*Artificial Intelligence for Engineering Design, Analysis and Manufacturing*

**cambridge.org/aie**

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

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

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

**Research Article**

**Cite this article:** Uribe S, Belmonte A, Moreno F, Llorente Á, López JP, Álvarez F (2019). New access services in HbbTV based on a deep learning approach for media content analysis. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **33**, 399–415. https://doi.org/10.1017/S0890060419000350

Received: 30 November 2018
Revised: 19 March 2019
Accepted: 1 April 2019
First published online: 4 December 2019

**Key words:**
Computer vision; deep learning; face detection; media accessibility

**Author for correspondence:**
Silvia Uribe, E-mail: sum@gatv.ssr.upm.es

**Abstract**

Universal access on equal terms to audiovisual 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 face detection for magnification and character identification. 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.

## 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 [e.g., in France (CSA, 2017)] or even provides less accessible content than that [e.g., 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 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 multiplatform, 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 Internet Protocol television (IPTV), over-the-top television (OTT-TV), or hybrid broadcast-broadband television (HbbTV) (Malhotra, 2013), the introduction of companion screens (CS) for multi-screening may define the new consumer behavior (Vinayagamoorthy *et al.*, 2012), where users may

© Cambridge University Press 2019

CAMBRIDGE UNIVERSITY PRESS

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.

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 accessibility services.

In the other 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 (Sáez Trigueros *et al.*, 2018; Voulodimos *et al.*, 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 paper 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 paper 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 the one hand, the new television paradigm for content provision, that is, the hybrid television due to its own capabilities and, on the other hand, the application of different artificial intelligence tools for content processing.

In this regard, these innovative services represent a new user-side centered 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 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 center 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 paper is outlined as follows: The section "Related work and problem formulation" 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 the section "Video analysis for facial information detection", including the explanation about the applied algorithms and the obtained results for our approach. Then, the section "New access services for the HbbTV environment based on the image analysis" provides a wide presentation of the final services, indicating how they are implemented and how the previous results are presented to the final users. The section "Discussion" includes a discussion about the obtained results. Finally, the section "Conclusions and next steps" is focused on presenting the conclusions and the related future work.

## 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 behavior 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 fulfill 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 analyzing the related research problem and contextualizing our approach.

### *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. Figure 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 the United States population [millennials and teens (McNally and Harrington, 2017)].

In this regard, there are two main facts that give an idea of the current multimedia market evolution: on the one hand, the increase of the Smart TVs purchase that has tripled in the last 5 years (Statista, 2017) and, on the other hand, 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 allow fulfilling 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.

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 digital terrestrial television (DTT) 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, personalized advertisement, statistical data information, or accessibility services (subtitles, sign language, audio description or others that facilitate access to content for

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

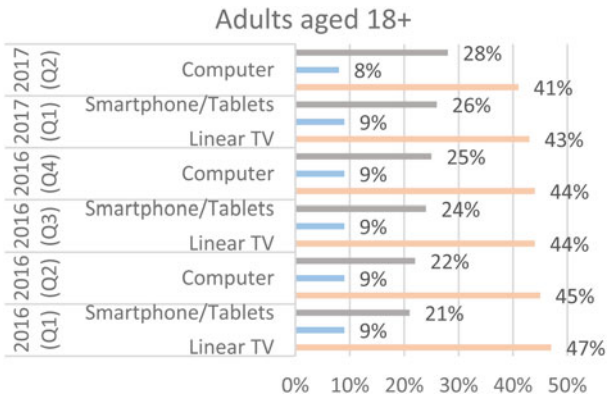



**Fig. 1.** Average audience composition by platform among adults aged 18+ [from Nielsen report Q1 2016 (NIELSEN a, 2017), Q2 2016 (NIELSEN b, 2017), Q3 2016 (NIELSEN c, 2017), Q4 2016 (NIELSEN d, 2017), Q1 2017 (NIELSEN e, 2018), Q2 2017 (NIELSEN f, 2018)].

people with visual or hearing disabilities, or elderly people). This provides a disruptive scenario where the enhancement of the user experience will be the center of the industry.

In relation to the accessibility aspect in this scenario, people with disabilities like watching TV, as 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 and 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 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 Regulators have imposed media accessibility to broadcasters, but the specification about actual goals or the 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 analyzing their fulfillment 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 *et al.*, 2015), Prosperity4All (Prosperity4All Project), ImAc (Immersive Accessibility Project), and EasyTV (EasyTV Project), where the work done for this paper is framed.

## *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 (dynamic adaptive streaming over HTTP) (Sodagar, 2011) service provision.
- User-centric approach: EasyTV will propose a methodology based on a continuous user feedback 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. Moreover, these tools will be integrated into a complete platform to be directly used by any broadcaster. Regarding the user requirements gathering, the performance of focus groups with final users is vital for our approach and helps us to understand which are the main demands in the real scenario. In this context, the next section is focused on presenting some of them.

## *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 know 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.

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 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.

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

**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 | – Difficulties to access the TV content<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 |

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

## *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 EU will reach 120 million by 2020 and World Health Organization (WHO and 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 audio-visual 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 aspects that are the basis of our approach.

### *Deep learning techniques for visual information gathering*

With this objective in mind, it is important to say that nowadays the application of different deep learning methods are 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 allows deep learning algorithms to improve their performance, so the chance is absolutely 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 the 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) (Xu *et al.*, 2011; Ahmed *et al.*, 2017), Haar cascade classifiers for face detection (Padilla *et al.*, 2012; Cuimei *et al.*, 2017), and semantic segmentation of objects in a scene (Yuheng and 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 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, such as faster region-based convolutional neural network (faster R-CNN; Ren *et al.*, 2015), single-shot detector (SSD; Liu *et al.*, 2015), you only look once (YOLO; Redmon *et al.*, 2015), and RetinaNet (Lin *et al.*, 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, such as kernelized correlation filter (KCF; Henriques *et al.*, 2014), minimum output sum of squared error (MOSSE; Danelljan *et al.*, 2014), discriminative correlation filter with channel and spatial reliability (DCF-CSRT; Lukezic *et al.*, 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 *et al.*, 2016), GOTURN (Held *et al.*, 2016), and Re3 (Gordon *et al.*, 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 *et al.*, 2016) [scale-invariant feature transform (SIFT), speeded-up robust features (SURF), oriented FAST and

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

rotate BRIEF (ORB)] 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 *et al*., 2016) or DensePose (Güler *et al*., 2018) are now at the top of body keypoint detection algorithms, and Mask R-CNN (He *et al*., 2017) is able to perform object segmentation.

### Face detection problem in image analysis

Face detection is a well-known problem in computer vision. In this regard, several methods have been already proposed in order to extract patterns from images and detect faces from them (Chakraborty and Das, 2014). In general, all these methods consist of three parts: firstly, an algorithm to inspect parts of the images (sliding window and region proposal); secondly, the obtaining of extracted features from these parts (Haar features, HoG features, and deep learning features); and finally, their classification whether they are 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 (Wang and Deng, 2018) show that it is 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 and 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 R-CNN 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 provided less accuracy results than faster R-CNN 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 and Zisserman, 2014) and ResNet (He *et al*., 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 used 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" (Luo *et al*., 2014; Fiaz *et al*., 2018), 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 *et al*., 2018), and Detect and Track (D&T) (Feichtenhofer *et al*., 2017).

## Proposed approach

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 preprocessed content, considering their application to broadcast delivered content as a possible future research work, which will be fully dependent 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:

1. Firstly, and as mentioned before, two focus groups were arranged to collect the users' requirements. Derived from them, two main services focused on the image processing were defined: image magnification and character recognition.
2. 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 with 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 to our custom dataset in order to achieve real or very near real-time processing. Our framework demonstrated that all these algorithms can work

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

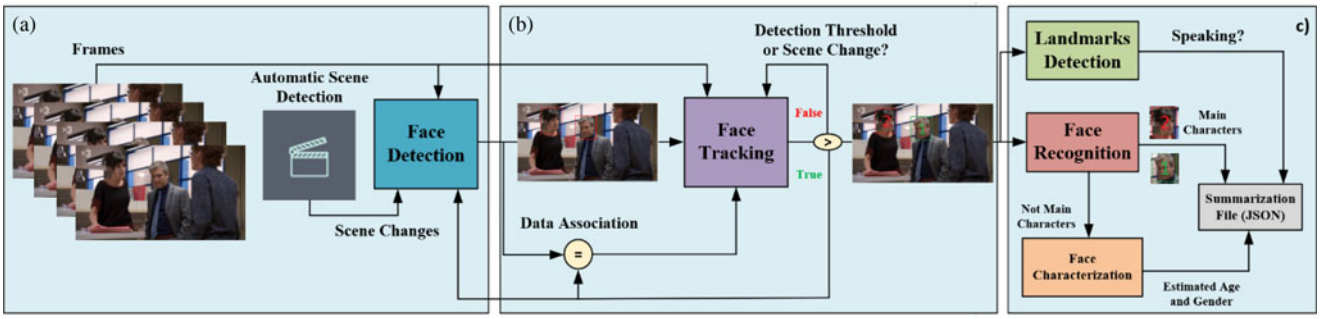



**Fig. 2.** Main architecture for data extraction. (a) Face detection, (b) face tracking, and (c) information extraction, including face recognition, characterization, and speaking detection.

together and provide better access solutions when retraining these networks with the targeted new data.

3. 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 with end users will be set during next months to obtain their feedback in order to improve them for a final version (the tests and their results are out of the scope of this paper, and they are described as a future work line).

## Video analysis for facial information detection

The modular architecture for the data extraction for face detection service can be seen in Figure 2, 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 with 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 accessibility services. It includes character recognition, face 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 is integrated into 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 the eighth generation, and one Nvidia GeForce 1080 Ti GPU.

### Face detection

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 and 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 has 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, Figure 3 shows some examples of the detected faces over different video frames.

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.

**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 300 × 300) with VGG16 backbone | 46 | 74.3 |
| SSD300 (input 300 × 300) with ResNet10 backbone | 39 | 77.8 |
| YOLOv2 (original implementation) | 57 | 64.4 |

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

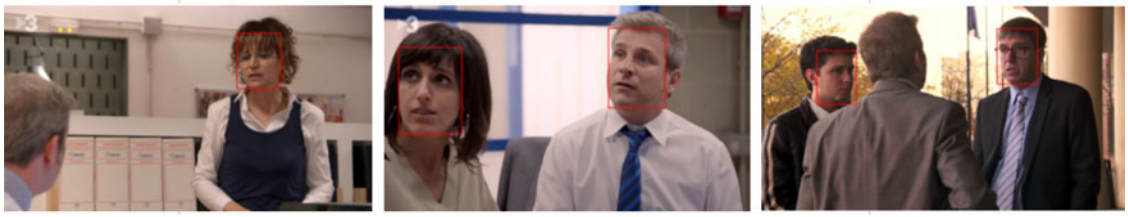


**Fig. 3.** Example of detected faces in the video.

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

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

### *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 as follows:

- Bounding box size: since it is not fixed for the face detector along with 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 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 the previous information. In this context, KCF makes the use of an adaptive threshold approach, which includes a KCF method and a Kalman filter algorithm to make tracking faster and more accurate. CSRT tracker takes the advantage of DCFs 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 for 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 tests. 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.

According to the information in Table 3, the 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 specific faces dataset. Finally, KCF presents high accuracy and the processing time increases slowly with the number of

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

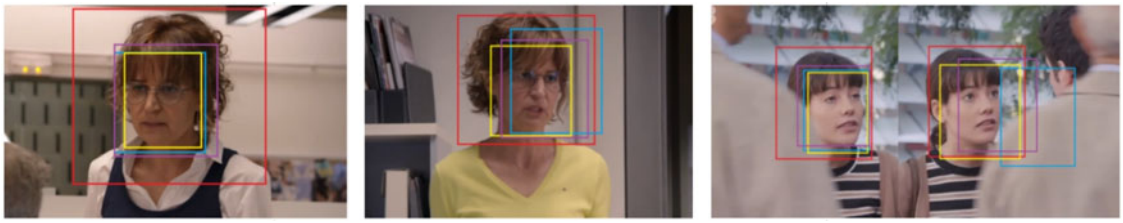


**Fig. 4.** Failure examples during tracking with tested algorithms. GOTURN (red), CSRT (yellow), KCF (blue), and Re3 (magenta). Slow motion (left), fast motion (center), and occlusion (right).

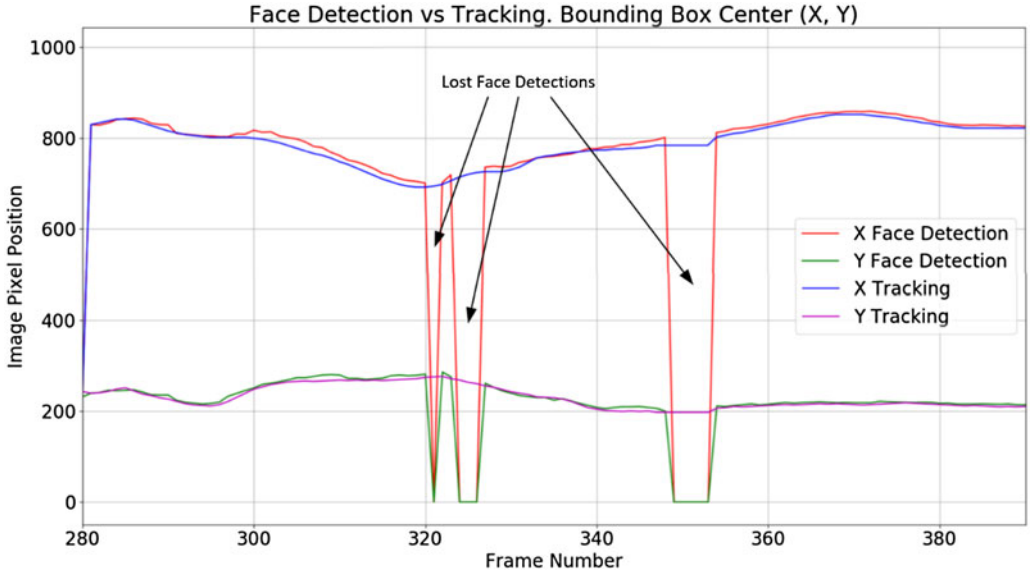



**Fig. 5.** Performance comparison between the face detector applied in all frames and the tracking over the first detected face.

trackers, but its main problem is the high probability to fail in the presence of important occlusions that hide a wide area of the tracked face.

Figure 4 shows some examples of failure with each algorithm. The left image represents a situation with slow motion, where all algorithms work well but GOTURN bounding box is bigger than the others. The 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, the right image presents the situation before and after occlusion. Re3 tracker bounding box has suffered a little displacement but KCF confidence decreases 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 5 shows an example of the comparison of the results obtained by using face detection versus 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.

## *Information extraction*

### *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 CNN 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.

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press




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

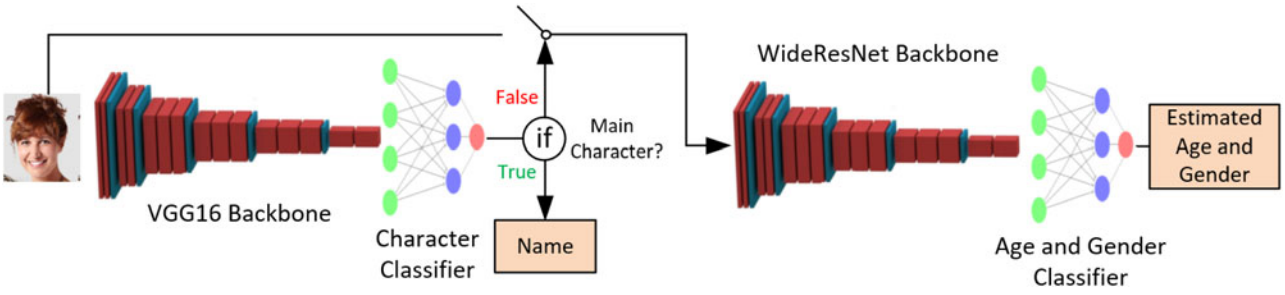



**Fig. 7.** Architecture for face recognition and characterization network.

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. Figure 6 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 and Qi, 2017) (more than 20,000 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 *et al.*, 2018) (more than 460,000 images of famous people annotated with age and gender).

The network architecture is divided into two branches, where the second branch is activated only when a detected face is not the main character, as shown in Figure 7. To compose face recognition network, different backbones have been tested with a classification network, which contains two neurons layers of 1024 neurons each, and a final SoftMax activation to get the final prediction. The input has been resized to a fixed size of $224 \times 224$.

The second network used for face characterization uses the WideResNet network (Zagoruyko and 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 (0–99 years) and two 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 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 no 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.

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

**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 (ms) |
|---|---|---|---|
| VGG16 (Recognition) | 95.2 | 92 | 2 |
| VGG19 (Recognition) | 97.5 | 88 | 3 |
| MobileNet (Recognition) | 89.1 | 88.2 | 2 |
| ResNet50 (Recognition) | 98.2 | 56.3 | 4 |
| WideResNet (Characterization) | 86.3 | 80.4 | 7 |

### *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 to 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, the 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 (Monzo *et al.*, 2010; Liu *et al.*, 2011).

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 *et al.*, 2016). It contains re-annotated landmark points (68 in total) in several available datasets, such as LFPW (Belhumeur *et al.*, 2013), AFW (Zhu and Ramanan, 2012), HELEN (Le *et al.*, 2012), and XM2VTS (Messer *et al.*, 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 that 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 Figure 8.

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. Figure 9 shows the difference between 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%.

## New access services for the HbbTV environment based on the image analysis

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

### *Modular system architecture*

The modular architecture of the proposed services can be split into two different sides, as shown in Figure 10, according to a client-server model:

- The server side contains the image analysis algorithms described in the section "Video analysis for facial information detection" 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 playing different video and other contents such as the developed accessibility tools presented in this paper: face detection and character recognition. As in the HbbTV environment, this CS will be synchronized with the TV hybrid terminal.

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

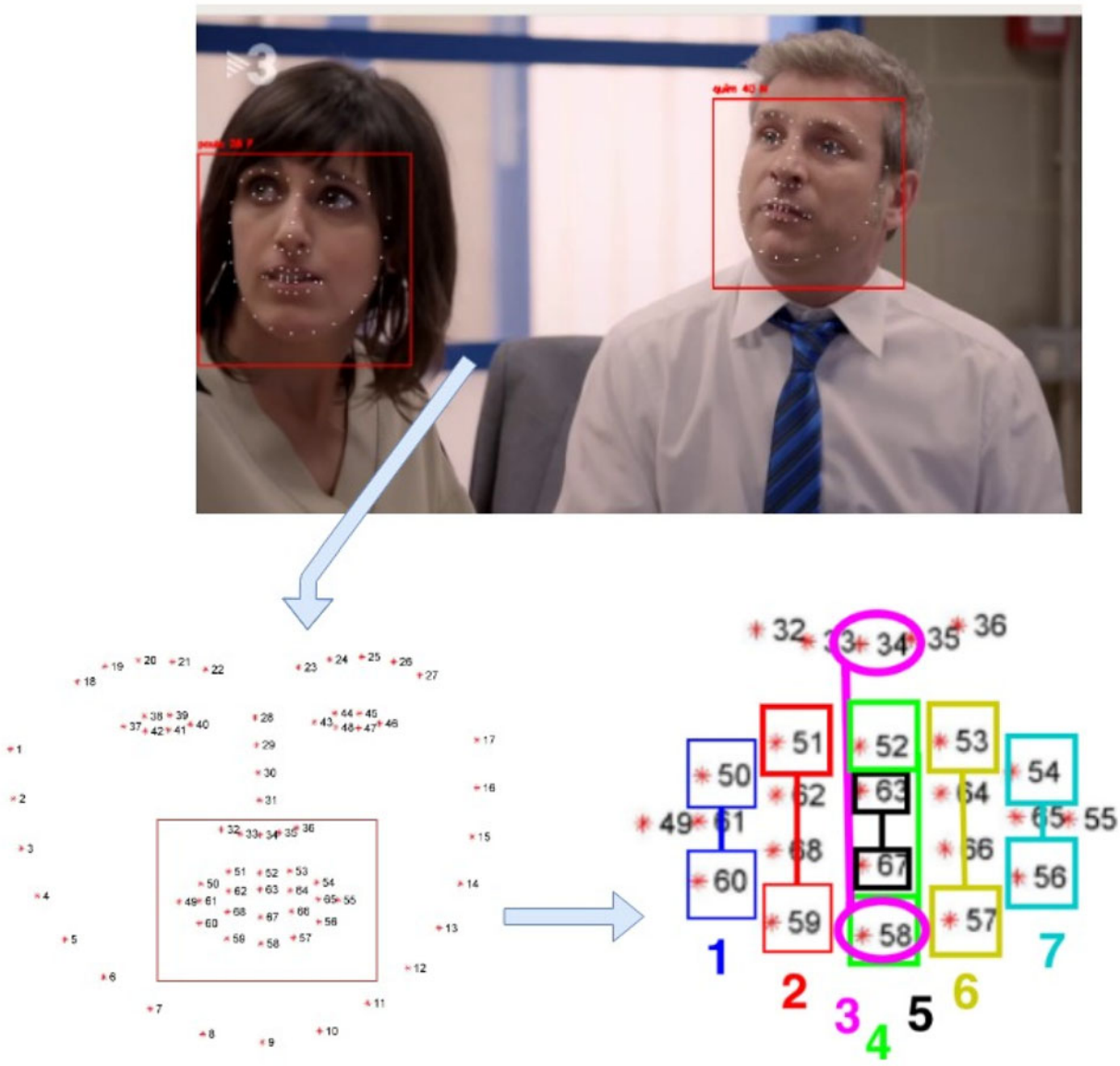


**Fig. 8.** Landmark detection and key points pairs difference.

The structure of the output from the image analysis module is shown in Figure 11. 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.

## Access services in the CS app

### 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 Figure 12. 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.

### 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 Figure 13. 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, and in order to provide a complete access service, a text to speech service

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

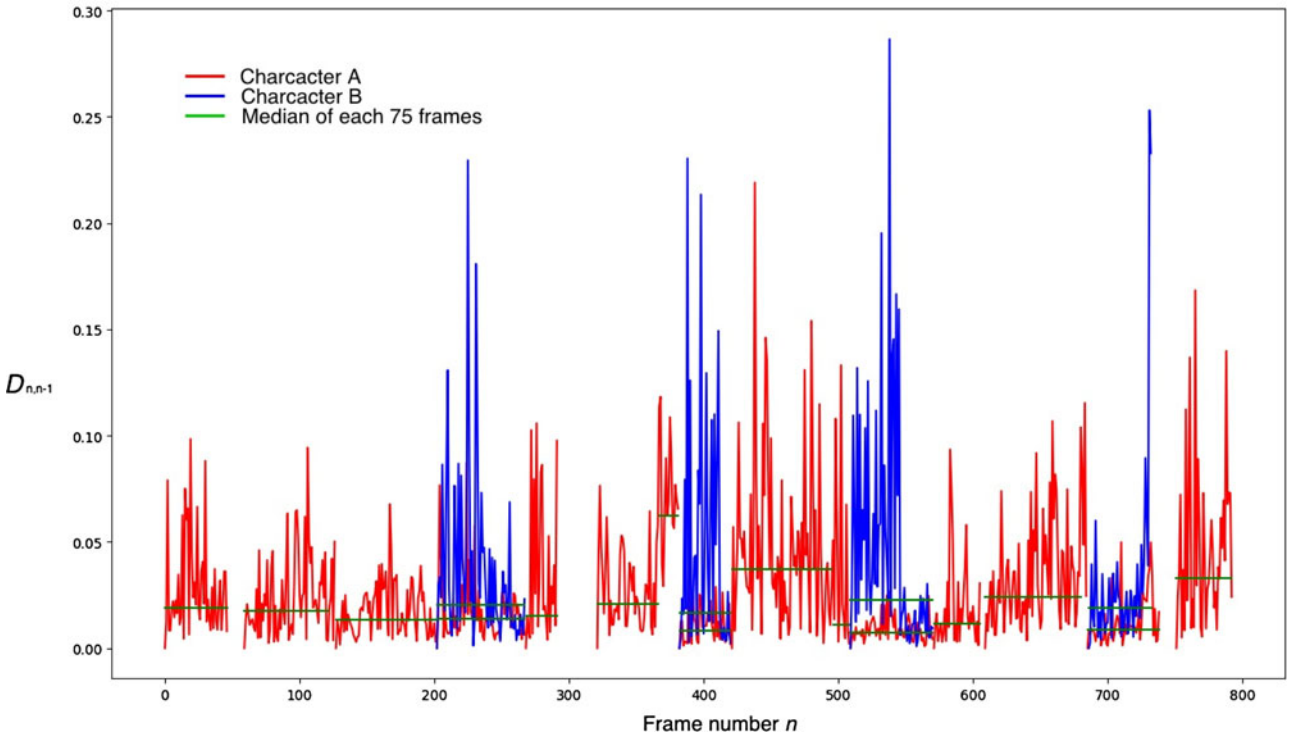



**Fig. 9.** Mouth landmark points difference along consecutive frames of two characters.

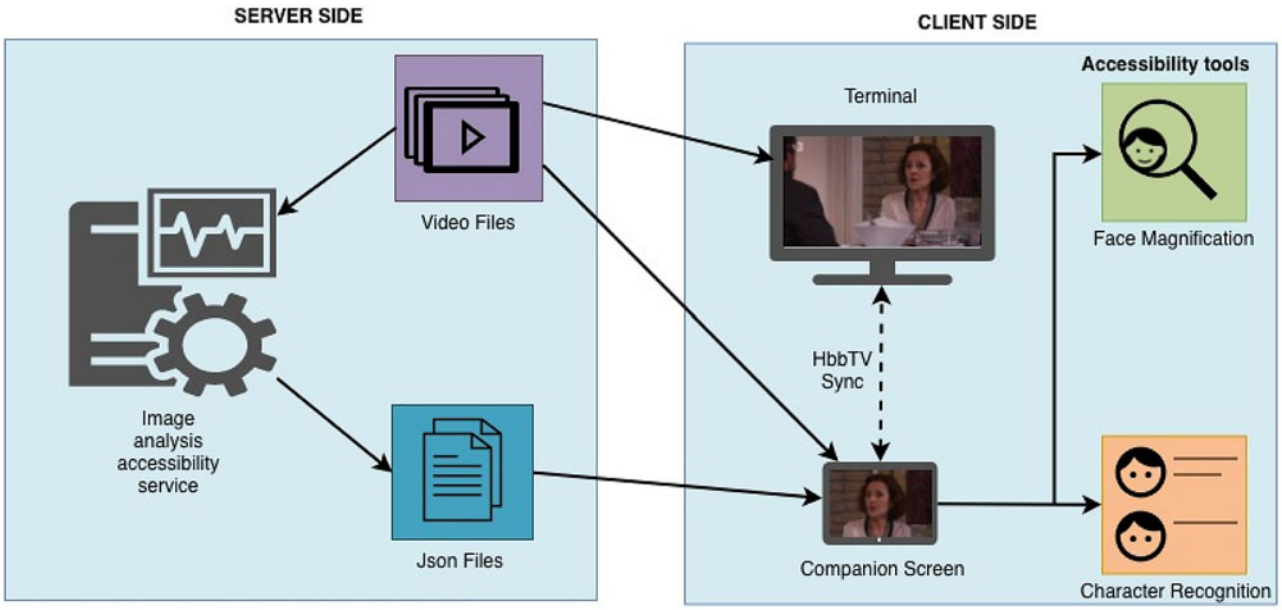



**Fig. 10.** Modular system architecture for new accessibility services.

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 provided by the image analysis module on the server side.

## Discussion

Regarding the main purpose of this study, throughout this paper we have presented two innovative accessibility services for audio-visual content in the 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 of 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 (Luo *et al.*, 2014; Fiaz *et al.*, 2018), where the first step, as in our case, is to detect all objects using an object detector in each frame, and then associate them

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

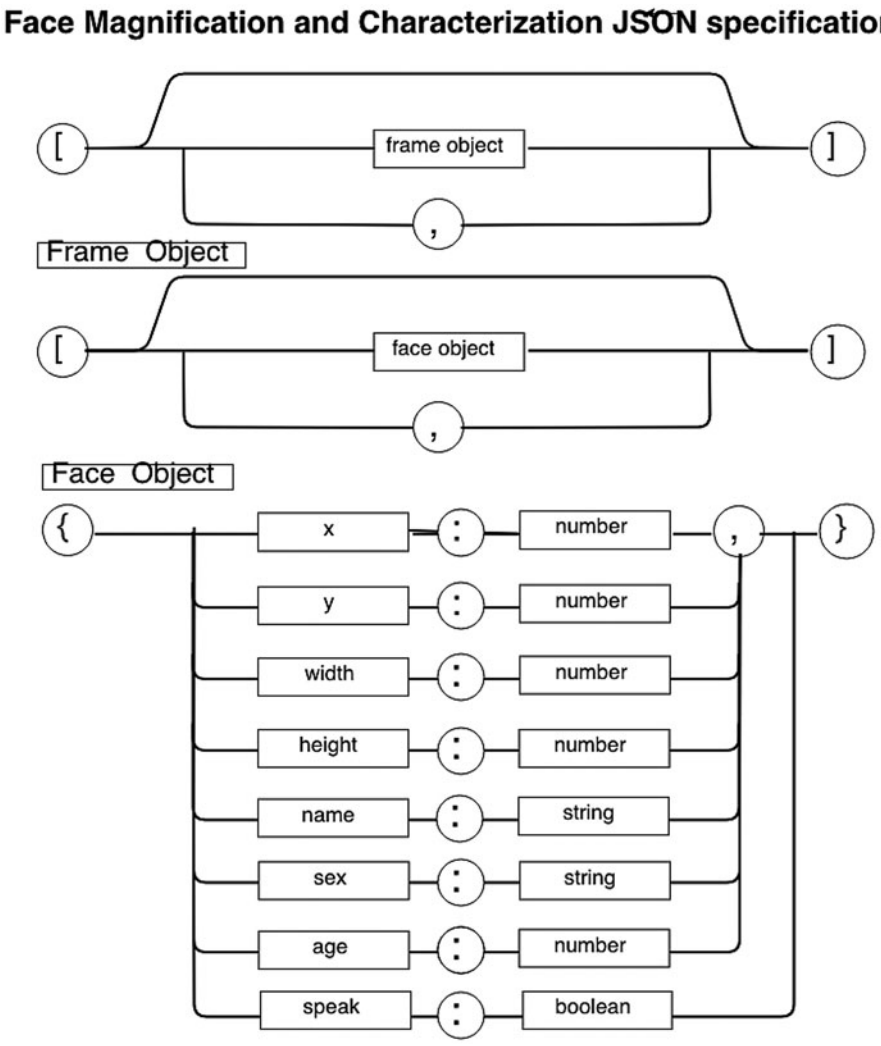



**Fig. 11.** Structure of the JSON file generated for face analysis service.

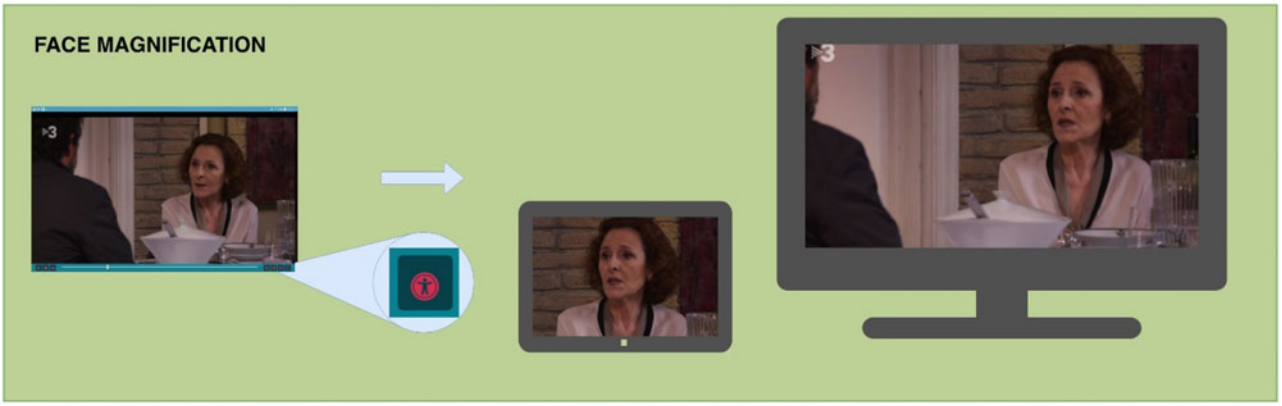



**Fig. 12.** Automated face magnification service in the CS app.

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 (Feichtenhofer *et al.*, 2017; Ning *et al.*, 2017; Broad *et al.*, 2018; Fiaz *et al.*, 2018) 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

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

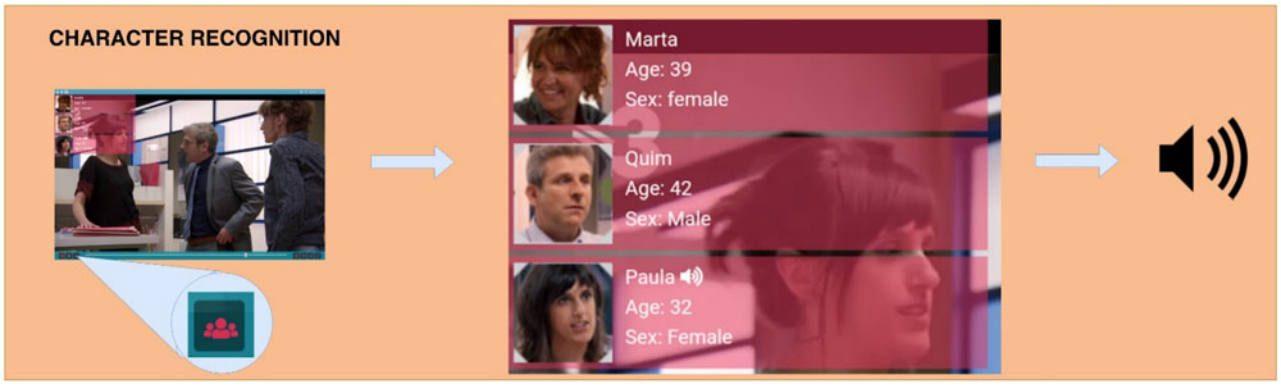



**Fig. 13.** Automated character recognition service in the CS app.

the environment evolves, new deep learning solutions have 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, such as ROLO (recurrent YOLO) (Ning *et al.*, 2017), Mf-SSD (Broad *et al.*, 2018), and D&T (Feichtenhofer *et al.*, 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 the one hand, the inclusion of an 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 and 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 *et al.*, 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 behavior. 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 to 2%.

Finally, these new access services give broadcaster and end users with additional valuable functionalities: broadcasters may save annotation resources by the automatication 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.

## Conclusions and next steps

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, in this paper 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 the section "Lack of media content accessibility in TV consumption"). 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 three-step image analysis work flow, 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 accessibility service composition, as explained in the section "New access services for the HbbTV environment based on the

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

image analysis", 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, such as speaking, singing, screaming, and eating, 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 is planned in order to obtain their feedback regarding the proposed services and allow their improvement.

**Acknowledgments.** 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 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 and Rothe R** (2017) Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA, pp. 87–94.

**Ahmed AH, Kpalma K and Guedi AO** (2017) Human detection using HOG-SVM, mixture of Gaussian and background contours subtraction. *2017 13th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pp. 334–338. doi:10.1109/SITIS.2017.62

**Belhumeur PN, Jacobs DW, Kriegman DJ and Kumar N** (2013) Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 2930–2940.

**Bertinetto L, Valmadre J, Henriques JF, Vedaldi A and Torr PHS** (2016) Fully-convolutional Siamese networks for object tracking. *CoRR, abs/1606.0*. Available at http://arxiv.org/abs/1606.09549

**Broad A, Jones M and Lee T-Y** (2018) Recurrent multi-frame single shot detector for video object detection. *British Machine Video Conference (BMVC),* Newcastle, UK.

**Cao Z, Simon T, Wei S-E and Sheikh Y** (2016) Realtime multi-person 2D pose estimation using part affinity fields. *CoRR, abs/1611.0*. Available at http://arxiv.org/abs/1611.08050

**Castellano B** (2018) Pyscenedetect. Available at https://pyscenedetect.readthedocs.io

**Chakraborty S and Das D** (2014) An overview of face liveness detection. *CoRR, abs/1405.2*. Available at http://arxiv.org/abs/1405.2227

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

**CNMC** (2017) Informe sobre el seguimiento de las obligaciones impuestas en materia de accesibilidad correspondiente al año 2016. Available at 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. Conseil Supérieur de L'audiovisuel. Rapport annuel 2016.

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

**Danelljan M, Häger G, Khan FS and Felsberg M** (2014) Accurate scale estimation for robust visual tracking. *British Machine Vision Conference (BMVC),* Nottingham, UK.

**Domínguez A, Agirre M, Flörez J, Lafuente A, Tamayo I and Zorrilla M** (2018) Deployment of a hybrid broadcast-internet multi-device service for a live TV programme. *IEEE Transactions on Broadcasting* **64**, 153–163. doi:10.1109/TBC.2017.2755403

**EasyTV Project** (n.d.) EasyTV project website. Available at https://easytvproject.eu/

**eMarketer** (2017) US simultaneous media users: eMarketer's estimates for 2017. Available at 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. Available at 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. Available at https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0636:FIN:en:PDF

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

**Fiaz M, Mahmood A and Jung SK** (2018) Tracking noisy targets: a review of recent object tracking approaches. *ArXiv Preprint ArXiv:1802.03098*.

**Gordon D, Farhadi A and Fox D** (2017) Re3: real-time recurrent regression networks for object tracking. *CoRR, abs/1705.0*. Available at http://arxiv.org/abs/1705.06368

**Güler RA, Neverova N and Kokkinos I** (2018) DensePose: dense human pose estimation in the wild. *CoRR, abs/1802.0*. Available at http://arxiv.org/abs/1802.00434

**Hassaballah M, Abdelmgeid AA and Alshazly HA** (2016) Image Feature Detectors and Descriptors. In Awad Ali Ismail and Hassaballah Mahmoud (eds), *Image Feature Detectors and Descriptors*. Springer International Publishing (Verlag), pp. 11–45.

**He K, Zhang X, Ren S and Sun J** (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 770–778.

**He K, Gkioxari G, Dollár P and Girshick RB** (2017) Mask {R-CNN}. *CoRR, abs/1703.0*. Available at http://arxiv.org/abs/1703.06870

**Held D, Thrun S and Savarese S** (2016) Learning to track at 100 {FPS} with deep regression networks. *CoRR, abs/1604.0*. Available at http://arxiv.org/abs/1604.01802

**Henriques JF, Caseiro R, Martins P and Batista J** (2014) High-speed tracking with kernelized correlation filters. *CoRR, abs/1404.7*. Available at http://arxiv.org/abs/1404.7584

**Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, 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. Available at http://www.imac-project.eu/

**Jain V and Learned-Miller E** (2010) *FDDB: a benchmark for face detection in unconstrained settings*.

**Le V, Brandt J, Lin Z, Bourdev L and Huang TS** (2012) Interactive facial feature localization. *European Conference on Computer Vision*, Florence, Italy, pp. 679–692.

**Lin T-Y, Goyal P, Girshick RB, He K and Dollár P** (2017) Focal loss for dense object detection. *CoRR, abs/1708.0*. Available at http://arxiv.org/abs/1708.02002

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

**Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C-Y and Berg AC** (2015) SSD: Single Shot MultiBox Detector. *CoRR, abs/1512.0*. Available at http://arxiv.org/abs/1512.02325.

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

**Lukezic A, Vojir T, Cehovin L, Matas J and Kristan M** (2016) Discriminative correlation filter with channel and spatial reliability. *CoRR, abs/1611.0*. Available at http://arxiv.org/abs/1611.08461

**Luo W, Xing J, Milan A, Zhang X, Liu W, Zhao X and 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**, 10–16. doi:10.1109/MCE.2013.2251760

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

**McNally J and Harrington B** (2017) How millennials and teens consume mobile video. *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*. New York, NY, USA: ACM, pp. 31–39. doi:10.1145/3077548.3077555.

**Messer K, Matas J, Kittler J, Luettin J and Maitre G** (1999) XM2VTSDB: The extended M2VTS database. *Second International Conference on Audio and Video-Based Biometric Person Authentication*, Washington, DC, USA, Vol. 964, pp. 965–966.

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

**NIELSEN a** (2017) The Nielsen comparable metrics report, Q1-2016. Available at 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. Available at 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. Available at 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. Available at 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. Available at 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. Available at 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 and He Z** (2017) Spatially supervised recurrent convolutional neural networks for visual object tracking. *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–4.

**Orero P, Martín CA and Zorrilla M** (2015) HBB4ALL: deployment of HbbTV services for all. *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, Baltimore, Maryland, USA, pp. 1–4, doi:10.1109/BMSB.2015.7177252.

**Padilla R, Filho C and Costa M** (2012) Evaluation of Haar cascade classifiers designed for face detection. *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering* **6**, 466–469

**Prosperity4All Project** (n.d.) Prosperity 4All project website. Available at http://www.prosperity4all.eu/

**Redmon J, Divvala SK, Girshick RB and Farhadi A** (2015) You only look once: unified, real-time object detection. *CoRR, abs/1506.0*. Available at http://arxiv.org/abs/1506.02640

**Ren S, He K, Girshick RB and Sun J** (2015) Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR, abs/1506.0*. Available at http://arxiv.org/abs/1506.01497.

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

**Sáez Trigueros D, Meng L and Hartnett M** (2018) Face recognition: from traditional to deep learning methods. *CoRR, abs/1811.00116*.

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

**Simonyan K and 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**, 62–67. doi:10.1109/MMUL.2011.71

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

**Vinayagamoorthy V, Allen P, Hammond M and Evans M** (2012) Researching the user experience for connected Tv: a case study. *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 589–604. doi:10.1145/2212776.2212832.

**Voulodimos A, Doulamis N, Doulamis A and Protopapadakis E** (2018) Deep learning for computer vision: a brief review. *Computational Intelligence and Neuroscience* **2018**, 7068349, 13 pages.

**Wang M and Deng W** (2018) Deep face recognition: a survey. *ArXiv Preprint ArXiv:1804.06655*.

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

**Woods RL and Satgunam P** (2011) Television, computer and portable display device use by people with central vision impairment. *Ophthalmic and Physiological Optics* **31**, 258–274

**World Health Organization and others** (2013) Universal eye health: a global action plan 2014-2019.

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

**Yuheng S and Hao Y** (2017) Image segmentation algorithms overview. *CoRR, abs/1707.0*. Available at http://arxiv.org/abs/1707.02051

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

**Zhang Zhifei SY and Qi H** (2017) Age progression/regression by conditional adversarial autoencoder. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawai, USA.

**Zhu X and Ramanan D** (2012) Face detection, pose estimation, and landmark localization in the wild. *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Provicende, Rhode Island, USA, 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. *2013 IEEE Third International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, pp. 1–5. doi:10.1109/ICCE-Berlin.2013.6697990.

**Silvia Uribe** received the degree in Telecommunication Engineering in 2008, the Master in Communications Technologies and Systems in 2010, the Master in Telecommunication Management in 2013, and finally, the PhD in 2016 by the Universidad Politécnica de Madrid (UPM). She is a member of the Visual Telecommunication Application Group since 2006. Her professional interests include interactivity technologies, content personalization technologies, and big data. She has been participating with technical responsibilities in some national (Buscamedia, Ciudad 2020, LPS-BIGGER, Repara 2.0) and European (eASLA, LASIE, EasyTV) projects, and she is author and co-author of several papers in international conferences and journals.

**Alberto Belmonte** received the degree in Telecommunication Engineering in 2014 and the Master's degree in 2016, focused on communication systems,

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press

from the Universidad Politécnica de Madrid (UPM). He worked in EVERIS Spain SL using the JAVA programming language to deploy web services. Currently, he is a member of the Visual Telecommunication Application Group, and he is PhD candidate. His main interests are the new communications technologies, IoT, sensors, wireless communications, computer vision, and image processing. He is working actively on detection and tracking algorithms for accurate localization and classification with cameras and sensors using both Machine and Deep Learning techniques.

**Francisco Moreno** received his BSc in Computer Science Engineering in 2008 on the Universidad Politécnica de Madrid. In 2007, he finished his final degree project called "Information Retrieval In Email Fields" on 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 the UK in 2012 where he worked 2 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** got the Bachelor of Engineering in Telecommunication Technologies and Services at Universidad Politécnica de Madrid in 2016. In 2018, he got the Master in Telecommunication Engineering at UPM, with his end of Master Project focused on the development of applications for a multi-device HbbTV 2.0.1 scenario with synchronous OTT services. Since 2015, he collaborates with the GATV in different national and international research projects and with the Chair of RTVE in UPM. His professional interests include the broadcast of digital television, accessibility services, interactive TV, and new video formats and audio-visual technologies.

**Juan Pedro López** is a Telecommunication Engineer since 2007, he got the International Doctor's Degree at Universidad Politécnica de Madrid in 2016 with his thesis focused on 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. 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 from UNED University in 2017.

**Federico Álvarez** is a Telecom Engineer (2003) and received PhD (2009), both by the Universidad Politécnica de Madrid. He is working as assistant professor lecturing in UPM, and he develops his research within the GATV. 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, and FI-PPP project XIFI, and coordinated the projects nextMEDIA, INFINITY, and FI-LINKS. He is author and co-author of (70+) papers in journals, congresses, and books.

https://doi.org/10.1017/S0890060419000350 Published online by Cambridge University Press