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), Havana, Cuba
{fbecerra,amorales,hmendez}@cenatav.co.cu

Jean-Luc Dugelay
Digital Security Department, EURECOM, France
{Jean-Luc.Dugelay@eurecom.fr}

1 Introduction

Demographic soft biometrics (e.g. gender, age, ethnicity) are among the most frequently used traits for improving and complementing the performance of biometric systems [1].

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 a pipeline for the automatic estimation of gender, ethnicity and age in videos.

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 [2] and a final video prediction is obtained with a majority voting strategy.

2 Proposal

Quality Assessment

We associate the relevance of a frame within a sequence to 12 quality parameters relatives to Pose, Illumination, Occlusion, Resolution, Sharpness, Mouth State, Eyes State, Gaze, Color Leveling, Face Centering, Red Eyes and Uniform Background.

Some quality measures could have more or less impact over the relevance of a frame, depending on the specific demographic attribute to classify.

We employed Random Forest (RF) classifiers [3] to learn the relations between the quality measures and each classification task (gender, age, ethnicity).

The output score of the 12 quality measures were concatenated and the resulting features for good and bad classification samples were used to train each RF quality classifier.

Final video prediction was obtained with a majority voting strategy among best quality frames selected by the RF classifier.

Demographic Estimators

We used fine-tuned MobileNet models to make the real-time demographic estimation of the selected video frames as efficient as possible.

Training was performed on three publicly uncontrolled image datasets: IMDB-Wiki Dataset, UTKFace Dataset and LFW.

Formalization

Given a sequence of video frames $F = \{f_1, f_2, \ldots, f_m\}$ and an attribute $\alpha$ with $L = \{l_1, l_2, \ldots, l_k\}$ 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 $q^n_k \subseteq q^n$ as the set of the $f \in q^n$ for which the classification according to $\alpha$ (from now on defined as $C_\alpha$) corresponds to the $l_k$ label:

$$ q^n_k = \{ f \in q^n | C_\alpha(f) = l_k \} \quad (1) $$

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

$$ l_i \in L \quad (1 \leq i \leq k) \quad \forall l_j \in L \quad (1 \leq j \leq k), j \neq i, |q^n_i| < |q^n_j| \quad (2) $$
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), Havana, Cuba
{fbecerra,amorales,hmendez}@cenatav.co.cu

Jean-Luc Dugelay
Digital Security Department, EURECOM, France
{Jean-Luc.Dugelay@eurecom.fr}

3 Experiments

We performed experiments in the selected datasets by comparing several frame combination strategies:
- Individual frames: Considers frames as single independent images.
- Sequence (all frames): Performs a majority voting among all frames in a sequence.
- Sequence - quality \( N \) frames: Performs the majority voting on the \( N \) top relevant frames.
- Sequence - random \( N \) frames: Performs the majority voting on \( N \) random frames.

Dataset Selection

1. **UvA-Nemo**:
   Fairly good quality video collection with gender and age annotations, created to analyze
   the change in smile dynamics across different ages.

2. **EURECOM Augmented**:
   Dataset representing frames of a video, fully annotated with demographic data and augmented
   with added noise to simulate a video scenario with several frames of different qualities.

3. **Youtube Faces (YTF)**:
   Uncontrolled video collection for which we mapped the Labelled Faces in the Wild (LFW) gender
   and ethnicity labels to its corresponding identities.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>Af</td>
</tr>
<tr>
<td>UvA-Nemo</td>
<td>185</td>
<td>215</td>
<td>400</td>
</tr>
<tr>
<td>EURECOM</td>
<td>14</td>
<td>38</td>
<td>20</td>
</tr>
<tr>
<td>YTF</td>
<td>149</td>
<td>285</td>
<td>315</td>
</tr>
</tbody>
</table>
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), Havana, Cuba
{fbecerra,amorales,hmendez}@cenatav.co.cu

Jean-Luc Dugelay
Digital Security Department, EURECOM, France
{Jean-Luc.Dugelay@eurecom.fr}

Poster presentation at the 25th International Conference on Pattern Recognition

Results and Discussion

Gender classification results in the “Deliberate” and “Spontaneous” subsets from UvA-Nemo dataset, and also in the “Entire” collection:

- 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.
- The experiments allows to compare our baseline method to other state-of-the-art algorithms.

- The proposed quality assessment is effective without dependence on the dataset or the classifier.
- 100% of classification accuracy in EURECOM.
- The quality assessment strategy favored the results of the minority class (Females).

Gender classification results in EURECOM and YTF datasets:

- The gender classification results in the “Deliberate” and “Spontaneous” subsets from UvA-Nemo dataset, and also in the “Entire” collection:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Strategy</th>
<th>Deliberate Accuracy (%)</th>
<th>Spontaneous Accuracy (%)</th>
<th>Entire dataset Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet (Ours)</td>
<td>Individual frames</td>
<td>96.27</td>
<td>89.11</td>
<td>83.90</td>
</tr>
<tr>
<td></td>
<td>Sequence (all frames)</td>
<td>96.51</td>
<td>88.84</td>
<td>83.29</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 5 frames</td>
<td>91.28</td>
<td>89.69</td>
<td>84.55</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 10 frames</td>
<td>95.87</td>
<td>88.95</td>
<td>83.29</td>
</tr>
<tr>
<td>DNN [30]</td>
<td>Individual frames</td>
<td>88.75</td>
<td>87.68</td>
<td>83.12</td>
</tr>
<tr>
<td></td>
<td>Sequence (all frames)</td>
<td>89.72</td>
<td>88.50</td>
<td>83.73</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 5 frames</td>
<td>89.72</td>
<td>88.73</td>
<td>83.50</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 10 frames</td>
<td>89.73</td>
<td>87.56</td>
<td>83.00</td>
</tr>
<tr>
<td>Decteda and Rimonel [17]</td>
<td>Sequence (all frames)</td>
<td>-</td>
<td>84.53</td>
<td>76.02</td>
</tr>
</tbody>
</table>

Blumski et al. [32] | Sequence (all frames) | - | - | - | - | - | - | - | - | - | - | - | - |

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Strategy</th>
<th>EURECOM Accuracy (%)</th>
<th>YTF Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet (Ours)</td>
<td>Individual frames</td>
<td>76.18</td>
<td>81.07</td>
</tr>
<tr>
<td></td>
<td>Sequence (all frames)</td>
<td>91.35</td>
<td>92.88</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 5 frames</td>
<td>86.82</td>
<td>88.96</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 10 frames</td>
<td>87.59</td>
<td>91.01</td>
</tr>
<tr>
<td></td>
<td>Sequence quality 5 frames</td>
<td>90.00</td>
<td>93.34</td>
</tr>
<tr>
<td></td>
<td>Sequence quality 10 frames</td>
<td>90.04</td>
<td>93.36</td>
</tr>
<tr>
<td>DNN [30]</td>
<td>Individual frames</td>
<td>80.35</td>
<td>77.81</td>
</tr>
<tr>
<td></td>
<td>Sequence (all frames)</td>
<td>92.31</td>
<td>84.52</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 5 frames</td>
<td>91.35</td>
<td>82.86</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 10 frames</td>
<td>84.52</td>
<td>71.43</td>
</tr>
<tr>
<td></td>
<td>Sequence quality 5 frames</td>
<td>95.04</td>
<td>98.20</td>
</tr>
<tr>
<td></td>
<td>Sequence quality 10 frames</td>
<td>98.08</td>
<td>96.36</td>
</tr>
</tbody>
</table>
### Results and Discussion

- Ethnicity classification results in EURECOM and YTF datasets:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Strategy</th>
<th>EURECOM Accuracy (%)</th>
<th>YTF Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>G-Mean</td>
</tr>
<tr>
<td>MobileNet (Ours)</td>
<td>Individual frames</td>
<td>63.57</td>
<td>68.84</td>
</tr>
<tr>
<td></td>
<td>Sequence (all frames)</td>
<td>85.58</td>
<td>89.25</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 5 frames</td>
<td>75.00</td>
<td>79.22</td>
</tr>
<tr>
<td></td>
<td>Sequence - random 10 frames</td>
<td>79.81</td>
<td>84.17</td>
</tr>
<tr>
<td></td>
<td>Sequence - quality 5 frames</td>
<td>99.04</td>
<td>99.37</td>
</tr>
<tr>
<td></td>
<td>Sequence - quality 10 frames</td>
<td>97.12</td>
<td>98.07</td>
</tr>
</tbody>
</table>

- We were not able to find an available state-of-the-art pre-trained model for this task.
- By using the quality strategy, in the EURECOM dataset:
  - The minority classes Asian and Other achieved 100% of accuracy, showing great improvements.
  - The majority class Caucasian also was largely favored.
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), Havana, Cuba
{fbecerra,amorales,hmendez}@cenatav.co.cu

Jean-Luc Dugelay
Digital Security Department, EURECOM, France
{Jean-Luc.Dugelay@eurecom.fr}

Poster presentation at the 25th International Conference on Pattern Recognition

Results and Discussion

- MAE in the estimation of exact age in UvA-Nemo dataset:
  - For 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
  - For DEX [30]:
    - Frames: 4.13 (± 0.88) | 3.19 | 1.33 | 1.43 | 6.21 | 6.98 | 5.89 | 5.99 | 4.76 | 1.15
    - Seq (all frames): 4.09 (± 1.03) | 3.02 | 1.23 | 1.37 | 6.18 | 6.74 | 5.69 | 5.91 | 4.69 | 0.95
    - Seq random 5: 4.18 (± 0.98) | 3.23 | 1.22 | 1.41 | 6.13 | 7.23 | 5.82 | 5.94 | 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
  - Number of samples:
    - MobileNet: 1,240
    - DEX: 158
    - DEX [30]: 158
    - Our MobileNet: 2,175
    - DEX: 250
    - DEX [30]: 66

- MAE in the estimation of exact age in EURECOM dataset:
  - For MobileNet (Ours):
    - Frames: 7.19 | 8.36 | 6.31 | 6.22 | 6.42 | 7.60 | 8.05 | 12.85 | 9.04 | 7.05 | 13.20 | 12.20
    - Seq (all frames): 5.56 | 8.86 | 4.22 | 5.86 | 3.75 | 6.64 | 9.83 | 5.70 | 10.80 | 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 | 2.90 | 5.07 | 4.21 | 4.45 | 7.17 | 8.10 | 13.50 | 8.33 | 5.50 | 14.50 | 10.50
    - Seq quality 5: 4.75 | 5.56 | 2.33 | 4.07 | 3.33 | 3.95 | 6.87 | 6.67 | 13.50 | 6.67 | 5.50 | 13.50 | 11.00
    - Seq quality 10: 4.21 | 5.57 | 2.43 | 3.25 | 3.18 | 6.17 | 4.50 | 11.00 | 7.00 | 5.50 | 14.00 | 11.00
  - For 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): 6.87 | 8.23 | 6.64 | 9.07 | 7.42 | 9.73 | 7.33 | 10.2 | 12.00 | 12.07 | 14.00 | 17.00 | 11.00
    - Seq random 5: 7.93 | 8.73 | 7.86 | 6.64 | 6.54 | 8.50 | 7.50 | 8.70 | 12.00 | 10.00 | 8.50 | 15.00 | 11.50
    - Seq random 10: 8.91 | 9.40 | 6.36 | 7.86 | 7.21 | 10.45 | 10.50 | 10.20 | 12.00 | 12.00 | 14.00 | 12.00 | 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 | 4.37 | 7.21 | 7.50 | 4.38 | 6.68 | 5.83 | 6.60 | 7.00 | 5.83 | 2.00 | 2.50 | 1.00

- We were not able to train an age quality estimator for this collection, as explained before.
- DEX classifier was slightly more accurate for age groups over 50 years old; however, DEX is not suitable for real-time video applications due to its larger processing time.
- DEX classifier showed better performance in the classification of people over 30 years old; our MobileNet was more accurate in the classification of the younger.
4 Conclusions
The quality strategy works with different classifiers and under different conditions, allowing:
✓ Less number of frames to be classified.
✓ Less processing time.
✓ Improved estimation accuracy.
✓ Bias mitigation on specific gender, ethnicity and age.

5 Future Work
- Explore other video frame combination beyond majority voting.
- Deeper analysis regarding quality problems affecting specific demographic attributes.

References