2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Attribute-based quality assessment for demographic estimation in face videos

Fabiola Becerra-Riera, Annette Morales-González and Heydi Méndez-Vázquez
Advanced Technologies Application Center (CENATAV)
7A ♯21406 Siboney, Playa, P.C.12200, Havana, Cuba
Email: {fbecerra,amorales,hmendez}@cenatav.co.cu

Jean-Luc Dugelay
Digital Security Department, EURECOM
Campus Sophia Tech, 450 route des Chappes
F-06410 Biot Sophia Antipolis, France
Email: Jean-Luc.Dugelay@eurecom.fr

***Abstract*—Most existing works regarding facial demographic estimation are focused on still image datasets, although nowadays the need to analyze video content in real applications is increasing. We propose to tackle gender, age and ethnicity estimation in the context of video scenarios. Our main contribution is to use an attribute-specific quality assessment procedure to select most relevant frames from a video sequence for each of the three demographic modalities. Selected frames are classified with fine-tuned MobileNet models and a final video prediction is obtained with a majority voting strategy. Our validation on three different datasets and our comparison with state-of-the-art models, show the effectiveness of the proposed demographic classifiers and the quality pipeline, which allows to reduce both: the number of frames to be classified and the processing time in practical applications; and improves the soft biometrics prediction accuracy.**

I. Introduction

Demographic soft biometrics (*e.g.* gender, ethnicity, age) are among the most frequently used traits for improving and complementing the performance of biometric systems. Their effectiveness in the context of biometric-related applications involving access control, video surveillance, person re-identification or human-computer interaction among others, has been set through several works in literature [1], [2].

Despite their already established convenience in video protection domains, there are just a few works regarding demographic soft biometrics estimation in videos and a reduced number of public video datasets are available to validate results [3]. Most of these available collections also have the additional problem of unbalanced classes (usually, more Males than Females, more Caucasians than other representative ethnicities, or more Middle-Age people than Children and Seniors).

Captured frames from real video applications, characterized by uncontrolled environments and non-cooperative subjects of unknown identities, usually go through a loss of visual information; occlusions; variations in terms of pose, illumination and facial expressions. Under these conditions, the automatic estimation of demographic attributes becomes significantly complex, and existing still images classifiers decrease their performance.

Motivated by the lack of works dealing with demographic estimators in the domain of video applications, and the increase of uploaded video content to the Internet, we propose a pipeline for the automatic estimation of gender, ethnicity and age in videos. The main contribution in this pipeline is a quality assessment procedure for each specific soft-biometric modality, including 12 quality measures, designed to select key frames from a video sequence. The most discriminant frames are classified by means of a fine-tuned MobileNet architecture [4] and a majority voting strategy is performed to obtain the final video prediction. The entire procedure reduces both: the number of frames to be classified and the processing time; and improves the estimation accuracy of using all frames.

II. Related Work

Most existing works on demographics estimation are focused on still images, and just a few approaches propose a solution for estimating gender, ethnicity and age at once [5], [6]. Gender estimation is the most studied demographic attribute. State-of-the-art approaches already obtain very good results in most of the still image datasets [7], [8]. For ethnicity, on the other hand, despite state-of-the-art accuracy is good, there is no standard consensus of ethnic labels and datasets to validate results from one paper to another: authors define their own ethnicity classes according to the practical interest of the research, in some cases depending on people skin color instead of their region of origin [9], [10]. The problem of age estimation remains the most complex one. State-of-the-art algorithms for exact age estimation rounds the 3 or 4 years of Mean Absolute Error (MAE) in most cases [11], while there is still a large margin of improvement for accuracy in the estimation of age groups [12].

A detailed review on demographic estimation algorithms can be found in [13], where authors discussed the lack of works dealing with demographic estimation in videos, as well as the lack of proper datasets for its evaluation in this type of scenarios. Analyzing the works that address soft-biometrics in videos, we find that the proposal of [14] consists on a demographic estimation strategy that encodes and exploits the correlation between face images of a video sequence through manifold learning. In [15], pixel intensity-based and Biologically Inspired Model (BIM) features are employed, along with Support Vector Machine (SVM) classifiers, to obtain a prediction of gender and ethnicity in video frames from surveillance scenarios. The works of [16], [17] employ smile-

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

dynamic features for age and gender respectively, exploring the influence of some facial expressions in the estimation of demographic attributes. The proposal of [18] represents the temporal dependencies on a video sequence by means of a probabilistic graphical model. The gender estimation from a face at a specific time depends on features extracted from previous frames of the sequence. A temporal coherent face descriptor for video gender recognition is proposed in [19]. Authors build a single video representation by concatenating pixel-intensity features from all frames and a SVM classifier is used to obtain video prediction at once. The use of intermediate features of a CNN was explored in [20], with a component-based face representation methodology that exploits the gender information provided by different face parts. The intermediate features extracted from video key frames are combined with two different strategies in order to preserve the temporal information and obtain the final gender prediction with a Random Forest (RF) classifier. These works explore different types of frame combination to obtain a single video prediction, but they do not address the problem of the face quality difference that can affect frames in a sequence. This quality difference may produce prediction mistakes when using classifiers trained in still image scenarios.

The closest work to our proposal is the one from Selim *et al.* [21], in which each frame of a video sequence is classified by means of a deep and compact CNN and a quality assessment step is introduced in the pipeline. A quality criterion that quantifies frame blurriness is applied, and specific CNN models are trained on faces with similar blurriness to obtain the prediction. In contrast to that work, we propose to evaluate the quality of each video frame, and select only the best quality ones for prediction combination. Another difference with Selim *et al.*'s proposal is that they only evaluate the quality in terms of blurriness, while we employ 12 quality measures comprising illumination, head pose, resolution, occlusions, among others.

The works of [14], [15] validated their results in video datasets with demographic attribute annotations which are not under public domain. Authors of [16], [17] evaluate the performance of their proposals in the UvA-NEMO Smile Database [16], a controlled face video collection fully annotated for gender and age, with fairly good quality/resolution videos. In [18], [19], [20], [21] the McGill Faces dataset [22] is used to validate results for gender classification in videos, but in [20] the prediction accuracy reached 100%, which makes this dataset limited to more extensive experimentation. To the best of our knowledge, these two datasets are the only face video collections with demographic annotations that are publicly available.

III. Proposal

One of the advantages of dealing with video sequences instead of still images is the information availability. Although it is a much more complex problem, the information present in different frames of a video sequence can be combined in one single representation, describing all the sequence variability; or use a majority voting strategy among frames to decrease error probabilities in the final video prediction. However, in real-time video applications there is no chance for processing an entire sequence. Therefore, the initial selection of frames that better represents a video will be a key step in the subsequent classification procedure.

Selecting the best frames of a video sequence could be done in many ways, depending on what is considered good. In the context of uncontrolled environments, where several factors influence the frame capture, choosing just the least affected frames should be a good strategy. We could then associate the goodness of a frame to certain quality parameters; for instance, illumination, resolution, facial expressions, among others.

In this paper, we propose a quality assessment for key frames selection that will depend on the specific demographic attribute: certain frames could have good quality to accurately classify the gender of a person, but not be suitable for age estimation, for instance. To evaluate the quality of a frame, the outputs of different quality measures could be combined to obtain a single value, but then, the definition of certain combination weights would be necessary. Since we do not know beforehand which quality measures are more important for the estimation of specific attributes, we employ a classifier to learn their relevance for each classification task.

Given a video sequence and a specific demographic attribute (*e.g.* gender), we classify each frame according to the attribute and we label it as "Good" or "Bad" depending on if it was correctly classified or not. We consider correctly classified frames as having Good Quality and wrongly classified ones as having Bad Quality for the particular task of gender estimation in this case.

Once we have identified Good and Bad Quality frames, we represent them by a vector of 12 components corresponding to the output score of 12 different quality measures relatives to: Pose, Illumination, Occlusion, Resolution, Sharpness, Mouth State, Eyes State, Gaze, Color Leveling, Face Centering, Red Eyes and Uniform Background; which were introduced in the work of [23] to determine the identification value of face images according to the standard ISO/IEC 19794-5 [24]. This quality ground truth with its corresponding quality features is used to train a Random Forest (RF) classifier [25] to learn to discriminate Good and Bad Quality frames based on these 12 quality measures. Final video prediction is obtained with a majority voting strategy among best quality frames selected by the RF classifier:

Formally, given a sequence of video frames $F = \{f_1, f_2, \dots f_m\}$ and an attribute α with $L = \{l_1, l_2, \dots l_k\}$ possible labels, we define $q^n \subseteq F$ as the set of the n best quality frames of the sequence $(1 \leq n \leq m)$ and $q^n_{l_k} \subseteq q^n$ as the set of the $f \in q^n$ for which the classification according to α (from now on defined as C_α) corresponds to the l_k label:

$$q^n_{l_k} = \{f \in q^n \mid C_\alpha(f) = l_k\} \tag{1}$$

Then, the l_i resulting after applying the majority voting strategy over the sequence F given the α attribute, responds to the following formulation:

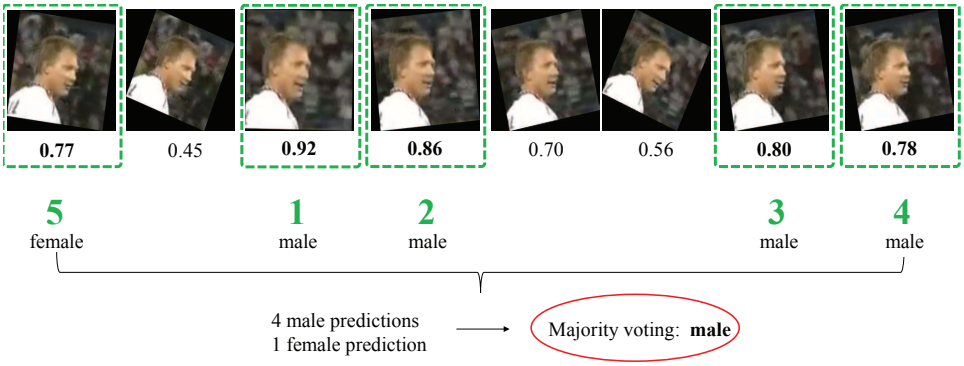

Fig. 1. Example of quality assessment procedure for gender estimation.

$$l_i \in L(1 \leq i \leq k) \mid \forall l_j \in L(1 \leq j \leq k, j \neq i), \mid q_{l_j}^n \mid < \mid q_{l_i}^n \mid \quad (2)$$

Figure 1 shows an example of a gender quality assessment procedure. In the figure, given 8 frames and the probability of each one to be a Good Quality frame for gender estimation, we select the n best ones (in this case we set $n = 5$) and obtain the video gender prediction with the majority voting strategy.

IV. Experiments

A. Dataset selection

As mentioned before, there are several still image datasets for face analysis with available demographic attribute labeling, but they are not suitable for evaluating the performance of soft-biometric estimators in video scenarios. Current image datasets do not possess the variability or noise present specifically in videos, like low resolution, motion blur, arbitrary poses, occlusions, bad and varying illumination problems, etc. Even though we can count now with several video datasets for face analysis tasks, to the best of our knowledge, just the McGill and the UvA-Nemo datasets present some demographic annotations. For the former, a very small dataset, gender classification accuracy already reached 100% in the work of [20], therefore is unnecessary to conduct more extensive experiments on it. For the latter, we found that it contains fairly good quality videos, so the slight difference between those frames tagged as "Good" and those tagged as "Bad" by our models, leave little or no room for a quality procedure. Taking all this into account, we adopted two evaluation strategies in order to show the suitability of using a quality assessment within the video classification pipeline: (1) We took a face image dataset representing frames of a video, fully annotated with demographic data, and we augmented it adding noise to simulate a video scenario with several frames of different qualities. (2) We used the uncontrolled Youtube Faces Database (YTF), a video collection for which we mapped the Labelled Faces in the Wild (LFW) gender and ethnicity labels to its corresponding identities.

UvA-NEMO Smile Database [16]. This dataset is a large-scale collection created to analyze the change in dynamics of smiles for different ages varying from 8 to 76 years. It consist of 1240 smile RGB videos (597 spontaneous and 643 posed) from 400 subjects (185 female, 215 male), recorded with a resolution of 1920×1080 pixels under controlled illumination conditions. Videos are annotated with gender, age, and specific smile spontaneity. Subjects are predominantly Caucasian males (no other ethnicities were identified) and most of them are "child-teenagers" or "middle-age". Distribution of gender and age groups can be found in Table I and sample images from one subject of the database can be seen in the first row of Figure 2.

EURECOM Kinect Face Dataset [26]. This dataset consists on the multimodal facial images of 52 people (14 females, 38 males) obtained by Kinect. The data was captured in two sessions spaced for half a month. In each session, the dataset provides the facial images of each person in nine states of different facial expressions, lighting and occlusion conditions: neutral, smile, open mouth, left profile, right profile, occlusion eyes, occlusion mouth, occlusion paper and light on. We have selected this dataset, because these nine images per subject are actually frames extracted from a video. Although we do not have the entire sequence, this configuration can be approximated to an actual video scenario. Information about the gender, year of birth, ethnicity, glasses presence and capture time of each session is also available. Sample images from one subject can be seen in the second row of Figure 2.

In order to simulate a longer video sequence with more quality variability, we augmented each RGB sequence from the EURECOM Kinect Face Dataset by adding noisy and low resolution images. For each one of the nine original images per sequence, we incorporated two generated images for a total of 27: the first one was obtained by randomly adding 1 of 9 different types of noise to the original image; the second one was created by down-scaling the original image to 20% of its size, and then resizing it up to simulate poor resolution conditions[1].

The dataset class distribution of gender and ethnicity can be

[1] http://rgb-d-2.eurecom.fr/

found in Table I. As can be seen, there is an imbalance that indicates more males than females and more Caucasians than people from other ethnicities, which is a common problem in most of the image collections with demographic annotations, as we found for the UvA-Nemo dataset. The age values range from 25 to 38 years old: 45 subjects are considered "young" (bellow 30 years old) and only 7 belong to "middle-age" class. Hence, in this case we decided to estimate exact age instead of age groups.

YTF Dataset [27]. This dataset was created to evaluate unconstrained face recognition in videos. It contains 3 425 videos of 1 595 different people obtained from YouTube. These videos are affected by many of the factors mentioned above (see third row of Figure 2).

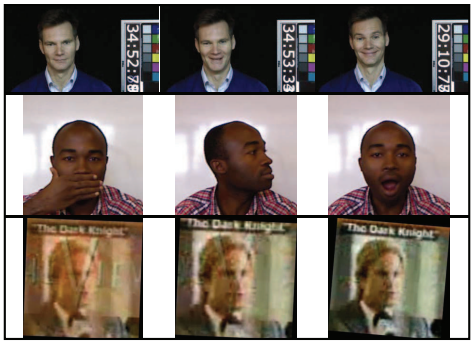

Fig. 2. Sample frames from one video sequence in UvA-Nemo (first row), EURECOM (second row) and Youtube Faces (third row) datasets.

The YTF dataset was constructed starting from the subjects in the popular Labeled Face in the Wild (LFW) database [28]. Hence, we have mapped the 73 attribute labels obtained for LFW in [29] with the subjects on YTF. From those attributes, we kept the ones related to gender and ethnicity. We employed a subset of the YTF dataset, comprising 25% of the total dataset, where we manually checked the provided labels for correctness, since the attribute labels available are those automatically computed in [29] with an approximate error of 10%. Given the task that we are trying to solve, it was mandatory for us to have a fully correct ground-truth information. In Table I, it can be seen the class distribution for gender and ethnicity in the YTF dataset: there is also a large imbalance in the data, similar to that found in EURECOM and UvA-Nemo datasets. The predominant subject type is Caucasian Male, for a large margin over other classes.

For the YTF, age labels were not directly transferable from LFW. This is due to the fact that, in LFW, each subject is assigned a single age-related label but, in YTF dataset, many individuals have several video sequences from different moments of their life, therefore, it is incorrect to make an association of a single age group to all the sequences of the same person.

B. Implementation details

In order to obtain individual demographic estimators for gender, ethnicity and age, we trained CNN classifiers using the MobileNet model [4], which is based on a streamlined architecture to build light-weight deep neural networks, to make the real-time estimation of a video sequence as efficient as possible (MobileNet network structure is described in detail in [4]). Training was performed on three publicly uncontrolled image datasets: IMDB-Wiki Dataset [30], UTKFace Dataset [31] and LFW [28]. We joined all three datasets to obtain a larger and varied training set and we employed the same face alignment and crop used in [30]. We considered exact age range from 1 to 100 years old and four different classes in ethnicity estimation: Caucasians, Africans, Asians and Others (including Hispanics, Latinamericans and Indians).

For the quality classifier that should decide if a frame is Good or Bad according to its 12 quality scores, we employed a RF classifier, given its good results for classification tasks and the possibility of interpreting their outcome in terms of variable importance [25]. This feature can provide an idea of how individual quality problems affect each specific soft-biometric estimation.

TABLE I
DATASETS DISTRIBUTION OF **GENDER**: FEMALE (F), MALE (M); **ETHNICITY**: CAUCASIAN (C), AFRICAN (AF), ASIAN (AS), INDIAN/LATINAMERICAN (I/L); AND **AGE**: CHILD-TEENAGERS (0-18), YOUNG (19-30), MIDDLE-AGE (31-59), SENIOR (60-)

Datasets	Gender		Ethnicity				Age			
	F	M	C	Af	As	I/L	(0-18)	(19-30)	(31-59)	(60-)
UvA-Nemo	185	215	400	-	-	-	150	81	150	19
EURECOM	14	38	20	3	11	18	-	45	7	-
YTF	149	285	315	33	15	72	-	-	-	-

C. Results

In order to validate our proposal, we performed experiments in the selected datasets by comparing several frame combination strategies. "Individual frames" strategy just considers frames as single independent images. The strategy named "Sequence (all frames)" performs a majority voting among all frames in a sequence. "Sequence - quality N frames" performs the majority voting only taking into account the top N most relevant frames in the sequence (for the present case $N = 5, 10$), obtained with the attribute-specific quality assessment. "Sequence - random N frames" performs the majority voting on N random frames of the sequence ($N = 5, 10$), and it is intended to depict the case where quality information is not used for frame selection.

In our experiments we used the accuracy of classification as the evaluation metric. We show overall accuracy, the accuracy for predicting individual classes and the geometric mean (G-Mean). G-Mean is a better overall estimation of performance when the datasets are too unbalanced, as our case, since it takes into account individual classes accuracy, disregarding the number of samples in each class. For the case of exact age estimation, we employed Mean Absolute Error (MAE), which is the most used metric for this task, since exact age is a continuous variable. In our case, MAE is an average of

absolute errors between the ground-truth and the predicted exact age.

1) Gender estimation: We performed gender estimation experiments in 2 different subsets of the UvA-Nemo dataset (the subset containing deliberate smile videos and the one containing spontaneous smiles) and also in the entire collection (see Table II). Our results were compared to those obtained by the authors of [17] and [32], which are, to the best of our knowledge, the only ones that validate gender estimation in this video dataset. We also included performance accuracy of a state-of-the-art classifier (called DEX) proposed by [30], which is based on the popular VGG-16 architecture. Although the original DEX model is devoted to estimate apparent age from face images, the released gender estimation model[2] shows impressive results, but at the cost of a considerable increase in processing time (7 times slower than MobileNet).

For the experimentation in the two subsets (Deliberate and Spontaneous smiles), we took the best performing method reported in [17], which corresponds to the combination of COTS and Smile Dynamics (Bagged Trees, PCA), and we used it in our comparison. Columns "Deliberate" and "Spontaneous" from Table II shows a 15 folds average accuracy over these two subsets respectively, considering two age groups: subjects with less than 20 years old and subjects older than 19.

The proposal of [32] was validated in the entire UvA-Nemo dataset. Authors reported average accuracy through 5 folds and two age groups with a small variation with respect to the ones defined in [17]: young people was considered from 20 or less years old, and adult people older than 20, in this case. Results are reported in column "Entire dataset" from Table II.

Since a large amount of videos in the dataset were considered as good quality videos, we were not able to train a gender quality estimator for this collection due to the lack of bad quality samples and their slight differences with the good quality ones, as we explained in section IV-A. Our MobileNet classifier was mostly accurate and stable in the classification of individual frames, and that's why considering the majority voting strategy among all frames don't really change individual frame's results and even by selecting random frames of the sequence, the results remains consistent. The performance accuracy of our model was always superior to the one reported by [17] for both, deliberate and spontaneous subsets. In the case of the entire dataset's results, the proposal of Bilinski et al. [32] was 2.98% more accurate in the estimation of people younger than 20 years old, but 3.08% less effective than our model for adult people, resulting in a slightly better G-Mean value for them (88.62% for [32] and 88.54% for us). In comparison with the results obtained with the DEX model, our MobileNet classifier was more accurate in most of the experiments, showing superior overall and G-Mean values in all cases. Even for the classification of people with spontaneous smiles, when the DEX classifier obtained a result 1.01% better than ours in the adults subset, the MobileNet model save the difference by being 3.89% more accurate in the estimation of younger people, showing a better G-Mean performance. Best results by column are highlighted in bold. Although we were not able to analyze the quality strategy performance in this case, we thought it was useful to show these results in order to compare our baseline methods to other state-of-the-art works evaluated in this dataset.

In order to validate the hypothesis that a quality assessment could indeed improve the results obtained by using all frames of a video sequence in more problematic scenarios, we continue the experimentation in the other two datasets. Table III show the results for gender estimation in the EURECOM and YTF datasets. Our main goal is to show that the proposed quality strategy for selecting frames in the video soft-biometric estimation works with different classifiers and different conditions, and that's why best results by column are highlighted for each classifier individually.

As can be seen in Table III, the proposed quality assessment is effective without dependence on the dataset or the classifier. For the EURECOM dataset, the strategy of selecting the top 5 more discriminant frames outperformed all other strategies, yielding 100% of classification accuracy using our MobileNet classifier. The strategy of selecting the top 10 more discriminant frames obtained the second best results. For the DEX classifier this behaviour was confirmed. For the YTF dataset, the best strategy was selecting the top 10 more relevant frames, followed by the selection of the top 5, confirmed by both classifiers. It is worth noting that in both datasets and for the 2 tested classifiers, our quality assessment strategy favored the results of the minority class (Females) over the majority class (Males).

2) Ethnicity estimation: In Table IV we show the results for ethnicity estimation. We were not able to find an available state-of-the-art pre-trained model for this task, therefore, we used our MobileNet classifier alone. The results are very similar to those displayed for gender estimation in both EURECOM and YTF datasets: best performance by column can be seen highlighted in bold. Once again, the best strategy was the one selecting the top 5 more discriminant frames and the second best was obtained selecting the top 10. It is interesting to see that using our quality strategy, in the EURECOM dataset, the minority classes Asian and Other achieved 100% of accuracy, showing great improvements, and the majority class (Caucasian) also was largely favored by these quality strategies.

3) Age estimation: We validated our 6 different combination strategies in the UvA-Nemo collection for the task of exact age estimation, and our results were compared to the ones achieved by the authors of [16], who reported their best performance using the fusion of appearance and dynamic features. DEX classifier results were also included. UvA-Nemo was split into 10 folds and a 10-fold cross validation procedure was performed . In table V, best results by column are highlighted in bold. Overall MAE represents the average MAE by folds and ours is not better than the one obtained in [16]; however, the MAE of individual age groups show that our classifier is more consistent throughout all groups and

[2]https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

TABLE II
GENDER CLASSIFICATION RESULTS IN THE DELIBERATE AND SPONTANEOUS SUBSETS FROM UVA-NEMO DATASET, AND ALSO IN THE ENTIRE COLLECTION

		Deliberate Accuracy (%)				Spontaneous Accuracy (%)				Entire dataset Accuracy (%)			
Classifier	Strategy	Overall	G-Mean	< 20	> 19	Overall	G-Mean	< 20	> 19	Overall	G-Mean	≤ 20	> 20
MobileNet (Ours)	Individual frames	90.27	89.11	83.90	94.64	89.25	**88.62**	**83.61**	93.92	89.59	88.47	83.20	94.08
	Sequence (all frames)	90.51	88.84	83.29	94.77	88.96	87.89	82.21	93.97	**89.76**	88.54	83.32	**94.09**
	Sequence - random 5 frames	**91.28**	**89.69**	**84.35**	**95.37**	88.46	87.45	81.94	93.32	89.19	87.83	82.12	93.94
	Sequence - random 10 frames	90.67	88.95	83.29	95.00	**89.29**	88.12	82.44	94.20	89.59	88.31	82.89	**94.09**
DEX [30]	Individual frames	88.75	87.68	83.12	92.49	87.96	86.73	79.72	94.35	88.19	86.75	80.61	93.36
	Sequence (all frames)	89.72	88.50	83.71	93.56	88.09	86.52	79.08	94.66	88.95	87.51	81.68	93.75
	Sequence - random 5 frames	89.72	87.83	82.50	93.50	87.59	86.06	78.47	94.38	88.87	87.39	81.43	93.79
	Sequence - random 10 frames	89.73	87.56	80.80	94.89	88.09	86.49	78.76	**94.98**	88.79	87.24	81.05	93.91
Dantcheva and Brémond [17]	Sequence (all frames)	-	84.53	76.92	92.89	-	84.58	76.92	93	-	-	-	-
Bilinski *et al.* [32]	Sequence (all frames)	-	-	-	-	-	-	-	-	-	**88.62**	**86.30**	91.01

TABLE III
GENDER CLASSIFICATION RESULTS IN EURECOM AND YTF DATASETS

		EURECOM Accuracy (%)				YTF Accuracy (%)			
Classifier	Strategy	Overall	G-Mean	Female	Male	Overall	G-Mean	Female	Male
MobileNet (Ours)	Individual frames	76.18	81.07	94.84	69.30	94.30	92.24	86.97	97.83
	Sequence (all frames)	91.35	92.88	96.43	89.47	94.66	92.67	86.63	99.15
	Sequence - random 5 frames	84.62	88.86	**100.0**	78.95	94.55	92.75	87.23	98.64
	Sequence - random 10 frames	87.50	91.04	**100.0**	82.89	94.55	92.67	86.93	98.81
	Sequence - quality 5 frames	**100.0**	**100.0**	**100.0**	**100.0**	95.75	94.29	**89.67**	99.15
	Sequence - quality 10 frames	99.04	99.34	**100.0**	98.68	**95.86**	**94.37**	**89.67**	**99.32**
DEX [30]	Individual frames	89.35	77.81	60.58	99.95	91.57	89.01	82.66	95.84
	Sequence (all frames)	92.31	84.52	71.43	**100.0**	90.73	88.13	80.55	96.43
	Sequence - random 5 frames	91.35	82.38	67.86	**100.0**	90.84	88.16	81.46	95.41
	Sequence - random 10 frames	92.31	84.52	71.43	**100.0**	91.38	89.09	82.32	96.43
	Sequence - quality 5 frames	**99.04**	**98.20**	**96.43**	**100.0**	92.49	90.27	84.37	**96.60**
	Sequence - quality 10 frames	98.08	96.36	92.86	**100.0**	**92.87**	**91.03**	**85.91**	96.46

TABLE IV
ETHNICITY CLASSIFICATION RESULTS IN EURECOM AND YTF DATASETS

		EURECOM Accuracy (%)						YTF Accuracy (%)					
Classifier	Strategy	Overall	G-Mean	Caucasian	African	Asian	Other	Overall	G-Mean	Caucasian	African	Asian	Other
MobileNet (Ours)	Individual frames	63.57	68.84	57.69	91.98	66.67	63.48	75.67	68.55	83.20	89.91	82.84	35.64
	Sequence (all frames)	85.58	89.25	72.50	**100.0**	95.45	91.67	76.63	67.89	83.90	87.67	76.67	37.67
	Sequence - random 5 frames	75.00	79.22	75.00	**100.0**	72.73	72.22	76.30	65.49	84.20	87.67	70.00	35.62
	Sequence - random 10 frames	79.81	84.17	77.50	**100.0**	86.36	75.00	76.20	65.11	84.20	89.04	70.00	34.25
	Sequence - quality 5 frames	**99.04**	**99.37**	**97.50**	**100.0**	**100.0**	**100.0**	**78.59**	**70.19**	**85.54**	**93.15**	**76.67**	**39.73**
	Sequence - quality 10 frames	97.12	98.07	92.50	**100.0**	**100.0**	**100.0**	78.37	69.85	85.39	**93.15**	**76.67**	39.04

that's why our standard deviation is considerable lower, from 4.87 that report the authors to 0.73 in our best experiment. This little margin of difference in overall results is explained by the imbalance problem in this dataset, as can be seen in the last row of Table V. That's why we included a column showing the G-Mean for age groups, in order to take into account the unbalanced classes. In that column it is possible to see that our results are much better than the one of [16]. DEX classifier was slightly more accurate than our MobileNet model, especially for group ages over 50 years old; however, as explained before, DEX is no suitable for video applications due to its larger processing time. There was no room for a quality procedure in this experiment, due to the intrinsic good quality of UvA-Nemo dataset.

In Table VI we can see the validation for exact age estimation in the EURECOM dataset using our MobileNet classifier and DEX. The Overall MAE shows an improvement for the strategy of selecting the 10 top most relevant frames for MobileNet, and for DEX the largest improvement was obtained by selecting the 5 best ones. In general, the MAE for individual ages also exhibits the same behavior. In bold, best results by column for each classifier are highlighted.

D. Discussion

According to the experiments performed in the previous section, we can confirm that the quality analysis in the soft biometrics prediction is a relevant step to improve the classification results in video scenarios. We could see that the performance of the three demographic attribute estimators exhibit similar behavior when selecting the most relevant frames in each case.

TABLE V
MAE IN THE ESTIMATION OF EXACT AGE IN UVA-NEMO DATASET

Classifier	Strategy	MAE (years) Overall	G-Mean	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79
MobileNet (Ours)	Frames	4.94 (± 0.76)	3.92	0.62	2.49	7.23	7.02	6.98	6.75	6.6	2.27
	Seq. (all frames)	4.88 (± 0.73)	3.77	**0.53**	2.55	7.27	**6.6**	6.59	7.2	6.32	2.08
	Seq. random 5	4.95 (± 0.81)	3.79	0.55	2.52	7.25	6.81	6.81	6.83	6.43	2.09
	Seq. random 10	4.93 (± 0.75)	3.76	**0.53**	2.54	7.17	6.69	6.81	7.12	6.23	2.05
DEX [30]	Frames	4.13 (± 0.88)	3.19	1.33	1.43	6.21	6.98	5.89	3.99	4.76	1.15
	Seq. (all frames)	**4.09 (± 1.03)**	**3.02**	1.23	1.37	6.18	6.74	5.69	3.91	**4.69**	**0.95**
	Seq. random 5	4.18 (± 0.98)	3.23	1.22	1.41	6.13	7.23	5.82	3.94	4.85	1.38
	Seq. random 10	4.09 (± 1.06)	3.11	1.18	**1.35**	6.31	6.74	5.74	**3.78**	4.87	1.22
Dibeklioğlu *et al.* [16]	Seq. (all frames)	4.81 (± 4.87)	5.96	2.73	2.99	**5.45**	6.83	**4.35**	8.45	10.87	13.18
Number of samples	-	1 240	-	158	333	215	171	250	66	30	17

TABLE VI
MAE IN THE ESTIMATION OF EXACT AGE IN EURECOM DATASET

Classifier	Strategy	MAE (years) Overall	G-Mean	25	26	27	28	29	30	31	32	33	36	38
MobileNet (Ours)	Frames	7.19	8.36	6.31	6.90	6.22	6.42	7.60	8.05	12.85	9.04	7.05	13.20	12.20
	Seq. (all frames)	5.56	6.86	4.22	5.86	3.75	4.64	9.83	5.70	10.50	8.17	5.50	14.00	10.00
	Seq. random 5	6.13	6.91	3.57	6.14	4.71	6.77	6.33	7.30	14.00	9.00	**5.00**	**10.50**	**8.00**
	Seq. random 10	5.45	6.74	2.50	5.07	4.21	4.45	7.17	8.10	13.50	8.33	5.50	14.50	10.50
	Seq. quality 5	4.36	**5.56**	2.50	4.07	3.33	3.95	**5.83**	**4.10**	**9.00**	**6.67**	5.50	13.50	11.00
	Seq. quality 10	**4.21**	5.57	**2.43**	**3.50**	**3.25**	**3.18**	6.17	4.50	11.00	7.00	5.50	14.00	11.00
DEX [30]	Frames	9.58	10.01	9.24	8.53	9.00	9.82	9.54	10.50	12.57	11.10	10.74	11.26	8.59
	Seq. (all frames)	8.87	8.23	6.64	9.07	7.42	9.73	7.33	10.2	12.00	12.67	14.00	17.00	**1.00**
	Seq. random 5	7.93	8.73	7.86	**6.64**	6.54	8.59	7.50	8.70	12.00	10.00	8.50	10.00	11.50
	Seq. random 10	8.91	9.49	**6.36**	7.86	7.21	10.45	10.50	10.20	12.00	12.00	14.00	12.50	5.50
	Seq. quality 5	**5.78**	**4.11**	7.14	7.57	5.29	**6.59**	8.33	**2.80**	**6.50**	**3.33**	**2.00**	3.00	**1.00**
	Seq. quality 10	6.59	4.57	7.21	7.50	**4.38**	9.68	**5.83**	6.60	7.00	5.83	**2.00**	**2.50**	**1.00**

It is important to notice that in video applications where an online prediction is desired, it is not possible to use the "Sequence (all frames)" strategy, because a response is required before having the entire sequence of a subject. This provides additional importance to the frame selection strategy that should be used in this type of scenarios.

One of our main interests when we introduced the quality strategy in the context of videos, was to show that the decrease in the number of frames to be classified was also coupled to a decrease in the processing time. With the aim to corroborate that the selection and further classification of discriminant frames is less consuming than a standard strategy, we measured the average time of evaluating the 12 quality measures, the average time of a frame to be classified with our RF attribute-specific quality estimator, and the average prediction time of our MobileNet classifiers. Our experiments were conducted in Windows, in a PC with 8GB of RAM, I7-4770 Intel processor to a maximum speed of 3.40GHZ, with 4 cores and 8 logical processors. Results showed a value of 20 milliseconds (ms) for using the overall quality procedure over a frame and 45 ms for the MobileNet prediction. Therefore, if we consider that the amount of 10 best quality frames is required to accurate represent a video sequence, we could define the entire time (ms) of classification by using quality as $20 \cdot F + 45 \cdot 10$, and the entire time (ms) of classification without using quality as $45 \cdot F$, with F being the total number of frames in the sequence. In this case, our quality approach by using 10 frames will be effective for videos that satisfy the inequality: $20 \cdot F + 45 \cdot 10 \leq 45 \cdot F$, which means that our strategy is suitable if we need to classify 18 or more video frames per sequence (since $18 \leq F$), or a minimum of 9 frames if we substitute 10 by 5 discriminant frames.

Another interesting analysis is the relation of quality problems with the correct classification of soft biometrics in a given image or frame. Using the variable importance measure, proposed by Breiman for the RF classifier [25], we could see that the Illumination and Color Leveling quality scores were among the most relevant features for selecting good frames for classification. In the same way, Mouth state (Open/Close), Eyes state (Open/Close) and Gaze scores were the least important features, which might be an indication that, in terms of quality, environmental conditions could affect more than facial expressions in the estimation of these demographic attributes.

V. CONCLUSIONS AND FUTURE WORK

In this paper we addressed the problem of classifying demographic soft-biometrics in face videos, specifically gender, ethnicity and age. We proposed to take into account several quality issues that affect video sequences, with the hypothesis that given the benefit of information redundancy provided by several frames of the same individual, it is possible to select some single frames with less quality problems to improve the video sequence classification.

Our experiments, conducted in two datasets and with two different classifiers, showed that using this quality assessment

was useful for improving the overall sequence prediction in all cases and, in several cases, the minority classes showed a larger improvement over the others. This might be an indication that an appropriate frame selection can mitigate the biases of specific gender, ethnicity and age classification.

As future work, we plan to conduct more extensive experiments and to explore other video frame combinations beyond majority voting, that may also benefit by these quality selection strategy. Also, a deeper analysis regarding the quality problems that affect each specific soft biometric modality would be interesting.

ACKNOWLEDGMENT

This research work has been partially supported by a grant from the European Commission (H2020 MSCA RISE 690907 IDENTITY).

REFERENCES

[1] U. Park and A. K. Jain, "Face matching and retrieval using soft biometrics," *Trans. Info. For. Sec.*, vol. 5, no. 3, pp. 406–415, Sep. 2010.

[2] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon, "Soft biometrics and their application in person recognition at a distance," *Trans. Info. For. Sec.*, vol. 9, no. 3, pp. 464–475, Mar. 2014.

[3] G. Azzopardi, A. Greco, A. Saggese, and M. Vento, "Fast gender recognition in videos using a novel descriptor based on the gradient magnitudes of facial landmarks." in *Advanced Video and Signal Based Surveillance*, 2017, pp. 1–6.

[4] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[5] P. Carcagnì, M. D. Coco, D. Cazzato, M. Leo, and C. Distante, "A study on different experimental configurations for age, race, and gender estimation problems," *EURASIP Journal on Image and Video Processing*, vol. 2015, pp. 1–22, 2015.

[6] R. Gupta, S. Kumar, P. Yadav, and S. Shrivastava, "Identification of age, gender and race smt (scare, marks, tattoos) from unconstrained facial images using statistical techniques," *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pp. 1–8, 2018.

[7] M. Afifi and A. Abdelhamed, "Afif4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 77 – 86, 2019.

[8] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Effective training of convolutional neural networks for face-based gender and age prediction," *Pattern Recognition (PR)*, vol. 72, no. C, pp. 15–26, 2017.

[9] S. E. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, and A. Hadid, "Pyramid multi-level features for facial demographic estimation," *Expert Systems with Applications*, vol. 80, no. C, pp. 297–310, 2017.

[10] T. Vo, T. Nguyen, and C. T. Le, "Race recognition using deep convolutional neural networks," *Symmetry*, vol. 10, no. 11, p. 564, 2018.

[11] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "Ssr-net: A compact soft stagewise regression network for age estimation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 7 2018, pp. 1078–1084.

[12] P. Terhrst, M. Huber, J. N. Kolf, I. Zelch, N. Damer, F. Kirchbuchner, and A. Kuijper, "Reliable age and gender estimation from face images: Stating the confidence of model predictions," in *10th International Conference on BTAS*. IEEE, 10 2019.

[13] F. Becerra-Riera, A. Morales-González, and H. Méndez-Vázquez, "A survey on facial soft biometrics for video surveillance and forensic applications," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1155–1187, Aug 2019.

[14] A. Hadid and M. Pietikinen, "Demographic classification from face videos using manifold learning," *Neurocomputing*, vol. 100, pp. 197 – 205, 2013, special issue: Behaviours in video.

[15] M. Demirkus, K. Garg, and S. Guler, "Automated person categorization for video surveillance using soft biometrics," in *Biometric Technology for Human Identification VII*, vol. 7667. SPIE, 2010, pp. 236 – 247.

[16] H. Dibeklioğlu, T. Gevers, A. A. Salah, and R. Valenti, "A smile can reveal your age: Enabling facial dynamics in age estimation," in *Proceedings of the 20th ACM International Conference on Multimedia*, ser. MM 12. New York, NY, USA: Association for Computing Machinery, 2012, p. 209218.

[17] A. Dantcheva and F. Brémond, "Gender estimation based on smile-dynamics," *IEEE Trans. Information Forensics and Security*, vol. 12, no. 3, pp. 719–729, 2017.

[18] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, "Hierarchical spatio-temporal probabilistic graphical model with multiple feature fusion for binary facial attribute classification in real-world face videos," *IEEE TPAMI*, vol. 38, pp. 1185–1203, 2016.

[19] W.-C. Wang, R.-Y. Hsu, C.-R. Huang, and L.-Y. Syu, "Video gender recognition using temporal coherent face descriptor," in *16th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing, SNPD*. IEE, 2015, pp. 113–118.

[20] F. Becerra-Riera, A. Morales-González, and H. Méndez-Vázquez, "Exploring local deep representations for facial gender classification in videos," in *International Workshop on Artificial Intelligence and Pattern Recognition (IWAIPR)*, ser. Lecture Notes in Computer Science, vol. 11047. Springer, 2018, pp. 104–112.

[21] M. Selim, S. Sundararajan, A. Pagani, and D. Stricker, "Image quality-aware deep networks ensemble for efficient gender recognition in the wild," in *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018)*, vol. 5, 2018, pp. 351–358.

[22] M. Demirkus, D. Precup, J. J. Clark, and T. Arbel, "Hierarchical temporal graphical model for head pose estimation and subsequent attribute classification in real-world videos," *Computer Vision and Image Understanding*, vol. 136, pp. 128–145, 2015.

[23] H. Méndez-Vázquez, L. Chang, D. Rizo-Rodríguez, and A. Morales-González, "Evaluación de la calidad de las imágenes de rostros utilizadas para la identificación de las personas," *Computación y Sistemas*, vol. 16, pp. 147 – 165, 06 2012.

[24] "Face Recognition Format Data Interchange, Version 2.0," InterNational Committee for Information Technology Standards (INCITS) Secretariat. Information Technology Industry Council, Standard, 2006.

[25] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[26] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 07 2014.

[27] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity." in *CVPR*. IEEE Computer Society, 2011, pp. 529–534.

[28] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[29] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE TPAMI*, vol. 33, no. 10, pp. 1962–1977, Oct 2011.

[30] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision (IJCV)*, 2016.

[31] S. Y. Zhang, Zhifei and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[32] P. T. Bilinski, A. Dantcheva, and F. Brémond, "Can a smile reveal your gender?" in *2016 International Conference of the BIOSIG*, ser. LNI, vol. P-260. GI / IEEE, 2016, pp. 27–38.