# Cross-modal, Cross-domain Biometric Systems

Ali Akbari, Ammarah Farooq, Syed Safwan Khalid, Chiho Chan, Junaid Awan, Zhenhua Feng,
Tiangyang Xu, Soroush Fatemifar, Lei Ju, Muhammad Awais, Josef Kittler
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK

***Abstract*—We demonstrate three biometric systems, including cross-domain age estimation, cross-resolution face recognition, and cross-modal person re-identification (ReID). Recently, various facial age estimation systems have been developed and deployed. However, their performance tends to degrade in unseen scenarios. We present a robust age estimation system, showing better performance in real-world scenarios. For cross-resolution face recognition, we use a distillation based approach to train deep neural networks. Our proposed approach outperforms many state-of-the-art solutions for low-resolution and cross-resolution face recognition tasks. For cross-modal ReID we learn a joint embedding space for vision and language to perform cross-modal search. Interestingly, the multimodal approach significantly enhances the performance of the vision modality.**

## I. Cross-Domain Age Estimation Demo

With the emergence of Deep Neural Networks (DNN), various facial age estimation systems have been proposed. However, the performance of these methods has been shown to degrade in practical applications, where the system needs to perform well on any arbitrary input image. In fact, there are many factors which affect the estimation performance, such as gender, race, illumination conditions, image quality, makeup, lifestyle, and cosmetic surgery [3]. In existing ageing datasets, some of these factors are well represented, and consequently the models learnt using such training data are likely to be biased, and generalise poorly to unseen data.

There are three approaches to addressing this problem.

The ideal solution is to collect a balanced dataset, representing all these factors. However, the cost of such an exercise is prohibitive, and only marginal improvement can be made by rebalancing existing datasets.

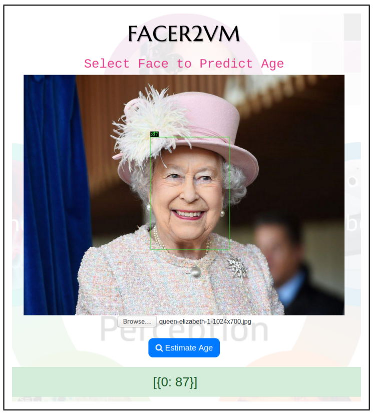
The lack of generalisation is also aggravated by the choice of a learning algorithm adopted for the age estimation system design  [1]. Finally, the generalisation problem can be mitigated by minimising any cohort bias by bias correction or reformulating the learning problem.

We will demonstrate an age estimation system that adopts two of the above three strategies to improve age estimation performance. The dataset shortcomings are mitigated by a targeted augmentation of the training data to fill the most glaring deficiencies in its characetristics. By collecting samples to complement unrepresentative categories we create a Balanced AGeing dataset (BAG), as described in our paper [2]

The second measure is to train a DNN using a robust loss function that is shown to exhibit enhanced generalisation properties. The de-biasing work is planned for our future research.

### A. The System Design

We model the age estimation problem as the label distribution learning framework [7], by which the semantic correlation among the face images of nearby ages are emphasised during training stage. Then, the proposed loss function in [2] and BAG dataset are utilised for training the VGG model [9]. The input of the network is constituted by facial images with resolution $224 \times 224$ pixels. Optimisation is done by stochastic gradient descent using mini-batches of size 80. The momentum and weight decay coefficients are set to 0.9 and 0.0005, respectively. The learning rate decreases exponentially (with the exponential growth $-1$) for 30 epochs from 0.001 to 0.00001. All faces are automatically detected and aligned with respect to five facial landmarks (eyes centre, nose tip and mouth corners) extracted by the RetinaFace face detector [4] and the Rectified Wing Loss [6]. The face image is then extracted and resized to $256 \times 256$ pixels. Standard data augmentation techniques (random flipping, cropping and colour jittering) are employed during the training phase. At the inference step, the central-cropped image is used as the input of the network and the subject's age is predicted by taking the most probable value of the network's output, as $\hat{y} = \operatorname{argmax}_i p_i$.



queen-elizabeth-1-1024x700.jpg

[{0: 87}]

Fig. 1. Age estimation demo interface.

### B. Demo

In the demo session, we will show an interactive demo and demo video of the system. The video demo shows a recorded video of people in the age range from 0-100 years old. The interactive demo has been developed as a visitor interface, where a visitor looks at the camera and the predicted age appears on the screen. The subject should be looking directly into the camera within a distance less than 2m from camera. An illustration of the demo interface is shown in Fig. 1.

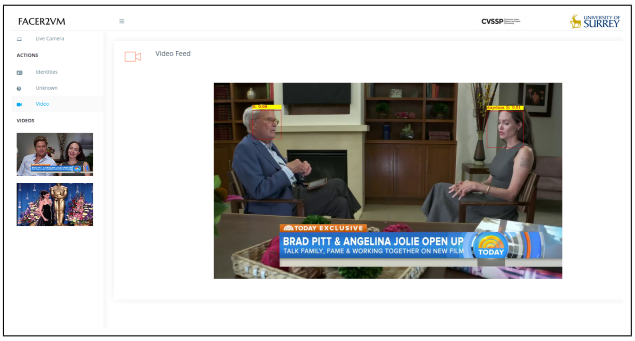


Fig. 2. Cross-resolution face recognition demo interface.

## II. Cross-Resolution Face Recognition Demo

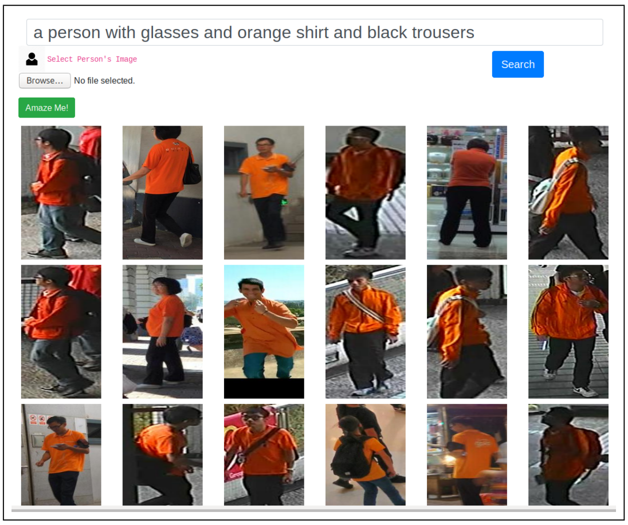
Face recognition systems use DNNs to obtain facial representations/embeddings. These modern solutions give excellent performance in terms of verification and identification when tested on high resolution (HR) facial images. However, when tested on low resolution (LR) facial images these systems suffer from significant degradation in performance. A naive solution is to train a DNN using a mixture of LR and HR images. Although this kind of approach will give better results on low resolution, however, it would suffer from performance loss at high resolution. The demo is built upon our paper [8] on resolution invariant face recognition. The main idea is to train a network using both low and high resolution images under the guidance of a fixed network, pretrained on high resolution face images. The DNN is trained by minimising the KL-divergence between Softmax probabilities of the pre-trained and the trainable network. The pre-trained model acts as a "teacher" while the trainable model acts as a "student" network. The weights of the Softmax layer of the teacher and student networks are also shared. Our approach sets a new state-of-the-art across various resolutions on multiple datasets including FaceScrub, MegaFace, TinyFace and SCFace.

### A. Demo

The cross-resolution face recognition system is trained by using loss functions proposed in [8]. The pipeline starts by applying a face and landmark detector [4]. The detected face is aligned, cropped and resized to 112 × 112 and used as an input to the DNN model. The DNN model computes the face embeddings and compare it against a gallery. In the demo session, we will show an interactive real time on a recorded video or live camera feed. An illustration of the demo interface is shown in Fig. 2.

## III. Cross-Modal Person ReID Demo

Conventional person re-identification (ReID) approaches work on the assumption that given an image of a person of interest, the person can be found across different cameras. However, in many cases this assumption is not valid. For example, for a crime there could be multiple witnesses but no visual information. For a missing person there may not be any picture taken on the particular day. In such cases the only source of information about the person of interest is the descriptions provided by the witnesses. This demo deals with all cross-modal, multi-modal as well as intra-model person ReID scenarios using vision and language. This is achieved by learning a joint DNN embedding space for both vision and language as presented in our paper [5]. To learn the joint embedding space, a two stream DNN supervised by cross entropy loss with shared weights for softmax layer is used. To enhance the correlation between two modalities in the joint embedding space Canonical Correlation Analysis (CCA) is performed. This approach set a new state-of-the-art for cross-modal, multi-modal person ReID sceanrios as demonstrated on CUHK-PEDES and CUHK-SYSU datasets.


a person with glasses and orange shirt and black trousers

No file selected.

Fig. 3. Cross-Modal Person ReID demo interface.

### A. Demo

The cross-modal person ReID system is trained by using loss functions proposed in [5]. The demo have multiple option to query person ReID database. The query can be an image of a person of interest like in conventional person ReID (intra-modal setting), or it can be a natural language description given by the user at run time (cross-modal setting) or it can be both image of a person of interest and corresponding description of person (multi-modal setting). Given an image or a description or both corresponding to a person of interest the system computes the embedding in the joint space, performs CCA and does the matching with the gallery. An illustration of the demo interface is shown in Fig. 3.

## IV. Conclusion

In the proposed demo, we show three novel DNN based biometric systems, namely, cross-domain age estimation, cross-resolution face recognition, and cross-modal person ReID. Our systems are deployable in real-world applications.

## Acknowledgment

This work was supported in part by the EPSRC Programme Grant (FACER2VM) EP/N007743/1 and the EPSRC/dstl/MURI project EP/R018456/1.

## References

[1] A. Akbari, M. Awais, Z. Feng, A. Farooq, and J. Kittler. Distribution cognisant loss for cross-database facial age estimation with sensitivity analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[2] A. Akbari, M. Awais, Z. Feng, A. Farooq, and J. Kittler. A flatter loss for bias mitigation in cross-dataset facial age estimation. In *International Conference on Pattern Recognition (ICPR)*, 2021.

[3] A. Akbari, M. Awais, and J. Kittler. Sensitivity of age estimation systems to demographic factors and image quality: Achievements and challenges. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6, 2020.

[4] 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*, pages 5203–5212, 2020.

[5] A. Farooq, M. Awais, J. Kittler, A. Akbari, and S. S. Khalid. Cross modal person re-identification with visual-textual queries. In *Proceedings of the International Joint Conference on Biometrics*, 2020.

[6] Z.-H. Feng, J. Kittler, M. Awais, and X.-J. Wu. Rectified wing loss for efficient and robust facial landmark localisation with convolutional neural networks. *International Journal of Computer Vision*, pages 1–20, 2019.

[7] X. Geng, C. Yin, and Z. Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, Oct 2013.

[8] S. S. Khalid, M. Awais, Z.-H. Feng, C.-H. Chan, A. Farooq, A. Akbari, and J. Kittler. Resolution invariant face recognition using a distillation approach. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):410–420, 2020.

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.