The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Using Text-Spotting to Query the World

Ingmar Posner and Peter Corke and Paul Newman.

Abstract— **The world we live in is labeled extensively for the benefit of humans. Yet, to date, robots have made little use of human readable text as a resource. In this paper we aim to draw attention to text as a readily available source of semantic information in robotics by implementing a system which allows robots to read visible text in natural scene images and to use this knowledge to interpret the content of a given scene. The reliable detection and parsing of text in natural scene images is an active area of research and remains a non-trivial problem. We extend a commonly adopted approach based on boosting for the detection and optical character recognition (OCR) for the parsing of text by a probabilistic error correction scheme incorporating a sensor-model for our pipeline. In order to interpret the scene content we introduce a generative model which explains spotted text in terms of arbitrary search terms. This allows the robot to estimate the relevance of a given scene with respect to arbitrary queries such as, for example, whether it is looking at a bank or a restaurant. We present results from images recorded by a robot in a busy cityscape.**

I. Introduction

Text, by design, is a rich source of semantic information which often cannot be inferred otherwise from the current vantage point, or at all, using our senses alone. Human-readable text is plentiful in man-made environments. Outdoors, street signs, bus stops, and shop fronts all provide good quality text that is rich in information about function and location. Shop fronts are particularly rich in text which provides information about the nature of the shop and is potentially queryable using internet search resources to determine the shop location. Street signs can provide important navigational cues. Indoors, where GIS and/or GPS information may be denied or unavailable, text can provide similarly vital clues. Oftentimes, objects and places are labelled directly: key words like "push" or "pull" can be indicative of doors, areas are marked as "kitchen", and so on. However, despite its apparent utility, text has so far been largely ignored as a source of information for robots. In this paper we describe an approach to enable autonomous agents to leverage this valuable and under-exploited resource to determine the relevance of a given scene with respect to an arbitrary query. For example, a restaurant might (in the best case) be indicated by the observed word "restaurant", but it may also be indicated by synonyms such as "bistro" or words that denote the cuisine ("Chinese", "Thai") or the food specialty ("seafood", "pizza", "steak"). We describe a generative probabilistic model which explains spotted text with respect to a search term and thus enables a robot to establish a direct connection between a place in a map and an abstract semantic concept.

I. Posner and P. Newman are with the Mobile Robotics Group, Department of Engineering Science, Oxford University, UK, {ingmar, pnewman}@robots.ox.ac.uk

P. Corke is with the School of Engineering Systems, Queensland University of Technology, Australia, peter.corke@qut.edu.au

Fig. 1. A typical example output of our text-spotting pipeline. P-values indicate the value of the posterior probability $p(w|z)$. See Section II-D for details.

Fig. 2. The data acquisition robot used in this work. Images were captured using the Bumblebee camera mounted on a pan-tilt head.

The core of our system consists of a text-spotting engine which robustly detects and parses text in the environment (see, for example, Figure 1). Despite the long history of automatic text recognition, the application beyond printed documents remains an active research problem (e.g. [1]–[4]). The challenges with wild text include the lack of contrast between text and its background, the rich diversity of fonts and character sizes, highly variable horizontal and vertical alignment of characters and related words, and perspective distortion due to non fronto-parallel viewing.

The contributions of this work consist of a robotic system which exploits a valuable but thus far unused navigational and informational resource using vision and optical character

recognition (OCR). A generative model is introduced, which explains the subject of a scene in terms of detected text. The remainder of this section describes related prior work. The core components of the text-spotting engine, text detection and OCR are described in Section II. The generative probabilistic model used to select images relevant to arbitrary search terms is described in Section III. Experimental results are presented in Section IV. We conclude in Section V.

A. Related Work

The potential of exploiting human-readable text in robotics has been recognised by several researchers in the past. However, to the best of our knowledge no prior art exists where text-spotting in natural scene images has been implemented and deployed in a robotics context. The use of OCR with robots is suggested, but not implemented, in [5]–[7]. In [5] a small robot with onboard DSP-based computation is proposed that would read signs and licence plates. It is not clear how far this work has progressed. The authors of [6] discuss OCR and propose its application to robotic navigation. In [8] a book-manipulation robot uses OCR to confirm the title of the book to be taken from a shelf. The authors of [9] describe an indoor mobile robot that performs OCR, although the extracted text is not utilised. More recent work has explored the exploitation of direction signs in robotics. In [10] the authors approach this task using object recognition techniques predicated upon a prior knowledge of a set of signs of interest. Signs are recognised by virtue of the geometry of their constituent parts. Crucially, neither text detection, parsing nor understanding are brought to bear. In contrast, our work aims to enable the recognition and understanding of any text in a scene, which provides for a much broader spectrum of applications of which sign-following is but one. No prior knowledge of signs of interest is assumed.

An important part of our system is the extraction of text from natural scene images. This is an area of current research interest (e.g. [1]–[4]). ICDAR[1] has organised two competitions (2003 and 2005) for the robust detection of wild text based on a standard set of labelled images. The results are summarised in [11], [12]. Other non-document OCR applications include detecting text in television streams [13], licence plate recognition [14]–[16], and assistive devices for the visually impaired [17], [18].

II. The Text-Spotting Tool Chain

At the heart of our system lies a text-spotting engine. Commonly, this problem is decomposed into stages: the detection of text in the image, recognition of characters, and the grouping of characters into coherent units of text (such as words or sentences). With few exceptions (see, for example, [3]) these individual steps are considered independent, sequential processes and no information is shared between them. Our text-spotting implementation follows this classical approach to the problem. The principal elements are:

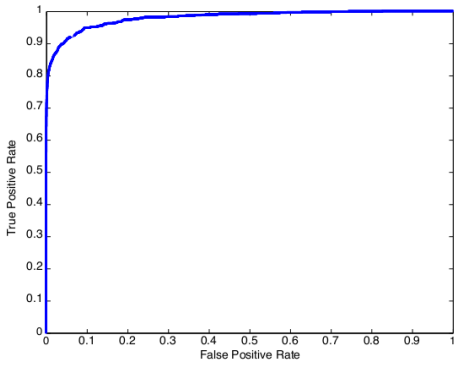


Fig. 3. Performance of a single boosted classifier after 1,000 rounds of training using both the training partition of the ICDAR data and the Weinman data.

1) Text detection. Determine regions of the input image that are likely to contain text.
2) Layout analysis. Text regions with similar sized characters that are close and aligned, horizontally or vertically, are merged.
3) Optical character recognition (OCR). Convert these image regions to character strings, typically words.
4) Text filtering and spelling correction. The output from the OCR stage is very noisy, often containing spurious characters and many character substitution errors.

A. Detecting Text in Natural Scene Images

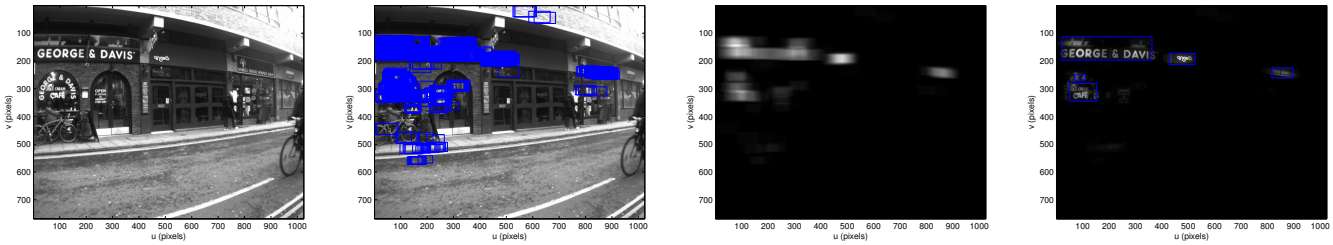
The aim of this stage is to efficiently detect instances of text in a given image. Boosting techniques [19] coupled with an attentional cascade, introduced in [20], provide a straightforward means to this end and have a successful track record in text detection [2], [13], [21]. In this work we apply GentleBoost [22] with the base classifiers consisting of decision stumps operating on a set of Haar-like features. These features are obtained by sliding predefined block patterns over an image and computing features as functions of statistics such as mean and variance of each of the individual blocks.

Chen et al [2] note that image gradient information captures a distinctive characteristic of text. We follow [21] in our selection of features and use feature channels based on x- and y-gradient and gradient magnitude in addition to mean and variance. We compute 22 features from each of five feature channels giving a total of 110 feature dimensions to be considered.

Two independent third-party data sets were employed for training of our text detector. The first dataset is provided publicly as part of the ICDAR 2003 challenge on robust reading and text locating[2]. It consists of a training and a test set each comprising 250 hand-labelled images drawn from indoor and outdoor environments. Since our focus is on outdoor applications we augmented these data with a subset of the data used by Weinman [3] comprising 300 images taken in outdoor urban settings and including a higher

[1]International Conference on Document Analysis and Recognition

[2]http://algoval.essex.ac.uk/icdar/RobustReading.html



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

Fig. 4. Stages of the text-spotting pipeline. (a) the original image, (b) with overlaid detection rectangles for scales 48, 57 and 69 (c) the text likelihood map, (d) the detected text regions for this scale range.

proportion of natural scene clutter as well as instances of multiple lines of text per label.

To investigate the efficacy of the features we trained a single monolithic boosted classifier based on 450 positive and 2,000 negative examples of text randomly sampled from a combination of the *training partition* of the ICDAR data and the complete Weinman data. The trained classifier was evaluated using a hold-out set of 996 positive and 38,000 negative data sampled from the same datasets. The classifier performance on the validation set after 1,000 rounds of training is presented in Figure 3. The number of training rounds was set arbitrarily large, designed to guarantee convergence to a stable validation error. Figure 3 indicates an adequate separation of the classes.

In order to provide an efficient classification framework with a suitably low false positive rate we deploy a cascade of boosted classifiers rather than a single monolithic one. The training was conducted using text regions randomly sampled from a combination of the *training partition* of the ICDAR data and the complete Weinman data. Each stage of the cascade was trained using 400 positive and 1,000 negative examples. The negatives were continuously sampled out of a pool of 35,000 data. The validation set consisted of 1,046 positive and 5,000 negative examples. The final output of the cascade yielded a detection rate of 79.4% while only 1.6 out of a thousand detections are spurious.

B. Region extraction

The output of the previous stage are lists of rectangles, one list for each scale, which are classified as containing text, see Figure 4(b). A typical image will have hundreds of rectangles at each of a number of scales. The rectangles are overlapping and at each scale we look for rectangles that have support, that is they overlap with at least N other rectangles (we use $N = 3$). It is highly unlikely that wild text will match the scale steps exactly so we consider the supported rectangles in a sliding window of M adjacent scales (we use $M = 3$). Each rectangle votes for the pixels that it contains and the votes are tallied in a voting array the same size as the original image, see Figure 4(c). The voting array is thresholded at 25% of the maximum value and bounding boxes for the regions are computed. The selected regions, at this scale, are shown in Figure 4(d).

Good bounding boxes are important for success in subsequent stages of the pipeline and, while our current simplistic approach to layout analysis allows for a reasonable number of recognitions, it often results in bounding boxes that are too tight or too loose.

C. Optical character recognition

Today OCR packages are very reliable for printed text which exhibits high contrast, simple background, uniformity in font and character size, and horizontal alignment of characters — characteristics not shared by wild text. We evaluated two open-source OCR packages: GOCR and Tesseract [23] and chose the latter. Tessearct deals well with skewed baselines which is advantageous when dealing with perspective distortion due to non fronto-parallel viewing.

The main mode of failure is misrecognition of characters and intercharacter spacing. Single character substitution errors are common (eg. zero for oh, one for ell, five for ess). Spaces can appear between adjacent characters, or spaces between words are sometimes not seen — both cases are problematic. The root cause is the wide range of fonts that are found in outdoor signage.

D. Probabilistic Error Correction

The output of the OCR engine can be improved considerably by constraining it to a set of meaningful words. A simple dictionary check would discard any word not found. This is unsatisfactory for the common case of single character substitution errors. Instead we use probabilistic inference over the true word present in the scene, w, given a possibly erroneous detection, z , $p(w|s)$.

Let $\mathcal{Z}$ denote the set of all possible OCR detections such that $z \in \mathcal{Z}$. Furthermore, let $\mathcal{V}$ denote the set of all terms in the English language such that $w \in \mathcal{V}$. We think of z as a noisy translation of some unknown generating word w. The posterior distribution over all words in the set $\mathcal{V}$ can be expressed as

$$p(w|z) = \frac{p(z|w)p(w)}{p(z)} \quad (1)$$

$$= \frac{p(z|w)p(w)}{\sum_{w \in \mathcal{V}} p(z|w)p(w)}. \quad (2)$$

Evaluation of this expression requires the determination of $p(z|w)$ — the distribution of text detections given a correctly spelt and complete observation-generating word w. Intuitively, the "closer" z is to a word, the more likely that word is to explain the detection. We use the Levenshtein edit

Fig. 5. Examples of wild text found by the robot. The annotations are the raw Tesseract output without any error correction applied.

Fig. 6. Examples of wild text found by the robot after error correction. P-values indicate the value of the posterior probability $p(w|z)$.

Fig. 7. Examples of incorrect detections of wild text due to *texture words*. P-values indicate the value of the posterior probability $p(w|z)$.

distance $\phi(z, w)$ to capture this sense of distance between detected text z and word w and write

$$p(z|w) = \alpha \quad e^{-\alpha\phi(z,w)}. \tag{3}$$

Here α is a free parameter encoding the accuracy of the text detection system. For the results presented in this paper α was set by hand using random spelling mistakes. No data contained either in the training or test sets were used. In future work we intend to learn this parameter from a large training set. Finally, Equation 2 requires the specification of the prior probability of a given word w occurring in a scene. We use word frequencies obtained from the British National Corpus [24], a collection of approximately 100×10^6 words encompassing ca. $130,000$ unique terms.

III. Relating Text to Subjects

We now derive a model which explains the subject of an image in terms of the detected text it contains. Importantly, because of the use of a large corpus of text, we need not limit ourselves to a finite set of subjects chosen *a-priori*. We apply this model to execute subject searches in which a robot will return a list of places and views which relate semantically to the search term. Specifically, we require that searching for the subject *mobile phone* would return geographic coordinates of views containing text like "nokia", "samsung", "broadband", etc. — evidence that the scene captured in an image has something to do with mobile phones. Note that we do not expect or demand flawless text detection since, due to the detector model introduced in Section II-D, we can handle incorrect detections like "nqkio", "smssag", or "roodbond".

Given a corpus of images, let $\mathcal{Z}$ denote the set of all detections of text throughout the corpus. Furthermore, let $\mathcal{S}$ denote the set of all possible scene subjects. Our goal is to explain a particular subject term $s \in \mathcal{S}$ with respect to a given particular text detection $z \in \mathcal{Z}$. In a probabilistic sense we can express this as the task of finding the posterior probability of the search term given the detection

$$p(s|z) = \frac{p(z|s)p(s)}{p(z)}. \tag{4}$$

The partition function $p(z)$ is the probability distribution over all possible detections and can be expanded in terms of a marginalization over subject terms of the joint distribution $p(z, s)$. If we take all subjects to be equally likely, Equation 4 reduces to

$$p(s|z) = \frac{p(z|s)p(s)}{\sum_{s\in\mathcal{S}} p(z|s)p(s)} \tag{5}$$

$$= \frac{p(z|s)}{\sum_{s\in\mathcal{S}} p(z|s)}. \tag{6}$$

The term $p(z|s)$ is the likelihood of the OCR returning a string z when the underlying scene subject is s. We leverage the detector model introduced in Equation 3 to account for the noise in the detection and parsing of text. We introduce a layer of now hidden variables $w \in \mathcal{V}$, where once again $\mathcal{V}$ denotes the vocabulary of the English language and each w

lunch *taxi* *bank*

Fig. 8. Images related to the query subjects *lunch*, *taxi* and *bank*.

is a word. By marginalising over the $\mathcal{V}$ our desired likelihood term $p(z|s)$ can be expanded in terms of the hidden words

$$p(z|s) = \sum_{w \in \mathcal{V}} p(z|w, s)p(w|s). \quad (7)$$

If we take detection noise to be independent of subject, we can express the likelihood $p(z|s)$ as

$$p(z|s) = \sum_{w \in \mathcal{V}} p(z|w)p(w|s), \quad (8)$$

which requires the determination of the detector model $p(z|w)$. The remaining term in Equation 8 is $p(w|s)$ — the probability of a bonafide word w occurring in a corpus of words on subject s. We assume an internet connected robot and launch a web search for the subject string s. The words in the returned documents are aggregated into a single *subject document*. For the results presented here we searched the websites of the BBC News, the New York Times and the Guardian Newspaper. The construction of the subject document allows $p(w|s)$ to be estimated directly by counting the number of times word w occurs.

IV. Experimental Results

We used the robot Marge, an iRobot ATRV-JR equipped with a variety of sensors (Figure 2). Images were captured with a Bumblebee stereo head that provides 1024×768 greyscale images with a $60\,\mathrm{deg}$ field of view. Only images from the left camera in the stereo pair are considered here.

Figures 5 - 7 show a small selection of typical results of applying our text-spotting pipeline to the collected dataset of 941 images[3]. Figure 5 presents the raw OCR output before error correction is applied. Note that a number of words are misspelt and that, for the middle two frames, the bounding box has truncated a word. Figure 6 shows the same scenes with successfully corrected words. Our system recovers some of the misspelt words or discards those that were truncated. As well as the extracted words the system provides a confidence level $p(w|z)$ — computed as per Equation 2 — as to how well the inferred word w explains the observation z . This posterior probability over generating words provides a natural and intuitive way of thresholding system output. Figures 6 - 8 only show detections with a confidence greater than 90%.

[3]Full-size versions of all the results presented here, an extended set of results, the labelled data used for evaluation (Figure 9) as well as other resources for text-spotting in robotics can be found at [http://www.robots.ox.ac.uk/ mobile/wikisite/pmwiki/pmwiki.php?n=Main.TextSpotting].

lunch		*taxi*		*bank*	
term	$p(s\|z)$	term	$p(s\|z)$	term	$p(s\|z)$
restaurant	**0.0186**	**telephone**	**0.0112**	**barclays**	**0.1131**
barclays	0.0052	queue	0.0092	george	0.0060
queue	0.0035	february	0.0051	street	0.0047
children	0.0033	street	0.0042	february	0.0043
keep	0.0032	over	0.0024	telephone	0.0041

TABLE I

The top 5 words extracted from the dataset ranked by likelihood. Words rendered in bold exceed the threshold.

The failure cases shown in Figure 7 provide examples of what we call *texture words*. In these cases, scene texture such as fences, vertical window edges, brickwork, architectural features and adornments, etc. elicit a positive response from the text detection stage and OCR zealously assigns characters — typically, letters from the set "ILETUCMWA".

We applied our subject-relevance model, querying in turn for the subjects *lunch, taxi* and *bank*. In the first instance the output of the system consists of a ranking of all the terms extracted from the corpus of images based on the posterior probability $p(s|z)$. The top five returns per subject are shown in Table I together with the probability of the topic given the observed word. In every case the system manages to successfully extrapolate from the query to semantically related terms. We apply a threshold at 1%. The images corresponding to our query terms are shown in Figure 8. The collection of a subject document relevant to a query incurs a computational expense. In practice this information can be cached and provides an ever growing body of knowledge for the robot. For particular problem domains the relevant subject documents can be pre-retrieved.

Figure 9 provides a quantitative performance analysis of our text-spotting engine when applied to a corpus of 300 city-centre images recorded with a hand-held camera. The corpus contains 3,935 manually labelled words. Depending on the threshold on the word posterior, our system achieves recall rates between 6-8.8% with precision in the range 60-94% while constantly outperforming the uncorrected OCR output. While adequate precision is achieved, the relative recall figures are low (on average one word per frame). However, text does not occur uniformly throughout an environment: some scenes contain no text at all while, in others, text is abundant. Our experiments indicate that the amount of text correctly retrieved in practice is sufficient to perform tasks such as the determination of scene relevance.

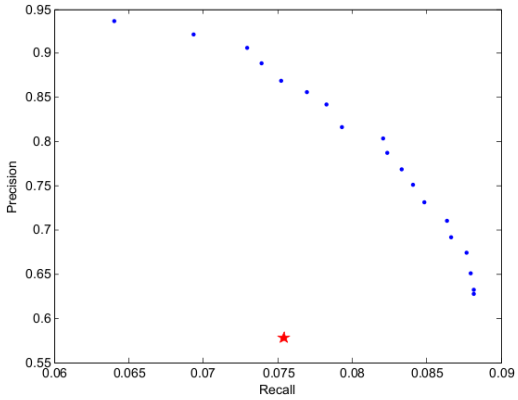


Fig. 9. Precision-recall curve for text retrieved using our system from a corpus of 300 city-centre images taken with a hand-held camera. The red star indicates performance without error correction. The blue dots indicate performance as the results are thresholded according to the posterior probability $p(w|z)$. Note the difference in scale ranges.

Our system does not presently return individual word boundaries but rather detections and parsings of *blocks* of text. Accordingly, 100% detection recall could be achieved trivially by drawing a bounding box around an entire image, though in the majority of these cases the OCR would fail. To exclude this as a factor in our analysis we have verified that in 75% of cases the areas of the detections are commensurate with those of the hand-labelled annotations. The overall largest bounding box recovered spanned ca. 60% of the image area.

V. Conclusions

We have described a robotic system that is capable of detecting and reading wild text, a rich source of semantic information indigenous to man-made environments. Our work demonstrates the potential of this resource for robotics applications by investigating query-based navigation where an arbitrary, abstract search term is related to relevant scene images and, by extension, places in a map.

This is early work in the field of literate robotics and our work is progressing on several fronts. Firstly, we are integrating the system presented here into a 3G-connected robot that can implement these techniques online. Secondly, we are constantly seeking to improve our text-spotting capability. In particular, we are investigating means to improve the performance of the OCR step, which is currently exhibiting a very high error rate. We are also investigating opportunities to improve performance by exploiting the contiguous nature of the workspaces traversed by robots and the additional sensor modalities available in this domain. Thirdly, we are investigating a variety of robotics applications including text-based localisation — where textual clues are used in conjunction with an internet-based geocoding service — as well as the integration of textual cues into object detectors.

VI. Acknowledgements

The authors would like to thank Jerod Weinman for making his data available for use in this work. The work reported here was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence and has partly been supported by the EC under FP7-231888-EUROPA.

References

[1] 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.

[2] X. Chen and A. L. Yuille, "Detecting and Reading Text in Natural Scenes," *Computer Vision and Pattern Recognition, IEEE Computer Society Conf. on*, vol. 2, pp. 366–373, 2004.

[3] J. J. Weinman, *Unified Detection and Recognition for Reading Text in Scene Images*. PhD thesis, University of Massachusetts Amherst, 2008.

[4] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform," in *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] G. Engel, D. Greve, J. Lubin, and E. Schwartz, "Space-variant active vision and visually guided robotics: Design and construction of a high-performance miniature vehicle," in *Intl. Conf. on Pattern Recognition*, pp. 487–487, IEEE Computer Society Press, 1994.

[6] M. Mirmehdi, P. Clark, and J. Lam, "A non-contact method of capturing low-resolution text for OCR," *Pattern Analysis & Applications*, vol. 6, no. 1, pp. 12–21, 2003.

[7] A. Carbone, A. Finzi, A. Orlandini, F. Pirri, and G. Ugazio, "Augmenting situation awareness via model-based control in rescue robots," in *Proc. of IROS-2005 Conf.*, Citeseer, 2005.

[8] R. Ramos-Garijo, M. Prats, P. Sanz, and A. Del Pobil, "An autonomous assistant robot for book manipulation in a library," in *Proceedings of the IEEE Intl. Conf. on Systems, Man, and Cybernetics*, pp. 3912–3917, 2003.

[9] J. Samarabandu and X. Liu, "An edge-based text region extraction algorithm for indoor mobile robot navigation," *Intl. Jrnl. of Signal Processing*, vol. 3, no. 4, pp. 273–280, 2006.

[10] J. Maye, L. Spinello, R. Triebel, and R. Siegwart, "Inferring the semantics of direction signs in public places," in *Proc. of The IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2010.

[11] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proceedings of the Seventh Intl. Conf. on Document Analysis and Recognition*, vol. 2, pp. 682–687, Citeseer, 2003.

[12] S. Lucas, "ICDAR 2005 text locating competition results," in *Proceedings of the Eighth Intl. Conf. on Document Analysis and Recognition, ICDAR05*, pp. 80–84, Citeseer, 2005.

[13] M.Lalonde and L. Gagnon, "Key-text spotting in documentary videos using adaboost," in *Proceedings of SPIE*, vol. 6064, pp. 507–514, 2006.

[14] Y. H. T. Wing Teng Ho, Hao Wooi Lim, "Two-stage licence plate detection using gentle adaboost," in *First Asian Conf. on Intelligent Information and Database Systems*, 2009.

[15] N. Ben-Haim, "Task specific image text recognition," Master's thesis, University of California, San Diego, 2008.

[16] L. Dlagnekov, "Video-based car surveillance: License plate, make, and model recognition," Master's thesis, University of California, San Diego, 2005.

[17] X. Chen and A. Yuille, "A time efficient cascade for realtime object detection: with applications for the visually impaired," in *Proceedings of the CVAVI05, IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, 2005.

[18] S. Hanif, L. Prevost, and P. Negri, "A cascade detector for text detection in natural scene images," in *Pattern Recognition, 2008. ICPR 2008. 19th Intl. Conf. on*, pp. 1–4, 2008.

[19] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting," *Jrnl. of Computer and System Sciences*, vol. 1, no. 55, pp. 119–139, 1997.

[20] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Intl. Jrnl. of Computer Vision*, vol. 2, no. 57, pp. 137–154, 2004.

[21] S. Escalera, X. Baró, J. Vitrià, and P. Radeva, "Text Detection in Urban Scenes," in *Proc. Conf. on Artificial Intelligence Research and Development*, pp. 35–44, 2009.

[22] J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," *Annals of Statistics*, vol. 28, 1998.

[23] R. Smith, "An overview of the Tesseract OCR engine," in *In Proc. of Intl. Conf. Document Analysis and Recognition (ICDAR)*, vol. 2, pp. 629–633, 2007.

[24] J. H. Clear, "The British national corpus," in *The digital word: text-based computing in the humanities*, pp. 163–187, MIT Press, 1993.