
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n: 761999


**EasyTV: Easing the access of Europeans with disabilities to converging media and content.**

# EasyTV tools for improvement of graphical interfaces

**EasyTV Project**

*H2020. ICT-19-2017 Media and content convergence. – IA Innovation action.*

**Grant Agreement n°: 761999**

Start date of project: 1 Oct. 2017

Duration: 30 months

Document. ref.: D4.1

## Disclaimer



The EasyTV Consortium consists of the following partners:

| | Partner Name | Short name | Country |
|---|---|---|---|
| 1 | Universidad Politécnica de Madrid | UPM | ES |
| 2 | Engineering Ingegneria Informatica S.P.A. | ENG | IT |
| 3 | Centre for Research and Technology Hellas/Information Technologies Institute | CERTH | GR |
| 4 | Mediavoice SRL | MV | IT |
| 5 | Universitat Autònoma Barcelona | UAB | ES |
| 6 | Corporació Catalana de Mitjans Audiovisuals SA | CCMA | ES |
| 7 | ARX.NET SA | ARX | GR |
| 8 | Fundación Confederación Nacional Sordos España para la supresión de barreras de comunicación | FCNSE | ES |
| 9 | Sezione Provinciale di Roma dell'Unione Italiana dei ciechi e degli ipovedenti | UICI | IT |

| | |
|---|---|
| **PROGRAMME NAME:** | **H2020. ICT-19-2017 Media and Content Convergence – IA Innovation Action** |
| **PROJECT NUMBER:** | 761999 |
| **PROJECT TITLE:** | EASYTV |
| **RESPONSIBLE UNIT:** | UPM |
| **INVOLVED UNITS:** | CERTH |
| **DOCUMENT NUMBER:** | D2.2 |
| **DOCUMENT TITLE:** | EasyTV tools for improvement of graphical interfaces |
| **WORK-PACKAGE:** | WP 2 |
| **DELIVERABLE TYPE:** | Prototype |
| **CONTRACTUAL DATE OF DELIVERY:** | 31-01-2020 |
| **LAST UPDATE:** | |
| **DISTRIBUTION LEVEL:** | PU |

**Distribution level:**

**PU** = *Public,*

**RE** = *Restricted to a group of the specified Consortium,*

**PP** = *Restricted to other program participants (including Commission Services),*

**CO** = *Confidential, only for members of the LASIE Consortium (including the Commission Services)*

# Document History

| VERSION | DATE | STATUS | AUTHORS, REVIEWER | DESCRIPTION |
|---|---|---|---|---|
| v. 0.1 | 28/11/2019 | Draft | UPM | Table of Contents definition and document structure |
| V 1.0 | 17/12/2019 | Draft | UPM | First document version |
| V 1.1 | 19/12/2019 | Draft | UPM | Added subtitles adaptation and services indicators sections |
| V 1.2 | 20/12/2019 | Draft | UPM | Added Executive Summary and last corrections |
| V 1.3 | 22/01/2020 | Draft | Georgios Gerovasilis (CERTH/ITI)<br>Nikolaos Kaklanis (CERTH/ITI)<br>Konstantinos Votis (CERTH/ITI)<br>Dimitrios Tzovaras (CERTH/ITI) | Chapter 8 |

# Definitions, Acronyms and Abbreviations

| ACRONYMS / ABBREVIATIONS | DESCRIPTION |
|---|---|
| CSS | Cascading style sheets |
| DR | Danish Broadcasting Corporation |
| HbbTV | Hybrid Broadcast Broadband Television |
| CS | Companion Screen |
| QoE | Quality of Experience |
| SIFT | Scale Invariant Feature Transform |
| SURF | Speeded-up Robust Features |
| ORB | Oriented Fast and Rotate Brief |
| KCF | Kernelized Correlation Filter |
| MOSSE | Minimum Output Sum of Squared Error |
| CSRT | Discriminative Correlation Filter with Channel and Spatial Reliability |
| RCNN | Region-Based Convolutional Neural Network |
| SSD | Single Shot Detector |
| YOLO | You Only Look Once |
| IoU | Intersection over Union |
| FPS | Frames Per Second |
| mAP | Mean Average Precision |
| DoW | Description of Work |

# Table of Contents

# List of Figures

# List of Tables

# Executive Summary

This document corresponds to the deliverable D2.2 “EasyTV tools for improvement of graphical interfaces” of the WP2. This deliverable collects all the information relative to several tools incorporated in the EasyTV platform which improve the quality of experience and the information retrieval for person with disabilities.

Section 1 (Introduction) presents some relevant aspects and an analysis about the necessity of this tools and the inclusion of deep learning solutions to solve some concrete problems related to video information.

Section 2 describes the tool for subtitles adaptation.

Section 3 collects all the information related to the services indicators where some easy understanding icons are used to know which services are activated at each moment.

Section 4 includes all the development related to image magnification and character characterization using deep learning techniques. Lots of algorithms have been tested and combined in a final architecture able to detect and track faces to finally apply algorithms to detect not only which person is the detected persons, also to provide information about unknown characters in the scene including estimated age and gender.

Section 5 present the services included in the app to perform the face magnification and the characters information presentation using the data extracted by the deep learning algorithms using the proposed framework.

Finally, Section 6 and 7 draw some final conclusions and future work to improve the proposed solutions.

# 1. INTRODUCTION

In today's society, television consumption is still one of the main activities of everyone's life. It can serve different purposes such as entertainment, information and education, thus being considered as an essential tool in building inclusive societies. In this regard, while the provision of media content in terms of television coverage is nearly complete, many people who live with some form of disability are still unable to enjoy it, due to access problems related to the content, the information and/or the devices necessary for accessing it.

Current strategies for allowing people with disabilities to perceive what is happening on the TV are usually based on common services such as closed captioning and signing for the deaf, audio descriptions and audio captions for the blind and accessible remote-control devices for people with reduced capabilities. The cost of providing these kinds of solutions, which includes not only their implementation, but also the research work that has to be done previously, has been usually presented as one of the main reasons for limiting the amount of accessible content in the television programming, making that it only achieves the legal level in the best cases (for example, in France (CSA, 2017)) or even provides less accessible content than that (for example in Spain (CNMC, 2017)). For this reason, the definition of new low-cost efficient access services will promote their spread in the multimedia content delivery scenario.

In another aspect, and as it is going to be explained later, sometimes current solutions are not enough for assuring a fully access to the entire content in the terms of user's needs and expectation, thus making more difficult for people with disabilities to participate in different aspects of social and cultural activities related to the television consumption. In this regard, additional solutions must be defined, and they have to be focused not only on "how" the content is provided, but also on "what" additional content can be delivered to enhance the user experience for all.

The digital transformation of the traditional TV landscape within the current audio-visual sector provides a new innovative area for facing these accessibility challenges. In this context, the increase of the use of smart devices while watching TV with Internet capabilities as the same time that the typical broadcast environment for content delivery is alive to compose a new scenario where innovative services can be the answer to new expectations (Claudy, 2012): the hybrid television environment.

This evolution, from typical stand-alone TV set scenario to complex hybrid multi-platform, boots the improvement of several aspects that can be considered as essential for increasing the accessibility level, such as the content hyper personalization and the user interaction. In this context, with the emergence of different technologies such as IPTV (Internet Protocol Television), OTT-TV (Over The Top Television) or HbbTV (Hybrid Broadcast Broadband Television)(Malhotra, 2013), the introduction of companion screens (CS) for multi-screening may define the new consumer behaviour (Vinayagamoorthy, Allen, Hammond, & Evans, 2012), where users may access additional content and services related to the main content in TV in a synchronized way in order to complement it and even to replace it in some cases, creating new users' experiences (Boronat, Marfil, Montagud & Pastor, 2018).

Taking this into account, this new scenario will be used for the "how" aspect of our approach, since it will help to guarantee an optimal provision of the content not only in terms of delivery, but also in terms of presentation, due to the use of the CS as a new mean for the presentation of additional access services.

In another hand, the "what" aspect of our solution will be focused on the content processing for obtaining low cost-efficient access services that can beneficiate from the application of new algorithms and processing techniques such as deep learning (Trigueros & Daniel, 2018 and Voulodimos & Athanasios, 2018). These methods can be applied to obtain new services that provide a higher accessibility level in the media content provision landscape. The automated detection of particular objects inside the content may not only increase the information to be presented to the user, but also improve the way it is presented. According to this idea, in this project we present a

complete method for automated face detection which will allow to develop different solutions for enhancing the access to the content as magnification for helping in the lips reading process for deaf people and character identification to provide additional information to complement current solutions (improving the subtitles and the audio description).

Considering the aforementioned concepts, the main objective of this part of the project is to present two innovative access services for helping the inclusion in the TV environment based on the combination of these two different dimensions: on one hand, the new television paradigm for content provision, that is, the hybrid television due to its own capabilities and, on the other, the application of different artificial intelligence tools for content processing.

In this regard, these innovative services represent a new user-side centred approach focused on obtaining additional information from media content by the application of different image analysis algorithms and presenting it to the final user through the hybrid TV environment, enhancing the accessibility level of the content provided. From these facts we can conclude that our approach may allow to improve the quality of experience (QoE) related to the TV watching experience for people with disabilities, facilitating the transformation of the current no fully accessible landscape into a more inclusive scenario, where interaction, immersion and customization for all will become the new centre of the TV content provision. Nevertheless, this claim should be confirmed by the final users during the testing phase, which has been defined as a future work in the conclusion section.

To describe the above-mentioned contributions, this part of the project is outlined as follows: Section 2 justifies the application of different deep learning techniques for media content processing for new access services and their provision through hybrid TV environment together with their theoretical background. Our methodology for image analysis based on deep learning techniques for face detection is presented in Section 3, including the explanation about the applied algorithms and the obtained results for our approach. Then, Section 4 provides a wide presentation of the final services, indicating how they are implemented and how the previous results are presented to the final users. Section 5 includes a discussion about the obtained results. Finally, Section 6 is focused on presenting the conclusions and the related future work.

Similar to what has been happening in other fields, digital revolution has come to shake the pillars of the entire TV landscape, causing a deep change in consumer behaviour and in-service provision possibilities. In line with this, the evolution from standard machine learning methods to more complex deep learning algorithms has opened a wide range of processing opportunities, especially in the image analysis field, thus helping the definition of innovative services to increase the content accessibility and fulfil user's needs and preferences. Next sections present the background of these two main areas and explain today's main problems in media content accessibility with the purpose of analysing the related research problem and contextualizing our approach.

## 1.1. Hybrid television and content access problems

Regardless the high penetration that broadcast services still maintain nowadays, there is a clear setback of their importance within the media content consumption. Fig. 1 shows how this consumption is decreasing over standard broadcasted TV while the view time via Internet is continuously increasing, especially among younger consumers averaged across United States population (millennials and teens (McNally & Harrington, 2017)).

In this regard, there are two main facts that give an idea of the current multimedia market evolution: on one hand, the increase of the Smart TVs purchase, that has tripled in the last five years (Statista, 2017), and, on the other, the increase of the consumers performing activities on a second screen as part of their viewing (81% during the traditional TV viewing, 72% during digital video streaming (eMarketer, 2017)), either to complement the content or to create new audio-visual experiences (Ziegler, 2013).

As a result of the above facts, a new disruptive environment opens up to facilitate new solutions that allows to fulfil consumers' needs and expectations. In this regard, the emergence of hybrid content provision technologies represents a relevant framework that considers these main factors and provides innovative solutions, especially with the definition and development of the new HbbTV standards, which combines broadband and broadcast delivery for digital multimedia content.

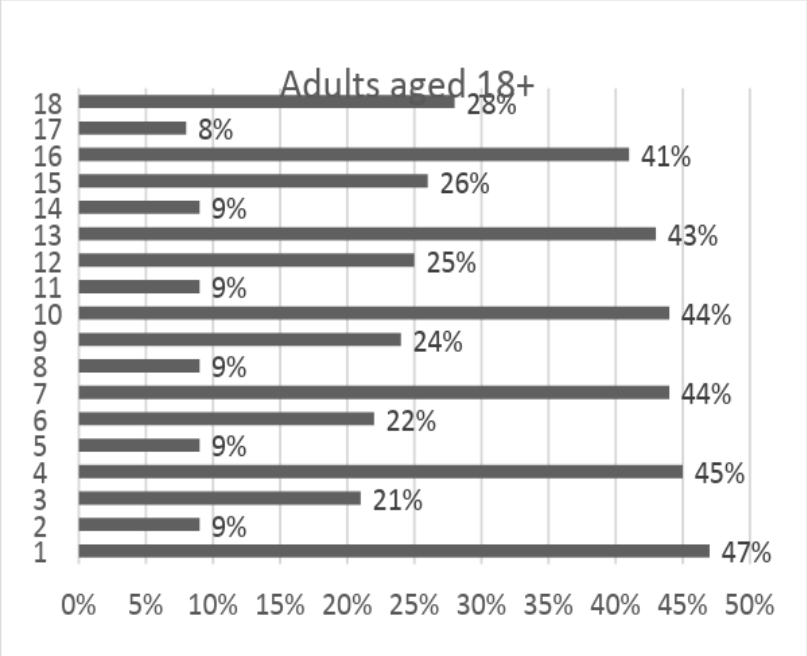


**Figure 1 - Average audience composition by platform among adults aged 18+ [from Nielsen report Q1 2016 (NIELSEN a, 2017), Q2 2016 (NIELSEN b, 2017), Q3 2016 (NIELSEN c, 2017), Q4 2016 (NIELSEN d, 2017), Q1 2017 (NIELSEN e, 2018), Q2 2017 (NIELSEN f, 2018**

The rise of Smart TVs penetration has also involved an increase of the HbbTV adoption all over Europe since almost the 70% of the TV sets are compatibles with hybrid television (Domínguez et al., 2018). Moreover, the adoption of HbbTV 2.0.1 (ETSI, 2016) as the new specification to be deployed, allows the connection between a hybrid television and second screen devices when they are connected to the same network (typically, using a WiFi connection). In this way, it is possible to create a different user experience with synchronized content in all devices and offer personalized services in the second screen applications. They are known as multi-device scenarios and there are many different use cases contemplated. The most typical one is when a user is watching a specific program through the DTT (Digital Terrestrial Television) channel on TV and, on the second screen and thanks to a broadband channel, the user can also access to synchronized additional services such as alternatives audios, multiview content, personalised advertisement, statistical data information or access services (subtitles, sign language, audio description or others that facilitate access to content for people with visual or hearing disabilities, or elderly people). This provides a disruptive scenario where the enhancement of the user experience will be the centre of the industry.

In relation with the accessibility aspect in this scenario, people with disabilities like watching TV, even considering it as an important family time activity. In this regards, their great desire would be to improve the television experience (as in the case of blind and visually impaired people (Woods & Satgunam, 2011), which can only be done by means of an efficient provision of complete access services for assuring an optimal media content access on an equal basis for all.

At European Union level, this issue has been recognized as an essential user right in the audio-visual sector, so the setting of different actions for its implementation has been defined as one of the priorities of the European Disability Strategy 2010-2020 (European Commission, 2010). In this context, European Union (EU) Regulators have imposed media accessibility to broadcasters, but the

specification about actual goals or number of hours of programming to be compliant to the directives have not been defined yet. Nevertheless, there are some countries such as Spain and France, as mentioned above, that have already provided specific regulation regarding this issue and they are making important efforts in analysing their fulfilment by the different broadcasters.

## 1.2. Lack of media content accessibility in TV consumption

Considering the current situation of the access services in TV scenario that has been briefly explained before, during the first step of the project we have defined and performed two specific focus groups that have helped us to better define our proposal, one for deaf people and one for visually impaired people. These focus groups were completed by a total of six super end users each, where super end user means experts that knows both the disability and the technology, with the aim of trying to avoid any biased result (more details about the methodology, the ethic requirements and the participants are provided in (Matamala et al., 2018). Following this approach, Table 1 shows some of the main frustration that blind and deaf people find when using the TV that have been collected within these focus groups.

**Table 1 - Most typical problems with the multimedia content access in TV obtained from the focus groups in EasyTV project.**

| | | |
|---|---|---|
| Blind and low vision people | Current experience using TV | • Not easy to access the TV<br>• Very difficult to use the remote control without audio feedbacks |
| | Frustrations advices | • Not enough audio descriptions<br>• It would be useful to have audio description on mobile devices (CS)<br>• Audio description is not useful for all programs (e.g. music programs)<br>• Teletext is not accessible.<br>• Overlay text in the content is not accessible<br>• It would be useful to slow the scrolling text and read it to the user.<br>• Possibility of stopping the image to see properly what is on the screen at this moment.<br>• It would be useful to magnify specific portion of the screen for a better recognition. |
| Deaf and hard of hearing people | Current experience using TV | • Low amount of content with associated subtitles (including emergency emissions)<br>• Low quality of the provided subtitles |
| | Frustrations advices | • Regarding the subtitles: not enough subtitles, no contextual information, lack of literality, delays and synchronization problems, low quality of the linguistic interpretation, small size, etc.<br>• Regarding sign language captions: incorrect placement or overlapping with on-screen signs, impossibility of switch on-off the sign language window. |

The services proposed in this project are derived from some of the lacks indicated by the users with visual impairments in the focus groups, since they are focused on the image analysis of the content. In this respect, tools to improve the visualization like subtitles adaptation tis presented in order to provide to the users some easy interactions to present the subtitles in a comfortable way. Some

icons definitions have been incorporated to the screen to know at each moment which tool is working. The face detection tool will be used for magnification purposes in order to help people with vision deficit to improve their perception of these specific elements in the content. Moreover, the application of different deep learning algorithms on the face detected will lead us to obtain information automatically that will improve the audio description information by identifying the people in the scene and their specific characteristics in terms of gender and age.

Once the definition of these new services has been justified, next section will present the theoretical background of the image analysis with deep learning techniques that are going to be applied for obtaining them.

According to the “European Disability Strategy” Report (European Commission, 2010), the number of people with disabilities in the EU will reach 120 million by 2020 and World Health Organization (Organization & others, 2013) estimated that 39 million people were blind in 2010 from a total of 285 million people visually impaired. For that reason, the promotion of affordable access to services in environments such as education, health or employment is a key factor for reaching equality among the citizens of the EU countries. The development of technologies for improving the accessibility is an important responsibility for developers and content distributors, because it is necessary for ensuring that people with disabilities have access to goods, information and multimedia services. And considering the current capabilities that the deep learning techniques are bringing to the audiovisual automatic information gathering, we should take advantage of applying them into this environment in order to obtain innovative services that may advance to break the access barriers. In this regard, and for this problem statement, next sections are going to present the different aspect that are the basis of our approach.

## 1.3. Deep learning for information extraction

With this objective in mind it is important to say that nowadays the application of different deep learning methods is outperforming good several results in this area, representing a powerful solution to be considered in this environment. Moreover, the availability of a huge quantity of annotated information is growing exponentially and is the key concept that allow deep learning algorithms to improve their performance, so the chance is absolutely clear.

The application of computer vision tools for analysing different human behaviour and characteristics have been studied along the time, resulting in three main research areas: human detection, human recognition and human tracking. And, based on these three areas, it is possible to extract important information to perform more complex tasks as human activity recognition or pose estimation.

Computer vision algorithms are mainly based on a mathematical background and sequential steps that provide the final estimated result for the problem under study. Some years ago, some of these algorithms started to include machine learning techniques to learn about the extracted features from the algorithm in order to perform a final classification. Examples of these applications can be used for background subtraction (for movement detection in consecutive frames) in combination with support vector machines (to classify the detected movement in different categories) (Ahmed, Kpalma, & Guedi, 2017; Xu, Xu, Li, & Wu, 2011), Haar cascade classifiers for face detection (Cuimei, Zhiliang, Nan, & Jianhua, 2017; Padilla, Filho, & Costa, 2012), semantic segmentation of objects in a scene (Yuheng & Hao, 2017), etc.

All above examples have been achieved with the use of computer vision in combination with machine learning techniques, providing interesting results in a wide amount of particular applications. Nowadays, all these results have been improved by the application of deep learning algorithms in order to solve complex computer vision tasks. In this regard, a huge variety of object detection algorithms (Faster-RCNN -Faster Region-based Convolutional Neural Network- (Ren, He, Girshick, & Sun, 2015), SSD (Single Shot Detector) (W. Liu et al., 2015), YOLO (You Only Look Once)(Redmon, Divvala, Girshick, & Farhadi, 2015), RetinaNet (Lin, Goyal, Girshick, He, & Dollár, 2017)) are providing very accurate results in detection tasks.

In the field of tracking with computer vision and deep learning techniques, there are two main principal approaches: online tracking, which is related to the estimation of the next state based on the previous state by a direct mathematical analysis (KCF -Kernelized Correlation Filter-(Henriques, Caseiro, Martins, & Batista, 2014), MOSSE (Minimum Output Sum of Squared Error)(Danelljan, Häger, Khan, & Felsberg, 2014), CSRT (Discriminative Correlation Filter with Channel and Spatial Reliability) (Lukezic, Vojir, Cehovin, Matas, & Kristan, 2016)) or offline tracking, which is based on the previous extraction of different patterns about the objects under tracking to be later used in a new sequence of frames (Siamese Fully Convolutional networks (Bertinetto, Valmadre, Henriques, Vedaldi, & Torr, 2016), GOTURN (Held, Thrun, & Savarese, 2016), Re3 (Gordon, Farhadi, & Fox, 2017).

Finally, another important research area is related to keypoints detection and image segmentation. In computer vision several algorithms for feature extraction and keypoints detection (Hassaballah, Abdelmgeid, & Alshazly, 2016) (SIFT -Scale Invariant Feature Transform-, SURF-Speeded-up Robust Features-, ORB-Oriented Fast and Rotate Brief-) in combination with machine learning techniques (for optimization, regression) have been applied. For image segmentation, algorithms like k-means clustering, mean-shift clustering or interactive image segmentation can be used. Furthermore, deep learning methods are solving these complex tasks thanks to the availability of wide datasets with annotated keypoints and masks for image segmentation. OpenPose (Cao, Simon, Wei, & Sheikh, 2016) or DensePose (Güler, Neverova, & Kokkinos, 2018) are now at the top of body keypoint detection algorithms and Mask-RCNN (He, Gkioxari, Dollár, & Girshick, 2017) is able to perform object segmentation.

# 2. SUBTITLES ADAPTATION

The subtitles adaptation is a service integrated on the Companion Screen App. It is designed to let the visual impaired or color blindness users to adapt the text size, the color and the combination of foreground and background colors of the subtitles or adapt them automatically depending on their capabilities detected on the user model.

The adaptation of the subtitles is achieved technically by using the CSS technology combined with javascript. By default, the settings are established by the user model, but the users are able to change these settings manually to better fit their needs.

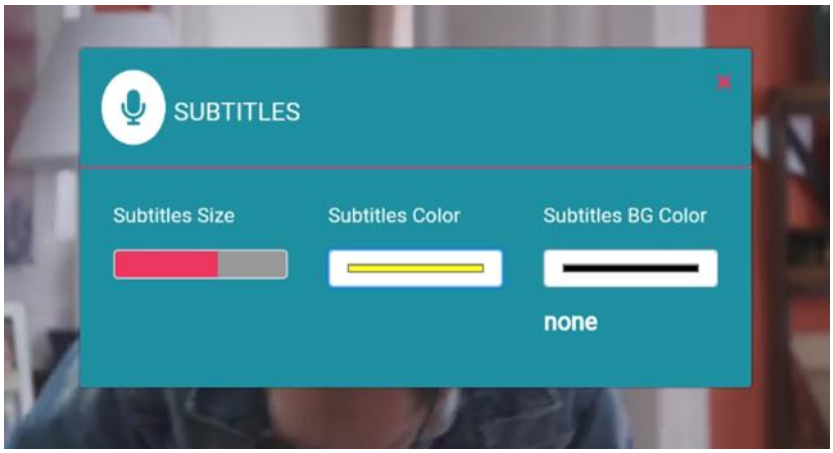


**Figure 2 - Subtitles Customization pop up.**

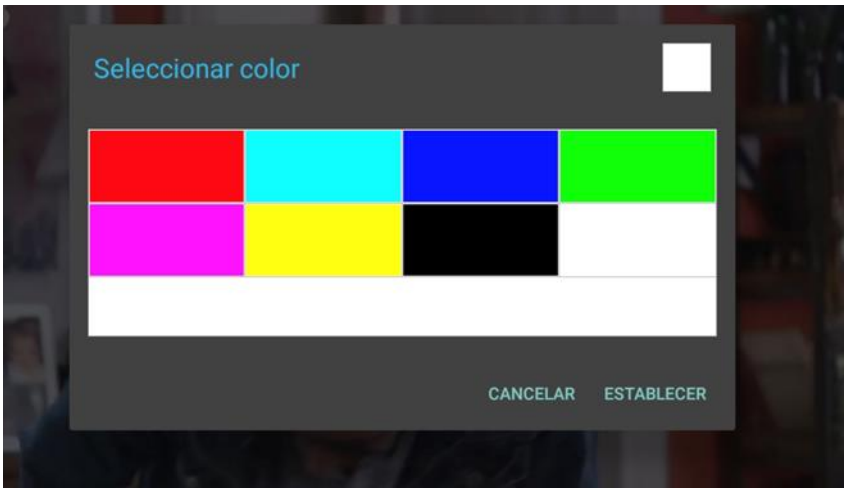


**Figure 3 - Font-color and background color pop up,**

# 3. SERVICES INDICATORS

To let the user, know which accessibility services are available for each content we are using the four standardized icons designed by Heidi Sivebæk, (https://www.dr.dk/om-dr/about-dr/smart-icons-design-common-european-standardization) described below:

- Subtitles: this indicator is an equal sign inside square brackets [=].
- Spoken subtitle: this indicator is an “o” inside square brackets [o].
- Audio description: this indicator is two dots inside square brackets.
- Sign Language: this indicator is a less-than sign inside square brackets [<].

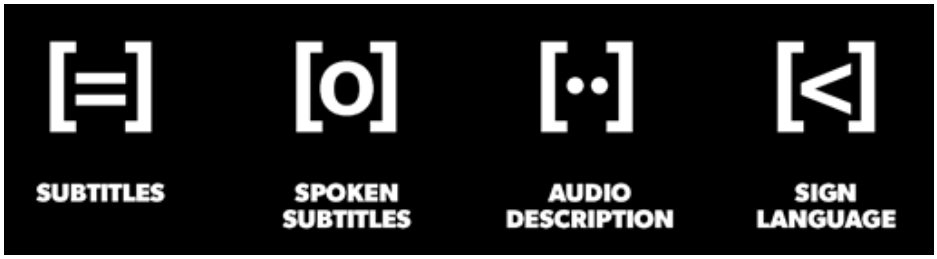


**Figure 4 - DR icons for accesibility services.**

For the new services developed on the EasyTV project; face magnification, text detection, character recognition and sound detection, we have designed four new icons:

- Face magnification: this indicator is a colon symbol and close parenthesis inside square brackets [:)] and try to represents a face.
- Text Detection: this indicator is an “o” followed by a dash inside square brackets [o-] and try to represents a magnification glass.
- Character recognition: this indicator is a question mark inside square brackets, and try to represents the idea of the question of who is in the scene.
- Sound Detection: this indicator is a mayor-than sign followed by a close parenthesis and try to represents the amplification of sound on an ear.


**Figure 5 - EasyTV designed icons.**

Instructions of How to type an icon

- Select Arial or Hevetical typography in bold version
- Use background color white and font-colour or other contrasting colors
- Type one of this:

[=] AltGr + 8 Shift + 0 AltGr + 9

[o] AltGr + 8 Shift + o AltGr + 9

[··] AltGr + 8 Alt + 0183 Alt + 0183 AltGr + 9

[<)] AltGr + 8 < AltGr + 9

[:)] AltGr + 8 Alt + 58  Shift + 9 AltGr + 9

[o-] AltGr + 8  o Alt + 45 AltGr + 9

[?] AltGr + 8 Alt + 63 AltGr + 9

[>] AltGr + 8 Shift + <  Shift + 9 AltGr + 9

# 4. DEEP LEARNING TECHNIQUES FOR VISUAL INFORMATION GATHERING

## 4.1. Proposed system for data extraction

Bringing together the feedback from the focus groups about which kind of service could be interesting for final users and the capability of innovative deep learning techniques for visual information gathering, we propose two main services for enhancing the accessibility of media content in the TV environment. we have defined a specific scenario in which users with disabilities can consume TV content with people without disabilities on equal basis by using a combination of advance image processing methods and HbbTV. In this scenario a TV is connected via HbbTV to a companion screen in which the user can make use of these accessibility services. At this stage it is important to note that these solutions are defined to be applied to broadband pre-processed content, considering their application to broadcast delivered content as a possible future research work which will be fully dependant on the existing software and hardware computational capabilities.

As it was mentioned before, we have followed a user-centric process for defining our approach. In this regard, the methodology applied can be divided into three main stages as follows:

Firstly, and as mentioned before, two focus groups were arranged to collect the user's requirements. Derived from them, two main services focused on the image processing were defined: image magnification and character recognition.

Secondly, a complete framework was created to extract image information from the multimedia broadband delivered content to feed the two access services. The proposed architecture uses several artificial intelligence and deep learning techniques trained with existing datasets together with information from our media contents to better fit the main information retrieval techniques:

- Face detection: this part is intended to find the location of the character's faces over the video.
- Face tracking: the goal of this component is to achieve a smoothed tracking over faces for the image magnification service.
- Face recognition: the main purpose of this module is to detect and recognize different characters along the video.
- Face speaking: this part is in charge of estimating which character is speaking.

For obtaining these modules, different combinations of existing fine-tuned frameworks were applied into our custom dataset in order to achieve real or very near real-time processing. Our framework demonstrated that all these algorithms can work together and provide better access solutions when retraining these networks with the targeted new data.

With the results of these modules, new access contents have been generated and included in a real companion screen application to provide the magnification and character recognition services. Iterative testing's with end users will be set during next months to obtain their feedback in order to improve them for a final version.

## 4.2. Video analysis for facial information detection

The modular architecture for the data extraction for face detection service can be seen on Fig. 6, and it can be divided into three main phases:

- Face detection: this step is in charge of applying specific deep learning networks in order to find the faces within the different frames of the video. It includes a scene detector module for detecting the changes between scenes, which may help the face tracking along the video.
- Face tracking: this step is in charge of applying deep tracking over the detected faces for improving the computational time related to the detection process.
- Information extraction: this is the final step and it is in charge of analyzing different aspects of the faces in order to provide additional information to deploy the access services. It includes character recognition, face characterization and speaking detection.

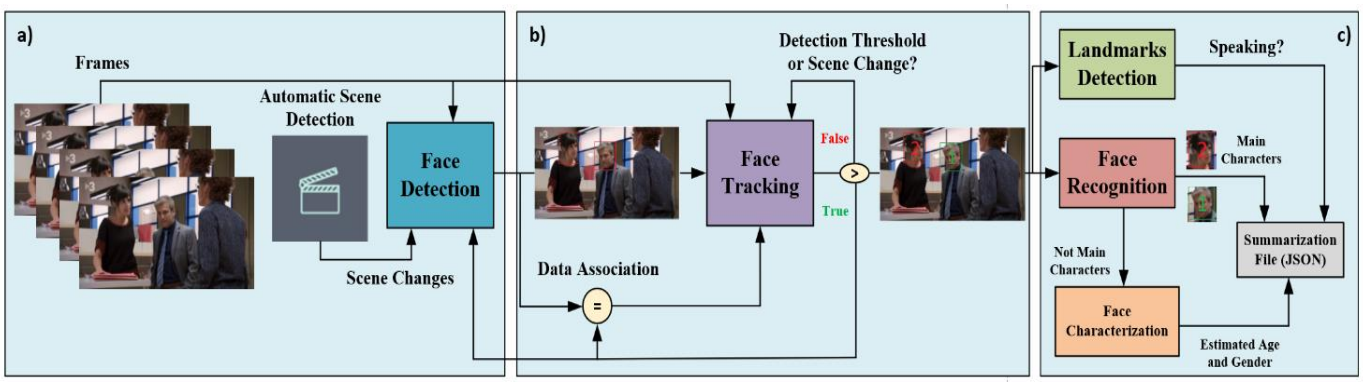


**Figure 6 - Main architecture for data extraction. a) Face detection b) Face tracking c) Information extraction, including face recognition, characterization and speaking detection.**

All the video contents are processed frame by frame to extract information from the whole video. Detection parts are applied at frame level and tracking parts are applied at frame level but taking into account previous and next frames information. After the extraction of all the involved data another process cleans the information using temporal connections between each scene in the video.

On this basis, the contents are preprocessed and the obtained accessibility content are integrated in the final broadband flow in order to allow the broadcasters to provide the access services to the end users.

To apply all these techniques some important hardware requirements are needed in order to process the information according to the main ideas presented in this work. All the results have been obtained using one dedicated server with 64GB of RAM, Intel core I7 processor of 8 generation and one Nvidia GeForce 1080 Ti GPU.

## 4.3. Face Detection

Face detection is a well-known problem in computer vision and it represents the center of our approach. During the last years, several methods have been already proposed in order to extract patterns from images and detect faces from them (S. Chakraborty & Das, 2014). In general, all these methods consist of three parts: first, an algorithm to inspect parts of the images (sliding window, region proposal), second, the obtaining of extracted features from these parts (Haar features, Histogram of Oriented Gradients features, deep learning features) and finally their classification whether they're a face or not using machine learning (support vector machine, adaboost…) or deep learning (Artificial Neural Networks).

Traditional methods rely on how we can model the features or the patterns manually trying to find edges, blobs or other interesting patterns with the aim of defining different features and classify them. However, recent studies such as (Wang & Deng, 2018) show that it's better to delegate those tasks to the computer and let them learn by themselves. In this context, convolutional neural networks (CNNs) are changing the concept of feature extraction in computer vision, since they are able to learn during the training process which parameters are needed to extract complex features that can define different types of objects in a way that is almost impossible manually.

Deep learning object detectors are the key point for finding faces in images efficiently (Wang & Deng, 2018) and rely on the use of GPUs processing and of wide and good-annotated datasets. However, the object detector selection depends heavily on the task to be performed in order to achieve high accuracy in detection with final high Intersection over Union (IoU) and with less computational cost and processing time. In this field, Faster-RCNN was the first popular object detector based on a complete deep learning architecture, but its main problem is related to the high computational and time cost related to the Region Proposal Network (RPN) and the final classification and regression branches in the algorithm. Subsequent architectures that have been defined to deal with these problems have provide less accuracy results than faster-RCNN but with a low computational cost and time. These are the cases of SSD which combines the task of selecting Regions of Interests (RoIs) in a single end to end architecture and YOLO which is a good option if requirements related to real time processing are high.

Backbone for feature extraction (convolutional part in the network) selection in these networks is also important because more complex networks can extract more complex features using more computational resources. VGG (Simonyan & Zisserman, 2014) and ResNet (He, Zhang, Ren, & Sun, 2016) are the most commonly used pertained backbones on ImageNet dataset. In order to perform face detection, only one type of object will be classified so not a very deep network is necessary to perform feature extraction, thus restricting the backbone use to just ResNet10 or VGG16. Another important consideration is to achieve real or near real time face detection processing where the use of SSD and YOLO are the best options for this purpose.

On the basis of this information, and taking into account the results obtained from the application of the deep learning techniques for face detection, our approach applies them to define new services for improving the intelligibility of these image areas, where the information of the content can be found. In this regard, along the discussion section we present a comparison with other existing frameworks that use the concept of "Tracking by Detection" (Fiaz, Mahmood, & Jung, 2018; Luo et al., 2014), which applies an unique network for tracking and detection at the same time, such as ROLO (recurrent YOLO) (Ning et al., 2017), Mf-SSD (Broad, Jones, & Lee, 2018) and D&T (Detect & Track) (Feichtenhofer, Pinz, & Zisserman, 2017).

This is the main part of the entire architecture, since if no faces are detected, no postprocessing can

be done to extract related information from the video under analysis. For this reason, and in order to enhance the process, a well-known dataset such as FDDB (Jain & Learned-Miller, 2010) for face detection has been used for training the algorithm.

Additionally, two object detectors have been selected in order to find an optimal balance between accuracy and computational consumed time. SSD and YOLO have been used as the fastest algorithms to perform object detection. SSD have been trained using VGG16 and ResNet10 backbones for decreasing the complexity in the network, achieving a good tradeoff between time and accuracy. From YOLO, version 2 has been selected, due to its better performance in comparison with version 1 and a simpler architecture than in version 3. The results after training are presented in Table 2, where Frames Per Second (FPS) rate and Mean Average Precision (mAP) values are collected. Moreover, Fig. 7 shows some examples of the detected faces over different video frames.

**Table 2 - Results after training for SSD with two different backbones and YOLOv2 with their original architecture**

| Networks / Metrics | Frames Per Second (FPS) | Mean Average Precision (%) |
|---|---|---|
| SSD300 (input 300x300) with VGG16 Backbone | 46 | 74.3 |
| SSD300 (input 300x300) with ResNet10 Backbone | 39 | 77.8 |
| YOLOv2 (original implementation) | 57 | 64.4 |

The next step is to sample the video for detecting where the scene change appears. This is an important task for characterizing each scene with the appropriate information. Moreover, it also gives essential information about when the tracking algorithm should stop and start over with a new face detection in the next scene. Our approach uses an existing framework known as PySceneDetect (Castellano, 2018) that is able to detect scene changes in videos and automatically splits the video into separate clips. This tool offers several detection methods from simple thresholding to advanced content aware fast-cut detection.

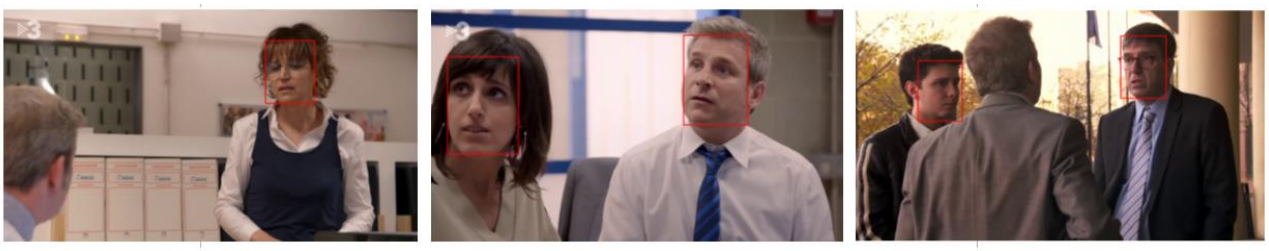

**Figure 7 - Example of detected faces in video.**

## 4.4. Face Tracking

This second step is related to one of the main purposes of the proposed architecture, which is to provide information about where the faces are located for their magnification. If it was only based on the face detection algorithm, different important problems would appear, such as:

- Bounding box size: since it is not fixed for the face detector along the scene, the face magnification process will be noisy due to the fact that the center of this box is the one selected for performing the magnification.
- Detection loss: face detector is able to detect faces in single images but it may present different problems in complex situations, such as when the face is not pointing directly to the

camera, reducing its accuracy along the scene time. Tracking over the detected faces solves this problem by maintaining where the face is located if no detection is achieved.

- Computational time, since deep learning detectors are usually slower than other solutions such as deep tracking. In fact, while face detector is about 39 frames per second (FPS), tracking algorithms are able to work up to 100 FPS, thus increasing the processing speed for the entire architecture.

Considering these main issues, our proposed architecture includes tracking algorithms for speeding up the process by performing face detection only every fixed number of frames or after a scene change. In this regard, once a face is detected, it is linked with a tracker from previous frames if possible, if not, a new tracker is created, avoiding duplicating trackers. Moreover, when a scene change happens, the previous trackers are deleted.

Four trackers have been tested in order to select the best performance with less computational time. The first two options are considered as online trackers, which work directly over the frame sequence by estimating the next state using previous information. In this context, Kernelized Correlation Filters (KCF) makes use of an adaptive threshold approach, which includes a kernelized correlation filter method and Kalman filter algorithm to make tracking faster and more accurate. CSRT tracker takes the advantage of discriminative correlation filters and provides a novel learning algorithm for its efficient and seamless integration in the filter update and the tracking process.

With the aim of comparing the performance and results of these two trackers, two more offline trackers have been tested. The first is known as GOTURN tracker and is the first deep learning Tracker using CNNs that showed accurate results with a huge variety of objects. The last algorithm tested is known as Re3, a real-time deep object tracker capable of incorporating temporal information into its model and one of the fasters tracking algorithm for single tracking.

Our experiments use pretrained models created with GOTURN and Re3 regression networks, which use annotated motion datasets that contain many objects including faces for training, as there are not specific datasets for tracking faces. Our own dataset has been annotated using the face detection information between consecutive frames. A dataset with 1000 pairs of faces have been used to retrain the models to better fit in the face tracking task. For the case of KCF and CSRT, and as they are online trackers, there is no need of additional data. In another aspect, several sequences have been annotated in our own videos to test the accuracy in the trajectories obtained with all the algorithms. A total of 100 video sequences from different contents have been used to perform the testing's. Table 3 presents some metrics to evaluate the performance of these algorithms, considering that mean pixel error have been obtained using the Euclidean distance between the annotated bounding box center and the estimated bounding box center given by the tracking algorithms.

**Table 3 - Mean pixel error and processing time depend on the number of trackers at the same time**

| | Mean pixel error | Processing Time (1 Tracker) | Processing Time (2 Trackers) | Processing Time (≥4 Trackers) |
|---|---|---|---|---|
| **KCF** | 18 pixels | 12 ms | 13 ms | 16 ms |
| **CSRT** | 16 pixels | 40 ms | 49 ms | 61 ms |
| **GOTURN** | 35 pixels | 27 ms | 28 ms | 30 ms |
| **Re3** | 23 pixels | 8 ms | 14 ms | 25 ms |

According to the information in Table 3, CSRT algorithm provides the best results in terms of accuracy, due to its efficiency with slow motion objects and with occlusions problems but it is slower compared to others algorithms, thus losing the real time processing capabilities. On the other hand, GOTURN tends to cover a wide area around the face after some frames, which results in a lower accuracy and a smoothing tracking results provision. Re3 gives high accuracy in the trajectories with the best processing time if only one face is detected in the image but it linearly increases with the number of detected faces. Nevertheless, the performance of these two algorithms can be improved by training both with a specific face's dataset. Finally, KCF presents high accuracy and the processing time increases slowly with the number of trackers, but its main problem is the high probability to fail in presence of important occlusions that hide a wide area of the tracked face.

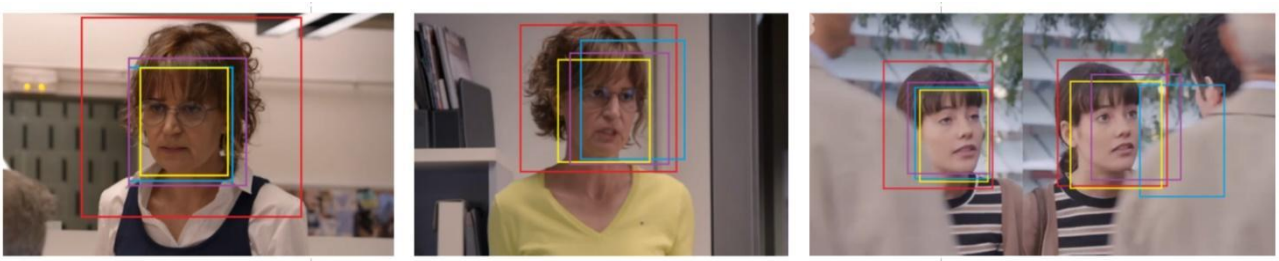

**Figure 8 - Failure examples during tracking with tested algorithms. GOTURN (red), CSRT (yellow), KCF (blue) and Re3 (magenta). Slow motion (left), fast motion (centre) and occlusion (right).**

Fig. 8 shows some examples of failure with each algorithm. Left image represents a situation with slow motion, where all algorithms work well but GOTURN bounding box is bigger than the others. Center image presents one frame after a fast movement of the face along some frames, where KCF and Re3 trackers have been displaced and now are not centered in the face, while CSRT and GOTURN maintain the bounding box centered. Finally, right image presents the situation before and after an occlusion. Re3 tracker bounding box has suffered a little displacement but KCF confidence decrease dramatically, losing the tracker and waiting for new face detections to start again...

According to these results, KCF has been selected to be included in our approach due to the obtained accuracy, processing time and performance. Considering that if the tracker is lost, a new face detection is needed, this is a good solution for solving occlusions in the video...

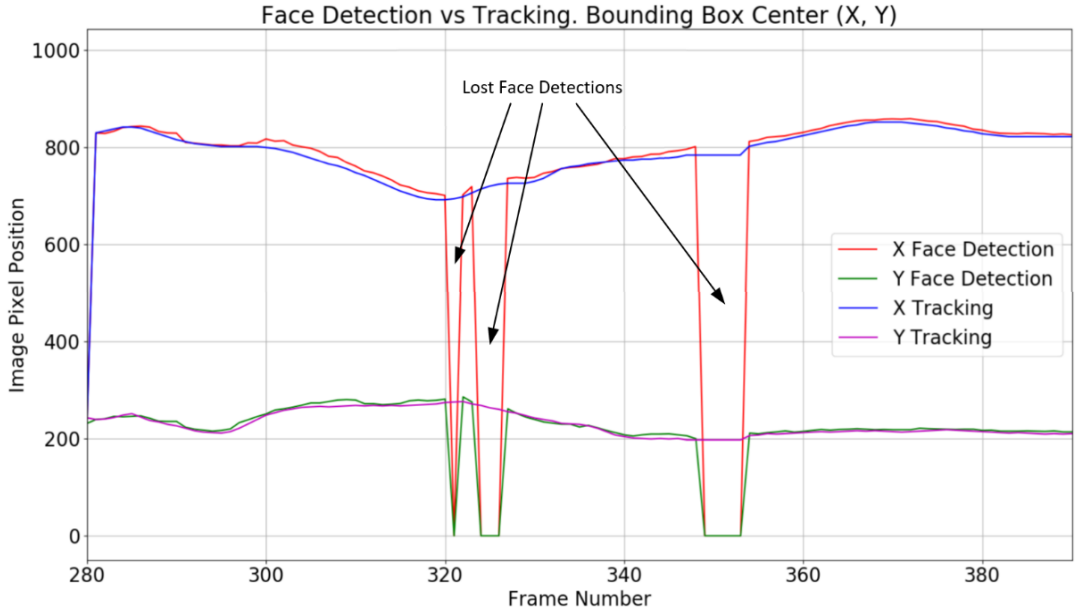


**Figure 9 - Performance comparison between the face detector applied in all frames and the tracking over the first detected face.**

Fig. 9. shows an example of the comparison of the results obtained by using face detection vs face detection plus tracking. The curves present (x,y) coordinates of the center of the detected face of both approaches, where the first one gets the coordinates of the detected face frame to frame while the second one starts with a face detection in the first frame of a scene, followed by face tracking. When using face detector, some faces are lost for some frames during a scene. This makes that the faces will be continuously reduced and magnified, giving a bad experience to the user. Regarding the face detection plus tracking algorithm, two important results can be observed: firstly, if the face is not detected by the face detector, tracking algorithm is able to continue detecting it. Secondly, the tracking curve is smoother than in the first solution. These two facts make the algorithm filter better the trajectory avoiding very noisy transitions.

## 4.5. Face Recognition and Characterization

In order to provide more information about what happens in a scene, the proposed architecture includes a module that is able to perform information extraction that includes face recognition and characterization.

The input for this module is the cropped face from the previous estimations after tracking algorithm, and it uses a convolutional neural network trained for classification tasks over the faces. The output is the same that the labels in the applied network, which are the names of the main characters and one more label identified as “unknown” for others. Considering this, when one of the main characters is recognized, the presentation module will display his related information. In contrast, if some other faces are detected and labeled by the network as “unknown”, these faces pass along another branch of the network in order to estimate the age and gender to collect a general information for each scene.

To train this network, a set of different faces of the same person in different situations have been collected and annotated. In this regard, our dataset includes more than 60.000 images for training (divided into 17 different classes) and 15.000 for testing. Fig. 10 presents an example of some of the main characters together with a set of different images of only two of them.

For train the second branch of the network, which is in charge of providing specific information such as the age and the genre of the detected face, some existing datasets were collected and mixed for improving the results. Particularly, the datasets used were UTKFace (Zhang Zhifei & Qi, 2017) (more than 20000 images labeled with age, gender and race), APPA-Real (Agustsson et al., 2017) (7591 images with several labels per image with the apparent age voted by 38 persons) and finally IMDb-Faces (Rothe, Timofte, & Van Gool, 2018) (more than 460000 images of famous people annotated with age and gender).


**Figure 10 - Examples of the main characters (top) and some images in our dataset for some of them (bottom).**

The network architecture is divided into two branches where the second branch is activated only when a detected face is not a main character, as shown in Fig. 11. To compose face recognition network, different backbones have been tested with a classification network, which contains two neurons layer of 1024 neurons each, and a final SoftMax activation to get the final prediction. The input has been resized to a fixed size of 224x224.

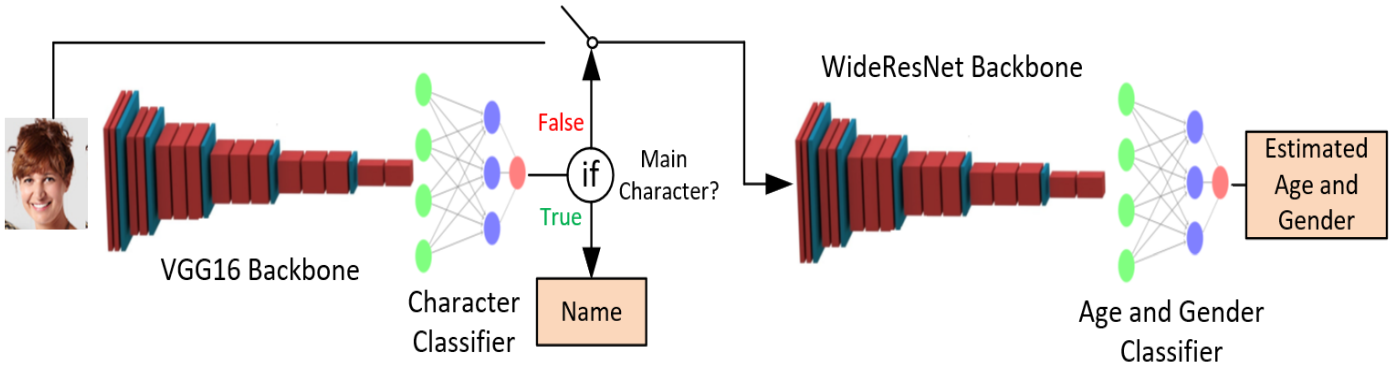


**Figure 11 - Architecture for face recognition and characterization network.**

The second network used for face characterization uses the WideResNet network (Zagoruyko & Komodakis, 2016) as backbone, due to its power to work with a huge number of classes and its capacity to reduce computational cost when training with big datasets. It uses the same concept as the ResNet architectures but improving the training speeds to get similar results on complex datasets with much more classes. This network has been trained with 100 classes (0 to 99 years) and 2 more classes for gender estimation (male and female). The classification network has been developed to concatenate directly the age and gender predictions in order to get just one output instead of creating two separate classification networks for each task.

**Table 4 - Training, validation and prediction time results for tested backbones on face recognition network and results for face characterization network**

| Backbone / Train-Test Precision | Train Precision | Validation Precision | Prediction Time |
|---|---|---|---|
| VGG16 (Recognition) | 95.2% | 92% | 2 ms |
| VGG19 (Recognition) | 97.5% | 88% | 3 ms |
| MobileNet (Recognition) | 89.1% | 88.2% | 2 ms |
| ResNet50 (Recognition) | 98.2% | 56.3% | 4 ms |
| WideResNet (Characterization) | 86.3% | 80.4% | 7 ms |

Table 4 shows the results retrieved after training the network for face recognition using different backbones and the results achieved after training the Face Characterization Network. The selected backbone for face recognition network was VGG16. Analyzing the final values, it is possible to extract some valuable information about the behavior of each network. VGG16 reaches high accuracy in

both training and validation sets. This network is less complex than other tested networks and there was not overfitting problem during the training process. VGG19 starts to present overfitting. This information can be extracted directly by the validation precision achieved. This network is more complex, does not generalize right all the classes and also performs worst at validation time. MobileNet (Howard et al., 2017) is a less complex architecture that works well but does not achieve the accuracy presented by the previous ones. ResNet50 is the clear example of very deep network that overfits quite quickly and performs poorly at validation time over this dataset which is not very complex.

After training and validating the algorithm in the presented datasets, some real tests have been performed over the custom content with known characters' age and gender. The mean deviation obtained in these tests is 3.47 years for the age, while the gender classification shows an accuracy of 97%.

According to the obtained results, face characterization network achieves a very high accuracy in both training and validation, thus confirming that WideResNet was a good selection for this problem.

## 4.6. Face speaking detection

Detecting which character is speaking in a scene is an important information and could make the difference for users with visual disabilities. The knowledge of who is speaking, for someone who is just listening the media content, could make him to better understand the scene. This knowledge is also very useful in order to magnify the face of the speaking character. Once the faces in the scene have been detected, next step is to know who is speaking automatically. In our approach, this process is done by making use of image analysis, focusing on detecting the mouth and its movements using facial landmarks.

Landmark detection in faces is a well-known problem. Before deep learning algorithms started to solve these problems offering an increase on accuracy, landmark detection was performed by using a combination of traditional computer vision techniques to extract facial features with some optimization or machine learning algorithms to learn where the points should be located (A. Liu et al., 2011; Monzo, Albiol, Albiol, & Mossi, 2010).

The availability of several huge datasets containing the location of these points in faces provides the necessary data to use deep learning in order to obtain high accuracy predictions. The dataset used in this work is known as 300 Faces In-The-Wild Challenge (Sagonas, Antonakos, Tzimiropoulos, Zafeiriou, & Pantic, 2016). It contains re-annotated landmark points (68 in total) in several available datasets (LFPW (Belhumeur, Jacobs, Kriegman, & Kumar, 2013), AFW (Zhu & Ramanan, 2012), HELEN(Le, Brandt, Lin, Bourdev, & Huang, 2012) and XM2VTS (Messer, Matas, Kittler, Luettin, & Maitre, 1999)) using their own annotation tool.

The network used is MobileNet, where traditional convolutions have been replaced with Depthwise Convolutions (Howard et al., 2017). This greatly reduces the number of parameters that are required while still keeping efficiency and not destroying cross-channel features. This is a good starting point to work in landmark detection task due to our labels are the pixel coordinates of each point then not very complex feature information is needed from the Convolutional layers. The top part of the network was replaced with a neural layer with 136 neurons (total of coordinates for the 68 points) and trained minimizing the Smooth L1 loss in order to minimize the distance between the estimated points and the real ones.

Our approach determines when a character is speaking by tracking the changes of the mouth in a face between frames. When a mouth is making fast movements in a succession of frames on a scene, we make the assumption that this face is speaking. As the faces are already been tracked, we can assure if a mouth is the same in two consecutive frames. To detect changes in a mouth, we check the landmark detection points of the mouth. We use the normalized distance between two key points from the mouth in a frame, and compare it with the same key point from the same mouth in

the previous frame. We have considered seven distances that interpret the position of the mouth as shown in Fig. 12.

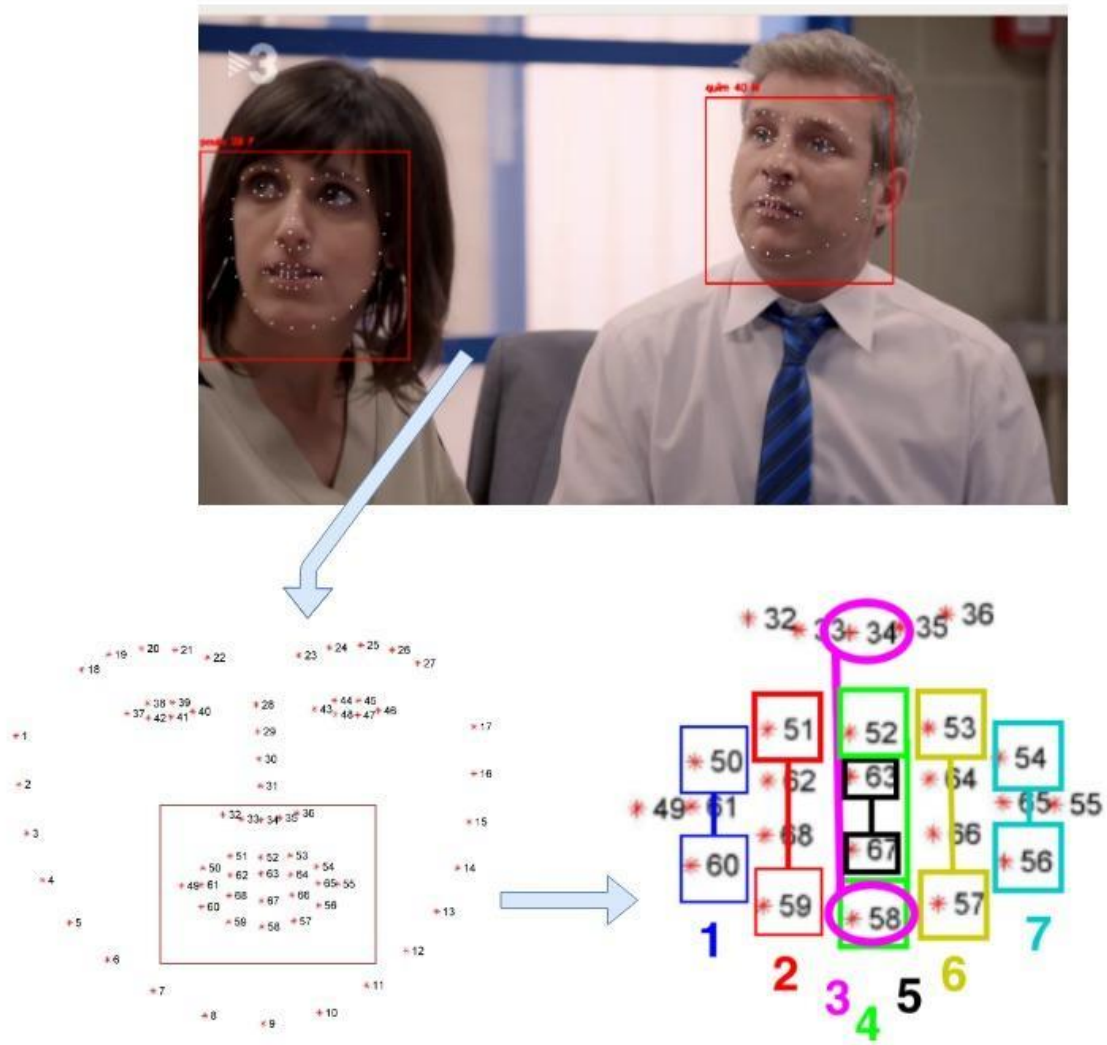


**Figure 12 - Landmark detection and key points pairs difference.**

We have defined the difference between two mouths in two consecutive frames D_(i,j) as the sum of the absolute values of the differences between the distances where j=i-1 for all i>0.

$$D_{i,j} = \sum_{k=1}^{7} |d_{i,k} - d_{j,k}| \forall i > 0$$

Knowing the difference of mouths between pairs of consecutive frames of a full scene can give an idea of the behavior of a mouth during that scene. We made an experiment with a video where some characters appear on the scene and are speaking between them. Fig. 13 shows the difference of two mouths D(An, (n-1)) and D(Bn, (n-1)) along the frames of the video, where A is the character A, B is the character B and n is the frame number. The line in red shows the difference of the mouth of the character A, the blue line shows the difference between the mouth of the character B, and the green lines indicate the median of the values of each character during 75 frames. It can be observed that when the character is speaking the green line is higher than when he is not. In the first approach we selected the highest median of a not speaking mouth as the threshold to determine if a character is speaking. After 10 experiments with different scenes, we adjusted the threshold by using the average of all the thresholds retrieved to 0.01335. We assumed that medians below the threshold will be considered not speaking faces and medians above will be considered speaking faces. The

average threshold used in 10 different videos detected 8245 correct frames from a total of 9075, achieving an accuracy of 90.85%.

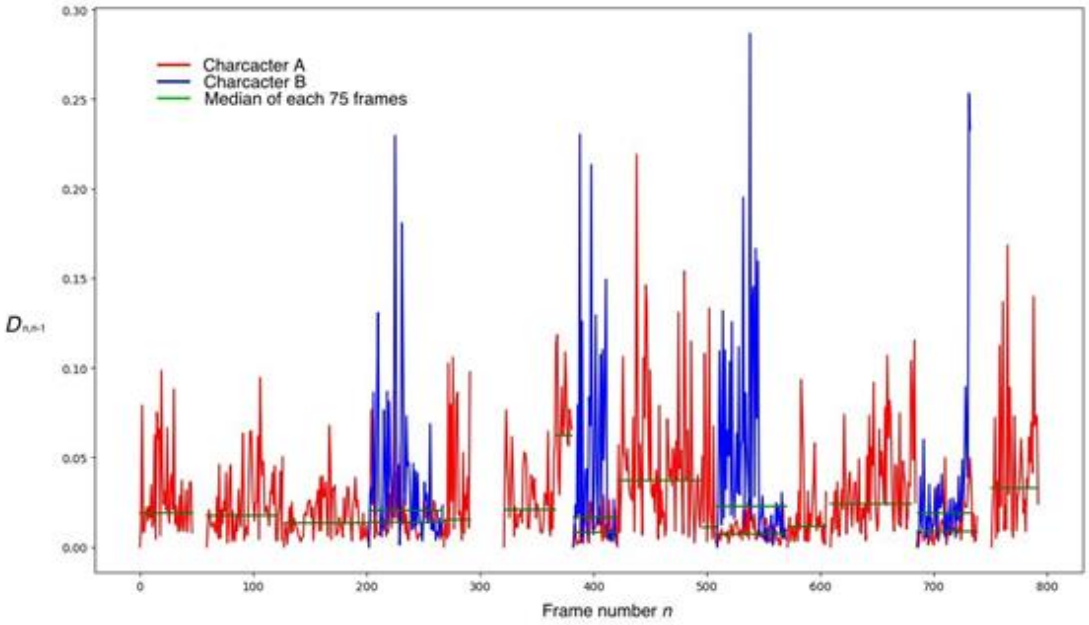


**Figure 13 - Mouth landmark points difference along consecutive frames of two characters.**

## 4.7. Information storage

### 4.7.1. Data structure

The final extracted data is stored to be used by the user application to show and perform all the information and required actions that que user needs at each moment. The structure of the information in the main server is presented in Fig. 14 where the folders and files are shown in a tree.

Initially, a main folder named “media” contains all the EasyTV contents in terms of series, news and each additional content that need the processing step in order to extract the main useful information to show in the user application.

Inside a folder which represents a series or other content there are another set of folders representing each chapter. Additionally, one folder called “Dataset” contains all the data necessary to train the recognition algorithm in each concrete set of contents and will contain the trained model after finish the annotation process. This folder is created after annotate several contents with the provided tool in the main interface. Regarding the faces dataset it is composed by one folder for each main character annotated and around 1000 face images related to the concrete character. Furthermore, another folder with the character “Unknown” represents all the characters that are different of the main annotated ones.

Finally, inside each chapter there are three main JSON files related to the process of the faces for magnification and recognition. The first one contains all the extracted information about the complete video. The second one provides a resumed JSON file with the necessary data to presents correctly the information in the annotation tool. The last JSON file contains all the information necessary for the user application related to face magnification and faces recognition in a summarized way. The content of this JSON file is described in the next section.

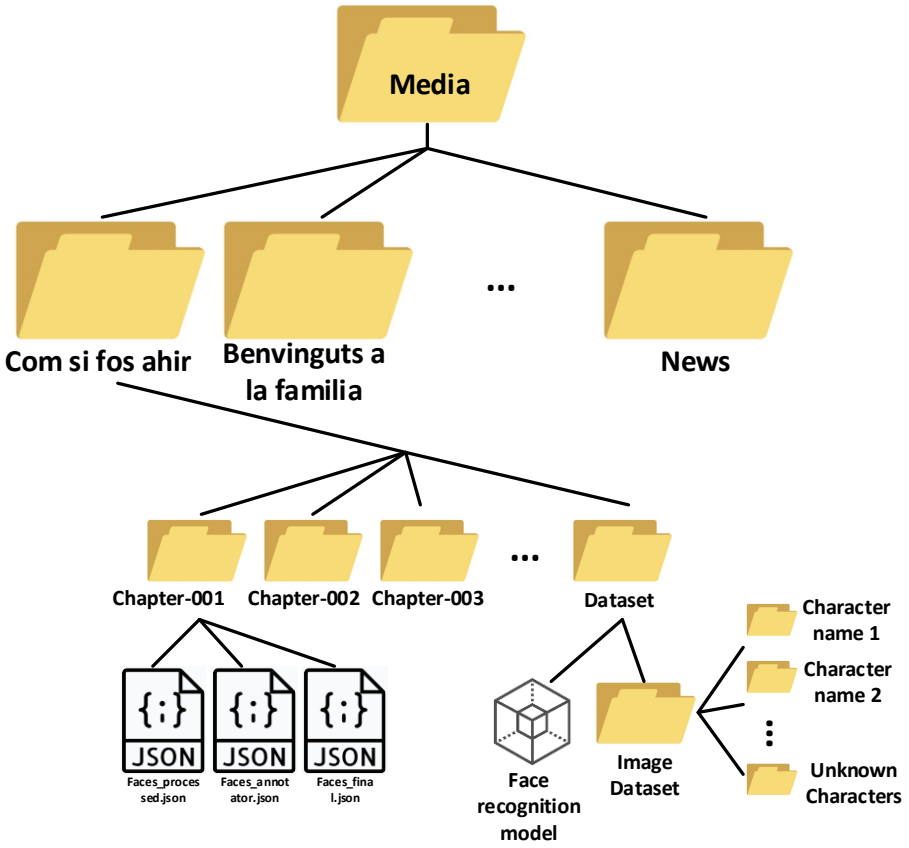


**Figure 14 - Saved information structure.**

### 4.7.2. **Annotation tool for faces data extraction**

Following the structure presented in the previous section it is possible navigate along the media contents in order to annotate some of the videos to train the character recognition network to detect main character in each different tv show, program or news. Fig. 15 shows an example of the main media content folder where several TV shows and news programs are available.

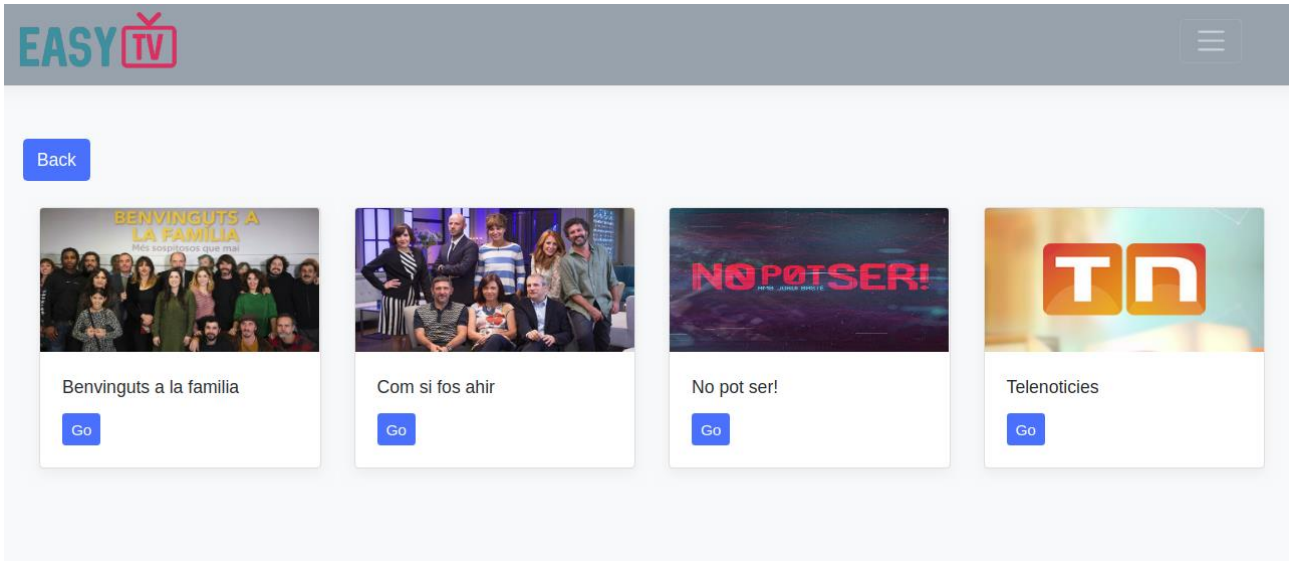


**Figure 15 - Main media contents in the EasyTV access services interface.**

Inside a concrete TV show (for example “Com si fos ahir”, Fig. 16) are available a set of different chapters that can be annotated easily by the developed tool for this purpose.

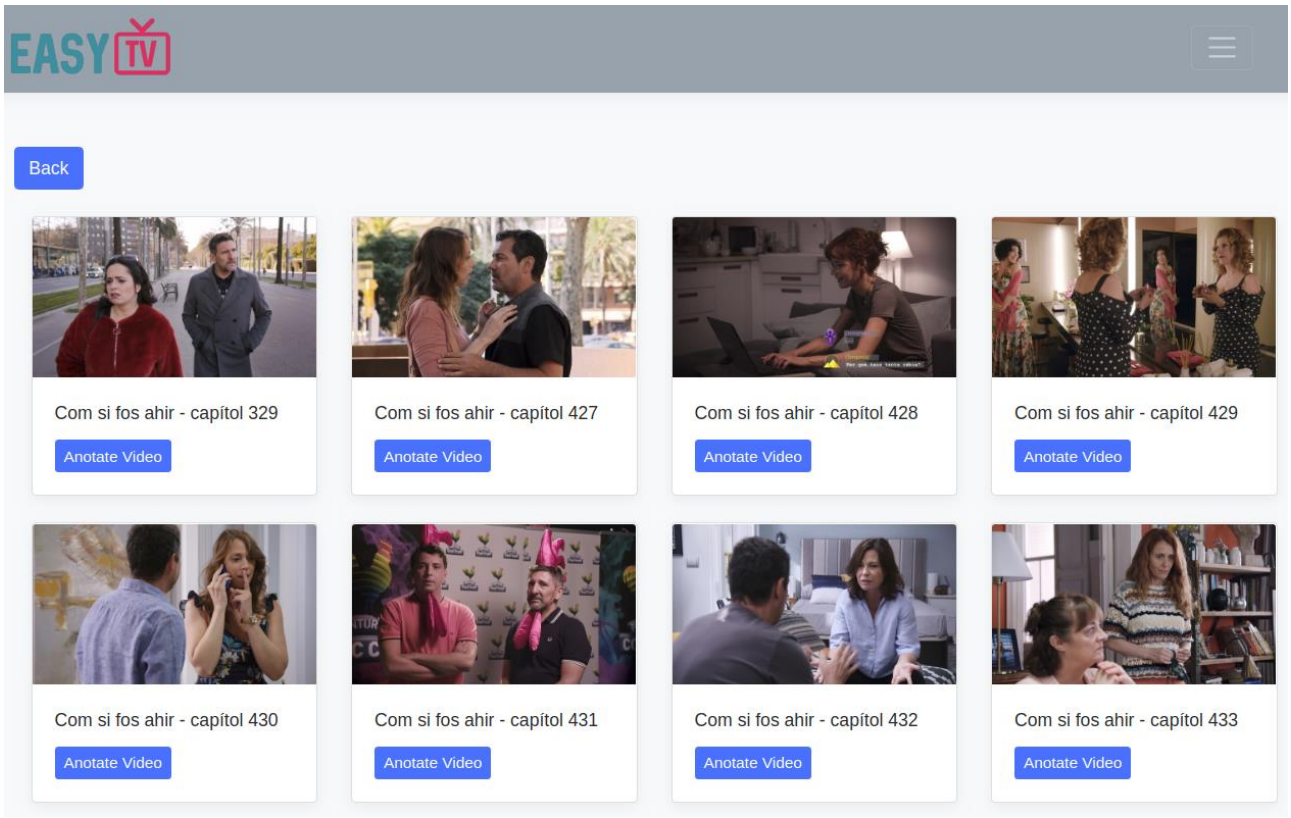


**Figure 16 - Chapters of a TV show inside EasyTV access services platform.**


**Figure 17 - Annotation Tool for character recognition dataset creation.**

Finally, clicking on the “Annotate Video” button the main interface to annotate is accessible. This tool is divided in three main parts. In the top part a time bar is divided in some cut parts where a face has been detected and identified as a unique person due to the tracking algorithm. At the left part, a simple text input is included in order to introduce all the names of the main characters that the user wants annotate during the video. This input is available all the time in order to include new characters when they appear during the video sequence. Finally, at the right part the video sequence for the cut parts is presented and can be played. To annotate a character, you only need move the text label with the name over the video and automatically this character will be annotated. It is possible modify and delete the labels all the time. The dataset will be created when the system detects that there is a right number of data annotated for each of the main characters selected. An example of the interface of the annotation tool is presented in Fig. 17.

# 5. ACCESS SERVICES FOR THE HBBTV ENVIRONMENT BASED ON THE IMAGE ANALYSIS

## 5.1. Modular system architecture

Once the different algorithms for image processing have been described, next step is to define the specific access services that will be provided to the user in order to improve the content accessibility. Next sections are in charge of presenting them.

The modular architecture of the proposed services can be split into two different sides, as shown in Fig. 18, according to a client-server model:

- The server side contains the image analysis algorithms described in section 3 to preprocess the media files and to provide the accessibility information in the form of a JSON file that contains the information retrieved from the image analysis for face detection and character recognition. This file is directly sent to the client application, which is in charge of interpreting it for showing the additional information to final users in an accessible way through a CS app.
- In the client side, the final user app in the CS allows to play different video and other contents such as the developed access tools presented: face detection and character recognition. As in a HbbTV environment, this CS will be synchronized with the TV hybrid terminal.

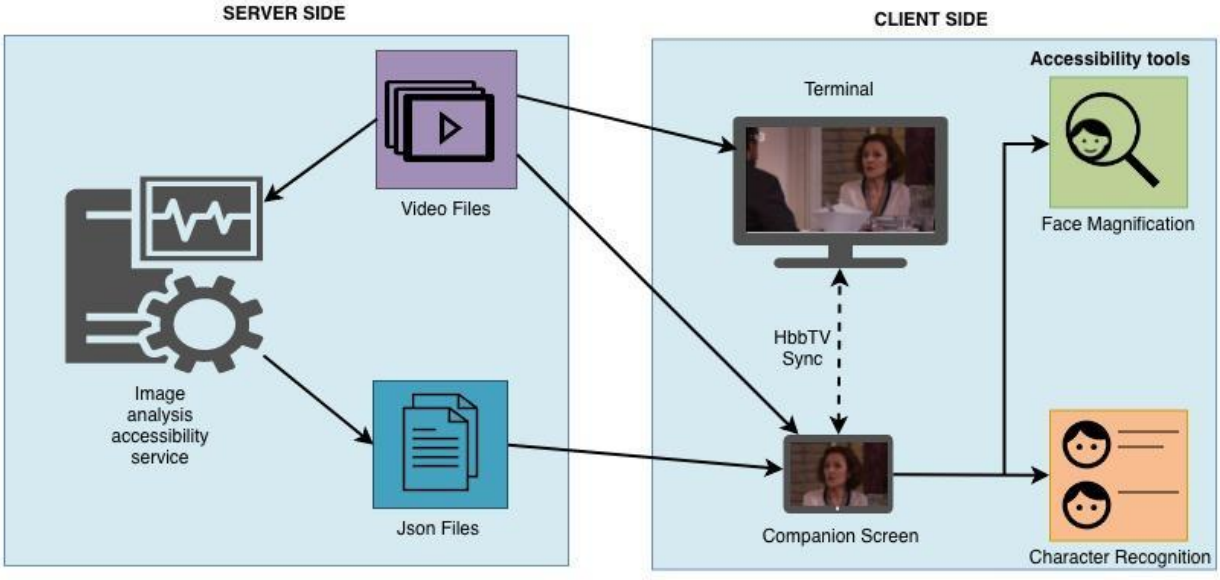


**Figure 18 - Modular system architecture for new access services.**

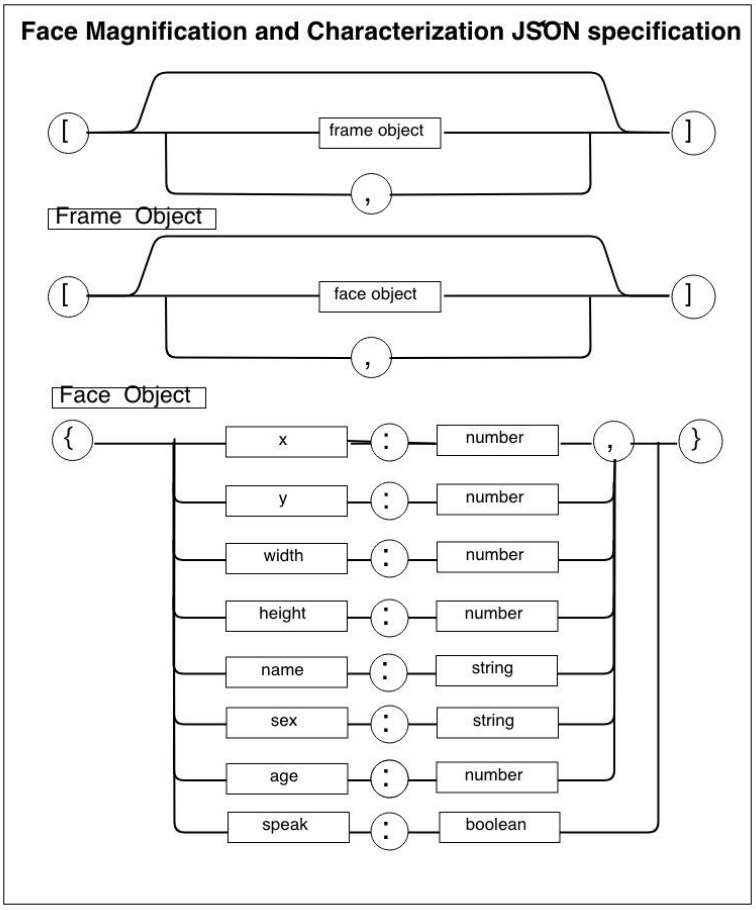


**Figure 19 - Structure of the JSON file generated for face analysis service.**

The structure of the output from the image analysis module is shown in Fig. 19. As can be seen, this JSON is composed by an array of frame objects that contains an array of face objects of each particular frame with the information retrieved and processed of the video, including the information of the character that has been recognized: name, age and sex. It also includes the information about the position of the face ((x,y) coordinates of the center of the face and the weight and height of the area). Finally, it also contains a boolean variable that determines if the person (detected by the face) is speaking or not.

## 5.2. Access services in the CS application

### 5.2.1. Face magnification service

This tool is aimed to allow the end-user to better access specific areas of the content such as faces for improving the intelligibility on the video. This tool can be directly activated from the specific menu in the CS app, as can be seen in Fig. 20. When the option “face magnification” is enabled, the app will automatically zoom in and out the video in real time, using the (x,y) coordinates in the JSON to focus the face and the height and width to adjust the zoom scale. The app will zoom it either the character is speaking or not. If two or more faces are detected, the app will zoom the character that is speaking, while if no one is speaking or more than one is speaking the app will zoom out to show all faces.

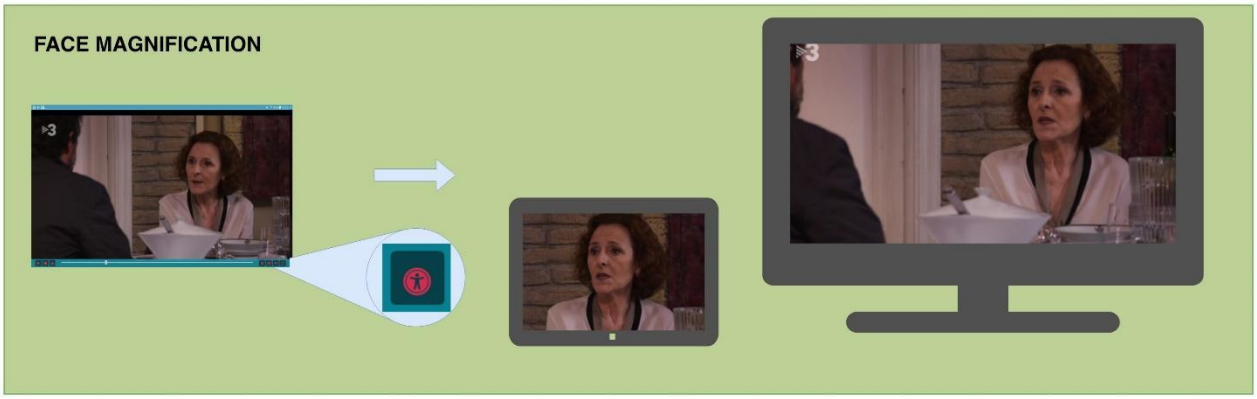


**Figure 20 - Automated face magnification service in the CS app.**

### 5.2.2. Character recognition

This tool is aimed to allow the user to know which characters are on the scene at a specific moment. The function is activated by pressing a button on the app's player controls as indicated in Fig. 21. When the button is pressed the video is set to pause, then a box is shown with the characters' information: picture, age and sex, and the person who is talking. If needed, the age can be presented as a number or as a defined age interval such as children, young, adult and elderly. At the same time a text to speech service is used to read this information out loud. When the locution ends the video is resumed. To retrieve the information, the app gets the data of the paused frame from the JSON file provided by the image analysis module on the server side.

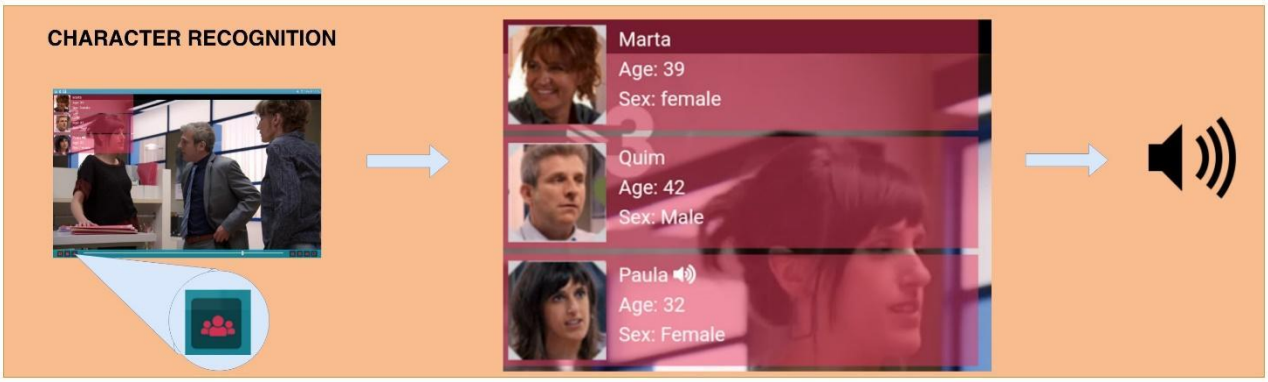


**Figure 21 - Automated character recognition service in the CS app.**

# 6. USER INTERFACE ADAPTATION

EasyTV framework contains the hyper-personalization framework, which is a component that implements the personalization. Part of the personalization process is automatic image adaptation. As declared in the DoW, image adaptation refers to automatic adaptation of settings to improve the accessibility of the content. The parameters available for image enhancement are those listed in in section 2 that is subtitles size, language, color and background color. Accessibility services to improve the audio-visual content are, face detection, image magnification, and character recognition. To be part of the personalization process these parameters have been included in the user profile and a user can initialize their values through the user model editor. The table below indicates these preferences. As part of the user model, the below preferences also take part in the personalization process

**Table 5 Image adaptation related preferences**

| Preference | Data | Value range | Short description |
|---|---|---|---|
| Text detection | Boolean | [ True, false ] | Text detection service |
| Face detection | Boolean | [ True, false ] | Face detection service |
| Character recognition | Boolean | [ True, false ] | Character recognition service |
| Image magnification | Integer | [1.5, 3.0] with step 0.5 | Manual magnification of the image |
| Subtitle language | String | en,ca,it,el,es | Subtitle language |
| Subtitle font size | Integer | [1 – 50] | Subtitles font size |
| Subtitle font color | String | range | Subtitle font color |
| Subtitle background color | String | range | Subtitle background color |

## 6.1. Indicative use case

We present in this section a use case that is indicative of the personalization working related to the preferences described above. A user with visual impairment has the following profile (Table 6). As the user profile indicates, the user prefers a UI font size of 23 and a cursor size of 3.0 (both values corresponds to large values of the corresponding preference). In addition, the user has set his/her subtitle font size to 15 (a small value in comparison with the other values). The user preferences' values suggest that the user may have some visual impairments and he/she would benefit from enabling face and text detection.

**Table 6 User profile example**


```
{
"user_preferences": {
        "default": {
           "preferences": {
                "http://registry.easytv.eu/application/cs/accessibility/faceDetection": false,
                "http://registry.easytv.eu/application/cs/accessibility/textDetection": false,
                "http://registry.easytv.eu/application/cs/accessibility/characterRecognition ": false,
                "http://registry.easytv.eu/application/cs/cc/subtitles/fontColor": "#39dc2",
                "http://registry.easytv.eu/application/cs/cc/subtitles/fontSize": 15,
                "http://registry.easytv.eu/application/cs/cc/subtitles/backgroundColor": "#ee6243",
                "http://registry.easytv.eu/application/cs/cc/subtitles/language": "CA",
                "http://registry.easytv.eu/application/cs/ui/text/size": "23",
                "http://registry.easytv.eu/common/display/screen/enhancement/cursor/Size": 3.0
          }
        }
    },
    "user_context": {
           "http://registry.easytv.eu/context/device": "tablet",
           "http://registry.easytv.eu/context/light": 10,
           "http://registry.easytv.eu/context/proximity": 20,
           "http://registry.easytv.eu/context/location": "es",
           "http://registry.easytv.eu/context/time": "09:47:00"
    },
```

```
    "user_content": {
        "media": "Com_si_fos_ahir",
        "episode": "com_si_fos_ahir_capitol_428"
    }
}
```

Given the user profile, the personalization process would suggest enabling the face and text detection services; however, assuming that only text detection service is available, the outcome would be to suggest text detection and a service that can substituted the face detection. A substitution service for face detection is manual magnification. Figure 22 shows personalization suggestions shown in the CSapp. The suggested value for manual magnification is 3.0, which is in accordance with the user needs. In addition, to increasing the subtitle size value to 30 rather than 15 and to enable text detection service.

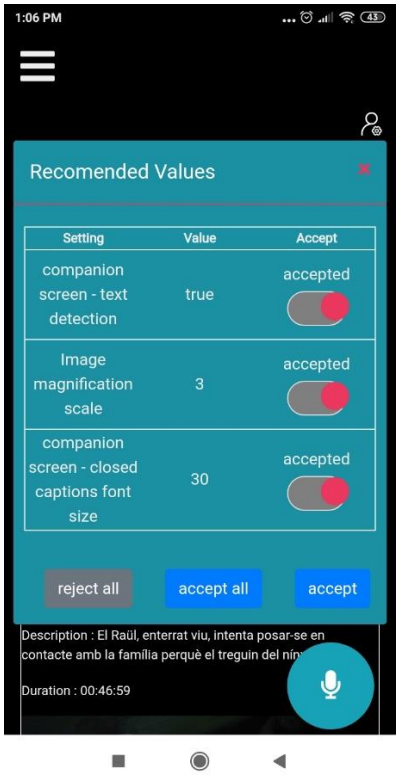


**Figure 22 Suggestions presented to the user through CSapp**

# 7. CONCLUSIONS

Regarding the main purpose of this study, we have presented two innovative access services for audio visual content in HbbTV environment based on the application of different deep learning techniques for visual information extraction.

As it has been explained, our approach is initially composed by two different visual analysis modules, one for object detection and another for tracking, allowing an easier control over the entire process. This has been previously proposed in the tracking-by-detection paradigm (Fiaz, Mahmood, & Jung, 2018; Luo et al., 2014), where the first step, as in our case, is to detect all objects using an object detector in each frame, and then associate them between frames by using different features such as location and appearance. According to this idea, it is clear that the performance of tracking-by-detection models is heavily dependent on the accuracy of the detection model, so the selection of a good detector was vital for our approach. The application of SSD with ResNet Backbone at this step has provided a good tradeoff between time and accuracy, allowing a reduced frame-rate even in real-time scenarios. Related to this, it is important to mention that recent studies (Fiaz, Mahmood, & Jung, 2018; Ning et al., 2017; Broad, Jones, & Lee, 2018 and Feichtenhofer, Pinz, & Zisserman, 2017) are working in the merge of both analyses into one single process, but some problems keep appearing related to complexity and time processing.

At this point, and in order to evaluate our architecture, other solutions from the state of the art are going to be presented. As the environment evolves, new deep learning solutions are gained the attention within tracking-by-detection paradigm, considering architectures that combines the power of CNN object detectors and the ability of RNN to predict future states in time. Some new approaches as ROLO (recurrent YOLO) (Ning et al., 2017), Mf-SSD (Broad, Jones, & Lee, 2018) and D&T (Detect & Track) (Feichtenhofer, Pinz, & Zisserman, 2017) are very interesting ideas to compare with our scheme.

In ROLO's paper authors present an interesting graph that contains how the processing speed is reduced after each algorithm iteration. In this regard, after 10 iterations the computation time is reduced to near 30 ms per frame (30 FPS) but accuracy also decreases over time steps (0.4 – 0.45 mAP).

In the case of Mf-SSD, two main ideas have been considered. On one hand, the inclusion of a RNN (similarly to ROLO) or even replacing it with a Multi-Fusion stage for improving the final mAP accuracy (0.75 – 0.81 mAP). Results show that computational time per frame is around 20 - 70 ms (50 – 14 FPS respectively) and it directly depends on the use of fusion techniques or the RNN. In this case, while the accuracy is reduced, the processing time increases, so it is not an optimal solution for real-time applications.

Finally, D&T is one of the latest architectures that use this approach in order to track objects maintaining a very high accuracy (0.76 – 0.83 mAP), but the main related problem is the computational cost which is between 127 – 141 ms per frame (7 FPS) on NVIDIA Titan X GPU. This architecture is not intended for real time or very long video sequences.

Coming back to our approach, the training process is also vital for obtaining a good performance, so a completely annotated face tracking dataset is needed to learn not only to detect faces but also to track them. The availability of these datasets is not huge in comparison with tracking datasets for objects, but one of the main examples to mention is YouTube Faces (Wolf, Hassner, & Maoz, 2011). In our case, our architecture avoids this problem by training object detector separately by using annotated faces and then training (if needed) the tracking algorithm in common objects, given that the starting bounding boxes are given by the object detector.

It is important to say that tracking algorithms are achieving very high computation speed as shown in Table 3, as well as decreasing the total time consumed even when at the same time the faces are being detected in every frame. Furthermore, our work preserves the accuracy level when there are no occlusions or complex fast behaviors during scenes. As it takes more computational time, the object extraction is done only once every fixed number of frames while the tracking algorithm keeps

making predictions all the time, obtaining a faster solution in our approach than the ones explained above.

On this basis, our approach is able to work in real time during the preprocessing step in order to extract the final information used by the new access services. In comparison, the other discarded algorithms need more computational resources to get a similar behaviors. This resource optimization is very valuable when processing several media contents continuously with the aim of providing the new access service as soon as possible. Moreover, after the retraining of these models with our contents, their accuracy has improved a 2%.

Finally, these new access services give broadcaster and end users with additional valuable functionalities: broadcasters may save annotation resources by the automatization of the task and both visual and hear impaired users may get new access information and tools for a better understanding of the existing multimedia content.

# 8. NEXT STEPS AND FUTURE WORK

As it has been explained, although there is a lack of accessibility in the television environment, new scenarios such as HbbTV provide important capabilities for breaking this barrier, enhancing the user experience. Moreover, the application of deep learning techniques over video content for information extraction allows the definition and deployment of innovative services that may help the access to media content for all.

According to these ideas, we proposed two main services based on image analysis for helping the access to content for blind and visually impaired people, according to some of the suggestions given by real users (information obtained from the EasyTV focus groups, as explained in section 5.2). In this regard, It is important to note that the deep learning techniques that have been selected already exist in the state of the art, but their combination for obtaining these innovative access services is a complete original work.

The implementation of these services is based on a 3-step image analysis workflow, which provides good results in terms of accuracy and computational cost as explained in the discussion section. Firstly, a face detection phase based on an SSD object detector is applied, followed by a face tracking step based on KCF tracker. Finally, the information extraction is done thanks to different algorithms according to the pursued objective: character recognition, face characterization and face speaking detection. The output from this analysis module is finally used for access service composition, as explained in section 4, providing a new way for automated content contextualization by means of additional information.

According to the obtained results we can assure that deep learning techniques can be used to generate accessible content automatically. On this basis, Hybrid TV, and in particular HbbTV, together with a thorough definition of innovative services where the accessible content can provide competitive strengths will help the enhancement of the access to media content for all.

Nevertheless, some future work lines can be defined. First of all, there is margin to improve in the detection of which faces are speaking on a scene. At this moment, this decision is done by calculating a median and using a threshold, but the use of deep learning algorithms over a specific dataset with mouth movements could improve the results. Even more, this dataset could be also tagged to categorize the mouth movements as speaking, singing, screaming, eating, etc. in order to apply deep learning algorithms for additional behaviors detection. On the other hand, an additional future line is related to the combination of the image analysis for character recognition with the audio processing of the voices in the content to get more accurate results. Finally, the evaluation with final users are planned in order to obtain their feedback regarding the proposed services and allow their improvement.

# 9. REFERENCES


Voulodimos, Athanasios et al. “Deep Learning for Computer Vision: A Brief Review.” Comp. Int. and Neurosc. (2018).

Trigueros, Daniel S'aez et al. “Face Recognition: From Traditional to Deep Learning Methods.” CoRR abs/1811.00116 (2018).

Wang, Mei and Weihong Deng. “Deep Face Recognition: A Survey.” CoRRabs/1804.06655 (2018).

Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., & Rothe, R. (2017). Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on* (pp. 87–94).

Ahmed, A. H., Kpalma, K., & Guedi, A. O. (2017). Human Detection Using HOG-SVM, Mixture of Gaussian and Background Contours Subtraction. In *2017 13th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)* (pp. 334–338). https://doi.org/10.1109/SITIS.2017.62

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2930–2940.

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. S. (2016). Fully-Convolutional Siamese Networks for Object Tracking. *CoRR*, *abs/1606.0*. Retrieved from http://arxiv.org/abs/1606.09549

Boronat, F., Marfil, D., Montagud, M., Pastor, J. (2018). HbbTV-Compliant Platform for Hybrid Media Delivery and Synchronization on Single and Multi-Device Scenarios. *IEEE Transactions on Broadcasting*, 64(3), 721-746.

Broad, A., Jones, M., & Lee, T.-Y. (2018). Recurrent Multi-frame Single Shot Detector for Video Object Detection.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2016). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR*, *abs/1611.0*. Retrieved from http://arxiv.org/abs/1611.08050

Castellano, B. (2018). Pyscenedetect. Retrieved from https://pyscenedetect.readthedocs.io

Chakraborty, S., & Das, D. (2014). An Overview of Face Liveness Detection. *CoRR*, *abs/1405.2*. Retrieved from http://arxiv.org/abs/1405.2227

Claudy, L. (2012). The broadcast empire strikes back. *IEEE Spectrum*, *49*(12), 52–58. https://doi.org/10.1109/MSPEC.2012.6361764

CNMC. (2017). Informe sobre el seguimiento de las obligaciones impuestas en materia de accesibilidad correspondiente al año 2016. Retrieved from https://www.cnmc.es/sites/default/files/1855187_9.pdf

CSA. (2017). L’accessibilité des programmes de télévision aux personnes handicapées et la représentation du hándicap à l’antenne.

Cuimei, L., Zhiliang, Q., Nan, J., & Jianhua, W. (2017). Human face detection algorithm via Haar cascade classifier combined with three additional classifiers. In *2017 13th IEEE International Conference on Electronic Measurement Instruments (ICEMI)* (pp. 483–487). https://doi.org/10.1109/ICEMI.2017.8265863

Danelljan, M., Häger, G., Khan, F. S., & Felsberg, M. (2014). Accurate Scale Estimation for Robust Visual Tracking. In *BMVC*.

Domínguez, A., Agirre, M., Flörez, J., Lafuente, A., Tamayo, I., & Zorrilla, M. (2018). Deployment of a Hybrid Broadcast-Internet Multi-Device Service for a Live TV Programme. *IEEE Transactions*

*on Broadcasting*, *64*(1), 153–163. https://doi.org/10.1109/TBC.2017.2755403

EasyTV Project. (n.d.). EasyTV project website. Retrieved from https://easytvproject.eu/

eMarketer. (2017). US Simultaneous Media Users: eMarketer's Estimates for 2017. Retrieved from https://www.emarketer.com/Report/US-Simultaneous-Media-Users-eMarketers-Estimates-2017/2002163

ETSI. (2016). Hybrid Broadcast Broadband TV ETSI Standard TS 102 796 2016. Retrieved from https://www.etsi.org/deliver/etsi_ts/102700_102799/102796/01.04.01_60/ts_102796v010401p.pdf

European Commission. (2010). European Disability Strategy 2010-2020: A Renewed Commitment to a Barrier-Free Europe. Retrieved from https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0636:FIN:en :PDF

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2017). Detect to track and track to detect. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3038–3046).

Fiaz, M., Mahmood, A., & Jung, S. K. (2018). Tracking Noisy Targets: A Review of Recent Object Tracking Approaches. *ArXiv Preprint ArXiv:1802.03098*.

Gordon, D., Farhadi, A., & Fox, D. (2017). Re3 : Real-Time Recurrent Regression Networks for Object Tracking. *CoRR*, *abs/1705.0*. Retrieved from http://arxiv.org/abs/1705.06368

Güler, R. A., Neverova, N., & Kokkinos, I. (2018). DensePose: Dense Human Pose Estimation In The Wild. *CoRR*, *abs/1802.0*. Retrieved from http://arxiv.org/abs/1802.00434

Hassaballah, M., Abdelmgeid, A. A., & Alshazly, H. A. (2016). Image features detection, description and matching. In *Image Feature Detectors and Descriptors* (pp. 11–45). Springer.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. (2017). Mask {R-CNN}. *CoRR*, *abs/1703.0*. Retrieved from http://arxiv.org/abs/1703.06870

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Held, D., Thrun, S., & Savarese, S. (2016). Learning to Track at 100 {FPS} with Deep Regression Networks. *CoRR*, *abs/1604.0*. Retrieved from http://arxiv.org/abs/1604.01802

Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2014). High-Speed Tracking with Kernelized Correlation Filters. *CoRR*, *abs/1404.7*. Retrieved from http://arxiv.org/abs/1404.7584

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv Preprint ArXiv:1704.04861*.

Immersive Accessibility Project. (n.d.). Immersive Accessibility project website. Retrieved from http://www.imac-project.eu/

Jain, V., & Learned-Miller, E. (2010). *Fddb: A benchmark for face detection in unconstrained settings*.

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. In *European conference on computer vision* (pp. 679–692).

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *CoRR*, *abs/1708.0*. Retrieved from http://arxiv.org/abs/1708.02002

Liu, A., Du, Y., Wang, T., Li, J., Li, E. Q., Zhang, Y., & Zhao, Y. (2011). Fast facial landmark detection using cascade classifiers and a simple 3D model. In *Image Processing (ICIP), 2011 18th IEEE International Conference on* (pp. 845–848).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C.-Y., & Berg, A. C. (2015). {SSD:} Single Shot MultiBox Detector. *CoRR*, *abs/1512.0*. Retrieved from

http://arxiv.org/abs/1512.02325

Lukezic, A., Vojir, T., Cehovin, L., Matas, J., & Kristan, M. (2016). Discriminative Correlation Filter with Channel and Spatial Reliability. *CoRR*, *abs/1611.0*. Retrieved from http://arxiv.org/abs/1611.08461

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Zhao, X., & Kim, T.-K. (2014). Multiple object tracking: A literature review. *ArXiv Preprint ArXiv:1409.7618*.

Malhotra, R. (2013). Hybrid Broadcast Broadband TV: The Way Forward for Connected TVs. *IEEE Consumer Electronics Magazine*, *2*(3), 10–16. https://doi.org/10.1109/MCE.2013.2251760

Matamala, A., Orero, P., Rovira-Esteva, S., Casas Tost, H., Morales Morante, F., Soler Vilageliu, O., … Tor-Carroggio, I. (2018). User-centric approaches in access services evaluation: profiling the end user. In *Proceedings of the Eleventh International Conference on Language Resources Evaluation (LREC 2018)* (pp. 1–7).

McNally, J., & Harrington, B. (2017). How Millennials and Teens Consume Mobile Video. In *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video* (pp. 31–39). New York, NY, USA: ACM. https://doi.org/10.1145/3077548.3077555

Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. (1999). XM2VTSDB: The extended M2VTS database. In *Second international conference on audio and video-based biometric person authentication* (Vol. 964, pp. 965–966).

Monzo, D., Albiol, A., Albiol, A., & Mossi, J. M. (2010). A comparative study of facial landmark localization methods for face recognition using hog descriptors. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (pp. 1330–1333).

NIELSEN a. (2017). The Nielsen Comparable Metrics Report, Q1-2016. Retrieved from https://www.nielsen.com/us/en/insights/reports/2016/the-comparable-metrics-report-q1-2016.html

NIELSEN b. (2017). *The Nielsen Comparable Metrics Report, Q2-2016*. Retrieved from https://www.nielsen.com/us/en/insights/reports/2016/the-comparable-metrics-report-q2-2016.html

NIELSEN c. (2017). The Nielsen Comparable Metrics Report, Q3-2016. Retrieved from https://www.nielsen.com/us/en/insights/reports/2017/the-comparable-metrics-report-q3-2016.html

NIELSEN d. (2017). The Nielsen Comparable Metrics Report , Q4-2016. Retrieved from https://www.nielsen.com/us/en/insights/reports/2017/the-comparable-metrics-report-q4-2016.html

NIELSEN e. (2018). The Nielsen Comparable Metrics Report, Q1-2017. Retrieved from https://www.nielsen.com/us/en/insights/reports/2017/the-nielsen-comparable-metrics-report-q1-2017.html

NIELSEN f. (2018). The Nielsen Comparable Metrics Report, Q2-2017. Retrieved from https://www.nielsen.com/us/en/insights/reports/2017/the-nielsen-comparable-metrics-report-q2-2017.html

Ning, G., Zhang, Z., Huang, C., Ren, X., Wang, H., Cai, C., & He, Z. (2017). Spatially supervised recurrent convolutional neural networks for visual object tracking. In *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on* (pp. 1–4).

Orero, P., Martín, C. A., & Zorrilla, M. (2015). HBB4ALL: Deployment of HbbTV services for all. In *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting* (pp. 1–4). https://doi.org/10.1109/BMSB.2015.7177252

Organization, W. H., & others. (2013). Universal eye health: a global action plan 2014-2019.

Padilla, R., Filho, C., & Costa, M. (2012). Evaluation of Haar Cascade Classifiers Designed for Face Detection. In *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering* (Vol. 6).

Prosperity4All Project. (n.d.). Prosperity 4All project website. Retrieved from http://www.prosperity4all.eu/

Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. *CoRR*, *abs/1506.0*. Retrieved from http://arxiv.org/abs/1506.02640

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster {R-CNN:} Towards Real-Time Object Detection with Region Proposal Networks. *CoRR*, *abs/1506.0*. Retrieved from http://arxiv.org/abs/1506.01497

Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, *126*(2–4), 144–157.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, *47*, 3–18.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Sodagar, I. (2011). The MPEG-DASH Standard for Multimedia Streaming Over the Internet. *IEEE MultiMedia*, *18*(4), 62–67. https://doi.org/10.1109/MMUL.2011.71

Statista. (2017). Smart TV shipments worldwide. Retrieved from https://www.statista.com/statistics/461561/smart-tv-shipments-worldwide-by-region/

Vinayagamoorthy, V., Allen, P., Hammond, M., & Evans, M. (2012). Researching the User Experience for Connected Tv: A Case Study. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (pp. 589–604). New York, NY, USA: ACM. https://doi.org/10.1145/2212776.2212832

Wang, M., & Deng, W. (2018). Deep Face Recognition: A Survey. *ArXiv Preprint ArXiv:1804.06655*.

Wolf, L., Hassner, T., & Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 529–534).

Woods, R. L., & Satgunam, P. (2011). Television, computer and portable display device use by people with central vision impairment. *Ophthalmic and Physiological Optics*.

Xu, Y., Xu, L., Li, D., & Wu, Y. (2011). Pedestrian detection using background subtraction assisted Support Vector Machine. In *2011 11th International Conference on Intelligent Systems Design and Applications* (pp. 837–842). https://doi.org/10.1109/ISDA.2011.6121761

Yuheng, S., & Hao, Y. (2017). Image Segmentation Algorithms Overview. *CoRR*, *abs/1707.0*. Retrieved from http://arxiv.org/abs/1707.02051

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. *ArXiv Preprint ArXiv:1605.07146*.

Zhang Zhifei, S. Y., & Qi, H. (2017). Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 2879–2886).

Ziegler, C. (2013). Second screen for HbbTV — Automatic application launch and app-to-app communication enabling novel TV programme related second-screen scenarios. In *2013 IEEE Third International Conference on Consumer Electronics ¿ Berlin (ICCE-Berlin)* (pp. 1–5). https://doi.org/10.1109/ICCE-Berlin.2013.6697990