Ensembling of local and global CNNs for apparent age estimation

September 14, 2015

1 Team details

- **Team name**: Enjuto
- **Lead**: Cristian Canton Ferrer
- **Address**: One Microsoft Way, Redmond, WA
- **Phone**: (+1) 206 458 9148
- **Email**: cristian.canton@microsoft.com
- **Affiliation**: Microsoft Corporation (Applied Vision and Imaging group)

2 Contribution details

- **Title of the contribution**: Ensembling of local and global CNNs for apparent age estimation
- **Final score**: TBD by the organizers
- **General method description**: Given several different convolutional neural networks (CNN) architectures trained to estimate age, we fine-tuned them to estimate the apparent age using the data provided by the organizers. These CNNs have been grouped into two families: a first group where faces are weakly aligned using rotation and scale and the final image is directly ingested by a CNN, hence accounting for the global appearance of the face; and, a second one, where each part of the face (eyes, mouth, nose and forehead/hair) is locally aligned, crop and feeded into a multi-input CNN. The output of the local and global and fine-tuned and not fine-tuned CNNs is combined using a linear model to produce the final apparent age estimation.

3 Face Detection Stage

The face detection stage of this system relied on the combination of two state-of-the-art methods. First, the algorithm presented by [5] relying on an integral channel detector together with rigid templates; the second face detection algorithm is the one present in the DLIB library [4] which is based on the classic Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid, and sliding window detection scheme. These two algorithms have been selected in order to fulfill the constrain imposed on this competition stating that all used code must be made public. For both face detection components, the respective authors have provided source code available online.

3.1 Face Detections Combination

The set of detections provided by the two aforementioned face detection algorithms are combined using a simple heuristic that consists in:

1. Merge the overlapping detections when there is over 90% of overlap.
2. Select the largest detected face out of the merged set

3.2 Method complexity

Complexity of the face detection stage is the direct sum of the complexities of each of the employed face detectors.

4 Face Landmarks Detection Stage

Once the a face is detected on the image, another state-of-the-art method for landmark detection is applied [3]. Again, given the constrain of this competition to provide source code for the end-to-end system, we opted for this method. This particular method proved fast and robust when analyzing unconstrained images. In particular, this method uses an ensemble of regression trees to estimate the face's landmark positions directly from a sparse subset of pixel intensities.

4.1 Method complexity

This particular method is known to be extremely efficient and fast enough to deliver real-time performance.

5 Global Method Description

- **Total method complexity**: The proposed method relies on the ensemble of of the output of several CNN models (6 models in the final implementation). All the employed CNN models have a complexity comparable to the standard AlexNet architecture. If we add up all the complexities for each of the stages (face detection, face alignment, CNN evaluation), the system achieves a speed of 0̃.75 fps (more accurate measurements to be provided in the final paper version).

- **Which pre-trained or external methods have been used**: Several CNN models have been trained to estimate real age directly from images. In order to train them, we employed several online available datasets (Adience [2], Cross-Age Celebrity Dataset [1], FG-NET Aging Dataset) plus another non-public dataset automatically gathered from the web[1]. All data amounted to 2̃00K face images with its corresponent age. This pretrained models where the essential building blocks to the final fine-tuned models.

- **Which additional data has been used in addition to the provided ChaLearn training and validation data**: Already discussed in the previous point.

- **Qualitative advantages of the proposed solution**: Two main advantages: first, the usage of a large amount of data to train the real age CNN models and, second, the ensembling of local and global face aligned CNN outputs.

- **Novelty degree of the solution and if is has been previously published**: This end-to-end system has not been published before.

6 Other details

- **Language and implementation details**: C++ code has been the used for the face detection and alignment and CNNs processing (Caffe). The rest of the system architecture has been developed in Python. The whole system runs a Linux machine; the memory footprint is small (1GB) but the system requires a high-end GPU, in our case one NVIDIA Titan X.

- **Human effort required for implementation, training and validation**: Code development took 4 weeks. Human effort for training and validation is none since there is no step in the processing chain that requires human supervision. This system has been architected to run unsupervised when train/validation data are provided at input.

- **Training/testing expended time**: Fine tuning of CNN models requires around 45 minutes per model (4̃.5 hours). Running the end-to-end system for the test dataset requires 4̃5 minutes.

[1] More details on this private dataset and instructions on how to generate it will be disclosed in the final paper version.

- **General comments and impressions of the challenge? what do you expect -from a new challenge in face and looking at people analysis?** This challenge is interesting since it attempts to delve into the problem of apparent age, instead of real age. There are several psychological factors that are involved in the perception of age and the proposed dataset provide interesting data to further understand these factors. Something we found missing in the competition would be to also provide the real age of subjects in the images; that would had allowed to really study the mapping between real age and perceived age in more detail.

References

[1] B.C. Chen, C.S. Chen, and W.H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, 2014.

[2] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9:12:2170–2179, 2014.

[3] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[4] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[5] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014.