The Improved Faster R-CNN for Detecting Small Facial Landmarks on Vietnamese Human Face Based on Clinical Diagnosis

Ho Nguyen Anh Tuan
Human Anatomy Department, University of Medicine Pham Ngoc Thach, Ho Chi Minh City, Vietnam
Email: drhotuan.pnt@gmail.com

Nguyen Dao Xuan Hai
Faculty of Mechatronic, Ho Chi Minh University of Technology and Education, Ho Chi Minh City, Vietnam
Email: ndxuanhai@gmail.com

Nguyen Truong Thinh
Ho Chi Minh University of Technology and Education, Ho Chi Minh City, Vietnam
Email: thinhnt@hcmute.edu.vn

***Abstract*—Facial landmarks are places on the human face that are used to explain morphology. Facial landmarks are employed in anthropology to minimize the quantity of data utilized in morphological analysis, and they have become one of the pillars of anthropological study over the years. Their applications in facial analysis range from identifying morphological changes throughout facial development to identifying facial similarities. Despite its extensive use, most of the research in automatic facial landmark recognition has been motivated by the necessity for automation of security activities such as face identification and verification. Approaches using 2D pictures models have been widely researched in order to offer reliable face landmark recognition in a variety of real-world scenarios. Many current techniques, for example, rely on landmark recognition on low resolution data with significant out-of-plane rotation1 and occlusions. This is in sharp contrast to the data utilized in anthropology, which is comprised of high quality 2D photometry. The data is frequently collected from consenting participants and adjusted to minimize non-shape facial changes, such as varying size owing to the subject's distance from the capturing equipment or head rotation. The datasets in anthropology purpose are unavailable in comparison to those used for training current state-of-the-art facial landmark detection approaches based on Improved Faster Region Convolutional Neural Networks (Faster R-CNN). Therefore, an improved in data of anthropometric images by augmentation and change Roi Pooling by Roi Align that could promise an optimized and enhance the accurate with less time.**

***Index Terms*—faster R-CNN, small facial landmark, Vietnamese face, landmarks detector, anthropology**

I. Introduction

Anthropology is a field of study that is constantly a hot issue in developing countries. It is a basis for developing surgical procedures or redesigning components for a variety of objectives. Since the development of researching of anthropology in medical. Rhinoplasty has been significant changed with the help of specialist doctors and data collected using manual measuring devices (clamps, angle measure, etc). Since then, it has provided the best alternatives for cosmetic surgery, nasal lifting, and repairing cancer-related nose damage. This task requires a significant amount of time and effort, but the recover results are slow, significant divergence. As a result, in order to have a dependable precision, we must measure several times, which takes a lot of money and time-consuming. With the rapid developing of technology, the typical approach is to indirectly measure just by the photo using photographs by supporting of Cephalometric imaging system. By this way, it makes easier in studies the relationships between bony and soft tissue landmarks [1], [2]. It is often used in order to diagnose facial growth abnormalities prior to treatment, in the middle of treatment to measure progress, or at the end of a treatment to determine if the treatment goals have been met. However, penetrating form of high-energy electromagnetic radiation is a highly impacted on human health, which has often raised concerns of facial anthropology proposed as alternative methods. Therefore, several hospitals, including a well-known research facilities have used 3D image scanning such as 3dMDface or Vectra H1 to collect and analyzing 3D facial images with excellent precision [3]. Despite the fact that technology and methods have resulted in more accurate and less deformed handle measurement results, this has been accompanied with a surprisingly high cost investment and high data collecting making it difficult to deploy globally. Furthermore, managing landmark locations and taking measurements is not only time-consuming, but also given subjective results by human factor, particularly in concealed areas, when use of

Manuscript received November 17, 2021; revised April 11, 2022.

doi: 10.18178/joig.10.2.76-81

common measuring equipment for many sensitive locations, such as eyes and lips, is at danger of transmitting undesired source illnesses, thus each measure must be conducted by each person. From previous studies as well as with the desire to contribute a few research results in the process of automatic anthropometric identification for the most simplest and safest way to minimize costs and times. The diagnostic time, this study has achieved some good results assessed to the results due to the manual measurement and previously troubleshooting methods.

The datasets available in anthropology are usually small in comparison to those used for training current state-of-the-art facial landmark detection approaches based on Convolutional Neural Networks (CNN). For example VGGFace2 [3], a CNN-based approach used for face recognition across different poses and ages, used a dataset of over 3.3 million images, with over 9000 unique subjects captured. This is mainly due to the different demands for facial landmark detection in anthropology. Many of the state-of-the-art facial landmark detection approaches are not driven by anatomical principals, but rather focus on detecting landmarks that would best describe the facial shape. These landmarks are not placed in biologically defined positions, and as such, they do not guarantee homology between the studied subjects. Facial landmarks are merely a stepping stone to a number of different tasks in anthropology, most of which rely on statistical analysis to quantitatively describe a morphology of an individual, or a studied group. Especially in case a large number of faces, such as is the case in biological anthropology, there's a requirement for the detected landmarks to be homologous between the studied faces.

II. Facial Landmarks Based on Anthropometry

This study has been exempt from ethical approval, based on a written reply from the ethical commission. After obtaining permission, 2200 photogrammetry training samples and 182 testing samples were selected and collected data from the records of the Human Anatomy Department, Pham Ngoc Thach University of medicine, based on the following criteria: Firstly, all Vietnamese origin, as a general training group data with the age mean was 35.09±11.56 year-old. The testing group with the age mean was 22.01±1.39. Secondly, all participants are well visibility and nasal structures never do surgery or damage. Prior to capture photo for this study, all the volunteers were examined for scars or facial deformities or trauma, no previous maxillofacial or plastic surgery and no nasal dermatitis. All participants declared to have at least three generations of Asian ancestry. Data were collected between 2019 and 2020. Photos are captured as standard head position which each patient looking forward made sure to keep the Frankfurt in horizontal plane.

For evaluating the model in testing data, at first, the operator found and noted the usual landmarks on the cutaneous surface with thorough scrutiny and palpation. During the landmarks marking, the respondents sat comfortably in a position that allowed them to correctly identify face characteristics. The accepted definitions are developed on 2D pictures and originate from Farkas [4] under the supervision of conventional procedures and are based on anthropometric measures in the craniofacial area. Images were collected using a Canon digital camera (Canon, Inc, Tokyo, Japan). The training file is taken from 2200 samples by the frontal view and lateral view which were augmentation and labeled for supervised learning.

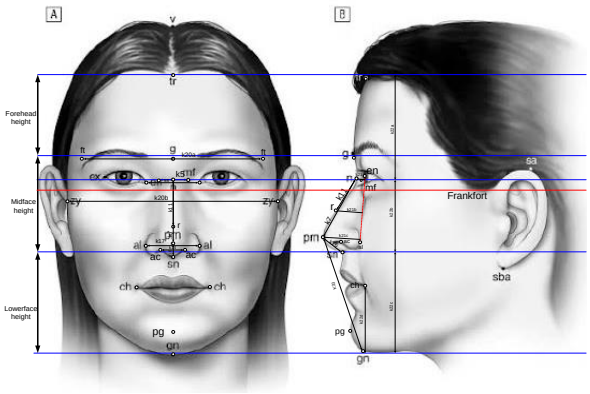


Figure 1. Ground truth position of the anthropometric landmarks on the face in opposite of profile view and from the right side of lateral view.

In the lateral view, linear measurements were as follows: nasal height (n-sn), nasal bridge length (n-prn), nasal tip protrusion (sn-prn) (see Table I).

TABLE I. Name of Lanmarks Shown in Fig. 1

Abbreviation	Name
tr	trichion
Eyes	
en	Endocanthion Right and Left
ex	Exocanthion Right and Left
Nose	
g	Glabella
sn	Subnasale
mf	maxillofrontale
al	Alare Right and Left
n	Nasion
r	Rhinion
prn	Pronasale
sn	
Mouth and Chin	
ch	cheilion
pg	Pogonion
gn	Gnation

In the frontal view, linear measurements were as follows: intercanthal width (en'-en'), nasal width (al'-al'), nasal base width (ac_{right}-ac_{left}), ala length (ac'-prn, ac'-prn), Angular measurements were as follows Fig. 1: (1) Nasofrontal angle, which is measured between the proximal nasal bridge contour and the anterior surface of the forehead below the glabella; (2) Nasal tip angle, which is formed by the lines following the general direction of the columella and the nasal bridge; and (3) Nasolabial angle, which is measured between the surfaces of the columella and of the upper lip skin.

III. Improved Faster R-CNN for Facial Landmarks Detection

Girshick, Ross developed Fast R-CNN (2015) [5] takes a step forward. Instead than applying CNN to recommended areas 2,000 times, it just sends the original picture to a pre-trained CNN model once. Search selection method is generated using the preceding step's output feature map. The ROI pooling layer is then utilized to guarantee that the output size is standard and pre-defined. As inputs, these valid outputs are delivered to a fully linked layer. Finally, a softmax classifier is utilized to predict the observed target, and a linear regressor is used to modify bounding box localizations. Faster R-CNN advances faster than Fast R-CNN. Region Proposal Network (RPN) replaced Search selective process. As the name revealed, RPN is a network to propose regions.

Faster R-CNN is a supervised learning with labels, which can be divided into 3 stages: feature extraction networks called Backbone, Region Proposal Network (RPN), and RoI network named Classifer. In this study, the principle architecture of target landmarks detection using Improved Faster R-CNN is shown in Fig. 2.

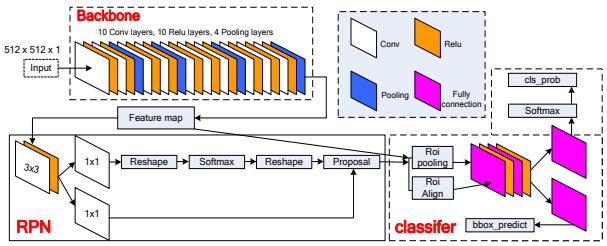


Figure 2. Faster R-CNN architecture using.

1) Input an image of any size, and perform feature extraction through a feature extraction network, including Convolutional layers, Batch Normalization (BN) layers and Pooling layers. Backbone used Convolutional Neural Network is an important stage to identification and image classification. Multi-layer Convolution uses the characteristics of the image. With each different kernel extracted one specific characteristic of the image, therefore each convolutional layer was transferred information of feature, at the end, many properties of the image in feature map were being moved to RPN stage. Pooling layers are usually used between Convolutional Layers to principal component of data but still retain the important characteristics of the image. Therefore, Faster-CNN is faster than predecessor.

2) Use the RPN to generate high-quality proposals. About 300 proposals are generated on each image and map the proposals from the feature map;

3) The feature maps and proposals are sent into the RoI network. The RoI Pooling layer is used to transfer the coordinates of the feature map proposals back to the source picture. Accurate class judgment and position are output after the convolutional layers, classification, and regression. RoI Pooling uses closest interpolation to get integer pixel location coordinates on the source picture. The upward restoration will produce a position shift of more than ten pixels for small items in the image, resulting in significant inaccuracies. To overcome this problem, the new technique proposed in this study uses RoI Align [6] in mask R-CNN instead of RoI Pooling. RoI Align uses bilinear interpolation to get the floating pixel position coordinates on the original picture, which improves the identification of tiny landmarks.

The problem of landmark localization is the most difficult part of object detection. In the following description, the performance of the proposed scheme is evaluated in terms of accuracy, precision, sensitivity, specificity, and F-measure [7]. The sensitivity is the fraction of all positive examples divided that reflects the classifier's ability to detect positive cases, and the recall is the same as the sensitivity. The specificity indicates the fraction of all split negative instances and measures the classifier's ability to recognize negative situations. The precision reflects the fraction of examples that are truly positive when split into positive cases. The F-Measure is a comprehensive evaluation indicator, and its high value indicates that the classification model is more effective. The study uses multiple measurement indicators: precision, recall rate, mean precision and mean average accuracy. The assessment criterion assesses the pros and cons of the relevant attribute according to the value of the evaluation index.

In the work of region patches classification, there are only four possible outcomes of applying the classifier on any instance. These outcomes are intersection over union (IoU) follows:

$$precision = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \tag{1}$$

$$IoU = \frac{A_{\text{detectResult}} \cap A_{GroundTruth}}{A_{\text{detectResult}} \cup A_{GroundTruth}} \tag{2}$$

where T_P, T_N, F_P, and F_N are real landmarks predicted as true positions, false landmarks predicted as true positions, real landmarks predicted as false objects, and false landmarks predicted as false objects, respectively; $A_{\text{detectResult}}$ is the ground truth of landmark boxes; $A_{GroundTruth}$ is the test result of landmarks; N is the number of figures in the test set. In these measurement indexes, the higher the measure, the better the associated attribute. Recall and mAP formular are defined as follows:

$$recall = \frac{T_P}{T_P + T_N} \tag{3}$$

$$\text{mAP} = \sum_{k=1}^{Q} \text{Q}(k)\Delta \text{r(k)} \tag{4}$$

$Q(k)$ represents the value of precision when k figures can be recognized; and Δ r(k) represents the change of the recall value when the number of recognized figures changes from k-1 to k.

Training RPNs, the model evaluates a binary class label (of being an object or not) to each anchor. Anchors that

are neither positive nor negative do not contribute to the training objective. With these definitions, by minimizing an objective function following the multi-task loss in Fast R-CNN. Our loss function for an image is defined as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (6)$$

here, i is the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an object. The ground-truth label p_i^* is 1 if the anchor is positive, and is 0 if the anchor is negative. t_i is a vector representing the 4 parameterized coordinates of the predicted bounding box, and t_i^* is that of the ground-truth box associated with a positive anchor. The classification loss L_{cls} is log loss over two classes (true landmarks and false landmarks) follows as formula (7):

$$L_{cls}(p_i, p_i^*) = -\log[p_i^* p_i + (1-p_i^*)(1-p_i)] \quad (7)$$

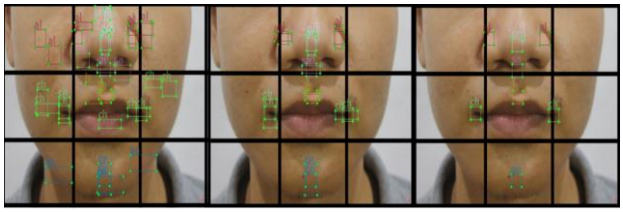
For the regression loss, calculated by the fomular (8):

$$L_{reg}(t_i, t_i^*) = S_{L1}(t_i - t_i^*) \quad (8)$$

where S_{L1} is the robust loss function (smooth L_1) defined in [5]. The term $p_i^* L_{reg}$ means the regression loss is activated only for positive anchors (p_i^* =1) and is disabled otherwise (p_i^* =0). The outputs of the *cls* and *reg* layers consist of { p_i } and { t_i } respectively.

IV. Experiments and Results

There have been two components of advancement on Improved Faster RCNN: Firstly, the input data, one of which is optimized by aligning the Frankfurt horizontal, the others might not by augmentation data with brightness, rotating, edge enhancement, and so on. Secondly, RoI Pooling is modified by RoI Align. The upgraded Faster R-CNN outperforms the original algorithm and methods with only one enhancement. Table II shows the experimental settings. Table III depicts the impact of changes.



(a) (b) (c)

Figure 3. Evaluate results shown (a) bounding box in which roi pooling and align picture; (b) bounding box in which roi align and align picture; (c) bounding box in which roi align and augmented picture is given the highest prediction probability.

TABLE II. Impoved Setting in Experiments

No	Condition	Experiments
1	Aligned picture,	Roi Pooling, Extraction network: VGG16
2	Augmented picture	Roi Align, Extraction network: VGG16
3	Aligned picture	Roi Align, Extraction network: VGG16
4	Augmented picture	Roi Pooling, Extraction network: VGG16

The results of experiment 1 and 3 were shown that the performance of VGG16 with aligned pictures is lower than with augmented pictures. Compared to aligned pictures, augmentation data enhanced much data for training three times of quantity. However, it is difficult for a model to understand all features without additional data that could lead missing area being blinded or covered. Therefore, the network becomes complicated and the detection becomes not untruth shown as in Fig. 3, the accuracy of the detection of scratches and sand inclusion defects becomes lower due to the increase of parameters. So Aligned picture is not suitable for landmark detection in this study. Augmentation data with VGG16 is chosen as the feature extraction network. The detection speed of trials does not change much because the same feature extraction network is employed. When comparing experiment 3 to experiment 1, it is clear that the ROI alignment improves the quality of proposals and the capacity to identify sand inclusion. Because the sizes of sand inclusion faults all fall into the same category, the aspect ratio distribution of the scratch bounding box is quite unequal, with some extremely large proportions. This has no discernible effect on the ability to identify scratches. Because RoI Align reduces the inaccuracy of mapping back to the source image and improves the detecting ability of tiny pots. The use of Roi Align, on the other hand, has no discernible influence on the identification of scratches. Because the bounding box of the scratch is so large, even if an error occurs after mapping back to the original picture, it has little effect on the detection of the scratch. Experiment 5 demonstrates that the upgraded Faster R-CNN has considerably improved the detection performance for scratches and sand inclusion flaws in two ways. The algorithm's mAP may reach 93.892%. When compared to direct detection speeds of several seconds, the detection speed may approach 110 secs/pic, making it more suited for industrial inspection. This paper's enhanced method is appropriate for detecting landmarks based on accurate facical landmark anthropology for measurement and regeneration.

TABLE III. Results of Evaluate Experiments on 200 Models

No	AP (scratch)	mAP	Detection speed (ms/pic)
1	81.573%	86.853%	132
2	83.283%	93.892%	150
3	86.748%	90.643%	110
4	84.396%	92.295%	178

From the table can conclude, the model has shown a good result with loss function reached at 1,435% with

multiple complex and hidden landmarks. But when the input image is affected by the exterior and light, brighten or shadow the model's accuracy decreases sharply to 89.489%. Comparing automatic calculation point data with data calculation on the Pytorch is shown in Table III and Fig. 4.

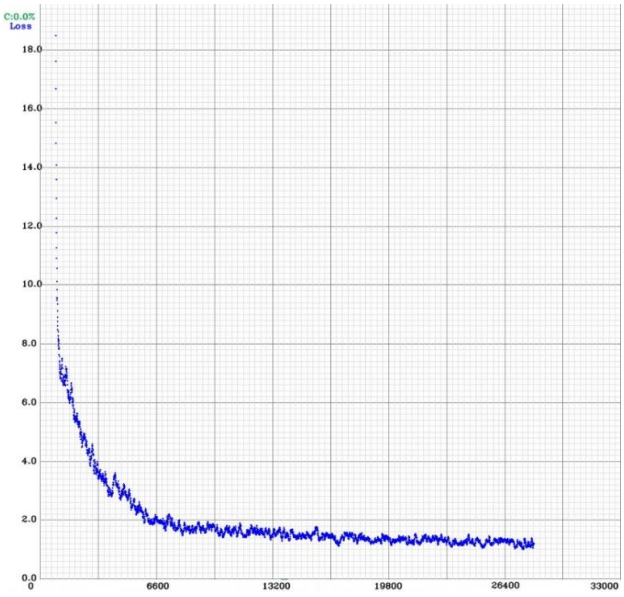


Figure 4. Curve of loss function during train and validation over 28100 iteration and 379 hours.

Figure 5. Nasal anthropology landmarks predicted in lateral view.

The facial landmarks detecting line in Fig. 4 illustrates the modification in the loss function between the train and validation datasets. The loss reduces promptly at the start of the training, suggesting that the learning rate is appropriate and the weight changes rapidly. After a few epochs of training, the loss function converges, suggesting that the weights begin to fine tune and achieve optimum. This demonstrates how the trained network eventually learns to more precisely characterize the picture and increase classification accuracy. The model detection result is given in Fig 5; Fig. 6 and Fig. 7 are evaluated visually with excellent accuracy.

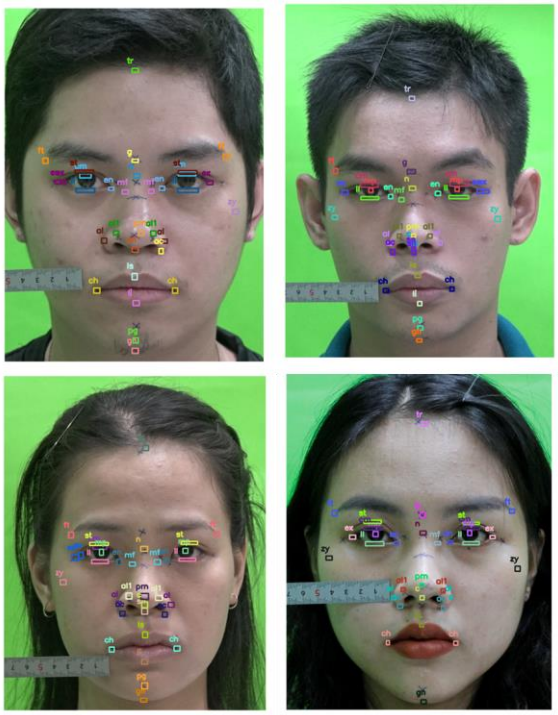


Figure 6. Landmarks were automatically detected by the profile view.

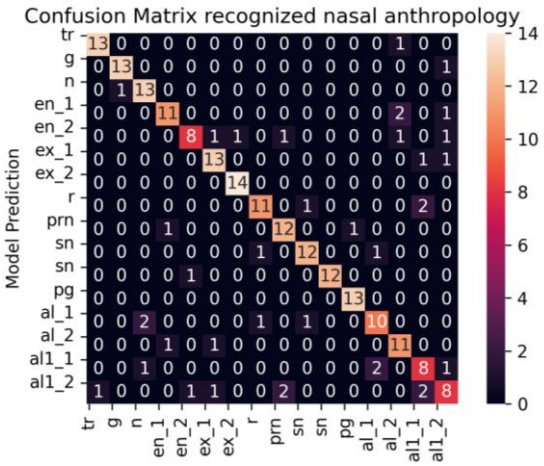


Figure 7. Evaluated the accuracy of data.

V. Conclusion

The method of identifying landmarks in anthropology was effectively handled and extremely accurate in this study. This method allows for the computation of 3D facial restructuring using specified coordinates. Since then, it is feasible to construct an interpolation method of nasal structure, using the data and the detected point coordinates in order to correctly adjust the face. Furthermore, this technique aids in the survey of environmental influences that impact each individual or application into the reformation of missing body part.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

H. N. A. Tuan performed the document preparation, data collection; N. D. X. Hai is a co-first author which analyzed the results and wrote the first draft of the manuscript; the manuscript was revised by N. T. Thinh. All authors contributed to conceptualization and design of the study structure and content; all authors had approved the final version.

ACKNOWLEDGMENT

The authors would like to specially thank for the support of Pham Ngoc Thach University of Medicine and Ho Chi Minh City University of Technology and Education in experiments and dataset

REFERENCES

[1] A. Jankowska, J. Janiszewska-Olszowska, and K. Grocholewicz, "Nasal morphology and its correlation to craniofacial morphology in lateral cephalometric analysis," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, p. 3064, 2021.

[2] J. D. White, *et al.*, "Sources of variation in the 3dMDface and Vectra H1 3D facial imaging systems," *Scientific Reports*, vol. 10, no. 1, pp. 1-10, 2020.

[3] Q. Cao, *et al.*, "Vggface2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018.

[4] L. G Farkas, *Anthropometry of the Head and Face*, Raven Press, 1994.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE International Conference on Computer Vision*, 2015.

[6] K. He, *et al.*, "Mask r-cnn," in *Proc. the IEEE International Conference on Computer Vision*, 2017.

[7] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," arXiv preprint arXiv:2010.16061, 2020.

Ho Nguyen Anh Tuan, M.D., Ms.C holds as deputy Head of the Academic Affairs. He received Bacherlor degree - Pham Ngoc Thach University of Medicine Hochiminh City, Vietnam in 2008. He pursued Master of science in Human Anatomy in University of Medicine and Pharmacy of Ho Chi Minh City in 2012. His interested in applying the EMP software for evaluation the theory courses in anatomy department, applying newest technology method in human anatomy and morphological study.

Nguyen Dao Xuan Hai was graduated B.S of Mechanical Engineering at Ho Chi Minh City University of Technology and Education, Vietnam. His fields are service robotic, medicine robot, machine learning applied to machine vision and manufacture. Currently, he is head project manager of a Mechatronics Laboratory in Ho Chi Minh City University of Technology and Education - VietNam. He also got many scientific research awards.

Nguyen Truong Thinh is an Associate Professor in Mechatronics. He obtained his PhD in 2010 - Mechanical Engineering from Chonnam National University. His work focuses on Robotics and Mechatronic system. Projects include: Service robots, Industrial Robots, Mechatronic system, AI applying to robot and machines, Agriculture smart machines.