Journal of KIIT. Vol. 19 No. 3 pp. 103-112, Mar. 31, pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.3.103

Auto-encoder 기반의 촬영 각도 무관 동작 임베딩 생성 모델을 활용한 영상 속 동작 유사도 측정

박 종 혁*

Motion Similarity Measurement using Auto-Encoder based Camera Angle-Agnostic Motion Embedding Generation Model

Jonghyuk Park*

요 약

사람의 동작을 분석하여 유사도를 측정하는 것은 다양한 문제에서 활용되고 있다. 하지만 사람의 동작은 카메라를 통해 촬영되기 때문에, 촬영 각도가 달라지면 2차원 관절 좌표의 시계열로 표현되는 동작도 변화하게 된다. 동작 유사도 측정 시 이러한 요소가 고려되지 않으면 기계는 촬영 각도가 달라진 동작을 같은 동작이라고 결정내리지 못하게 된다. 본 연구에서는 다른 각도에서 촬영된 두 동작의 유사도 측정 성능을 개선하기 위하여 촬영 각도 요소를 배제한 동작 임베딩을 생성하는 auto-encoder 기반의 모델을 활용하였다. 또한 생성된 동작 임베딩으로부터 유사도를 측정할 수 있는 방법을 제안하여 촬영 각도와 무관하게 동작 유사도 측정이 가능하도록 하였다. 실험은 같은 동작을 다양한 각도에서 투영한 동작들의 2차원 관절 좌표로 수행되었고, 촬영 각도를 고려하지 않은 모델로부터 생성된 동작 임베딩으로 측정한 유사도보다 정확한 결과를 보여주었다.

Abstract

Analyzing a person's motion to measure similarity is used in the various problem. However, since the motion is captured through a camera, the motion expressed as a time series of two-dimensional joint coordinates in a video also changes if the camera angle is varied. If this factor is not taken into account when measuring motion similarity, the machine will not be able to determine motions with different camera angles as the same motion. In this study, in order to improve the similarity measurement performance in this situation, an auto-encoder based model that generates motion embedding independent of the camera angle factor was utilized. Furthermore, through the proposed algorithm that can measure the similarity from the generated motion embedding, it is possible to measure the motion similarity regardless of the camera angle. The experiment was conducted with two-dimensional joint coordinates of motions projecting from various angles and showed more accurate results than the similarity measured by motion embedding generated from a model that did not consider the camera angle.

Keywords

motion similarity, human analysis, auto-encoder, metric learning, deep learning, computer vision

* 서울대학교 산업공학과 박사과정(교신저자)
- ORCID: https://orcid.org/0000-0003-4283-1155

· Received: Feb. 15, 2021, Revised: Mar. 19, 2021, Accepted: Mar. 22, 2021

· Corresponding Author: Jonghyuk Park
Dept. of Industrial Engineering, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826, Korea
Tel.: +82-2-880-7631, Email: chico2121@snu.ac.kr

Ⅰ. 서 론

2차원 관절 좌표의 시계열로 표현할 수 있는 사람의 동작을 분석하여 유사도를 측정하는 개념은 관련 인공지능 연구에서 널리 활용되고 있다. 구체적으로, 사람 동작 인식[1]-[3], 동작 수행 능력 평가[4]-[6], 이상 동작 탐지[7]-[9] 등 사람의 동작과 관련된 여러 연구 분야에서 유사도를 직간접적으로 측정하여 이용하고 있다.

동작 사이의 유사도를 측정하는 것이 이렇게 널리 활용됨에도 불구하고, 동작 유사도 측정만을 위한 연구는 많지 않다. 그 이유 중 하나는 영상을 촬영하는 카메라의 위치에 따라 같은 동작이라 할지라도 다른 관절 좌표의 시계열로 표현되기 때문에 정확한 유사도 측정에 한계가 있기 때문이다. 영상을 구성하는 이미지 프레임이 3차원이 아닌 2차원이기 때문에 이미지의 2차원 좌표로 표현되는 동작은 촬영하고 있는 카메라의 위치에 따라 다르게 나타나고, 이로부터 측정되는 동작 유사도는 부정확하게 된다.

따라서 본 연구에서는 다른 각도에서 촬영되는 동일한 동작의 유사도 측정 성능을 보정할 수 있는 유사도 측정 방법을 제안한다. 이를 위해 촬영 각도에 구애받지 않는 동작 임베딩을 생성하는 auto-encoder 기반 딥러닝 모델[10]을 활용하였다. Auto-encoder는 입력 값과 출력 값으로부터 차원이 축소된 은닉 임베딩을 학습하는 딥러닝 모델로, 입력 값의 특징을 표현하는 임베딩을 생성하여 다양한 목적을 위해 이를 활용한다[11]-[14].

본 연구에서 사용한 [10]의 모델은 2차원 관절 좌표의 시계열로부터 동작, 신체 구조, 카메라 촬영 각도를 나타내는 임베딩을 분리하여 학습한다. 이를 통해 학습된 동작 임베딩에는 사람의 신체 구조 및 카메라 촬영 각도에 해당하는 특징이 반영되지 않고, 오직 동작 요소에 해당하는 특징만 반영된다.

위 모델을 통해 동작 임베딩을 생성할 때, 세밀한 유사도 측정을 위해 전체 시계열을 sliding window 기법을 사용하여 패치(Patch) 단위로 나누어 각 패치의 동작 임베딩을 만든다. 그리고 두 시계열의 패치들에 대한 동작 임베딩을 DTW(Dynamic Time Warping)[15]를 사용하여 정렬시켜 동작 수행 속도에 차이가 있더라도 비슷한 특징을 가지고 있는 패치끼리 짝지어 유사도를 측정할 수 있도록 한다.

본 연구에서 제안한 동작 유사도 측정 방법의 효과를 평가하기 위해서 MADS(Martial Arts, Dancing and Sports) 데이터셋[16]을 활용하였다. HipHop (Dancing), Jazz(Dancing), Kata(Martial arts), Taichi (Martial arts), Sports의 다섯 가지 동작 분류로 구성된 MADS 데이터셋은 해당 동작에 대한 관절 좌표를 3차원 좌표로 제공한다. 이를 회전 변환을 통해 회전시켜 2차원에 투영시켜 같은 동작에 대해 다양한 각도에서 촬영한 것과 같은 관절 좌표 시계열을 생성하였다. 그리고 이 시계열들로부터 유사도를 측정하여 다양한 각도로 나타나는 같은 동작을 높은 유사도로 측정할 수 있는지 평가하였다. 유사도 측정 결과, 본 연구에서 제안한 방법은 다른 모델로부터 생성된 임베딩을 사용한 유사도 측정과 달리, 투영 각도에 구애받지 않고 같은 동작으로부터 생성된 시계열들을 높은 유사도로 측정하는 것을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구로 동작 유사도 측정 연구에 대해서 논하고, DTW 알고리즘에 대해 설명한다. 또한, 본 논문에서 사용한 auto-encoder 기반 동작 임베딩 생성 모델[10]을 소개한다. 3장에서는 학습된 동작 임베딩으로부터 동작 유사도를 측정하는 방법에 대해 논한다. 4장은 구체적인 실험 설정에 대해 언급하고, 실험 결과에 대해 기술한다. 마지막으로 5장에서는 본 연구의 의의와 향후 연구 방향을 제시하며 논문을 마무리한다.

Ⅱ. 관련 연구

2.1 동작 유사도 측정 관련 연구

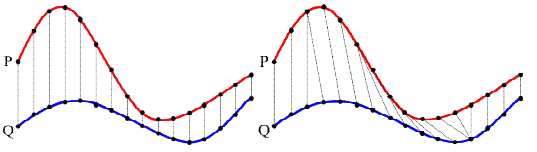
두 동작 사이의 유사도를 측정하는 것은 서론에서 언급한 것처럼, 다양한 분야의 연구에 활용되고 있지만 유사도 측정만을 위한 연구는 많지 않다. Kim[6]은 3차원 관절 좌표와 관절의 각도를 사용하여 두 춤 동작 사이의 유사도를 측정할 수 있는 알고리즘을 제안하였다. Shen[17]은 물체와 사람 또는

사람과 사람 사이에 상호 작용이 있는 상황에서 동작의 유사도를 측정할 수 있는 방법론을 제시하였다. [6]과 [17]은 각자의 문제 상황에서 동작 사이의 유사도 측정을 시도하였으나 모두 한계가 존재한다. [6]의 경우 3차원 관절 좌표를 사용하기 때문에 관절 좌표의 측정이 부정확할 수 있다는 한계가 존재한다. 또한 유사도를 측정하는 단위가 동작(여러 프레임)이 아니라 자세(한 프레임)이기 때문에 여러 프레임으로 표현되는 동작을 고려하기 어렵다. [17]은 상호 작용이 있는 상황에서만 측정 가능한 방법론이기 때문에 범용적으로 사용될 수 없다는 한계가 있다. 한편, Coskun[1]은 layer 정규화가 포함되어 있는 LSTM(Long Short Term Memory) 기반의 동작 임베딩 생성 모델을 제안하였다. 이때 저자는 최대 평균 불일치(Maximum mean discrepancy)[18]을 통해 유사도 학습의 negative mining에 필요한 계산량을 절약하였으며, 학습된 동작 임베딩은 동작 인식 및 검출 분야에서 우수한 성능 기록함을 실험을 통해 보였다. 본 논문에서는 [1]의 모델을 통해 생성된 동작 임베딩 또한 유사도 측정에 사용하여, 제안 방법의 동작 임베딩과의 비교 실험을 수행하였다.

2.2 DTW

DTW[15]는 두 시계열 사이 최적의 정렬 방법을 찾기 위해 제안된 알고리즘이다. 두 시계열을 각각 $P=(p_1, p_2, \dots, p_{T_P})$, $Q=(q_1, q_2, \dots, q_{T_Q})$ 라고 한다면, DTW는 동적 프로그래밍을 사용하여 비용 행렬 $C\in\mathbb{R}^{T_P\times T_Q}$을 계산한다. 여기서 $C_{ij}=d(p_i, q_j)$, $i\in[1:T_P], j\in[1:T_Q]$이고, $d(\cdot)$는 Euclidean 거리 같은 거리 측정 함수를 의미한다. 이때, 최적의 시계열 정렬은 C_{11}부터 $C_{T_PT_Q}$까지 비용의 합이 가장 적은 경로이다.

그림 1은 두 시계열을 같은 시간 지점끼리 정렬하는 방법인 Euclidean matching(그림 1(a))과, DTW를 사용하여 두 시계열을 정렬했을 경우(그림 1(b))를 비교한 그림이다. Euclidean matching을 통한 정렬과 달리 DTW를 사용한 정렬은 비슷한 패턴에 해당하는 지점들끼리 서로 짝지어진 것을 확인할 수 있다.



(a) (b)

그림 1. Euclidean matching과 dynamic time warping matching의 비교

Fig. 1. Comparison between Euclidean matching and dynamic time warping matching, (a) Euclidean matching, (b) Dynamic time warping matching

2.3 Auto-Encoder를 활용한 동작 임베딩 생성

Aberman[10]은 모델을 학습하기 위해서 Adobe Mixamo[19]에서 인공 데이터를 수집하여 사용하였다. 인공 데이터는 같은 동작을 수행하는 여러 애니메이션 캐릭터로 구성되어 있는데, 저자들은 이로부터 3차원 관절 좌표의 시계열을 추출하였다. 그리고 이를 회전 변환을 통해 이를 시계 방향 혹은 시계 반대 방향으로 회전시켜 2차원에 투영하였다. 결과적으로 같은 동작을 수행하는 여러 캐릭터, 촬영 각도의 2차원 관절 좌표를 생성하게 되었고, 이를 모델 학습에 활용하였다.

동작 임베딩에서 신체 구조 요소 및 촬영 각도 요소를 배제하기 위해서 Aberman[10]은 cross reconstruction 손실 함수를 통해 auto-encoder를 학습시켰다. m_1, m_2가 각각 다른 동작 요소를 의미한다고 하자. 마찬가지로 s_1, s_2는 각각 다른 신체 구조 요소, v_1, v_2는 각각 다른 촬영 각도 요소를 의미한다고 하자. 그러면 m_1, s_1, v_1 요소로 구성된 2차원 관절 좌표는 $X_{m_1,s_1,v_1}\in\mathbb{R}^{n_J\times 2\times n_T}$으로 정의할 수 있다. 여기서 n_J는 관절의 개수, n_T는 시계열의 길이이다. 이 2차원 관절 좌표를 동작 encoder E_M, 신체 구조 encoder E_S, 촬영 각도 encoder E_V의 입력 값으로 하여 각각에 해당하는 임베딩 $E_M(X_{m_1,s_1,v_1})=z_{m_1}$, $E_S(X_{m_1,s_1,v_1})=z_{s_1}$, $E_V(X_{m_1,s_1,v_1})=z_{v_1}$을 생성한다. 같은 방법으로, m_2, s_2, v_2 요소로 구성된 2차원 관절 좌표는 X_{m_2,s_2,v_2}로부터 임베딩 z_{m_2}, z_{s_2}, z_{v_2}를 생성한다. 여기서 z_{m_1}과 z_{s_2}, z_{v_2}를 결합하고, 이를 decoder D에 입력하여 m_1, s_2, v_2 요소에 대한 추정 2차원 관절 좌표인 $\hat{X}_{m_1,s_2,v_2}$를 만

들어내어 ground truth 2차원 관절 좌표인 X_{m_1,s_2,v_2}와 비교, cross reconstruction 손실을 계산한다. 같은 방법으로, m_2, s_1, v_1 요소에 대한 추정 2차원 관절 좌표인 $\hat{X}_{m_2,s_1,v_1}$를 만들어내어 ground truth 2차원 관절 좌표인 X_{m_2,s_1,v_1}와 비교, cross reconstruction 손실을 계산한다. 그리고 두 cross reconstruction 손실을 역전파하여 모든 encoder와 decoder의 파라미터를 갱신하는 방식으로 모델을 학습시킨다. 위와 같은 과정이 그림 2에 나타나 있다. 이러한 과정을 거쳐서 모델을 학습하면 각 요소에 해당하는 특징만 임베딩에 포함하게 되기 때문에 궁극적으로 동작 임베딩에 신체 구조 및 촬영 각도 요소가 배제될 수 있다. 이러한 학습 방식은 각 요소들의 여러 조합에 대한 ground truth 관절 좌표를 모두 인공적으로 생성할 수 있었기 때문에 가능하였다.

III. 제안 방법

앞서 언급한 것처럼, 동작 임베딩에서 영상의 촬영 각도 요소를 배제하기 위해 본 논문에서는 [10]의 auto-encoder 모델을 사용하였다. 본 장에서는 [10]의 auto-encoder 모델로부터 생성된 동작 임베딩을 통해 동작 유사도를 계산하는 방법을 설명한다.

동작 유사도를 측정하기 위한 과정은 크게 세 단계로 나눌 수 있다. 첫 번째는 2차원 관절 좌표의 시계열을 패치 단위로 나누는 단계이고, 두 번째는 생성된 패치를 동작 인코더의 입력 값으로 하여 동작 임베딩을 생성하는 단계이며, 마지막 단계는 생성된 동작 임베딩으로부터 유사도를 측정하는 단계이다.

먼저 2차원 관절 좌표의 시계열을 작은 시간 단위로 쪼개어 패치의 집합을 생성한다. 이때 sliding window 기법을 사용하여 앞뒤로 생성되는 패치에 겹치는 시간 지점이 존재하도록 시계열을 나눈다. 시간 축으로의 패치 길이를 w, sliding 길이를 l 이라고 하자. 그러면 전체 시간 길이가 T_X인 2차원 관절 좌표 시계열 $X \in \mathbb{R}^{n_J \times 2 \times T_X}$의 x번째 패치 $X^x \in \mathbb{R}^{n_J \times 2 \times w}$는 $l(x-1)$ 프레임에서 $l(x-1)+w$ 프레임까지의 시계열이다. 이런 방식으로 패치를 생성하면 전체 동작을 작은 시간 단위의 동작으로 비교할 수 있기 때문에 세밀한 유사도 측정이 가능해진다.

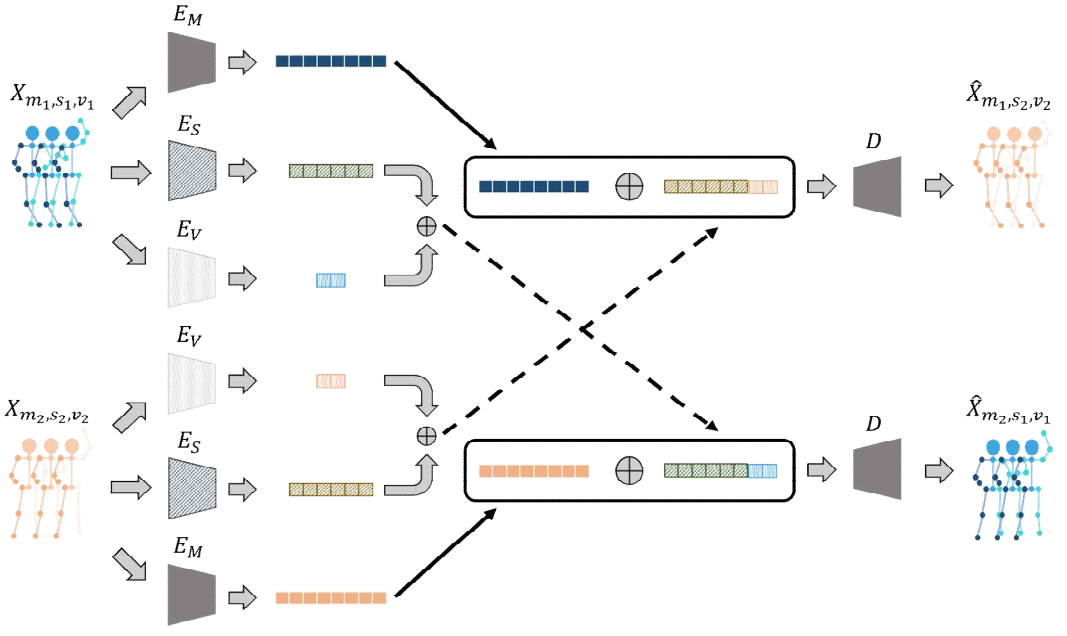


그림 2. 동작 임베딩에서 신체 구조와 카메라 촬영 각도 요소를 배제하기 위한 [10]의 auto-encoder 구조
Fig. 2. Auto-encoder structure of [10] to exclude body structure and camera angle factors from motion embedding

다음으로는 생성된 각각의 시계열 패치를 동작 인코더의 입력 값으로 하여 총 n_X개의 동작 임베딩으로 구성된 집합 M_X을 만든다. 이를 수식으로 표현하면 다음과 같다.

$$M_X = \{E_M(X^1), E_M(X^2), \cdots, E_M(X^{n_X})\} \quad (1)$$

$$where\ n_X = \left\lfloor \frac{T_X - w + l}{l} \right\rfloor \quad (2)$$

같은 방식으로 전체 길이가 T_Y인 2차원 관절 좌표 시계열 $Y \in \mathbb{R}^{n_J \times 2 \times T_Y}$의 각 패치에 대한 동작 임베딩 집합 M_Y를 생성한다. 그리고 DTW를 사용하여 두 동작의 각 패치에 대한 동작 임베딩을 서로 짝지어 정렬시켜 비용의 합이 가장 적은 경로 $path$를 구한다.

$$path = \{(1,1), \cdots, (x,y), \cdots, (n_X, n_Y)\} \quad (3)$$

같은 동작이라 할지라도 수행하는 사람이 달라지면 그 동작의 수행 속도나 발생 시점이 달라질 수 있기때문에 비슷한 움직임을 가진 동작 패치끼리 비교할 수 있도록 DTW를 활용한다.

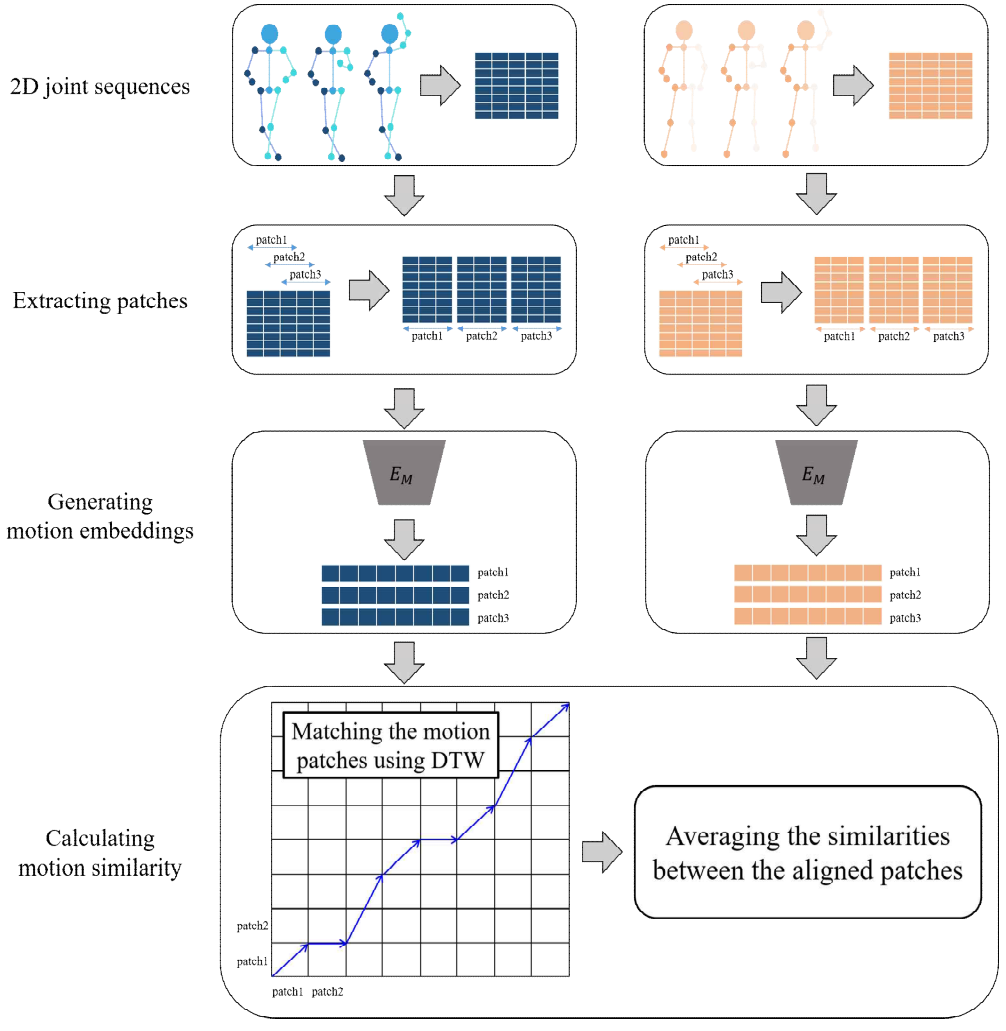


그림 3. 동작 임베딩을 활용한 동작 유사도 측정
Fig. 3. Measurement of motion similarity using motion embedding

마지막으로 짝지어진 두 동작의 패치 사이의 유사한 정도를 코사인 유사도를 통해 계산한다. 그리고 DTW를 통해 정렬된 모든 패치 쌍에 대해 계산된 유사도를 평균내어 두 시계열 사이의 최종 동작 유사도를 결정한다. $path$ 의 짝지어진 동작 임베딩 쌍 중 하나를 p 라고 하고, 이에 대한 코사인 유사도를 $CosineSim(p)$ 이라고 하자. $path$ 에 짝지어진 동작 임베딩 쌍이 총 n_p 개 존재한다고 하면, 최종 동작 유사도 sim 는 다음 수식과 같이 계산된다.

$$sim = \frac{\sum_{p \in path} CosineSim(p)}{n_p} \quad (4)$$

이러한 과정은 그림 3에 나타나 있다.

IV. 실 험

4.1 평가 데이터셋

본 연구에서 제안한 동작 유사도 측정 방법의 성능을 검증하기 위해서 MADS 데이터셋[16]을 활용하여 평가 데이터셋을 구축하였다. MADS 데이터셋은 무도에 해당하는 Kata, Taichi 클래스와 춤에 해당하는 HipHop, Jazz, 스포츠에 해당하는 Sports 클래스로 구분된다. 각 클래스는 6개의 개별 동작으로 구성되어 있으며, 총 30개의 동작 중 부정확한 좌표를 포함하고 있어 사용하지 않을 것을 권고받고 있는 7개의 동작을 평가 데이터셋에서 제외하였다. 따라서 평가 데이터셋은 6개의 HipHop 및 Jazz 동작, 5개의 Kata 동작, 2개의 Taichi 동작, 4개의 Sports 동작으로 구성되었다. 각 클래스의 동작 예시는 그림 4에 나타나 있다.

MADS 데이터셋은 해당 동작의 3차원 관절 좌표의 시계열을 함께 제공하기 때문에 이를 활용하여 같은 동작의 2차원 회전 시계열을 생성하였다. 구체적으로, 3차원 좌표의 회전 변환을 통해 시계 방향으로 45°, 시계 반대 방향으로 45° 회전하여 2차원에 투영한 관절 좌표 시계열을 생성하여 같은 동작으로부터 생성된 시계열을 높은 유사도로 측정할 수 있는지 평가하였다. 그림 5는 같은 동작의 2차원 회전 시계열 중 한 시점을 표현한 그림으로, 시계 방향으로 45° 회전한 관절 좌표(그림 5(a)), 회전하지 않은 관절 좌표(그림 5(b)), 시계 반대 방향으로 45° 회전한 관절 좌표(그림 5(c))가 나타나 있다.

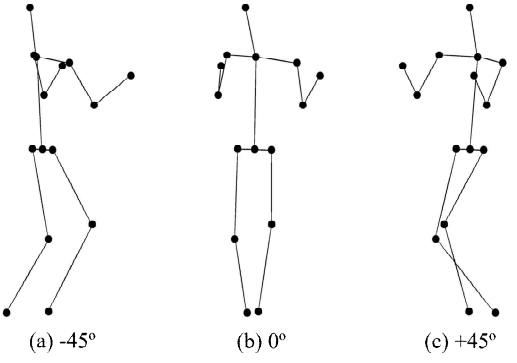


그림 5. 회전한 관절 좌표 시계열의 예시
Fig. 5. Examples of rotated joint coordinates

그림 4. MADS 데이터셋[16]의 동작 예시
Fig. 4. Motion examples of MADS dataset[16]

4.2 실험 상세

제안 모델의 동작 임베딩 생성을 위한 auto-encoder 모델의 파라미터는 [10]의 공식 홈페이지 [20]에서 제공하는 모델 checkpoint를 통해 획득하였다. [10]과 같은 방법으로, 관절은 총 15개(코, 목, 오른쪽 어깨, 오른쪽 팔꿈치, 오른쪽 손목, 왼쪽 어깨, 왼쪽 팔꿈치, 왼쪽 손목, 오른쪽 엉덩이, 오른쪽 무릎, 오른쪽 발목, 가운데 엉덩이, 왼쪽 엉덩이, 왼쪽 무릎, 왼쪽 발목)를 사용하였다. 그 중 가운데 엉덩이 관절을 기준으로 상대 좌표화하여 모델의 입력 값으로 사용하였다.

한편, 생성된 동작 임베딩에서 촬영 각도가 잘 배제되었는지 평가하기 위하여, [1]의 LSTM 기반 모델에서 생성된 동작 임베딩 또한 실험의 비교군으로 사용되었다. 이 모델은 우리가 아는 한, 여러 관련 연구 중 2차원 관절 좌표로부터 유사도 학습을 통해 동작 임베딩을 생성하는 유일한 모델이다. 앞에서 언급한 것처럼, 이 모델로부터 생성된 동작 임베딩은 동작 인식 및 검출에서 우수한 성능을 보이는 것이 확인되었기 때문에, 동작 임베딩에 충분한 동작 관련 정보가 포함되었을 것이라 판단할 수 있다. 따라서 [1]의 모델을 비교군으로 채택하였으며, 공정한 비교를 위해 [10]의 모델과 똑같이 Adobe Mixamo 데이터셋으로 모델을 학습시켰다.

4.3 성능 평가

평가 데이터셋에 대한 실험 결과가 표 1과 표 2에 각각 나타나 있다. 회전시키지 않은 시계열과 시계 방향으로 회전시킨 시계열(-45°), 회전시키지 않은 시계열과 시계 반대 방향으로 회전시킨 시계열(45°)의 유사도를 측정하여 유사도를 1에 가깝게, 즉 같은 동작에 가깝다고 판정할 수 있는지 확인하였다. 시간 축으로의 패치 길이는 64 프레임, sliding 길이는 2 프레임을 사용하여 패치를 생성하고 유사도를 측정하였다.

Auto-encoder에서 생성된 동작 임베딩을 사용하여 측정된 유사도는 LSTM 기반의 모델에서 생성된 임베딩을 통해 측정된 유사도보다 전반적으로 1에 더 가까운 수치를 보여주었다. 이는 auto-encoder의 동작 임베딩이 촬영 각도 요소를 잘 분리하여, 동작 정보 위주로 임베딩을 형성하였음을 의미한다.

한편, HipHop의 동작 유사도가 다른 동작보다 전반적으로 더 낮은 값으로 측정되는 것을 확인할 수 있는데, 이는 HipHop의 동작이 역동적인 움직임으로 구성되어 있는 것에서 기인한다. 일반적이지 않은 동작의 경우 학습 데이터셋인 Adobe Mixamo 인공 데이터셋에 포함되어 있지 않을 가능성이 크다. 이런 경우 모델이 해당 동작의 정보를 온전히 포함한 동작 임베딩을 생성하기 힘들다.

4.4 패치의 하이퍼 파라미터 분석

패치 생성 시, 두 하이퍼 파라미터인 시간 축으로의 패치 길이와 sliding 길이에 따라, 나누어지는 패치의 수가 달라진다.

표 1. 다른 각도로 투영된 동일 HipHop 동작의 유사도 측정 결과
Table 1. Similarity measurement result of the same HipHop motions projected from different angles

Motions / Models	HipHop1		HipHop2		HipHop3		HipHop4		HipHop5		HipHop6	
	-45°	45°	-45°	45°	-45°	45°	-45°	45°	-45°	45°	-45°	45°
Auto-encoder[10]	0.930	0.875	0.921	0.878	0.917	0.844	0.903	0.836	0.920	0.839	0.920	0.850
LSTM[1]	0.861	0.800	0.849	0.823	0.869	0.849	0.889	0.831	0.863	0.811	0.867	0.794

표 2. 다른 각도로 투영된 동일 Jazz 동작의 유사도 측정 결과
Table 2. Similarity measurement result of the same Jazz motions projected from different angles

Motions / Models	Jazz1		Jazz2		Jazz3		Jazz4		Jazz5		Jazz6	
	-45°	45°	-45°	45°	-45°	45°	-45°	45°	-45°	45°	-45°	45°
Auto-encoder[10]	0.922	0.882	0.922	0.874	0.912	0.869	0.927	0.878	0.912	0.858	0.929	0.896
LSTM[1]	0.858	0.843	0.858	0.809	0.856	0.843	0.866	0.821	0.859	0.848	0.878	0.829

표 3. 다른 각도로 투영된 동일 Kata 동작의 유사도 측정 결과
Table 3. Similarity measurement result of the same Kata motions projected from different angles

Motions / Models	Kata1		Kata2		Kata3		Kata4		Kata5	
	-45°	45°	-45°	45°	-45°	45°	-45°	45°	-45°	45°
Auto-encoder[10]	0.929	0.941	0.951	0.953	0.937	0.951	0.951	0.963	0.943	0.952
LSTM[1]	0.811	0.767	0.641	0.699	0.800	0.792	0.828	0.755	0.797	0.820

표 4. 다른 각도로 투영된 동일 Taichi 동작 및 Sports 동작의 유사도 측정 결과
Table 4. Similarity measurement result of the same Taichi and Sports motions projected from different angles

Motions / Models	Taichi1		Taichi2		Sports1		Sports2		Sports3		Sports4	
	-45°	45°	-45°	45°	-45°	45°	-45°	45°	-45°	45°	-45°	45°
Auto-encoder[10]	0.887	0.947	0.917	0.962	0.919	0.933	0.919	0.931	0.921	0.919	0.919	0.916
LSTM[1]	0.847	0.879	0.860	0.847	0.822	0.831	0.852	0.837	0.822	0.814	0.767	0.761

또한, 패치 길이가 커질수록 더 긴 시간에 대한 2차원 관절 좌표 시계열로부터 동작 임베딩을 생성하는 것이기 때문에 동작에 대한 포괄적인 정보가 임베딩에 포함된다. 반대로 패치 길이가 작아지면 해당 구간의 움직임에 대한 특징이 잘 표현된 임베딩이 생성된다.

그림 6은 두 하이퍼 파라미터 설정에 따른 4.3절의 유사도 측정 결과를 그래프로 도식화한 것으로, 각 클래스의 개별 동작에 대한 유사도를 평균내어 그래프에 표현하였다. 동작 임베딩 생성 모델로는 auto-encoder 기반 생성 모델을 사용하였다. 그래프를 통해, 패치 길이가 커질수록 다른 각도에서 투영된 동일 동작들에 대한 유사도가 더 커지는 것을 확인할 수 있다. 패치 길이가 커지면 동작 임베딩이 더 긴 구간에 대한 정보를 포함하고 있어 이상치의 영향이 작아진다. 이로부터 더 정확한 유사도 측정이 가능해진 것으로 해석된다. Sliding 길이의 경우, 전반적으로 sliding 길이가 작을수록 더 높은 유사도를 보여주었다. Sliding 길이가 커 상대적으로 적은 수의 패치로부터 임베딩을 생성할 경우, DTW 정렬 임베딩 쌍의 동작 유사도가 작게 측정되면 이를 보정할 수 있는 방법이 없게 된다.

반면, sliding 길이가 작을 경우 많은 수의 패치로부터 임베딩을 생성하기 때문에 DTW에 의해 더 거리가 가까운 임베딩끼리 유사도를 측정할 수 있게 되어 결과적으로 높은 유사도를 얻을 수 있게 된다.

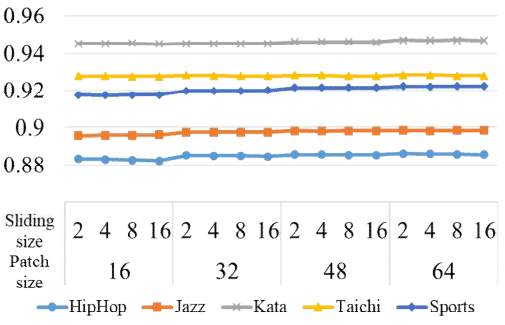


그림 6. 패치 길이와 sliding 길이에 따른 동작 유사도 측정 결과
Fig. 6. Measurement result of motion similarity according to patch size and sliding size

V. 결론 및 향후 과제

본 연구에서는 auto-encoder 기반의 동작 임베딩 생성 모델을 활용하여 촬영 각도에 영향을 덜 받는 동작 유사도 측정 방법을 제안하였다. 사용한 auto-encoder 모델은 cross reconstruction 손실을 통해 모델이 학습되기 때문에 촬영 각도 요소로부터 분리된 동작 임베딩을 생성할 수 있었다. 이를 활용한 제안 동작 유사도 측정 방법은 DTW를 통해 패치 단위로 나뉘어진 두 동작의 임베딩들을 정렬시킴으로써 더 정확한 유사도 측정이 가능하게 하였다. 제안 방법의 성능을 검증하기 위하여 MADS 데이터셋의 동작을 다양한 각도에서 2차원에 투영한 시계열들에 대해 유사도를 측정하는 실험을 수행하였다.

실험 결과, 제안 방법이 다른 임베딩을 사용한 유사도 측정 결과 대비 0.088 높은 동작 유사도를 계산함으로써 촬영 각도로부터 강건한 동작 유사도 측정이 가능함을 보여주었다.

본 연구의 향후 과제는 15개의 관절 좌표보다 더 많은 관절을 사용하여 세밀한 유사도를 측정하는 것이다. 예를 들어 춤의 경우, 손가락의 움직임 또한 중요한 요소 중 하나이기 때문에 손가락의 움직임을 유사도 측정에 반영할 수 있다면 활용 가치가 더욱 커질 것이다. 골프 스윙의 경우 기본적인 동작 자체는 모두 비슷하기 때문에 더 많은 관절을 사용해야 정확한 유사도 측정이 가능해질 것이다.

References

[1] H. Coskun, D. J. Tan, S, Conjeti, N. Navab, and F. Tombari, "Human Motion Analysis with Deep Metric Learning", The 15th European Conference on Computer Vision (ECCV), Munich, Germany, pp. 667-683, Sep. 2018.

[2] Z. Gao, L. Guo, T. Ren, A. Liu, Z. Cheng, and S. Chen, "Pairwise Two-Stream ConvNets for Cross-Domain Action Recognition With Small Data", IEEE Transactions on Neural Networks and Learning Systems, pp. 1-15, Nov. 2020. https://doi.org/10.1109/TNNLS.2020.3041018

[3] Z. Gao, L. Guo, W. Guan, A. Liu, T. Ren, and S. Chen, "A Pairwise Attentive Adversarial Spatiotemporal Network for Cross-Domain Few-Shot Action Recognition-R2", IEEE Transactions on Image Processing, Vol. 30, pp. 767-782, Nov. 2020. https://doi.org/10.1109/TIP.2020.3038372

[4] F. Chin-Shyurng, S. E. Lee, and M. L. Wu, "Real-Time Musical Conducting Gesture Recognition based on a Dynamic Time Warping Classifier using a Single-Depth Camera", Applied Sciences, Vol. 9, No. 3, pp. 528, Feb. 2019. https://doi.org/10.3390/app9030528

[5] A. Elaoud, W. Barhoumi, E. Zagrouba, and B. Agrebi, "Skeleton-Based Comparison of Throwing Motion for Handball Players", Journal of Ambient Intelligence and Humanized Computing, Vol. 11, No. 1, pp. 419-431, May 2019.

[6] Y. Kim and D. Kim, "Real-Time Dance Evaluation by Markerless Human Pose Estimation", Multimedia Tools and Applications, Vol. 77, No. 23, pp. 31199-31220, Dec. 2018.

[7] K. W. Cheng, Y. T. Chen, and W. H. Fang, "Video Anomaly Detection and Localization using Hierarchical Feature Representation and Gaussian Process Regression", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 2909-2917, Jun. 2015.

[8] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 11996-12004, Jun. 2019.

[9] X. Zhang, S. Yang, X. Zhang,W. Zhang, and J. Zhang, "Anomaly Detection and Localization in Crowded Scenes by Motion-Field Shape Description and Similarity-based Statistical Learning", arXiv:1805.10620, May 2018.

[10] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or, "Learning Character-Agnostic Motion for Motion Retargeting in 2d", ACM Transactions on Graphics (TOG), Vol. 38, No. 4, Art. No. 75, Jul. 2019.

[11] V. H. Trong, G. H. Yu, D. T. Vu, J. H. Lee, N. H. Toan, and J. Y. Kim, "A Study on Weeds Retrieval based on Deep Neural Network Classification Model", Journal of Korean Institute of Information Technology, Vol. 18, No. 8, pp. 19-30, Aug. 2020. https://doi.org/10.14801/jkiit.2020.18.8.19

[12] C. Kim, "Scheme of a Classic Control-Based Program Model with Non-Symmetric Deep Auto-Encoder of Actor-Critic", Journal of Korean Institute of Information Technology, Vol. 18, No. 7, pp. 15-20, Jul. 2020. https://doi.org/10.14801/

jkiit.2020.18.7.15

[13] J. N. Lee and K. C. Kwak, "Classification of Horse Gaits Based on Wavelet Packet and Auto-Encoder for Self-Coaching System", Journal of Korean Institute of Information Technology, Vol. 15, No. 5, pp. 1-9, May 2017. https://doi.org/10.14801/jkiit.2017.15.5.1

[14] W. J. Choi, "Multi-MEMS-Sensor Based Machinery Fault Diagnosis System of Mechanical Device", Journal of Korean Institute of Information Technology, Vol. 18, No. 6, pp. 17-23, Jun. 2020. https://doi.org/10.14801/jkiit.2020.18.6.17

[15] R. Bellman and R. Kalaba, "On Adaptive Control Processes", IRE Transactions on Automatic Control, Vol. 4, No. 2, pp. 1-9, Nov. 1959. https://doi.org/10.1109/TAC.1959.1104847

[16] W. Zhang, Z. Liu, L. Zhou, H. Leung, and A. B. Chan, "Martial Arts, Dancing and Sports Dataset: A Challenging Stereo and Multi-view Dataset for 3d Human Pose Estimation", Image and Vision Computing, Vol. 61, pp. 22-39, May 2017. https://doi.org/10.1016/j.imavis.2017.02.002

[17] Y. Shen, L. Yang, E. S. L. Ho, and H. P. H. Shum, "Interaction-based Human Activity Comparison", IEEE Transactions on Visualization and Computer Graphics, Vol. 26, No. 8, pp. 2620-2633, Aug. 2020. https://doi.org/10.1109/TVCG.2019.2893247

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A Kernel Two-Sample Test", The Journal of Machine Learning Research (JMLR), Vol. 13, No. 25, pp. 723-773, Mar. 2012.

[19] Adobe Mixamo. Adobe Systems Inc., 2021. [Online]. Available: https://www.mixamo.com. [accessed: Feb. 13, 2021]

[20] Learning Character-Agnostic Motion for Motion Retargeting in 2D. R. Wu and K. Aberman. 2019. [Online]. Available: https://github.com/ChrisWu1997/2D-Motion-Retargeting. [accessed: Feb. 13, 2021]

저자소개

박 종 혁 (Jonghyuk Park)

2015년 2월 : 서울대학교 산업공학과(공학사)
2015년 1월 ~ 2016년 2월 : ㈜삼성전자 메모리 사업부
2016년 3월 ~ 현재 : 서울대학교 산업공학과(석박사통합과정) 재학 중

관심분야 : 컴퓨터 비전, 딥러닝, 머신 러닝 응용