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

**Deliverable D3.2.2**

# Final 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: | M24 – 31 October2015 |
| Actual Delivery Date: | M24 – 31 October2015 |
| Suggested Readers: | All project partners |
| Version: | 1.0 |
| Keywords: | video annotation; final prototype |

Disclaimer



| | |
|---|---|
| Full Project Title: | xLiMe – crossLingual crossMedia knowledge extraction |
| Short Project Title: | xLiMe |
| Number and<br>Title of Work package: | WP3<br>Cross-lingual Multimedia Semantic Annotation |
| Document Title: | D3.2.2 - Final Prototype for Video Annotation |
| Editor: | Nicu Sebe, UNITN |
| Work package Leader: | Nicu Sebe, UNITN |

# 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 the final prototype 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 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 D1.4.2, we needed the further perfect the the visual object recognition component, developed for the early prototype in year I of the project. According to the requirements, we need to develop tools to perform lightweight approximate annotation of video streams in terms of appearance of several different classes of general objects, as well as specific brand logo appearances.

The early prototype was able to annotate the IPTV streams with frame level annotations for 1000 objects, but was, due to the limitation of data it was trained on, sensitive to the scale of the objects and not suitable for object detection in complex scenes. It was also unable to achieve object localisation within the frame.

In the second year we have focused on enabling the prototype to detect and localise objects in complex scenes. The computational requirements of our use cases, combined with the available state-of-the-art hardware, allow us to do so for 20 general categories of objects. In addition to extracting the information about the presence and localisation of these common objects, the prototype is able to provide similar information regarding the appearance of specific brand logos.

 

# Table of Contents



 

# 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 new technologies related to content recognition in multimedia (images and video) and use them to improve the early xLiMe prototype of the system that was used to do annotate multimedia data within the scope of the xLiMe in year I.

## 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 and 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 behaviour.
- **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.

The requirements for the final prototype oof the video annotation component are derived from all the 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 visual-object related data as possible from the frame, providing valuable input for all use cases.
The prototype harnesses 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 PASCAL VOC challenge data [18]. 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].

In year two, we continued the practice of incorporating and adapting state-of-the-art technology and based the final prototype on the Fast-RCNN approach [19], which allows us to do not only frame based classification, but also real time object detection and localisation.

 

# 3. Early prototype for Visual Multimedia Annotation

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

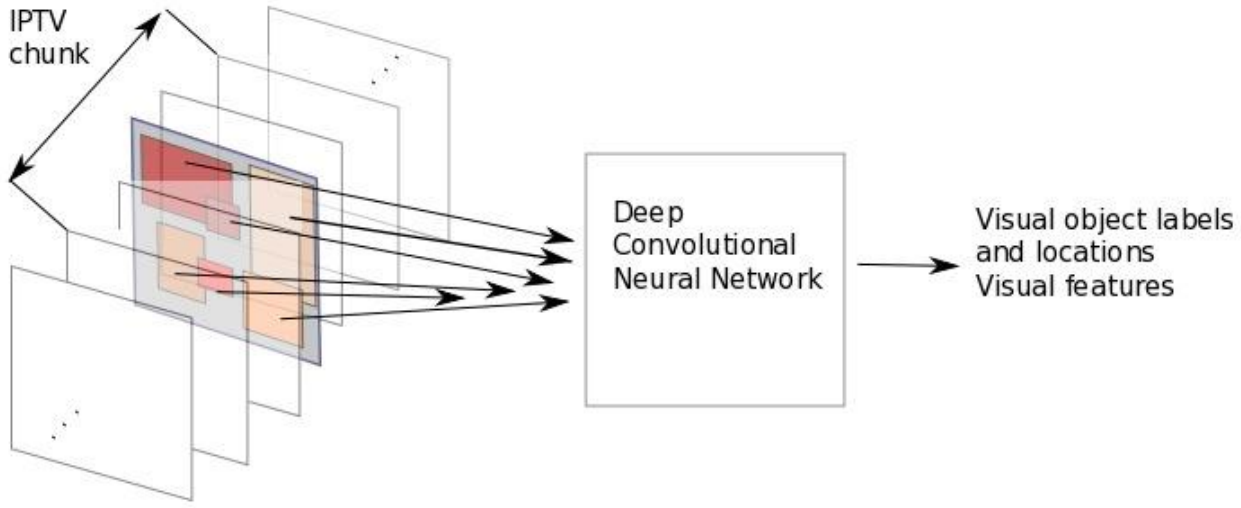


**Figure 1: Pipeline of the final prototype for video annotation**

The early prototype relied on two different annotation pipelines for visual multimedia annotation. In year two, these separate modules have been joined to create a single annotation pipeline which recognises and detect 20 types of general objects, present in the Pascal VOC dataset and the additional Deutsche Telekom logo, of interest for our use-cases. The object classes detected include: ‘aeroplane’, ‘bicycle’, ‘bird’, ‘boat’, ‘bottle’, ‘bus’, ‘car’, ‘cat’, ‘chair’, ‘cow’, ‘diningtable’, ‘dog’, ‘horse’, ‘motorbike’, ‘person’, ‘pottedplant', ‘sheep', ‘sofa’, ‘train’, ‘tvmonitor' and ‘dt’(Deutsche Telecom).

The final prototype of xLiMe is essentially based on the same CNN structure as the early prototype. The seminal Krixhevsky net [5], which has ben modified to speed up the region based object detection by Girshick [19]. The network contains 650,000 neurons architecture, organized in eight layers, five convolutional and the three fully connected.

Unlike the early prototype, in which frames were fed directly to the CNN, the final prototype relies on a object proposal generation methodology to extract a number of regions (object proposals) from a single frame, that could correspond to an object. The regions are rectangles deemed likely to contain a single object and can be of any scale. For each frame, up to 10,000 object proposals is generated using the EdgeBoxes methodology [20]. These proposals are are then fed to a Fast-RCNN [19].

For our prototype and training of the detectors, we rely on the open source Fast-RCNN implementation, which extends the CAFFE deep learning framework [6], used for the early prototype. The 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 regions 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, object proposals are extracted and processed. The output of the detection is sent to the xLiMe Apache Kafka stream, under the topic ”tv-annotate”.

 

For the the general purpose classes, the weights of the Fast-RCNN detector pre-trained on Pascal VOC data can be used directly. To detect a specific brand logo, the detector needs to be finetuned using manually annotated data, that contains the bounding-boxes ground truth information about the location of the logo in the images.

To alleviate the problem of generating such a dataset, we have developed an online annotation tool, allowing several people to share the annotation load. In addition, to reduce the amount of effort needed to create a data set required to train a detector for a specific logo/object, a semi-automatic annotation procedure has been devised:

1. The use case partner collects a set of positive example images obtained using a key-word search from a readily available search engine (e.g. Google).
2. The set of images is used to train a CNN classifier similar to the logo-detector used within the early xLiMe prototype.
3. Region proposals are extracted for the set of positive images and fed into the classifier trained in phase two. The highest scoring proposal is deemed the initial location of the ground-truth bounding box.
4. The automatically generated bounding boxes obtained in step 3 are then provided to the users to verify/modify thorough the annotation tool.

To validate the proposed procedure, we used the dataset collected for the Deutsche Telekom logo within year I. Sample images from our dataset are shown in Figure 2.


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

We used the 520 images, obtained from Google Images, running a keyword search for 'Deutsche Telekom', as a starting point and used the annotation procedure to create the ground truth. The manual verification and modification of the initial proposed bounding boxes took environ 20 seconds per image. Using the online tool, the job was done within an hour by 5 annotators. The ground truth was then used to fine-tune the Fast-RCNN to detect the additional DT logo class. The final ground truth dataset contains 423 images, which remained after the removal of the duplicate images and images that did not feature a prominent logo. The annotation tool and the procedure are described in more detail in D3.1.2.

 

# 4. Evaluation

For the annotator the performance is evaluated in terms of object detection following the Fast-RCNN evaluation procedure for Pascal VOC. To enable the evaluation of the performance of the annotator on the DT logo detection task, we split our positive training data set randomly into training (90% of the data) and validation (10% of the data). The negatives were obtained from the Pascal VOC dataset. The classifier achieves 55% mean average precision (MaP) over the 21 classes (20 Pascal VOC classes augmented with the DT logo class).

# 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 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. During year II we have improved the performance of the prototype, addressed the limitations in terms of sensitivity to scale, and attempted 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] Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." International journal of computer vision 88.2 (2010): 303-338.

[19] Girshick, Ross. "Fast R-CNN." arXiv preprint arXiv:1504.08083 (2015).

[20] Zitnick, C. Lawrence, and Piotr Dollár. "Edge boxes: Locating object proposals from edges." Computer Vision–ECCV 2014. Springer International Publishing, 2014. 391-405.


 