# New accessibility services in HbbTV based on a deep learning approach for media content analysis

**Authors:**

Silvia Uribe, Alberto Belmonte, Juan Pedro López, Francisco Moreno, Álvaro Llorente, Federico Álvarez

Grupo de Aplicación de Telecomunicaciones Visuales, ETSIT, Universidad Politécnica de Madrid

{sum, abh, jlv, fmg, alg, fag}@gatv.ssr.upm.es

**Corresponding author:**

Silvia Uribe

Avenida Complutense 30, 28040, Madrid, ETSIT, Edif. D, Despacho 103.

+34 910672632

sum@gatv.ssr.upm.es



**Short title:**

Deep learning in video for accessibility

**Manuscript content:**

Number of pages: 25

Number of tables: 4

Number of figures: 12

# New accessibility services in HbbTV based on a deep learning approach for media content analysis

**Abstract:**

Universal access on equal terms to audio visual content is a key point for the full inclusion of people with disabilities in activities of daily life. As a real challenge for the current Information Society, it has been detected but not achieved in an efficient way, due to the fact that current access solutions are mainly based in the traditional television standard and other not automated high cost solutions. The arrival of new technologies within the hybrid television environment together with the application of different artificial intelligence techniques over the content will assure the deployment of innovative solutions for enhancing the user experience for all. In this paper a set of different tools for image enhancement based on the combination between deep learning and computer vision algorithms will be presented. These tools will provide automatic descriptive information of the media content based on two main applications: face detection for magnification and character identification and text detection for embedded information extraction. The fusion of this information will be finally used to provide a customizable description of the visual information with the aim of improving the accessibility level of the content, allowing an efficient and reduced cost solution for all.

**Keywords:** media accessibility, deep learning, computer vision, face detection, text detection,

# 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 dexterity. The cost of providing these kind of solution, 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 reason 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 delivery content 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 broadcasted environment for content delivery is alive 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 hyperpersonalization and the user interaction. In this context, with the emergence of different technologies such as IPTV, OTT-TV or HbbTV (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 synchronous way in order to supplement it and even to replace it in some cases.

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 un terms of presentation, due to the use of the CS as a new mean for the submission of additional accessibility services.

In another hand, the "what" aspect of our solution will be focused on the content processing for obtaining low cost efficient accessibility services that can beneficiate from the application of new algorithms and processing techniques such as deep learning. 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 paper we present a complete method for detecting two different elements inside the video:

- Firstly, the automated face detection which will allow developing different solutions for

enhancing the access to the content as magnification for helping in the lips reading process for deaf people and character identification for providing additional information to supplement current solutions (improving the subtitles and the audio description)
- Secondly, the automated text detection for textual information embedded in the audio visual content. This information is not fully accessible, so people with visual disabilities will miss it unless they have any tool for reading it.

Considering the aforementioned concepts, the main objective of this paper is to present a new set of innovative accessibility services for helping the total 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 improving the 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.

To describe the above-mentioned contributions, this paper 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 and text 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. Finally, Section 5 is focused on presenting the main limitation of our solution, the related future work and the conclusions.



# 2. RELATED WORK AND PROBLEM FORMULATION

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 scenario, 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.

## 2.1. Hybrid television and content accessibility 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 (millennials and teens (McNally & Harrington, 2017)).

**FIG. 1**

In this regard, there are two main facts that gives 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' need and expectation. 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 HbbTV standards.

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. In this way, it is possible to create a different user experience with synchronized content in all devices and offer personalised 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, he can also access to synchronized additional services such as alternatives audios, multiview content, personalised advertisement, statistical data information or accessibility 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 as much as the rest of the population, even considering it as an important family time activity. Furthermore, 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 accessibility services for assuring an optimal media content access on an equal basis for all.



At European Union level, this issue has been recognised as an essential user right in the audiovisual sector, so the setting of different actions for its implementation has been defined has one of the priorities of the European Disability Strategy 2010-2020 (European Commission, 2010). In this context, 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.

Furthermore, UE has supported the deployment of different accessibility solutions for this scenario by funding several research and innovation projects such as HBB4ALL (Orero, Martín, & Zorrilla, 2015), Prosperity4All (Prosperity4All Project), ImAc (Immersive Accessibility Project) and EasyTV (EasyTV Project), where the work done for this paper is framed.

### 2.1.1. **EasyTV project: an innovative approach for multimedia content accessibility**

EasyTV is a 30-month length project aiming to offer easier access to converging media content to persons with disabilities by offering new mechanisms for content delivery, interaction and presentation based on four main pillars:

- Universal multimedia access: EasyTV will ensure improved access to multimedia content through the provision of different services that will enrich the visual and sound experience for all. To do so, it will be focused on providing new graphical interfaces adapted to users' needs, on performing an innovative audio processing for intelligibility improvement and finally on defining new tools for enhancing the image of the content, such as the services presented in

this paper. EasyTV will also deploy efficient tools for universal interaction by means of multimedia devices.

- Novel technologies for breaking the sign language barriers: the project will be focused on the creation of a multilingual crowdsourcing platform for the sharing and generation of accessible content such as subtitles, the definition of an ontology for the link and translation of different sign language concepts and, finally, the advance in the capture and generation of sign language content by means of hyper realistic avatars.
- Personalization: EasyTV will delve into an improved customization of the user experience thanks to the generation of adaptive menus and graphical interfaces, the provision of service recommendation tools and content based on user models and finally to the access adaptation through MPEG-DASH (Sodagar, 2011) service provision.
- User centric approach: EasyTV will propose a methodology based on a continuous user consultation for a clear idea of the existing needs (for user requirements gathering) and for a reliable opinion about the deployed services (for user validation). This has been an essential input for the definition of the services presented in this paper, as it is going to be explained later.

As can be seen, EasyTV is an ambitious and widespread project where new tools are going to be developed for improving the access to the contents of people with disabilities according to their own indications. Regarding the user requirements gathering, the performance of focus groups with final users is vital for our approach, as help us to understand which are the main demands in the real scenario. In this context, next section is focused on presenting some of them.

#### 2.1.2. **Lack of media content accessibility in TV consumption**

Although the explanation about the focus group that have been done within the first step of the project is out of this paper scope (including the methodology, the ethic requirements and the participants) (Matamala et al., 2018), their results are vital for understanding the definition of our accessibility services. 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**

The services proposed in this paper 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, the face detection tool will be firstly 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 automated tools that will improve the audio description information by identifying the people in the scene and their specific characteristics in terms of gender and age. On the other hand, the text detection and identification is directly related to one of the particular problems indicated by the users: how to access the overlay text that is not accessible in order to obtain additional information that may help to enrich the context and so the user experience.

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.

## 2.2. Need for a higher level of accessibility: image analysis based on deep learning for improving the content access

According to the “European Disability Strategy” Report (European Commission, 2010), the number of people with disabilities in the European Union (EU) will reach 120 million by 2020 and World Health Organization (WHO) (Organization & others, 2013) estimated that 39 million people were blind in 2010 from a total of 285 million people visually impaired. On the other hand, the expected

volume of IP traffic for that date is expected to reach one million minutes of multimedia content streamed per second (Cisco, 2017). his amount of transmitted contents stem from different sources, containing text that is not accessible for all, especially for people with low vision. 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.

With this objective in mind, and taking into account that our services are based in video analysis for face and text detection, it is important to say that nowadays the application of different deep learning methods are actually 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 crystal clear.

The application of computer vision tools for analyzing different human behavior 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 the use of background subtraction (for movement detection in sequential frames) in combination with support vector machines (to classify the detected movement classification) (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)…



All above examples have been achieved with the use of computer vision in combination with machine learning techniques, providing interesting results in a big variety 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 (Ren, He, Girshick, & Sun, 2015), SSD (W. Liu et al., 2015), YOLO (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 (Henriques, Caseiro, Martins, & Batista, 2014), MOSSE (Danelljan, Häger, Khan, & Felsberg, 2014), CSRT (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 FC networks (Bertinetto, Valmadre, Henriques, Vedaldi, & Torr, 2016), GOTURN (Held, Thrun, & Savarese, 2016), Re3 (Gordon, Farhadi, & Fox, 2017)**¡Error! No se encuentra el origen de la referencia.**).

Finally, others 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, SURF, ORB) in combination with machine learning techniques (for optimization, regression) have been applied. For image segmentation can be used algorithms like k-means clustering, mean-shift clustering or interactive image segmentation. Furthermore, deep learning methods are solving these complex tasks thanks to the availability of huge 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.

Regarding the second aspect of our approach, Optical Character Recognition (OCR) for text and image recognition in scanned documents is considered by many researchers as a solved problem (Weinman, Learned-Miller, & Hanson, 2009) but it is still a challenge in video due to the unpredicted conditions of resolutions, encoding and impairments introduced. For that reason, it is necessary to apply computer vision and pattern recognition techniques for improving the work of the OCR and exploiting temporal redundancy and tracking the textual objects for detecting the timing of the text in the scene. To avoid false positives in the result it is necessary to develop a process of text verification. The state-of-the-art in this field (Ye & Doermann, 2015) presents several techniques including fuzzy systems for clustering algorithms (FCM) and transferred deep CNN classifiers (Lu, Sun, Chu, Huang, & Yu, 2018).

### 2.2.1. **Face detection problem**

Face detection is a well-known problem in computer vision. In this regard, Several methods have been already proposed in order to extract pattern 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 images (sliding window, region proposal), second, the obtaining of extracted features from this parts (Haar features, HoG 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 (ANNs).

Traditional methods rely on how we can model the features or the pattern manually trying to find edges, blobs or other interesting patterns with the aim of defining different features and classify them. However, recent studies development 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 Single Shot Detector (SSD) which combine the task of selecting Regions of Interests (RoIs) in a single end to end architecture and You Only Look Once (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 backbones pretrained on ImageNet dataset selecting the number of layers to use in each one. 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 use to ResNet10 or VGG16 backbones. Other 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 these purposes.

Párrafo adicional

#### 2.2.2. **Text detection problem**

As derived from the focus groups, people with low vision present difficulties in the access to text overexposed in video sequences, such as headliners and other types of graphics text. Optical Character Recognition (OCR) techniques are very common in text detection and recognition in static image, but in video is a challenging task because it requires the preparation of the frame to ensure enough contrast of the characters from the background before the process of detection and recognition, but can also benefit from the temporal redundancy of information.

Different techniques are used for improving the labour of the OCR's result adapting the video contents in the detection and recognition of artificial text and natural text in scenes. The development of tools for detecting text in video is a challenging task due to the video encoding process, the variety of resolutions, the complex backgrounds and the diversity of fonts. Anthimopoulos et al (Anthimopoulos, Gatos, & Pratikakis, 2010) propose a two-stage system for text detection, where text line are detected in the first phase and the use of features related to a SVM (Support Vector Machine) classifier constitute the second phase. Other proposals are focused on the unpredictable blur and distortion affecting to the video using video quality assessment tools for identifying the impaired image using devoconvolutional models (Khare, Shivakumara, Raveendran, & Blumenstein, 2016) (D. Chakraborty, Roy, Saini, Alvarez, & Pal, 2018); fractal properties in the gradient domain (Shivakumara et al., 2017) or a bayessian classifier through image binarization (Roy et al., 2015). Applying preprocessing to the natural image for improving the result employs different techniques include evaluation of MSER (Maximally Stable Extremal Regions) and intersection with Canny filter detected edges (Islam, Mondal, Azam, & Islam, 2016).

For that reason, we propose the inclusion of deep-learning techniques for creating a pseudo-subtitling file including the recognized text with its corresponding timing that can be easily transformed into audio as an innovative accessibility tool. A dataset of video sequences was specifically created for this purpose including variation of characteristics of the text and background included in the scene, such as contrast, color, font type, video encoding and text in motion with techniques such as horizontal and vertical scrolls. Results reveal the efficiency of this kind of techniques in text recognition for the improvement of accessibility tools.



# 3. METHODOLOGY

## 3.1. Image analysis for face detection

Based on all the available frameworks, algorithms and the increase of the computational capacity due to the use of GPUs, deep learning approaches have been used in this paper in order to perform some simple and complex tasks. Additionally, a comparison with computer vision algorithms have been presented in order to clarify the selection of different approaches in our main architecture.

**FIG. 2**

Our proposed architecture takes the advantage of these concepts and algorithms and learn to detect faces in images as the main important part of the complete architecture. If no faces are detected no postprocessing can be done to extract related information from the video under analysis. For this purpose, to train the algorithms a well-known dataset (FDDB (Jain & Learned-Miller, 2010)) for face detection was used.

Two object detectors have been selected looking always the best accuracy with the less computational consuming time. SSD and YOLO have been used as the quicker algorithms to perform object detection. SSD have been trained using VGG16 and ResNet10 backbones in order to decrease the complexity in the network achieving a good tradeoff between time and accuracy. From YOLO have been selected version 2 due to its better performance in comparison with version 1 and

a simpler architecture in contrast as the presented by version 3. The results after training are presented in Table 2 where Frame Per Second (FPS) rate and Mean Average Precision (mAP) values are collected:

**TABLE 2**

**FIG. 3.**

### 3.1.1. **Scene Detection**

Next step in the proposed architecture is sampling the video looking where the scene change appears. This is an important task in order to characterize each scene with the appropriate information along the time and give information about when tracking algorithm should stop to start with new faces in the next scene. Our approach uses an existing framework known as PySceneDetec (Castellano, 2018) that is able to detect scene changes in videos and automatically splitting the video into separate clips. The tool offers several detection methods from simple thresholding to advanced content aware fast-cut detection. The scene changes in our videos are previously annotated by this tool and are provided to our architecture.

### 3.1.2. **Tracking Faces**

One of the main purposes with the proposed architecture is provide information about where the faces are located to perform magnification along the main important face if is speaking. If the algorithm is based only on the face detection some important problems appear. The first problem is the bounding box size estimated by the face detector. These bounding boxes are not fixed along time then if the center is selected as the point where perform magnification on the face the results along time will be noisy. The second difficulty is the loss of detection along time. Face detector is able to detect faces in single images but along the time could appear complex situations where the face of the person is not pointing directly to the camera failing in the detection task. Perform tracking over the detected faces solves the problem maintaining where the face is located if no detection is achieved. The last problem to comment is related to the computational time due to object detectors are slower than tracking algorithms. If the frame rate of the face detector is about 39 FPS and the video is recorded at 25 FPS real time is achieved but after face detection more processes are performed increasing time. Tracking algorithms are able to work up to 100 FPS increasing the speed in the complete architecture and video processing time is highly reduced.



Proposed architecture incorporates tracking algorithms to speed up the process and perform face detection only every fixed number of frames or after a scene change. After every face detection previous trackers are associated with the new detections to avoid the creation of new trackers or trackers are deleted if a scene change happens.

Four trackers have been tested in order to select the best performance during tracking with the less computational time. By one hand, the first two trackers are considered inside the online trackers group. This type of trackers works directly over the frame sequence estimating the next state using information collected along the time. Kernelized Correlation Filters (KCF) contains merits of adaptive threshold approach, kernelized correlation filter method, and Kalman filter algorithm to make tracking faster and more accurate at the same time. CSRT tracker take the advantage of discriminative correlation filters and provide a novel learning algorithm for its efficient and seamless integration in the filter update and the tracking process.

By the other hand, two other offline trackers have been tested to compare both types. 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.

In our experiments the pretrained models for GOTURN and Re3 have been used due to have been trained using annotated motion datasets that contains faces between all the possible objects and the lack of other datasets created specially to track faces. KCF and CSRT are online trackers and no need additional data. Several sequences have been annotated in our own videos to test the accuracy in the trajectories obtained with all the algorithms and regarding the possible most common mistakes that can happens during tracking. Table 3 presents some metrics to evaluate the performance of these algorithms.

**TABLE 3**

Mean pixel error have been obtained using the Euclidean distance between the annotated bounding box center and the estimated bounding box center by the tracking algorithms. The best result is obtained with CSRT algorithm due to its very efficient with slow motion objects and works well in presence of occlusions but regarding the processing time is very slow in comparison with others algorithms losing the real time processing capabilities.

Deep learning trackers present worst mean pixel error along the experiments performed. GOTURN tends to cover a big area around the face after some frames losing accuracy and smoothing tracking results. Another related problem is the fact that this big area can concentrate the focus in other objects outside the face and estimate a new bounding box with center far from the annotated bounding box center. Re3 works well and gives high accuracy in the trajectories with the best processing time if only one face is detected in the image. The problem appears in cases where more than one person is in the scene increasing the time lineally with the number of faces detected. The performance of these two algorithms can be improve it training both with a concrete dataset of only faces.

In contrast KCF presents high accuracy and the processing time increase slowly with the number of trackers. The main problem with this algorithm is the high probability to fail in presence of important occlusions that hide a big area of the tracked face.



Fig. 4 shows some examples of failure with each algorithm. Left image represents a situation with slow motion in the image and the face near the same position along the scene. All algorithms work well but can be observed that GOTURN bounding box is bigger than the others. Center image present one frame after a fast movement of the face along some frames. KCF and Re3 trackers have been displaced and now are not centered in the face. CSRT and GOTURN maintain the bounding box centered at the face. Right image presents the situation before and after an occlusion. Re3 tracker bounding box has suffered a little displacement but KCF has totally lost the face.

**FIG. 4**

Finally, KCF have been the algorithm selected to take part in the whole architecture due to the accuracy obtained, processing time and regarding that fast movements and occlusions happens few times in the video when the person is speaking.

**FIG. 5**

An example of the results obtained by the use of the tracking algorithm can be shown in Fig. 5. Estimated bounding box center coordinates (x, y) are presented using face detector only along all frames and applying tracking algorithm over the face detected at the initial frame after a scene change. Using face detector only some detections are lost and the magnification will finish at this time to be applied again after few frames giving a bad sensation to the observer. Regarding the tracking algorithm estimation can be observed two important results. Even if the face is not detected by the face detector, tracking algorithm is able to continue detecting the face in the image. The second important result is the smooth curve in comparisons with the curve presented by the face detector. Tracking algorithm is able to filter better the trajectory avoiding very noisy transitions.

### 3.1.3. **Face Recognition and Characterization**

In order to provide more information about what is happen in a scene the proposed architecture include a module that is able to perform face recognition over the face of the main characters related to the concrete content displayed at the moment. It is possible then provide more information about the scene, not only where the face is located, also who is in the scene.

This module takes as input the cropped faces from the previous estimations after tracking algorithm and use a convolutional neural network trained for classification tasks over the faces. The labels in this network are the names of the main characters and one more label identified as “unknown” due to in a scene not always are only the main characters.

If one of these principal characters is recognized the presentation module will display all the necessary information about it. 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 of these persons to collect a general complete knowledge at each scene.

To train first network in face recognition tasks our own dataset was collected with faces of the same person in different situations and were annotated with the name of the character as labels. The number of images for training is more than 60000 images separated in 17 classes (approximately 3500 images per class) and 15000 images for testing. For the unknown class the images collected are from any face that is not a main character. Fig. 6 present some of the main characters and different images in our dataset of two of them. For train the second branch of the network some existing datasets were collected and mixed to have more training data in order to improve the results. 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 IMDb-Faces (Rothe, Timofte, & Van Gool, 2018) (more than 460000 images of famous persons annotated with age and gender).



**FIG. 6**

The network architecture is divided in two brands where the second branch is activated only when a detected face is not a main character Fig. 7. To compose Face Recognition Network different backbones have been tested with a classification network that contains two neurons layer of 1024 neurons each one and a final SoftMax activation to get the final prediction. The input has been resized to a fixed size of 224x224.

Second network used for Face Characterization use as backbone the WideResNet network (Zagoruyko & Komodakis, 2016)**¡Error! No se encuentra el origen de la referencia.**, a modification of ResNet architectures, in order to reduce the number of layers using the same concept but improving highly the training speeds offering similar results on complex datasets with lots of 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 in order to concatenate directly the prediction of the age and the prediction of the gender in only one output instead of create two separate classification networks for each task. The selection of the WideResNet backbone is justified due to the power offered by ResNet networks to work with a huge number of classes effectively and reducing the computational cost at training phase for big datasets.

**FIG. 7**

Table 4 collects all results after train the network for face recognition using different backbones and

results after train Face Characterization network. The selected backbone for face recognition network was VGG16. Analyzing the final values, it is possible can extract some valuable information about the behavior of each network. VGG16 reach high accuracy in both training and validation sets. This network is less complex than other tested network and no overfitting problem appear during training process. VGG19 start to present overfitting. This information can be extracted directly by the validation precision achieved. This network is more complex and probably don't generalize well all the classes performing worst at validation time. MobileNet (Howard et al., 2017) is a less complex architecture that works well but don't achieve the accuracy presented by the previous ones. ResNet50 is the clear example of very deep network that overfit quite quickly and performing poor at validation time over this not very complex dataset.

Face Characterization network achieve very high accuracy in both training and validation regarding that this classification task was more complex than the previous affirming that WideResNet was a good selection for this problem. After some experiments the error in the estimated age is most of the times between a little interval around the real person age.

**TABLE 4**

### 3.1.4. **Face Speaking**

Detect which character is speaking in a scene is an important information and could make the difference for users with visibility disabilities. The knowledge of who is speaking, for someone who is just listening the media content, could make him to better understand the scene. We already know how is on the scene as the algorithms are tracking and detecting the faces in the scenes as seen in previous sections, but to know who is speaking automatically using only image analysis our approach set the focus on detecting the mouth and its movements using landmarks detection.



#### *3.1.4.1* Landmarks detection

Landmark detection in faces is a well-known problem. Before Deep Learning algorithms started to solve these problems offering an increase in accuracy, Landmark detection was performed using a combination of traditional computer vision technique to extract facial features in combination 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 huge several datasets containing the location of these points in faces provide all 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 landmarks 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.

#### *3.1.4.2* Speaking detection

## 3.2. Image analysis for text detection

- Esquema funcionamiento
- Comparativa
- Resultados

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

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

## 4.1. Modular system architecture

The proposed architecture can be split on two different sides, as shown in Fig. 8:

**FIG. 8**

- The server side contains the image analysis service that uses the algorithms described in section 3 to preprocess the media files and to provide the accessibility information in the form of two different JSON files: the first one contains the data retrieved from the image analysis for text detection, and the second one the information retrieved from the image analysis for face detection and character recognition. These files are sent to the client application which is in charge of interpreting them and showing the additional information in an accessible way to final users through an app.
- In the client side the end user has a HbbTV environment consisting in a TV terminal together with a companion screen which are synchronized. This CS also contains an app that allows the user to play different video and other contents such as the developed accessibility tools presented in this paper: face detection, text detection and character recognition.



The structures of the JSON files is shown in **¡Error! No se encuentra el origen de la referencia.**. For the text case, the file is composed by two main parts: a frame objects array where the position corresponds to the frame number and the frame object that contains the information of the text detected in a single frame: the string of the detected text in the image and the position ((w.y) coordinates of the left upper corner position of the text on the screen, and the height and width of the text).

Regarding the face detection and characterization service, the data retrieved is stored in the JSON file, which is firstly composed by an array of frame objects again. Then, the frame object contains an array of face objects of that specify frame. Finally, the face object contains 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 upper left corner of the face and the weight and height of the area). Finally, it also contains a boolean variable that determines if the user is speaking or not.

**FIG. 9**

## 4.2. Access services in the CS app

### 4.2.1. Face detection 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 **¡Error! No se encuentra el origen de la referencia.**. When the option "face detection" 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.

**FIG. 10**

### 4.2.2. Text Detection

This tool is aimed to enlarge detected text on videos and present them in accessible colours. As in the previous case, to activate this tool there is a menu in the app with an on/off switch as shown in **¡Error! No se encuentra el origen de la referencia.**. When the option "text detected" is enabled the app will automatically enlarge the detected texts. The app will periodically get the frame information from the JSON and check if there is a text detected for the frame. When a frame contains a text, the app will show a box with the enlarged text and will wrap the detected text with vivid colour borders. The font size, font colour and background of the box are customizable. The text can be remove by swiping the box out of the screen.

**FIG. 11**



### 4.2.3. 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 **¡Error! No se encuentra el origen de la referencia.**. 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. In order to provide a complete accessible service, a text to speech service can be applied 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.

**FIG. 12**

# 5. DISCUSSION

Object detection and Tracking - Comparison with other methods that perform both stages at the same time.

There is a particular tracking paradigm, called tracking-by-detection (Fiaz, Mahmood, & Jung, 2018; Luo et al., 2014). Although all tracking systems need to detect the objects at some stage, tracking by-detection methods make a clear distinction between the detection and the tracking of objects. The general idea for each frame, is first localize all objects using an object detector and then associate them between frames using features such as location and appearance. However, the performance of tracking-by-detection models is heavily dependent on the accuracy of the detection

model. This fact indicates that CNN-based object detectors could greatly boost the tracking performance, even with simple tracking models but, having a powerful detector, can come at the cost of reduced frame-rate in real-time scenarios.

In order to compare our architecture some other architectures are presented. New Deep Learning solutions are gained the attention in this paradigm with architecture 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. Can be shown that after 10 iterations, computation time is reduced to near 30 ms per frame (30 FPS) achieving both detection and tracking but accuracy is also decreasing over time steps (0.45 – 0.4 mAP).

Mf-SSD exploit two different ideas, by one side, the inclusion of a RNN (similarly to ROLO) or replace this network by a Multi-Fusion stage in order to improve 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 depends directly on the use of fusion technique or the RNN reducing accuracy at cost of increase speed or reverse.

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). The main problem relies on 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.

Our approach maintains separate blocks for object detection and tracking and differs from the presented works. The motivation to continue using separate blocks is due to the availability to control both parts more easily. Merge both in only one framework is a recent field of study and is achieving some interesting results, but looking at the papers some problems keep appearing related to complexity and time processing. Another important part to take into consideration is the training, where a completely annotated Face Tracking Dataset in video is needed to learn not only detect faces but also track them. The availability of these datasets is not huge in comparison with tracking datasets for objects. An example of dataset is YouTube Faces (Wolf, Hassner, & Maoz, 2011). Our architecture avoids this problem by training object detector separately using annotated faces and training (if needed) the tracking algorithm in common objects, because the starting bounding boxes are given by the object detector. Tracking algorithms alone are achieving very high computation speed as shown Table X in our experiments and decreasing the total time consumed if more tasks have been done after tracking faces every frame. Our work preserves the accuracy all the time if there are no occlusions or complex fast behaviors during scenes. As it takes more computational time, the objet extraction is done once every fixed number of frames while the tracking algorithm keeps making predictions all the time. This approach is faster than the presented in other works. All our experiments were performed using NVIDIA GeForce 1080 Ti GPU.



TEXT detection comparison

# 6. CONCLUSIONS AND NEXT STEPS

On this bases, Hybrid TV, and in particular HbbTV, together with a thorough definition of innovative services where deep learning can provide competitive strengths will help the enhancement of the access to media content for all.

## ACKNOWLEGMENT

This work has been partially funded by the EC H2020-ICT-19-2016-2 76199 “EasyTV: Easing the access of Europeans with disabilities to converging media and content”. The authors would like to acknowledge CCMA (Corporació Catalana de Mitjans Audiovisuals) broadcaster for the cession of contents used for this research.

# REFERENCES


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

Anthimopoulos, M., Gatos, B., & Pratikakis, I. (2010). A two-stage scheme for text detection in video images. *Image and Vision Computing*, *28*(9), 1413–1426.

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

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, D., Roy, P. P., Saini, R., Alvarez, J. M., & Pal, U. (2018). Frame selection for OCR from video stream of book flipping. *Multimedia Tools and Applications*, *77*(1), 985–1008.

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

Cisco, V. N. I. (2017). Cisco Visual Networking Index: Forecast and Methodology 2016--2021.(2017).

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/

Islam, M. R., Mondal, C., Azam, M. K., & Islam, A. S. M. J. (2016). Text detection and recognition using enhanced MSER detection and a novel OCR technique. In *2016 International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 15–20).

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

Khare, V., Shivakumara, P., Raveendran, P., & Blumenstein, M. (2016). A blind deconvolution model for scene text detection and recognition in video. *Pattern Recognition*, *54*, 128–148.

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

Lu, W., Sun, H., Chu, J., Huang, X., & Yu, J. (2018). A Novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network. *IEEE Access*, *6*, 40198–40211. https://doi.org/10.1109/ACCESS.2018.2851942

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.

Roy, S., Shivakumara, P., Roy, P. P., Pal, U., Tan, C. L., & Lu, T. (2015). Bayesian classifier for multi-oriented video text recognition system. *Expert Systems with Applications*, *42*(13), 5554–5566.

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.

Shivakumara, P., Wu, L., Lu, T., Tan, C. L., Blumenstein, M., & Anami, B. S. (2017). Fractals based multi-oriented text detection system for recognition in mobile video images. *Pattern Recognition*, *68*, 158–174.

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 by lalala. 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*.

Weinman, J. J., Learned-Miller, E., & Hanson, A. R. (2009). Scene Text Recognition Using Similarity and a Lexicon with Sparse Belief Propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(10), 1733–1746. https://doi.org/10.1109/TPAMI.2009.38

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

Ye, Q., & Doermann, D. (2015). Text Detection and Recognition in Imagery: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(7), 1480–1500. https://doi.org/10.1109/TPAMI.2014.2366765

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

# FIGURES

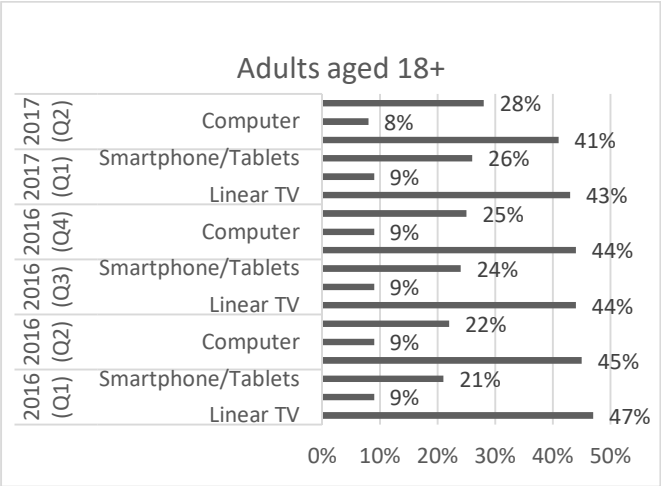


**Fig. 1. Average audience composition by platform among adults aged 18+ [from Nielsen report Q1 2016** (NIELSEN a, 2017)**, Q2 2016** (NIELSEN b, 2017)**, Q3 2016** (NIELSEN c, 2017)**, Q4 2016** (NIELSEN d, 2017)**, Q1 2017** (NIELSEN e, 2018)**, Q2 2017** (NIELSEN f, 2018)**].**

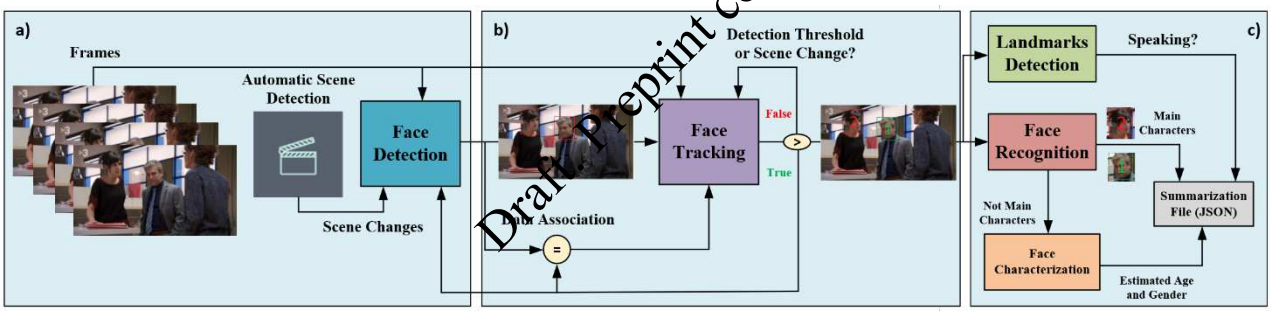




**Fig. 2. Main architecture to data extraction. a) Face detection with deep learning networks is implemented to find faces in video. b) Deep Tracking over the detected faces achieve follow the same faces along the time improving highly the computational time. c) Detected faces pass over several networks in order to extract main information to provide the most important features for accessibility purposes.**

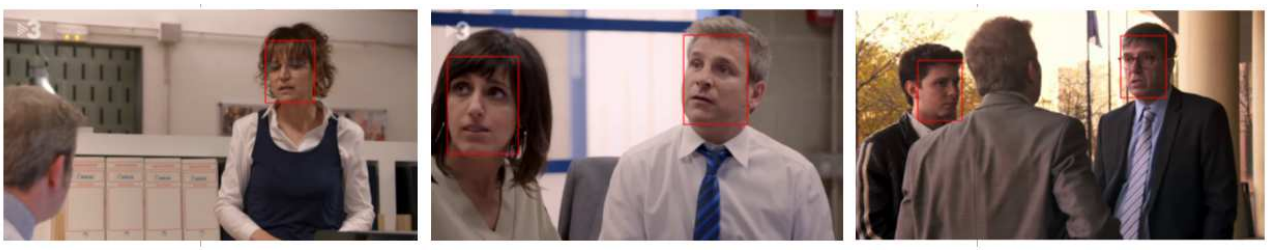

**Fig. 3. Example of detected faces in video**

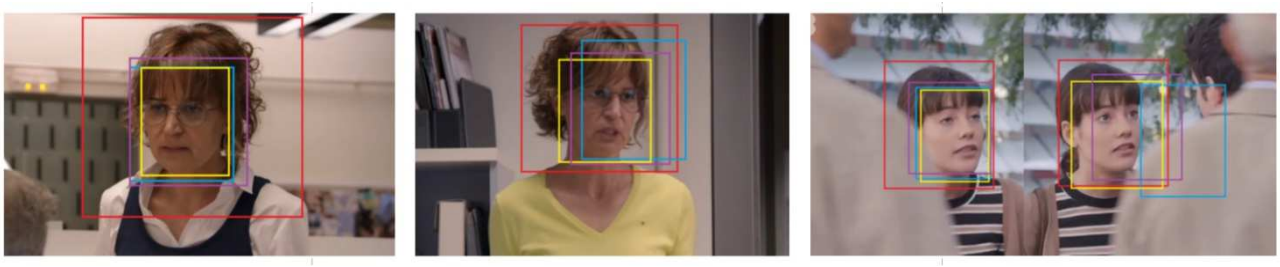

**Fig. 4. Failure examples during tracking with tested algorithms. GOTURN (red), CSRT (yellow), KCF (blue) and Re3 (magenta). Slow motion (left), fast motion (centre) and occlusion (right)**

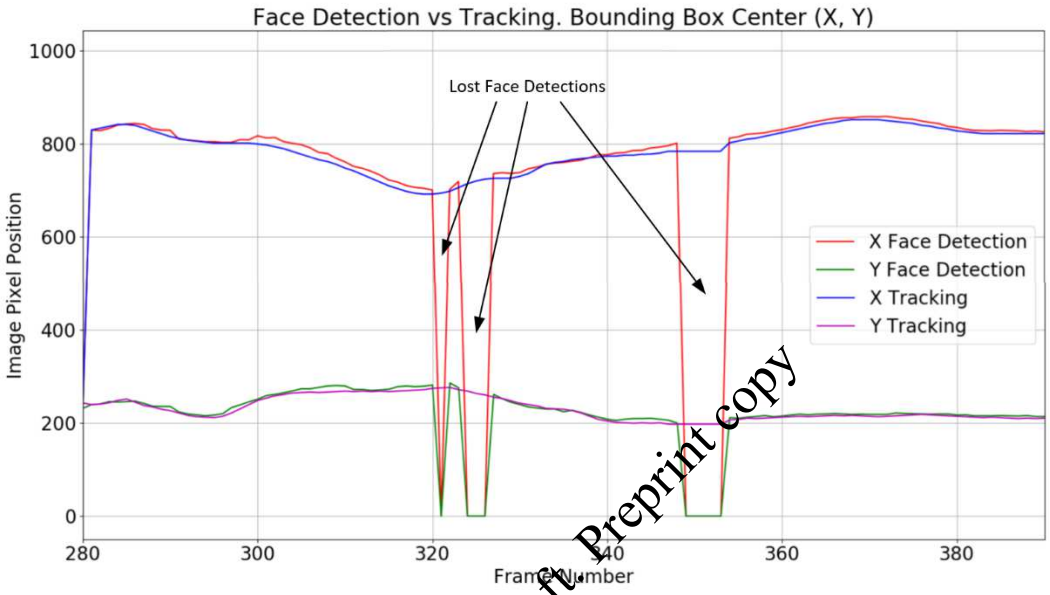




**Fig. 5. Performance comparison between the face detector applied in all frames and the tracking over the first detected face.**


**Fig. 6. Examples of the main characters (top) and some images in our dataset for the same characters (bottom)**

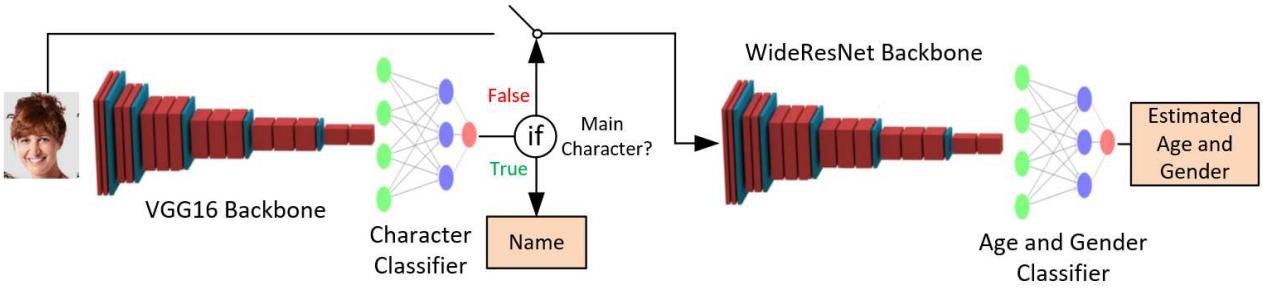


**Fig. 7. Architecture for face recognition and characterization network**

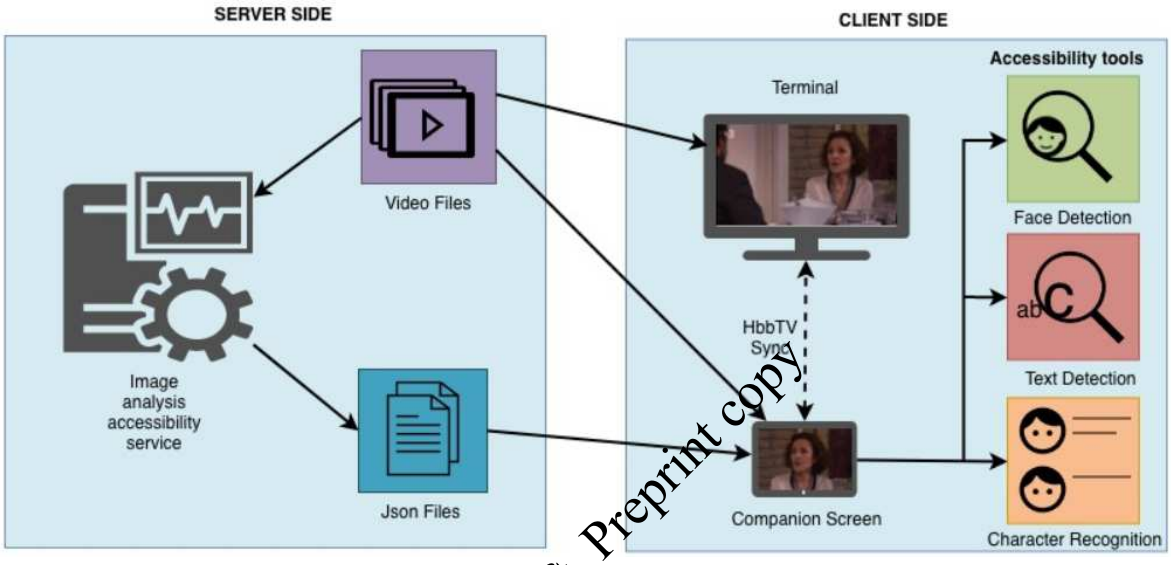




**Fig. 8. Modular system architecture for new accessibility services**

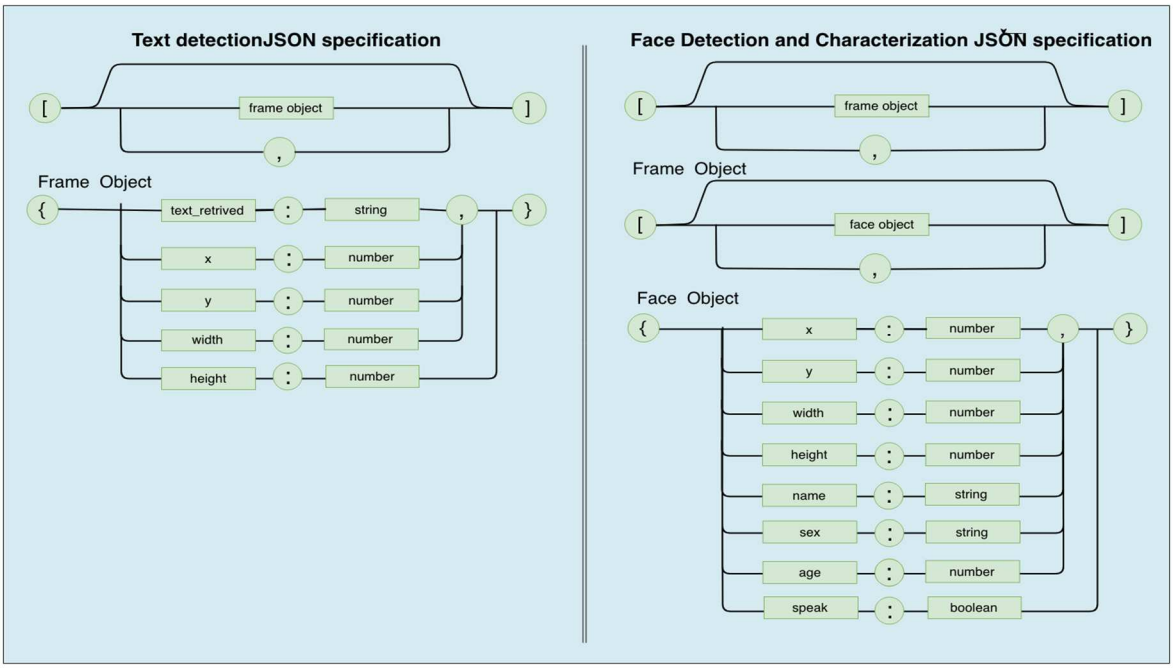


**Fig. 9 Structure of the JSON files from text analysis service (left) and face analysis service (right**

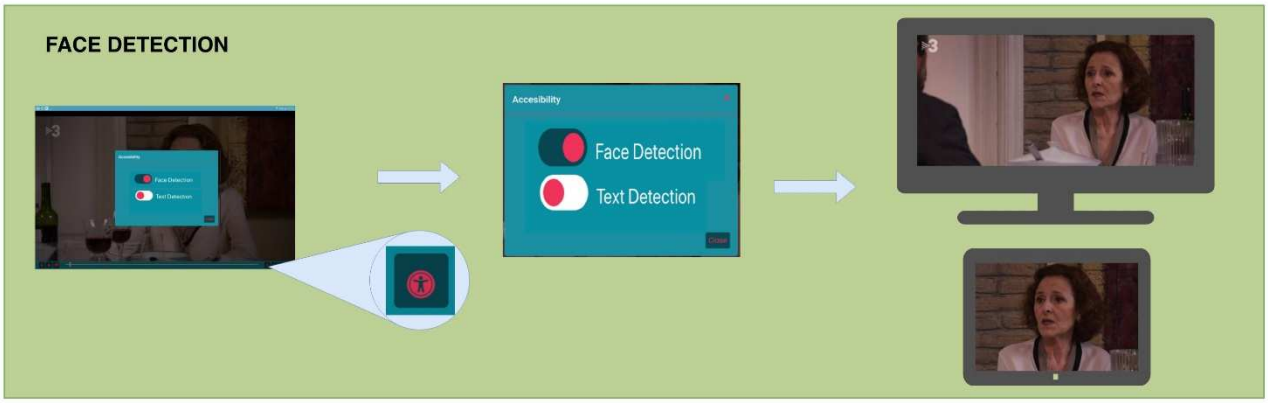


**Fig. 10 Automated face detection and magnification service in the CS app**


**Fig. 11 Automated text detection and magnification service in the CS app**




**Fig. 12. Automated character recognition service in the CS app**

# TABLES

**Table 1. Most typical problems with the multimedia content accessibility 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. |



**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 |

**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 |

**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 |

# AUTHOR BIOGRAPHIES

**Silvia Uribe**

Silvia Uribe received the Telecom Engineer degree (Hons) in February 2008, the Master in Communications Technologies and Sistems in September 2010, the Master in Telecommunication Management in 2013 and finally the Ph.D degree (cum laude) in 2016 by the “Universidad Politécnica de Madrid”(UPM). She is a member of the Visual Application Telecommunication Group(G@TV) since 2006.
Her professional interests include interactivity technologies, content personalization technologies and big data. Related to this, she has been participating with technical responsibilities in some national (Buscamedia, Ciudad2020, LPS-BIGGER, Repara 2.0) and European (eASLA, LASIE, EasyTV) projects, and she is author and co-author of several papers and scientific contributions in international conferences and journals.

**Alberto Belmonte**

Alberto Belmonte Hernández received the degree in Telecommunication Engineering (2014) and the Master’s degree (2016), focused on communication systems, from the Technical University of Madrid (UPM). He worked in EVERIS SPAIN SL using Liferay programming tool and JAVA programming language to deploy webservices and web pages. Currently he is working for the research group in the Visual Telecommunications Applications group (GATV) of UPM and is PhD candidate. His main interests are the new communications technologies, Internet of Things (IoT), sensors, wireless communications, computer vision and image processing. He is working actively on the development of indoor tracking algorithms, accurate localization and image detection and classification with cameras and sensors combining the information from both and using Machine and Deep Learning algorithms approaches.



**Juan Pedro López**

Juan Pedro López is a Telecommunication Engineer since 2007, he got the International Doctor’s Degree at Universidad Politécnica de Madrid (UPM) in February 2016 with his thesis in the field of quality assessment for 2D and 3D stereoscopic video.
His professional interests include video encoding, compression formats, high and ultra-high definition television, signal processing and innovation that relates technology with environments such as accessibility, education, healthcare and art.
During his more than ten years of work in research and development he has developed applications in different programming languages and platforms, including C++, C#, Android or JavaScript.
Since 2008 he complements teaching with working in national and international research projects about video encoding, quality of experience, broadcasting, accessibility and HbbTV technologies. He got the Bachelor of History of Art of UNED University in July 2017.

**Francisco Moreno**

Francisco Moreno received his BSc in computer science engineering in 2008 on the Technical University of Madrid. In 2007 he finished his final degree project called “Information Retrieval In Email Fields” on the Roskilde University (Denmark). From 2007 to 2012 he worked for different companies developing webs, where he got to be the Chief Developer in the advertisement agency Srburns. He moved to United Kingdom in 2012 where he worked two years and a half as mobile developer. Since October 2015 he is working on the GATV group where he works developing webs and mobile apps in different research projects.

**Álvaro Llorente**

Álvaro Llorente got the Bachelor of Engineering in Telecommunication Technologies and Services at Escuela Técnica Superior de Ingenieros de Telecomunicación (ETSIT) of Universidad Politécnica de Madrid (UPM). In July 2016 he carried out the Final year Project, "Design and execution of subjective quality tests on Ultra High Definition TV". Since 2015 he collaborates with the Grupo de Aplicación de Telecomunicaciones Visuales (GATV) of Universidad Politécnica de Madrid and with the Chair of RTVE in UPM. Currently, he studies the Master in Telecommunication Engineering at ETSIT UPM.

**Federico Álvarez**

Dr. Federico Alvarez is Telecom Engineer with honours (2003) and Ph. D. (2009), both by the “Universidad Politécnica de Madrid”. He is working as assistant professor lecturing in the “Telecommunication Systems” and “Visual Communications” area in UPM. He develops his research within the research group in the Visual Telecommunications Applications group (GATV) of the "ETS Ingenieros de Telecomunicación" of the "Universidad Politécnica de Madrid". He is nowadays the coordinator of EasyTV and FI-GLOBAL, and technical coordinator of ICT4LIFE in H2020. He has been in the last 10 years also leading the UPM participation in several EU-funded projects, such as the SEA, SIMPLE, AWISSENET, RESCUER, FI-PPP project XIFI (on experimentation infrastructures), .... and coordinated the projects nextMEDIA, INFINITY and FI-LINKS. He worked as expert for the European Institute for Prospective Technological Studies for mobile search. He had taken part in standardisation bodies such as DVB-ETSI or CENELEC TC206 and is author and co-author of (70+) papers in journals, congresses and books in the field of ICT technologies. He is serving in the Programme Committee of several congresses and as reviewer of scientific journals. He organised the Future Internet Assembly in Madrid in December 2008.