who have shown that: specific visual features of texts, rather than classic low-level salient features, are the main attractors of attention; texts placement partially contributes to this effect; the meaningfulness of texts does not increase attentional capture.

Such experimental findings provide the rationale behind the model introduced in the following section

3 The model

Assume that the input is in the form of either a picture (static image) $\mathbf{I}$, or a video, that is a time-parametrized sequence of images. The general aim of a computational model of visual attention is to answer the question *Where to Look Next?* by providing, at the computational theory level an explanation of the mapping *viewed scene* $\mapsto$ *gaze sequence*, say $\{\mathbf{r}(1), \mathbf{r}(2), \cdots\}$, together with a procedure that implements such mapping.

In analogy with other aspects of motor behaviour and action selection, the guidance of eye movements is likely to be influenced by a hierarchy of several interacting control loops, operating at different levels of processing [10]. Each level exploits the most suitable representation $\mathcal{R}$ of the viewed scene for its own level of abstraction: Schütz, et al. [10], in a plausible portrayal, have singled out salience, objects, values, and plans.

A great deal of approaches that qualify as computational models of visual attention are incomplete with respect to the mapping *viewed scene* $\mapsto$ *gaze sequence*. They mostly account for the mapping from an image, or less frequently from an image sequence, to a representation $\mathcal{R}$, typically a saliency map s. The saliency map is then quantitatively evaluated by comparing with eye movement data according to some evaluation measure [3]. Thus, a partial mapping $\mathbf{I} \mapsto s$ is provided. Clearly, even though the mapping $\mathbf{I} \mapsto s$ is taken for granted, yet the next step $s \mapsto \{\mathbf{r}(1), \mathbf{r}(2), \cdots\}$ is a long way off.

In this perspective, our model of attentive text location/exploration grounds in an action-perception loopy interaction between the perceptual/inferential process and the guidance process under a given task. At a glance (cfr., Fig. 1), the model relies upon three processing components: the visual front-end component handles low-level representations of the scene; the perceptual component deals with mid level and higher level descriptions; the guidance component provides the appropriate gaze shift dynamics as a function of the current gaze location and the perceived scene.

In an analysis-by-synthesis formulation, the context or *gist* $\mathbf{G}$ of a scene influences the appearance and possible locations $\mathbf{L}$ of certain kinds of *objects* $\mathbf{O}$ [19]. Main objects considered here are textual objects. $\mathbf{G}$ and $\mathbf{O}$ together, generate a mid-level *proto-objects* representation. In our model proto-objects $\mathbf{W}$ are dynamic feature-based descriptions of the salient and value based portion

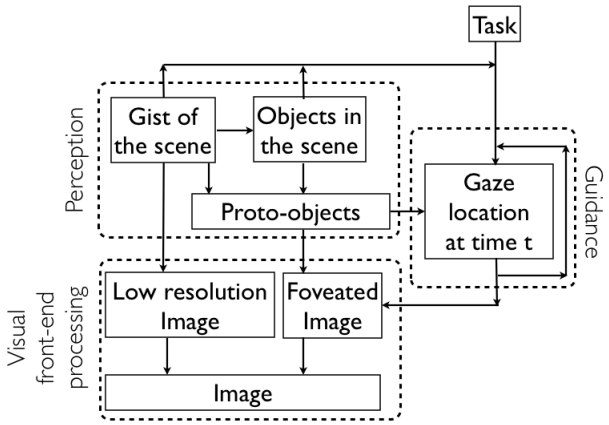

Fig. 1. The model at a glance. Dashed line blocks denote processing components; solid line blocks stand for the different kinds of representations handled by processing components

of the *foveated image* $\widehat{\mathbf{I}}$ available at the current time t. For what concerns the visual front-end representation, the low-resolution peripheral representation of the visual field $\widetilde{\mathbf{I}}$ depends on the gist $\mathbf{G}$, whilst the foveated image $\widehat{\mathbf{I}}$ depends on the dynamic mid-level representation and the current gaze location, the FOA center $\mathbf{r}_{FOA}(t)$. The input image or frame $\mathbf{I}$ can eventually be generated from the low-resolution representation $\widetilde{\mathbf{I}}$ together with the foveal high resolution representation $\widehat{\mathbf{I}}$.

The next gaze position $\mathbf{r}_{FOA}(t+1)$ is dynamically determined in terms of the given task, the global gist, the knowledge about objects, the current setting of the mid-level representation and the current FOA $\mathbf{r}_{FOA}(t)$. The gaze guidance mechanism is the component able to select visual information, grabbing in turn a few proto-objects and making them available for further processing. In order to account for the several latent factors in the perception/action loops involved in the guidance of eye-movements [10], together with the oculomotor biases and the "internal" noise (related to perceptual and motor systems) [17], we assume that the above representations are shaped in the form of probability distributions and that the gaze sequence is generated by an underlying stochastic process. In particular, the guidance mechanism is conceived following a "foraging eye" metaphor and designed as a constrained composite random walk [1], [2], on the dynamic visual landscape here represented by $\mathbf{W}$. A basic composite random walk is one where the forager can be engaged in one of two possible behaviors or motor actions $\mathbf{A}(t)$: (i) local feeding on food patches or (ii) engaging in longer displacement or flight (typically, a Lévy flight, [1], [2]) to encounter new patches. Local feeding corresponds to fixating; note, that a fixation is not simply the maintenance of the visual gaze on a single location but rather a slow oscillation of the eye (minimum 50 milliseconds duration) within a circumscribed region

(typically $0.5° - 2.0°$ degrees of visual angle), [5]. Long displacements stand for saccades. The resulting random walk is the visual scanpath. Such stochastic formulation has been shown to suitably describe and mimic inter- and intra-subject scanpath variability [1],[2].

The bulk of the model outlined above is represented by the interaction between proto-objects and eye-guidance. Our notion of proto-objects is inspired by the Coherence Theory of Attention [9] where they are conceived as the dynamic interface between high-level and low-level processing, a "quick and dirty" interpretation of the scene. They are volatile, being replaced when any new stimulus appears at their retinal location: focused attention acts then as a metaphorical hand that grasps a small number of proto-objects endowing them the coherence of an object.

In our framework, such a dynamic mid-level interpretation of the scene is formalized as follows. Following [19], the observed scene can be described in terms of local and contextual features $F = \{F_L, F_C\}$. Assume that gist features $\mathbf{F_C}$ have been suitably computed (e.g., [19]). Then, at a certain location $\mathbf{L}$, local features $\mathbf{F_L}$ are generated from the object being present at that specific location. Since in our case we are dealing with text localization, $\mathbf{O}$ represents a binary random variable in the set $\{text, \neg text\}$ and the distributions $P(\mathbf{F_L}|\mathbf{O} = text)$ and $P(\mathbf{F_L}|\mathbf{O} = \neg text)$ can be preventively learned in a training stage (cfr, Section 4).

At observation time t, we can define the probabilities of a scene point $\mathbf{L}$ to be salient/ non salient, $P(\mathbf{S}|\mathbf{L}, \mathbf{F_L}, \mathbf{F_C})$, and to bear an information value as regards it being embedded or not in a text region, $P(\mathbf{V}|\mathbf{L}, \mathbf{F_L}, \mathbf{F_C}, \mathbf{O} = text)$. Here, $\mathbf{S}$ and $\mathbf{V}$ are the saliency and value binary random variables, respectively.

To account for the very notion of proto-objects, we assume that, at any time t, the dynamic proto-object map $\mathbf{W}(t)$ is formed by the foraging eye, taking into account the saliency and text information value according to the current oculomotor action $\mathbf{A}(t)$. Such a "constantly regenerating flux" [9] can thus be summarized in the following sampling steps.

1. Sampling the proto-objects:

$$\mathbf{S}(t) \sim P(\mathbf{S}(t)|\mathbf{L}(t), \mathbf{F_L}(t), \mathbf{F_C}(t)), \tag{1a}$$

$$\mathbf{V}(t) \sim P(\mathbf{V}(t)|\mathbf{L}(t), \mathbf{F_L}(t), \mathbf{F_C}(t), \mathbf{O}(t) = text), \tag{1b}$$

$$\mathbf{W}(t) \sim P(\mathbf{W}(t)|\mathbf{S}(t), \mathbf{V}(t), \mathbf{A}(t)), \tag{1c}$$

2. Sampling where to look next:

$$\mathbf{A}(t) \sim P(\mathbf{A}(t)|\mathbf{A}(t-1)), \tag{2a}$$

$$\mathbf{r}_{FOA}(t+1) \sim P(\mathbf{r}_{FOA}(t+1)|\mathbf{r}_{FOA}(t), \mathbf{A}(t), \mathbf{W}(t)). \tag{2b}$$

At each time step t, $\mathbf{r}_{FOA}(t+1)$ represent the point in the scene where to look next. When the gaze is shifted to such point, at time $t+1$, a new foveated image $\widehat{\mathbf{I}}(t+1)$ is generated, and as a consequence the new proto-object landscape $\mathbf{W}(t+1)$, according to Eqs 1a,1b and 1c.

4 Simulation and results

A publicly available dataset (`http://algoval.essex.ac.uk/icdar/Datasets.html`) has been used for testing the behavior of the model's simulation. This consists of 307 color street view pictures of sizes ranging from 1360×1024 to 1024×768 pixels. The text content is embedded in the scene in the form of shop names, street signs or advertisements and it is usually not located at the center of the image, nor covering a large region of the image, so as to make the localization problem more difficult and calling for an exploration of the scene.

In the current version of the model, sampling steps have been simulated as follows. According to [19], gist features $\mathbf{F_C}$ are computed from the low resolution image $\widetilde{\mathbf{I}}$, the lowest level of a 4 level Gaussian pyramid decomposition of the input image $\mathbf{I}$, and locations $\mathbf{L}$ within the image spatial support Ω of the scene where objects are likely to appear, are determined. Then, given a fixation point $\mathbf{r}_{FOA}(t)$ (the center of the image, for $t = 0$), the foveation process is simulated by blurring $\mathbf{I}$ through an isotropic Gaussian function centered at $r_{FOA}(t)$, whose variance is taken as the radius of a FOA, $\sigma = |FOA|$, approximately given by $1/8 \min[w, h]$, where $w \times h = |\Omega|$, $|\Omega|$ being the dimension of support Ω. The foveated image $\widehat{\mathbf{I}}(t)$ is used to compute the feature matrix $\mathbf{F_L}$, through a locally data-adaptive kernel density estimator of the distributions $P(\mathbf{F_L}|\mathbf{O})$ [11]. Saliency is estimated subsequently as follows. By using Bayes' rule, we can write the r.h.s. of Eq. 1a as $P(\mathbf{S}(t)|\mathbf{L}(t), \mathbf{F_L}(t), \mathbf{F_C}(t)) \propto P(\mathbf{F_L}(t)|\mathbf{S}(t), \mathbf{L}(t), \mathbf{F_C}(t))$ $P(\mathbf{S}(t)|\mathbf{L}(t), \mathbf{F_C}(t))$. The prior probability $P(\mathbf{S}(t)|\mathbf{L}(t), \mathbf{F_C}(t))$ is obtained as the result of the gist procedure. The likelihood $P(\mathbf{F_L}(t)|\mathbf{S}(t), \mathbf{L}(t), \mathbf{F_C}(t))$ is computed by resorting to the Self-resemblance algorithm [11].

The value map distribution $P(\mathbf{V}(t)|\mathbf{L}(t), \mathbf{F_L}(t), \mathbf{F_C}(t), \mathbf{O}(t) = text)$, Eq. 1b, is computed from $\widetilde{\mathbf{I}}$ as a rough, pre-attentive estimation of the probability of a location to contribute to a text / non text region. This is obtained by partitioning the low resolution image in 50×50 pixels square patches, and by using a probabilistic binary classifier, namely a Relevance Vector Machine [18], to assign each patch the probability of supporting text or non-text objects. RVM learning was performed off-line on a data set different from the test data set, and comprising 100 urban street view pictures created for training purposes.

Proto-object sampling, Eq.1c, is shaped as the sampling from a Beta-Bernoulli distribution describing the choice between $\mathbf{V}(t)$ as opposed to $\mathbf{S}(t)$ of serving as the foraging landscape, where the prior distribution on the choice parameter is set as a function of the current oculomotor action $\mathbf{A}(t)$: if the motor action corresponds to local foraging (micro-saccades and fixation), then $\mathbf{S}(t)$ is used; otherwise, for long gaze shifts (saccades), $\mathbf{V}(t)$ becomes the actual constraining landscape. Once $\mathbf{W}(t)$ has been sampled, the subsequent sampling steps specified by Eqs. 2a and 2b can be realized through the constrained random walk on the proto-object landscape, via the composite information foraging mechanism described[1] in [1], [2], where the local information feeding stage (fixation)

[1] Matlab code can be freely downloaded from `http://homes.di.unimi.it/~boccignone/GiuseppeBoccignone_webpage/Stochastic_Gaze_Shift.html`

is complemented by site depletion before saccading, in order to implement the Inhibition-of-Return mechanism.

To assess the plausibility of the model generated oculomotor behaviors, eye-tracking experiments have been conducted on 6 subjects using a video-based SMI RED eye tracker (120Hz. sampling rate). Each subject, in a contact-free setup, was asked to look at pictures presented on a 1024×768 pixels screen, in free viewing conditions. Stimulus luminance was linear in pixel values. The distance between the screen and the subject was 70 cm. Pictures were presented in randomized order and each picture was shown for 5 seconds.

Examples qualitatively comparing humans' and model's scanpaths, and representative of results achieved on the test data-set, are provided in Fig. 2.

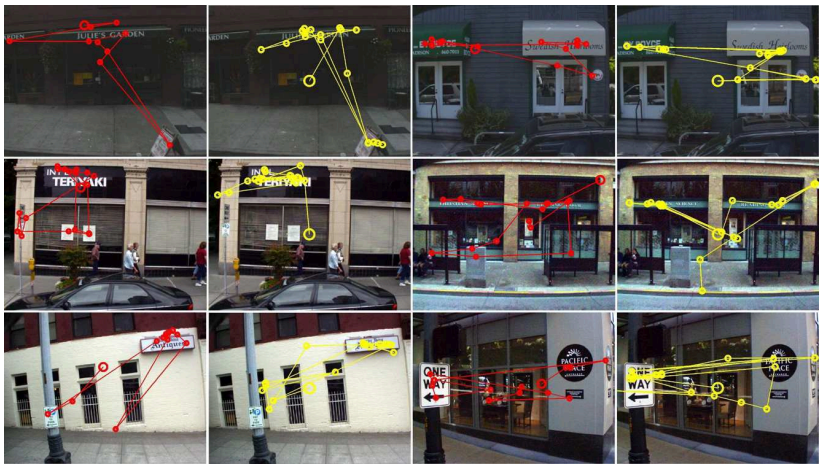

Fig. 2. Left to right, each pair of images shows the human scanpath recorded by eye trackers (left, in red) and the model generated scanpath (right, in yellow)

More quantitatively, we have compared the gaze shift amplitude empirical distributions of model scanpaths with those estimated from the eye-tracked human observers [17],[16]. In 71 % of cases the null hypothesis that both distributions came from the same distribution was accepted by the two-sample Kolmogorov-Smirnov test ($\alpha = 0.05$).

References

1. Boccignone, G., Ferraro, M.: Feed and fly control of visual scanpaths for foveation image processing. Annals of Telecommunications pp. 1–17 (2012), `http://dx.doi.org/10.1007/s12243-012-0316-9`
2. Boccignone, G., Ferraro, M.: Gaze shift behavior on video as composite information foraging. Signal Processing: Image Communication pp. 1–18 (2012), `http://dx.doi.org/10.1016/j.image.2012.07.002`

3. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence In Press (2012), http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.89
4. Cerf, M., Frady, E., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision 9(12) (2009)
5. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye tracking: a comprehensive guide to methods and measures. Oxford University Press, Oxford, UK (2011)
6. Karaoglu, S., van Gemert, J., Gevers, T.: Object reading: Text recognition for object recognition. In: Proc. ECCV 2012 Workshop on Information Fusion in Computer Vision for Concept Recognition (2012)
7. Meng, Q., Song, Y.: Text detection in natural scenes with salient region. In: Proceedings of the 2012 10th IAPR International Workshop on Document Analysis Systems. pp. 384–388. IEEE Computer Society (2012)
8. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) Computer Vision ACCV 2010, Lecture Notes in Computer Science, vol. 6494, pp. 770–783. Springer Berlin / Heidelberg (2011)
9. Rensink, R.A.: The dynamic representation of scenes. Vis. Cognit. 7, 17–42 (2000)
10. Schütz, A., Braun, D., Gegenfurtner, K.: Eye movements and perception: A selective review. Journal of Vision 11(5) (2011)
11. Seo, H., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. Journal of Vision 9(12), 1–27 (2009)
12. Shahab, A., Shafait, F., Dengel, A.: Bayesian approach to photo time-stamp recognition. In: Proc. International Conference on Document Analysis and Recognition (ICDAR 2011). pp. 1039–1043. IEEE (2011)
13. Shahab, A., Shafait, F., Dengel, A., Uchida, S.: How salient is scene text? In: Proc. 10th IAPR International Workshop onDocument Analysis Systems (DAS, 2012). pp. 317–321. IEEE (2012)
14. Sumathi, C., Santhanam, T., Priya, N.: Techniques and challenges of automatic text extraction in complex images: a survey. Journal of Theoretical and Applied Information Technology 35(2) (2012)
15. Sun, Q., Lu, Y., Sun, S.: A visual attention based approach to text extraction. In: 20th International Conference onPattern Recognition (ICPR 2010). pp. 3991–3995. IEEE (2010)
16. Tatler, B., Hayhoe, M., Land, M., Ballard, D.: Eye guidance in natural vision: Reinterpreting salience. Journal of vision 11(5) (2011)
17. Tatler, B., Vincent, B.: The prominence of behavioural biases in eye guidance. Visual Cognition 17(6-7), 1029–1054 (2009)
18. Tipping, M.: Sparse bayesian learning and the relevance vector machine. The Journal of Machine Learning Research 1, 211–244 (2001)
19. Torralba, A., Oliva, A., Castelhano, M., Henderson, J.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review 113(4), 766 (2006)
20. Wang, H., Pomplun, M.: The attraction of visual attention to texts in real-world scenes. Journal of Vision 12(6) (2012)
21. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 1457–1464. IEEE (2011)