Morphometry landmarks detection by convolutional neural network

LE Van Linh and BEURTON-AIMAR Marie

September, 2017

Abstract

In this work, we study the two methods that used to detect the landmarks on 2D images: **Deep Convolutional Network Cascade for Facial Point Detection** proposed by Yi Sun et al[1] and **Automatic ear detection and feature extraction using Geometric Morphometrics and Convolutional neural networks** from Celia Cintas et al[2]. Both articles have shown the convolution neural network (CNN) to study the landmarks on 2D images: Yi Sun studied the human face while Celia worked on human ears. Then, we show the results when applying the methods on pronotum of beetle. Finally, we propose a specific network to work on pronotum and compare the results.

1 Model 1: Facial point detection by CNN

Yi Sun et al[1] focused on five facial keypoints: *left eye center*(LE), *right eye center*(RE), *nose tip*(N), *left mouth corner*(LM) and *right mouth corner*(RM) (called landmarks). A model with 3-levels of networks are proposed to study from high-level to low-level of the landmarks.

1.1 Data

The dataset with 13466 face images includes 5590 images from LFW [3] and remaining images are downloaded from the web. The dataset is randomly divided into the training set with 10000 images and validation set with 33466 images. Each face is labeled with five landmarks and the bounding box is created around the face.

1.2 Architecture

The proposed architecture includes 3-levels of CNN: three networks at the first level, and ten networks for each remaining level(Fig.1). The networks at level 1 is designed to detect multiple landmarks while two last levels are designed for working on each landmark.

At the first level, three CNNs are employed to study the location of the facial points: F1, EN1, NM1 whose input regions cover the whole face. F1 is studied all the position of five landmarks; EN1 is worked on the eyes and nose while NM1 worked on nose and the mouth corners. Each network predicts the landmarks corresponding with the region that it covers. At the end of level 1, the coordinate of each landmark is averaged of coordinates that predicted from three networks. Fig.2 illustrates the deep structure of the networks at level 1, which contains four convolutional layers followed by max pooling and two dense layers. F1, EN1, and NM1 take the same structure but with different size of the input and different output at full-connected layers to suitable with the number of predicted landmarks.

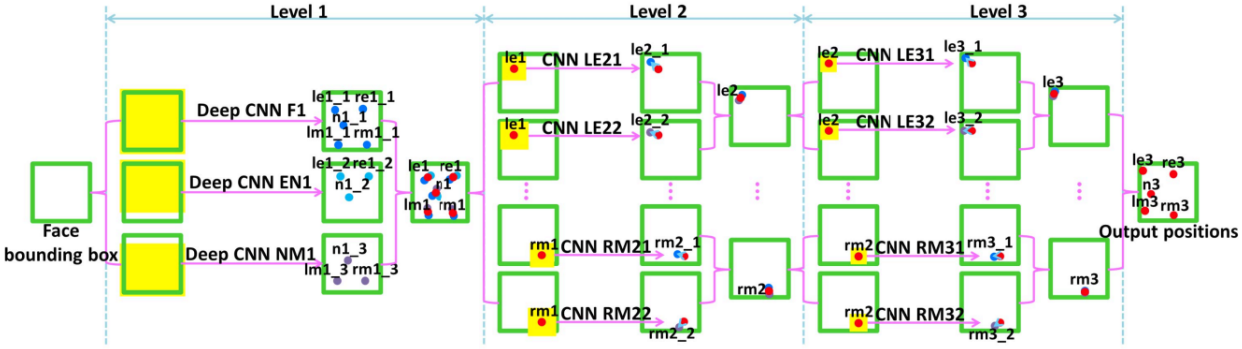


Figure 1: The 3-levels proposed architecture

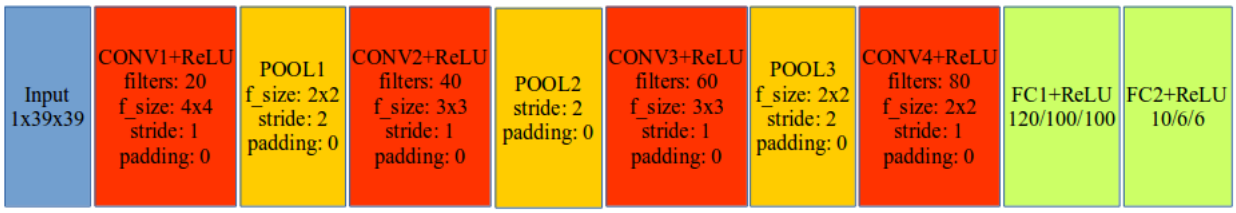


Figure 2: The structure of the networks in level 1

The networks at the second and third levels take local patches centered at the predicted position of previous levels as input. Besides, they allowed to make small changes to previous prediction. The size of patches are also reduced along with the cascade model. For each position, two networks are used to predict the new positions. The last predicted position is average of the new positions. Fig.3 illustrates the architecture of the networks in level 2 and level 3. Basically, the networks in last two levels are similar, the only difference is the way to choose the patch around the landmark. A padding is added to the coordinates of the landmark to make the change of the patches i.e $0.16, 0.18$ in level 2 and $0.11, 0.12$ in level 3. Then, the patch is resized to the size of 15×15 before giving the networks.

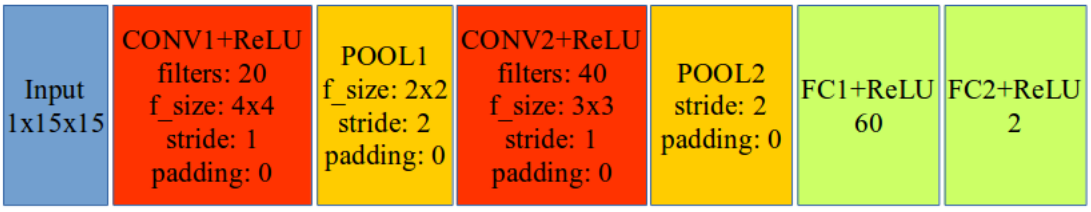


Figure 3: The structure of the networks in level 2, level 3

With 3-levels model, the purpose of the networks at the first level are estimated the landmark positions with large errors; the networks at last two levels are designed to achieve high accuracy.

1.3 Experiments

1.3.1 Training

At the first level, $F1$ takes the whole face as input (size of 39×39 of the network and outputs the position of all the five points. $EN1$ takes the top and middle part of face as input (size of 31×39) and outputs the positions of two eye centers and nose tip. $NM1$ takes the bottom and middle part of the face to predict the positions of nose tip and two corners of mouth.

All the networks at level 2 and level 3 take a small squares (15×15)centered at the predicted position by the previous level as the input and output the incremental prediction. The last predicted positions at each level are average of corresponding positions from all the networks in each level.

During training, the size of the patches is decreased for each level. The learnable parameters include weight w, the gain g and the bias b which are initialized by small random number and learned by stochastic gradient descent.

The detection error on each facial point is measured by Eq.1. If the error is greater than 5%, it is considered as failure.

$$err = \sqrt{(x - x')^2 + (y - y')^2}/l \tag{1}$$

Where:

- l is the width of the bounding box around the face.
- (x, y) is ground truth facial point
- (x', y') is predicted position

1.3.2 Testing

The model is tested with a dataset of 2557 face images. The image with the bounding box of the face is used as the input of the model. At the end, the predicted position is estimated from the model. By using the way in Eq.(1) to evaluate the model, the error statistic on each level is obtained (Figs. 4, 5, 6).

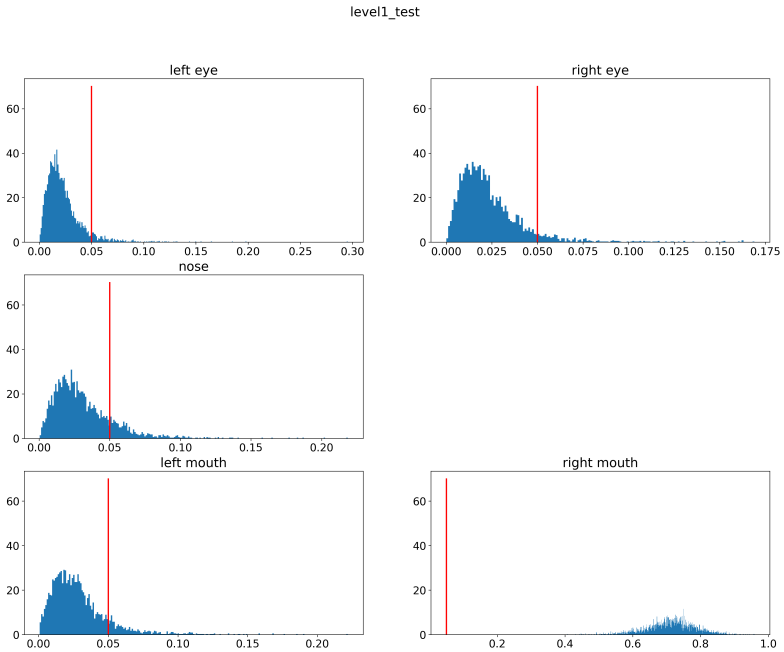


Figure 4: The error on each landmark in level 1

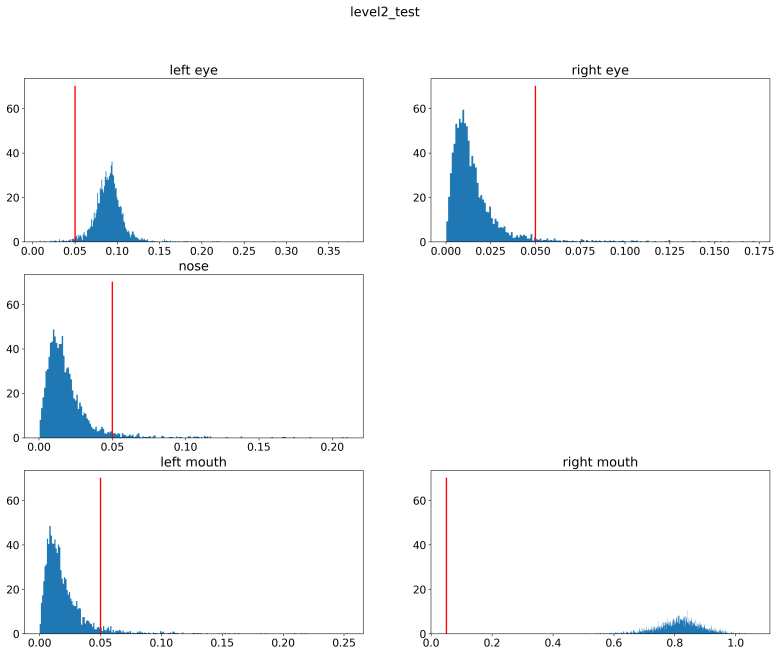


Figure 5: The error on each landmark in level 2

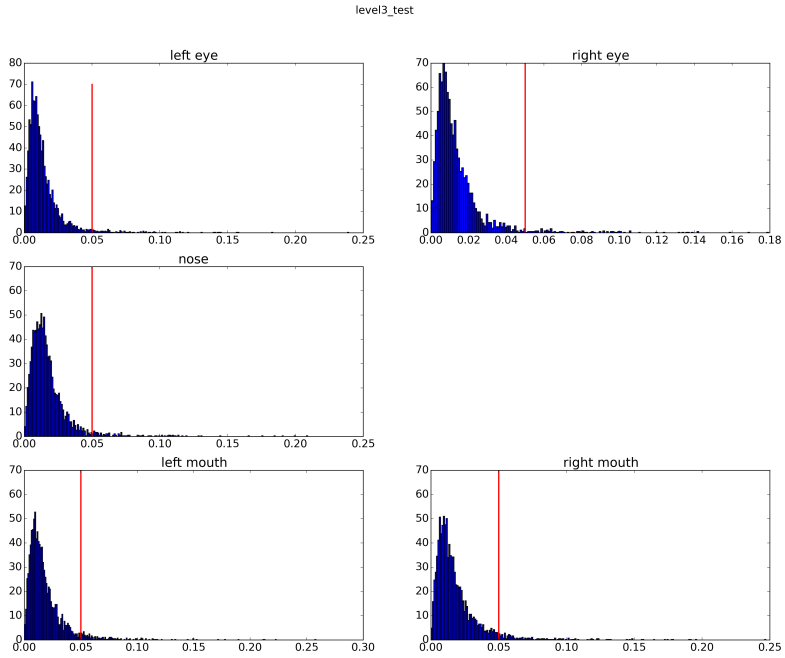


Figure 6: The error on each landmark in level 3

2 Model 2: Automatic ear landmarks detection by CNN

Celia Cintas et al[2] proposed a method based on geometric morphometric and deep learning for automatic ear detection and feature extraction in the form of landmarks. The convolutional neural network was trained with a set of manually landmarks examples. The network is able to provide the morphometric landmarks on ear image automatically.

2.1 Dataset

The image and manual landmarks belong to the CANDELA initiative [1], a project includes geneticists, bioinformatics and social-anthropologists interested on Latin American. CANDELA contains 7500 images with the size of 2136×3216. The provided dataset contains 2753 images which extracted from the CANDELA dataset. For each image, a set of 45 landmarks and semi-landmarks provided by human operators. The dataset was split into a training set with 2051 images (75%) and a validation set of 684 images (25%).

2.2 Network

Three models were designed and trained for performing the automatic landmarks task. These architectures are different in the number of convolution layers, the filter sizes, and the learning rate. An image with a single channel of the size 96×96 with brightness scaled to $[0, 1]$, is taken as the input of the network. The target (landmarks coordinates) is scaled to $[-1, 1]$. **Fig.7** shows the best architecture. In this architecture, a structure of two convolutional layers with the filters, followed by maximum pooling and dropout layer. This structure is repeated three times to obtain features at different levels with different size of filters(i.e 4×4 and 3×3). After extraction the features, two fully connected linear layers with 1500 units each and a dropout layer is hired. The output layer contains 90 output units corresponding with 45 landmarks for the predicted position of the landmarks. The implementation used Python and the Lasagne library[4].

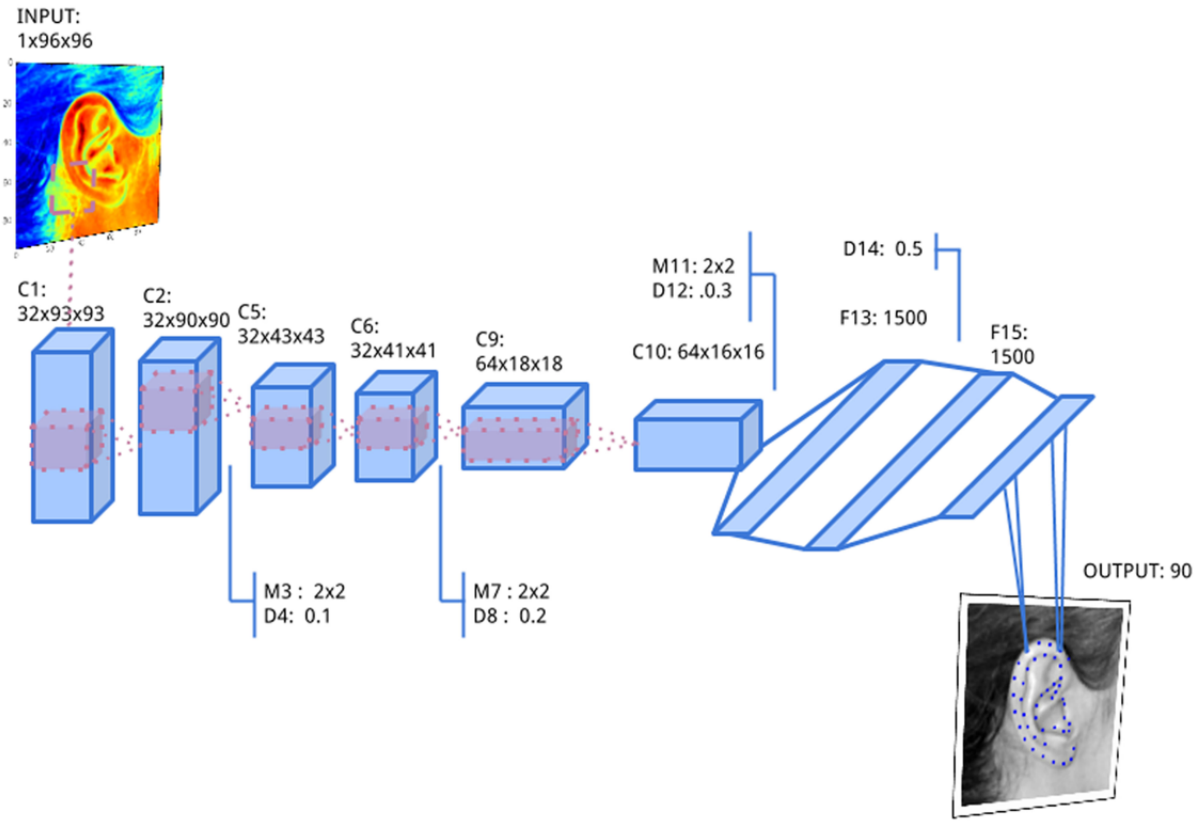


Figure 7: The best architecture for automatic ear's landmarks detection

[1] https://www.ucl.ac.uk/candela

2.3 Experiments

Because the CANDELA dataset is not published. Another dataset was chosen to study the network. The new dataset was used for the Facial Keypoint Detection including 7049 gray-scale images (96×96). For each image, we are supported learn to find the position of 15 landmarks. After dropping some missing data, the dataset remains 2140 images. All the images with coordinates of manual landmarks is stored in csv file and fetched into the network. During the training and validation, the usual quality metrics for regression problems is used, in particular, root mean square error (RMSE). Fig.8 shows the learning curves of the model on training and validation set error. The training is finished after 3000 iterations with the loss arround 10^{-3}. Fig.9 shows some test on the real facial images.

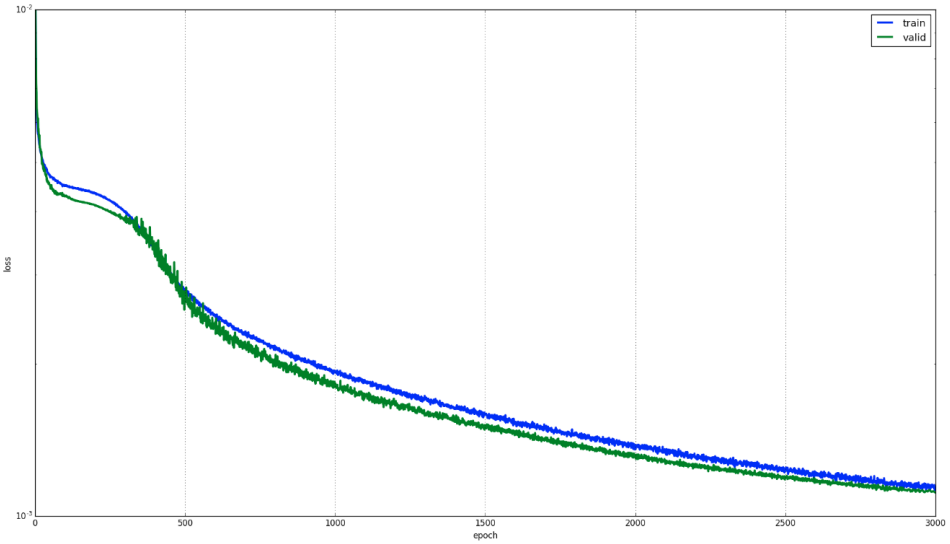


Figure 8: The loss of Ear-CNN network on facial point dataset

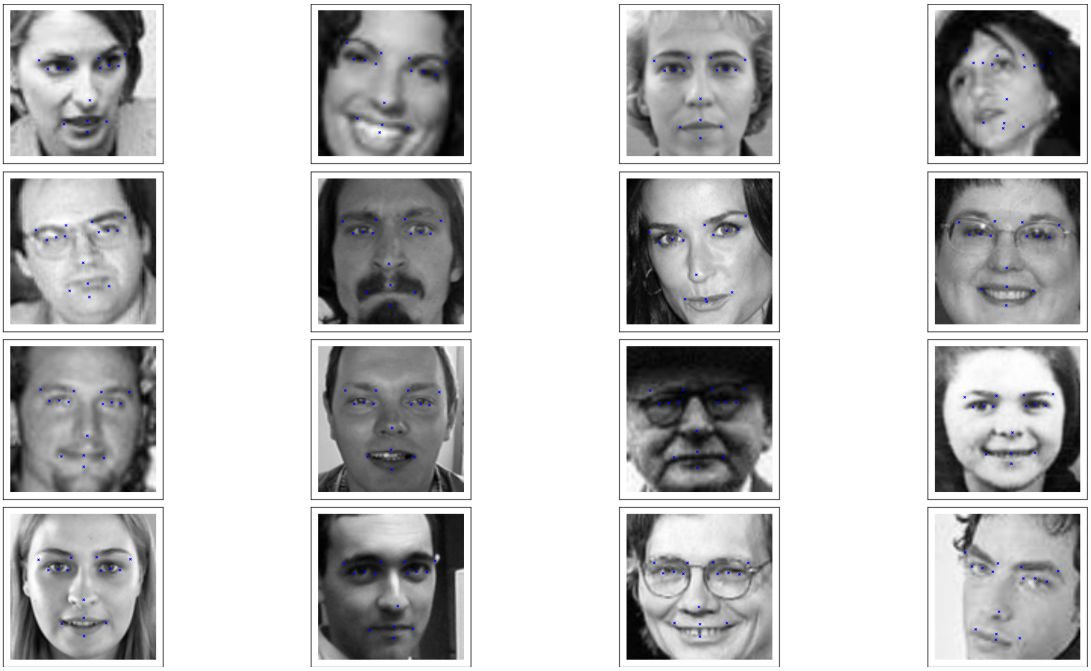


Figure 9: The prediction landmarks on human face

3 Model 1 and model 2 on pronotum

3.1 Dataset preparing

The dataset includes 293 pronotum images. The images are divided into three subsets: the training set (200 images), the validation set (60 images) and the testing set (33 images). Because the dataset is limited and the models are worked on gray-scale images, we applied some ways to en-large the dataset. Firstly, for each original image in RGB, each channel is modified by adding some values. Secondly, the channels of the original image are split. So, we have obtained 1400 images for the training set and 420 images for the validation set. At the end, the images are down-sampled with the size of 256×192 before giving to the networks.

3.2 Model 1 and pronotum landmarks

The networks in the first level are modified to suitable with the prediction of landmarks on the pronotum (8-landmarks). For each pronotum, eight manual landmarks have been set. The bounding box is created depending on the coordinate of the manual landmark and kept with the same size. The networks in the first level are used as followed:

- F1 network recognizes whole pronotum bounding box with eight landmarks.
- EN1 network predicts the location of the first five-landmarks i.e [1..5].
- NM1 network is used to estimated the position of last four-landmarks and the first landmark i.e $[5..8, 1]$.
- At the end, the position of each landmark is average of the predicted position in the networks.

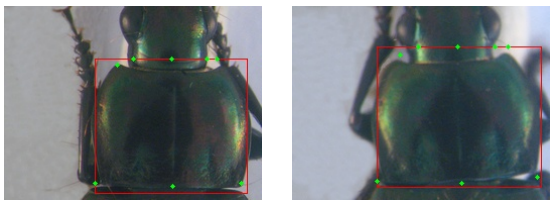
3.2.1 Testing

During training, the Euclidean distance (sum of squares) is used to compute the loss of the networks. The error rate of each network during training is shown in the Table.1:

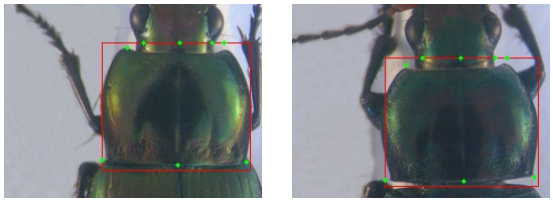
Network	Loss
F1	0.013
EN1	0.47
NM1	0.5

Table 1: The loss of the networks in Model 1 on pronotum dataset

From Table 1, the errors of EN1 and NM1 are still high. That errors make the prediction result of level 1 do not enough good. Besides, the networks at level 2 and level 3 used the prediction at level 1 as the input data to predict the new position. So, we can not continue with the level 2, 3 until the result at level 1 is improved. Perhaps, the model in model 1 is not suitable to detect the landmarks on pronotum. Fig. 10 shows the prediction landmarks on four images. Followed that, the networks can detect the landmarks at positions 4, 5 and 7; the different positions still not good.



(a) Test 1 (b) Test 2



(c) Test 3 (d) Test 4

Figure 10: The pronotum with predicted landmarks at level 1

3.3 Model 2 and pronotum landmarks

The dataset is kept the same with model 1 (1400 images for training and 420 images for validation) but having some changes. Firstly, the ways to choose the data(to train and validate) is changed. All images are combined. Then, the network will automatically choose 75% data to train and 25% for validation. Secondly, the inputs that given to the network are just the image and landmarks(without the coordinates of the bounding box).

The network is run 3000 iterations with the learning rate begin from 0.08 to 0.01. During training, the learning rate is changed to fit with the remaining iterations[5]. Fig.11 shows the first 700 iterations during training. The loss did not have many changes after 100^{th} iteration.

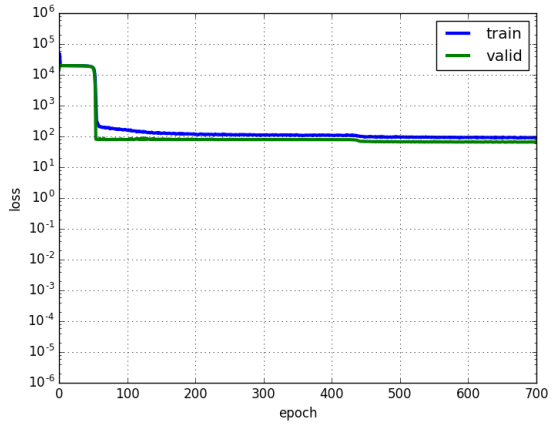


Figure 11: The losses of model 2 on pronotum dataset

Fig.12 shows the prediction landmarks on 16 images. Following, the prediction landmarks from the network of model 2 are closed with the pronotum but the location is still inaccurate.

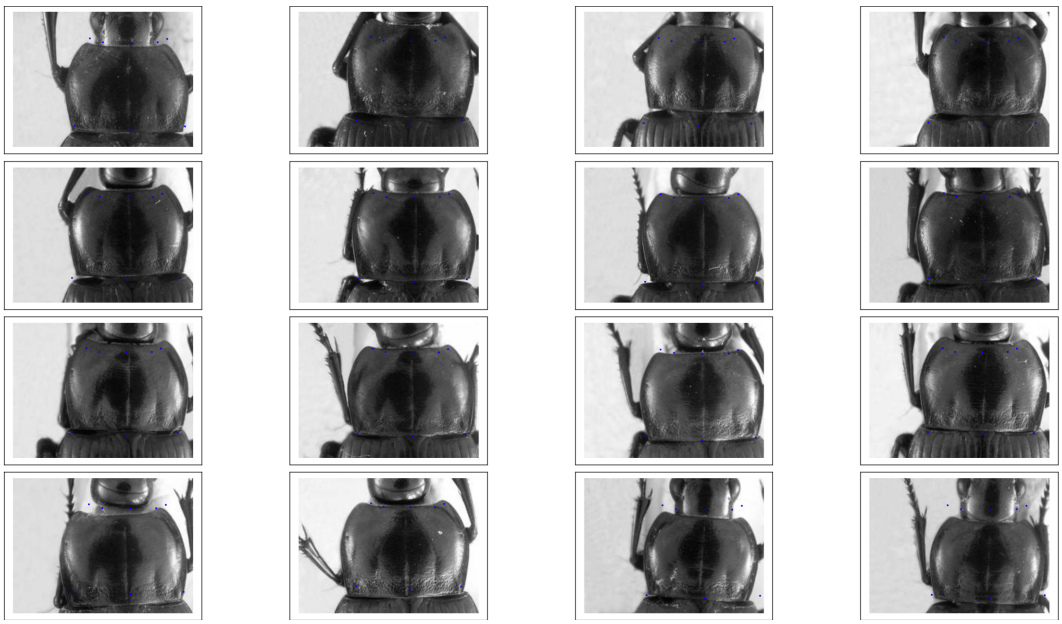


Figure 12: The prediction landmarks on pronotum of model 2

4 Proposed architecture

4.1 Model and parameters

From the tutorial of Daniel Nouri[2] about using CNN to detect facial key points. We propose a CNN to detect the landmarks on pronotum. The proposed network includes three convolutional layers followed by three maximum pooling layers and three full connected layers(Fig.13). The network receives the gray-scale image (256×192) as the input. The deep of convolutional layers is increased from $32, 64$, to 128 with different size of filter. The size of filters in pooling layers are kept in the same size of 2×2. At the end of network, three full-connected layers with the size of $500, 500$, and 16 are set up to predict the positions of landmarks. Besides, the model is designed with a small sharing learning-rate and the momentum. The learning-rate and the momentum are changed overtime of training.

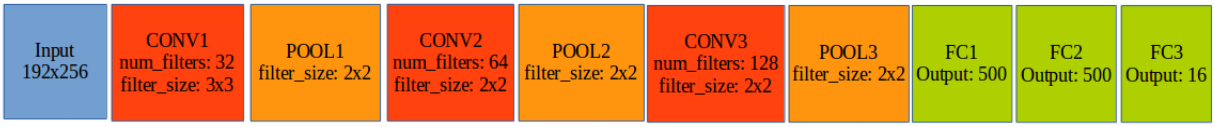


Figure 13: The architecture of proposed model

4.2 Training and experiments

The model is trained with 1820 images in 5000 iterations. The images are normalized before giving to the network by scaling the intensity value to [0,1], instead of 0 to 255. The target values (x and y coordinates) is kept as original. During the training, the root-mean-square error (MSE) is used to calculate the loss.

To evaluate the stability and confidence of the model, we proposed several rules to select the images for the network, as Table.2:

[2]http://danielnouri.org/

Rule	Test set	Training and validation set
rule_1	From 1^{st} image to 33^{rd} image	Remaining images
rule_2	From 260^{th} image to 293^{rd} image	Remaining images
rule_3	Random	Random
rule_4	From 90^{th} image to 122^{nd} image	Remaining images
rule_5	From 200^{th} image to 232^{nd} image	Remaining images

Table 2: The rules to choose the data for the network

Following the rules to choose the data, the loss of training and validation is shown in Table 3. From the results in the table, the training losses in the cases of *rule_4, rule_5* are smaller than other rules; but the validation losses are stability. It means the overfitting is appeared clearly in the case of *rule_4* and *rule_5*. In which, the smallest difference value between training and validation loss is belong to **random** case.

Rule	Training loss	Validation loss
rule_1	0.12739	0.63681
rule_2	0.15204	0.59480
rule_3	0.16694	0.55584
rule_4	0.08798	0.61934
rule_5	0.0918	0.52843

Table 3: The training loss and validation loss following each rule to choose the data

Fig.14 shows the training curve loss and validation curve loss of the model on each rule to choose the data. From the beginning, the loss is not changed. When the training is longer and the learning rate is improved, the loss is decreased and a large distance between training and validation is appeared (over-fitting) but it is not that bad.

The model is tested on the test datasets(five rules). Then, correlation between the manual landmarks and predicted landmarks is computed by applying the correlation methods (see Table.4, 5, 6)

Rule	x correlation	y correlation
rule_1	0.9953877	0.9941767
rule_2	0.9968787	0.9960827
rule_3	0.9966784	0.9957729
rule_4	0.9975662	0.9985097
rule_5	0.9972048	0.9976416

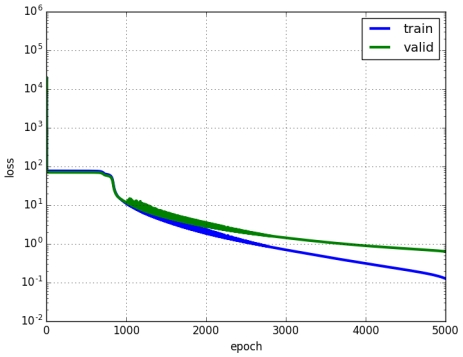
Table 4: The correlation between manual and predicted landmarks by Pearson[6] method

Rule	x correlation	y correlation
rule_1	0.9893943	0.9289319
rule_2	0.992556	0.9444423
rule_3	0.9913126	0.9565425
rule_4	0.9943106	0.9789221
rule_5	0.9920646	0.9864683

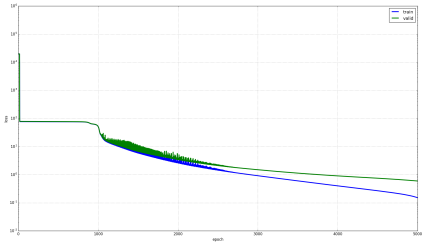
Table 5: The correlation between manual and predicted landmarks by Spearman[7] method

Rule	x correlation	y correlation
rule_1	0.913517	0.7498531
rule_2	0.9303295	0.8231899
rule_3	0.9273002	0.8273057
rule_4	0.9419902	0.8904413
rule_5	0.9299128	0.9051508

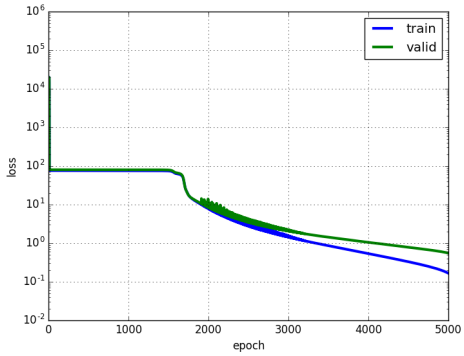
Table 6: The correlation between manual and predicted landmarks by Kendall[8] method



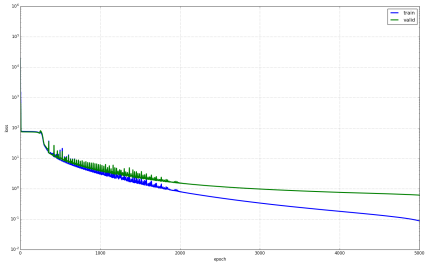
(a) rule_1



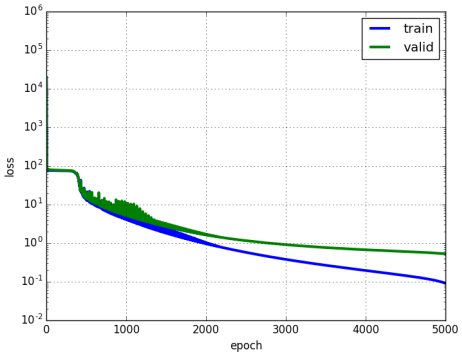
(b) rule_2



(c) rule_3



(d) rule_4



(e) rule_5

Figure 14: The loss during training and validation followed each rule to choose data

Fig.15 show the predicted positions on test dataset followed rule_3:

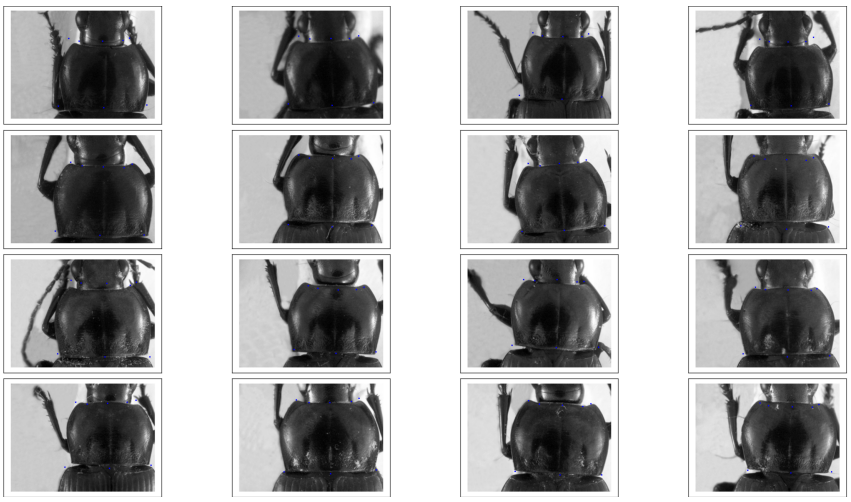


Figure 15: The prediction landmarks on 16-pronotum images

5 Conclusions

In this studied, two methods that used to predict the landmarks on 2D gray-scale images are studied. For each case, the model is suitable with different dataset but the results are still not good when we change the data (pronotum). Besides, we proposed a network to learn and detect the landmark positions on pronotum. The accuracy of the model is greater than 98%. The correlation coefficient between manual and predicted landmarks have been evaluated enough good to precise.

References

[1] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[2] Celia Cintas, Mirsha Quinto-Sánchez, Victor Acuña, Carolina Paschetta, Soledad de Azevedo, Caio Cesar Silva de Cerqueira, Virginia Ramallo, Carla Gallo, Giovanni Poletti, Maria Catira Bortolini, et al. Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks. *IET Biometrics*, 6(3):211–223, 2016.

[3] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[4] Sander Dieleman, Jan Schlter, Colin Raffel, Eben Olson, Sren Kaae Snderby, Daniel Nouri, et al. Lasagne: First release., August 2015.

[5] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[6] Julie Pallant. *SPSS survival manual*. McGraw-Hill Education (UK), 2013.

[7] Jerome L Myers, Arnold Well, and Robert Frederick Lorch. *Research design and statistical analysis*. Routledge, 2010.

[8] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

ANNEX

A The parameters in the networks of each model

	layer 0	layer 1	layer 2	layer 3	layer 4	layer 5	layer 6	layer 7	layer 8	layer 9
F1	I(1,39,39)	CR(20,4,1)	P(2)	CR(40,3,1)	P(2)	CR(60,3,1)	P(2)	CR(80,2,1)	F(120)	F(10)
EN1 & NM1	I(1,31,39)	CR(20,4,1)	P(2)	CR(40,3,1)	P(2)	CR(60,3,1)	P(2)	CR(80,2,1)	F(100)	F(6)
level2 & level 3	I(1,15,15)	CR(20,4,1)	P(2)	CR(40,3,1)	P(2)	F(60)	F(2)			

Table 7: The parameters in the networks of Model 1

layer 0	layer 1	layer 2	layer 3	layer 4	layer 5	layer 6	layer 7	layer 8	layer 9	layer 10	layer 11	layer 12	layer 13	layer 14	layer 15	layer 16
I(1,96,96)	C(32,4,1)	C(32,4,1)	P(2)	D(0.1)	C(32,3,1)	C(32,3,1)	P(2)	D(0.2)	C(64,3,1)	C(64,3,1)	P(2)	D(0.3)	F(1500)	D(0.5)	F(1500)	F(90)

Table 8: The parameters in the networks of Model 2

layer 0	layer 1	layer 2	layer 3	layer 4	layer 5	layer 6	layer 7	layer 8	layer 9	layer 10	layer 11	layer 12	layer 13	layer 14	layer 15
I(1,245,326)	CR(25,3,1)	CR(25,3,1)	P(2)	D(0.1)	CR(25,3,1)	CR(25,3,1)	P(2)	D(0.1)	CR(50,3,1)	P(2)	D(0.1)	FR(500)	D(0.2)	FR(500)	F(1)

Table 9: The parameters in the networks from Nicolas

layer 0	layer 1	layer 2	layer 3	layer 4	layer 5	layer 6	layer 7	layer 8	layer 9
I(1,192,256)	C(32,3,1)	P(2)	C(64,2,1)	P(2)	C(128,2,1)	P(2)	F(500)	F(500)	F(16)

Table 10: The parameters in the networks of proposed model

Which:

- I: Input(depth, height, width)
- C: Convolutional(output, filter kernel size, stride)
- CR: Convolutional + ReLU (output, filter kernel size, stride, padding = 0)
- P: Maximal pooling(filter kernel size)
- F: Full connected(output)
- FR: Full connected + ReLU