A USER-CENTRIC APPROACH TO FACIAL RECOGNITION FOR TV CONTENT

Alexandre Rouxel[1], Alberto Messina [2], Sébastien Ducret [3], Pierre Fouché [1]

[1] European Broadcasting Union, Switzerland, [2] Radiotelevisione Italiana, Italy, [3] Radio Télévision Suisse, Switzerland

ABSTRACT

The increasing relevance of facial recognition technology in the broadcasting industry raised the need for a consistent evaluation framework designed specifically for video content. Addressing this challenge, the European Broadcasting Union (EBU) has developed a benchmark and state-of-the-art AI models tailored for facial recognition in television programmes. This initiative involved the extensive annotation of a diverse video dataset guided by user-centric metrics that prioritise the accurate retrieval of relevant personalities, as defined by documentalists. Our machine learning models employ a unique approach that selectively identifies personalities active in the TV programme and deliberately excludes incidental characters, maximising user-centric metrics and enhancing the relevance and quality of the metadata. This strategy improves the overall performance of the facial recognition system while addressing privacy concerns by complying with General Data Protection Regulation (GDPR) and ensures ethical and responsible use of facial recognition technology in the media sector.

INTRODUCTION

The increasing applicability of facial recognition technology (FRT) in the broadcast and media industries necessitates a standardised evaluation framework specifically designed for video content. The absence of such a framework poses challenges in the decision-making process regarding the implementation of facial recognition systems, as reliance on conventional Machine Learning (ML) metrics may lead to suboptimal choices. In fact, these metrics prioritise performance optimization, which can inadvertently overlook user-centric properties essential for practical applications and result in masking critical user-centric properties, such as the relevance and accessibility of the metadata produced. To address this gap, the EBU has developed a benchmark tailored for facial recognition in television

programming, accompanied by a state-of-the-art AI model optimised for this framework. This initiative involved the extensive annotation of a video dataset guided by user-centric metrics, which prioritise the accurate retrieval of relevant personalities in accordance with the requirements of documentalists. This framework has been specifically designed for the management of the open set use case.

In the open set setting, a model must not only discriminate between the classes of the training set or, more generally, the set of classes to be identified, but also determine whether a class has not yet been encountered. Conversely, in the closed set setting, a model is tasked with recognising a fixed set of classes that remains consistent throughout both the training and testing phases. Both closed and open sets can involve a zero-shot learning approach, where the model recognises categories that it has not explicitly seen during training, but which are part of the set of classes to be identified.

To set the context, the concept of open-set identification is widely used in the media, particularly for TV shows and films. Here, the facial recognition system must determine whether a person in a given image or frame is known and part of an existing dataset of personalities to be recognised, or unknown, i.e., not recorded in this dataset. This is in contrast to closed-set identification, where the system assumes that all detected persons are from a pre-existing dataset of individuals.

Identifying individuals in an open set is particularly relevant to the media, as television programmes and films often introduce both recurring characters and new, unpredictable ones. Consequently, in practical applications, the facial recognition system must constantly adapt and decide whether faces represent new individuals that should be added to the database.

In summary, open-set identification in the context of facial recognition for media needs to accommodate the constant evolution of media content and ensure that systems are both effective in recognition and able to identify unknown individuals who should be recognized. To establish the terminology in the context of FRT, the reference database comprises two primary elements: the thesaurus and the gallery. The thesaurus contains the names or identifiers of individuals to be recognised, while the gallery consists of a collection of images for each individual.

In general, facial recognition algorithms compute one-to-many similarity scores to determine the specific identity of a person in a probe image. In the context of facial recognition, the individual in the probe image may or may not be previously known to the system. If the probe face does not meet the acceptance criterion, such as having a similarity score below a predefined threshold, the person is considered unknown. In such cases, subject to legal and ethical considerations, the system may be enriched with the new individual's information, which would be added to both the thesaurus and the gallery.

The classical ML performance metrics for FRT on open-set are detailed in 'Phillips et all (2)' and illustrated in Figure 1. The Detection and Identification Rate (DIR) is the proportion of people in the thesaurus that passed the acceptance criteria and are correctly identified. The False Alarm Rate (FAR) is the proportion of unknown persons identified as persons in the thesaurus. The Receiver Operating Characteristic (ROC), the DIR as a function of the FAR, gives a complete view of the performance. Classically, two single values provide a more synthetic view of performance, the Equal Error Rate (EER) and the Area Under the Curve (AUC). The AUC is the area that assesses performance across the full range of operating thresholds, a higher AUC corresponds to higher performance. The False Negative Identification Rate (FNIR) is the proportion of persons in the thesaurus that are not identified, i.e., rejected, by definition FNIR = 1-DIR. The EER is the point where FAR equals FNIR. A lower EER value indicates better performance as it reflects a lower probability of error in both FAR and FNIR at a given operating point.

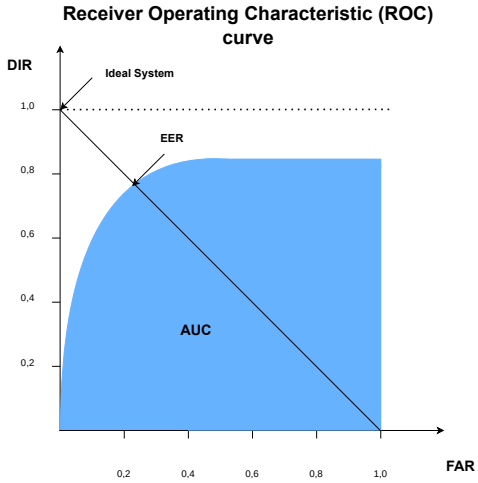

Figure 1 : AUC and EER

While these metrics provide valuable insights into absolute performance, they do not account for the practical utility of the tool. The user-centric metrics that we developed therefore, focus on the specific needs of documentalists. Furthermore, a comprehensive evaluation of facial recognition technology must consider a broader spectrum of factors, including computational complexity and adherence to legal standards such as the General Data Protection Regulation. Our approach simplifies the implementation by focusing on key operational and legal considerations, ensuring that the deployment of facial recognition technologies is both effective and compliant with privacy regulations.

USER CENTRIC METRICS

User-centric metrics are designed to meet the specific needs of documentalists who are responsible for managing annotations and efficiently identifying relevant content.

Role of the documentalists

Documentalists perform two critical functions: annotation management and relevant content retrieval. During the annotation process, documentalists are responsible for verifying and naming individuals who should be added to the thesaurus. This role requires ensuring that the system prioritises individuals who are actively contributing to the content of the programme over random bystanders or background characters, and who can be lawfully added to the database in light of the GDPR.

During the search phase, documentalists need access to precisely annotated intervals rather than continuous frame-by-frame data that can overwhelm the system with irrelevant information. This flood of data not only obscures the search process, but also increases the likelihood of errors, requiring further post-processing to eliminate false automatic annotations.

Performance metrics

To adapt the performance metrics to the specific needs of the documentalists for the annotation and gallery management, we introduce the Gallery Candidate Precision (GCP). GCP measures the precision with which a facial recognition system identifies and suggests candidates for inclusion in the reference database. GCP is calculated as the ratio of Valid Gallery Candidates (VGC) to the total number of candidates proposed for gallery inclusion. VGC refers to individuals correctly identified by the system as suitable for addition to the gallery due to their significant presence in the content. Whereas Invalid Gallery Candidates (IGC) refers to individuals the system inaccurately suggests for inclusion in the database. Typically, IGCC are bystanders or incidental characters whose presence in the gallery would not add value and could potentially clutter the database. The GCP can thus be expressed as: $\mathrm{GCP} = \frac{\mathrm{VGC}}{\mathrm{VGC+IGC}}$.

Optimizing the GCP significantly reduces the workload of documentalists in managing the gallery. In the following sections, we will explain how the algorithms are built to optimize this criterion.

Refining data annotation through intervals

During the search phase, documentalists need access to precisely annotated intervals rather than continuous frame-by-frame data that can overwhelm the system with irrelevant information. By adopting an interval-based annotation approach, the system is more closely aligned with the operational needs of users, ensuring that only meaningful data is highlighted and stored. This approach ensures optimal performance by emphasising the relevance of detections without compromising the efficiency of the system.

In conclusion, user-centric metrics in facial recognition for media applications must prioritise the specific operational needs of documentalists. Key metrics such as the GCP and the strategic use of interval-based annotations are critical to making these systems practical, effective and user-centric.

ANNOTATED DATASET

To evaluate the performance of FRT in the context of video content, an annotated dataset is essential. While there are numerous open-source image datasets, such as VGG Face 2 ‘Cao et all (5)’ and LFW ‘Huang et all (4)’, for evaluating the quality of facial recognition systems on images only, there is a lack of annotated video datasets specifically designed to evaluate the unique characteristics of FRT for television content. The EBU has addressed this gap by developing a dataset based relevant intervals rather than individual frames. The EBU annotated video dataset is composed of 335 videos, contributed by European Broadcasters, and encompasses over 78 hours of diverse television content from RTS, RAI, BBC, and France Television. The dataset covers a wide range of programmes, including news, TV shows, debates and music shows. The associated gallery features 700 distinct personalities, each assigned a unique identifier, which is the Wikidata ID when available.

Interval-based annotation approach

An interval in the EBU dataset is defined by two criteria: it must be less than 30 seconds in duration and contain a maximum of three distinct active individuals. This interval-based annotation approach ensures that the dataset is aligned with the operational needs of documentalists, emphasising the relevance of the recognized faces without compromising the efficiency of the system.

During the semi-automatic annotation process, features are generated for all detected faces. These features are then clustered based on their similarity scores. These scores are determined by factors such as detection score, resolution, blur and similarity. For each interval and each distinct group of personalities, the frames containing the faces with the highest scores are extracted and manually verified. In a second step, the personalities in the selected frames are manually annotated using the part of the gallery associated with the programme. This approach significantly reduces the cost of annotation compared to frame rate annotation, while maintaining the relevance of the data for practical applications.

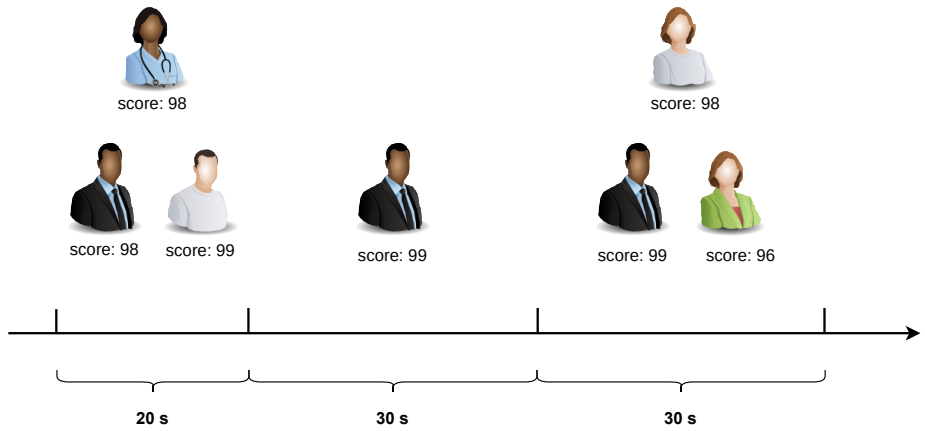

Figure 2: Example of annotation intervals

The EBU annotated video dataset is a comprehensive and interval-based resource specifically created to assess the performance of facial recognition systems in the context of television content. By focusing on user-centric metrics that meet the needs of documentalists, this dataset is valuable for benchmarking and optimising facial recognition models.

FACIAL RECOGNITION IN TV SHOW

In this section, we will explore the process of facial recognition technology for video and highlight the specificity of the algorithms developed by the EBU to optimise user-centric performance metrics. The classical processing bocks are detailed in ‘Du et all (1)’ and ‘Mendes et all (3)’, starting with frame extraction, covering the transformation of detected faces into high-dimensional vectors using deep neural networks, and the subsequent clustering and matching process to associate these faces with individuals in the gallery.

The specific component developed by the EBU, the face selection block, will be discussed in detail, highlighting its role in capitalising on a priori knowledge of the shooting style to slightly simplify the computational requirements and optimise the user-centric metrics.

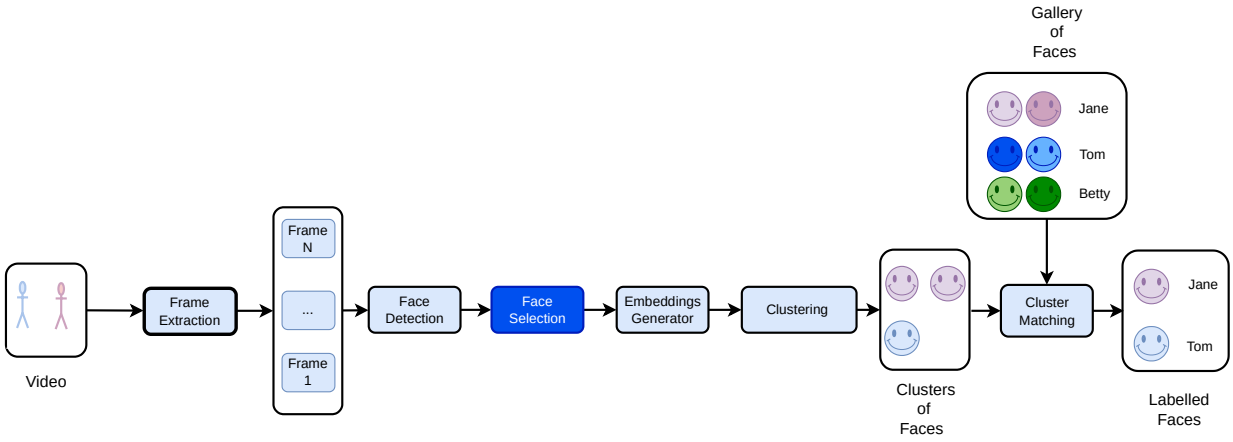

Figure 3: Facial recognition processing blocks for video

Figure 3 illustrates the process of facial recognition on video, after the frame extraction, the processing begins by detecting and cropping faces from video content. The face selection block, which will be explained in the next section, is the core block for exploiting the a priori knowledge of the shooting style. A deep neural network (DNN) is then employed to generate embeddings for each detected face. This DNN transforms raw images into high-dimensional vectors, known as embeddings. This is the block in the processing chain that requires the most computing power. The embeddings for the facial images in the gallery, which correspond to the individuals to be recognized, are precomputed. Once all the faces are represented by their respective embeddings, an unsupervised clustering algorithm is applied to group similar faces from the video. Subsequently, cluster matching is performed to identify the pre-labelled clusters from the gallery of faces that are closest to the unlabelled clusters obtained from the video content. This process enables the recognition of individuals appearing in the video by associating them with the corresponding gallery clusters.

Leveraging filming style for efficient face selection

In television programmes and films, the style of the cameraman has a significant impact on the essence of the production. The shooting style is designed to emphasise the importance of key characters or personalities, making them visually dominant on the screen. These key characters are the active participants in the programme who should be present in the metadata. The processing pipeline, shown in Figure 3, uses the face selection block to identify relevant faces based on several factors. It filters detected faces based on their area and distance from the bottom of the frame. First, faces with detection scores below a threshold are eliminated. The algorithm then calculates the area and distance of the remaining faces from the centre of the bottom edge of the frame. To determine the type of shot (close-up or far distance), the algorithm uses a parameter called "Area-to-Squared-Distance Ratio" (ASDR), which is the ratio of the area divided by the square of the distance to the centre of the bottom edge. Based on the dominant face with the maximum ASDR, the algorithm calculates a selection threshold. Finally, it selects faces with ASDR values above this threshold. Figure 4 illustrates this process in a TV show.

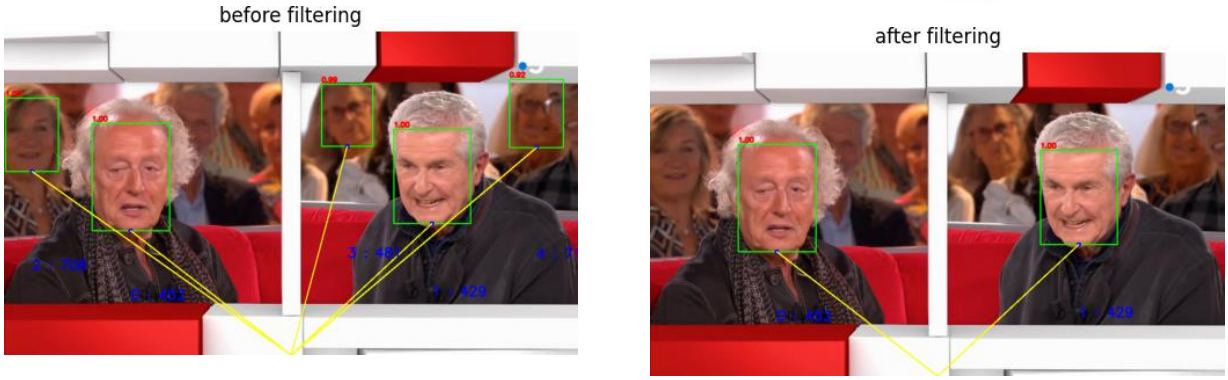

Figure 4: Face selection processing before and after filtering

Impact of the Face Selection algorithms on user-centric performance metrics

The face selection algorithm aims to optimize user-centric metrics, enhance identification performance, and reduce complexity. Figure 5 illustrates the effectiveness of the algorithm in a TV show where the public is often present in the background of the guests and hosts. With the face selection activated, no Invalid Gallery Candidates (IGC) are detected, compared to 11 IGC without the filter.

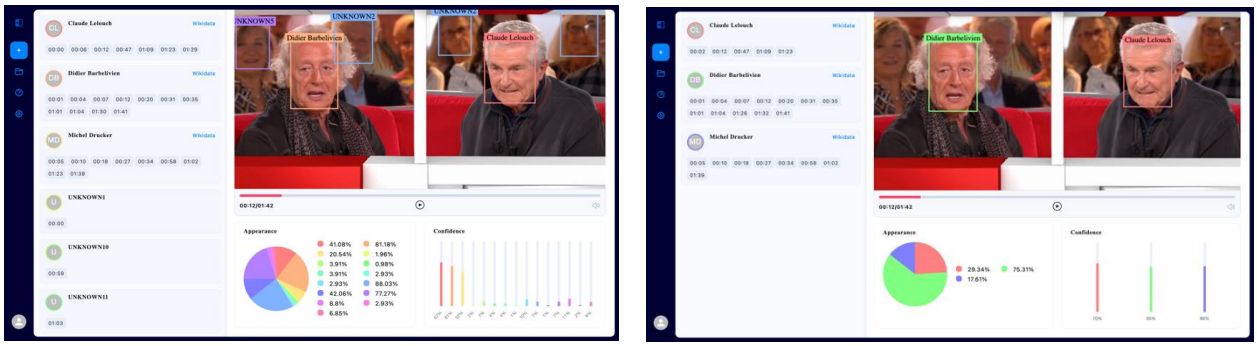

Figure 5: Effect of the face filtering in a TV show

In terms of identification performance, activating the filter significantly enhances cluster purity by removing IGC before clustering, ensuring that clusters contain only relevant data. This simplification results in more accurate cluster matching, thereby reducing the risk of identification errors. Through our observations, we have found that commercial solutions with their large galleries can lead to a higher probability that an IGC may resemble someone already present in the gallery. This increased likelihood of similarity can result in incorrect identification and attribution of names in the thesaurus to unknown individuals.

Figure 6 illustrates the number of elements per cluster before and after filtering for the TV show shown in Figure 5. After filtering, the clusters representing the three personalities of interest become clearly distinguishable. So, a straightforward post-processing step based on the number of faces per cluster enables the removal of residual noise, which is typically caused by incorrect face detection (on objects) or poor quality of cropped celebrity faces, especially in distant shots.

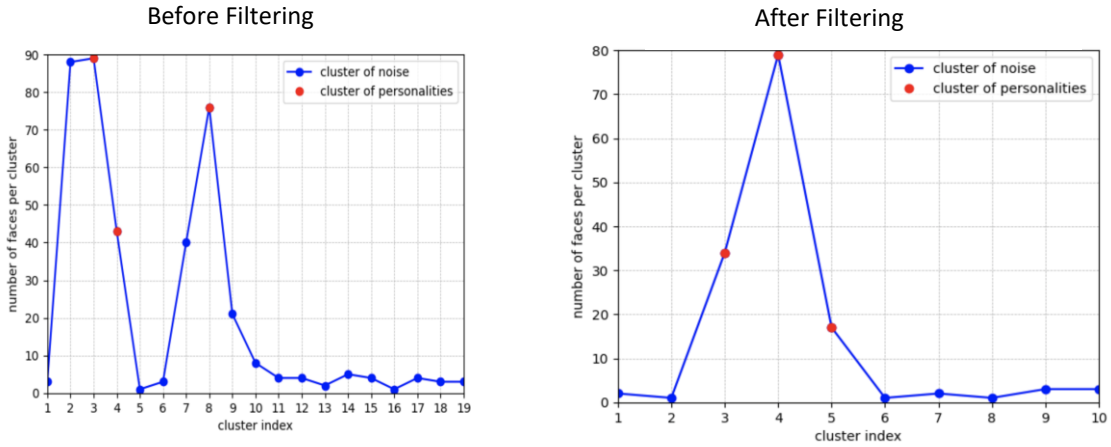

Figure 6: Effect of the face selection algorithm on the clusters.

In the context of user-centric performance metrics, the primary objective is to accurately identify key individuals within specific time intervals in a video. The system's filtering process enhances precision by minimizing the number of false positives (incorrectly identified individuals). Moreover, assessing the system at the interval level, rather than at the frame level, contributes to improving recall. As long as the system can accurately identify the active individuals within the designated time intervals, it is deemed to be functioning effectively according to user-centric performance metrics.

Facial recognition and GDPR compliance

The use of facial recognition technology on spectators or inactive participants at public events, such as television programmes, raises significant privacy concerns under the EU's General Data Protection Regulation (GDPR), 'Regulation (EU) (6)'. While the GDPR allows limited exemptions for journalistic purposes (Article 85), this cannot serve as a blanket exemption for indiscriminate biometric processing without balancing data protection rights and implementing appropriate safeguards. Furthermore, as biometric data processing for identification is considered high-risk, a Data Protection Impact Assessment (Article 35) is mandatory to evaluate and mitigate potential risks to individuals' rights and freedoms. However, developing robust algorithms to reliably detect and filter out the facial data of non-participants from any biometric identification processing can help uphold multiple GDPR principles. Considering that active participants in the TV programmes like guests and hosts give their consent for storing their biometric data or are covered by the journalistic exception, the EBU processing adheres to data minimization by preventing excessive collection (Article 5(1)(c)), provides a lawful basis by excluding non-consenting individuals (Articles 6, 9), aligns with purpose limitation by processing only data of consenting participants (Article 5(1)(b)), and constitutes data protection by design by integrating necessary safeguards into the processing itself (Article 25). This privacy-preserving approach enables the intended facial recognition functionality for TV content, while preventing GDPR violations related to

consent, lawful basis, and individual rights that would arise from indiscriminate biometric processing of audiences without justification. In the case of celebrities who are not actively participating in a TV show, the application of journalistic exemption should be restrictive and based on legitimate public interest. This approach prioritizes the protection of privacy over satisfying curiosity about a celebrity's private life.

CONCLUSION

We have developed a user-centric facial recognition approach tailored for television content, addressing the need for a consistent evaluation framework in the broadcasting industry. This initiative involved creating a benchmark, state-of-the-art AI models, and an extensively annotated video dataset guided by user-centric metrics that prioritize the accurate retrieval of relevant personalities. The introduction of user-centric metrics, such as the Gallery Candidate Precision (GCP), and the adoption of an interval-based annotation approach enhance the practicality and effectiveness of facial recognition systems in TV programs. Our machine learning models focus on selectively identifying active personalities and excluding incidental characters, maximizing these user-centric metrics and enhancing metadata quality. This strategy improves overall performance while addressing privacy concerns and ensuring compliance with GDPR regulations by processing only the facial data of active participants. In summary, the user-centric approach to facial recognition for TV content prioritizes the specific needs of documentalists, ensuring that the technology effectively serves its intended users while adhering to legal standards and privacy regulations.

REFERENCES

1- Du, H., Shi, H., Zeng, D., Zhang, X.P. and Mei, T., 2022. The elements of end-to-end deep face recognition: A survey of recent advances. ACM Computing Surveys. pp.1-42.

2- Phillips, P.J., Grother, P. and Micheals, R., 2011. Evaluation methods in face recognition. Handbook of face recognition, pp.551-574.

3- Mendes, P.R.C., Busson, A.J.G., Colcher, S., Schwabe, D., Guedes, Á.L.V. and Laufer, C., 2020, November. A Cluster-Matching-Based Method for Video Face Recognition. In Proceedings of the Brazilian Symposium on Multimedia and the Web. pp. 97-104.

4- Huang, G.B., Mattar, M., Berg, T. and Learned-Miller, E., 2008, October. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition.

5- Cao, Q., Shen, L., Xie, W., Parkhi, O.M. and Zisserman, A., 2018, May. Vggface2: A dataset for recognising faces across pose and age, 2018 In 13th IEEE international conference on automatic face & gesture recognition. pp. 67-74

6- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

AKNOWLEDGEMENTS

The authors express their gratitude to the EBU Members for their valuable participation in the EBU AI-Benchmarking working group, where the core of this study was performed.