Who's in My Archive? An End-to-End Framework for Automatic Annotation of TV Personalities

Maurizio Montagnuolo
Centre for Research, Technological Innovation and Experimentation
Rai Radiotelevisione Italiana
Turin, Italy
Email maurizio.montagnuolo@rai.it

Fulvio Negro
Centre for Research, Technological Innovation and Experimentation
Rai Radiotelevisione Italiana
Turin, Italy
Email fulvio.negro@rai.it

Alberto Messina
Centre for Research, Technological Innovation and Experimentation
Rai Radiotelevisione Italiana
Turin, Italy
Email alberto.messina@rai.it

Angelo Bruccoleri
Centre for Research, Technological Innovation and Experimentation
Rai Radiotelevisione Italiana
Turin, Italy
Email angelo.bruccoleri@rai.it

Roberto Iacoviello
Centre for Research, Technological Innovation and Experimentation
Rai Radiotelevisione Italiana
Turin, Italy
Email roberto.iacoviello@rai.it

***Abstract*—Knowledge about the presence of people in a video is a valuable source of information in many applications, such as video annotation, retrieval and summarisation. The contribution of this paper goes in the direction of demonstrating how AI-based face processing technologies can be profitably used to perform video annotation of television content. To validate our vision, we developed the Face Management Framework (FMF), which implements an end-to-end pipeline for face analysis and content annotation based on few-shot or zero-shot face embedding extraction models. The results of the test campaign of the system show that the key performance indicators that we defined were exceeded by a wide margin, demonstrating how media workflows could greatly benefit from the tool and the efficiency improvements it brings.**

***Index Terms*—metadata, face recognition, clustering, knowledge management applications, media archive**

I. Introduction

Content preservation, high-quality production and process automation are at the core of the current transformation of Public Service Media (PSM) from its traditional business to the modern digital era. Nevertheless, accessing, retrieving and consuming desired content (and metadata) can be a challenging and nontrivial task. This is where artificial intelligence (AI) comes into play, providing solutions helping users to extract knowledge and organise data more efficiently and effectively. Emerging AI-based technologies can assist PSM in this transition by providing capabilities that facilitate the organisation and exploitation of retained assets. Sample tasks are those aimed at AI-based metadata extraction (e.g., recognition of TV celebrities or geographical landmarks in broadcasts or archival media assets) and content enhancement (e.g., video super resolution, video denoising). The broadcaster's archives have a leading role to play in these processes, being an invaluable source of information to run the business. On the one hand, archive content could be used to train and fine tune AI systems, thus helping to overcome their current limitations. On the other hand, these AI systems could be used for the enrichment, refinement, and improvement of the execution of work processes.

This paper describes a system for automated annotation of TV personalities in video streams and archives. The system addresses the open set face identification problem [1], in which a hybrid approach making use of archival annotations, archive content, and external knowledge bases are employed to build the galleries of reference personalities. Reference galleries can be created, edited, and updated in an opportunistically (e.g., based on production or archival metadata), or systematically (e.g., from an a priori list of celebrities), in order to ensure the flexibility and adaptability of the identification process. Annotating the TV personalities in the videos is done by a deep learning pipeline using state-of-the-art algorithms for face detection, representation, grouping and labelling. Face detection is used to find the regions in the video depicting human faces. A descriptor is extracted by each detected face. These descriptors, also called embeddings, provide compact representations that retain the unique characteristics of human faces. They can therefore be used to distinguish between different people in the video, or estimate other characteristics like the biological gender (i.e., sex) [2], age or emotions. Face clustering is then applied to group faces into clusters of different persons. Finally, a face identification process is performed to name the clustered faces using the galleries of reference identities.

This work is supported by European Union's Horizon 2020 research and innovation programme under grant agreement number 951911 - AI4Media (https://www.ai4media.eu/).

979-8-3503-2445-7/23/$31.00 ©2023 IEEE

The "dynamic" nature of the reference galleries makes it possible to mitigate the weakness of existing facial recognition systems. Although many approaches to unconstrained face recognition in video have been presented [3], they suffer from some inherent limitations that affect their usability in a real operational environment. First, nearly all of them are trained on datasets that are built using only very popular celebrities [4]–[8]. However, it is often desirable to be able to deal with less familiar entities such as minor league players, supporting performers or emerging personalities. This is especially true for archival material that is highly heterogeneous in regards to age, quality, and other conditions. Next, they have been designed in terms of either breadth (i.e., many people but few images for each person) or depth (i.e., few people but many images for each person) of data, when they should instead combine both. Moreover, they are prone to technological (e.g., camera settings or lighting conditions) or demographic (e.g., ethnicity, biological gender, age) biases that negatively impact the ability of AI models to generalise across datasets [9], [10]. Finally, existing datasets are affected by incorrect annotations, which dramatically increase along the dataset size [11], [12].

The remain of the paper is organised as follows. Section II presents the state of the art and related work. Section III describes the system we developed for face analysis and content annotation. Section IV provides the experiments that were performed to validate the system performance in a real world application scenario. Section V concludes the paper with final remarks and sketches some future directions.

II. Related Work

This section describes the related work and is organised in two main parts. The former examines how media experts perceive the challenges and possibilities in implementing AI technologies in their respective organisations. The latter overviews existing literature on the topic of face analysis.

A. AI in the Media Industry

Media professionals strongly believe in AI technology and its application in the media domain, although it is still not clear how far we are from fully operational and high quality functionalities. Despite a considerable effort to integrate AI functionalities at different levels of maturity [13]–[15], numerous challenges still remain unsolved, in particular those related to performance and ethics [16], [17]. These are the main findings of a survey we conducted among experts working in the media industry from different European countries [18]. The interviewees, comprising managers, archivists, editors, journalists, and technicians, were requested to identify the primary issues and difficulties they face in their daily tasks. Additionally, they were asked to share their perspectives on the degree of applicability, maturity, usefulness, and trustworthiness of AI in those activities. Among the replies and comments received in the survey, the most important results can be summarised as follows:

- (Processes and workflows) There is a lack of metadata integration and media information along the value chain. There is a need for a common terminology to define the problems. It is difficult to define clear and measurable business indicators.
- (Resources and budget) It is hard to implement the transition to next-generation systems due to legacy issues, shortage of human resources and need for rapid return of investments. Data overload and time constraints pose challenges that need to be addressed.
- (Awareness and fears) There is a lack of understanding of the potential of AI. The complexity of the pipelines, the software development and the deep network optimisation can be an obstacle. Open-source solutions must be carefully assessed and monitored for stability, usability, and reliability.
- (Impact and importance) AI-driven tool integration would primarily increase process efficiency (i.e., less time, better quality) rather than reduce or optimise costs.
- (Trustworthiness and credibility) Trustworthy AI capabilities are a crucial factor influencing the widespread integration of AI in the media industry, particularly in respect to safeguarding privacy and ensuring legal compliance.

Archives are the domain where expectations are higher and difficulties deeper [19]. In particular, processes where AI technologies could significantly improve efficiency relate to content search and retrieval, quality verification, and metadata annotation [20]–[22]. Since without a good user experience, even the best-performing tool won't be effective, methodologies to browse, inspect and analyse the results of AI-based automated content analysis have been proposed [23], [24].

B. Face Analysis

Face analysis deals with the detection, representation, clustering and labelling of facial images. Labelling includes subprocesses such as recognising the face identity, or estimating face attributes like head pose, expression, age and biological gender.

Basic approaches on face detection focused on the localisation of frontal human faces [25]. More challenging approaches focused on the problem of rotated multi-view face detection, i.e., faces under different poses, orientations and lighting conditions [26], [27]. In recent years, face detection have been dominated by deep learning based methods [28]–[31]. Among them, the RetinaFace [32], and the SCRFD [33] have shown the most promising performance, with average precision of over 90% on the WIDER FACE benchmark dataset [34].

Face clustering adopts algorithms in which faces are represented as numerical vectors in a high dimensional space. Any similarity function defined in this space can be used for grouping faces sharing similar features [35]. In the Local Binary Pattern method [36] face images are split into small blocks of pixels, and from each block shape and texture histograms are extracted. Clustering is done using the chi-square measure and a nearest neighbour classifier. In order to ensure the invariance of illumination and geometric deformation the Scale Invariant Feature Transform (SIFT) and the Speeded-Up Robust Features (SURF) methods can be used

[37], [38]. The aim of these methods is to extract a set of key-points that represent the key features of the depicted faces. The grouping of similar faces can be then achieved by means of a hierarchical clustering technique [39], [40]. As for the detection task, modern approaches for face representation are based on deep network architectures [41]–[45]. Among them, ArcFace outperformed the state of the art methods by introducing an additive angular margin loss to improve the discriminative ability of the extracted face embeddings [43].

The objective of face identification (also called face recognition) is to correctly identify probe faces that are present in a gallery of reference faces, while rejecting probe faces that do not belong to the gallery. To this end, deep neural network models have been shown to be highly effective, achieving over 98% accuracy [43], [46]–[49]. Face embeddings extracted from deep face recognition networks can also be used to estimate face attributes, of which biological gender is a very important one [50]–[57]. In fact, the guarantee of gender equality in the media is one of the main pillars of public service broadcasting. The analysis and reporting of how women and men participate in radio and TV programmes is becoming increasingly important. For this purpose, national and international Government policies have been put in place. AI-based techniques can be used to speed up the process of monitoring the streams that are being broadcast [58], [59].

III. The Face Management Framework

Person annotation is the process of analysing TV streams or content from the archives to locate, group and identify individuals of potential interest who may be present in the scene at any time. This has extreme value in a broadcast production environment, where the need to create ever new entertainment or in-depth programmes makes it essential to perform archival research and filtering on specific well-known personalities. As an example, given a video clip acquired from DTT (Digital Terrestrial Television) broadcast or retrieved from the archives, the editor would like to access the exact time points of the clip where a certain person appears.

The Face Management Framework (FMF) implements an end-to-end pipeline for face verification and content annotation based on few-shot or zero-shot face embedding extraction models [60]. This workflow allows archivists and editorial staff to analyse TV streams or content from the archives in order to detect and identify persons of interest who are present on the scene from time to time. Additional metadata can be retained, including details such as the date, time, and channel of the appearance of individuals.

The workflow of the FMF is a process made of three components:

1) The gallery manager to create, update and manage the galleries of known identities and the related metadata.
2) The TV personality recogniser to detect, group and label unannotated video streams using the built reference galleries.
3) The stream monitor to check the conformity of broadcasts to PSM values, including, for example, diversity analysis, and content reliability assessment.

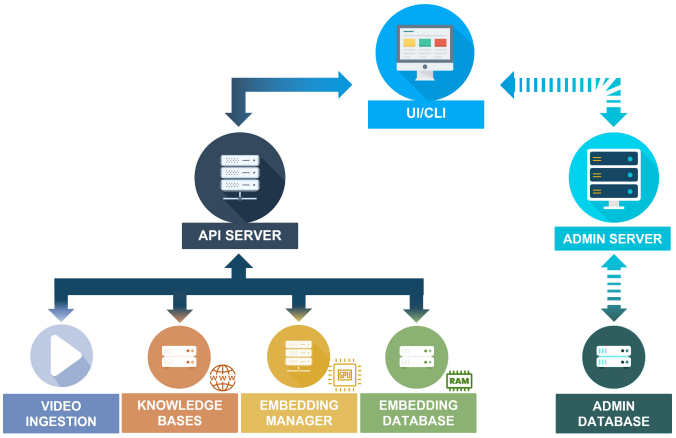

Fig. 1. Building blocks of the Face Management Framework.

The building blocks of the framework are shown in Fig. 1. The system is accessible via a graphical Web interface, command line interface and REST API, enabling both interactive and programmatic interaction with it. The API server receives the requests from the user interface and interacts with the following components:

- The video ingestion service, which imports images and video into the system, and performs shot detection and selection of keyframes. Content can be loaded from both local and network file folders, retrieved from open Web repositories, or manually uploaded from the user's laptop.
- The knowledge base manager, which interacts with online knowledge repositories to enrich metadata about a certain identity. Imported metadata are persisted in the system using a custom data model, ensuring effective long-term storage.
- The embedding manager, which detects, extracts and processes the facial features from the selected keyframes. Face detection is based on the RetinaFace network [32]. ArcFace embeddings are used to represent detected faces. The facial features extracted through ArcFace also show notable differentiation between the biological gender classes [53]. The embedding manager makes use of a multilayer perceptron for this purpose.
- The embedding database, which stores the extracted ArcFace embeddings and the reference with the corresponding gallery identities (and metadata). It is based on the OpenSearch engine,[1] which provides a highly scalable system for efficient vector-based search and retrieval.

Furthermore, the architecture includes an administration module to manage user authorisation and authentication. In the following subsections, details about the core components of the gallery manager and of the video annotator are given.

[1]https://opensearch.org/ (last accessed oct 2023).

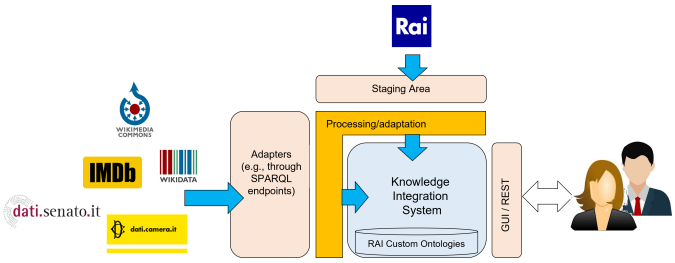

Fig. 2. High level architecture of the face gallery management system.

A. Gallery Manager

Fig. 2 shows the high-level architecture of the system we developed to manage the gallery of known identities and the related metadata. The system is highly flexible in terms of the data inputs it supports; it can process content from Rai's archives, public Web resources and various knowledge repositories, ensuring the long-term validity of the technical solutions developed. The main features are as follows:

- Import of metadata about a *person* from the selected knowledge bases. A person is a media personality who is notable for being either internationally (e.g., a movie star), nationally (e.g., a TV programme presenter) or locally (e.g., a representative of a municipal body or institution) acknowledged. Persons are identified by name or, if they are homonyms, by additional distinguishing information retained by the knowledge bases.
- Import of metadata about a *person group* from the selected knowledge bases. A person group is a collection of persons sharing some interests, memberships, or other criteria defined by the user (e.g., the members of a political party).
- Addition of *custom persons*, *custom person groups*, and *thematic collections* to create identities, groups, and communities not originally included in the imported knowledge bases. This allows users to extend and personalise the gallery of reference identities according to their specific need, such as the inclusion of the cast of a reality show before the first airing of the programme.
- Automatic retrieval, filtering, and association of faces to persons. A *face* is a single visual representation of a person. A face is referenced by means of the image file showing it, the rectangle within the image in which it is positioned (i.e., bounding box), and a numerical vector representing it in mathematical form (i.e., the ArcFace embedding).

The backbone of the framework is the Knowledge Integration System, which retrieves metadata from open knowledge repositories via the corresponding API, structures them in a personalised ontology, and stores them in a Neo4j graph database[2] for optimal accessibility and retrieval. Several knowledge bases and open data repositories have been considered for integration within the system, among which Wikidata, and those published by the Italian Chamber of Deputies and the Italian Senate.[3,4,5]

The custom ontology models in a unified way the information and relations between all the entities stored in the system. As shown in Fig. 3, the core entity of the ontology is the `HUMAN` class that describes an individual and includes details about that person like id, birthdate, deathdate, etc. Humans relate to other humans, such as whether they are married (i.e., `SPOUSE` relation) or have sons and daughters (i.e., `CHILDREN` relation). A human participates in one or more activities (i.e., `OCCUPATION` class), which provide information about their role, job or responsibility. Examples for occupations are a musician, a composer, or a politician. Humans also have one or more nationalities (i.e., `COUNTRY` class), and one or more pseudonyms (i.e., `ALIAS` class). Information about person groups is described by the `GROUP` class that includes details like the the date of formation and the date of dissolution, if any. Humans are linked to person groups by a membership relation. On top, there are the communities (i.e., `THEMATIC_COLLECTION` class), which may include groups as well as individuals. An example of the GUI for the creation of the thematic collection named "Italian jazz musicians" is shown in Fig. 4. The `FACE` class stores information about the facial characteristics of an individual, including the numeric representation and coordinate points of the rectangle surrounding their face. Finally, the `IMAGE` class references the source from where faces have been detected.

Once individuals have been imported or created, faces can be linked to them. This is a clustering process that runs fully automated or with various levels of human supervision. The core of the procedure is the DBSCAN algorithm [61], which is applied to the ArcFace embeddings to group similar faces. Compared to other clustering algorithms, DBSCAN is fast, and is able to detect clusters of arbitrary shapes and sizes while excluding outlying data points, without the need of defining in advance the number of desired clusters [35]. The variety of input data sources (i.e., images from the Internet, images from a local repository, or videos from the Rai archive), and the properties of DBSCAN, allows users to build smart, meaningful face galleries, settings only few parameters, such as the input sources and the way output clusters are pre-selected (i.e., all the created face clusters or the largest face cluster) through the dedicated GUI or programmatic API. This method offers a high degree of flexibility and versatility for different situations, such as face galleries where there are homonyms or more than one individual (e.g., all the members of band). Fig. 5 shows an example of a person and related metadata imported from Wikidata.[6] The faces associated to the person have been extracted from a collection of videos taken from the Rai archives. This allows galleries to be built with of range of different characteristics, including face size, head pose, age, and make-up.

[2]https://neo4j.com/ (last accessed oct 2023).

[3]https://www.wikidata.org/ (last accessed oct 2023).
[4]https://data.camera.it/data/en/datasets/ (last accessed oct 2023).
[5]https://dati.senato.it/sito/home (last accessed oct 2023).
[6]https://www.wikidata.org/wiki/Q128297 (last accessed oct 2023)

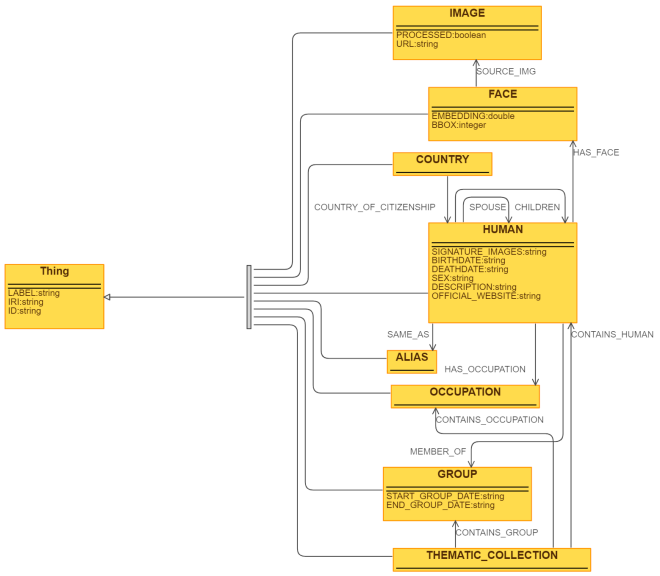

Fig. 3. Ontology of the Face Management System.

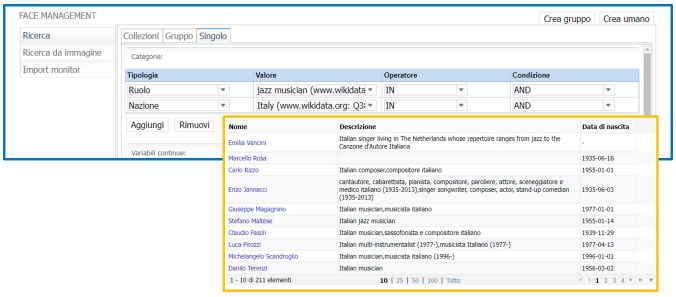

Fig. 4. Example of the creation of the thematic collection named "Italian jazz musicians". The user selects the person's role (i.e., jazz musician), and nationality (i.e., Italy). The system queries the underlying knowledge graph database to get the list of matching persons and returns it back to the user. For each person, some core metadata such as name, description and birthdate, are also shown.

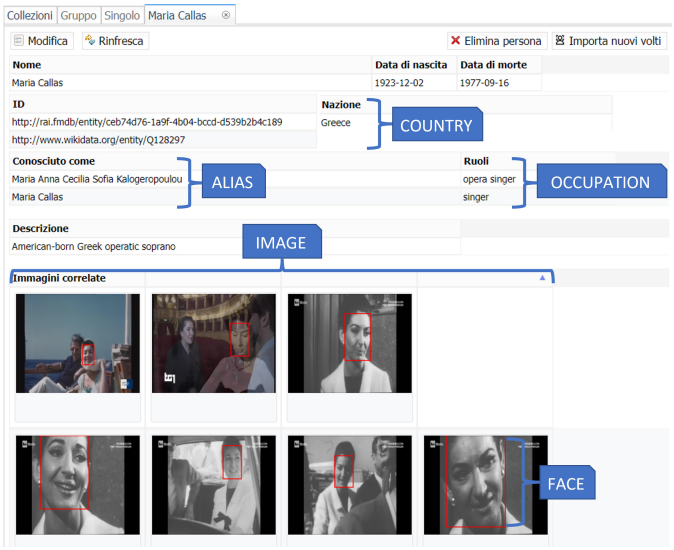

Fig. 5. Example of the metadata and faces associated to a person in the gallery of known identities.

B. TV Personality Recogniser

As stated earlier, the TV personality recogniser workflow is built on a deep learning pipeline, whose tasks include face finding, representation, grouping and labelling. The input comprises a collection of keyframes taken from the analysed video. Various strategies are available. These include a fixed subsampling rate, e.g., one frame per second, or a shot detection and keyframe selection procedure. Each keyframe is processed to detect faces and extract facial features using the RetinaFace [32] and the ArcFace networks [43], respectively. After that, a graph is built where the nodes are the ArcFace embeddings, and the edges are the Cosine similarity between them. The Chinese Whispers graph clustering algorithm [62] is then used to group faces of the same person based on their similarity. The Chinese Whispers algorithm was chosen over DBSCAN due to its superior performance when dealing with clusters of similar densities, and its faster data processing capabilities. In fact, while it is acceptable to assume the existence of one (or very few) predominant cluster in the gallery creation process (i.e. the one corresponding to the target individual), this may not be true when processing a generic video featuring multiple individuals with varying frequencies of appearance. Qualitative evaluations show that this clustering approach is very promising, being able to group face images of the same person over different conditions [63]. Next, face labelling applies a retrieval-based open-set face identification strategy to assign each cluster the identity of the corresponding person. This is implemented through the Hierarchical Navigable Small World (HNSW) library [64], an efficient algorithm to perform approximate K-Nearest Neighbor (KNN) search. The matches whose score is above a minimum threshold are labelled with the corresponding identity, and each cluster is assigned to the identity label whose occurrence is maximum among those included in the cluster. Finally, information about the annotated identities, e.g., channel, date, time, face bounding box, identity, is saved in a metadata repository.

Extracted annotations offer valuable insights for different users, including documentalists, journalists, program editors, and data analysts. Here are some notable examples:

- Annotated video browsing: exploring annotated videos through keyframes visualisation where bounding boxes underline the position of the referenced personalities.
- Person-based search: searching the archive for videos containing a given identity, with smart data integration (e.g., biographies) coming from reliable knowledge bases.
- Video insights analysis: visualising statistics and charts (e.g., audience score, programme schedule and identities) for a video stream by broadcast metadata and TV personality appearance.

An example of the output of the video annotation process is illustrated in Fig. 6. The interface displays the list of identities that are synchronised with the timeline of the video. The

Fig. 6. Example of the output of the TV personality recogniser. The video timeline is synchronised and annotated with the names of the identified TV personalities.

TABLE I
NEURAL NETWORK ARCHITECTURE FOR BIOLOGICAL GENDER ESTIMATION

Block	Description	#Params
Input	ArcFace → Normalize	-
Backbone	FC → ReLU → BN → DR → FC → ReLU	526,336
Neck (1)	BN → DR → FC → ReLU	132,352
Neck (2)	BN → DR → FC → ReLU (x2)	132,608
Neck (3)	BN → DR → FC → ReLU	33,408
Neck (4)	BN → DR → FC → ReLU (x2)	33,536
Neck (5)	BN → DR → FC → ReLU	8,512
Neck (6)	BN → DR → FC → ReLU (x2)	8,576
Neck (7)	BN → DR → FC → ReLU	2,208
Head	FC → SM	66
	Total Parameters:	**877,602**

combined use of clustering and labelling makes it possible to maximise the temporal extension of the identifications (i.e., strengthening the overall recall of the process) while still maintaining a high level of identification accuracy. This improvement is significant even in cases where there are variations in the age of the individuals or image quality. In addition, the system presents data on unnamed individuals. These are individuals whose faces have been grouped by clustering, but whose identities have not been determined (e.g., because they are missing from the reference galleries). In these scenarios, the user can manually assign an identity and modify the reference gallery. This is a circular process of learning, wherein the users' annotations serve as sources for the reference galleries and vice versa.

C. Stream Monitor

The stream monitor module makes use of facial-related features to examine the presence of women and men in TV programmes. This is a classification problem, where machine learning algorithms may be employed to estimate the biological gender on the basis of the respective face embeddings extracted for each face.

The choice of the most appropriate method is crucial, as it has a direct impact on the overall performance and efficacy [51], [52]. Classical statistical pattern recognition classifiers, such as decision trees and support vector machines, are easy to implement and understand. However, their efficiency depends on the class separability of the input data. In contrast, neural networks are powerful in realising complex nonlinear problems. In addition, they do not require any a priori assumptions about the characteristics of the input data, are robust to noise and provides fast evaluation of unknown data. Based on [54], the stream monitor uses a multilayer perceptron neural network, with the ArcFace normed embeddings serving as the input. The output is a vector including the predicted biological gender class and the uncertainty of the neural network output.

Table I shows the network architecture that counts 877,602 trainable parameters. The backbone of the architecture is a fully connected layer (FC), followed by a rectifier layer (ReLU), and coupled with a cascade combination of batch normalisation (BN), dropout (DR), FC and ReLU layers. The use of a batch normalisation layer followed by a dropdown layer has been shown to improve training efficiency of a neural network [65]. Then, there are seven blocks, which again consist of a sequence of layers BN → DR → FC → ReLU, repeated twice for even blocks and once for odd blocks. The architecture ends with a fully connected layer coupled with softmax activation function (SM). Our implementation differs from [54] in the input normalisation and the different hyperparameters, specifically batch size, learning rate, weight decay and learning rate scheduling.

Fig. 7 shows the correlation between the output of each neural network layer and the biological gender classes, mapped to a 2D feature space using the t-SNE algorithm for dimension reduction [66]. Each point on the graphs represents an individual, whose face was detected within some keyframes taken from Rai's TV channels and manually labelled as female (blue markers) or male (orange markers). The top left graph plots the input ArcFace normalised embeddings. Neighbouring points represent with good approximation faces belonging to the same individual. The others graphs show (top to bottom, left to right) the output of the backbone layers (block 0, actual size of the embeddings 512), and the outputs of the neck layers (blocks 1 to 7, actual size of the embeddings 256, 256, 128, 128, 64, 64, 32, respectively). It is interesting to note that the network does a satisfactory job of distinguishing between the two classes from the intermediate blocks.

IV. EXPERIMENTAL RESULTS

This section describes the experiments performed to validate the effectiveness of the system, including both objective evaluations based on standard information retrieval and data mining measures, and subjective user studies addressed to establish whether and to what extent the introduction of the FMF pipeline would improve the workflows of media companies.

A. Face Clustering and Identification

As previously described, we used the Chinese Whispers to group faces of the same individual, as this algorithm objectively grouped faces into coherent clusters better than other commonly used algorithms, such as DBSCAN, HDBSCAN and AHC (Agglomerative Hierarchial Clustering). For the evaluation, we used a dataset made of 1,099 keyframes extracted from a video clip depicting 30 international and

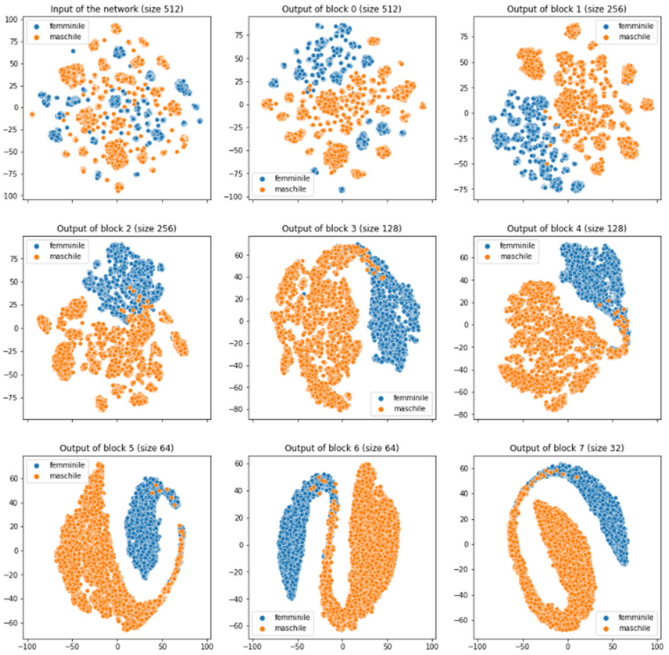

Fig. 7. Correlation between neural network layers' output and biological gender categories.

Fig. 8. Examples of the output of the face clustering.

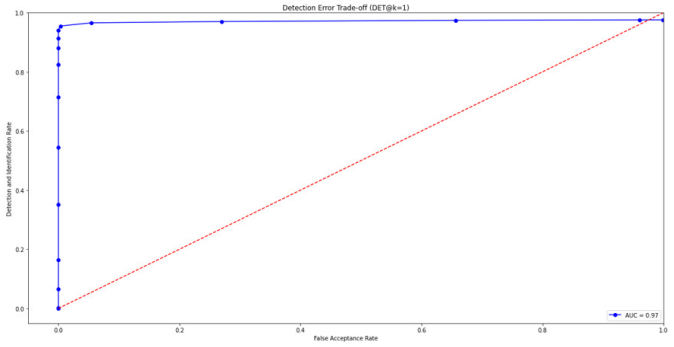

Fig. 9. DIR vs. FAR curve describing the trade-off for rank one identification and false alarms for the face labelling task. The best balance between DIR and FAR is got for Cosine similarity in the range $(0.4, 0.5)$.

national TV personalities from three Rai's TV channels. The performance was measured by computing the cluster completeness, the cluster homogeneity and the V-measure of the generated clusters [67]. Completeness measures whether all faces of an individual are grouped in the same cluster. Homogeneity measures whether a cluster contains only the faces of the same individual without impostors. They score between 0.0 and 1.0, where 1.0 stands for perfectly complete / homogeneous clustering. Completeness, homogeneity and V-Measure have similar meaning with that of recall, precision and F-Measure in information retrieval. On average, we got the completeness of 0.92, the homogeneity of 0.99, and the V-measure of 0.95. This confirms the goodness of the approach and the good ability to group faces of the same individual, even when they are shown with different age, pose and size (see Fig. 8 for same examples).

The capability of labelling the face clusters with the corresponding identities was tested using a gallery of 66 Rai newsreaders, and a probe set of about 10,000 faces detected in Rai's newscasts. The Cosine similarity was set as the distance metric. The performance was measured computing the Detection and Identification Rate (DIR) versus the False Alarm Rate (FAR) [68] for the rank K equal to one and Cosine similarity varying from zero to one. Fig. 9 draws the achieved chart. The red dotted line represents a system that is no better than random guessing. The solid blue line represents the measured values. The AUC (Area Under the Curve) score is 0.97, denoting excellent performance.

B. Biological Gender Estimation

In accordance with [54], we adopted a 2-step approach to build the neural network model. First, we trained the model using the IMDB-WIKI dataset [69], one of the largest datasets of face images from IMDB and Wikipedia with age and gender labels. Then, we fine tuned the model using the Adience dataset [70], and a 5-fold strategy for refinement and validation. Both datasets were preprocessed to filter out uninformative samples, i.e., images containing zero or more than one face, and images whose ground truth was absent. Differently from [54], we did not make any assumption on the face quality, and retained all the images found by the detection procedure. We believe this better reflects the application context in which we operate, which is characterised by great variability of the processed content, such as image quality (e.g., the oldest digitised material, sampling artefacts)

TABLE II
PERFORMANCE OF THE FINE-TUNING NEURAL NETWORK ARCHITECTURE FOR BIOLOGICAL GENDER ESTIMATION

	Mean Loss (σ)	**Mean Accuracy (σ)**
Fine-tuning	0.035 (0.002)	0.99 (0.0007)
Validation	0.068 (0.014)	0.98 (0.004)
Test	0.219 (0.026)	0.92 (0.014)

TABLE III
COMPARISON OF DIFFERENT BIOLOGICAL GENDER ESTIMATION TOOLS APPLIED TO TV STREAMS

Library	**Precision**	**Recall**	**F-score**
InsightFace [71]	0.962	0.962	0.962
FaceLib [72]	0.924	0.924	0.923
DeepFace [73]	0.942	0.939	0.938
Commercial	0.983	0.983	0.983
This	0.979	0.979	0.979

and size (e.g., long shots, very long shots). In total, we used $410,910$ images from IMDB-WIKI (78% of the original, $165,404$ females, $245,506$ males) and $13,099$ images from Adience (67% of the original, $7,116$ females, $5,983$ males). Table II shows the mean and standard deviation (σ) of loss and accuracy for the fine-tuning, validation and test on Adience.

In order to strengthen the experimentation in a real life application scenario, we collected a set of $7,345$ keyframes extracted from Rai's programmes of various TV genre, including in-depth journalism, talk show and entertainment. Each keyframe contains only one face that was manually annotated as either female ($2,530$ in total) or male ($4,815$ in total). Table III shows the results in terms of weighted precision, recall and F-score of our architecture compared to several libraries for gender recognition freely or commercially available on the market. We calculated the weighted form in order to take into account the imbalance between the two classes in the test set. This can result in an F-score that is not between precision and recall. The commercial system achieves the best results. Among the free tools, our architecture goes beyond the state of the art, exhibiting minimal differences from commercial alternatives.

C. Subjective Evaluations

The aim of the subjective evaluation activity is to assess the impact and advantages brought by the introduction of the person annotation pipeline into media companies' workflows, along with getting opinions from professionals about its usability and desirable improvements. For this purpose, we set up a series of user trials with representatives from different Rai departments (i.e., archives, news, editorial, marketing) and with different skills (e.g., documentalist, data analyst, technician, digital marketing). A qualitative approach was taken to the evaluation process. After a demo session, some questions and open discussion, we collected user feedback through a questionnaire structured in four parts. The first part (user data) collects information about the professional background of the interviewees. The second part (human-computer interaction) asks for feedback on several aspects of the graphical user interface (GUI), such as its intuitiveness and interactivity. The third part (tool functionalities) requests feedback about the usefulness or impact of each functionality on existing workflows following a 5-point rating scale (i.e., strongly disagree, disagree, neutral, agree, strongly agree). The last part (conclusions) is designed to elicit suggestions for further refinements, adaptations, and improvements and to measure the results of the survey against the target Key Performance Indicator (KPI).

The general feedback is positive, and the responses gathered through the questionnaires are satisfactory and rich. The insights collected clearly indicate usefulness, relevance, and attractiveness of the demonstrated pipeline. On the other hand, results showed how it would be beneficial to improve the graphical user interface to allow for easier and enhanced interaction by the users. The feedback from the users on the look and feel of the GUI is as follows: the GUI is easy and intuitive, but the user experience should be improved. More options for filtering based on broadcast metadata, like for example date, channel, or programme's title, should be provided. Furthermore, displaying advanced statistics about a person, such as the percentage of time appearance versus the video duration, or the face size against the keyframe size, would be considered a plus. Finally, cross-referencing the face annotations with other information, such as speech transcriptions, would be of significant help for some more in-depth analyses. As for the development of further functionalities, the users would like to be able to query the system for people and context (e.g., a certain person on the street), recognise the type of camera shot (e.g., close-up, very close-up, etc.), search by facial attributes such as hair colour or type of glasses, and look for more than one person at a time.

To measure the KPI, we asked the participants to give their opinion on the following statements:

1) The system provides more comprehensive and more accurate information than those provided by existing systems.
2) The system enables smarter and more efficient processes than do current workflows.
3) The system enables additional or novel insights on content and its impact.

The KPI was met if more than 30% of participants agreed or strongly agreed with the statements. Collected opinions clearly indicate that the KPI has been fully achieved (see Fig. 10), being exceeded by close to three times the target threshold.

V. CONCLUSIONS

This paper described a system for annotating TV personalities in video streams and archives. The system addresses the open set face identification problem, in which a hybrid approach making use of archival annotations, archive content, and external knowledge bases can be employed to build the galleries of reference personalities. The results of the system's experimental campaign show that the defined KPIs were exceeded by a wide margin, demonstrating how existing

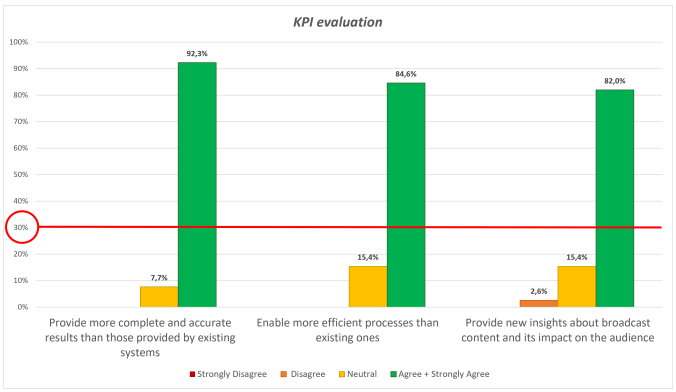

Fig. 10. Illustration of the KPI evaluation for the purpose of detection and annotation of personalities in TV programmes.

media workflows could greatly benefit from the tool and the efficiency improvements it brings.

Driven by the observation that the analysis and reporting of how persons of different gender participate TV programmes is becoming increasingly important, we also investigated the use of facial-related gender features (i.e., face embeddings extracted by deep networks for face recognition) to examine the presence of women and men in TV programmes. The developed approach was evaluated and compared with existing solutions, using Rai programmes of various TV genres. The outcomes of this study could serve as one of the elements for developing advanced analytical tools.

Future research may involve the investigation of additional features in the system, as requested by the participants who conducted the evaluation. These could include, but are not limited to, improvements such as the ability to estimate more attributes of the detected faces, or recognise the environment where the identified individual is played.

REFERENCES

[1] W. N. I. Al-Obaydy and S. A. Suandi, "Open-set face recognition in video surveillance: a survey," in *InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019*. Springer, 2020, pp. 425–436.

[2] Sex and gender. Council of Europe. (last accessed nov 2023). [Online]. Available: https://www.coe.int/en/web/gender-matters/sex-and-gender

[3] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[4] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[5] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," *Advances in face detection and facial image analysis*, pp. 189–248, 2016.

[6] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks," in *2017 IEEE international joint conference on biometrics (IJCB)*. IEEE, 2017, pp. 464–473.

[7] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7044–7053.

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[9] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the ieee/cvf international conference on computer vision*, 2019, pp. 692–702.

[10] A. Popescu, L.-D. Ștefan, J. Deshayes-Chossart, and B. Ionescu, "Face verification with challenging imposters and diversified demographics," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3357–3366.

[11] F. Wang *et al.*, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 765–780.

[12] Y. Zhang *et al.*, "Global-local gcn: Large-scale label noise cleansing for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7731–7740.

[13] G. Amato *et al.*, "Ai in the media and creative industries," *arXiv preprint arXiv:1905.04175*, 2019.

[14] M.-F. de Lima-Santos and W. Ceron, "Artificial intelligence in news media: current perceptions and future outlook," *Journalism and media*, vol. 3, no. 1, pp. 13–26, 2021.

[15] F. Tsalakanidou, "AI technologies and applications in media: State of Play, Foresight, and Research Directions," CERTH, Report D2.3, feb 2022, (last accessed oct 2023). [Online]. Available: https://www.ai4media.eu/reports/roadmap-ai-technologies-and-applications-in-media/

[16] B. Sančanin and A. Penjišević, "Use of artificial intelligence for the generation of media content," *Social Informatics Journal*, vol. 1, no. 1, pp. 1–7, 2022.

[17] C. Trattner *et al.*, "Responsible media technology and ai: challenges and research directions," *AI and Ethics*, vol. 2, no. 4, pp. 585–594, 2022.

[18] A. Bruccoleri, R. Iacoviello, A. Messina, S. Metta, M. Montagnuolo, and F. Negro, "AI in vision: High quality video production & content automation," Rai – Radiotelevisione Italiana, White Paper, oct 2023, (last accessed oct 2023). [Online]. Available: https://www.ai4media.eu/whitepapers/ai-in-vision-high-quality-video-production-content-automation/

[19] A. Schjøtt, R. Bocyte, J. Oomen, L. Dutkiewicz, and G. Thallinger. (2023, sep) AI for audiovisual archives. (last accessed oct 2023). [Online]. Available: https://www.ai4media.eu/wp-content/uploads/2023/09/Factsheet_AIforAudiovisualArchives_final.pdf

[20] A. Mercier, S. Ducret, C. Bürki, and L. Bouchet, "Examples of uses of artificial intelligence in video archives," in *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, 2019, pp. 49–50.

[21] E. L. De Quintana and A. L. Carpio, "Artificial intelligence for a role change in television archives: The atresmedia–etiqmedia experience," *Journal of Digital Media Management*, vol. 10, no. 2, pp. 177–187, 2021.

[22] V. Bazán-Gil, "Artificial intelligence applications in media archives," *Profesional de la información*, vol. 32, no. 5, 2023.

[23] V. A. de Jesus Oliveira, G. Rottermanner, S. Größbacher, M. Boucher, and P. Judmaier, "Requirements and concepts for interactive media retrieval user interfaces," in *Nordic Human-Computer Interaction Conference*, 2022, pp. 1–10.

[24] V. A. de Jesus Oliveira *et al.*, "Taylor–impersonation of AI for audiovisual content documentation and search," in *International Conference on Multimedia Modeling*. Springer, 2023, pp. 751–757.

[25] A. Majumdar and P. Nasiopoulos, "Frontal face recognition from video," in *Advances in Visual Computing: 4th International Symposium, ISVC 2008, Las Vegas, NV, USA, December 1-3, 2008. Proceedings, Part II 4*. Springer, 2008, pp. 297–306.

[26] B. H. Jeon, S. U. Lee, and K. M. Lee, "Rotation invariant face detection using a model-based clustering algorithm," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, vol. 2. IEEE, 2000, pp. 1149–1152.

[27] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, pp. 137–154, 2004.

[28] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 354–370.

[29] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 192–201.

[30] C. Zhang, X. Xu, and D. Tu, "Face detection using improved faster rcnn," *arXiv preprint arXiv:1802.02142*, 2018.

[31] J. Li *et al.*, "Dsfd: dual shot face detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5060–5069.

[32] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[33] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," *arXiv preprint arXiv:2105.04714*, 2021.

[34] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.

[35] E. Bijl, "A comparison of clustering algorithms for face clustering," Ph.D. dissertation, University of Groningen, 2018.

[36] G. Heusch, Y. Rodriguez, and S. Marcel, "Local binary patterns as an image preprocessing for face authentication," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 6–pp.

[37] J. Križaj, V. Štruc, and N. Pavešić, "Adaptation of sift features for robust face recognition," in *Image Analysis and Recognition: 7th International Conference, ICIAR 2010, Póvoa de Varzim, Portugal, June 21-23, 2010. Proceedings, Part I 7*. Springer, 2010, pp. 394–404.

[38] G. Du, F. Su, and A. Cai, "Face recognition using surf features," in *MIPPR 2009: Pattern recognition and computer vision*, vol. 7496. SPIE, 2009, pp. 593–599.

[39] P. Antonopoulos, N. Nikolaidis, and I. Pitas, "Hierarchical face clustering using sift image features," in *2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing*. IEEE, 2007, pp. 325–329.

[40] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B.-L. Lu, "Person-specific sift features for face recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2. IEEE, 2007, pp. II–593.

[41] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[42] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[43] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[44] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, "Sub-center arcface: Boosting face recognition by large-scale noisy web faces," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 741–757.

[45] X. An *et al.*, "Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4042–4051.

[46] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi, "Ghostfacenets: Lightweight face recognition model from cheap operations," *IEEE Access*, 2023.

[47] G. G. Chrysos, S. Moschoglou, G. Bouritsas, J. Deng, Y. Panagakis, and S. Zafeiriou, "Deep polynomial neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 8, pp. 4021–4034, 2021.

[48] J. Jung, S. Lee, H.-S. Oh, Y. Park, J. Park, and S. Son, "Unified negative pair generation toward well-discriminative feature space for face recognition," *arXiv preprint arXiv:2203.11593*, 2022.

[49] X. An *et al.*, "Partial fc: Training 10 million identities on a single machine," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1445–1449.

[50] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "Beyond identity: What information is stored in biometric face templates?" in *2020 IEEE international joint conference on biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[51] A. Swaminathan, M. Chaba, D. K. Sharma, and Y. Chaba, "Gender classification using facial embeddings: A novel approach," *Procedia Computer Science*, vol. 167, pp. 2634–2642, 2020.

[52] Y. Lin and H. Xie, "Face gender recognition based on face recognition feature vectors," in *2020 IEEE 3rd International conference on information systems and computer aided education (ICISCAE)*. IEEE, 2020, pp. 162–166.

[53] M. Farzaneh, "Arcface knows the gender, too!" *arXiv preprint arXiv:2112.10101*, 2021.

[54] T. Kim, "Generalizing MLPs with dropouts, batch normalization, and skip connections," *arXiv preprint arXiv:2108.08186*, 2021.

[55] M. Kuprashevich and I. Tolstykh, "Mivolo: Multi-input transformer for age and gender estimation," *arXiv preprint arXiv:2307.04616*, 2023.

[56] E. Bonet Cervera, "Age & gender recognition in the wild," B.S. thesis, Universitat Politècnica de Catalunya, 2022.

[57] A. Hast, "Sex Classification of Face Images using Embedded Prototype Subspace Classifiers," *Computer Science Research Notes*, vol. 3301, pp. 43–52, 2023.

[58] D. Doukhan, G. Poels, Z. Rezgui, and J. Carrive, "Describing gender equality in french audiovisual streams with a deep learning approach," *VIEW Journal of European Television History and Culture*, vol. 7, no. 14, pp. 103–122, 2018.

[59] M. Bazin and C. Méadel, "Les SHS dans le projet Gender Equality Monitoring," GEM, Tech. Rep., oct 2022, (last accessed oct 2023). [Online]. Available: https://univ-pantheon-assas.hal.science/hal-03877033

[60] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[61] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[62] C. Biemann, "Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems," in *Proceedings of TextGraphs: the first workshop on graph based methods for natural language processing*, 2006, pp. 73–80.

[63] L. Chang, A. Pérez-Suárez, M. González-Mendoza *et al.*, "Effective and generalizable graph-based clustering for faces in the wild," *Computational Intelligence and Neuroscience*, vol. 2019, 2019.

[64] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 4, pp. 824–836, 2018.

[65] G. Chen, P. Chen, Y. Shi, C.-Y. Hsieh, B. Liao, and S. Zhang, "Rethinking the usage of batch normalization and dropout in the training of deep neural networks," *arXiv preprint arXiv:1905.05928*, 2019.

[66] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[67] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007, pp. 410–420.

[68] P. J. Phillips, P. Grother, and R. Micheals, "Evaluation methods in face recognition," *Handbook of face recognition*, pp. 551–574, 2011.

[69] R. Rothe, R. Timofte, and L. Gool, "Imdb-wiki-500k+ face images with age and gender labels," *Online] URL: https://data. vision. ee. ethz. ch/cvl/rrothe/imdb-wiki*, vol. 4, 2015.

[70] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on information forensics and security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[71] (last accessed noc 2023). [Online]. Available: https://insightface.ai/

[72] (last accessed noc 2023). [Online]. Available: https://github.com/sajjjadayobi/FaceLib

[73] (last accessed noc 2023). [Online]. Available: https://github.com/serengil/deepface