3-D Human Pose Estimation by Convolutional Neural Network in the Video Traditional Martial Arts Presentation

Tuong-Thanh Nguyen[1], Van-Hung Le[2], Thanh-Cong Pham[1]*
[1] *Hanoi University of Science and Technology, No. 1, Dai Co Viet, Hai Ba Trung, Hanoi, Viet Nam*
[2] *Tan Trao University, Km6, Trung Mon, Yen Son, Tuyen Quang, Viet Nam*
Received: May 11, 2019; Accepted: November 28, 2019

Abstract

Preservation and maintenance of traditional martial arts and teaching martial arts are very important activities in social life. It helps preserving national culture, train health, and self-defense for people. However, traditional martial arts have many different postures and activities of the body and body parts. In this paper, we are proposed using deep learning with Convolutional Neural Network (CNN) for estimating key points and joints of actions in traditional martial art postures and proposed the evaluation methods. The training set has been learned on the 2016 MSCOCO key points challenge classic database [21], the results are evaluated on 14 videos of traditional martial art performances with complicated postures. The estimated results are high and published. In particular, we presente the results of estimating key points and joints in 3-D space to support the construction of a traditional martial arts conservation and teaching application.

Keywords: Estimation of key points, deep learning, skeleton, dancing and teaching of traditional martial arts

1. Introduction

Estimation and prediction of the actions of the human body is a widely-studied issue in the community of robotics and computer vision. These studies are applied in many applications of human daily life such as detecting the patients falling in hospitals [1], or system for detection of falling cases for the elderly [2], [3]. These systems can use information from color images, depth images [1], or skeleton images [4] obtained from sensor types. Among them, Microsoft (MS) Kinect sensor version 1 (v1) is a common and cheap sensor that can collect information from the environment such as color images, depth images, skeleton [19]. However, there are many challenges in detecting actions such as falling [4], [20]. Currently, together with the strong development of deep learning in detection, recognition and prediction of actions are good approaches. Therefore, in this paper, we presented an experiment that uses deep learning to estimate and predict the skeleton of human on video data of martial arts presentation performed by martial arts instructors, students and evaluation methods for key points estimation. This approach is based on learning and estimating key points on the human skeleton model. In particular, this approach can estimate the human pose based on skeletons in the case of being hidden.

* Corresponding author: Tel: +(84) 914.092.020
Email: thanh1277@gmail.com

Currently, there are many studies on the detection, recognition and prediction of human actions. These studies have been applied in many practical applications for humans such as Rantz et al. [1] have proposed a system of automatic detection of falling events in hospital rooms. The system uses wireless accelerometers mounted on the patient's body which compared to the acceleration of data collected from a wall-mounted MS Kinect sensor. At the same time, the system also calculated the distance between the human and the bed to detect the patient's falling event. Especially in Vietnam [5], [6] as well as many countries in the world, like China [7] there are many martial arts postures or martial arts to be preserved and passed down to posterity. Preservation and maintenance in the era of technological development can be performed by the preservation of the martial arts instructor's actions in the form of joints.

Data obtained from MS Kinect sensor v1 usually contains a lot of noise and lost when obscured. Especially skeleton data of a human. Therefore, it is important to estimate the skeleton in which bone points are key points on the human body. Umer et al. [25] used Regression Forests to estimate the human direction with the depth image obtained from MS Kinect version 2. The training is performed on the human parts under ground truth, with 1000 samples of image point on depth images. However, the accuracy of the highest average result is only 35.77%.

Currently, with the strong development of deep learning, the estimation of key points on human bodies is widely implemented. Daniil et al. [26] introduced a new CNN for learning the features on the key point dataset such as the location of key points, the relationship between pairs of points on the human body. This new network is based on the OpenPose toolkit [15] and can be applied for learning on the CPU. In particular, convolutional neural networks are learned and evaluated on the 2016 COCO multi-population database [21]. This is a huge database under ground truth with over 150 thousand people, with 1.7 million ground truth for key points.

Kyle et al. [23] used CNN to learn from the data of the key points of the human body that was under ground truth and extracted from the connected data when projecting two cameras into people. And the results are then projected into 3-D space and used the minimum squared distance algorithm to evaluate the estimated results. Cao et al. [18] used the CNN to learn the position of key points on the human body and allowed the geometric transformations of the lines connecting the key points in connective relations on the human body. This article is evaluated on two classic databases, MPII [27] and COCO [21]. In particular, the database of COCO key points [8], [9] has been developed for many years. These databases are collected from many people and there are also many challenges for estimation of human activities.

2. Usage of deep learning for estimating human actions in traditional martial arts

2.1. Estimation on the map of key points and corresponding body parts

The action of the human body is detected, recognized and predicted, estimated based on the parts of the human body (body part). The parts are constituted based on the connection between the key points. Among them, each part is represented by a vector $\boldsymbol{Lc}$ in space 2-D (image space) in a set of vectors on the human body $\boldsymbol{S}$, and in the set of vectors $\boldsymbol{L= \{L_1, L_2, ..., L_C\}}$, there is $\boldsymbol{C}$ vector on human body $\boldsymbol{S}$. Among them, the human body **S** is represented by $\boldsymbol{J}$ key points), $\boldsymbol{S =\{S_1, S_2, ..., S_j\}}$. With an input image in the size $w \times h$, the position of key points may be $S_J \epsilon \boldsymbol{R}^{w \times h}$, $j \epsilon \{1,2,...,J\}$ as shown in Fig.3. Then is the matching between the corresponding parts on the body of different persons calculated according to the affine. In this paper, we are completely used the convolutional neural networks designed and calculated in [18] to perform the estimation of vectors in $\boldsymbol{L}$.

As shown in Fig.4, the CNN by Zhe et al. [18]. This CNN consists of two branches performing two different jobs. From input data, a set of feature maps $\boldsymbol{F}$ is created from analyzing the image then these confidence maps and affinity fields are detected at the first stage. The key points on the training data are displayed on confidence maps as shown. These points are trained to estimate key points on color images. The first branch (top branch) is used to estimate key points, the second branch (bottom branch) is used to predict the affinity fields matching joints on many people. In particular, the output of the previous stage is the input for the later stage and the number of stages in the architecture (as Fig.5) is usually equal to 3. This means that the results of the heatmaps prediction at this stage will be the input for training and predicting the heatmaps at the next stage. As shown in the Fig.6, the result of predicting the heat map is gradually converging. In which each heatmap is a candidate of a bone point in the skeleton of the human. These points are trained to estimate the key points on color images. The first branch (top branch) is used to estimate the key points, the second branch (bottom branch) is used to predict the affinity fields matching joints on many people.

2.2. Dataset of traditional martial arts

Traditional martial arts is a very important sport that helps people train health exercise and protect themselves. In many countries around the world, especially in Asia, there are many traditional martial arts handed down from generation to generation. With the development of technology, it is important to maintain, preserve and teach such martial arts [10], [11]. There are also many different types of image sensors that can collect information about martial arts teaching and learning of the schools of martial art. The MS Kinect sensor v1 is the cheapest sensor today. This type of sensor can collect a lot of information such as color images, depth images, skeleton, acceleration vector, sound, etc. From the collected data, it is possible to recreate the environment in 3-D space about teaching martial arts in the schools of martial art. However, in this paper, based on the information collected from the MS Kinect sensor v1, we are only used color, depth images for the construction of this study.

To obtain data from the sensor environment, the Microsoft Kinect SDK 1.8 is used to connect computers and sensors [12]. To perform data collection on computers, we are used a data collection program developed at MICA Institute [14] with the support of the OpenCV 3.4 libraries [13], C++ programming language. Between the sensors of color images, depth images, and the skeleton, there is a distance as shown in Fig.1. Therefore, it is recommended to make a calibration to take the data on color images and depth images, particularly, we

are applied the data calibration of Zhou et al. [22] and Jean et al. [24]. In these two calibration tools, the calibration matrix is used as in formula (1):

$$H_m = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_x & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

In which, (c_x, c_y) is the center of the image, (f_x, f_y) is the focus of the lens (distance from the sensor surface to the optical center of the lens system).

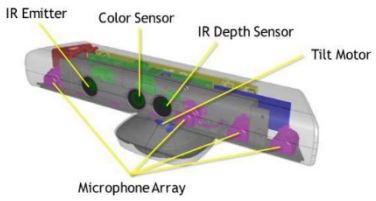

Fig. 1. MS Kinect sensor v1

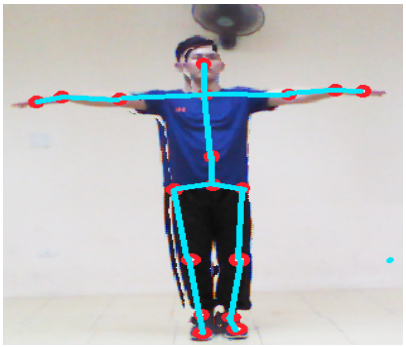

Fig. 2. Illustrations on ground truth for key points on image data of the human. Red points are key points on the human body. Blue segments show the connection between the parts of the human body.

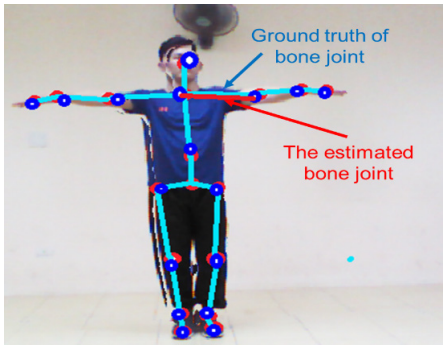

Fig. 3. Illustration of the estimated results of the key points. The blue points are estimated. Red joints are estimated.

MS Kinect sensor v1 can collect data at a rate of about 10 frames/s on a low-configuration Laptop. The obtained image resolution is 640×480 pixels. The obtained dataset consists of 14 videos of different postures, with the number of frames listed in Tab.1 and illustrated in Fig.3.

Table 1. Number of frames in martial arts postures.

Video	1	2	3	4	5	6	7
Number of frame	120	74	100	87	80	88	87
Video	8	9	10	11	12	13	14
Number of frame	74	71	90	100	97	65	68

We are prepared manual ground truths for key points with hands as illustrated in Fig.2 and Fig.3. This dataset only includes a human in each image. In this paper, we use a trained model on the 2016 MSCOCO key points challenge database [21]. The trained model based on the published Openpose [16]. To perform the training process, it is necessary to use the sets "caffe_train" and "VGG-19 model" boards; Details are shown in the papers [17], [18]. Among them, the model trained for estimation of key points is trained on annotation with 25 key points on the human body. Training toolkit is written in Python language and runs on the server's GPU. Testing tools can be implemented on Windows or Ubuntu operating systems with programming languages [16] such as C++, MatLab, Python.

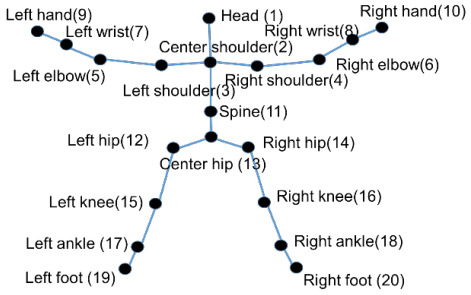

Fig. 4. Key points on the human body and the labels.

2.3. Evaluation Method

In order to perform and evaluate the results, a map of representative points and corresponding vectors of parts of the human body is estimated. We are changed the size of the input image from 640×480 pixels to 654× 368 pixels, to match the memory on the GPU. The testing process is performed on workstation computer with Intel (R) Xeon (R) CPU E5-2420 v2 @ 2.20 GHz 16GB RAM, GPU GTX 1080 TI-12GB Memory. The running process consists of two main parts: the first is the running time of the CNN, the second is the running time predicted on many persons. These two parts are evaluated in terms of complexity, respectively ***O(1)*** and ***O(n^2)***, where ***n*** is the number of persons in the image.

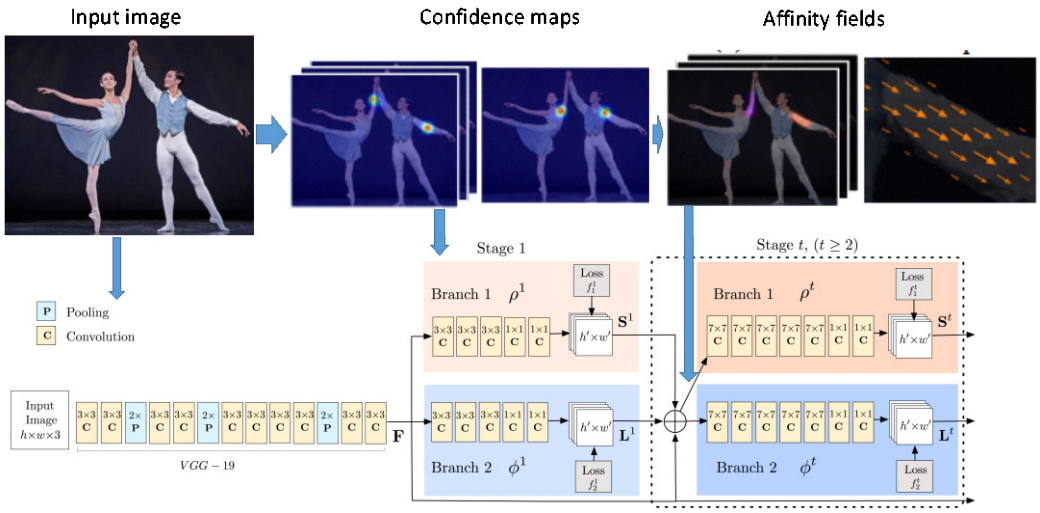

Fig. 5. The architecture of the two-branch multi-stage CNN for training the model estimation [18].

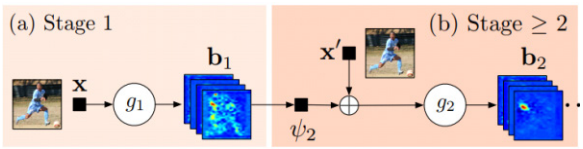

Fig. 6. Illustration of the training and prediction on the heatmaps. *x, x'* are the training blocks; g_1, g_2 are the predicting blocks.

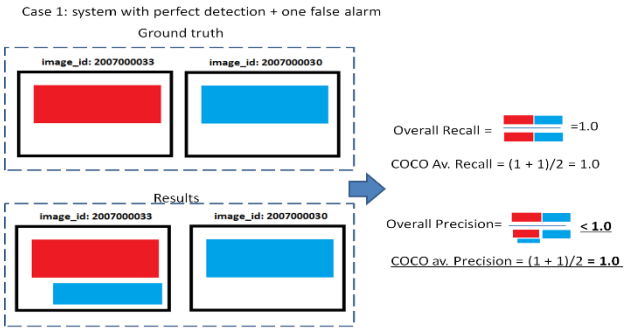

Fig. 7. Illustration on a matrix of assessment of the similarity of the key points [17].

Fig. 8. Illustration on the chain of estimation results of the key points and joints on videos of actions in traditional martial arts videos

As in [18], we evaluate the similarity of object key points similarity **(*OKS*)** and use average precision **(AP)** with threshold ***OKS*** = 0.5. This is calculated from the change in the size of the human body compared to the distance between the estimated key points and the points under ground truth.

The calculation of the ***OKS*** rate is performed on each joint on the estimated key points and calculated according to the formula in [17], as illustrated in Fig.7. In which, Fig.7 is detailed as in the equation (2).

$$OKS = \frac{abs(\|G_{ground} - R_{result}\|)}{G_{ground}} \quad (2)$$

where G_{ground} is the length of the ground truth vector, R_{result} is the length of the jointed vector that is estimated according to the predefined index. If ***OKS***> 0.5, is a difference greater than 50% of length, that is a false estimation, otherwise a true estimation.

At the same time, we also assessed the angle of deflection between the joint under ground truth (***VG***) and the estimated joint (***VE***) from the estimated key points (***AD*** (%)). The angle between the two vectors ***(A= argcos(V_G, V_E))***. If (***A<=10^0***) that is a true estimation, otherwise, it is a false estimation. The (***AD***) ratio is calculated by the correct estimation divided by the total number of joints. We evaluated the deviation of the location of key points (D_p); It is the average distance from the ground truth key point to the estimated key point. We computed only the estimated key points. The distance is computed according to formula (3) and the unit of the pixel.

$$D(p_g, p_e) = \sqrt{(x_g - x_e)^2 + (y_g - y_e)^2} \quad (3)$$

where D is the distance between two points (p_g, p_e), p_e is the estimated key point whose coordinates are (x_e, y_e), p_g is the ground truth key points whose coordinates are (x_g, y_g).

The input data of the system includes color photos, videos. The output data is the result of the estimation of the key points on the image while the joints between the key points are also shown. The data on ground truth and the location of the estimated key points are also saved in the files according to the predefined structure.

2.4. Results of estimation

The results of the joint estimation are evaluated and shown in Tab.2. The average result is 95.6%. This result is high because, on the test dataset, each image has only a human in the image. In the dataset [21] and [27], there are many humans in the image. In video #4, the result is 89.6%. This is the lowest result in the videos. In this video, the images contain a lot of noise and element broken and deflected in the process of calibration of color images and depth images. Especially, Fig.8 illustrates visually the results of estimating joints on the traditional martial dataset.

Table 2. The results of the estimation of the joints on the database collected about the postures of traditional martial arts.

Video	**1**	**2**	**3**	**4**	**5**
AP (%)	95.4	93.7	96.2	89.6	96.1
Video	**6**	**7**	**8**	**9**	**10**
AP (%)	92.8	97.4	98.8	96.9	94.5
Video	**11**	**12**	**13**	14	
AP (%)	96.9	96.2	95.7	98.2	

The estimated result is 25 key points on the human body [21]. However, in the data of key points ground truth, we made ground truth of only 20 key points, therefore, the assessment is only performed over 20 key points. It can be seen that the results estimation are highly accurate, although the training model is available on MSCOCO key points challenge data [21] and our test data contains a lot of noise. At the same time, we also show the predicted probability (**IOU**) on each key point, as shown in Fig.9. The *x-axis* is the number of estimated key points on videos. The *y-axis* is the probability distribution estimating the key points estimate with the trained model [18].

In Fig. 9, we showed the probability graph (**IOU**) that estimates key points in 3 videos. We can see that the probability concentrates at about 0.7 to 0.9. This means that the trained model in [15] has good predictability. Table 3 shows the accurate estimation results based on the deflection angle of the joints (***AD***). The estimation result has an average accuracy of 95.3%. Details of the estimated results are saved in this address: https://www.fshare.vn/file/Q3YA7XRP31KH?token=1556244489

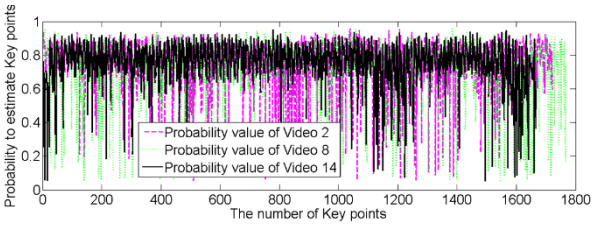

Fig. 9. The graph shows the probability distribution estimating the key points in 3 videos of the martial arts database.

The average results of the deviation of the estimated key points with the ground truth points (D_p) are shown in the Tab.4. The average deviation of the key points is estimated to be 14.73 pixels.

Table 3. Accurate estimation results are based on the angular deviation between joints under ground truth and the estimated joints on each video.

Video	1	2	3	4	5
AP (%)	93.7	94.6	92.8	90.9	95.3
Video	6	7	8	9	10
AP (%)	94.6	95.8	97.6	97.8	95.1
Video	11	12	13	14	
AP (%)	97.0	95.8	96.3	96.9	

Table 4. The average distance of the representative points is estimated with the original representative points.

Video	1	2	3	4	5
D_p (pixel)	21.2	18.6	9.7	25.9	13.8
Video	6	7	8	9	10
D_p (pixel)	15.7	9.4	15.4	12.4	10.1
Video	11	12	13	14	
D_p (pixel)	14.0	12.8	11.3	16.9	

In addition, we also render a 3-D environment of each video's scene. In particular, each frame includes results on a color image taken respectively to the depth image. And based on the intrinsic parameter of the Kinect sensor v1 and the PCL library [28], OpenCV[13], the point cloud data of scene and the results are projected into 3-D space. The real coordination (x_p, y_p, z_p) and color value of each pixel when projecting them from 2-D space to 3-D space (3-D data) are calculated as the equation (4). Illustration of a scene is shown in Fig.10.

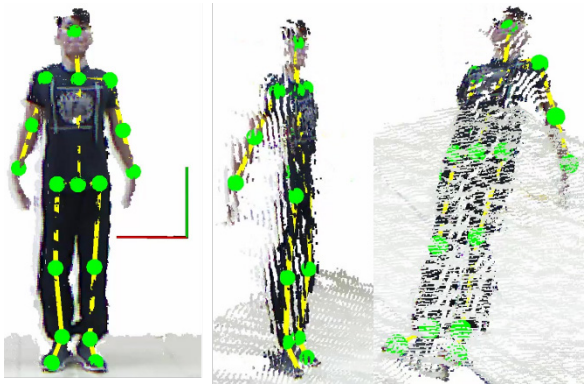

Fig. 10. Illustration of the estimated results of key points and joints in 3-D space of a frame.

$$
\begin{aligned}
x_p &= \frac{(x_a - c_x) * depthvalue(x_a, y_a)}{f_x} \\
y_p &= \frac{(y_a - c_y) * depthvalue(x_a, y_a)}{f_y} \\
z_p &= depthvalue(x_a, y_a) \\
c(r,g,b) &= colorvalue(x_a, y_a)
\end{aligned}
\tag{4}
$$

where *depthvalue* (x_a, y_a) is the depth value of a pixel (x_a, y_a) on the depth image, *colorvalue(r, g, b)* is the color value of a pixel (x_a, y_a) on the color image.

3. Conclusion and discussion

The preservation, storage and teaching of traditional martial arts are very important in preserving national cultural identities and training health and self-defense of people. However, the actions of the body (body, arms, legs) of a martial arts instructor are not always clear. There are many hidden joints. In this paper, we have proposed using CNN for estimating key points to predict the actions of martial arts instructor and traditional martial arts videos. At the same time, we have presented methods for evaluating the estimated key points and joints. Especially, we have presented the results in 3-D space. The points represent the amount, from which the joints can be drawn about those actions. Therefore, training martial arts by video becomes easier and more explicit.

However, there are some cases where the joints are obscured in videos that the model has not yet estimated. In the future, we will conduct studies to estimate obstructed joints. When there are sufficient joints, it is possible to build a visual martial arts teaching model and evaluate the performance of traditional martial arts representation.

Reference

[1]. Rantz, M., Banerjee, T., Cattoor, E., Scott, S., Skubic, M., & Popescu, M. Automated fall detection with quality improvement "rewind" to reduce falls in hospital rooms. J Gerontol Nurs, 40(1), 13-17, 2014.

[2]. Miguel, K. d., Brunete, A., Hernando, M., & Gambao, E. Home CameraBased Fall Detection System for the Elderly. Journal of Sensors, 17(12), (2017).

[3]. Ahmed, M., Mehmood, N., Adnan, N., Mehmood, A., & Rizwan, K. Fall Detection System for the Elderly Based on the Classiffication of Shimmer Sensor Prototype Data. Healthc Inform Res, 23(3),147-158, 2017.

[4]. IgualCarlos, R., Carlos, M., & Plaza, I. Challenges, Issues and Trends in Fall Detection Systems. BioMedical Engineering OnLine, 12(1), 147-158, 2013.

[5]. Dinh, T. B. Bao ton va phat huy vo co truyen Binh dinh: Tiep tuc ho tro cac vo duong tieu bieu.

http://www.baobinhdinh.com.vn/viewer.aspx?macm=12&macmp=12&mabb=88043.[Accessed; April, 4 2019], 2017.

[6]. Dinh, T. B. Ai ve Binh Dinh ma coi, Con gai Binh Dinh bo roi di quyen. http://www.seagullhotel.com.vn/du-lich-binh-dinh/vo-co-truyen-binh-dinh-5. [Accessed; April, 4 2019], 2019.

[7]. Chinese Kung Fu (Martial Arts). https://www. travelchinaguide.com/intro/martial_arts/. [Accessed; April, 4 2019], 2019.

[8]. ECCV2018. ECCV 2018 Joint COCO and Mapillary Recognition). http: //cocodataset.org/#home. [Accessed 18 April 2019], 2018.

[9]. 2017, M.. MSCOCO Keypoints Challenge 2017). https:// places-coco2017.github.io/. [Accessed 18 April 2019], 2017.

[10]. Dinh, T. B. (2011). Preserving traditional martial arts). http://www.baobinhdinh.com.vn/ culture-sport/2011/8/114489/.[Accessed 18 April 2019].

[11]. Chinese (2012). Traditional Chinese martial arts and the transmission of intangible cultural heritage).

https://www.academia.edu/18641528/Fighting_modenity_traditional_Chinese_martial_arts_and_the_transmission_of_ intangible_cultural_heritage.[Accessed 18 April 2019].

[12]. Microsoft. Kinect for Windows SDK v1.8. https://www.microsoft.com/en us/download/details.aspx?id= 40278. [Accessed 18 April 2019], 2012.

[13]. Opencv library. https://opencv.org/. [Accessed 19 April 2019], 2018.

[14]. MICA. International Research Institute MICA. http://mica.edu.vn/. [Accessed 19 April 2019], 2019.

[15]. Openpose. https://github.com/CMU-Perceptual-Computing-Lab/ openpose. [Accessed 23 April 2019], 2019.

[16]. Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. Realtime Multi Person Pose Estimation. https: //github.com/ZheC/Realtime_Multi-Person_Pose_Estimation.[Accessed 23 April 2019].

[17]. COCO. Observations on the calculations of COCO metrics.https://github.com/cocodataset/ cocoapi/issues/56. [Accessed 24 April 2019].

[18]. Cao, Z., Simon, T., Wei, S.-E., & Sheikh,Y. Realtime Multi-Person 2D PoseEstimation using Part A-nity Field, CVPR, 2017.

[19]. Kramer, J., Parker, M., Castro, D., Burrus, N., & Echtler, F. Hacking the Kinect. Apress. 2012.

[20]. Tao, X., & Yun, Z. Fall prediction based on biomechanics equilibrium using Kinect. International Journal of Distributed Sensor Networks, 13(4), 2017.

[21]. X, Z. A Study of Microsoft Kinect Calibration. Technical report Dept. of Computer Science George Mason University. 2012.

[22]. Brown, K. Stereo Human Keypoint Estimation. Stanford University,2017.

[23]. B., J.-Y. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/ bouguetj/calib_doc/. [Accessed 19 April 2019], 2019.

[24]. Ra, U., Gall, J., & Leibe, B. (2015). A semantic occlusion model for human pose estimation from a single depth image. In: CVPR Workshops (CVPRW).

[25]. Osokin, D.. Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose. Published in ArXiv, 2018.

[26]. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., & Schiele, B. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. CVPR 2016), 2016.

[27]. Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. Convolutional pose machines.

[28]. PCL, Point Cloud Library, http://pointclouds.org/. [Accessed 19 April 2019]