Journal of Korea Multimedia Society Vol. 25, No. 4, April 2022(pp. 575-582)
https://doi.org/10.9717/kmms.2022.25.4.575

Skeleton 정보와 LSTM을 이용한 작업자 동작인식

전왕수†, 이상용††

Motion Recognition of Workers using Skeleton and LSTM

Wang Su Jeon†, Sang Yong Rhee††

ABSTRACT

In the manufacturing environment, research to minimize robot collisions with human beings have been widespread, but in order to interact with robots, it is important to precisely recognize and predict human actions. In this research, after enhancing performance by applying group normalization to the Hourglass model to detect the operator motion, the skeleton was estimated and data were created using this model. And then, three types of operator's movements were recognized using LSTM. As results of the experiment, the accuracy was enhanced by 1% using group normalization, and the recognition accuracy was 99.6%.

Key words: Pose Estimation, Pose Classification, Hourglass, LSTM, CNN, Group Normalization

1. 서 론

오늘날 제조업은 인건비 절감, 생산성 증가와 안전성까지 확보해야 하는 과제에 직면해 있어서, 제조과정의 자동화가 필요하다. 로봇을 이용한 완전 자동화도 있지만, 사람과 협업하며 작업하는 협업 로봇(Collaborative Robot)의 수요도 증가하고 있어서 협업 로봇에 관한 연구도 증가하고 있다. 협업 로봇은 사람과 작업공간을 공유하고 물리적인 상호작용을 할 수 있는 것을 말한다.

기존의 현장에서는 로봇과 충돌하여 작업자가 다치는 경우가 있었었고, 이러한 사고를 최소화하기 위해 센서를 이용하여 작업자와 로봇의 충돌을 예방하는 연구가 많이 진행되었다. 그러나 로봇과 실질적인 협업을 하기 위해서는 로봇이 사람의 동작을 정확하게 알고 다음 동작을 예측할 수 있어야 한다. Alejandro[1] 등은 Fig. 1[2]과 같이 RGB 영상에서 사람 몸의 키포인트를 검출하여, 사람의 자세를 인식하였는데 이것을 확장하면 인간과 로봇 간의 상호작용이 가능하다.

작업자의 자세는 대부분 자이로 센서나 카메라를 이용하여 인식하였는데, 센서를 사용하면 동작 인식은 가능했지만, 매번 신체에 부착해야 하는 단점이 있다. 카메라를 사용하여 자세를 검출하는 연구는 정확도 측면에서 좋지 않았다. 그러나 최근 딥러닝 발전으로 인하여 정확도가 좋아졌고 동작 인식 속도도 실시간이 가능해졌다[3-4]. 심층신경망은 Stacked hourglass 모델을 많이 사용한다.

Stacked hourglass 모델은 모래시계 형태의 모듈을 반복적으로 쌓는 모델이다. 이 방법은 특징의 정보가 조정되는 이점이 있었으나, 모델의 깊이가 깊어짐에 따라 학습이 불안정하여 정확도 편차가 큰 문제

※ Corresponding Author : Sang Yong Rhee, Address: (51767) 7 Kyungnamdaehak-ro, Masanhappo-gu, Changwon, Korea, TEL : +82-55-249-2706, FAX : +82-55-248-2554, E-mail : syrhee@kyungnam.ac.kr
Receipt date : Mar. 22, 2022, Revision date : Apr. 2, 2022
Approval date : Apr. 5, 2022

† Dept. of Computer Engineering., Kyungnam University (E-mail : jws2218@naver.com)
†† Dept. of Computer Engineering., Kyungnam University
※ This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

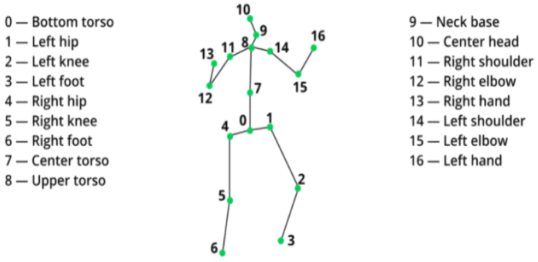


Fig. 1. Examples of body keypoints.

점 있었다. 본 논문에서는 현장에 반복적인 동작을 수행하는 작업자의 자세를 검출하기 위해 일반화가 잘될 수 있도록 그룹 정규화(Group Normalization)를 사용하여 모델을 개선한다. 이 모델을 이용하여 걸음(Walking), 이동(Moving), 집다(Grapping) 3가지 동작의 데이터를 생성하고 LSTM(Long Short Term Memory)을 이용하여 분류하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 동작 인식에 관한 기존의 연구를 소개한 다음, 3장에서 제안하는 동작 인식 방법에 관해 설명한다. 여기에서는 사용된 자세 검출 방법과 동작 분류 방법에 관해 설명한다. 그리고 4장에서는 제안한 방법의 실험방법과 결과를 비교하고 5장에서 결론을 맺는다.

2. 기존 연구

작업 도중에 발생은 부상은 작업자가 서투른 자세로 장시간 작업을 하거나 쉬는 과정에서 쪼그리는 등으로 인해 발생[5,6] 하므로, 작업자의 자세에 따른 부하 최소화를 위한 연구가 많이 진행되었다[7]. 이러한 연구들은 신체 관절의 각도 정보를 활용하며, 작업자의 자세 검출 중심의 연구가 대부분이다[8]. 이 방법들은 환경에 따라 자세 검출 정확도의 차이가 발생하였다. 그러나 Kinect의 3D depth 카메라를 사용하거나 CNN을 이용하면[9], 비교적 정확하게 자세 검출이 가능하다.

기존의 자세 검출 방법은 신체의 각 부위를 나누어 각도 등의 정보를 추출한 후, 이 정보를 결합하는 방법을 사용하였으며[10-12], 트리구조의 그래픽 모델을 사용하여 작업자의 동작을 모델링하여 나타내고 분석하였다[13-16]. Tompon[17] 등, Pfister[18] 등, 그리고 Weit[19] 등은 넓은 수용범위를 얻기 위해 atrous covnolution을 적용한 CNN 모델을 개발하여 자세 검출을 수행하였다.

인간 몸의 특성을 고려하여 2D 영상에서 3D 골격과 관절을 예측하고 시각화하기 위한 연구가 많이 진행되었고, openpose[20], stacked hourglass[1] 모델 등을 이용하여 작업자의 자세를 검출하였다. 그러나 이 방법들은 공장 환경에서 적용된 사례가 없으며, 로봇과 협업보다는 동작을 예측하거나 따라 하는 연구와 손동작을 인식하는 연구[21,22]가 대부분이다.

3. 작업자의 동작인식

본 논문에서는 Fig. 2와 같이 자세 검출(Pose estimation)에 많이 사용되는 stacked hourglass 모델을 이용하여 스켈레톤 추출한다. 스켈레톤 키포인트의 정확도 향상을 위해 그룹 정규화를 이용하여 모델을 개선한다. 그 후 이 모델을 이용하여 작업자의 3가지 동작의 데이터 셋을 만들고 LSTM을 이용하여 분류한다.

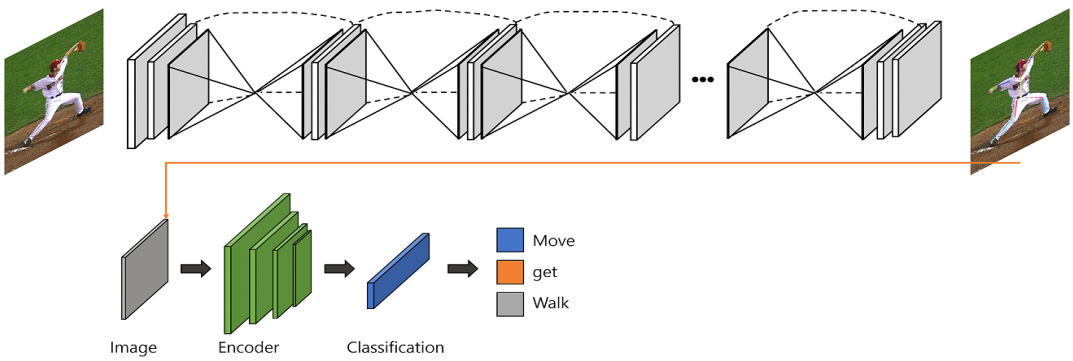


Fig. 2. System overview.

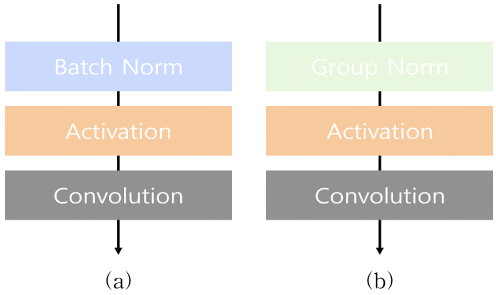


Fig. 3. Transformation result of conv module (a) with out group norm and (b) with group norm.

3.1 자세검출

본 연구에서는 자세 검출을 위해서 stacked hourglass 모델을 이용하여 추출한 스켈레톤을 사용한다. Fig. 3과 같이 배치 정규화를 그룹 정규화로 대체하여 학습을 수행하였다. 배치 정규화(Batch Normalization)는 배치단위로 계산을 수행하기 때문에 배치 크기에 영향을 받는다. 유사한 것으로 레이어 정규화(Layer Normalization)와 인스턴스 정규화(Instance Normalization) 등이 있고, Fig. 4와 같이 비교할 수 있다[23].

Fig. 4의 레이어 정규화는 각 채널과 영상전체를 모두 정규화하는 것이고 인스턴스 정규화는 채널별로 정규화시켜준다. 레이어 정규화와 인스턴스 정규화는 배치크기에 의존하지 않으므로 다른 모델에 적용했을 때 잘 동작하는 것을 볼 수 있다. 하지만 이 방법은 영상인식에서는 좋은 성능을 보이지 못하기 때문에 그룹 정규화를 사용한다[23]. 그룹 정규화는 각 채널을 N개의 그룹으로 나누어 정규화하는 방법으로 채널이 6개 있고, G가 2이면 한 그룹당 3개의 채널이 만들어진다. 본 연구에서는 식 (1)과 같이 그룹의 개수인 G의 값을 32로 기본 설정하여 사용한다.

이때, 식 (1)의 S_i는 계산된 결과값을 의미하고, k_N와 i_N는 특징맵을 의미하고, N은 배치사이즈 축이고, C는 채널축, H,W는 특징맵의 축을 의미한다.

$$S_i = \left\{ k \mid k_N = i_N, \left\lfloor \frac{k_C}{C/G} \right\rfloor = \left\lfloor \frac{i_C}{C/G} \right\rfloor \right\} \quad (1)$$

3.3 동작분류

앞 절에서 설명한 모델로부터 생성된 데이터를 이용하여 동작 분류를 수행한다. 연속적인 동작을 분석하기 위하여 LSTM을 사용한다. 배치 사이즈는 8이고 epoch는 100을 사용한다. optimizer는 adam을 사용하고 학습율(Learning Rate)는 0.0001로 설정하였다. LSTM은 Fig. 5와 같이 forget gate, input gate, output gate가 있다. forget gate에서는 이전 상태의 정보를 반영비율을 조절하고, input gate에서는 이전 상태와 입력 데이터를 현재 상태에 어느 정도 반영할지를 조절한다. 그 후 output gate에서는 최종값을 갱신하여 다음 상태로 전달한다.

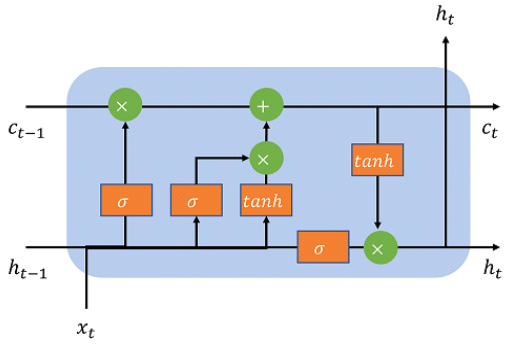


Fig. 5. LSTM model overview.

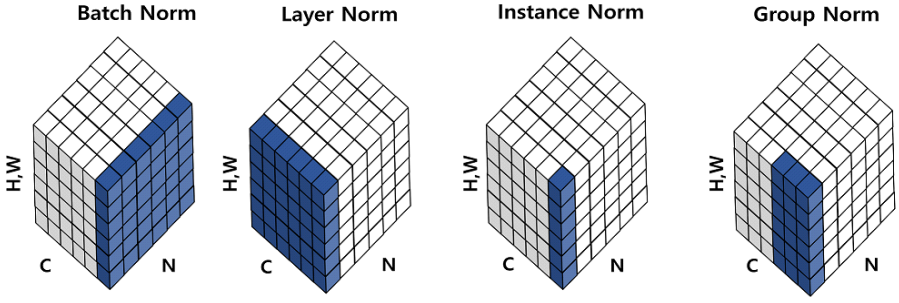


Fig. 4. Comparison of normalization.

4. 실험 결과 및 고찰

4.1 실험환경

본 논문의 실험환경은 다음과 같다. 운영체제는 Ubuntu 18.04.4 환경이고, CPU는 Intel Xeon Gold 5120을 사용하고 메모리는 128GB이다. 그리고 병렬처리 보드인 그래픽카드는 RTX TITAN X 24GB를 사용하고, 딥러닝 프레임워크는 Tensorflow 1.15.0을 사용한다. 실험에 사용된 데이터는 LSP(Leeds Sports Pose) 데이터[24]와 MPII(MPII Human Pose) 데이터[25]를 사용하여 실험한다. LSP는 스포츠 경기 중인 이미지를 수집하여 만든 데이터셋으로 2,000장의 사진을 가지고 있다. 각 이미지에는 14개의 관절 좌표를 가지고 있다. MPI 데이터는 유튜브 비디오에서 이미지를 추출하여 약 25,000장의 이미지로 구성되었다. 각 이미지에는 관절 좌표, 3D 좌표, 방향 정보가 있다.

본 실험에서는 모델의 안정적인 학습을 위해 기존의 모델에 그룹 정규화를 적용한 후 비교분석을 수행한다. 그 후에 개선된 모델을 이용하여 3개의 클래스인 걸음, 이동, 집다 데이터를 총 2,000장 생성한다. 이때, 이동은 작업자가 물건을 집은 후 이동하는 것을 말하고, 걸음은 그냥 걷는 것을 의미한다. 생성된 데이터를 이용하여 학습에는 1,700장을 사용하고 테스트에는 300장으로 LSTM에 사용하여 성능을 비교한다.

4.2 자세검출 정확도 비교결과

성능을 평가하기 위해 제안하는 방법과 hourglass 모델의 stack을 1과 8로 변경하여 비교하였다. Fig. 6과 같이 MPII 데이터를 이용하여 실험한 결과는 제안하는 방법이 안정적인 학습과 정확도를 보였고, stack의 크기를 조절해도 안정적인 그래프를 보였다. Fig. 7은 LSP 데이터를 이용한 결과를 나타내며 Fig. 6과 같이 제안하는 방법이 안정적인 학습을 보

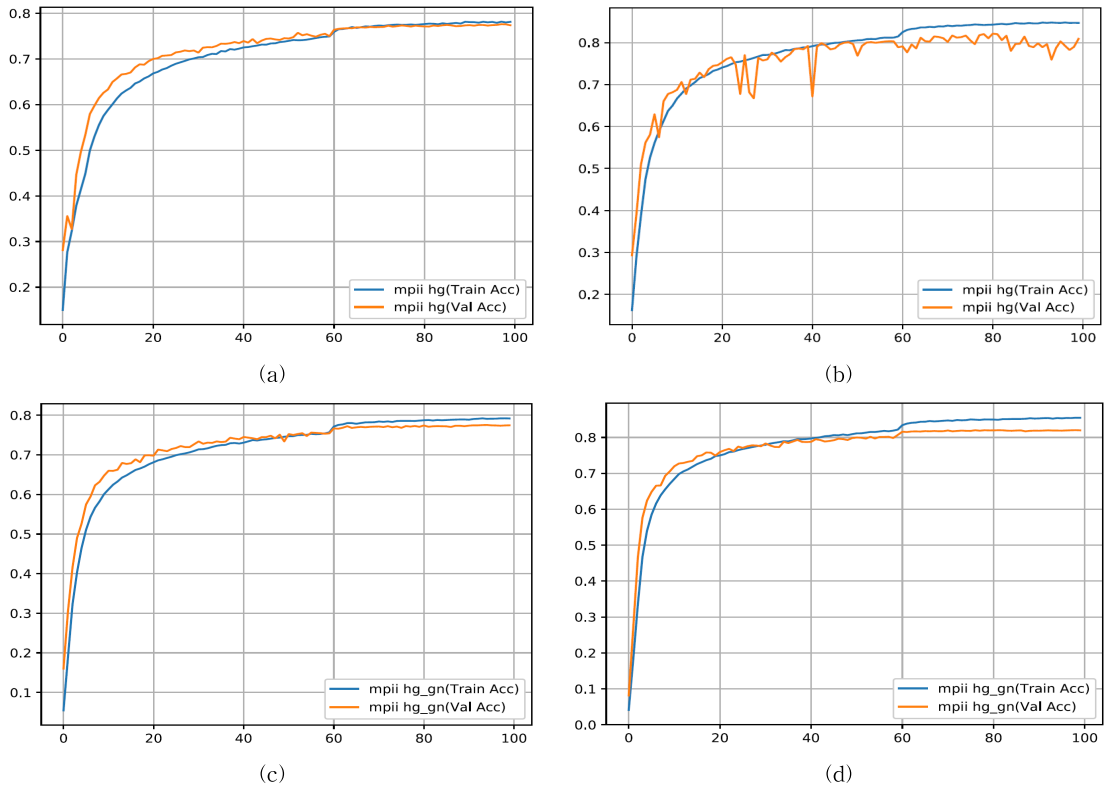


Fig. 6. Comparison of train and valid performance for MPII dataset (a) without Group normal model stack 1, (b) without Group normal model stack 8, (c) with Group normal model stack 1, and (d) with Group normal model stack 8.

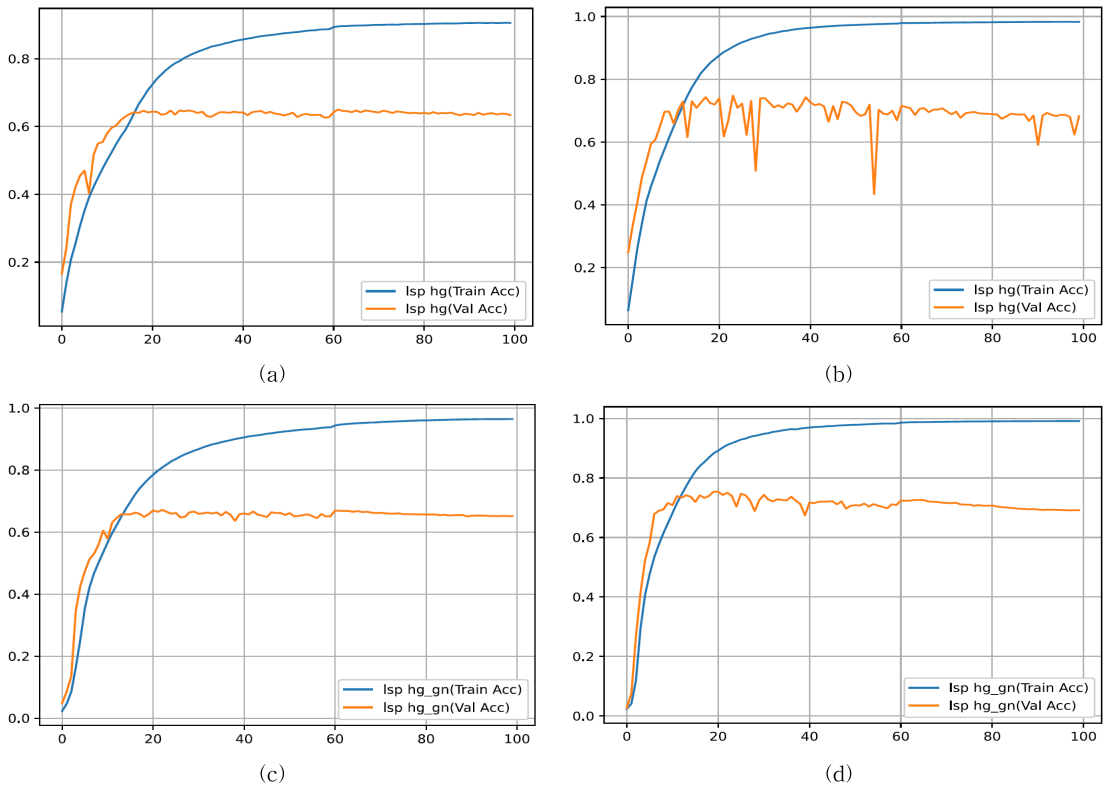


Fig. 7. Comparison of train and valid performance for LSP dataset (a) without Group normal model stack 1, (b) without Group normal model stack 8, (c) with Group normal model stack 1, and (d) with Group normal model stack 8.

Table 1. The performance for proposed algorithm by dataset.

	With Group normalization		Without Group normalization	
	Stack 1	Stack 8	Stack 1	Stack 8
MPII Dataset	77.4	81.9	77.3	80.9
LSP Dataset	36.4	68.2	65.1	69.2

였다. 이로 인해 그룹 정규화가 레이어가 깊어짐에 따라 안정적인 그래프를 보였다. 그리고 이 그래프의 지표를 Table 1에 나타내어 비교하였고, 기존의 방법보다 제안하는 방법이 정확도가 1% 높았다.

Table 2는 스켈레톤의 부위별 정확도를 나타낸 것이다. Table 2와 같이 머리(Head) 부분에서는 그룹 정규화가 낮은 정확도를 보였지만 나머지에서는 높은 정확도를 보였다. 그리고 평균적으로 0.77%의 정확도가 향상되었다.

4.3 동작분류 정확도 비교결과

4.2에서 언급한 자세검출 모델에서 생성된 데이터

Table 2. The performance for proposed algorithm by body parts.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
With Group normalization	96.11	93.50	84.47	79.61	84.92	80.45	76.95	85.26
Without Group normalization	96.32	93.48	83.79	79.29	82.29	79.00	76.62	84.49

Table 3. The performance for proposed algorithm.

Confusion Matrix			
	Moving	Grapping	Walking
Moving	97	0	3
Grapping	1	99	0
Walking	5	0	95
Classification score			
	LSTM		Mobilenet
Accuracy	99.6		99.8

를 이용하여 동작분류를 수행한다. 이때, Mobilenet [25]과 LSTM[26]을 이용하여 학습하고 테스트하였다. 테스트한 결과는 Table 3과 같으며 혼동행렬(Confusion matrix)을 비교하여보니 '걸음' 동작과 '이동' 동작이 유사하여 3~5개 정도 오차가 발생하였다. 그리고 '집다' 동작은 단순하여서 정확도가 높았다. 그리고 Mobilenet과 LSTM의 정확도를 비교한 결과 Mobilenet이 0.2% 높은 정확도를 보였고 LSTM 모델도 일반적인 시계열 모델들과 비교한다면 좋은 결과를 보인다는 것을 알 수 있다.

5. 결 론

본 논문에서는 작업자의 동작인식을 위해 Stacked hourglass 모델에 그룹 정규화를 사용하여 개선하였고, 학습 안정성을 높였다. 그룹 정규화를 적용한 결과 기존의 모델보다 0.7% 성능개선이 있었고, 학습과 테스트 과정에서 1% 성능이 개선됨을 볼 수 있었다. 그리고 이 모델을 이용하여 데이터를 생성하여 LSTM으로 3가지 동작을 분류하였다. 동작을 분류한 결과 99.6%의 정확도를 보였다.

차후 연구에서는 로봇과 사람이 협업할 수 있도록 몸의 각 부분의 이동할 위치를 추정할 수 있도록 연구를 진행할 예정이다.

REFERENCE

[1] N, Alejandro, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," *European Conference on Computer Vision*, pp. 483-499, 2016.

[2] 3D Human Pose Estimation Experiments and Analysis(2020), https://www.kdnuggets.com/2020/08/3d-human-pose-estimation-experiments-analysis.html (accessed March 14, 2022).

[3] B. Tekin, S.N. Sudipta, and P. Fua, "Real-Time Seamless Single Shot 6D Object Pose Prediction," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 292-301, 2018.

[4] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-Wise Voting Network for 6-DOF Pose Estimation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4561-4570, 2019.

[5] C. Jiayu, Q. Jun, and A. Changbum, "Construction Worker's Awkward Posture Recognition through Supervised Motion Tensor Decomposition," *Automation in Construction*, Vol. 77, pp. 67-81, 2017.

[6] K. Osmo, K. Pekka, and K. likka, "Correcting Working Postures in Industry a Practical Method for Analysis," *Applied Ergonomics*, Vol. 8, No. 4, pp. 199-201, 1977.

[7] W. Umer, H. Li, G.P.Y. Szeto, and A.Y.L. Wong, "Identification of Biomechanical Risk Factors for The Development of Lower-Back Disorders During Manual Rebar Tying," *Journal of Construction Engineering and Management*, Vol. 143, No. 1, 2017.

[8] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision*, Vol. 61, pp. 55-79, 2005.

[9] K. Adhikari, H. Bouchachia, and H.N. Charif, "Deep Learning Based Fall Detection Using Silplified Human Posture," *International Journal of Computer and Systems Engineering*, Vol. 13, No. 5, 2019.

[10] D. Ramanan, D.A. Forsyth, and A. Zisserman, "Strike a Pose: Tracking People by Finding Stylized Poses," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 271-278, 2005.

[11] M. Andriluka, S. Roth, and B. Schiele, "Mo-

nocular 3D Pose Estimation and Tracking by Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 623-630, 2010.

[12] Y. Wang and G. Mori, "Multiple Tree Models for Occlusion and Spatial Constraints in Human Pose Estimation," *European Conference on Computer Vision*, pp. 710-724, 2008.

[13] L. Sigal and M.J. Black, "Measure Locally, Reason Globally: Occlusion-Sensitive Articulated Pose Estimation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2041-2048, 2006.

[14] X. Lan and D. P. Huttenlocher, "Beyond Trees : Common-Factor Models for 2D Human Pose Recovery," *International Conference on Computer Vision*, pp. 470-477, 2005.

[15] L. Karlinsky and S. Ullman, "Using Linking Features in Learning Non-Parametric Part Models," *European Conference on Computer Vision*, pp. 326-339, 2012.

[16] J.J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation," *Advance in Neural Information Processing Systems*, pp. 1799-1807, 2014.

[17] T. Pfister, J. Charles, and A. Zisserman, "Flowing Convnets for Human Pose Estimation in Videos," *International Conference on Computer Vision*, pp. 1913-1921, 2015.

[18] S.E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724-4732, 2016.

[19] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *International Conference on Computer Vision*, pp. 2961-2669, 2017.

[20] Z. Cao, T. Simon, S.E. Wei, and Y. Sheikh, "Real Time Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291-7299, 2017.

[21] J.Y. Kim and J.I. Park, "HOG-HOD Algorithm for Recognition of Multi-cultural Hand Gestures," *Journal of Korea Multimedia Society*, Vol. 20. No. 8, pp. 1187-1199, 2017.

[22] K.J. Cheoi and J.H. Han, "A Novel Door Security System using Hand Gesture Recognition," *Journal of Korea Multimedia Society*, Vol. 19, No. 8, pp. 1320-1328, 2016.

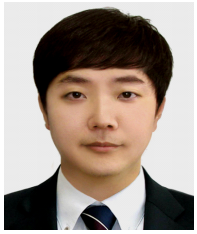
[23] M. Sandler, A. Howard, M. Zhu, W. Yuxin, and K. He. "Group Normalization," *European Conference on Computer Vision*, pp. 3-19, 2018.

[24] S. Johnson and M. Everingham, "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation," *British Machine Vision Conference*, Vol. 2, No. 4, pp. 1-11, 2010.

[25] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the art Analysis," *IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686-36893, 2014.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.C. Chen. "Mobilenet v2: Inverted Residuals and Linear Bottlenecks," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520, 2018.

[27] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, No. 9. Vol. 8, pp. 1735-1780, 1997.



전 왕 수

2016년 경남대 컴퓨터공학과 졸업
2018년 경남대 대학원 융합IT 공학과(공학석사)
2022년 경남대 대학원 융합IT 공학과(공학박사)
2022년~현재 경남대 컴퓨터공학부 박사 후 과정

관심분야 : 컴퓨터 비전

이 상 용

1982년 고려대 산업공학과 졸업
1984년 고려대 대학원 산업공학과 (공학석사)
1992년 포항공대 대학원 산업공학과(공학박사)
1992년~현재 경남대 컴퓨터공학부 교수

관심분야 : 컴퓨터 비전, 증강현실, 뉴로-퍼지, 인간-로봇 인터페이스