DELIVERABLE REPORT D4.3.2
"Mobile Text Detection and Recognition"

collaborative project

MASELTOV
Mobile Assistance for Social Inclusion and Empowerment of Immigrants with Persuasive Learning Technologies and Social Network Services

Grant Agreement No. 288587 / ICT for Inclusion

project co-funded by the
European Commission
Information Society and Media Directorate-General
Information and Communication Technologies
Seventh Framework Programme (2007-2013)

Due date of deliverable:	Dec. 31, 2013 (month 24)
Actual submission date:	July 23, 2014 (Revision 1)
Start date of project:	Jan 1, 2012
Duration:	36 months

Work package	**WP 4 – Multisensory Context Awareness**
Task	**4.3.2**
Lead contractor for this deliverable	**CTU**
Editor	**Lukas Neumann (CTU)**
Authors	**Lukas Neumann (CTU)**
Quality reviewer	**Nicoletta Bersia (TI), Graziella Spinelli (TI)**

<table>
<tr><td colspan="3">Project co-funded by the European Commission within the Seventh Framework Programme (2007–2013)</td></tr>
<tr><td colspan="3">Dissemination Level</td></tr>
<tr><td><b>PU</b></td><td>Public</td><td>X</td></tr>
<tr><td><b>PP</b></td><td>Restricted to other programme participants (including the Commission Services)</td><td></td></tr>
<tr><td><b>RE</b></td><td>Restricted to a group specified by the consortium (including the Commission Services)</td><td></td></tr>
<tr><td><b>CO</b></td><td>Confidential, only for members of the consortium (including the Commission Services)</td><td></td></tr>
</table>

MASELTOV partner			organisation name	country code
01	JR		JOANNEUM RESEARCH FORSCHUNGSGESELLSCHAFT MBH	AT
02	ATE		AUSTRIAN INSTITUTE OF TECHNOLOGY	AT
03	AIT		RESEARCH AND EDUCATION LABORATORY IN INFORMATION TECHNOLOGIES	EL
04	UOC		FUNDACIO PER A LA UNIVERSITAT OBERTA DE CATALUNYA	ES
05	OU		THE OPEN UNIVERSITY	UK
06	COV		COVENTRY UNIVERSITY	UK
07	CTU		CESKE VYSOKE UCENI TECHNICKE V PRAZE	CZ
08	FHJ		FH JOANNEUM GESELLSCHAFT M.B.H.	AT
09	TI		TELECOM ITALIA S.p.A	IT
10	FLU		FLUIDTIME DATA SERVICES GMBH	AT
12	FUN		FUNDACION DESARROLLO SOSTENIDO	ES
13	DAN		VEREIN DANAIDA	AT
14	MRC		THE MIGRANTS' RESOURCE CENTRE	UK
15	PP		PEARSON PUBLISHING	UK

CONTENT

1. Executive Summary 4
2. Text lens application for immigrants 5
2.1 Introduction 5
2.2 User interface 6
2.3 Technical description 8
3. Textspotter algorithm 11
3.1 Introduction 11
3.2 Our work 11
4. Summary 14
5. REFERENCES 14
APPENDIX 1: Scene Text Localization and Recognition with Oriented Stroke Detection 15
APPENDIX 2: On Combining Multiple Segmentations in Scene Text Recognition 16

1. EXECUTIVE SUMMARY

A mobile text detection and recognition application processes an image (or a video) taken by a mobile phone. It finds all areas in the image that would be considered as text by a human, marks boundaries of the areas (usually by rectangular bounding boxes) and outputs a sequence of (Unicode) characters associated with its content. This digital text format can be then further processed by any application on the mobile phone.

2. TEXT LENS APPLICATION FOR IMMIGRANTS

2.1 INTRODUCTION

Text Lens is a mobile application for Android devices which uses mobile phone camera to capture images and then detects and recognizes text using the scene text recognition method TextSpotter (see Section 3).

Immigrants encounter text which they do not understand in their daily life (see Figure 1). When facing such situation, the Text Lens application helps them by providing instant translation of the problematic text.

Unlike traditional OCR mobile applications (e.g. ABBYY Mobile OCR, Google Docs OCR App), which typically require text to be written on solid background and the camera pointing directly at the text, Text Lens is able to "read" text captured by a standard mobile phone camera, it can deal with situations when the text only occupies small part of the image (thanks to automatic text localization phase, TextLens is able to recognize text even in distance, e.g. a name of a shop), when the text is prospectively distorted or slanted (a user is not required to always take a "nice" picture with fronto-parallel view on the text) and it supports complex backgrounds and large variety of fonts (user can for example translate a sign written on a brick wall).

Text Lens runs fully on the mobile device without any need for an Internet connection (the only time Internet connection is required is when offline dictionary is not sufficient and the user requires complex translation by an online service – see below).

The user also has the ability to manually correct the recognized text (we call this step "text annotation") and post the annotated text to other MASELTOV applications.

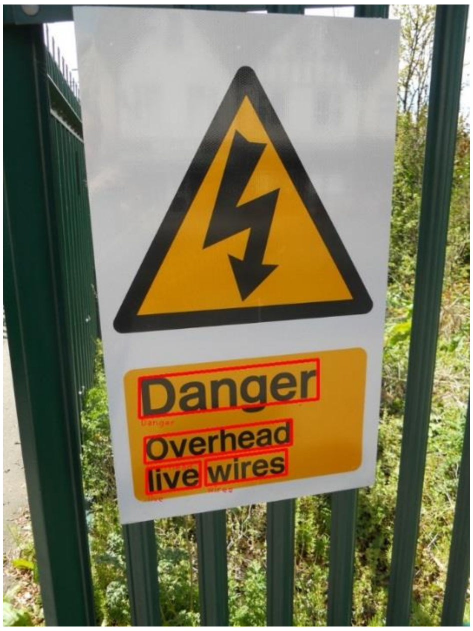


Figure 1: Immigrants encounter text which they do not understand in their daily life

2.2 USER INTERFACE

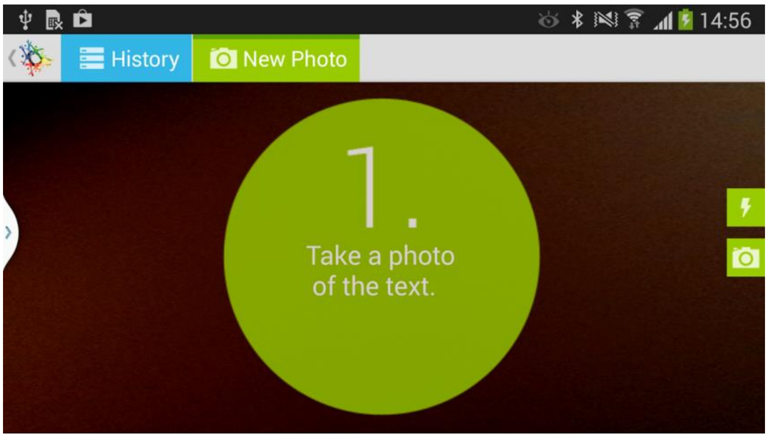
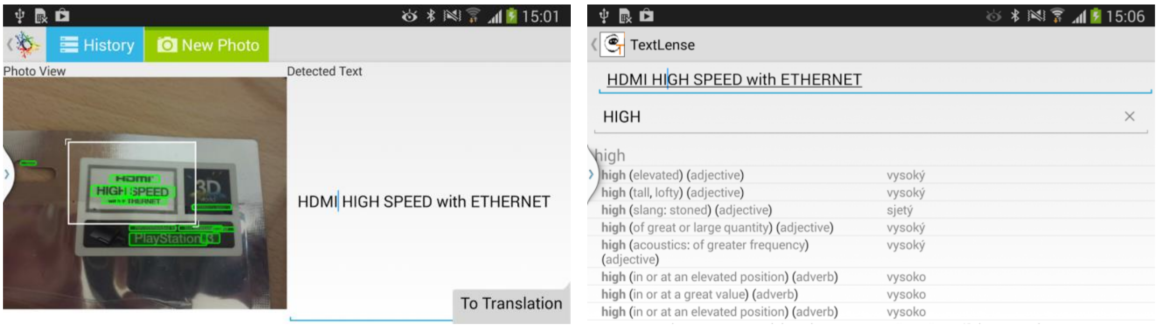


Figure 2: Text Lens mobile application. User takes a photo by simply taping the display or clicking the Capture button (far right). It is also possible to display previously taken photos (History button) or to take a new photo in case the current one is not a good one (New Photo button). User can also enable flash by clicking the Flash button.

A user launches Text Lens in MASELTOV app on his/her mobile phone, points the mobile phone camera on the text which he/she does not understand and takes a picture of the text. The text is automatically detected and localized using the text recognition method running directly on the device (see Section 3).

The user then can translate the detected text using a built-in dictionary (see Figure 3), get a translation of the whole text by an online translation service (Microsoft Bing Translate) or use the detected text with other MASELTOV Apps (e.g. post the text to the MASELTOV Forum requesting help). Additionally, Text Lens automatically sends all detected text to the Recommendation Engine (see Section 2.3), so that user can be automatically offered related content based on a set of predefined rules in the Recommendation Engine (e.g. a user takes a picture of an entrance sign or a form at a doctor and he is automatically offered information about health care system in the country). If the text is recognized incorrectly, the user has the ability to correctly annotate the text before it is further processed.



HDMI HIGH SPEED with ETHERNET

HDMI HIGH SPEED with ETHERNET

HIGH

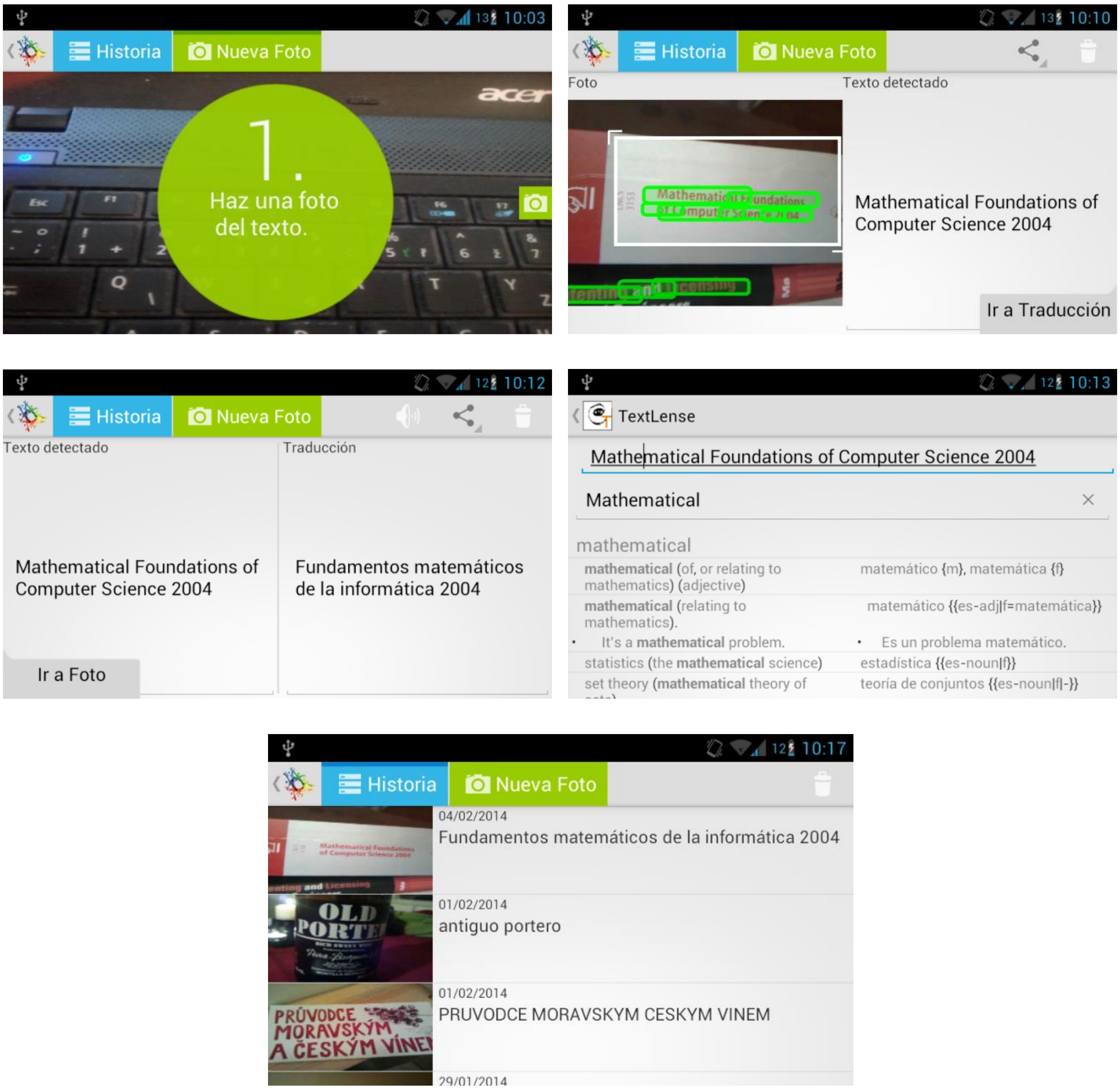
high

high (elevated) (adjective)	vysoký
high (tall, lofty) (adjective)	vysoký
high (slang: stoned) (adjective)	sjetý
high (of great or large quantity) (adjective)	vysoký
high (acoustics: of greater frequency) (adjective)	vysoký
high (in or at an elevated position) (adverb)	vysoko
high (in or at a great value) (adverb)	vysoko
high (in or at an elevated position) (adverb)	vysoko

Figure 3: Text detection and translation by Text Lens

The user interface is available in **English, Spanish, Turkish and Arabic**. Text Lens can detect English, German and Spanish texts and the translation is available to all languages supported by the UI, i.e. English, Spanish, Turkish and Arabic.

For more details about the functionalities of Text Lens user interface please refer to D9.2.3 („Iterative Evaluation of User Interfaces”).



Mathematical Foundations of Computer Science 2004

Fundamentos matemáticos de la informática 2004

Mathematical Foundations of Computer Science 2004

Mathematical

mathematical

mathematical (of, or relating to mathematics) (adjective)	matemático {m}, matemática {f}
mathematical (relating to mathematics).	matemático {{es-adj|f=matemática}}
• It's a **mathematical** problem.	• Es un problema matemático.
statistics (the **mathematical** science)	estadística {{es-noun|f}}
set theory (**mathematical** theory of sets)	teoría de conjuntos {{es-noun|f|-}}

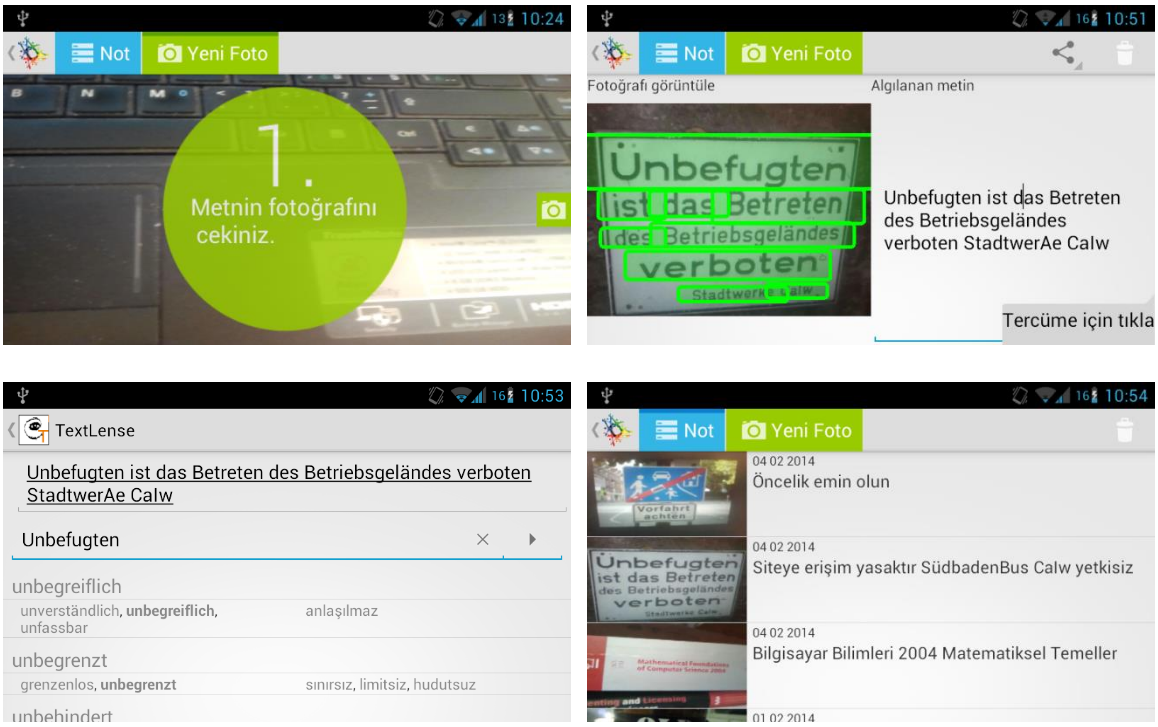
04/02/2014
Fundamentos matemáticos de la informática 2004

01/02/2014
antiguo portero

01/02/2014
PRUVODCE MORAVSKYM CESKYM VINEM

29/01/2014

Figure 4: English text recognition, its translation to Spanish and the History of translated texts



Unbefugten ist das Betreten des Betriebsgeländes verboten StadtwerAe Calw

Unbefugten

unbegreiflich	
unverständlich, **unbegreiflich**, unfassbar	anlaşılmaz
unbegrenzt	
grenzenlos, **unbegrenzt**	sınırsız, limitsiz, hudutsuz
unbehindert	

04 02 2014
Öncelik emin olun

04 02 2014
Siteye erişim yasaktır SüdbadenBus Calw yetkisiz

04 02 2014
Bilgisayar Bilimleri 2004 Matematiksel Temeller

01 02 2014

Figure 5: German text recognition, its translation to Turkish and the History of translated texts

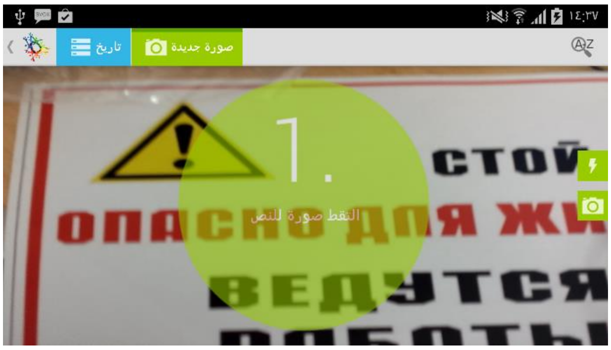


Figure 6: Text Lens in Arabic

2.3 TECHNICAL DESCRIPTION

TextLens is a stand-alone Android application, which consists of the following components (see Figure 7):

- **Main UI** – User Interface implemented in Java based on mock-up designs which were verified by CURE, the UI follows the MASELTOV App look and feel.
- **Offline dictionary** – contains dictionaries for all supported language combinations (translation from English, German and Spanish to English, Spanish, Turkish and Arabic), QuickDic Offline Dictionary (open-source) is used.
- **Machine Translation Library** – provides a unified API for text translation using an online service; it supports Bing Translator and Google Translate API, in current version Bing is used.
- **TextSpotterJava** – Java wrapper around the TextSpotter algorithm (see Section 3), which itself is implemented in C++.

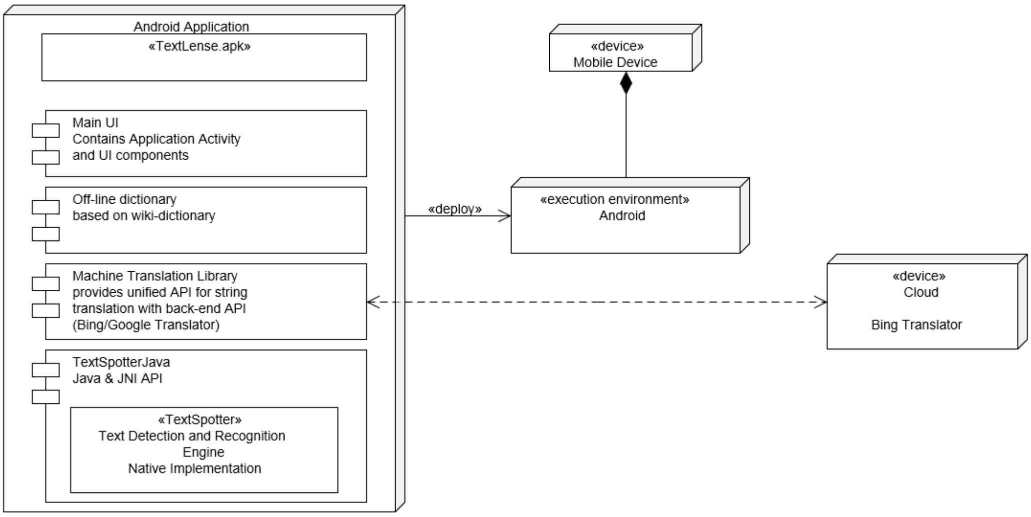


Figure 7: Text Lens component diagram

Text Lens sends the following Android events to other components on the mobile device (see Figure 8):

- **Text Detection** `("detectedText", <user corrected text in the image>)` – notifies the Recommendation Engine that certain text was detected.
- **Text Lens Usage** `("duration", <duration>)` – issued when user stops using TextLens so that user activity with TextLens can be monitored.
- **“Share” User Action** – a standard Android action which allows sharing text with all registered applications on the device; MASELTOV Forum is registered for this action, as well as other standard applications (E-mail, SMS, …).

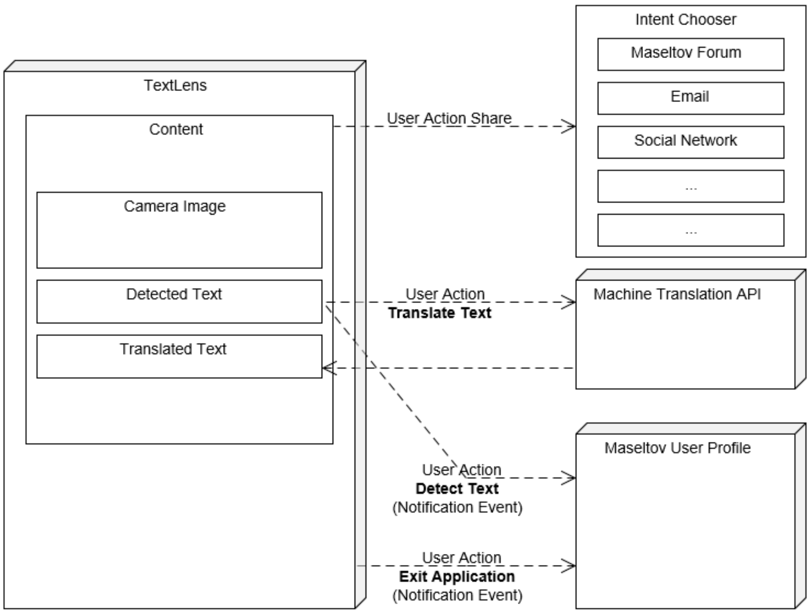


Figure 8: Text Lens data flow diagram

Additionally, if requested by the user the Text Lens application sends on-line **"Translate Text"** requests for online text translation. In current implementation, Microsoft Bing API (http://www.bing.com/dev/en-us/dev-center) is used for this purpose, however Google Translate API (https://developers.google.com/translate/) works in a very similar manner.

The **"Translate Text"** request is a standard HTTP request (where the text to be translated is submitted in the query parameters) and the results are obtained in a simple JSON string format. This minimizes data transfer overhead and the amount of data that has to be transferred over the mobile network. The use of JSON as a data exchange format is determined by the API publisher.

A sample communication to translate the text "General Practitioner" from English to Spanish is listed below:

Request

```
GET /V2/Ajax.svc/Translate?&from=en&to=de&text=general+practitioner%21
HTTP/1.1
Content-Type: text/plain; charset=UTF-8
Accept-Charset: UTF-8
Authorization:
User-Agent: Java/1.7.0_55
Host: api.microsofttranslator.com
Accept: text/html, image/gif, image/jpeg, *; q=.2, */*; q=.2
Connection: keep-alive
```

Response

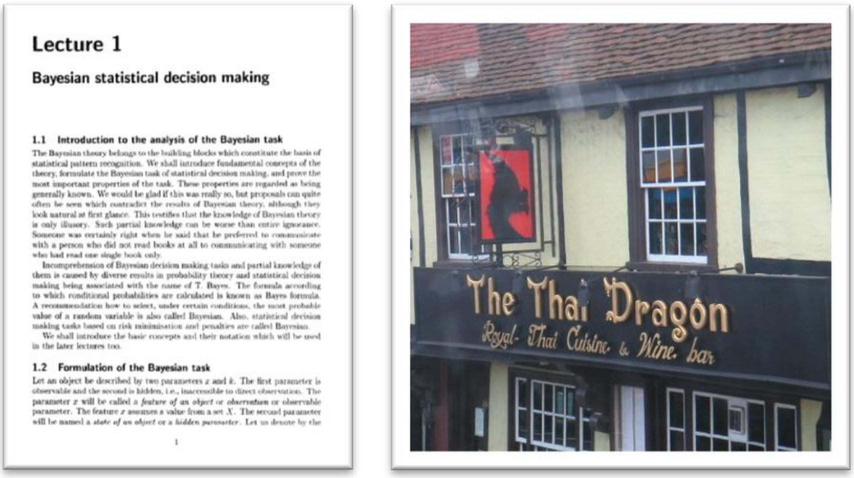
```
HTTP/1.1 200 OK
Cache-Control: no-cache
Pragma: no-cache
Content-Length: 23
Content-Type: application/x-javascript; charset=utf-8
Expires: -1
X-MS-Trans-Info: s=85005
Date: Fri, 13 Jun 2014 14:13:45 GMT

"médico general"
```

3. TEXTSPOTTER ALGORITHM

3.1 INTRODUCTION

Scene text localization and recognition (also known as *text localization and recognition in real-world images*, *nature scene OCR* or *text-in-the-wild problem*) is an open computer vision problem, unlike printed document recognition (OCR) where state-of-the-art systems are able to recognize correctly more than 99% of characters (see Figure 9). Factors contributing to the complexity of the problem include: non-uniform background, noise, blur, reflections, the need for compensation of perspective effects (for documents, rotation or rotation and scaling is sufficient). Moreover real-world texts are often short snippets written in different fonts and languages, text alignment does not follow strict rules of printed documents and many words are proper names which prevents an effective use of a dictionary.



Lecture 1

Bayesian statistical decision making

1.1 Introduction to the analysis of the Bayesian task

The Bayesian theory belongs to the building blocks which constitute the basis of statistical pattern recognition. We shall introduce fundamental concepts of the theory, formulate the Bayesian task of statistical decision making, and prove the most important properties of the task. These properties are regarded as being generally known. We would be glad if this was really so, but proposals can quite often be seen which contradict the results of Bayesian theory, although they look natural at first glance. This testifies that the knowledge of Bayesian theory is only illusory. Such partial knowledge can be worse than entire ignorance. Someone was certainly right when he said that he preferred to communicate with a person who did not read books at all to communicating with someone who had read one single book only.

Incomprehension of Bayesian decision making tasks and partial knowledge of them is caused by diverse results in probability theory and statistical decision making being associated with the name of T. Bayes. The formula according to which conditional probabilities are calculated is known as Bayes formula. A recommendation how to select, under certain conditions, the most probable value of a random variable is also called Bayesian. Also, statistical decision making tasks based on risk minimisation and penalties are called Bayesian.

We shall introduce the basic concepts and their notation which will be used in the later lectures too.

1.2 Formulation of the Bayesian task

Let an object be described by two parameters x and k. The first parameter is observable and the second is hidden, i.e., inaccessible to direct observation. The parameter x will be called *a feature of an object* or *observation* or observable parameter. The feature x assumes a value from a set X. The second parameter will be named a *state of an object* or a *hidden parameter*. Let us denote by the

1

Figure 9: Comparison of printed document OCR (left) and scene text recognition (right).

3.2 OUR WORK

As part of the deliverable D4.3.1, we proposed an end-to-end real-time scene text localization and recognition method TextSpotter [1, 2, 3], which achieves state-of-the-art results on standard datasets (we consider a text recognition method real-time if the processing time is comparable with the time it would take a human to read the text).

In the second year (deliverable D4.3.2) the method was further improved to achieve higher recognition accuracy and faster processing times.

In [4] (see Appendix 1) we proposed a novel approach for character detection and recognition which combines the advantages of sliding-window and connected component methods. Characters are detected and recognized as image regions which contain strokes of specific orientations in a specific relative position, where the strokes are efficiently detected by convolving the image gradient field with a set of oriented bar filters. The representation is robust to shift at the stroke level, which makes it less sensitive to intra-class variations and the noise induced by normalizing character size and positioning.

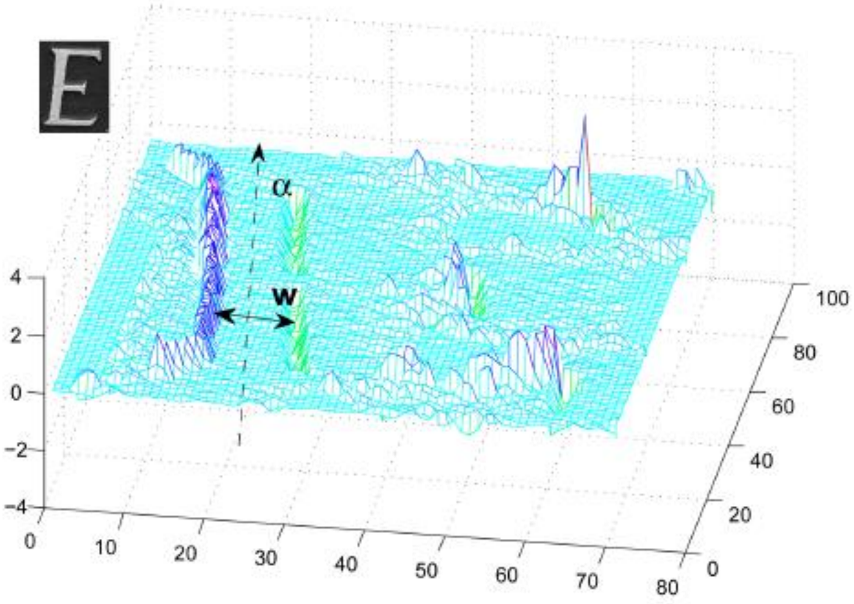


Figure 10: A stroke of direction is detected as two opposing ridges in the gradient (approximately) perpendicular to the stroke direction.

As a first contribution, we introduced a novel approach for character detection which combines the advantages of sliding-window and connected component methods. In the proposed method, the detected strokes induce the set of rectangles to be classified, which reduces the number of rectangles by three orders of magnitude when compared to the standard sliding-window methods. From the complexity perspective this makes the proposed method competitive with the methods based on connected components, but at the same time the robustness against noise and blur is maintained and the assumption that a character is a connected component is dropped, which allows for detection of joint or disconnected characters.

As a second contribution, a novel character representation efficiently calculated from the values obtained in the stroke detection phase was introduced. The representation is robust to shift at the stroke level, which makes it less sensitive to intra-class variations and the noise induced by normalizing character's size and positioning. The effectiveness of the representation is demonstrated by the results achieved in classification of real-world characters using a linear (approximative) nearest-neighbor classifier trained on synthetic data in a plain form (i.e. without any blurring or distortions). Additionally, the representation allows for efficient detection of rotated characters, as only a permutation of the feature vector is required (see Figure 11).

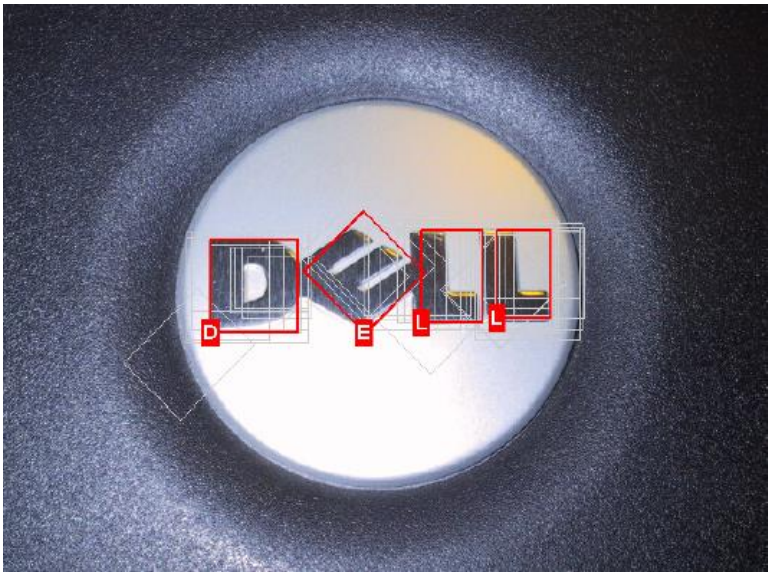


Figure 11: Character detection and recognition of connected and rotated characters. The character representation allows efficient detection of rotated characters, as only a permutation of the feature vector is required

In [5] (see Appendix 2) we demonstrated that keeping multiple segmentations of each character until the very last stage of the processing when the context of each character in a text line is known is beneficial for overall accuracy and we proposed an efficient algorithm for selection of character segmentations minimizing a global criterion. Additionally, we showed that, despite using theoretically scale-invariant methods, operating on a coarse Gaussian scale space pyramid yields improved results as many typographical artefacts (e.g. joint letters, characters consisting of small regions) are eliminated (see Figure 12).

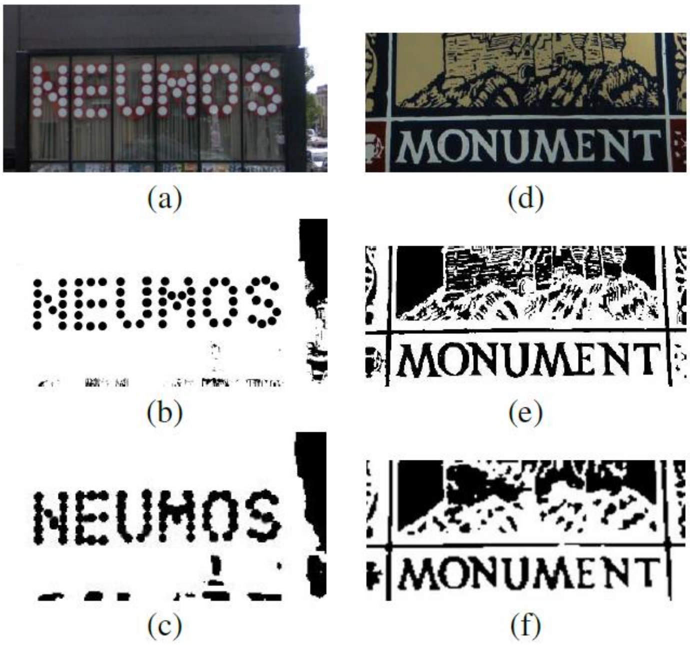


Figure 12: Processing with a Gaussian pyramid. Characters formed of multiple small regions (a,b) merge together and a single region corresponds to a single character (c). A single region which corresponds to characters ``ME'' (d) for which there does not exist any threshold in the original image (e) is broken into two and serifs are eliminated (f)

4. SUMMARY

A scene text detection and recognition method was proposed and successfully implemented as Text Lens mobile application for Android devices.

Thanks to such functionality, a user is able to quickly translate unknown text (such as signs, shop names, forms, menus in restaurants, etc.), which helps him/her with orientation in unknown environments.

The future work includes improvements of the speed and accuracy of the text recognition algorithm (see Section 3) and integration with other MASELTOV modules (e.g. integration with user profile to retrieve language preferences).

5. REFERENCES

[1] L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In ACCV 2010, volume IV of LNCS 6495, pages 2067-2078, November 2010.
[2] L. Neumann and J. Matas. Text localization in real-world images using efficiently pruned exhaustive search. In ICDAR 2011, pages 687-691, 2011.
[3] L. Neumann and J. Matas. Real-time scene text localization and recognition. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 3538-3545, 2012.
[4] L. Neumann and J. Matas. Scene Text Localization and Recognition with Oriented Stroke Detection, ICCV 2013, Sydney, Australia
[5] L. Neumann and J. Matas. On Combining Multiple Segmentations in Scene Text Recognition, ICDAR 2013, Washington D.C., USA

APPENDIX 1: SCENE TEXT LOCALIZATION AND RECOGNITION WITH ORIENTED STROKE DETECTION

Published as “L. Neumann and J. Matas. Scene Text Localization and Recognition with Oriented Stroke Detection, ICCV 2013, Sydney, Australia”

APPENDIX 2: ON COMBINING MULTIPLE SEGMENTATIONS IN SCENE TEXT RECOGNITION

Published as "L. Neumann and J. Matas. On Combining Multiple Segmentations in Scene Text Recognition, ICDAR 2013, Washington D.C., USA"