CNN-based Eye Landmark Estimation

Dana Kianfar
University of Amsterdam
dana.kianfar@student.uva.nl

Jose Gallego
University of Amsterdam
jose.gallegoposada@student.uva.nl

Abstract

Robust facial landmark estimation in-the-wild is a challenging task due to many variations in image conditions, pose and occlusion. Traditional research in landmark estimation uses feature-based methods and the extension of such approaches towards robust performance against large variations is challenging and computationally expensive. In this work we focus on estimating eye landmarks and present a convolutional neural network model for 2D and 3D landmark estimation. Furthermore, we define custom a evaluation metric and objective function which are more suitable for the task of landmark estimation. We show that through data augmentation, the network is able to generalize from synthetic images to real-world images and handle a variety of image, pose and gaze conditions, achieving a 93% 2D landmark estimation accuracy in test data. The quality of the 3D eye shapes reconstructed from estimated landmarks is acceptable despite the lack of depth information.

1. Introduction

Facial landmark estimation is a fundamental building block in computer vision tasks involving facial analysis and has a wide range of applications such as facial recognition, head pose and gaze estimation. Traditional research in landmark estimation has relied on feature-based approaches. Wang et al. [1] provides a review of different research on facial landmark estimation.

More recently, convolutional neural networks (CNNs) have been applied to this task with great performance and robustness. Wu et al. [3] show that features captured from higher layers of a CNN can be directly used as input for fully-connected neural networks in landmark coordinate regression tasks. Zhang et al. [4] propose a CNN architecture with several task-dependent early stopping criterion to alleviate the difficulty of reaching convergence.

We use a similar network architecture as in [4] and apply it on a synthetic dataset for the eye landmark regression task. As is common with deep learning models, they are often difficult to train and provide little diagnostic information. To this end, we define a new accuracy metric and objective function that is specifically suited for the task of landmark coordinate regression.

Although this research only explores eye landmarks, our approach could be extended to full facial landmark estimation. Zhang et al. [4] claim that estimating landmarks from any facial component, such as the mouth, should not be treated as an independent task from other components. Information about certain areas of the face can be useful in estimating landmarks for neighboring regions. Therefore, the performance of landmark estimations for each individual region of a face could potentially be improved when modelled jointly.

This paper is structured as follows: in Section 2 we describe the used dataset as well as our data augmentation process. Section 3 presents the methodology and structure of our solution. In Section 4 we display the results obtained on both synthetic and real data. Section 5 explains several application-related issues of our project. Finally, Section 6 summarizes the main conclusions of our work and provides possible future research directions.

2. Data

Figure 1. Examples images from the Syntheseyes dataset displaying the variations in gaze, pose, illumination conditions and skin features present in the training set.

In this work we use the Syntheseyes [2] dataset which consists of 11,382 synthesized photo-realistic close-up im-

ages of eyes. We perform an 80-20 train/test split and obtain 9105 training and 2277 test examples. The images are generated from 3D sythetic face models under different head poses, gaze directions, illumination conditions and facial features. There are 10 subjects representing both genders, different ethnicities and age groups. A selection of images from this dataset are presented in Figure 1.

Each 80×120 pixel image is labelled with 28 landmarks over the eye lids (12), iris (8) and pupil (8) in 2D and 3D. Figure 3 presents an example of a pre-rendered image with ground-truth landmarks. For the purposes of this research, we ignore other labels such as 3D gaze and head pose. The images in the dataset are rendered from high-quality models and are thus representative of natural images. However, we note that all images are centered on the pupil, of equal distance from the camera, and in high resolution.

Models trained only on the original synthetic dataset were not able to generalize to natural variations in the input image, such as displacement of the eye about the center of the image or lower resolutions. In order to account for such variations, we augment each image-landmark pair in our training set, effectively doubling the size of our dataset to 18210 training and 4554 testing images. Concretely, we perform an affine transformation where we scale, rotate, translate the image-landmark pair. Note that this transformation is only performed for 2D landmarks. For each pair, the rotation angle is randomly selected within $[-10°, +10°]$. Similarly, the translations along the x and y axes are selected randomly within $[-10, 10]$ pixels, and the scaling factor within $[0.9, 1.10]$. To account for different image resolutions, we blur the transformed image with a 5×5 Gaussian filter whose mean is zero and it's variance is randomly chosen within $[0.1, 5]$. Further experimentation on the transformation parameters was out of the scope of this work.

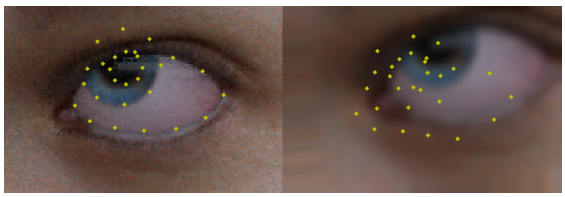

Figure 2. An example of an image from the dataset (left) and its transformed version (right). Landmarks are estimated from a model trained only on the synthetic dataset and are shown in yellow. While the predictions on the left image are accurate, the model was not able to generalize to the transformed image.

An example of an image and its transformed version is shown in Figure 2. Additionally, we displayed the estimated landmarks from a model trained only on the synthetic dataset. We can observe that the model has overfitted to eyes that are centered on the image. Consequently, the predicted landmarks for the transformed image are poor despite having accurate predictions for the original image.

3. Methodology

In order to evaluate the success of a particular model at the landmark estimation task in the 2D case we defined a metric called landmark accuracy. Given the true location of a landmark, we define an acceptance region of 5x5 pixels centered at the landmark, see Figure 3. Thus, the prediction for that landmark is classified as accurate if it is inside the acceptance region. Finally, the percentage of accepted predictions out of the 28 landmarks is the value of the landmark accuracy value for an image. Note that this metric is highly dependent on the acceptance region size. Further comments on this will be provided in Section 4.

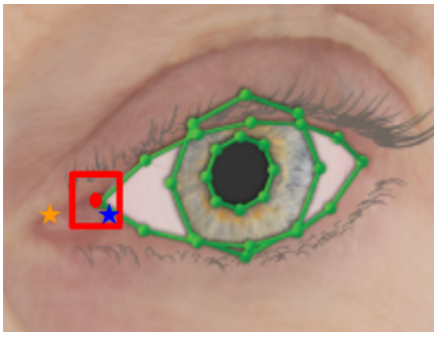

Figure 3. We set a 5x5 window centered around the ground truth position of a landmark (green points) as the acceptance region. In the figure, the blue star would be an accepted prediction, while the orange prediction would be rejected. The landmark accuracy measures the percentage of the 28 landmarks inside the corresponding acceptance regions in an image.

While our overall goal is a high landmark accuracy, optimizing for it directly is infeasible. For that reason, we employ squared, absolute and a custom loss based on the landmark-by-landmark prediction error, displayed in Figure 4. The shape of the custom function makes the loss indifferent to predictions inside the acceptance region described in Figure 3, thus resembling better the landmark accuracy metric. The landmark loss has the form $(y-\hat{y})^2 \cdot \sigma(\alpha|y-\hat{y}|-\beta)$, where y and $\hat{y}$ are the true and estimated locations of the landmark, σ is the sigmoid function, and α and β are scale and centering parameters.

Figure 4. Loss functions used for training. This function was applied to the norm of the offset vector between a estimated landmark and its true location, and averaged over the landmarks in an image. From left to right: mean absolute error, mean squared error, landmark loss.

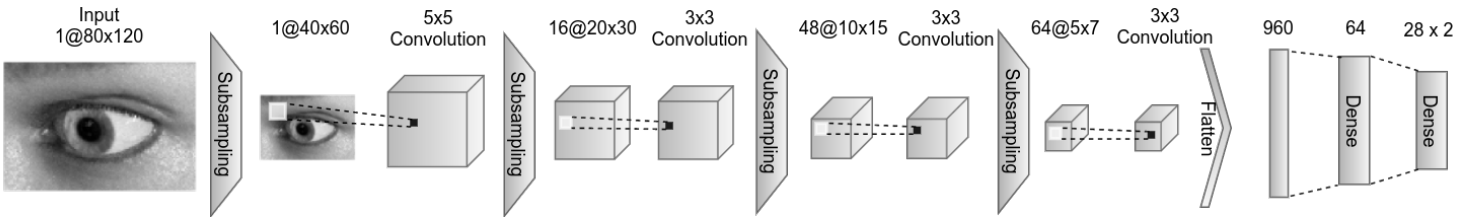

Figure 5. Illustration of the proposed CNN architecture. Note the use of a grayscale image at the input layer. All the subsampling layers are 2x2 max-pooling with stride of 2. All the activation functions are ReLU except for the linear activation in the final dense layer. For the 3D task, the final layer changed size from $56 = 28 \cdot 2$ to $84 = 28 \cdot 3$.

Figure 5 displays the structure of our proposed CNN. The landmark estimation begins with a gray-scale version of the eye region which is immediately subsampled to half of the size. This is followed by a sequence of convolution and pooling layers, which is then joined to a shallow multi-layer regressor. The output layer has the dimension of the number of landmarks (28) times the dimension of the task, either 2D or 3D. This architecture has around 130k parameters, which allows forward computations capable of real-time performance on modest devices.

Our approach for the 3D estimation is essentially the same as in the 2D case. However, in this case we do not have an explicit measure of accuracy and thus all the results provided in the following sections will be constrained to improvements in the training/test loss.

4. Results

4.1. 2D Landmark Estimation

We trained our model for 100 epochs using the Adam optimizer with a batch size of 32. The learning rate was initialized to $1e{-}3$, and other optimizer parameters were set as follows $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e{-}8$ and a decay rate of $1e{-}5$. We use the three objective functions displayed in Figure 4 and present the validation accuracy in Table 1. As we can observe in Table 1, the landmark loss (LL) performs at least as well as MSE and MAE.

Loss	Synthetic %	Augmented %	Combined %
MAE	96.4	88.8	92.6
MSE	**96.7**	89.3	**93.0**
LL	96.3	**89.6**	**93.0**

Table 1. Landmark accuracy across different subsections of the validation set. Both MSE and landmark loss (LL) objective functions result in a 93% accuracy which implies that on average 26 out of 28 landmarks were placed correctly within a 5×5 window centered on the ground-truth. As expected, the performance on the synthetic data is better than for the augmented data for all models.

For the remainder of this paper, we will present results from the model trained on with the LL objective. The performance of the LL model on the synthetic portion of the validation set is 96.3% implying that on average 27 out of 28 predicted landmarks are within a 5×5 window centered on the ground-truth. The landmark accuracy during training time is displayed in Figure 6.

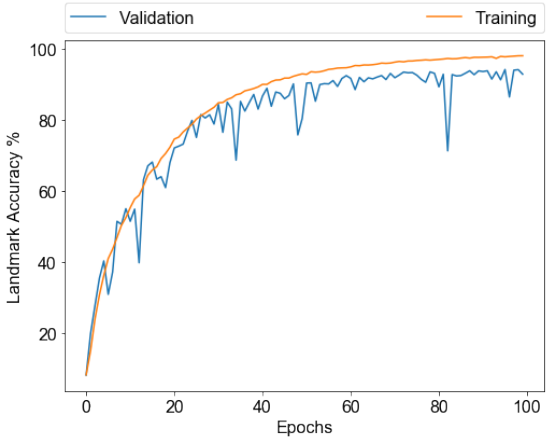

Figure 6. Accuracy of the LL model during training on the combined dataset

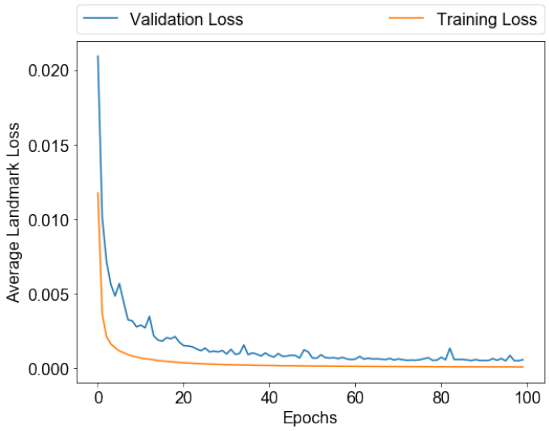

Figure 7. Loss of the LL model across 100 epochs of training on the combined dataset

As shown in Figure 6 the training accuracy smoothly increases until it saturates at 97.5 %. The validation accuracy varies strongly across training epochs until it becomes stable at 93 %. The landmark loss during training is displayed in Figure 7. As expected, the average training loss is considerably lower than the average validation loss.

To assess the performance of our LL model, we vary the window size of the landmark accuracy and display the results on different subsections of the validation set in Figure 8. As expected, the performance of the LL model on the synthetic dataset is always better than the augmented dataset. We can observe that the accuracy saturates to 100% on all subsections of the dataset when the window size is 8.

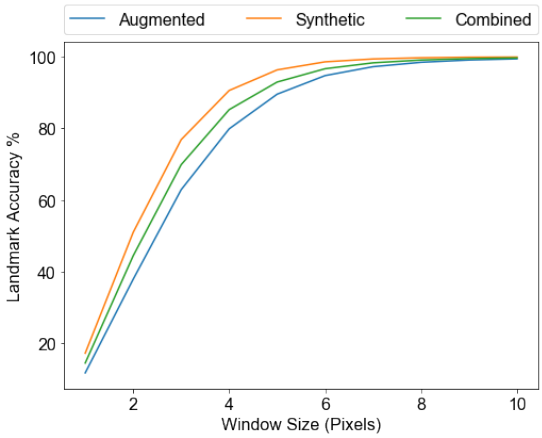

Figure 8. Landmark accuracy of the LL model across varying window sizes. Different plots refer to different subsections of the validation set.

In the original synthetic dataset, the size of the eyes do not vary greatly. Upon transforming the original images we randomly scale the images to a smaller or larger size. To account for this variation in our assessment of the LL model, we estimate the size of each eye in the validation set using its ground-truth landmarks, and plot the landmark accuracy as a ratio of the eye width in Figure 9. Using the results from Table 1, we can deduce that the LL model can achieve a validation performance of 93% on on a window as large as 8% of the eye width on average.

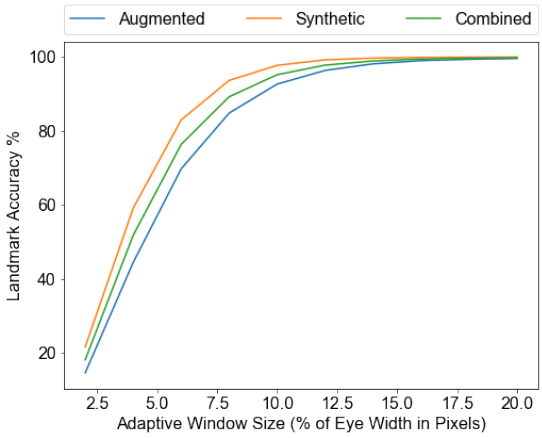

Figure 9. Landmark accuracy of the LL model across varying window sizes, where the window size is a ratio of the eye width. Different plots refer to different subsections of the validation set.

We present a few examples of our predictions on images from the validation set in Figure 10. Images from all columns except the first were transformed as described in Section 2.

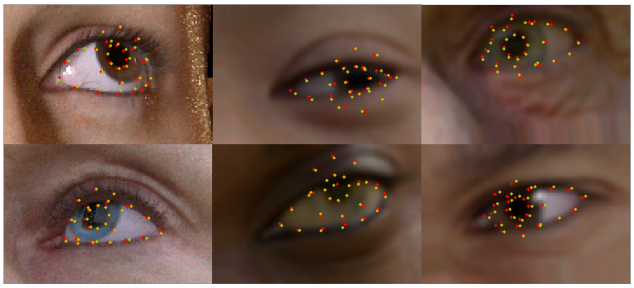

Figure 10. 2D landmark estimation performance on test set examples. Ground truth labels are displayed in red and estimated landmarks in yellow.

We display the landmark estimations of our LL model on a selection of real images from the internet in Figure 11. We can observe that qualitatively the model is able to generalize to real-world images. As expected, however, variations such as make-up or glasses which were not present in the dataset can result in faulty predictions. Surprisingly, the model is able to estimate landmarks with an acceptable precision on a drawing displayed in the bottom right corner of Figure 11.

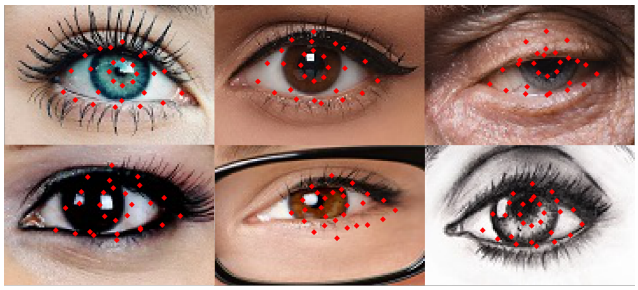

Figure 11. 2D landmark estimation on real eyes. The system performs well on real images, even under partial pupil/iris occlusion or extreme gaze directions. Low performance on eyes with glasses, drawings or excessive make up is expected, as these variations were not present during training.

Finally, we tested our model on a live video stream. As the feed-forward prediction of our network is very fast (0.5 seconds per 1000 batch of images on a laptop with a mid-range GeForce 940MX GPU), our model was able to estimate landmarks in real-time from a video stream. To provide the network with the correct input, we used Haar-cascade face/eye detector, and re-sized the eye regions to 80×120. An example of this shown in Figure 12

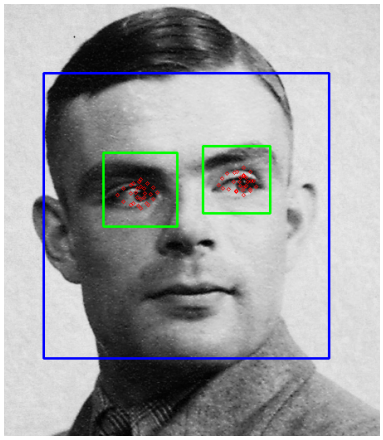

Figure 12. Performance of the 2D landmark estimation on a real image. Face and eye region proposals come from Haar detectors. The size of the proposed architecture allows for a frame-by-frame performance on live video.

Future work on conducting experiments with varying training parameters could yield insights into ways of increasing the performance of the system.

4.2. 3D Landmark Estimation

We used the same LL model and training procedure to estimate 3D landmarks. This is a more challenging task as both quantitative and qualitative evaluation is difficult without depth information of each image. Nevertheless, our model was able to produce visually appealing results. We visualize a few examples in Figure 13.

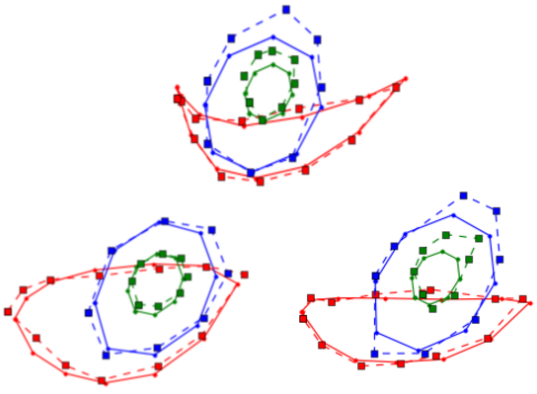

Figure 13. 3D landmark estimation performance on test set examples. The bold line contours created from ground truth landmarks. The dashed lines are the reconstructed eye shapes from estimated landmarks.

We can observe from Figure 13 that despite lacking the depth information, the estimated landmarks (dotted lines) follow the ground truth (solid lines) closely. The red lines represent the lids, the blue represent the pupils and the green the iris.

5. Application Remarks

The application of this kind of system to a real problem usually brings a number of concerns, for example *privacy*, *portability* and *robustness*, which we will discuss in this section.

Firstly, as with most computer vision applications which include a human interaction component, privacy is an important issue. Given that the training of our network was performed on publicly available synthetic data, privacy was not a concern. However, when using real data, an appropriate handling of sensitive user data is required for ensuring anonymity and respecting user rights, such as requesting permission for recording, implementing a secure storage of information, and taking safety considerations regarding the processing of the data in a local or third party platform.

Additionally a robust performance under several variation conditions is expected from our solution. As discussed earlier in this paper, we demonstrate that our model is able to generalize to many variations of input images such as illumination, occlusion, affine transformations, and low resolution by using data augmentation of our synthetic dataset.

Furthermore, our network contains only 130K parameters and is able to make predictions in real-time on a video stream. Given further experiments on the network architecture, we expect that the model complexity and network size could be shrinked further to enable it to run efficiently on a mobile device or custom video camera module.

6. Conclusions and Future Work

In this work we were able to build a robust end-to-end eye landmark estimation system based on the Syntheseyes dataset. Our network is able to generalize from synthetic images to real-world images and handle a variety of image, pose and gaze conditions, achieving a 93% 2D landmark estimation accuracy in test data. The quality of the 3D eye shapes reconstructed from estimated landmarks is acceptable in spite of the lack of depth information. Additionally, the size of the proposed architecture allows for a frame-by-frame performance on a live video stream.

Future research directions could be focused towards full face 2D and 3D landmark estimation and the corresponding 3D face reconstruction from 2D images. Additionally, the current landmark estimation pipeline could be used as a starting point for other eye-related tasks like gaze estimation or eye shape registration.

References

[1] N. Wang, X. Gao, D. Tao, and X. Li. Facial feature point detection: A comprehensive survey. *CoRR*, abs/1410.1037, 2014.

[2] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. Rendering of eyes for eye-shape registration

and gaze estimation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV 2015)*, 2015.

[3] Y. Wu and T. Hassner. Facial landmark detection with tweaked convolutional neural networks. *CoRR*, abs/1511.04031, 2015.

[4] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. *Facial Landmark Detection by Deep Multi-task Learning*, pages 94–108. Springer International Publishing, Cham, 2014.