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

Deliverable D3.2.1

Early Prototype for Video Annotation

Editor:	Nicu Sebe, UNITN
Author(s):	Dubravko Culibrk, UNITN; Nicu Sebe, UNITN
Deliverable Nature:	Prototype (P)
Dissemination Level:	Public (PU)
Contractual Delivery Date:	M12 – 31 October2014
Actual Delivery Date:	M12 – 31 October2014
Suggested Readers:	All project partners
Version:	1.0
Keywords:	video annotation; early prototype

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:	WP3 Cross-lingual Multimedia Semantic Annotation
Document Title:	D3.2.1 - Early Prototype for Video Annotation
Editor:	Nicu Sebe, UNITN
Work Package Leader:	Nicu Sebe, UNITN

Copyright notice

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 annotating video based on the content, which is the focus of task T3.2 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 visual object recognition component, able to extract the visual information present in IPTV videos obtained from ZATTOO. According to the requirements, we need to develop tools to perform lightweight approximate annotation of video streams in terms of brand logo appearance. The resulting annotation should annotate the video stream with information indicating the presence of a limited number of brands.

Table of Contents

Executive Summary 3
Table of Contents 4
Abbreviations........ 5
1 Introduction 6
1.1 Background and Motivation 6
2 Visual Recognition for Multimedia Annotation 8
3 Early prototype for Visual Multimedia Annotation........ 9
4 Evaluation........ 11
5 Conclusions 12
References........ 13

Abbreviations

CNN	Convolutional Neural Network
IPTV	Internet Protocol television
HLS	HTTP Live Streaming
GPU	Graphics Processing Unit
VOCR	Video Optical Character Recognition

1 Introduction

This deliverable outlines the results of the scientific investigation undertaken in order to identify existing technologies related to content recognition in multimedia (images and video) and use them to develop an early xLiMe prototype of the system that will be used to do annotate multimedia data within the scope of the xLiMe.

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 Use Case of ECONDA builds on the other Use Cases and starts in Year2 of the project, 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 video annotation 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 brands in the video stream.

While the focus of the early prototype is on brand-related data, the early prototype is also designed to recognize a variety of different objects in videos and images, 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 Annotation solution

2 Visual Recognition for Multimedia Annotation

The dominant methodology for visual recognition from images and video until recently relied on hand-crafted features [5][12]. Today, we are witnessing a paradigm shift and a growing interest in methods that learn features in both unsupervised and supervised settings.

Current research on deep learning suggests that there is significant potential in using large-scale Neural Networks (NNs) to address machine learning and, in particular, computer vision problems. The Google Brain project showed how an unsupervised AutoEncoder NN with 1 billion connections was able to learn to recognize common objects just by looking at a week's worth of YouTube videos [8]. In 2012, Krizhevsky *et al.*[7] showed how a Convolutional NN (CNN) with 650,000 neurons can be used to classify1.2M images in the ImageNet Large Scale Visual Recognition dataset into 1,000 classes [4], significantly advancing the state of the art.

Their approach has recently been successfully extended to object detection achieving beyond state-of-the-art results on the PASCALVOC challenge data [5]. Deep learning has also seen several successful applications in the domain of image classification [15][16] and content-based retrieval [14].

When it comes to learning from video data, using deep (convolutional) NNs, few approaches exist [8][12]. However, arguably the most prominent, 3D CNN [17], achieved the best performance in three human action recognition tasks of the TRECVID 2009 Evaluation for Surveillance Event Detection challenge [18], showing the significant potential for such approaches, when large amount of labeled data is available.

In the last couple of years, approaches relying on deep convolutional neural nets continue to dominate to dominate the field and achieve best results in relevant competitions (ImageNet and TRECVID). To ensure state-of-the-art performance and capitalize on these achievements, the early prototype for video segmentation is based on the approach of Krizhevsky*et al.*[7], as adapted in [5].

3 Early prototype for Visual Multimedia Annotation

The early prototype developed can operate on both frames extracted from videos and images. Therefore the terms “early prototype for video annotation” and “early prototype for visual multimedia annotation” are used interchangeably in the rest of the text. The typical pipeline used to recognize objects in images and video is shown in Figure 1.

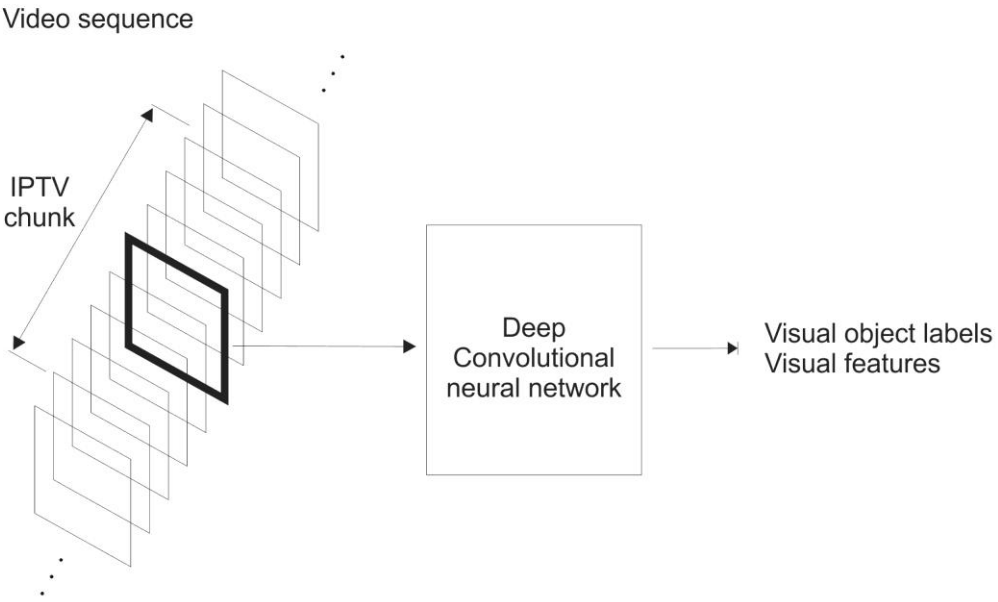

Figure 1 Pipeline of the early prototype for video annotation

For xLiMe we developed two different annotation pipelines that form the early prototype for visual multimedia annotation:

1. The **General Purpose (GP) annotation pipeline**, which attempts to recognize 1000 different objects that are represented in the IMAGENET data set, which can be used in both images and video.
2. **Logo Detection (LD) recognition pipeline** suitable for recognizing the occurrences of logos in “wild” images typically appearing on the Internet and logos appearing in standard locations in videos , to aid the add recognition process.

All the xLiMe prototype pipelines are based on the same CNN structure. The network contains 650,000 neurons architecture, organized in eight layers, five convolutional and the three fully connected.

For our prototype and training of the detectors, we rely on CAFFE deep learning framework [6]. CAFFE’s architecture allows for easy switching between the modes using the GPU and CPU for network training and image classification and enables us to process over 40M images per day with a single NVIDIA K40. On a typical CPU, the prototype is able to process an image in environ 20ms.

The prototype processes the 4 second chunks provided by the ZATTOO HLS stream. Currently, a single frame is extracted from each chunk, scaled to 256x256 pixels and processed. The output of the VOCR is sent to the xLiMe Apache Kafka stream, under the topic ”tv-annotate”.

For the GP pipeline, the pre-trained classifier provided within CAFFE is used directly. The output pushed to the xLiMe Kafka stream includes the labels of the top five scoring visual classes, as well as the activation values of the neurons in the last fully connected layers, as these have been shown to be discriminative features for further multimedia processing [5] and can be used for later, system-wide operations.

For the LD pipeline, we focus on detecting the logos of Deutsche Telekom (DT), as this is a client of interest to our use-case partner (VICO) and we fine tune the classifier pre-trained on IMAGENET data to be able to recognize the DT logo. The fine tuning is done by replacing the last softmax layer of the network with a layer containing just two neurons indicating the presence or the absence of the logo in an image or a video frame.

To train the classifier able to detect the appearance of the logo in images typically found on Internet we manually collected 1334 of such images from GoogleImages. Sample images from our dataset are shown in Figure 2.

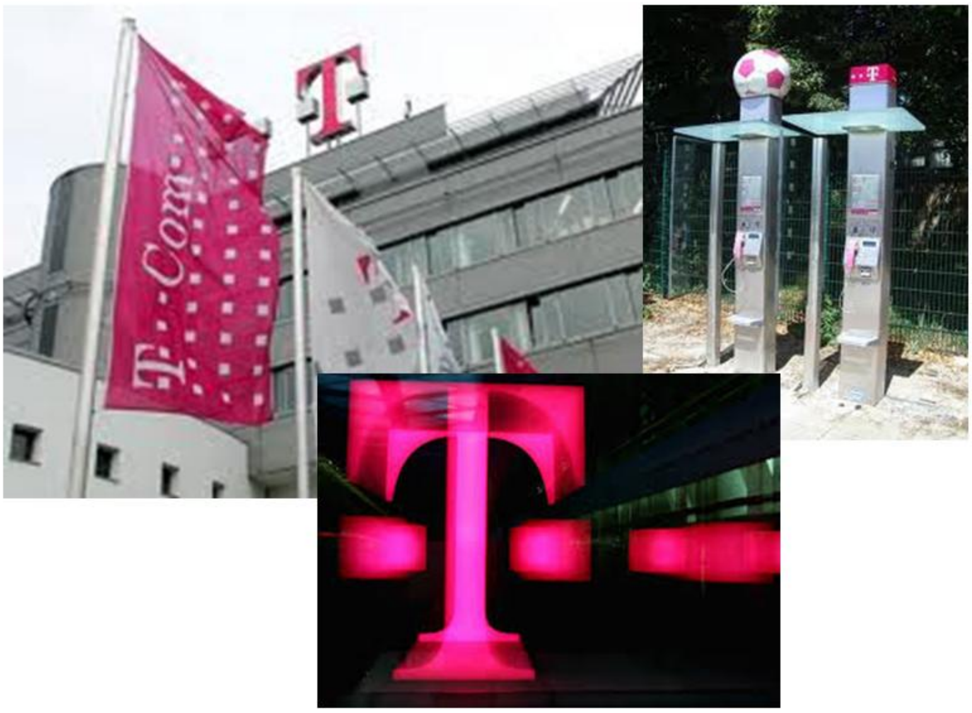

Figure 2 Sample images containing Deutsche Telekom logo from our data set

We extend our dataset based on three promotional DT videos (ads) provided to us by our partner. For this, we focus on the top right corner of the video frames (defined as the 20% of the frame size), where the logo is typically displayed in the ads and annotate manually all frames where the logo appears.

The results of the initial evaluation of running the pre-trained classifier on the DT data set, indicated that the classifier tends to confuse the DT logo with some IMAGENET classes sharing some visual similarity with the DT logo: lipstick and digital clock (see Figure 3). Therefore, we included samples of these classes into our dataset as negative examples. These are then extended by the crops from the frames of the promotional videos not containing the logo.

In total our training set contains 5613 samples, which include 914 positive samples, 814 obtained from the Internet and 100 from the promotional videos. We fine tune the net for 100000 iterations.

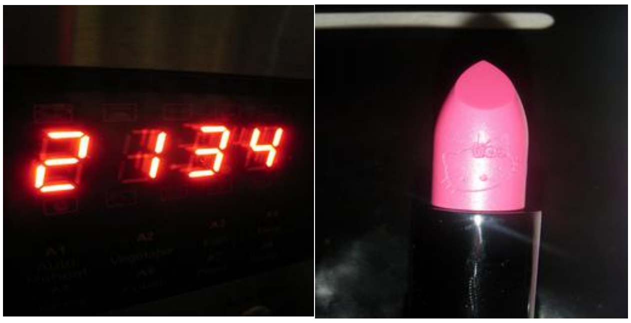

Figure 3 Examples of IMAGENET classes visually similar to DT logo

4 Evaluation

For the general-purpose annotator the performance is evaluated in terms of image classification following the classical IMAGENET procedure. The classifier achieves 15.3% top 5 error rate. The classifier is deemed wrong if the correct visual object class label is not in the top five labels selected for an image.

To evaluate the performance of the LD annotator on the Internet image logo detection task, we created a separate dataset test. It consists of:

- 520 positive examples of images containing the DT logo, obtained from the internet.
- Images for different logos provided by our other use-case partner (ECONDA) which include Adidas and CRYTEK logos.

The total size of the test data set is 1646 image, 520 positives and 1125 negatives.

The annotator achieves 99.5% accuracy on the test set.

Finally, to evaluate the usefulness of the LD for video ad detection, we ran the annotator on top-right crops for all the frames in the three promotional DT videos. Under these testing conditions, the annotator achieved 100% accuracy.

5 Conclusions

This deliverable describes the technologies used to develop the xLiMe early video annotation prototype, and the prototype itself. It also provides results of the initial evaluation of the prototype in terms of visual object recognition performance.

We have successfully developed a video annotation prototype based on state-of-the-art visual recognition 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 logo detection in wild video (unconstrained logo appearances in any type of video).

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] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. InCVPR, pages 248–255. IEEE, 2009.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524, 2013.

[6] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. InNIPS, volume 1, pages 4–9, 2012.

[8] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In CVPR,pages 3361–3368. IEEE, 2011.

[9] A. Ravichandran, R. Chaudhry, and R. Vidal. View-invariant dynamic texture recognition using a bag of dynamical systems. In CVPR, pages 1651–1657. IEEE, 2009.

[10] P. Saisan, G. Doretto, Y. Wu, and S. Soatto. Dynamic texture recognition. In CVPR, pages II–58–II–63. IEEE, 2001.

[11] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In MIR ’06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pages 321–330. ACM Press, 2006.

[12] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. InECCV, pages 140–153. Springer, 2010.

[13] I. H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. MorganKaufmann, San Francisco, 2005.

[14] P. Wu, S. Hoi, H. Xia, P. Zhao, D. Wan, and C. Miao. Online multimodal deep similarity learning with application toimage retrieval. In ACM MM, pages 153–162, 2013.

[15] Z. Yuan, J. Sang, and C. Xu. Tag-aware image classification via nested deep belief nets. In ICME, pages 1–6, 2013.

[16] S.-h. Zhong, Y. Liu, and Y. Liu. Bilinear deep learning for image classification. In ACM MM, pages 343–352, 2011.

[17] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. PAMI, 35(1):221–231, 2013.

[18] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In MIR ’06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pages 321–330. ACM Press, 2006.