Automatically landmarks prediction on Beetle's pronotum

Le Van Linh[1,3], **Beurton-Aimar Marie**[1], **Zemmari Akka**[1], **Parisey Nicolas**[2]

[1]LaBRI - CNRS 5800 Bordeaux University, France, van-linh.le/beurton/zemmari@labri.fr
[2]IGEPP - INRA 1349, France, nparisey@rennes.inra.fr
[3]ITDLU - Dalat University, Vietnam, linhlv@dlu.edu.vn

Keywords: Landmarks, convolutional neural networks, fine-tuning, recogntion, procrustes.

Abstract

In recent years, deep learning is known as a good solution for the difficult problems in computer vision. It appears in many fields such as classification, recognition, face detection. In this paper, we propose a scenario to predict the landmarks on 2D images, specify beetle's head images. The proposed method includes two stages: firstly, the landmarks are estimated by applying convolutional neural network; then, the estimated landmarks are verified to increase the accuracy. The method experimented on a set of 293 images. The accuracy of the method is evaluated by calculating the distance in pixels between the coordinates of the predicted landmarks and manual landmarks which were provided by the biologists.

1 Introduction

Morphometry landmark (or point of interest) is an important feature in many biological investigations. It was usually used to analyze the forms of whole biological organs or organisms. The analysis is mainly based on the coordinates of the landmarks. The collecting of enough the number of landmarks can help the biologists make a good estimate about organisms. Depending on the problem, the number of landmarks may be more or less; besides, the location of landmarks can be located on the shape (border) or inside the object, *for examples,* the landmarks on Drosophila wings have stayed on the veins of the wings but the landmarks on human ear can be located at the ear hole or inside. Recently, the landmarks were set manually by the biologist. This work is time-consuming and difficult to reproduce. Therefore, a method that proposes automatically the coordinates of landmarks could be a concern.

Based on the characteristics of the images, the images can be divided into two groups: the images that we can easy to segment the objects in the image, called segmented images; and the images that we can go in tight when segment the objects, called un-segmented images. For that reason, the methods that used to identify the landmarks automatically may be divided into two groups too. For segmented images, identification of landmarks on the shape can be finished by applying the image processing techniques such as HOG[**?**], SIFT[**?**], ... But for un-segmented images, defining the landmarks become a challenge and the image processing techniques seem to be inappropriate. This article introduces a scenario for automatic detection of the landmarks on biological images, specific beetle's head images, called *pronotum* images (Fig. 1). The method includes 2 stages: 1) the initially predicted landmarks are given by a convolutional neural network (CNN) [**?**]; 2) the predicted landmarks which located in the shape of pronotum will be refined the location to increase the accuracy of coordinates. In the first stage, the main idea is design and train a CNN with a set of images and their manual landmarks. The dataset includes 293 pronotum images and their manual landmarks which have been provided by the biologists. The images are presented in two dimensions and RGB color. After training, the trained network will be able to detect the initially predicted landmarks on the pronotum images. In the second stage, the predicted landmarks in the shape will be refined the coordinates by applying a Procrustes analysis[**?**]. For each manual landmark, a model is generated as a specific. Then, it is used to refine the corresponding predicted landmarks.

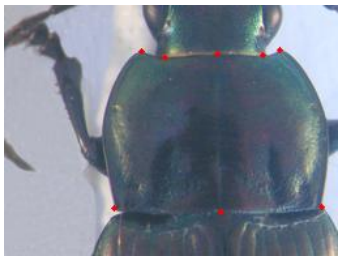


Figure 1: An example of pronotum images and its manual landmarks

In the next section, we present related works in domain automatically estimation landmarks on 2D images. In section 3, we present an overview about the stage that predict the initial automatically landmarks by applying CNN. The procedure apply to refine the predicted landmarks which provide by CNN will be presented in section 4. In the last section, we show all the experiments and analysing the results.

2 Related works

Landmarks or points of interest are one of the important characteristics in geometric morphometrics. Landmark studies have

traditionally analyzed on 2D images. Depending on which situation was stayed (segmented or un-segmented images), setting landmarks must apply the different methods.

When segmentation can be applied, Lowe et al. [**?**] have proposed a method to identify the key points in the 2D image. From the detected key points, the method is able to match two images. Palaniswamy et al. [**?**] have applied probabilistic Hough Transform to automatically estimate the landmarks in images of Drosophila wings. Adrien et al. [**?**] have extended Palaniswamy's method to detect landmarks automatically on beetles mandibles. Unfortunately, this method can not be applied to other parts of beetle that the segmentation has too many noises, such as pronotum images.

Recently years, machine learning is developing rapidly, specifically deep learning (CNN). It exists in most of the fields, especially in computer vision. We can finish a lot of difficult tasks with a deep convolution neural network such as classification [**?**], image recognition [**?**], speech recognition [**?**] and language translation [**?**]. Using CNN to determine landmarks on 2D images will produce good results and it may be a good solution for the un-segmented images. Yi Sun et al. [**?**] have proposed a cascaded convolutional network to predict the key points on the human face. Zhang et al. [**?**] optimizes facial landmarks detection with a set of related tasks such as head pose estimation, age estimation, ... Cintas et al. [**?**] have introduced a network to predict the landmarks on human ear images. In the same context, we have applied CNN to predict the landmarks on pronotum images. The predicted landmarks then refined to increase the accuracy of coordinates.

3 Automatic landmarks by using CNN

Deep learning presents a learning method with multiple levels of representation of connected layers (convolutional neural network). Data representation is transformed from a lower level to a higher level with many complex functions can be learned via backpropagation. In this section, we will present a CNN that we used to predict the landmarks on pronotum images. Besides, the techniques that applied to preprocess data before using for training the network.

3.1 Network architecture

Like the other networks [**?**], the proposed network consists of several common layers with different learnable parameters (Fig.2). It receives an input of $1 \times 256 \times 192$ to train, validate, and test. The network consists of three repeated-structure of a convolutional layer followed by a maximum pooling layer and a dropout layer. The depth of convolutional layers increases from 32, 64, and 128 with different size of the filter kernel: 3×3, 2×2, and 2×2. All the kernels of pooling layers have the same size of 2×2. The probability values used for dropout layers are $0.1, 0.2$, and 0.3. At the end, three full connected layers have been added to the network. The outputs of the full connected layers are 1000, 1000, and 16, respectively. The output of the last full-connected layer corresponds to 8 landmarks (x and y coordinates) which we would like to predict. Additional, to have a better control of overfitting, another dropout layer with a probability of 0.5 is inserted between the first two full connected layers [**?**].

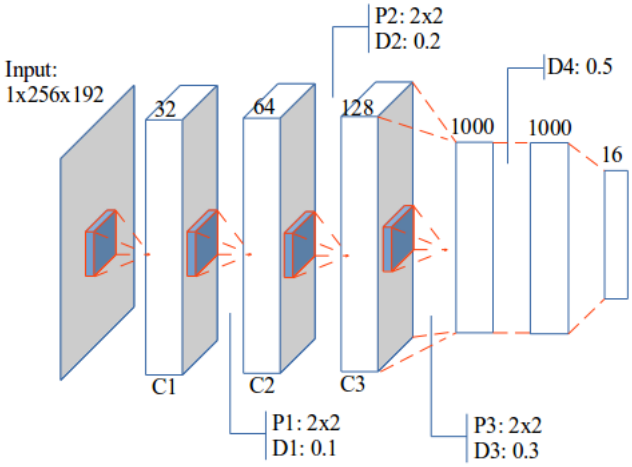


Figure 2: The architecture of the proposed convolutional neural network

During training, the values of learnable parameters have been updated to increase the accuracy of the network by applying gradient descent in backward phase. Therefore, the network is designed with a small sharing learning rate and momentum. Their values are updated over training time to fit with the number of epochs. The implementation of the network is done on Lasagne framework [**?**] which allows computing on GPU. The network has been trained on NVIDIA TITAN X cards.

3.2 Data processing

The dataset includes 293 pronotum images of beetles. All images are taken with the same camera in the same condition with a resolution of 3264×2448. Each image has been set 8 manual landmarks by biologists (Fig. 1). The dataset was split into two subsets: training (and validation) set contains 260 images and testing set includes 33 images. In most of CNNs [**?**], the size of the input was limited to 256 pixels. In our case, the resolution of input image seems that too large and it becomes a difficulty for the network. So, the images are down-sampling to a new resolution 256×192 before training and testing. Of course, the coordinates of manual landmarks are also scaled to fit with the new resolution of images.

Besides the size of the input, the number of images is also a challenge when applying CNN. Normally, training a CNN with a large dataset will give us the result better than when we training CNN on a small dataset. Moreover, working with a small dataset, we can meet a popular problem, *overfitting*. So, we need to enlarge the size of the dataset instead of 293. In image processing, we usually apply transform procedures (translation, rotation) to generate a new image but in fact, when we compute the value of the pixels, it does not change while CNN computes the values of the pixels. Therefore, we have

applied two other procedures to increase the number of images in the dataset. To address this problem, we have applied two procedures to enlarge the size of the dataset.

The first procedure was applied to change the value of each channel in the original image. According to this, a constant is added to a channel of RGB image and for each time, we just change the value of one of three channels. For example, from an original RGB image, if we add a constant c = 10 to the red channel, we will obtain a new image with the values at red channel by greater than the red channel of original image a value of 10. By this way, we can generate three new RGB images from a RGB image.

The second procedure is splitting the channels of RGB images. It means that we separate the channels of RGB into three gray-scale images. This work seems promising because the network works on single-channel images. At the end, we can generate six versions from an image, the total number of images used to train and validate is $260 \times 7 = 1820$ images (six versions and original image).

3.3 Training and experiments

The network was trained on a dataset of 1820 images. The number of images that used for training and validation is splitted randomly by a ratio (training: 80%, validation: 20%) that has been set during the network setup. During the training, the network learned the information through a pair of *(image, landmarks)* in training set. At the test phase, the image without landmarks was given to the trained network and the predicted landmarks will be given at the output. In practical of CNN, convergence is usually faster if the average of each input variable over the training set is close to zero. Moreover, when the input is set closed with zero, it will be more suitable with the sigmoid activation function [**?**]. According to [**?**], the brightness of the image is normalized to $[0; 1]$, instead of $[0; 255]$ and the coordinates of the landmarks are normalized to $[-1; 1]$, instead of $[0; 256]$ and $[0; 192]$ before giving to the network.

The training was finished in 5000 epochs[1]. The learning rate was initialized at 0.03 and stopped at 0.00001, while the momentum was updated from 0.9 to 0.9999. Because landmarks prediction can be seen as a regression problem in deep learning. Therefore, the root mean square error (RMSE) was used as a quality metric to evaluate the result and compute the losses of the proposed architecture.

Fig. 3 shows the training error and the validation error during training time. The blue curve presents RMSE error on training data. The green curve presents the validation error. Clearly, the losses are very different from the beginning. But, the difference is narrowed when the epoch increase.

Besides the losses during training, the accuracy on coordinates of predicted landmarks of the test images is also considered. Firstly, the trained model was used to predict the landmarks on all images in the test set. Then, the distance (in pixels) between manual and corresponding landmarks in each image was calculated as the error distance. Finally, the error distance per landmark was calculated for all test images. Table.1 shows the average error distance given on each landmark. With the size of the images is 256×192, if we accept an error around 3% of the image size (~ 3.5 pixels), the error distances are acceptable.

[1] An epoch is a single pass through the full training set

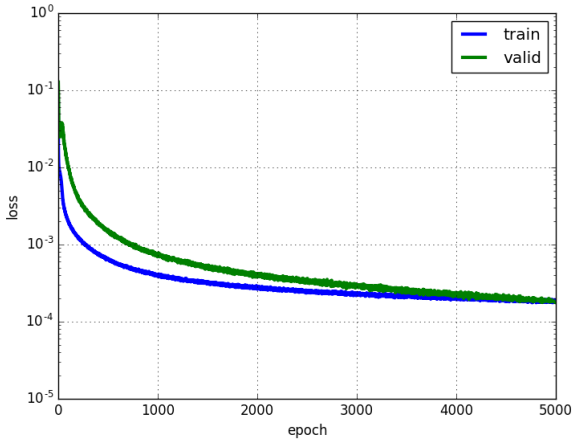
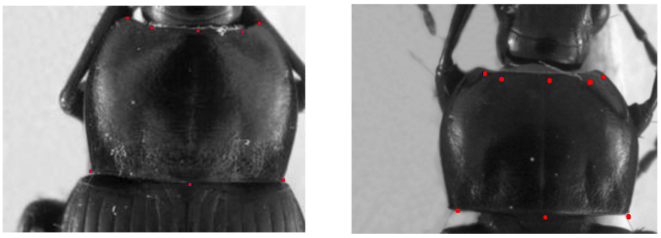


Figure 3: The loss curves during training of proposed network

#Landmark	Distance
1	4.002
2	4.4831
3	4.2959
4	4.3865
5	4.2925
6	5.3631
7	4.636
8	4.9363

Table 1: The average error distance per landmark

Fig.4 shows the predicted landmarks on two test images. When we consider the accuracy of predicted landmarks by calculating the distance between manual and corresponding predicted landmarks, the accuracy on coordinates of predicted landmarks on Fig.4a is 99% and the propotion on Fig.4b is 80%.



(a) Image with well-predicted landmarks

(b) Image with inaccuracy landmarks

Figure 4: The predicted landmarks on an image in test set. The read points present for the predicted landmarks

To have a better evaluation of the predicted landmarks, we have applied the standard deviation [**?**] to quantify the dispersion of a set of distances. In this case, a predicted landmark is considered as acceptable if its error distance to the corresponding manual landmark is less than the average error (per landmark) plus standard deviation value. Fig. 5 shows the ratio of acceptable per landmark. Most of landmarks have been predicted with the accuracy grater than 70%. In which, the lowest and highest prediction accuracies are 66.67% and 87.88%, respectively.

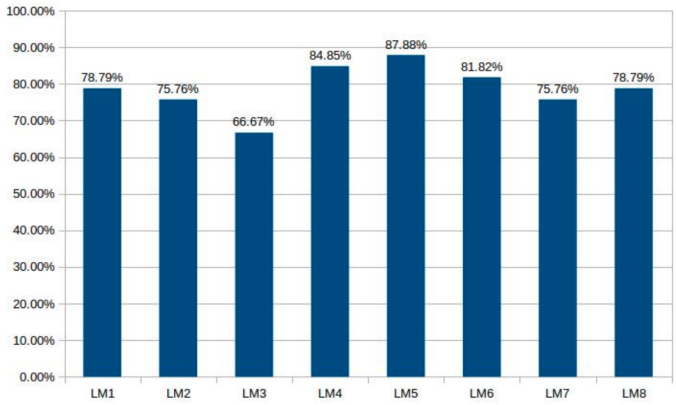


Figure 5: The proportion of acceptable predicted landmarks

As a result, the network is able to predict the landmarks on a test set of pronotum images. In statistic side, the predicted landmarks are acceptable. But in image processing side, we expect more about the accuracy (coordinates of predicted landmarks), and the result of CNN is still needed to improve. However, improving on all landmarks is difficult to perform because a number of landmarks have stayed inside the pronotum. Instead of, we will improve the result of the landmarks that stay in the shape of the pronotum.

4 Improving the predicted landmarks

The PDF format will be the final format under which the papers will appear in the Proceedings. Therefore you are required to submit your paper as PDF document. If this is not possible, Postscript format is also accepted as long as no fonts other than the recommended fonts are used.

You can use any of the popular free LaTeX editors (e.g. Kile).

5 Results

The submission process for ICPRS 2018 should be done on line at http://www.icprs.org

A PDF version of your final paper is required. It should be expected that after your submission, your paper is published directly from the file you send without any further proofreading. Therefore, it is advisable for the authors to print a hard copy of their final version and read it carefully.

6 Conclusions

The list of references should be ordered in the same order as first cited in the text. All references should be cited in the text, and using square brackets such as [**?**] and [**?**, **?**]. We recommend the use of IEEE Transactions style for references.

Acknowledgements

The acknowledgement for funding organisations etc. should be placed in a separate section at the end of the text.

Thank you for your cooperation in complying with these instructions.