xLiMe
FP7-ICT-2013-10
ICT-2013.4.1
Contract no.: 611346
www.xlime.eu

Deliverable D2.2.1

Early Text from Video Prototype

Editor:	Dubravko Culibrk, UNITN
Author(s):	Dubravko Culibrk, UNITN; Nicu Sebe, UNITN
Deliverable Nature:	Prototype (P)
Dissemination Level:	Public (PU)
Contractual Delivery Date:	M12 – 31 October 2014
Actual Delivery Date:	M12 – 31 October 2014
Suggested Readers:	All project partners
Version:	1.0
Keywords:	text from video; early prototype; VOCR

Disclaimer

This document contains material, which is the copyright of certain xLiMe consortium parties, and may not be reproduced or copied without permission.

In case of Public (PU):

All xLiMe consortium parties have agreed to full publication of this document.

In case of Restricted to Programme (PP):

All xLiMe consortium parties have agreed to make this document available on request to other framework programme participants.

In case of Restricted to Group (RE):

The information contained in this document is the proprietary confidential information of the xLiMe consortium and may not be disclosed except in accordance with the consortium agreement. However, all xLiMe consortium parties have agreed to make this document available to <group> / <purpose>.

In case of Consortium confidential (CO):

The information contained in this document is the proprietary confidential information of the xLiMe consortium and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the xLiMe consortium as a whole, nor a certain party of the xLiMe consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	xLiMe – crossLingual crossMedia knowledge extraction
Short Project Title:	xLiMe
Number and Title of Work Package:	WP2 Text Extraction from Multilingual Multimedia Natural Language
Document Title:	D2.2.1 - Early Text from Video Prototype
Editor:	Dubravko Culibrk, UNITN
Work Package Leader:	Blaž Novak, JSI

Executive Summary

The main goal of the xLiMe project is to enable extraction of knowledge from different media channels and languages and relating this knowledge to cross-lingual, cross-media knowledge bases. The functional requirements for an early prototype of a system able to do this have been gathered and systemized in deliverable (D1.4.1) of xLiMe. An integral part of the system described there is the module for extracting text from video, which is the focus of task T2.2 of the project.

To meet the functional requirements for early text from video, as specified in D1.4.1, we need to identify and adapt a suitable OCR (Optical Character Recognition) component, able to extract the textual information present in IPTV videos obtained from ZATTOO. According to the requirements, the component should focus on brand name detection from IPTV stream. The early prototype should provide annotations of xLiMe TV streams in terms of a select number of brand names detected.

Table of Contents

Abbreviations

OCR	Optical Character Recognition
IPTV	Internet Protocol television
HLS	HTTP Live Streaming

1 Introduction

This deliverable outlines the results of the scientific investigation undertaken in order to identify existing technologies related to the extraction of textual information from multimedia (images and video) and use them to develop an early xLiMe prototype that will be able to do this within the scope of the xLiMe project.

1.1 Background and Motivation

There are three use cases in xLiMe that will provide feedback end evaluation for the system to be developed. They will be run by the following partners:

- **ZATTOO [1]** is a pioneer of IPTV and leader in Switzerland and Germany, with additional presence in France, Spain, UK, Luxemburg, and Denmark. Zattoo earns income from ads, premium services, and B2B relationships. ZATTOO's proprietary technology assets include cloud-based recording which is currently configured to store over 200,000 hours (120 live TV channels are continuously recorded over the past 7 days and individual recordings for users in several countries) .
- **VICO Research & Consulting GmbH [2]** is concentrated on social media measurement and analysis, and the construction of social media monitoring systems as well as social media consulting. Their main customers are consumer goods manufacturers and marketing agencies. Clients amongst others are LG Electronics, Commerzbank, Symantec Europe, BMW, EnBW, Ferrero, Central, ENVIVAS, T-Systems, Mazda Europe, and Mindshare.
- **ECONDA GmbH [3]** focuses on web-analytics and recommendation solutions. For several years running, ECONDA is listed as one of the Top Five of web-analytics tools by independent experts. More than 1,000 satisfied e-business customers rely on ECONDA's web-analytics solutions. This includes customers such as retailers, textile specialists, manufacturers, brands, service providers, portals, publishers, price comparators, publishing houses, newspapers and NGOs. Since the ECONDA Use Case builds on the other Use Cases and starts in Year2, details and requirements will be covered in D1.4.2.

The use cases have been carefully selected in order to demonstrate the advantages of the technology developed within the project. They will focus on two different applications of interest to different stakeholders:

- **Cross-media content enrichment and search**: Providing multimedia content consumers with additional related content enhances the service provided by companies such as ZATTOO. The xLiMe project will develop applications, which will enable enrichment of the multimedia content of TV-channels watched by ZATTOO-users with related content, originating from other media sources (e.g., tweets, blog posts, YouTube videos, news articles, Wikipedia pages, etc.). The approach will be based on content, not on user behaviour.
- **Cross-media brand and topic monitoring**: The social media consulting process can be further enhanced by relating collections of social media documents on a topic to related TV-channels about the same topic. For instance to measure the coverage of topics in mainstream media which are trending in social media. From a business standpoint, brands are a topic of special interest for our use-case partners. The xLiMe project will provide tools to enable the annotation of mainstream multimedia streams with select advertisement presence and product placement data, which will then be used to establish the connection between social and mainstream media. The annotation information is also of use to multimedia content providers, as they are then able to analyse the product placement in the content. The annotation of the stream will be done by detecting logos, brands and ads in the multimedia stream and linking to the product shown in the ad. Initially, this will be done for a limited, predetermined number of logos, brands and ads.

The requirements for the early prototype of the text from video component are mainly derived from the second (cross-media brand and topic monitoring) use case, where the component will be used to detect the appearance of brand mentions in textual form, in the video stream.

While the focus of the early prototype is on brand-related data, the early prototype is designed to detect and extract as much textual data as possible from the frame, providing valuable input for both use cases.

The prototype harnesses the state-of-the-art technologies available and accessible to the consortium and builds upon them to provide a real-time performance Video OCR (VOCR) solution.

2 Text from Multimedia

While optical character recognition is a well-researched problem, which yielded numerous commercial solutions, extraction of text from images and videos (Video Optical Character Recognition - VOCR) is still very much an open research problem [4]. There is, to the best of our knowledge and at the time of writing this document, no open source VOCR solution available which would satisfy the requirements of the xLiMe project. For our early prototype we are, therefore, required to use a combination of technologies available to us.

For xLiMe we considered two different scenarios for extracting text from images/video:

1. Text "in the wild" – the problem of extracting text appearing in natural images and video frames regardless of the orientation, scale and location, that aims to match human observer performance (see Figure 1).
2. Overlaid text – the problem of extracting relatively flat text, printed over the frames, such as that in running titles (see Figure 2).

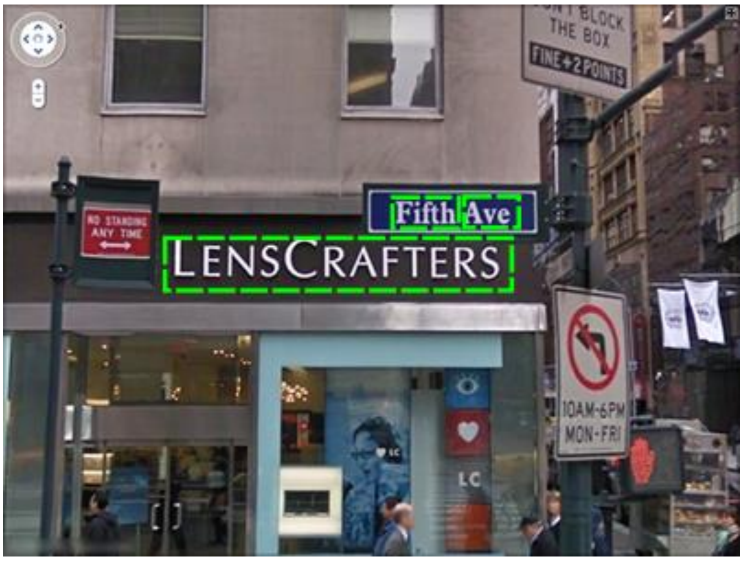

Figure 1: Text detection in the wild

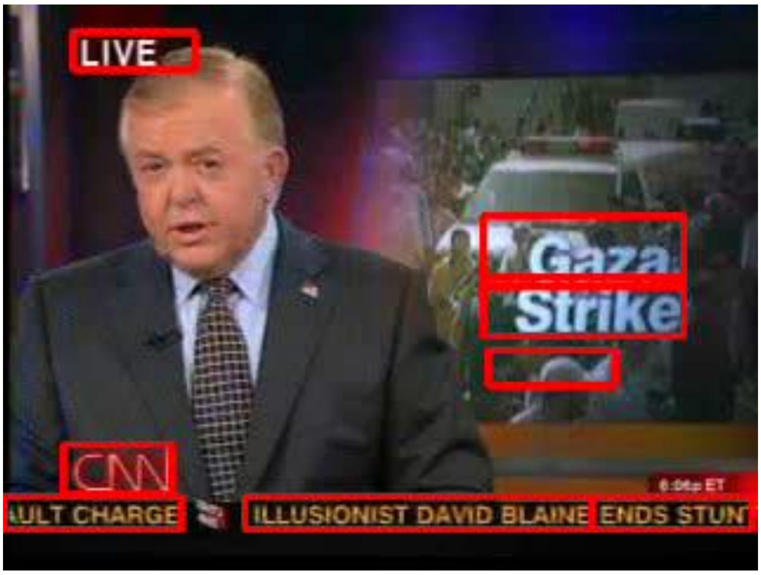

Figure 2: Overlaid text detection in a news video

We initially evaluated a recently proposed state-of-the-art solution for extracting text in natural images obtained from Google Street View [4]. The approach represents an end-to-end solution for extracting the text from images, but, unfortunately, the initial investigation revealed that the computational complexity does not allow for real-time performance within the scope of xLiMe use cases. Therefore, the early prototype is based on a more “classical” approach to VOCR. The Informedia approach proposed by Dipanjan *et al.* [7] has been shown to work well in a multimedia retrieval scenario and achieve real-time performance and forms the core of the technology in the early prototype.

The basic pipeline of the system is to first detect the frames of the video that contain text, then find and extract the text regions in each frame, followed by a separation phase where the text and its background get separated into a binary image that can be processed by any given OCR system, as shown in Figure 3.

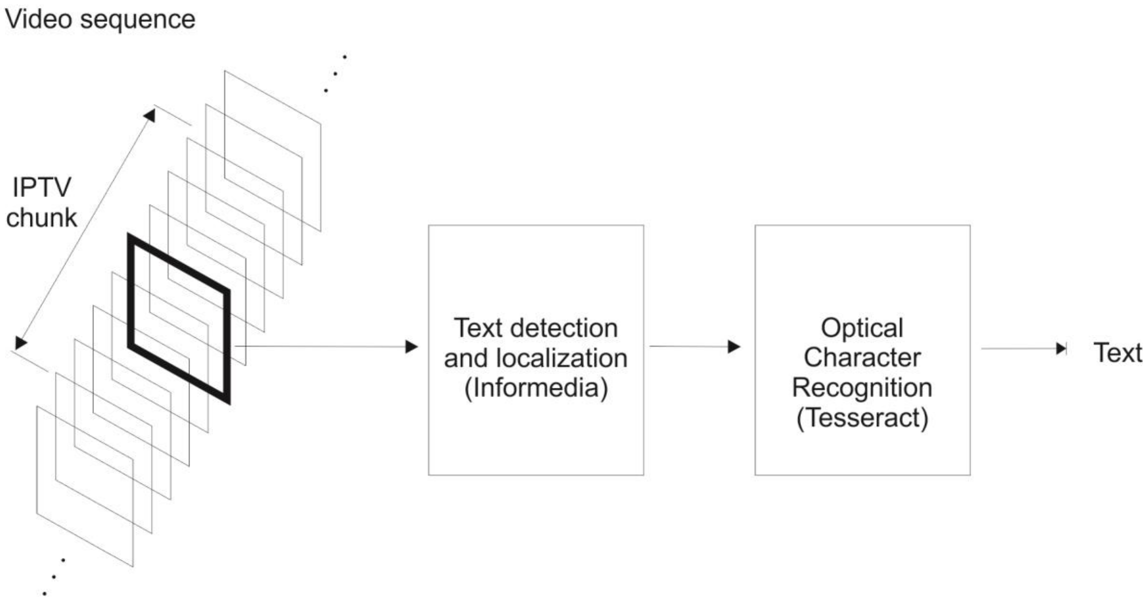
Video sequence
IPTV chunk
Text detection and localization (Informedia)
Optical Character Recognition (Tesseract)
Text

Figure 3: Early prototype VOCR pipeline

For our early prototype we opt to use the Informedia approach (described in Section 2.1.1) to detect and localize text. However, the Informedia system uses a commercial OCR solution for final text extraction, which is a potential limiting factor for the wider adoption of the xLiMe technology. Therefore, in the early prototype, this OCR system has been replaced by the open source Tesseract (described in Section 2.1.2) solution. The following subsections provide a more detailed description of the base technologies used to create the early prototype.

2.1.1 Informedia

The Informedia system is tailored for video information retrieval. The system assumes that text blocks consist of short edges in vertical and horizontal orientations. Moreover it assumes that those edges are connected to each other. The first step the system performs is a text localization step, using a Canny edge detector and applying morphological operators for both vertical and horizontal edge dilation. At this point the system has high recall, but quite a few false alarms, mainly caused by slanting stripes and small areas of the background or human faces which tend to cause sharp edges in the images. In order to reduce the number of false alarms and refine the locations of text in candidate regions that contain text connected with background objects, individual text lines are identified. The system then classifies extracted text lines into actual text regions, a process dubbed text verification. The final phase is the recognition, which is performed by a separate (commercial) OCR system. To ensure best results, before the text lines can be processed by such a system, the image needs to be binarized (converted to black and white). In the binarization step the text is extracted from the background using Otsu’s adaptive thresholding algorithm [8], which creates a histogram of the image and selects a threshold to maximize interclass variance. The resulting binary image is then passed to the OCR system, which, in the original Informedia system, is Textbridge OCR [9].

2.1.2 Tesseract

Tesseract is an open source OCR engine, originally developed by Hewlett-Packard in 1987 [10] as possible add-on for HP's line of flatbed scanners. The motivation behind this project was that commercial OCR engines were not able to handle anything but the best quality print. Even though the engine was significantly better than other commercial systems at that time, it never became a commercial product. In 1994 the development stopped and participated in the 1995 Annual Test of OCR Accuracy [11], where it proved to be the state of the art at that time. In 2005, HP released Tesseract for open source and currently it is under development by Google [12]. For a long time Tesseract was considered to be state of the art, but more recently commercial engines took over this position.

Historically Tesseract assumes that its input is a binary image where the text regions are already defined, because HP used page layout analysis technology in their products and therefore this was not part of Tesseract's job. First a connected component analysis is performed and outlines of the components are stored. This gives the advantage that inverse text can easily be recognized as well. Outlines are gathered together into blobs, which are organized into text lines that on their turn are broken into words. Today Tesseract is able to handle greyscale images, but the performance of its binarization stage is not good, so we opt for doing the binarization using Otsu's approach.

The recognition stage then first tries to recognize each word in order, which is given to an adaptive classifier as training data to be able to more accurately recognize upcoming words. As a second step, the adaptive classifier runs over the words again to try to refine the results.

Finally the recognized text is cleaned up by resolving fuzzy spaces. An overview of the Tesseract OCR engine can be found in [5].

As of version 3 (released in October of 2011), Tesseract can support multiple- language documents and already has language packs of varying quality for more than 30 languages, including: Arabic, English, Bulgarian, Catalan, Czech, Chinese (Simplified and Traditional), Danish, German (standard and Fraktur script), Greek, Finnish, French, Hebrew, Hindi, Croatian, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Dutch, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak (standard and Fraktur script), Slovenian, Spanish, Serbian, Swedish, Tagalog, Tamil, Thai, Turkish, Ukrainian and Vietnamese. Thus, Tesseract is able to support all languages of interest to xLiMe.

2.1.3 xLiMe VOCR Technical Details

Unlike the original Informedia solution, xLiMe VOCR is Linux and OpenCV 2.x [13] based. The prototype processes the 4-second-chunks provided by the ZATTOO HLS stream. Currently, a single frame is extracted from each chunk, and processed. The output of the VOCR is sent to the xLiMe Apache Kafka stream, under the topic ”tv-ocr”.

3 Evaluation

We evaluated the prototype both in terms of the performance of the text detection and localization module, as the performance of the Tesseract system is well documented. We initially followed the procedure used to evaluate the Informedia system and, subsequently, extended it to include an additional dataset manually annotated by our researchers.

The intermediate results of the text detection stage were initially evaluated using the ground truth generator proposed in [14]. The evaluation was based on a small dataset for which true locations of text boxes are known in every frame. The dataset consists of 45 video frames that have 158 text boxes in total, of which 128 of them are recognizable by humans. The maximum number of text boxes in one frame is 21 and there are also 3 frames that contain no text at all. The method makes use of a positive set and a negative set of indices at text box level, which evaluates the detection quality in terms of both location accuracy and fragmentation of the detected text boxes. Moreover a detection difficulty and a detection importance is taken into consideration for each ground truth text box.

To evaluate the sensitivity of the prototype to scale, we have also collected a second, larger, dataset, which contains 4225 images, extracted from video news broadcasts. The frames contain both text added by the news service and naturally occurring text. The images are all of 748x432 pixels size, double the size of those in initial data set. All instances of text that appear have been annotated by humans.

When evaluated on the ground truth generator dataset the system, not surprisingly matched the detection performance of the Informedia system, and achieved an accuracy of 59%.

On the higher-resolution additional dataset collected for the purposes of xLiMe, the accuracy was lower (30%), but the recall was 69%, while the precision was 30%. Since it is, from the point of view of the xLiMe goals, more favourable to have higher recall than precision in the text detection stage, and filter out the meaningless text subsequently, this result is satisfactory from the point of view the requirements for the early prototype.

In our experiments the system is able to process, on average, 6 frames per second on a common desktop computer.

4 Conclusions

This deliverable describes the technologies used to develop the xLiMe early text from video prototype and the prototype itself. It also provides results of the initial evaluation of the prototype in terms of text detection performance.

We have successfully developed a text from video prototype based on state-of-the-art open source OCR technology available. The prototype achieves performance sufficient to meet the goals of xLiMe.

In the future we will attempt to improve the performance of the prototype, address the limitations in terms of sensitivity to scale, and attempt to develop a solution that would be better suited to text detection in the wild (in natural images).

Further system-level evaluation of the performance of this component will also be conducted.

References

[1] http://corporate.zattoo.com

[2] http://www.vico-research.com

[3] http://www.econda.com

[4] Neumann, Lukas, and Jiri Matas. "Real-time scene text localization and recognition." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[5] R. Smith. An overview of the tesseract ocr engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), volume 2, pages 629–633, Washington, DC, USA, 2007. IEEE Computer Society.

[6] V. Manohar, P. Soundararajan, M. Boonstra, H. Raju, D. Goldgof, R. Kasturi, and J. Garofolo. Performance evaluation of text detection and tracking in video. In Lecture Notes in Computer Science, volume 3872, pages 576–587, January 2006.

[7] Das, Dipanjan, Datong Chen, and Alexander G. Hauptmann. Improving multimedia retrieval with a video OCR. Electronic Imaging 2008. International Society for Optics and Photonics, 2008.

[8] N. Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics, 9(1):62–66, January 1979.

[9] Nuance. Textbridge ocr. http://www.nuance.com/textbridge/.

[10] R. Smith. The Extraction and Recognition of Text from Multimedia Document Images. PhD thesis, University of Bristol, Bristol, England, November 1987.

[11] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fourth annual test of OCR accuracy. Technical report, Information Science Research Institute, July 1995.

[12] J. Schulenburg. Gocr. http://jocr.sourceforge.net/.

[13] Bradski, G., The OpenCV Library, Dr. Dobb's Journal of Software Tools, 2000.

[14] X.-S. Hua, L. Wenyin, and H.-J. Zhang. Automatic performance evaluation for video text detection. In Sixth Int. Conf. on Document Analysis and Recognition (ICDAR 2001), September 2001.

[15] Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722