ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 19-04-26
180
j.inst.Korean.electr.electron.eng.Vol.23,No.4,1295~1301,December 2019
http://dx.doi.org/10.7471/ikeee.2019.23.4.1295

딥 러닝 기반 휴먼 모션 디노이징
Deep Learning-Based Human Motion Denoising

김 성 욱*, 임 현 승*, 김 종 민*★

Seong Uk Kim*, Hyeonseung Im*, Jongmin Kim*★

Abstract

In this paper, we propose a novel method of denoising human motion using a bidirectional recurrent neural network (BRNN) with an attention mechanism. The corrupted motion captured from a single 3D depth sensor camera is automatically fixed in the well-established smooth motion manifold. Incorporating an attention mechanism into BRNN achieves better optimization results and higher accuracy than other deep learning frameworks because a higher weight value is selectively given to a more important input pose at a specific frame for encoding the input motion. Experimental results show that our approach effectively handles various types of motion and noise, and we believe that our method can sufficiently be used in motion capture applications as a post-processing step after capturing human motion.

요 약

본 논문에서는 어텐션 기법을 적용한 양방향 순환신경망을 이용하여 새로운 휴먼 모션 디노이징 방법을 제안한다. 본 방법을 이용하면, 단일 3D 깊이 센서 카메라에서 캡처된 노이즈가 포함된 사람의 움직임이 잘 교정된 자연스러운 움직임으로 자동 조정된다. 양방향 순환신경망에 어텐션 기법을 도입하면, 입력으로 들어온 움직임을 인코딩할 때 여러 자세 중에 더 중요한 자세가 있는 프레임에 더 높은 어텐션 가중치를 부여함으로써, 다른 딥 러닝 네트워크와 비교해 더 나은 최적화 결과와 더 높은 정확도를 보인다. 실험을 통해 본 논문에서 제시한 방법이 다양한 스타일의 움직임과 노이즈를 효과적으로 처리함을 확인하였으며, 제시한 방법은 모션 캡처 후처리 단계의 애플리케이션으로 충분히 사용 가능할 것으로 기대된다.

Key words : human motion, motion capture, motion denoising, attention, bidirectional recurrent neural network

Ⅰ. 서론

모션 캡처를 위해 액터의 움직임을 촬영하는 인기 있는 광학식 모션 캡처 시스템 중 하나는 적외선 링 라이트를 가지고 있는 여러 대의 카메라를 이용하여 반사가 잘 일어나는 마커를 액터에게 몇 개 부착한 뒤, 마커의 가상공간 안에서의 3D 위치를 추적하는 것이다. 광학식 모션 캡처 시스템은 비록 높은 정확도와 속도를 제공하지만, 일반적으로 그 규모가 크고 사전 준비 없이는 사용하기 어렵다는 단점이 있다. 이러한 문제점을 해결하기 위한 대안으로, 최근 마커를 부착하지 않는 마커리스

* Dept. of Computer Science, Kangwon National University
★ Corresponding author
E-mail : jongmin.kim@kangwon.ac.kr, Tel : +82-33-250-8447
※ Acknowledgment
This study was supported by 2017 Research Grant from Kangwon National University.
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1F1A1063467).
This is an extended version of the poster presentation that appeared in The 12th ACM SIGGRAPH Asia [21].
Manuscript received Dec. 10, 2019; revised Dec. 22, 2019; accepted Dec. 26, 2019.

모션 캡처 시스템이 각광받고 있다. 일례로, 저가형 3D 깊이 센서인 Microsoft 사의 키넥트 센서를 이용하면, 마커 부착 없이도 실시간으로 3D 움직임을 생성할 수 있다. 하지만 self-occlusion 등의 문제로, 생성된 움직임에 노이즈가 들어가거나 사람답지 않은 움직임이 만들어지는 등, 깊이 센서 기반 모션 캡처 특유의 문제점이 존재한다. 사람의 움직임은 시간과 공간에 서로 복잡하게 얽혀있기 때문에 기존 방법인 Kalman 필터나 Gaussian low-pass 필터 등으로 세부적인 움직임이나 그 맥락을 해치지 않으면서 움직임의 노이즈만 제거하는 것은 어려운 문제이다. 물론 수작업으로 노이즈가 있는 움직임을 교정할 수는 있으나, 이는 긴 시간과 노동력이 필요한 지루한 작업이다.

본 논문에서는 이처럼 키넥트 센서에서 캡처한 데이터에 필연적으로 존재하는 다양한 형태의 노이즈를 자동으로 제거하기 위해 새로운 딥 러닝 기반 아키텍처를 제안한다. 구체적으로, 노이즈에 오염된 모션 데이터를 자연스럽게 고치기 위해 미리 녹화된 모션 데이터로 학습된 양방향 순환신경망을 사용한다. 양방향 순환신경망을 사용하는 주된 이유는 양방향 순환신경망이 시간 축에 대해 정방향과 역방향의 입력을 모두 반영한 뒤 결과를 계산하기 때문에 다른 신경망에 비해 시계열 데이터 입력에 대해 더 좋은 결과를 보여주기 때문이다. 또한, 어텐션(Attention) 기법을 양방향 순환신경망에 도입하여 전체 입력 데이터의 가중치를 계산하여 선택적으로 중요한 특징들을 선별함으로써, 빠르게 네트워크를 학습하고, 어떤 특징이 다른 특징들과 비교해 네트워크에서 중요하게 작용하였는지 확인하였다. 추가로, 관절의 위치와 각도 등 다양한 조합의 입력 특징들을 실험하여 가장 좋은 특징 조합을 확인하였다. 또한, 제시한 방법을 다른 최신 인공신경망 구조와 정성적, 정량적으로 비교하였으며, 제시한 방법이 떨림이나 부자연스러운 모션 데이터를 생성하지 않고 가장 유망하고 경쟁력 있는 결과를 보여주는 것을 확인하였다.

II. 본론

1. 관련연구

가. 다양한 움직임에 대한 학습

사람의 움직임을 기록한 데이터에서 다양한 움직임을 생성하는 것은 컴퓨터그래픽스 및 기계학습 분야 연구자들의 주요 관심사 중 하나이다. 주성분 분석(PCA)을 이용하여 사람의 보행을 다양하게 생성하는 접근법은 성공적이었으나 다양한 움직임을 함께 다루기에는 적절하지 않았다. [1]에서는 국소적 주성분 분석을 이용하여 다양한 종류의 사람 움직임을 생성하였으며, 이를 저차원의 신호들로 변환한 후 움직임을 합성하는 데 사용하였다. [2]에서는 모션장(Field)이라 불리는 자료구조를 도입하여 사용자가 다양한 움직임을 상호작용하며 조절할 수 있게 하였다. 이러한 방법들은 각 프레임 사이의 거리를 재거나 데이터 처리에 걸리는 시간을 단축하기 위해 KD 트리를 이용하는 등, 상당히 많은 양의 데이터 전처리 과정을 포함하고 있다. [3]에서는 방사형 기저 함수를 이용한 예제 기반 Inverse Kinematics (IK)를 이용하여 신체 일부의 움직임을 전체에 매핑할 수 있게 하였다. [4]에서는 [3]에서의 연구를 발전시켜 Gaussian Process를 이용하여 더 부드러운 매핑을 만들어냈다. [5]에서는 Gaussian Process Latent Variable Model을 이용하여 저차원 데이터를 고차원으로 매핑하는 방법을 소개하였고, 이러한 방법은 [6]에서 사람의 움직임을 합성하는 데 사용되었다. 하지만, 방사형 기저 함수나 Gaussian Process 같은 비선형 기저 기반 방법은 공분산 행렬의 크기가 데이터 크기의 제곱에 비례해 커지므로 그 규모를 확장하기가 어렵다. 추가로, 이러한 연구들은 모두 이미 학습을 위해 설정된 움직임의 일부 구간만 인코딩할 수 있다는 한계가 있다.

나. 인공신경망과 모션 데이터에의 응용

Convolutional Neural Network (CNN)은 기계학습 및 그 응용 분야에서 효과적으로 사용되고 있다. CNN은 이미지 분류 문제에서 특히 효과적이며 [7, 8], 동영상 분류[9], 얼굴 인식[10], 동작 분류[11] 및 추적[12], 음성인식[13], 다양한 모션 데이터[14] 등에 대해서도 높은 성능을 보여주고 있다. 한편, Temporal Convolutional Neural Network (TCNN)은 실시간 음성 품질 개선[15] 및 동작 분할[16], 동작 인식[17], 비디오 초해상화[18] 등 다양한 후속 연구에 적용되고 있다. 본 연구에서도 이러한 관련 연구에서의 모션 데이터에 대한 접근법을 적용하였다.

2. 데이터 전처리

본 논문에서는 어텐션 기반 양방향 순환신경망을 학습시키기 위해 CMU 모션 캡처 데이터베이스[19]를 사용하였다. CMU 모션 캡처 데이터베이스에는 약 10시간에 달하는 2,605가지 종류의 모션 캡처 데이터가 저장되어 있다. 데이터 전처리 과정은 기존 연구[14]와 비슷하게 진행하였다. 모션 캡처 데이터의 뼈대는 Inverse Kinematics[20]를 계산하여 정규화된 길이의 21개의 관절을 가진 뼈대로 리타겟팅(retargeting) 된다. 이 과정에서 x-z 평면상의 뼈대 중심의 글로벌 이동과 뼈대의 y축 회전 또한 같이 제거된다. 제거된 글로벌 이동과 회전은 원본 데이터를 복원할 때에 사용할 수 있도록 입력 벡터에 추가되어 입력 벡터의 크기는 63 + 3이 된다. 네트워크의 학습을 위해 전체 모션 데이터를 240프레임 단위로 잘라내었으며, 실제 노이즈 입력에 대해 어떤 노이즈가 가장 잘 동작하는지 확인하기 위해 Gaussian 노이즈, Gamma 노이즈 등 다양한 종류의 노이즈를 인위적으로 학습데이터에 추가하였다. 하지만 위와 같은 노이즈를 사용하였을 경우 만족할만한 결과를 얻지 못하였는데, 이는 실제 키넥트 데이터의 노이즈 분포가 위에서 사용한 노이즈의 분포와 상당히 다르기 때문이다. 이러한 불규칙한 키넥트 노이즈를 수학적으로 정확히 모델링하는 것은 어려운 문제이다. 따라서 이를 해결하기 위해 본 논문에서는 광학식 모션 캡처 시스템과 키넥트를 동시에 사용하여 모션 캡처를 진행하였으며, 이를 통해 키넥트 데이터와 광학식 모션 캡처 데이터의 묶음을 생성하였다. 이후 두 데이터의 차이를 이용하여 노이즈를 모델링하였으며, 모델링된 노이즈의 일부는 랜덤하게, 일부는 광학식 모션 캡처 데이터와 유사성을 측정하여 가장 가까운 입력 데이터에 넣어주며 네트워크 학습을 진행하였다.

3. 네트워크

그림 1은 본 논문에서 제시된 양방향 순환신경망 기반의 네트워크 구조이다. 네트워크의 입력 $X=\{x_1,x_2,x_3,...,x_{66}\}\in R^{240\times 66}$는 노이즈 성분이 포함된 포즈의 관절 3D 위치값과 x-z 평면상의 중심 이동 및 y축 회전이 포함된 벡터이며, 네트워크 출력은 동일한 형태의 노이즈가 제거된 벡터 Y이다. 네트워크는 500개의 셀 상태를 갖는 Long Short-

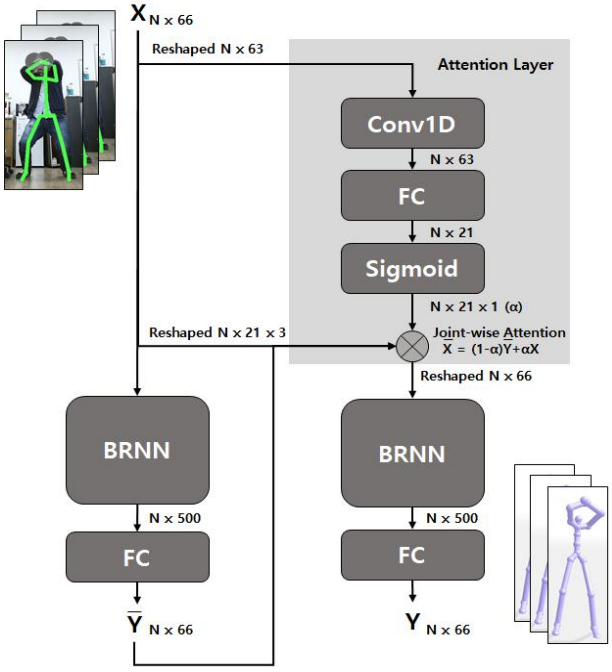

Fig. 1. System overview : We train an attention-based neural network to denoise human motion obtained from Kinect.

그림 1. 네트워크 개요 : 어텐션 기반의 인공신경망을 키넥트로부터 획득한 모션의 노이즈를 제거하도록 학습한다.

Term Memory (LSTM) 셀 3개와 완전 연결 레이어(fully-connected layer)로 이루어져 있으며, 관절 단위의 어텐션 또한 양방향 순환신경망과 함께 사용되었다. 시간 t의 i번째 관절의 중간 결과값인 $\overline{x'^{i_t}}$는 학습 가능한 파라미터인 어텐션 가중치 α_t^i를 이용하여 식 (1)과 같이 계산된다. 어텐션이 적용된 $\overline{X}$가 다시 LSTM 셀과 완전 연결 레이어를 거쳐 최종 결과인 Y가 만들어진다.

$$\overline{x'^{i_t}}=(1-\alpha^{i_t)}\overline{y^{i_t}}+\alpha^{i_t}\overline{x^{i_t}} \tag{1}$$

한편, 네트워크를 학습시키기 위해 네트워크 생성 결과와 목표 모션의 차이, 생성된 모션의 부드러운 정도를 목적함수로 주어 그 차이를 최소화하도록 학습하였다. 목적함수는 식 (2)와 같이 정의되며 λ_s와 λ_r은 사용자 정의 변수이다.

$$\phi_p(\hat{Y},Y)+\lambda_s\Sigma_t\phi_s(\hat{Y}_t,Y_{t+1})+\lambda_r\|\theta\|_1 \tag{2}$$

여기에서 Y_t와 $\hat{Y}_t$는 각각 현재 프레임 t에서 목표한 포즈와 네트워크 결과로 생성된 포즈이다. $\phi(\cdot)$

는 L2 norm distance로, 예측된 결과와 목표 결과 사이의 오차(거리)를 나타내는 함수이며 θ는 학습 가능한 전체 네트워크 파라미터이다. 실험에서 사용자 정의 변수는 각각 $\lambda_s = 0.01$과 $\lambda_r = 0.001$로 설정하였다.

4. 실험 결과

본 연구에서의 주요 목표는 노이즈 성분이 포함된 다양한 모션을 효과적으로 정제하여 부드러운 모션을 만들어내는 것이다. 키넥트에서 얻어진 오염된 모션은 많은 오류를 가지고 있으나, 본 논문에서 제안한 방법을 적용하면 섬세하게 조정된 딥 러닝 프레임워크를 통해 세부적인 움직임이나 의미를 해치지 않으면서도 오류를 제거하는 것이 가능하다.

본 논문에서 제안한 방법은 춤추거나 뛰는 등 역동적인 동작에 섞인 노이즈에 대해서도 잘 동작하며, 이러한 결과를 도출하기 위해 입력 포즈들에 대해 선택적으로 가중치를 부여하는 어텐션 메커니즘 또한 제안된 방법의 중요한 요소이다. 실험을 통해 본 논문에서 제안한 방법과 다른 딥 러닝 방법 간의 차이를 정성적(그림 2), 정량적(그림 3)으로 비교하였으며, 제시한 방법에 대해 다양한 입력 특징 조합(그림 4)에 따른 성능 차이도 비교하였다.

실험에서는 성능 비교를 위해 CNN 모델[14]과 Temporal Convolutional Network(TCN) 모델[16]을 사용하였다. CNN 기반 모델은 기존에 이미 많이 연구된 Convolutional Autoencoder 모델을 활용하였다. CNN 모델은 Gaussian 노이즈, Gamma 노이즈 등 여러 기존 노이즈 모델뿐만 아니라 키넥트에서 모델링한 노이즈에서도 잘 동작하였지만, 주어진 입력이 빠르게 움직이는 동작인 경우 그 세세한 움직임들이 사라져 잘 동작하지 않는 경우가 있었다. TCN 기반 모델은 기본적으로 [16]에서 제안한 방법을 적용하였으며, 비교를 위해 Maxpooling과 Upsampling 레이어가 없으며 Residual block과 Dilation을 통해 넓은 범위의 시간 정보를 얻기 위한 TCN의 핵심적인 요소만을 차용하여 네트워크를 구성하였다. TCN은 Residual block과 Dilation의 영향으로 입력 모션의 디테일과 그 의미가 어느 정도는 유지되었으나 그림 2에서 보여주는 것처럼 노이즈가 강할 시 전체 결과값이 노이즈값에 압도되어 교정이 제대로 이루어지지 않는 것을 확인하였다.

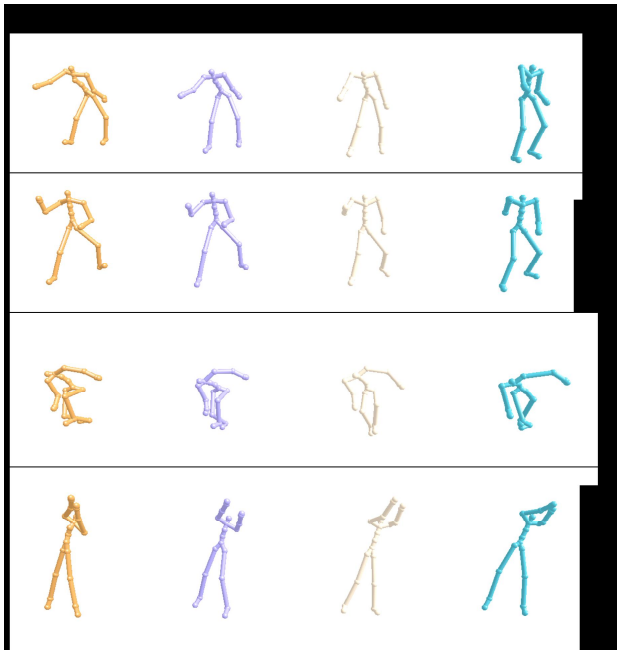

Fig. 2. Screenshots of the input and denoised poses from our method and other networks([14, 16]).
그림 2. 노이즈가 포함된 입력과 제시한 방법과 기존 네트워크들로 노이즈가 제거된 동작들

그림 3은 본 논문에서 제안한 Attention-based Bidirectional RNN (Attn-BiRNN) 모델과 CNN, TCN 모델의 다양한 모션(춤, 걷기, 달리기 동작)에 대한 실험 오차를 보여주는 그래프이다. 실험 결과 Attn-BiRNN이 모든 모션에 대해 실험 오차가 가장 작았으며, CNN과 TCN은 비슷한 성능을 보였다. 그림 4는 춤, 걷기, 달리기 모션에 대해 입력과 출력 모션 매핑을 위한 특징 조합에 따른 Attn-BiRNN의 성능 평가 결과이다. 실험 결과 위치에서 위치로의 매핑이 다른 입력 특징들보다 더 좋은 성능을 보였다.

III. 결론

본 논문에서는 광학식 마커 기반 모션 캡처 시스템이 가지고 있는 한계점에 주목하여, 이를 저가형 3D 깊이 센서인 키넥트 센서를 이용하여 해결하는 방법을 제안하였다. 비록 키넥트 센서로 생성된 움직임에는 노이즈가 들어가 있거나 사람답지 않은 움직임이 만들어지는 등, 깊이 센서 기반 모션 캡처 특유의 문제점이 있었으나, 이를 섬세하게 잘 설계된 딥 러닝 기반 프레임워크를 통하여 자연스

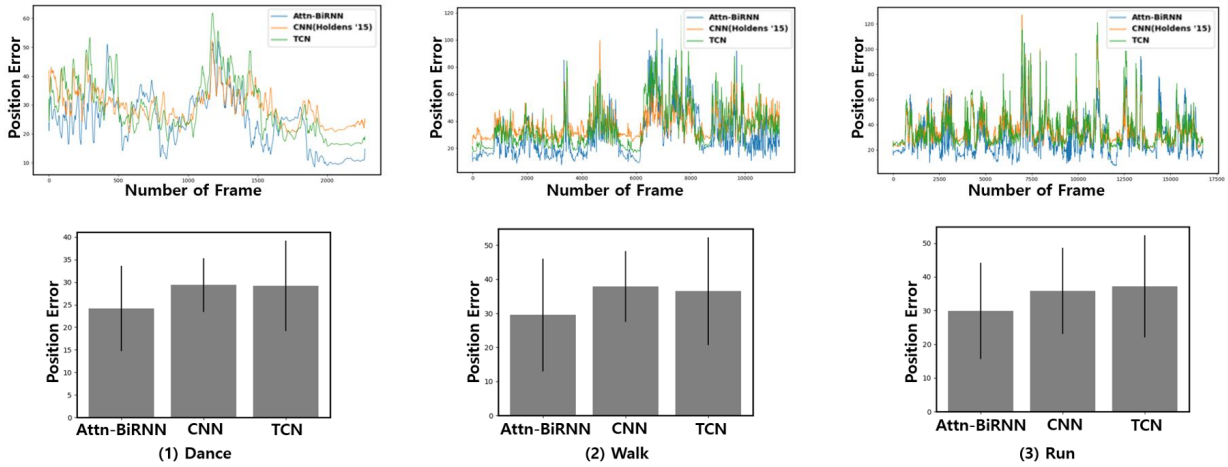

Fig. 3. Performance comparisons of the proposed method named Attn-BiRNN with other networks such as CNN [14] and TCN [16] for motion denoising.

그림 3. 본 논문에서 제안한 Attn-BiRNN 모델과 기존 모델인 CNN [14] 및 TCN [16]과의 모션 디노이징 성능 비교

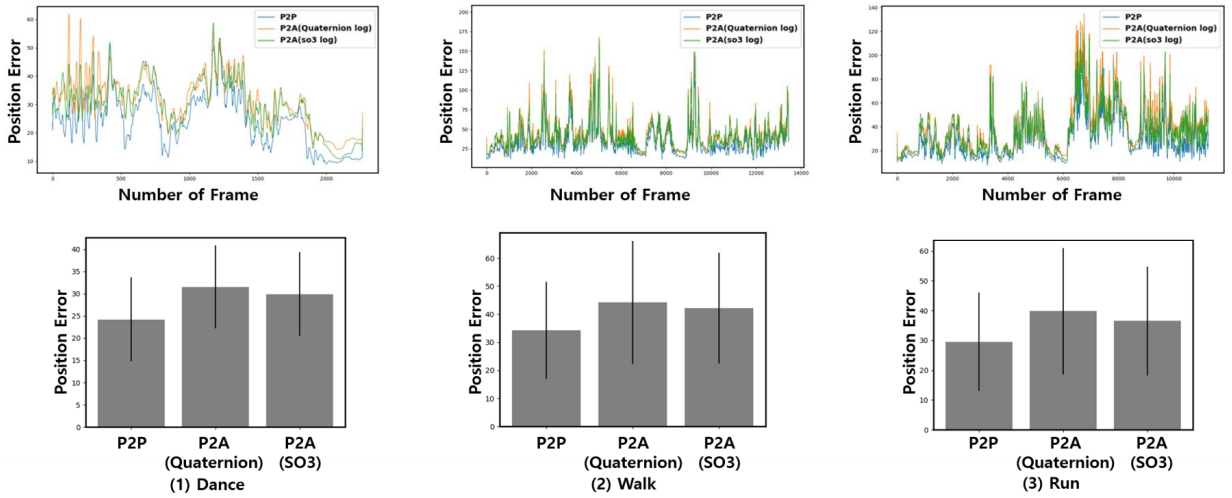

Fig. 4. Performance comparisons of feature combinations for mapping from input to output motion : position-to-position(P2P), position-to-angle(P2A) (quaternion), and position-to-angle(P2A) (SO3).

그림 4. 입력과 출력 모션의 매핑을 위한 특징 조합(위치-위치, 위치-각(Quaternion), 위치-각(SO3))에 따른 성능 비교

럽게 교정된 모션으로 정제할 수 있었다. 본 논문에서 제안한 어텐션 기반 양방향 순환신경망 모델은 기존의 모션 노이즈 제거 방법과는 다른 새로운 방법이며, 실험을 통해 제안한 모델이 세세한 움직임과 그 의미는 보존하면서도, 기존 광학식 모션 캡처 결과에 근사한 결과를 생성하는 것을 검증하였다.

References

[1] J. Chai and J. K. Hodgins, “Performance animation from low-dimensional control signals,” *ACM Trans. Graph,* Vol.24, no.3, pp.686-696, 2005. DOI: 10.1145/1073204.1073248

[2] Y. Lee, K. Wampler, G. Bernstein, J. Popović, and Z. Popović, “Motion fields for interactive character locomotion,” *ACM Trans. Graph,* 29, 6, Article 138, 2010.

[3] C. F. Rose III, P.-P. J. Sloan, and M. F. Cohen, “Artist directed inverse kinematics using radial basis function interpolation,” *Computer Graphics Forum,* Vol.20. No.3. pp.239-250, 2001. DOI: 10.1111/1467-8659.00516

[4] T. Mukai and S. Kuriyama, “Geostatistical

motion interpolation," *ACM Trans. Graph.,* vol.24, no.3, pp.1062-1070, 2005.
DOI: 10.1145/1073204.1073313
[5] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," In *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS'03),* pp.329-336, 2004.
[6] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popović, "Style-based inverse kinematics," *ACM Trans. Graph,* vol.23, no.3, pp.522-531, 2004.
DOI: 10.1145/1015706.1015755
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* pp.1097-1105, 2012.
DOI: 10.1145/3065386
[8] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," *arXiv preprint arXiv : 1202.2745,* 2012. DOI: 10.1109/CVPR.2012.6248110
[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp.1725-1732, 2014. DOI: 10.1109/CVPR.2014.223
[10] F. Nasse, C. Thurau, and G. A. Fink, "Face detection using gpu-based convolutional neural networks," *International Conference on Computer Analysis of Images and Patterns,* pp.83-90, 2009.
DOI: 10.1007/978-3-642-03767-2_10
[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.35, no.1, pp.221-231, 2013. DOI: 10.1109/TPAMI.2012.59
[12] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human Tracking Using Convolutional Neural Networks," *IEEE Transactions on Neural Networks*, vol.21, no.10, pp.1610-1623, 2010.
DOI: 10.1109/TNN.2010.2066286
[13] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.22, no.10, pp.1533-1545, 2014.
DOI: 10.1109/TASLP.2014.2339736
[14] D. Holden, J. Saito, T. Komura, and T. Joyce, "Learning motion manifolds with convolutional autoencoders," *SIGGRAPH Asia 2015 Technical Briefs,* Article 18, 2015.
[15] A. Pandey, and D. Wang, "TCNN: Temporal Convolutional Neural Network for Real-Time Speech Enhancement in The Time Domain," *2019 IEEE International Conference on Acoustics, Speech and Signal Processing,* pp.6875-6879, 2019. DOI: 10.1109/ICASSP.2019.8683634
[16] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp.156-165, 2017.
[17] L. Sun, K. Jia, D. Yeung, and B. E. Shi, "Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp4597-4605. 2015.
[18] J. Guo and H. Chao, "Building an end-to-end spatial-temporal convolutional network for video super-resolution," *Thirty-First AAAI Conference on Artificial Intelligence,* 2017.
[19] CMU, "Carnegie-Mellon Motion Capture Database," http://mocap.cs.cmu.edu/.
[20] S. R. Buss, "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods," *IEEE Journal of Robotics and Automation,* Vol.17, No.16, pp.1-19, 2004.
[21] S. U. Kim, H. Jang, and J. Kim, "Human Motion Denoising Using Attention-Based Bidirectional Recurrent Neural Network," In *SIGGRAPH Asia 2019 Posters (SA '19)*, Article 2, 2019.
DOI: 10.1145/3355056.3364577

BIOGRAPHY

SeongUk Kim (Member)

2014~ : BS student in Computer Science, Kangwon National University.

Hyeonseung Im (Member)

2006 : BS degree in Computer Science, Yonsei University.
2012 : PhD degree in Computer Science and Engineering, Pohang University of Science and Technology (POSTECH).
2012~2014 : Postdoc, Université Paris-Sud, France.
2014~2015 : Postdoc, Inria, France.
2015~ : Assistant Professor in Department of Computer Science, Kangwon National University.

Jongmin Kim (Member)

2006 : BS degree in Electrical Engineering and Computer Science, KAIST.
2014 : PhD degree in Electrical Engineering and Computer Science, Seoul National University.
2015~2016 : Research Professor, Hanyang University.
2016~2017 : Mocap Researcher, Weta Digital, New Zealand.
2017~ : Assistant Professor in Department of Computer Science, Kangwon National University.