November 20, 2018

A DATA AUGMENTATION AND PREPROCESSING METHOD FOR ROBUST FACIAL LANDMARK DETECTION WITH DIFFERENT FACE DETECTORS

HP INC

A Data Augmentation and Preprocessing Method for Robust Facial Landmark Detection with Different Face Detectors

Abstract: A technique for training data augmentation and preprocessing that trains the network without any specific face detector and makes the trained landmark detector robust to input face bounding boxes generated by different face detectors during testing. During testing time, the network can be cascaded with any arbitrary face detector which can output a reasonable face bounding box and still output good facial landmark estimation.

This disclosure relates to the field of facial recognition.

Facial landmark detection is critical in face alignment, face recognition, emotion recognition, 3D face model reconstruction and many other applications. Usually, facial landmark detection is implemented within the input face bounding box generated by a preceding face detector. As a result, the landmark detector needs to be trained and test with the particular face detector in order to get an accurate estimation. Degraded performance of a facial landmark detector results if the face detector used during testing varies from the one used in the training stage.

A technique for training data augmentation and preprocessing is disclosed that trains the network without any specific face detector and makes the trained landmark detector robust to input face bounding boxes generated by different face detector during testing.

According to the present disclosure, and as understood with reference to the Figure, the disclosed data augmentation and preprocessing method for landmark detection eliminates the need for the landmark detection networks to be trained with a specific face detector and allows them to be cascaded with different face detectors during testing time.

In the disclosed technique, an initial face bounding box is generated based on the ground truth landmarks. The initial rectangular face bounding box can be represented as (x_min, y_min, x_max, y_max), in which x_min, x_max are the ground truth landmark extremities along x axis of the image and y_min, y_max are the ground truth landmark extremities along y axis of the image. The rectangular face bounding box should always be within the original image for both the initial face bounding box and augmented preprocessed face bounding box.

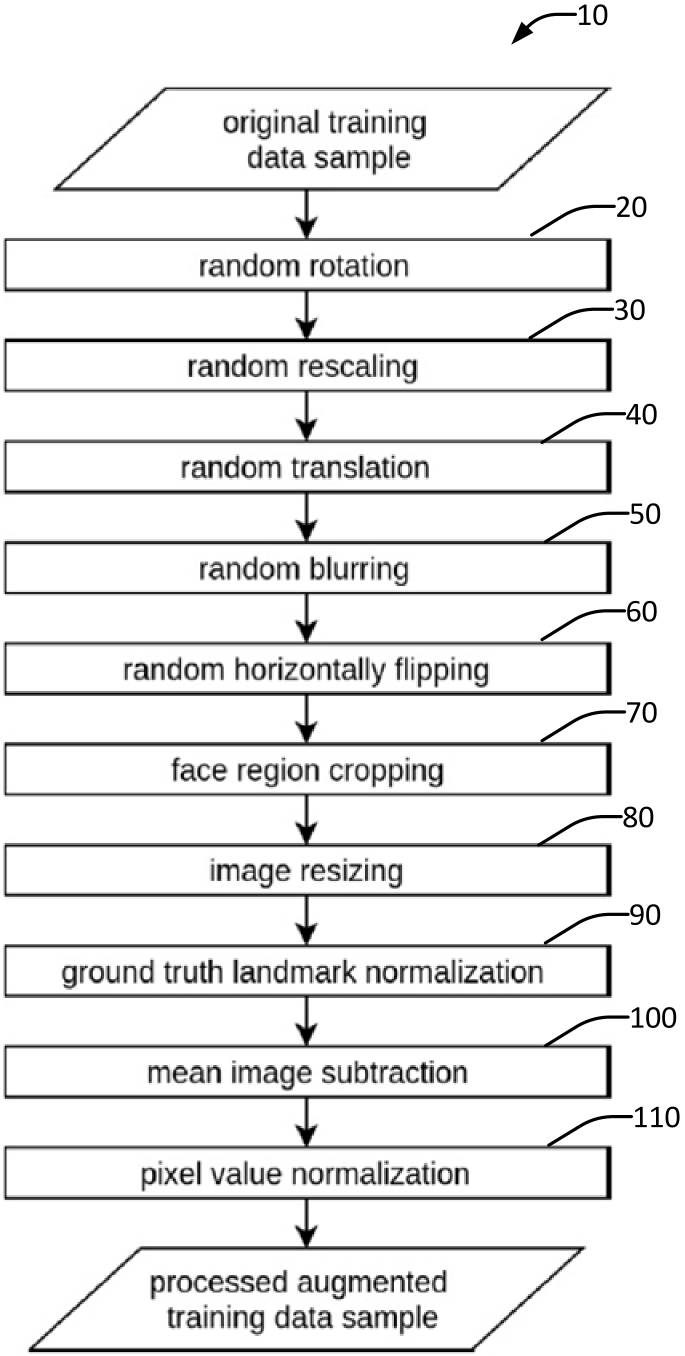
Each original training data sample is one sample input, and one augmented preprocessed training data sample is generated by the process 10. An augmentation number is predefined and the process 10 is implemented on each original training data sample multiple times so that the number of generated data samples equals the augmentation number. This process 10 virtually generates an infinite number of different training data samples, since the augmentation number can be set arbitrarily. Since some data samples including extreme poses and unusual expressions are more valuable for the training of the network, it can be useful to perform larger augmentation on these data samples. To do this, different augmentation numbers can be set on different original data sample sets; for example, larger augmentation numbers on the datasets that are more valuable for the network training.

The process 10 begins with a random rotation operation 20 in which the training image and corresponding ground truth landmarks are randomly rotated within a predefined range, e.g. (-max_degree, max_degree). At 30, random rescaling is performed in which the initial face bounding box is randomly expanded within a predefined range, e.g. (min_exp, max_exp), while keeping the bounding box center constant. Each side of the bounding box is expanded by the same portion of the bounding box width or height respectively so that no distortion of the original image will be made when doing the resizing. At 40, random translation is performed in which the face bounding box is randomly translated along x and y axes while keeping the size constant. While doing the random translation, all the facial landmarks are guaranteed to be within the new translated face bounding box. At 50, random blurring is performed in which the image is randomly Gaussian blurred within a predefined value range. Users can set the portion of processed images that are randomly blurred. At 60, random horizontal flipping is performed in which the training image and corresponding ground truth landmarks are randomly horizontally flipped. At 70, face region cropping is performed in which the face region is cropped according to the current face bounding box. At 80, image resizing is performed in which the cropped face image is resized to the same size of the neural network's input. At 90, ground truth landmark normalization is performed in which the ground truth landmark position values are normalized to (0, 1) according to the current face bounding box. At 100, mean image subtraction is performed in which a mean image is calculated over all augmented preprocessed training face images after resizing. In practice, it is done after obtaining all augmented preprocessed training data. At 110, pixel value normalization is performed in which all the pixel values of the training face images are normalized to be within (-1, 1). For example, in 24-bit RGB color system, all three channels of the training face images are divided by 255.

By using this process 10, the network no longer needs to be trained with a specific face detector during training. And during testing time, the network can be cascaded with any arbitrary face detector which can output a reasonable face bounding box and still output good facial landmark estimation.

The data augmentation process of the disclosed technique advantageously uses a large number of transformations, including scaling, cropping, rotation, blurring, translation, and flipping, to increase the number of total training data and make our models much more robust against variations in face bounding boxes. As a result, the trained landmark detection network can be deployed with different face detectors during testing without retraining the network. In addition, the disclosed training process generates training data directly from landmark points. Landmark detection networks can be trained without a specific face detector during training. This simplifies the training process and also reduces the dependency on the use of a specific face detector.

Disclosed by Ruiyi Mao, Qian Lin, and Jan Allebach, HP Inc.


10
original training data sample
20
random rotation
30
random rescaling
40
random translation
50
random blurring
60
random horizontally flipping
70
face region cropping
80
image resizing
90
ground truth landmark normalization
100
mean image subtraction
110
pixel value normalization
processed augmented training data sample