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

Deliverable D2.2.2

Final 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:	M24 – 31 October 2015
Actual Delivery Date:	M24 – 31 October 2015
Suggested Readers:	All project partners
Version:	1.0
Keywords:	text from video; final 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.

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

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:	D2.2.2 - Final Text from Video Prototype
Editor:	Dubravko Culibrk, UNITN
Work package Leader:	Blaž Novak, JSI

Copyright notice

© 2013-2016 Participants in project xLiMe

Executive Summary

The main goal of the xLiMe project is to enable the extraction of knowledge from different media channels and languages and relating this knowledge to cross-lingual, cross-media knowledge bases. The functional requirements for the final prototype (demonstrator) of a system able to do this have been gathered and systemized in deliverable (D1.4.2) 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.

A suitable Video OCR (VOCR) component has been developed in year one and was discussed in D2.2.1. Both the early and the final prototype provide transcriptions of flat text present in xLiMe TV streams. In year two, we focused on improving the early prototype to achieve a component which produces more stable results and improves precision at the expense of recall. We explored an alternative approach to VOCR, which increased precision, but at the expense of computational performance. In this deliverable we present the result of the experimental evaluation of both early prototype and the alternative developed during year II, as well as the description of the new approach. Both approaches have been developed to be interchangeable and can be used to annotate video streams depending on the goals of the application.

Since the relative improvement in terms of precision of the new approach is 15% relative (5% absolute), but at the expense of ~7 times slower processing, the final prototype of xLiMe will use the same approach developed for the early prototype (the code has been revised to remove bugs and support easier reuse by other researchers), while providing support to the use of the alternative for future applications and or when more powerful hardware is available.

Both approaches have been made available as open source, for the use of other researchers, as has our benchmark dataset and the evaluation code.

© xLiMe consortium 2013 - 2016

Table of Contents

 © xLiMe consortium 2013 – 2016

List of Figures

List of Tables

 © xLiMe consortium 2013 – 2016

Abbreviations

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

© xLiMe consortium 2013 - 2016

1 Introduction

This deliverable outlines the results of the scientific investigation undertaken in order to identify the existing technologies related to the extraction of textual information from multimedia (images and video) and use them to develop the xLiMe prototype that is 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 and were revised based on input from the year I review. They will focus on three different applications of interest to different stakeholders:

- Cross-media content enrichment: 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 behavior.
- Cross-media analytics: 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 of the stream will be done by detecting logos, brands and products 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 products/groups of products.
- Cross-media product recommendations:
  Once the cross-modal product placement data has been extracted from both mainstream and social media, this information can be used to enhance the performance of e-Commerce web sites (web shops) by providing better recommendations.

 © xLiMe consortium 2013 – 2016

The requirements for the final prototype of the text from video component are derived from all the use cases.

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

© xLiMe consortium 2013 - 2016

2 Text from Multimedia

While optical character recognition (OCR) is a well-researched problem, which yielded numerous commercial solutions, extraction of text from images and videos “in the wild” (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 final prototype we were, therefore, required to use a combination of technologies available to us and to develop new ones.

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

© xLiMe consortium 2013 – 2016

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 was 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. For our year II alternative approach we tried to follow the same basic principles in order to achieve improved performance.

The basic pipeline of our system is to first detect the frames of the video that contain text, then find and extract the text regions in each frame, which are then passed for processing by any given OCR system, as shown in Figure 3.

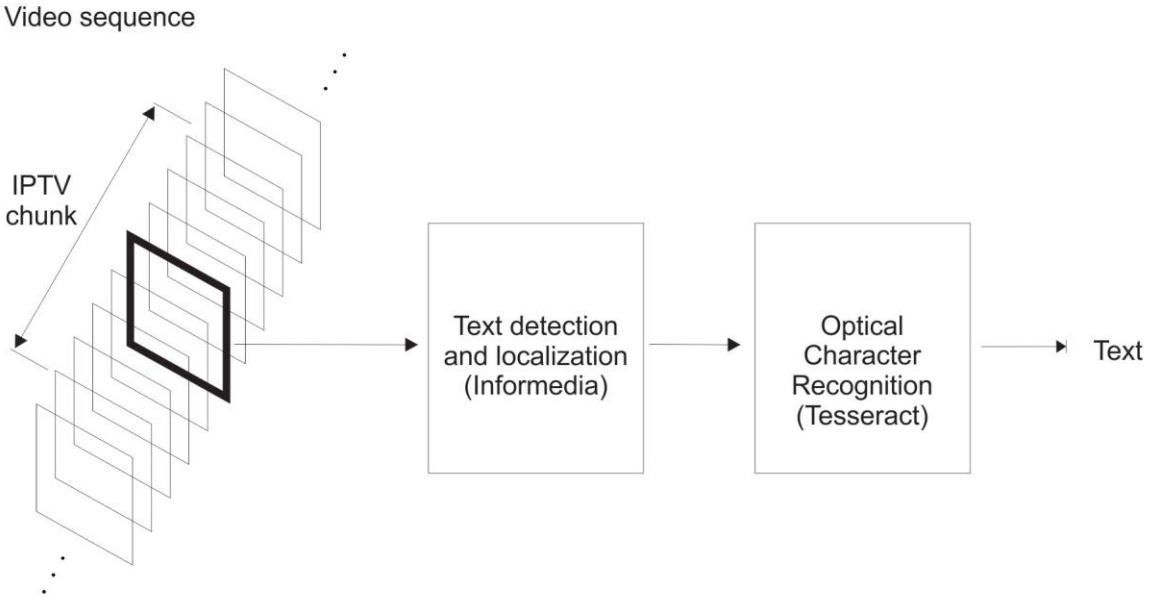

Figure 3: Early prototype VOCR pipeline

For our early prototype we opted 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 alternative approach developed in year II uses a different technology to detect the text regions, but the OCR system is still Tesseract. The following subsections provide a more detailed description of the base technologies used to create the early prototype text detection and the year II alternative (CaptionCapture described in section 2.1.2).

2.1 Early prototype text detection (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].

2.2 CaptionCapture

The approach developed in year II (CaptionCapture) relies not only on vertical and horizontal edge statistics, but also on filtering out regions that do not exhibit contours of significant size within the edges. Text detection is a two stage process, where first only the vertical edges are examined and the regions that exhibit contours of these edges larger than a predefined threshold are deemed to be candidates for potential text boxes. At this stage, all the regions that exhibit such properties are also assigned to groups based on their overlap. Sobel filtering is used for edge detection instead of the Canny approach used in Informedia.

Once the initial candidates are detected, a second stage of focuses on examining each candidate region individually. First, the edges in the region are detected using adaptive thresholding available within the OpenCV library. Next, contours are detected in the region and regions that do not contain contours large enough are discarded. For the remaining regions a count of edge pixels is derived and regions that contain less than a predefined portion of the edge pixels or are of inappropriate general shape (not horizontal rectangles of certain ratio of the dimensions) are eliminated.

Finally, for the remaining regions, group information extracted in the first stage is used to generate an estimate of the whole text region. These regions are then passed to the Tesseract OCR engine to extract the text.

The implementation of the whole system has been done in Python and the code is available on GitHub (https://github.com/holtzhau/captioncapture).

2.3 Tesseract

Tesseract is an open source OCR engine, originally developed by Hewlett-Packard in 1987 [10] as a 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 are then segmented into words. Today Tesseract is able to handle greyscale images. In our early prototype we opted for doing the binarization ourselves using Otsu's approach. For CaptionCapture detected regions this did not result in improved performance, so no binarization is performed before passing the images to Tesseract.

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].

 © xLiMe consortium 2013 – 2016

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.

2.4 xLiMe VOCR Technical Details

Unlike the original Informedia solution, both xLiMe VOCRs are Linux and OpenCV 2.x [13] based. The final 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”. The VOCR engine can easily be switched between the two alternative engines for any application.

© xLiMe consortium 2013 - 2016

3 Evaluation

We developed a new approach for text detection, in an attempt to improve the prototype performance between the early and final solution. We evaluated the performance of both prototypes in terms of the performance of the text detection and localization module.

The intermediate results of the text detection stage were evaluated using the benchmark collected for the purposes of the project and used to evaluate the year I early prototype for text from video. Of the two datasets used for evaluation in year I this one is significantly larger (4225 frames, compared to the 45 frames of the other dataset), more related to the goals of the project and proved to be much more challenging. Thus we feel that it represents a better benchmark within the scope of our project.

The dataset has been 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.

Table 1 shows the results obtained for the first year prototype and the second year alternative (CaptionCapture), in terms of recall, precision and speed. The alternative VOCR increased the precision, but lowered the recall. In addition, the speed of the alternative approach is significantly lower (by a factor of nearly 7). Based on the evaluation, the final text from video prototype remained based on the original text detection approach used in year I. However, the alternative approach developed within year two remains a viable alternative for applications in need of more precision and are not concerned with the speed so much, and is preferred for off-line processing.

	Precision	Recall	Speed (seconds per frame)
1st year prototype	30%	**69%**	**0.17**
2nd year alternative	**35%**	27%	1.16

Table 1: Evaluation results

© xLiMe consortium 2013 – 2016

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 evaluation of the prototype in terms of text detection performance, using the two text detection modules developed within the project.

We have successfully developed a text from video prototype based on state-of-the-art open source OCR technology available and our own developed solutions. The prototype achieves performance sufficient to meet the goals of xLiMe and can be tuned to favor precision or speed. Both approaches have been made available to other researchers as has been our benchmark dataset and the evaluation scripts used in our experiments.

The results presented here have been produced during the final evaluation of the component as a separate module, planned within the project. Further system-level evaluation of the performance of this component will be conducted, within the evaluation of the performance of the xLiMe pipeline as a whole.

© xLiMe consortium 2013 - 2016

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

 © xLiMe consortium 2013 – 2016