Automatize landmarks detection with Deep Learning

Le Van Linh[a,c,*], Beurton-Aimar Marie[a,1], Zemmari Akka[a], Parisey Nicolas[b,1]

[a] *University of Bordeaux, 351, cours de la Libration, 33405 Talence, France*
[b] *UMR 1349 IGEPP, BP 35327, 35653 Le Rheu, France*
[c] *Dalat University, Dalat, Lamdong, Vietnam*

Abstract

Landmark is one of the important concepts in Procrustes Analysis. Finding landmarks is not only helps us measure the shape of the object, but also determine the correspondence between the objects when applying the alignment methods. In biology, landmarks are widely used in the morphometric analysis to analyze the inter-organisms variations with the purpose to classify and to determine the evolution of an organism's family. Currently, the landmarks are mostly determined manually by the biologist. In this work, we applied Deep Learning, specify Convolutional Neural Network (CNN), to predict the landmarks on biological images. Whereby, we proposed a CNN architecture which was built from the "elementary blocks". Each block is made up of some popular layers of CNN. The network then will be trained and tested on a dataset includes five parts of beetle (head, elytra, pronotum, left and right mandibles). These works have also introduced another strategy to augment the dataset which can see a little bit small in our case. In the experiments, we apply two strategies to evaluate the network and to improve the obtained results: training from scratch and applying a fine-tuning step. The predicted landmarks from the network will be compared with the manual landmarks which provided by the biologists. The obtained results are considered to be statistically good enough to replace the manual landmarks. The complete workflow is implemented and freely available

*Corresponding author
Email address: van-linh.le@labri.fr (Le Van Linh)
[1] both authors contributed equally to this work.

on GitHub.

Keywords: Landmarks, deep learning, fine-tuning, CNN

1. Introduction

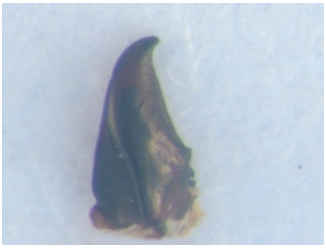
Anatomical landmarks are preferred to the points that represent a biological object. They have a biologically meaningful point in an organism and are widely used to ensure the correspondences within the same species. The anatomical landmarks are used in many biological studying, the biologist can estimate the shape from the collected landmarks, then they may evaluate the evolution of a group organism. Depending on which kind of analysis objects, the number of landmarks may different, as well as their positions can be defined along the shape or inside the object, i.e. Drosophila wings [] have 13 landmarks and they are located on the veins of the wings, while the head of a beetle has 10 landmarks and they stay inside the object (Fig. 15e).

Currently, the landmarks on anatomical images are set manually by the entomologist or semi-automatically by applying some image processing techniques. One thing can note that the manual work is very time-consuming and difficult to reproduce when the users change the operations. In another view, some methods have been applied to estimate the landmarks automatically on anatomical images []. In these methods, a sequence of image processing techniques have been hired to finish the estimation, i.e. SIFT, SURF [] are including 4 steps to give the estimated landmarks. One thing to note that if a step in these procedure provides a bad result, it will affect the final result and it will become a bottleneck of the method. For example, segmentation is the most important step in image processing. In some cases, the object in the image is easy to extract and we can analyze the object with the help of well-known image analysis procedures. But in other cases, when the image is noisy, applying segmentation will not give any hope. In the previous study [], we have studied two anatomical parts of the beetle: left and right mandibles. These parts just contain the interested object and pretty easy to segment. So, sequence algo-

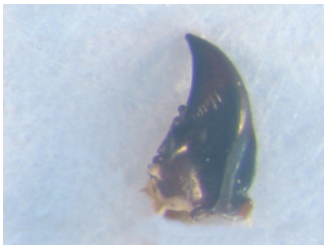
rithms have been applied to estimate the landmarks and they have provided good results. However, the results are not the same when we applied the same method to other parts of beetle, i.e. pronotum. So, considering a method to predict the landmarks without helping of segmentation step is needed.

In this work, we propose a CNN [1] which consists of several "elementary block". Each block is built from the popular layers in CNN, i.e. convolutional layer, max pooling layer, ... After designing, the proposed model will be trained on the dataset includes the images which are taken from 293 beetles. For each beetle, the biologists have taken images of five parts: *left and right mandibles, head, elytra, and pronotum* (Fig. 1). All the images are presented in the RGB color model with two dimensions. Along with each image, a set of landmarks has been marked by experts which can be used as ground truth to evaluate the predicted landmarks. During the experiments, the proposed network has been trained on the dataset by applying two strategies. In the first strategy, the network is trained from scratch on each dataset of each part; while in the second strategy, the training process has been modified to include a fine-tuning [2] stage. Besides, the value of Root Mean Square Error (RMSE) is used to compute the loss during two experiment processes. For more information about the model, you can see at our repository on GitHub: `https://github.com/linhlevandlu/CNN_Beetles_Landmarks`.

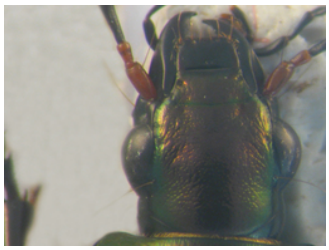
The rest of this paper is organized as followed: Section 2 discusses the related works of automatic estimation landmarks on 2D images. The, the dataset and the method which use to augment the dataset will be presented in Section 3. Section 4 shows the procedure of designing the network model. The first experiment of the network on each dataset is presented in Section 5. Section 6 presents a modification of training process and its experiment on datasets. Finally, the conclusion is given in Section 7.



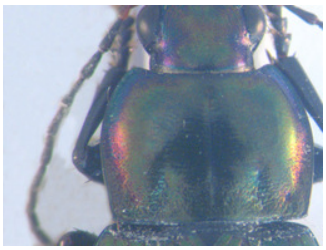
(a) Left mandible



(b) Right mandible



(c) Head



(d) Pronotum

(e) Elytra

Figure 1: The anatomical images of beetle are taken into account.

2. Related works

In geometric morphometry, landmarks (or points of interest) are important features to describe the shape. Depending on the complicacy of the objects in the image, setting automatic landmarks can rely on different methods. When the object can be segmented, the image processing techniques may applied to predict the landmarks. Lowe et al. [3] have proposed SIFT method to identify the keypoints on 2D images by extracting the distinctive features from the images. It is composed by four steps: (1) scale space extrema detection: a difference of Gaussian (DoG) function is applied to identify the interested points at all scales; (2) keypoint localization: the keypoint candidates are localized and refined by deleting the points which have the low contrast or not localized along the edge; (3) orientation assigment: for each keypoint candidates, the orientation and gradient magnitude are computed by considering their 4-neighborhoods; (4) keypoint descriptor are computed for each keypoint from its orientation and gradient magnitude. From the descriptors of the keypoints, they can be used to find the corresponding points between two images. SURF is another method which has been proposed by Herbert Bay et al. [4]. The SURF algorithm is the same principles with SIFT but details in each step are different. This algorithm mainly has three steps: (1) keypoints detection, (2) local neighborhood description and (3) matching. Different with SIFT, SURF uses a blob detector base on the Hessian matrix to find the keypoints. The Hessian matrix is used as a measure of local change around the point and chooses the points where this determinant is maximal. Then, the descriptors are computed around the key points by describing the intensity distribution of keypoint's neighborhoods. The matching points from different images are obtained by comparing the descriptors of the key points. Palaniswamy et al. [5] have proposed a method to automatically detect the landmarks on 2D images of Drosophila wings (Fig. 2). The method is mainly based on probabilistic Hough Transform. It includes four steps: (1) features detection of the fly wing structure (segmentation); (2) using pairwise geometric histogram (PGH) to record the compact invariant shape de-

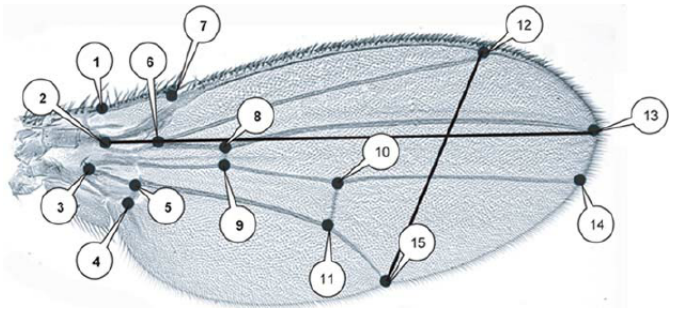


Figure 2: A Drosophila wing and its landmarks

scriptor; (3) estimating the global pose of wing using the probabilistic Hough transform; (4) and finally a template matching is applied to refine the correctly of individual features.

In recent years, deep learning [6] is known as a solution for the tasks in computer vision, especially for image analysis. Deep learning has been introduced in the middle of the previous century for artificial intelligence application but it has encountered several problems to take real-world cases. Luckily, the improvement of computing capacities, both in memory size and computing time with GPU programming, has opened a new perspective for deep learning. Many deep learning architectures have been proposed to sovle the problems of classification [7, 8], image recognition [9, 10, 11], speech recognition [12, 13], language translation [14, 15], ... Using deep learning, specifically CNN, to predict the landmarks on 2D images has achieved better results even if the images that can not segment. Yi Sun et al. [16] have proposed a cascaded CNNs to predict the facial points on the human face. Their model includes the networks which have separated into three levels of the cascade. The networks recognize the human face from the global to the local view to increasing the accuracy of predicted key points. Zhanpeng Zhang et al. [17] proposed a *Tasks-Constrained Deep Convolutional Network* to optimize facial landmarks detection. Their model detected the facial landmarks with a set of related tasks such as head pose estimation, gender classification, age estimation, or facial attribute inference. Cintas et al. [18] has introduced a network to predict the landmarks on human ears (Fig. 3).

After training, the network has the ability to detect 45 landmarks on human ears.

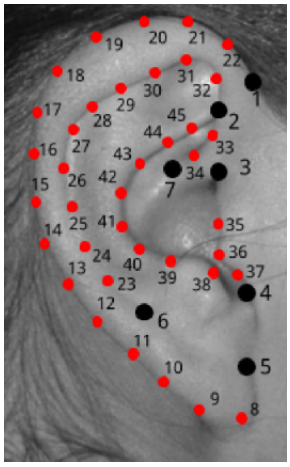


Figure 3: Landmarks on human ear in study of Cintas et al. [18]

As an association with the biologist, we study to develop an automatic method for predicting the landmarks on beetle's anatomical (Fig. 1). As we can see in this section, the methods for landmarking can be divided into two groups by considering the techniques that they have used. Most of the methods in the first group choose the image processing techniques while other methods are used deep learning. In a previous work [19], we have applied a series of algorithms to detect the landmarks automatically on beetles mandibles which are considered as the easied objects to segment (with an quality enough good for our need). In that work, the landmarks have been detected by registering two segmentations of images and then, using SIFT descriptor to refine the location of predicted landmarks. After the experiment, we have obtained good enough results on mandibles. Unfortunately, we have observed that the method did not provide good results when the segmentation is not precise, i.e. on pronotum or elytra images. This is explained why we have turned the automatically landmarking into another stage without any segmentation step. Using CNN for landmarking seems that a good choice for un-segmentable images.

In deep learning, the dataset is also an important component along with the network. Training a network on a large dataset will improve the learnable of the network because dataset has provided more different cases to learn. However, a large dataset is not always present in the practice. Instead of, the user applies some techniques to augment the dataset. Our case is not except. We work on a small dataset with 293 images of each part. This number is really modest with deep learning. So, a method has been through to augment the dataset. This procedure will be presenting in the next section.

3. Data augmentation

A characteristic of machine learning and deep learning is using a volume dataset to train the model. Of course, we are not always have enough data for training in practice. One way to solve this problem is to create the fake data from real data and to add it to the training set. Dataset augmentation has been a particularly effective technique for a specific problem. For example, in images classification problem, the operations like translating, rotating or scaling the images have also effective. The fake images may be generated by translating (rotating or scaling) in each direction. Besides, injecting noise in the input can also see as a form of data augmentation.

Our dataset includes 293 images of beetles (for each anatomical part). All the images are taken with the same camera in the same condition with a resolution of (3264×2448). Each image has a set of manual landmarks provided by biologists, i.e, each pronotum has 8 landmarks, each head has 10 landmarks. Applying CNNs to train each part with a small number of images to reach good results is impossible. So, we need to augment the dataset before training the networks. Firstly, we have found that the original solution of the images (3264×2448) are heavy for the neural network. For performance considerations, in most of CNNs [18, 1, 16], the size of the input is limited to (256×256) pixels, so we have decided to down-sampling the images to a new resolution (256×192) (to respect the ratio between x and y). Of course, the coordinates of manual landmarks

have been also scaled to fit with the new resolution of the images. In usual way, the transformations have been used to augment the dataset (i.e rotation, translation,...) but the analysis of image by CNN is most often translation and rotation invariant. Therefore, two other procedures have been imaged to increase the number of images in the dataset (256×192).

The first procedure is to change the value of a color channel in the original image to generate a new image. According to that, a constant is added to one of the RGB channels each time it is used for training. Each constant is sampled in a uniform distribution $\in [1, N]$ to obtain a new value caped at 255. For example, Fig. 4 shows an example when we added a constant $c = 10$ to each channel of an original image. Following this way, we can generate three version from an image.

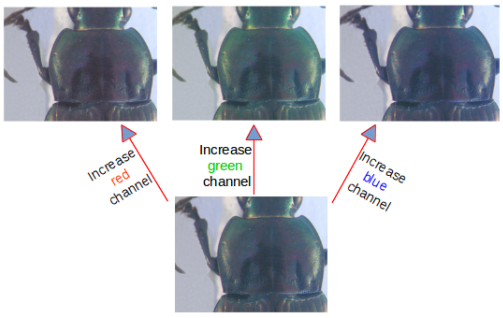


Figure 4: A constant $c = 10$ has been added to each channel of an original image

In the second procedure, we have applied the opposite procedure to the first one. Instead of adding the value, we separate the channels of RGB into three gray-scale images as the network works on single channel images (Fig. 5). At the end of the processes, we are able generate six versions from an original image. In total, we have $293 \times 7 = 2051$ images for each anatomical part of beetle (an original image and six generated images). However, we have not used all images for training and validation. So, we have chosen 260 original images and their generations (1820 images) of each dataset for training and validation processes, the remaining images (33 original images) are used for test process.

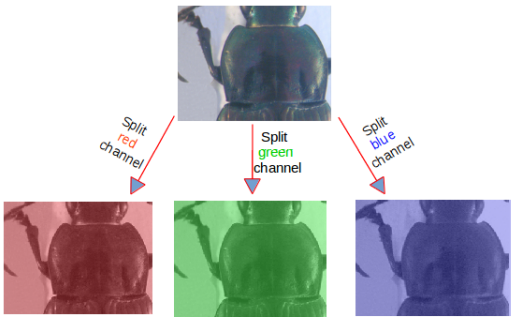


Figure 5: Three channels (red, green, blue) are separated from original image

In practical, to obtain a fast convergence during the computing, it is useful to normalize the brightness of the images to $[0, 1]$ instead of $[0, 255]$ and the coordinates of the landmarks have been also normalized [20].

4. Network architectures designing

In the process, we have tried three network models before deciding the final architecture for detecting the landmarks on beetle images. Like other CNN models, we have employed the classical layers to construct the models, i.e., convolutional layers, maximum pooling layers, dropout layers and full-connected layers.

The first architecture is very classical one, it receives an image with the size of $(1 \times 192 \times 256)$ as the input. Then, the network consists on three repeated strucutre of a convolutional layers followed by a maximum pooling layers. Most CNNs, the hyperparameters of convolutional layers have been set to increase the depth of the images from the first layer to the last layer. That is reflected in the setting of the number of filters at each convolutional layer. So, the depths of convolutional layers increase from $32, 64$, and 128 with different size of the kernels: (3×3), (2×2) and (2×2), respectively. Inserting pooling layers after a convolutional layers is a common periodcally. The pooling layer effects to progressively reduce the spatial size of the representation to reduce the number of parameters, computation in the network, and it also controls over-fitting.

The operations of pooling layers independent on every depth slice of the input. The most common form is a pooling layer with filters of size (2×2) and a stride of 2. It downsamples every depth by 2 along width and height of the input. Thefore, all the kernels of maximum pooling layers have the same size of (2×2) with a stride of 2 as usual. At the end of the model, three full-connected layers have been added to extract the global relationship between the features and to procedure the outputs. The first of two full-connected layers are set to non-linearity to make sure these nodes interact well and take into account all possible dependencies at the feature level. The outputs of the full-connected layers are $500, 500$ and 16. The output of the last full-connected layer corresponds to the coordinates (x and y) of 8 landmarks which we would like to predict. Fig. 6 shows details of the first model: The orange rectangles represent for convolutional layers while the yellow rectangles represent for maximum pooling layers and three full-connected layers with their parameters are presented at the end of the model.

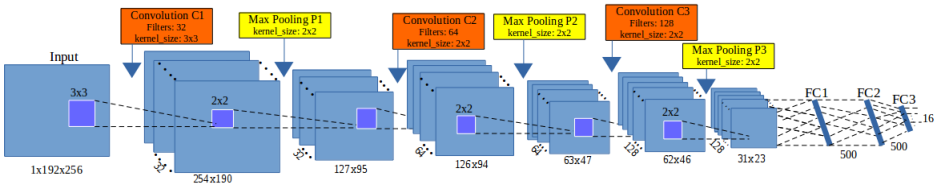


Figure 6: The architecture of the first model

The second architecture is modified from the first model. The layers are kept the same as the first one but the outputs of the first of two full-connected layers are changed from 500 (in the first model) to 1000 (Fig. 7). Increasing the value at full-connected layers is hoping to obtain more features from convolutional layer and to prevent the over-fitting.

To build the third architecture, we have used the definition of *elementary block*. An elementary block is defined as a sequence of convolution (C_i), maximum pooling (P_i) and dropout (D_i) layers (Fig. 8). This significantly reduces overfitting and gives major improvements over other regularization methods

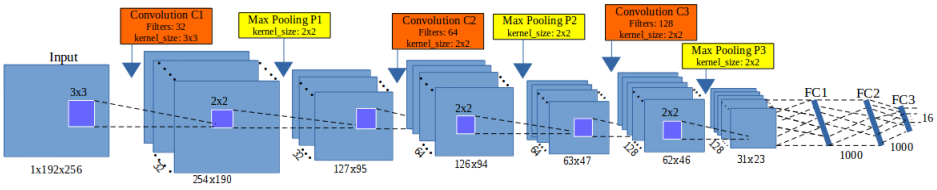


Figure 7: The architecture of the second model

[21]. The idea of dropout is to include some variations between different runs. During training phase, dropout samples are done from an exponential number of different "thinned" network. At test phase, it is easy to approximate the effect of averaging the prediction of all thinned networks by simply using a single unthinned network with smaller weights. So, we have modified the architecture by combining some *elementary blocks*.

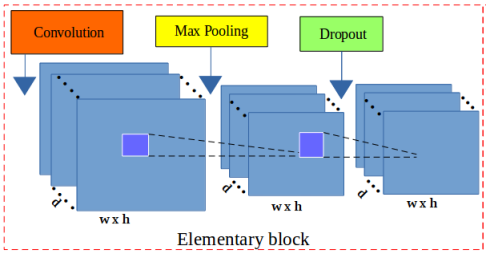


Figure 8: The layers in an elementary block

Fig. 9 illustrates the layers in the third architecture. For our purpose, we have assembled **3 elementary blocks**. The parameters for each layer in each elementary block are as below, the list of values follows the order of elementary blocks ($i = [1..3]$):

- CONV layers:
  - Number of filters: $32, 64$, and 128
  - Kernel filter sizes: $(3 \times 3), (2 \times 2)$, and (2×2)
  - Stride values: $1, 1$, and 1
  - No padding is used for CONV layers

- POOL layers:
  - Kernel filter sizes: $(2 \times 2), (2 \times 2)$, and (2×2)
  - Stride values: $2, 2$, and 2
  - No padding is used for POOL layers
- DROP layers:
  - Probabilites: $0.1, 0.2$, and 0.3

Three full-connected layers (FC) are kept the same as the second architecture: FC1 and FC2 have 1000 outputs, the last full-connected layer (FC3) has 16 outputs. As usual, a dropout layer is inserted between FC1 and FC2 with a probability equal to 0.5.

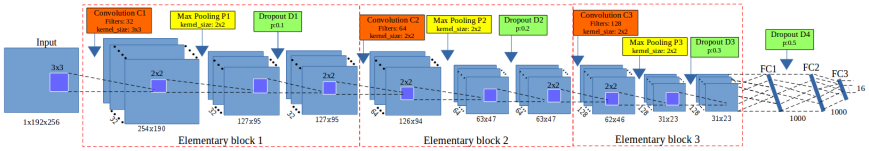


Figure 9: The architecture of the third model

The core of neural network is training over iteration. There are many ways to optimize the learning algorithm, but gradient descent [20] is currently a good choice to establish the way of optimizing the loss in neural network. The core idea is following the gradient until we statify with the results will remain the same. So, we have chosen gradient descent in the backward phase to update the values of learnable parameters and to increase the accuracy of the network. The networks are designed with a small sharing learning rate and a momentum. The learning rate is initialized at 0.03 and stopped at 0.00001, while the momentum is updated from 0.9 to 0.9999. Their values are updated over training time to fit with the number of epochs [2]. The implementation of the architectures have been done on Lasagne framework [22] by Python.

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

5. Experiments and results

Before widely applying to all anatomical parts, we have firstly tried with pronotum part to evaluate the performance. The networks have been trained in $5,000$ epochs on Ubuntu machine by using NVIDIA TITAN X cards. The set of images that used for training and validation are merged together. During the training, the images are chosen randomly from the dataset with a ratio of 60% for training and 40% for validation. The training step takes into account a pair of information (*images, manual landmarks coordinates*) as training data. In the context of deep learning, landmark prediction can be seen as a regression problem. So, we have used Root Mean Square Error (RMSE) to compute the loss of implemented architectures. At the test phase, images without landmarks are given to the trained network to produce output coordinates of the predicted landmarks. The results then evaluated by comparing with the manual landmarks coordinates provided by biologists which have been seen as ground truth.

Fig. 10 shows the training errors and the validation errors during traning phase of the first architecture. The blue curve presents the RMSE errors of training process while green curve is the validation errors. Clearly, over-fitting has appeared in the first model. The training losses are able to decrease but the validation losses are stable. In the second model (section 4), we have modified the parameters of full-connected layers to prevent the over-fitting but it seems that this solution is still not suitable.

Then, we have continued to train the third model on the same dataset of pronotum images. Fig. 11 illustrates the losses during the training of the third model. Like the previous figure (Fig. 10), the blue line is training losses, the green line is validation losses. In the opposite with two previous models, the losses are different (far) from the beginning but after several epochs, the values become more proximate and the over-fitting problem has been solved. This proves that adding dropout layers to build the elementary blocks have been effects to prevent over-fitting and contributory improve the accuracy of the

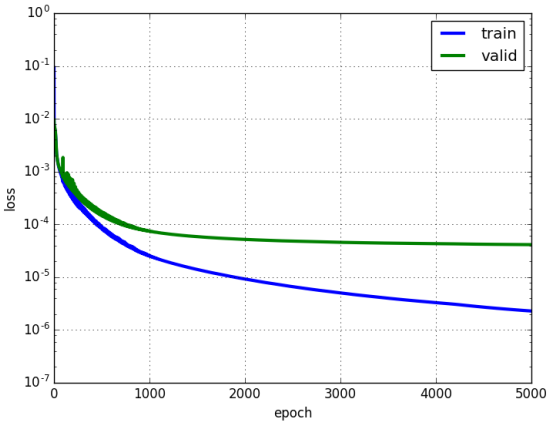


Figure 10: The training and validation losses of the first model

model. *So, we have decided to keep the architecture of the third model for our landmarking problem.*

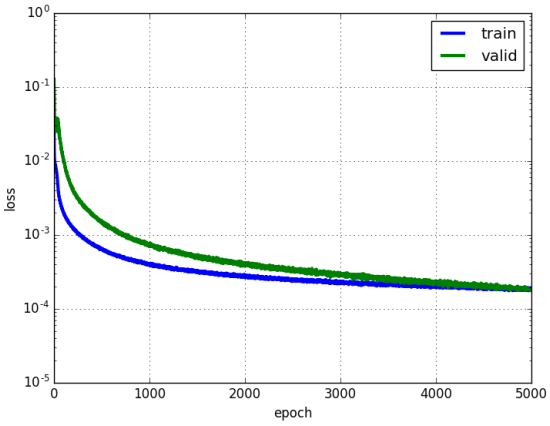


Figure 11: The training and validation losses of the third model

In order to have the predicted landmarks for all pronotum images (instead of only 33 images), we have applied *cross-validation* to choose the test images, called *round*. For each time, we have chosen a different fold of 33 images as testing images, the remaining images are used as training and validation images

(293/33 ≈ 9 rounds). Following that, the network will be trained with different datasets, then the trained model will be used to predicted the lanmarks on the images in the corresponding test set. Table. 1 resumes the losses of 9 rounds when we trained the third model on pronotum images. From the Table. 1, the losses among the rounds are stable (i.e ≈ 0.00020 for traning loss and ≈ 0.00024 for validation loss), they have a different but not so large. According that, the RMSE values are looking pretty good ($\approx 1.7 - 2.1$ pixels).

Round	Training loss	Validation loss
1	0.00018	0.00019
2	0.00019	0.00021
3	0.00019	0.00026
4	0.00021	0.00029
5	0.00021	0.00029
6	0.00019	0.00018
7	0.00018	0.00018
8	0.00018	0.00021
9	0.00020	0.00027

Table 1: The losses during training the third model on pronotum images

To evaluate the coordinates of predicted landmarks, the correlation metrics have been computed the correlation between the manual landmarks and their corresponding predicted one. Table. 2 shows the correlation scores of 3 metrics (using *scikit-learn* [23]), i.e, coefficient of determination (r^2), explained variance (EV), and Pearson correlation. All of three metrics have the same possibility. The best score is 1.0 if the correlation data is good, lower values are worse. It means that our predicted coordinates are very close with the ground truth. However, the measure is not enough good to provide a useful result to biologists. Moreover standing on the side of image processing, we are looking forward to seeing the predicted coordinates than the statistical results.

The main goal of computing is to predict the coordinates of landmarks,

Metric	$\mathbf{r^2}$	**EV**	**Pearson**
Score	**0.9952**	**0.9951**	**0.9974**

Table 2: Correlation scores between manual landmarks and predicted landmarks

so the distances (in pixels) between the coordinates of manual landmarks and corresponding predicted landmarks have been taken into account on all images. Then, the average of distances are computed by landmarks. Table. 3 shows the average distances by landmarks on all images of pronotum dataset. With images of resolution 256×192, we can consider that an error of 1% corresponds to 2 pixels that could be an acceptable error. Unhappily, our results exhibit average distance of 4 pixels in the best case, landmark 1 and more than 5 pixels, landmark 6. Other error distances are more than 2% pixels.

Landmark	**Distance** (in pixels)
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 3: The average distances on all images per landmark on pronotum images.

Fig. 12 shows the distribution of the distances on the first landmark of all images. The accuracy based on the distance in each image can be separated into three spaces: the images have the distance less than average value (4 pixels): 56.66%; the images have the distance from average value to 10 pixels (5% acceptable errors): 40.27%; and the images have the distance greater than 10 pixels: 3.07%.

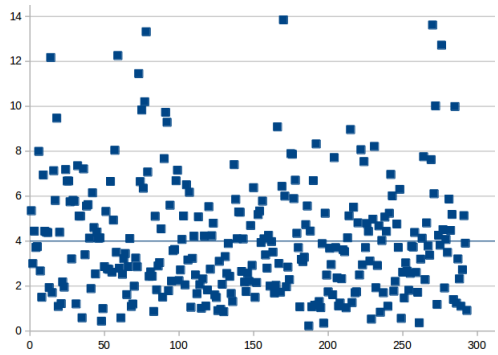
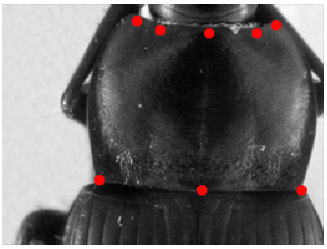
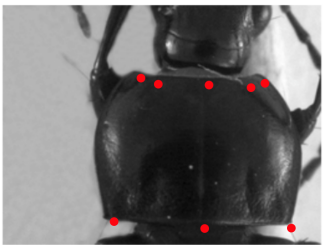


Figure 12: The distribution of the distances on the first landmark. The blue line is the average value of all distances.

To illustrate this purpose, Fig. 13 shows the predicted landmarks on two test images. One can note that even some predicted landmarks (Fig. 13a) are closed to the manual ones, in some case (Fig. 13b) the predicted ones are far from the expect results. This result explains why the average distance by landmarks are enough good while some predicted landmarks are so far from the manual one. So, the next step has been dedicated to the improvement of these results.



(a) Image with well-predicted landmarks



(b) Image with inaccuracy landmarks

Figure 13: The predicted landmarks, in red, on the images in test set.

From the success of the third architecture on pronotum dataset, we apply the same procedures (data augmentation, training,...) on other parts of beetle: *left and right mandibles, elytra and head*. However, we have modified the number of output of the last full-connected layer to adapt with each dataset before training.

According, the values at the last full-connected layer are set to 32, 36, 22 and 20 outputs corresponds to 16, 18, 11 and 10 landmarks on left mandible, right mandible, elytra and head, respectively. Of course, we have also applied *cross-validation* to select testing data to get all predicted landmarks for all images in each dataset. Then, the quality of predicted landmarks are evaluated by comparing with the corresponding manual landmarks (distance computation). Table. 4 shows the average distances on each landmark of elytra, head, left and right mandibles anatomical, respectively. Comparing with the average distances on the pronotum part, it seems that the proposed architecture provides more accurate predictions on the elytra dataset, but the results are opposite on other datasets.

6. Resulting improvement by fine-tuning

The proposed network (third architecture) presented in Section 4 have been trained from scratch on five datasets of beetles (left mandible, right mandible, pronotum, elytra, and head). At the first step, the network was able to predict the landmarks on the images. But as we have discussed, even if the strength of the correlation seems to validate the results, when we display the predicted landmarks on the images, the quality of the predicted coordinates are also not enough precise, and the average error are also still high (of course, we have the distances are higher than the average distances).

In order to reach more acceptable results for biologists, we have broadened model with another step of deep learning: **transfer learning**. That is a method enables to re-uses the model developed for a specific task/dataset to lead another task (called *target task*) with another dataset. This process allows rapid process and improves the performance of the model on the target task [24]. The most popular example has been given with the project ImageNet of Google [25] which has labeled several millions of images. The obtained parameter values which can be used in another context to classify another dataset, eventually very different dataset [26]. The name of this procedure to re-use parameters to pretrain a

Landmark	**Distance (in pixels)**			
	Right mandible	Left mandible	Elytra	Head
1	9.4981	9.1267	3.8669	5.528
2	7.1657	6.7198	3.973	5.1609
3	7.242	6.8704	3.9166	5.3827
4	7.0436	6.7719	3.8673	5.0345
5	7.1599	7.125	4.0151	4.8393
6	7.5699	6.9441	4.8426	4.4516
7	7.4251	7.3158	5.2125	4.7937
8	7.6636	7.4142	5.4685	4.5322
9	7.7906	7.5846	5.2692	5.1412
10	8.0197	7.6349	4.0709	5.0564
11	8.314	7.6873	3.9896	-
12	8.1564	8.4248	-	-
13	8.8879	7.9983	-	-
14	9.1842	7.4919	-	-
15	8.7875	7.7903	-	-
16	8.3141	8.5198	-	-
17	8.2866	-	-	-
18	8.8928	-	-	-

Table 4: The average distances on all images per landmark on left mandible, right mandible, elytra and head images.

model is currently called **fine-tuning**.

Fine-tuning does not only replace and retrain the model on the new dataset but also fine-tunes the weights of a trained model by continuing the backpropagation. Unfortunately, some rapid tests have shown that re-using ImageNet features has not been relevant for our application. We have designed a way to reproduce the method with our own data. It is worth noting that of course the size of data to pre-train has drastically decreased. For our pre-training step, the network has been trained on the whole dataset including the images of three parts of beetle *i.e pronotum, elytra and head*. Then, the trained model has been used to fine-tune and test on each dataset.

6.1. Data preparation and training

The images training dataset is combined from the images of three sets: *pronotum, elytra, and head* (after augmentation). When applying the training from the scratch, we have used cross-validation to select the data (9 folds). It means that for each dataset, we have some different training data and corresponding testing data. So, the images that use to train the model are just select from one of the folds in each dataset. Specifically, we have taken $1,820$ images of each part. In total, it includes $5,460$ images ($260 \times 7 \times 3$).

However, another problem has been appeared when we combined the images from different dataset. That is the different number of landmarks on each part: 8 *landmarks on pronotum part,* 10 *landmarks on head part, and* 11 *landmarks on elytra part*. Fig. 14 shows the possition of the landmarks on each part. Because of the meaning of landmarks on each anatomical part for biologists, we cannot insert the landmarks arbitrary. So, we have decided to keep the landmarks on pronotum as reference and to remove the landmarks on elytra and head parts instead of adding. We kept 8 (landmarks) as a reference number, then we have removed the supernumerary when it is unnecessary. Specifically, we have removed three landmarks on the elytra part ($1^{st}, 6^{th}, 9^{th}$), and two landmarks on the head part ($5^{th}, 6^{th}$).

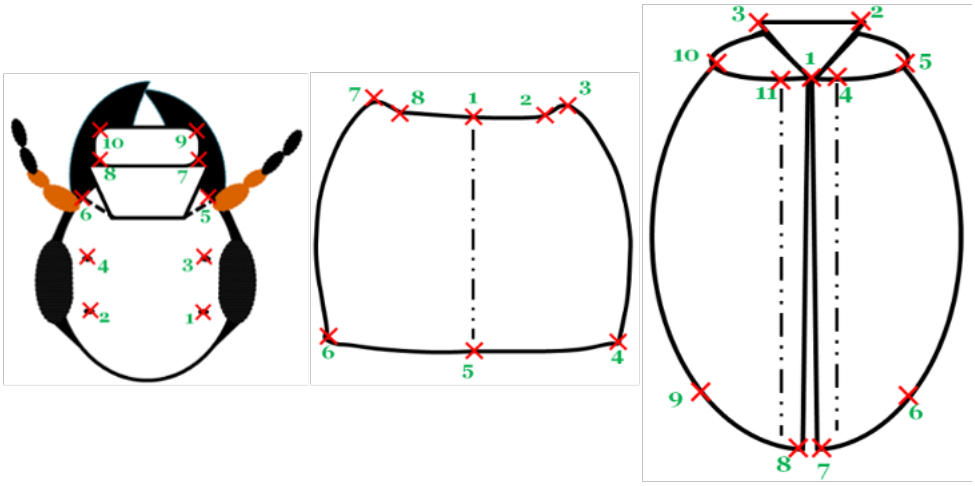


Figure 14: A presentation of head, pronotum and elytra part with corresponding manual landmarks

During training the proposed architecture on the combined dataset, the parameters of the network (learning rate, momentum, ...) are kept the same as training from scratch but the number of epochs are increased to $10,000$ instead of $5,000$ to achieve better learning on the parameters (weights). Additional, we have shuffled the training set. Because the neural network learns the faster from the most unexpected sample. It is advisable to choose a sample at each iteration that is the most unfamiliar to the system. Shuffling the examples will be helped the model works with different anatomical parts rather than the same anatomical samples in each training time.

6.2. Fine-tuning on each dataset

The combined dataset then used to train the third architecture with 16 outputs (8 landmarks). Then, the trained model is used to fine-tuning on each dataset. To compare the result with the previous one, we have also fine-tuned the trained model with different dataset by applying cross-validation. Firstly, we consider on the losses during fine-tuning. *For example*, Table. 5, 7, 9 show the losses during fine-tuning on pronotum, elytra, and head dataset, respectively. Comparing with the losses when we trained the model from scratch,

i.e. on pronotum, the validation losses of all round in this scenario have been significantly decreased (around 40%).

On each part, the landmarks are predicted on the test images. Then, the average error based on the distances between predicted and corresponding manual landmarks have been also computed. Table. 6, 8, 10, 11, and 12 show the average distances per landmark on pronotum, elytra, head, left and right mandibles dataset, respectively. **From scratch** columns remind the previously average distances. **Fine-tune** columns present the new average distances after applying fine-tuning on each part. It is clearly shown that the result of predicted landmarks with the help of fine-tuning is more precise than training from scratch. For example, the average distance at each landmark has decreased. Additional, when comparing the average distances between two processes, the worse case of fine-tuning process is still better than the best case of training from scratch.

Round	Training loss	Validation loss
1	0.00019	0.00009
2	0.00018	0.00010
3	0.00018	0.00010
4	0.00019	0.00008
5	0.00019	0.00009
6	0.00018	0.00008
7	0.00019	0.00008
8	0.00018	0.00006
9	0.00018	0.00009

Table 5: The losses during fine-tuning model on pronotum dataset

#LM	**From scratch**	**Fine-tune**
1	**4.00**	**2.49**
2	4.48	2.72
3	4.30	2.65
4	4.39	2.77
5	4.29	2.49
6	**5.36**	**3.05**
7	4.64	2.68
8	4.94	2.87

Table 6: The average error distance per landmark of two processes on pronotum images

In another view, Fig. 15 shows the comparation of the average distance distribution on each dataset in two procedures (from scratch and fine-tuning). In which:

Round	Training loss	Validation loss
1	0.00020	0.00006
2	0.00020	0.00006
3	0.00021	0.00006
4	0.00021	0.00006
5	0.00019	0.00006
6	0.00019	0.00006
7	0.00018	0.00005
8	0.00020	0.00006
9	0.00019	0.00006

Table 7: The losses during fine-tuning model on elytra dataset

#LM	From scratch	Fine-tune
1	**3.87**	2.34
2	3.97	2.27
3	3.92	2.27
4	**3.87**	**2.25**
5	4.02	2.27
6	4.84	3.14
7	5.21	3.14
8	**5.47**	3.29
9	5.27	**3.42**
10	4.07	2.49
11	3.99	2.30

Table 8: The average error distance per landmark of two processes on elytra images

Round	Training loss	Validation loss
1	0.00022	0.00007
2	0.00022	0.00007
3	0.00023	0.00008
4	0.00023	0.00008
5	0.00022	0.00008
6	0.00023	0.00007
7	0.00022	0.00008
8	0.00023	0.00007
9	0.00024	0.00008

Table 9: The losses during fine-tuning model on head dataset

#LM	From scratch	Fine-tune
1	5.53	**3.03**
2	5.16	2.94
3	**5.38**	2.96
4	5.03	2.88
5	4.84	2.76
6	**4.45**	2.67
7	4.79	2.29
8	4.53	**2.20**
9	5.14	2.57
10	5.06	2.44

Table 10: The average error distance per landmark of two processes on head images

#LM	From scratch	Fine-tune
1	9.1267	6.7655
2	6.7198	5.2952
3	6.8704	5.3468
4	6.7719	5.332
5	7.125	5.4391
6	6.9441	5.3004
7	7.3158	5.5314
8	7.4142	5.6486
9	7.5846	5.8864
10	7.6349	5.9245
11	7.6873	5.972
12	8.4248	6.5755
13	7.9983	6.1067
14	7.4919	5.6307
15	7.7903	5.8522
16	8.5198	7.174

Table 11: The average error distance per landmark of two processes on left mandible images

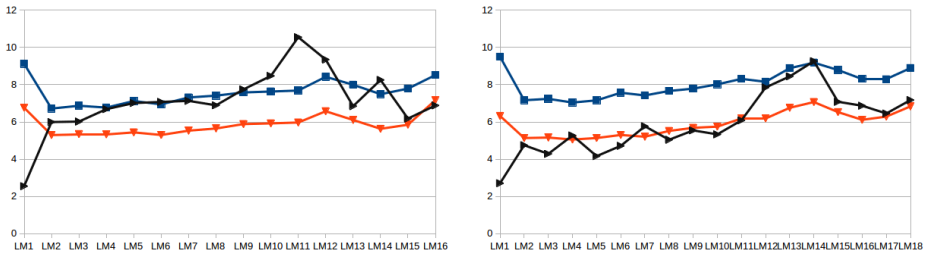
#LM	From scratch	Fine-tune
1	9.4981	6.3236
2	7.1657	5.1347
3	7.242	5.1613
4	7.0436	5.0537
5	7.1599	5.1372
6	7.5699	5.301
7	7.4251	5.2064
8	7.6636	5.5168
9	7.7906	5.6858
10	8.0197	5.7495
11	8.314	6.1975
12	8.1564	6.1898
13	8.8879	6.7612
14	9.1842	7.0694
15	8.7875	6.5293
16	8.3141	6.1147
17	8.2866	6.2881
18	8.8928	6.8367

Table 12: The average error distance per landmark of two processes on head images

- **Blue** curves: present for the average distances on each landmarks when we train the model from scratch.
- **Orange** curves: describe for the average distance on each landmark when we fine-tune the trained model.
- **Black** curves (in the case of left and right mandibles): illuslate for the average distances when we applied the image processing techniques to predict the landmarks on segmentable images.

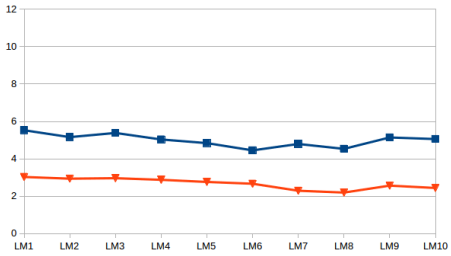
The fine-tuning process has improved the results of the proposed architecture on both 5 datasets: left, right mandible, pronotum, elytra and head. All the average distances are significantly decreased. Specially, the results have been improved $\approx 26.9\%$ on left mandible, $\approx 22.8\%$ on right mandible, $\approx 40.3\%$ on pronotum, $\approx 39.8\%$ on elytra, and $\approx 46.4\%$ on head part based on considering the average distances per landmark. Addition, in the cases of pronotum and head part, even if we plus the average distance and its standard deviation, the results are also less than the result when we trained the model from scratch. For segmentable images, we have a comparison between the results of deep learning and early method where we have applied image processing techniques to predict the landmarks. Clearly, the result with fine-tuning have improved the location of estimated landmarks, even when the average distances are still high when we trained the model from scratch: most of the average distance(or landmarks) of left mandibles are less than the results of the early method, while the average distances are very closed in the case of right mandibles.

To illustrate the final results, we display the distribution of the distances in both cases: the best and the worst results (resp. landmark 1 and 6 on pronotum dataset) of five datasets (pronotum, elytra, head, left and right mandible) in Fig. 16, 17, 18, 19, and 20, respectively. In these figures, the left images present for the results when we trained the model from scratch; while the right images shows the results after applying fine-tuning. The blue lines in the charts present the average distance values. Clearly, the results in the fine-tuning case have been improved significantly than training from scratch.

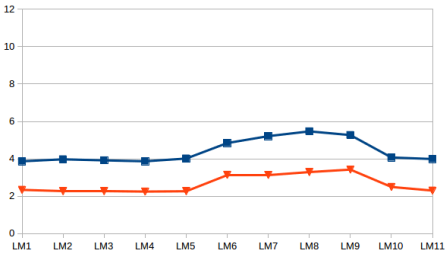


(a) Left mandible

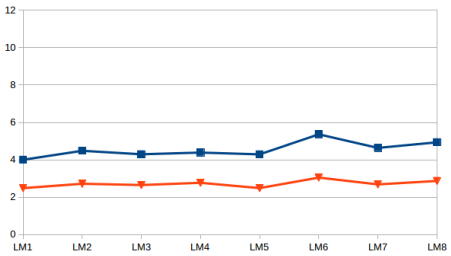
(b) Right mandible



(c) Head

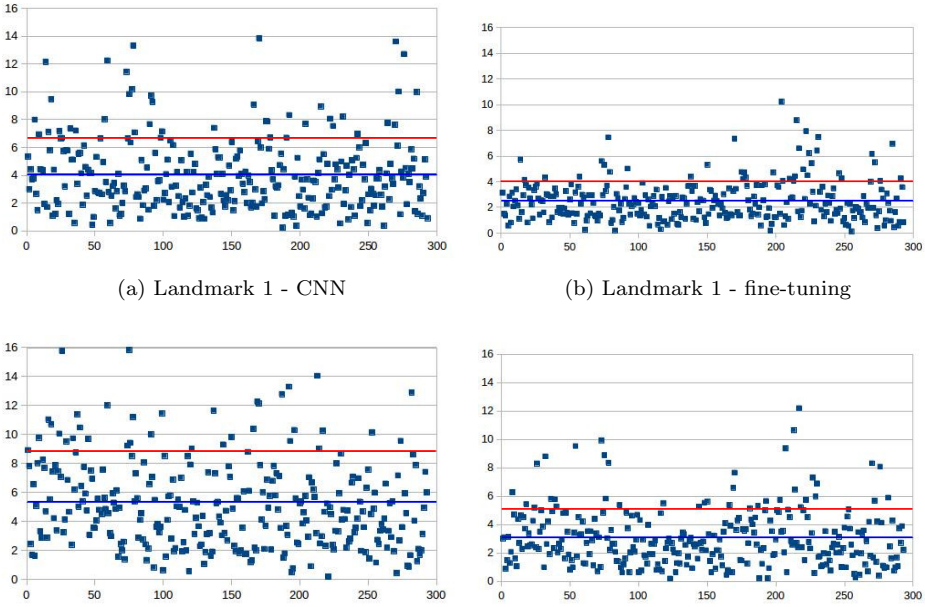


(d) Elytra



(e) Pronotum

Figure 15: The distribution of average distances on each landmark of each beetle's anatomical



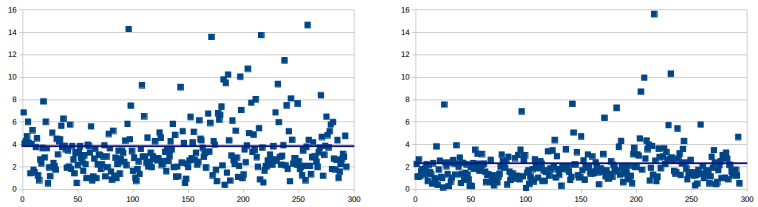
(a) Landmark 1 - CNN

(b) Landmark 1 - fine-tuning

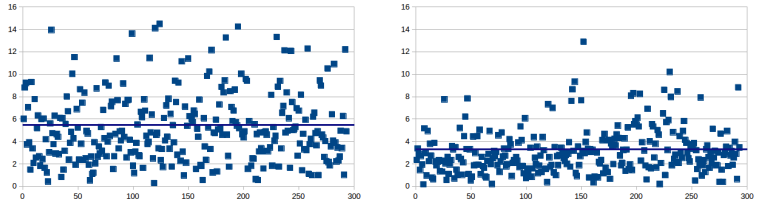
(c) Landmark 6 - CNN

(d) Landmark 6 - fine-tuning

Figure 16: The distribution of distances on 1^{st} and 6^{th} landmarks of all images in two testing steps (CNN and fine-tuning) (on **pronotum** dataset). The red line presents the standard deviation value.

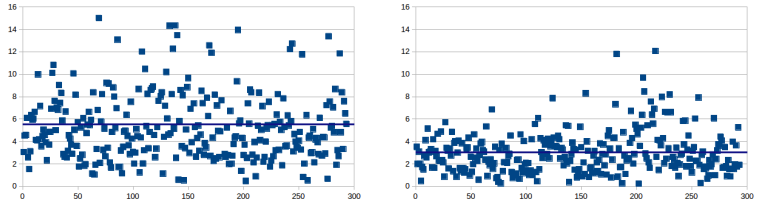


(a) Landmark 1 - CNN (b) Landmark 1 - fine-tuning

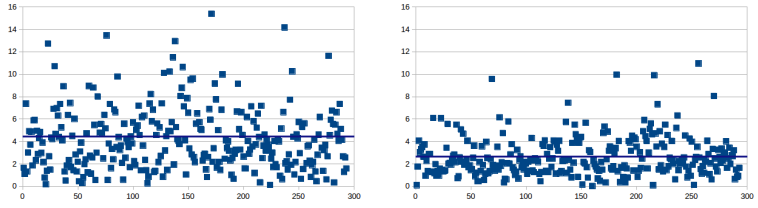


(c) Landmark 8 - CNN (d) Landmark 8 - fine-tuning

Figure 17: The distribution of distances on 1^{st} and 8^{th} landmarks of all images in two testing steps (CNN and fine-tuning) (on **elytra** dataset).

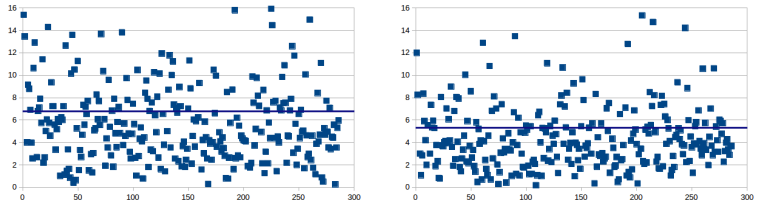


(a) Landmark 1 - CNN (b) Landmark 1 - fine-tuning

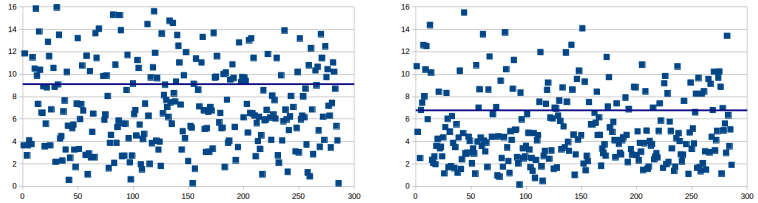


(c) Landmark 6 - CNN (d) Landmark 6 - fine-tuning

Figure 18: The distribution of distances on 1^{st} and 6^{th} landmarks of all images in two testing steps (CNN and fine-tuning) (on **head** dataset).

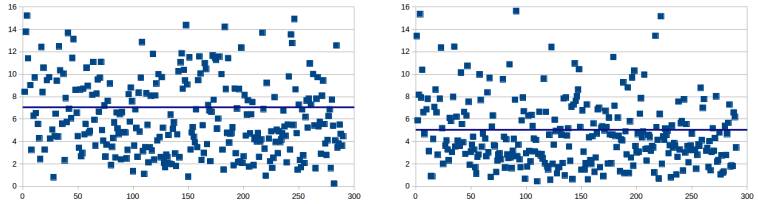


(a) Landmark 4 - CNN (b) Landmark 4 - fine-tuning

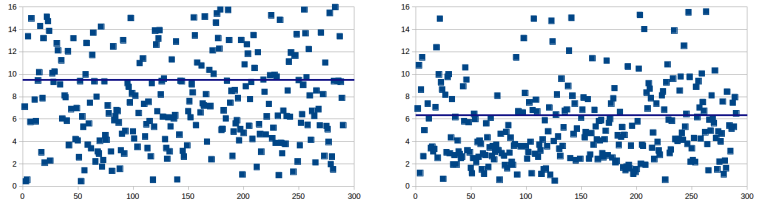


(c) Landmark 1 - CNN (d) Landmark 1 - fine-tuning

Figure 19: The distribution of distances on 1^{st} and 4^{th} landmarks of all images in two testing steps (CNN and fine-tuning) (on **left mandible** dataset).



(a) Landmark 4 - CNN (b) Landmark 4 - fine-tuning



(c) Landmark 1 - CNN (d) Landmark 1 - fine-tuning

Figure 20: The distribution of distances on 1^{st} and 4^{th} landmarks of all images in two testing steps (CNN and fine-tuning) (on **right mandible** dataset).

7. Conclusion

In this work, we have presented how to apply convolutional neural network to predict the landmark on 2D anatomical images of beetles. After going through many trial models, we have presented a convolutional neural network for automatic detection landmarks on anatomical images of beetles which includes the repeated of some elementary blocks (an elementary block consists of a convolutional layer, a max pooling layer, and a dropout layer) followed by fully connected layers. Then, the proposed model have been trained and tested by using two strategies: *train from scratch* and *fine-tuning*.

In our case, the size of dataset is limited. Therefore, we have applied the image processing techniques to augment dataset. The predicted landmarks have been evaluated by calculating the distance between manual landmarks and corresponding predicted landmarks. Then, the average of distance errors on each landmarks has been considered.

The results have been shown that using the convolutional network to predict the landmarks on biological images leads to satisfying results without need for segmentation step on the object of interest. The best set of estimated landmarks has been obtained after a step of fine-tuning using the whole set of images that we have for the project, i.e. about all beetle parts. The quality of prediction allows using automatic landmarking to replace the manual ones.

References

[1] Y. LeCun, K. Kavukcuoglu, C. Farabet, Convolutional networks and applications in vision, in: Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, IEEE, 2010, pp. 253–256.

[2] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Advances in neural information processing systems, 2014, pp. 3320–3328.

[3] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[4] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: European conference on computer vision, Springer, 2006, pp. 404–417.

[5] S. Palaniswamy, N. A. Thacker, C. P. Klingenberg, Automatic identification of landmarks in digital images, IET Computer Vision 4 (4) (2010) 247–260.

[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[8] D. Ciregan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 3642–3649.

[9] C. Szegedy, et al., Going deeper with convolutions, Cvpr, 2015.

[10] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE transactions on pattern analysis and machine intelligence 35 (8) (2013) 1915–1929.

[11] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325–5334.

[12] T. Mikolov, et al., Strategies for training large scale neural network language models, in: Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, IEEE, 2011, pp. 196–201.

[13] G. Hinton, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine 29 (6) (2012) 82–97.

[14] S. Jean, K. Cho, R. Memisevic, Y. Bengio, On using very large target vocabulary for neural machine translation, arXiv preprint arXiv:1412.2007.

[15] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Advances in neural information processing systems, 2014, pp. 3104–3112.

[16] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3476–3483.

[17] Z. Zhang, et al., Facial landmark detection by deep multi-task learning, in: European Conference on Computer Vision, Springer, 2014, pp. 94–108.

[18] C. Cintas, et al., Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks, IET Biometrics 6 (3) (2016) 211–223.

[19] V. L. Le, M. Beurton-Aimar, A. Krahenbuhl, N. Parisey, MAELab: a framework to automatize landmark estimation, in: WSCG 2017, Plzen, Czech Republic, 2017.
URL https://hal.archives-ouvertes.fr/hal-01571440

[20] Y. A. LeCun, et al., Efficient backprop, in: Neural networks: Tricks of the trade, Springer, 2012, pp. 9–48.

[21] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting., Journal of machine learning research 15 (1) (2014) 1929–1958.

[22] S. Dieleman, et al., Lasagne: First release. (Aug. 2015). doi:10.5281/zenodo.27878.
URL http://dx.doi.org/10.5281/zenodo.27878

[23] P. et al, Scikit-learn: Machine learning in python, Journal of machine learning research 12 (Oct) (2011) 2825–2830.

[24] L. Torrey, J. Shavlik, Transfer learning, Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques 1 (2009) 242.

[25] J. Deng, et al., ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

[26] J. Margeta, et al., Fine-tuned convolutional neural nets for cardiac mri acquisition plane recognition, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 5 (5) (2017) 339–349. arXiv:https://doi.org/10.1080/21681163.2015.1061448, doi:10.1080/21681163.2015.1061448.
URL https://doi.org/10.1080/21681163.2015.1061448