https://doi.org/10.2352/ISSN.2470-1173.2020.17.3DMP-004

Action recognition using pose estimation with an artificial 3D coordinates and CNN

Jisu Kim, Deokwoo Lee; Department of Computer Engineering, Keimyung University, Daegu, Republic of Korea

Abstract

Activity recognition and pose estimation are ingeneral closely related in practical applications, even though they are considered to be independent tasks. In this paper, we propose an artificial 3D coordinates and CNN that is for combining activity recognition and pose estimation with 2D and 3D static/dynamic images(dynamic images are composed of a set of video frames). In other words, We show that the proposed algorithm can be used to solve two problems, activity recognition and pose estimation. End-to-end optimization process has shown that the proposed approach is superior to the one which exploits the activity recognition and pose estimation seperately. The performance is evaluated by calculating recognition rate. The proposed approach enable us to perform learning procedures using different datasets.

Introduction

Activity recognition and pose estimation have received an important attention in the last years, not only because of their many applications, such as video surveillance and human-computer interfaces, but also because they are still challenging tasks. Pose estimation and action recognition are usually treated as independent problems [1]. Despite the fact that pose is related to action recognition, it is not commonly used to jointly solve the two problems in the benefit of action recognition. Therefore, we propose a CNN algorithm that can simultaneously process 2D and 3D human pose estimation and action recognition, as presented in Fig. 1. The overall architecture is similar to the work presented in [10]. Our work chiefly focuses on generating artificial third coordinates as detailed in Fig. 5.

One of the key advantages of deep learning is its ability to perform end-to-end optimization. As Kokkinos *et al.* [2] suggest, this is more evident for the CNN problem where the related tasks can benefit from each other. Recently, Convolutional Neural Network-based methods have achieved good results in 2D and 3D pose estimation due to the emergence of new architectures and the availability of large amounts of data [3, 4]. Similarly, activity recognition has recently been improved by using CNN from human pose [5]. Since most pose estimation performs heat map prediction, it is considered that pose estimation and activity recognition can not be combined with each other. This detection approach requires a non-differentiable argmax funtion to recover joint coordinates as a post processing stage that breaks the back-propagation chain required for end-to-end learning. We propose a method to solve the problem by supplementing the differentiated Soft-argmax [6, 7] for 2D and 3D pose estimation. It also improves performance by learning not only a simple 2D RGB image but also video of each frame together. This allows the end-to-end multitask train to be enabled by adding the pose estimation data to the activity recognition. Our algorithm has these advantages. First, the proposed pose estimation obtains the most accurate result of the regression method for 2D pose estimation and a good accurate result for 3D pose estimation. Second, the proposed pose estimation method is based on the still image, so we get the advantage of the image in the daily scene in 2D and 3D prediction. This is a very efficient way of learning visual features and is also very important in activity recognition. Third, the proposed activity recognition approach is based on RGB image and video extracting pose and visual data. Nonetheless, we have achieved excellent results in both 2D and 3D scenarios, compared to using ground-truth pose. Fourth, the pose estimation method can train different types of data sets at the same time so that 3D prediction can be generalized in 2D data. The structure of this paper is as follows. Section Activity recognition present algorithms for regression methods based on pose estimation and activity recognition, respectively. Section Experiments shows the experiment and section Conclusions concludes this paper.

Activity recognition

This section details the approach to the activity estimation that is one of the main contributions in the present work. The present work extends the aforementioned work in that 2D and 3D data are fully exploited using Soft-argmax function. Soft-argmax function is extended 2D and 3D pose regression in a unified way. One of the most important advantages of the proposed Method is the ability to integrate high level pose information with low level visual features in the CNN algorithm. This algorithm advantages allow sharing the same network architecture to pose estimation And visual feature extraction. In addition, visual features are learned using activity sequences and still images captured in real scenes with a very efficient way of learning robust visual representations.

As shown in Fig. 1, the proposed approach to activity recognition divided into two parts. The one, based on a body joint coordinates, is called pose-based recognition, and the other, on a sequence of visual features, is called appearance-based recognition. The results, generated from both parts, are combined to estimate the final activity label. In this section, we describe a detailed description of each recognition branch and detail the method for extraction of temporal information from a sequence of frames by extending single frame based pose estimation.

Pose estimation

The human pose regression problem is defined by the input RGB image ($\mathbf{I} \in R^{W \times H \times 3}$) the output estimated pose ($\hat{p} \in R^{N_J \times D}$) with N_J body joints of dimension D, and a regression function f_r, as given by the following equation [10] :

$$\hat{p} = f_r(I, \theta_r), \tag{1}$$

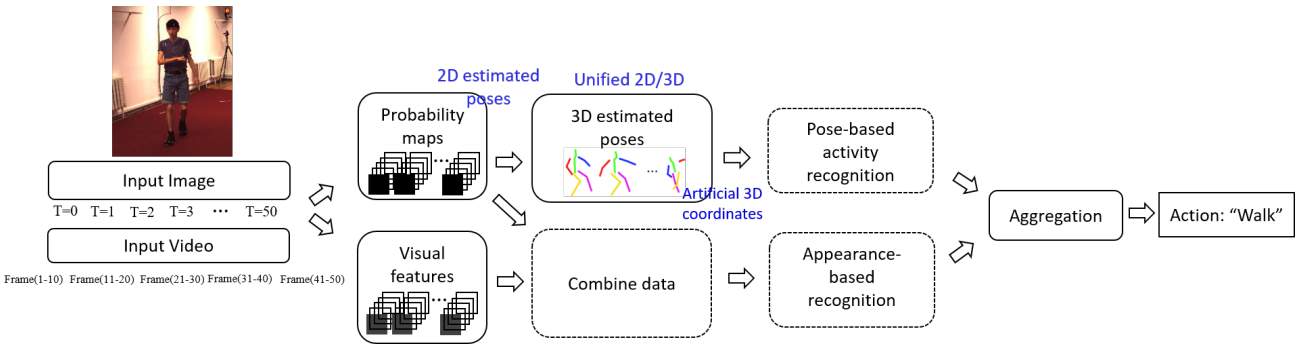


Figure 1. *The proposed multitask algorithm for pose estimation and action recognition. Pose and visual data are used to predict action.*

where r is a set of trainable parameters of function f_r. The objective is to optimize the parameters r in order to minimize the error between the estimated pose $\hat{p}$ and the ground truth pose p,

$$\hat{p} =_r \|p - \hat{p}\|. \tag{2}$$

In order to solve the function in Eq. (1), we use a deep CNN. As the pose estimation is the first part of our multitask approach, the function f_r has to be differentiable in order to allow end-to-end optimization. This is possible thanks to the Soft-argmax, which is a differentiable alternative to the argmax function and can be used to convert heat maps $M \subset R^2$ to joint coordinates (x,y) [6]. The network architecture of the neural network is shown in Fig. 2. As shown in Fig. 2, network architecture for feature extraction using Inception-V4 [9]. Similar to [6], to refine the result of the pose estimation, K prediction blocks, each of which is resulting from eight residual depth-wise convolutions and the K prediction block is represented as a set of probability maps, $p_1, p_2, ..., p_K$. Here the last prediction block p_K is used as the result of the pose estimation $\hat{p}$. In addition, we can access the intermediate joint probability maps that are indirectly learned thanks to the low-level visual features and Soft-argmax layer. In the proposed method for action recognition, visual features and joint probability maps are used to create appearance features as detailed in section Appearance based activity recognition.

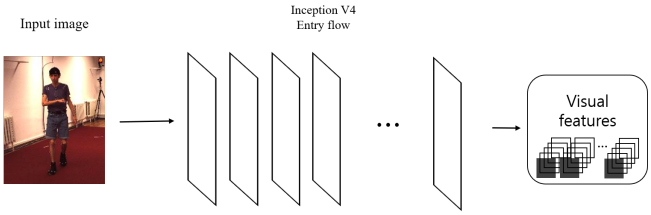


Figure 2. *Network architecture for feature extraction using Inception-V4.*

A graphical description of the Soft-argmax layer is shown in Fig. 4. Given an input data, to the layer, one of the main approach is to consider that the argument of maximum can be approximated by the expectation data of the input signal after being normalized to have particular distribution. In fact, for a leptokurtic distribution, expectation should be calculated by maximum a posteriori (MAP) estimation. The normalized exponential function (Softmax) is used because it alleviates undersirable influences of the maximum value and increases the pointiness of the resulting distribution. Using 2D heat map as an input, the normalized can be

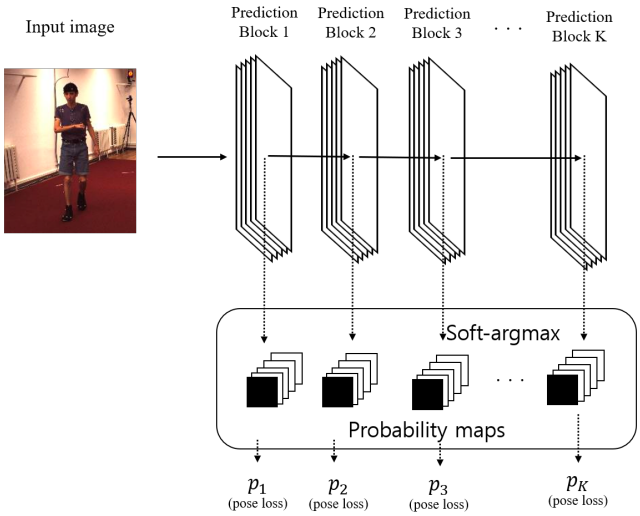


Figure 3. *The input image is through CNN consisting of one entry flow and k prediction blocks. The prediction is refined in each prediction block.*

interpreted as a probability map of the joint at position (x,y), and the expected value of the joint position is expressed as an expectation for the normalized, written by [10].

$$\left(\sum_{c=0}^{W_x} \sum_{l=0}^{H_x} \frac{c}{W_x} \Phi(\mathbf{X})_{(l,c)}, \sum_{c=0}^{W_x} \sum_{l=0}^{H_x} \frac{l}{H_x} \Phi(\mathbf{X})_{(l,c)} \right)^T, \tag{3}$$

where $\mathbf{X}$ is the input heat map with a size of $W_x \times H_x$, Φ is the Softmax normalization function, and $\Phi(\mathbf{X})_{(l,c)}$ is the value of Softmax n-function at position (l,c). The probability that certain joints appear in the image is computed as the sigmoid fuction that generated the maximum output the with the corresponding input heat map. Considering the pose p ($\sum_{i=1}^{K} p_i = 1$) with N_J joints, the joint vector is written by $\mathbf{v} \in R^{N_J \times 1}$. It should be noted that the visibility information included in $\mathbf{v}$ and the joint probability map $p = [p_1, p_2, ..., p_K]$.

Artificial 3D Coordinates

We extend 2D heat map to 3D volumetric representation, leading to extension of 2D pose regression to 3D scenario. As explained in section Pose Estimation, the depth of each block is defined as the depth resolution, that is a stack of the heat map, will be used for extension of 2D to 3D. Calculation of (x,y) coor-

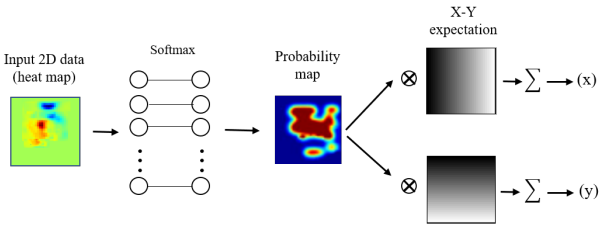


Figure 4. *Graphical representation of the Soft-argmax operation for 2D input data(heat map). The outputs are the coordinates x and y close to the maximum of the input data. (Σ : summation of all of the pixel values)*

dinates in the heat map is performed by applying a Soft-argmax operation to average heat map. Each probability map is composed of a set of heat maps, in this work, a number of heat map is eight. The i^{th} probability map p_i can be represented as $(p_i = [M_1^i, M_2^i, ..., M_8^i], 1 \leq i \leq 8$, in this case $K = 8)$. The average heat map is composed of an average values of the pixels each of which is included in the heat map. z coordinate is acquired from calculating cross variance of $\mathbf{x}$ and $\mathbf{y}$ ($\mathbf{x}$ and $\mathbf{y}$ are composed of x-coordinates and y-coordinates, respectively) followed by regression using Soft-argmax. This procedure is depicted in Fig. 5.

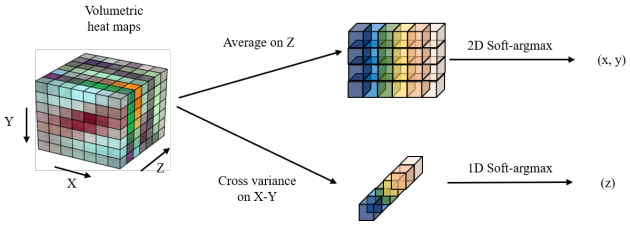


Figure 5. *Unified pose estimation by using volumetric heat maps*

The advantage of dividing the pose prediction into two geometric coordinate parts, (x, y) and z, is to keep the 2D heat map byproduct, which is useful for extracting appearance features as described in section Appearance based activity recognition. We can learn mixture of 2D and 3D data using the proposed unified approach. The gradient of 2D image is used for backprogation. As a result, the network can train both the high-precise 3D data in the motion capture system and the real-scene image collected in normal environment.

Pose based activity recognition

To explore the high level information encoded in the body joint position, we convert each sequence (composed of T poses) with N_J joint into an image-like representation. We let the temporal dimension encode the vertical axis, the coodinates of each point in 2D as a channel. Using this pproach, a pattern can be extracted from the body joint as a temporal sequence using a classical 2D convolution. Because the pose estimation method is based on the still image, time distributed abstraction is used to process the video clip. This is a simple technique to handle both single image and video sequence. We propose a fully convolutional neural network that extracts features from the input pose and generates an action heat map, as shown in Fig 6. The idea of this paper is that it is a very difficult learning problem because fully-connected layers will make unrelated joints zero for operations that depend only on few body joints, such as shaking hand. Conversely, 2D convolution is easier to learn because it implements a sparse structure without manually selecting joints. Also, other joints have very different coordinates and filter matching (eg, hand pattern) will not respond equally to the feet pattern. This pattern is then combined at a subsequent layer to produce discriminative activation until an action map of depth equal to the number of actions is obtained.

To generate an output probability for each action on a video clip, one need to do a pooling operation on the action map. Max pooling and Softmax activation are used to react more sensitively to the strongest response for each action. Inspired by the human pose regression method, the prediction is modified using a stacked architecture with an intermediate supervision function in the K prediction block. The action heat map at each prediction block re-enters Into the next action recognition block.

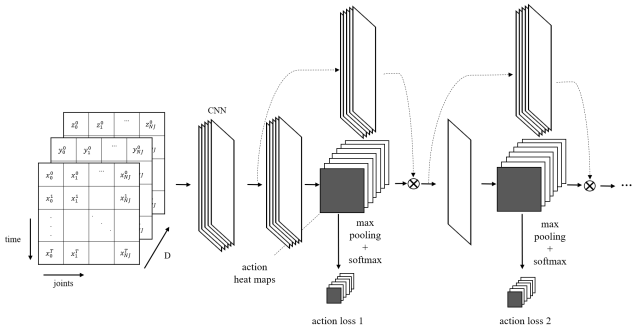


Figure 6. *Representation of the architecture for action recognition*

Appearance based activity recognition and Activity aggregation

The appearance based part is similar to the pose based part, except that it uses the local appearance features instead of joint coordinates. To extract the localized appearance features, the visual features $\mathbf{F}_t \in R^{W_f \times H_f \times N_f}$ obtained from the end of the global entry flow is multiplied to $\mathbf{F}_t$ to the probability map $\mathbf{M}_t \in R^{W_f \times H_f \times N_J}$ obtained at the end. $W_f \times H_f$ is the size of the feature map, N_f is the number of features, and N_J is the number of joints. Instead of performing multiplication to each value individually, as in the Kronecker product, multiplying each channel yields a tensor of $R^{W_f \times H_f \times N_J \times N_f}$. Then the spatial dimension is reduced to $N_J \times N_f$ to create an appearance feature of size $R^{N_J \times N_f}$ for a time duration t. For a series of frames, we have $t = \{0, 1, ..., T\}$ video clip appearance feature $\mathbf{v} \in R^{T \times N_J \times N_f}$. In Fig. 7, the process of extracting the appearance features above, is clarified. The appearance feature is fed into an action recognition network similar to the pose-based action recognition block shown in Fig. 6, and the function of the visual feature replaces the coordinate of the body joint. The approach has two benefits for the appearance based part of the multitask algorithm. First, most of the calculations are shared, so the computation efficiency is very high. Second, the extracted visual features are more robust because they are trained simultaneously for different tasks and datasets.

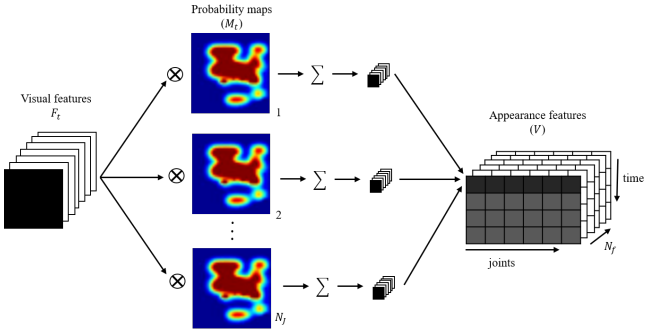


Figure 7. *Appearance feature is calculated by the result of convolution between visual features and probability maps.*

Some poses are high-level pose representations and are difficult to distinguish from other actions. For example, drinking and calling actions are similar when considering body joints. However, visual information corresponding to the target cup and phone is easily separated. To obtain the contribution of the pose and appearance based model, we provide the final prediction of the model by combining each prediction with a fully-connected layer including Softmax activation.

Experiments

Evaluation on activity recognition

We perform quantitative evaluation of the 2D pose estimation using the probability of the correct keypoint measure for head size (PCKh) as Table 1.

Table1. Comparison results on MPII. @0.5 is when the threshold = 50% of the head bore link.

Methods	PCKh@0.5	AUC@0.5
Recurrent VGG [11]	88.1	58.8
Heatmap regression [12]	89.7	59.6
DeepCut [13]	88.5	60.8
2D Soft-argmax	89.8	61.2

PCKh is a detected joint, which is correct if the distance between the predicted and the true joint is within a certain threshold. The MPII dataset for single person pose estimation consists of 25000 images, 15000 for train sample, 3000 for validation sample and 7000 for test sample. The results show that the Soft-argmax based regression method is a good approach, especially when considered under the accumulated precision given by the area under a ROC curve (AUC). AUC measures the entire two-dimensional area underneath the entire ROC curve.

In Human3.6M, we evaluate the proposed 3D pose regression method with the mean per joint position error (MPJPE), which is the most challenging in this dataset. For training, we use 50% data in MPII and Human3.6M. Our experimental results are shown in Table 2.

We evaluate the action recognition approach to 2D scenarios of the Penn Action dataset. We use mixed data of MPII (75%) and Penn Action (25%) for pose estimation. Results are shown in Table 3.

NTU's skeleton data trains pose estimation of NTU data 10%, 45% for MPII, and 45% for Human 3.6M because there are many noisy. Most of the previous methods use the pose provided by kinect. Our approach improves accuracy by using RGB frames and 3D predicted pose. This is shown in Table 4.

Conclusion

In this paper, we propose a CNN architecture for performing 2D and 3D pose estimation with activity recognition. Our model first predicts the 2D and 3D location of the body joints in the raw RGB frames. These locations are used to predict actions performed in the following two ways. In other words, we use visual information by using semantic information using temporal evolution of body joint coordinates and performing attention-based pooling on human body parts. We can solve this problem by sharing a lot of weight and feature in our model. Four tasks such as 2D pose estimation, 3D pose estimation, 2D action recognition, and 3D action recognition are performed in a single model very efficiently compared to dedicated approaches. We have conducted an extensive experiment that shows that our approach can be equal to or better than a dedicated approach to all these tasks.

References

[1] G. Ch'eron, I. Laptev, and C. Schmid. P-CNN: Pose-based CNN Features for Action Recognition. In ICCV, 2015.

[2] I. Kokkinos. Ubernet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. Computer Vision and Pattern Recognition (CVPR), 2017.

[3] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. European Conference on Computer Vision (ECCV), pages 483–499, 2016.

[4] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[5] F. Baradel, C. Wolf, and J. Mille. Pose-conditioned spatiotemporal attention for human action recognition. arxiv, 1703.10106, 2017.

[6] D. C. Luvizon, H. Tabia, and D. Picard. Human pose regression by combining indirect part detection and contextual information. CoRR, abs/1710.02322, 2017.

[7] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned Invariant Feature Transform. Euro-pean Conference on Computer Vision (ECCV), 2016.

[8] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[9] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR, abs/1602.07261, 2016.

[10] Diogo C. Luvizon, David Picard, Hedi Tabia. 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning (CVPR). 2018.

[11] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. CoRR, abs/1605.02914, 2016.

[12] A. Bulat and G. Tzimiropoulos. Human pose estimation via Convolutional Part Heatmap Regression. In European Conference on Computer Vision (ECCV), pages 717–732, 2016.

[13] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B.

Table2. Comparison Human3.6M[20, 21] evaluated on the averaged joint error.

Methods	Eat	Greet	Phone	Posing	Purchase	Sitting	Smoke	Walk
Pavlakos *et al.* [4]	66.7	69.1	71.9	65.0	68.3	83.7	71.4	59.1
Mehta *et al.*[14]	55.4	62.3	71.8	52.6	72.2	86.2	66.0	48.9
Ours	53.8	58.7	68.9	52.1	67.9	79.7	64.4	47.4

Table3. Comparison results on Penn Action. Result given as the percentage of correctly classified actions(Accuracy).

Methods	RGB	Estimated poses	Optical Flow	Accuracy
Nie *et al.* [15]	Used	-	Used	85.5
Iqbal *et al.*[16]	Used	Used	Used	92.9
Proposed	Used	-	Used	93.1

Schiele. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In European Conference on Computer Vision (ECCV), May 2016.

[14] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation using transfer learning and improved CNN supervision. CoRR, abs/1611.09813, 2016.

[15] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu. Joint action recognition and pose estimation from video. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[16] U. Iqbal, M. Garbade, and J. Gall. Pose for action - action for pose. FG-2017, 2017.

[17] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In CVPR, June 2016.

[18] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, ECCV, pages 816–833, Cham, 2016.

[19] S. Song, C. Lan, J. Xing, W. Z. (wezeng), and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In AAAI Conference on Artificial Intelligence, February 2017.

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru and Cristian Sminchisescu, Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, No. 7, July 2014.

[21] Catalin Ionescu, Fuxin Li and Cristian Sminchisescu, Latent Structured Models for Human Pose Estimation, International Conference on Computer Vision, 2011.

Author Biography

Jisu Kim received B.S. in Computer Engineering from Keimyung University, Daegu, South Korea in 2019. He is currently in master course in computer engineering at Keimyung University. His research area covers computer vision, machine learning, image and signal processing.

Acknowledgement

Following are results of a study on the ”Leaders in Industry-university Cooperation +” Project, supported by the Ministry of Education and National Research Foundation of Korea and was partly supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2016-0-00564, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding.

Table4. Comparison result with the existed work using NTU dataset.

Methods	Kinect poses	RGB	Estimated poses	Accuracy
Shahroudy *et al.* [17]	Used	-	-	62.9
Liu *et al.*[18]	Used	-	-	69.2
Song *et al.*[19]	Used	-	-	73.4
Proposed	-	Used	Used	78.4

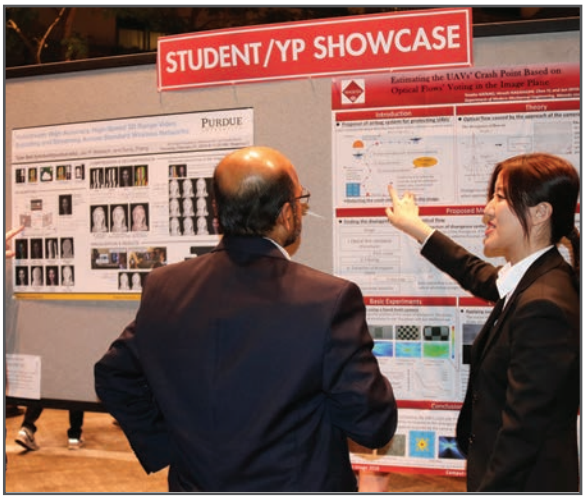
JOIN US AT THE NEXT EI!

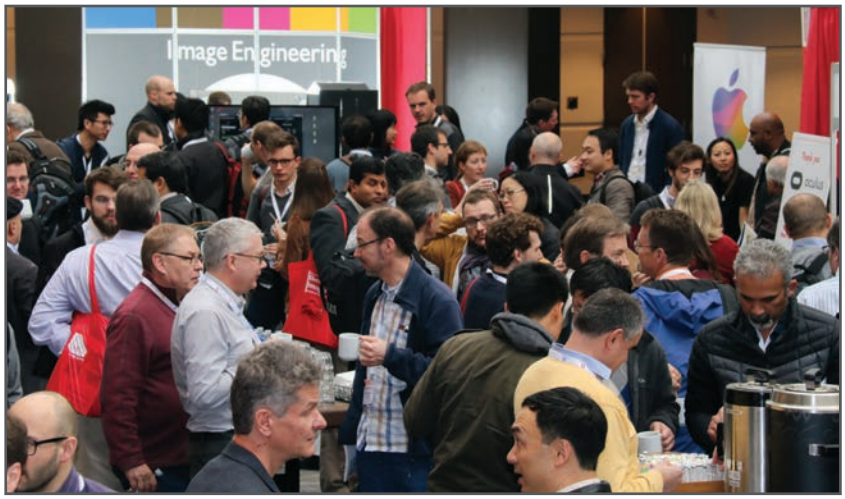
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org