논문 22-47-04-01 The Journal of Korean Institute of Communications and Information Sciences '22-04 Vol.47 No.04
https://doi.org/10.7840/kics.2022.47.4.575

특징점 검출 기반 정지 영상에서 판소리 발림의 식별

오 문 흠•, 이 혜 정*, 이 준 환°

Recognition of Pansori Motion in Still Images Based on Keypoint Detection

Wenqin Wu•, Hyejeong Lee*, Joonwhoan Lee°

요 약

MIR(Music Information Retrieval) 분야에서는 인공지능을 이용하는 음악 연구가 활발하게 진행되고 있다. 본 논문에서는 판소리의 주요 구성요소인 발림 동작 인식을 목적으로 동영상으로부터 이를 자동으로 분별하는 방법을 제안한다. 제안된 방법에서는 동영상을 구성하는 매 프레임에서 소리꾼 객체를 검출 시도하고 이를 포함하는 정지 영상에서 객체 영역의 발림 동작을 분별한다. 객체 검출 방법으로는 일정한 객체 크기에 대해 성능이 우수하고 검출객체 수가 작은 속도가 빠른 트랜스포머를 이용하였으며, 검출된 소리꾼의 관절과 부채 위치 등 특징점을 포함하는 골격 정보를 이용하여 강인하게 발림 동작을 분류한다. 부족한 판소리 발림 동작 영상 데이터 때문에 MS COCO 데이터를 활용하여 전이학습과 미세조정을 수행하였으며, 대 부류로 정리된 7개 클래스로 분류한 발림 동작의 분류에서 84.8%의 정확도를 달성하였다.

Key Words : Pansori, Balim, Artificial intelligence, Keypoint Detection, Skeletal Information, Transformer

ABSTRACT

Artificial intelligence research on music area in MIR(Music Information Retrieval) is being conducted very actively. This paper aims to recognize the Balim motions, an important component of Pansori, and proposes a method to automatically classify them from Pansori video. In the proposed method, the region of a singer is at first detected from every video frame, then the Balim motion is classified based on still image analysis in the detected region. For detecting singer object we use the transformer-based object detection, which has an excellent performance with high speed for the uniform size of small number of objects, and then try to classify Balim motion with skeletal information of key points including the singer's joint and fan position. Due to insufficient Pansori Balim motion data, transfer learning using MS COCO data and fine-tuning are performed. The accuracy of 84.8% has been achieved in the classification of Balim motion largely categorized into 7 classes.

※ 본 연구는 한국연구재단 중견연구과제(NRF-2021R1A2C2006895)에 의해 지원되었음.
◆ First Author : Jeonbuk National University Department of Computer Science and Engineering, ryanwu200@gmail.com, 학생회원
° Corresponding Author : Jeonbuk National University Department of Computer Science and Engineering, chlee@jbnu.ac.kr, 종신회원
* Jeonbuk National University Department of Korean Music, jina2747@naver.com
논문번호 : 202201-015-A-RE, Received January 28, 2022; Revised February 11, 2022; Accepted February 17, 2022

Ⅰ. 서 론

인류에게 노래는 오래전부터 존재해 왔던 하나의 문화 현상이며 감정표현의 수단이었다. 따라서 노래에 수반되는 감정을 표현하기 위해 가락이나 노래 내용에 따라 동작을 취하기도 한다. 판소리의 발림이란 소리꾼이 아니리나 소리를 부르며 취하는 동작으로 판소리 대목에 따라 또는 소리꾼에 따라 다양한 양상으로 나타난다[2-3]. 본 논문에서 판소리 동영상으로부터 이러한 발림 동작을 자동으로 찾아내는 방법은 판소리 소리꾼 발림 동작을 빠르게 분석하여 판소리 연구에 활용할 수 있으며 판소리 소리꾼의 동작 교육에도 활용할 수 있다.

소리꾼 발림 동작 인식방법에는 동영상 기반의 행동 인식 방법[18-22]과 동영상을 구성하는 매 프레임에서 소리꾼을 찾아내고 이 소리꾼 영상을 분류하는 방법을 활용할 수 있다. 그러나 일반적으로 발림은 소리꾼이 판소리 내용이나 가락 등에 따라 정형화된 부분도 있지만, 청중이나 전체적인 판의 상태에 따라 즉흥적으로 이루어지기도 한다. 따라서 동영상의 행동 인식의 경우 다양한 범주를 고려해야 하므로 어려운 문제이다.

이러한 이유로 본 논문에서는 비교적 단순한 방법인 소리꾼이 존재하는 프레임을 찾고 각 정지된 프레임을 기반으로 소리꾼의 동작을 주요 동작 7개의 범주[2-3]로 한정하고 자동으로 분류하는 방식으로 분석한다. 즉 동영상 프레임을 구성하는 매 프레임에서 소리꾼이 들어있는 프레임과 소리꾼 영역을 트랜스포머 기반의 객체 검출을 이용하여 자동으로 찾고, 소리꾼 영역의 특징 추출을 통하여 7개의 주요 동작으로 분류한다. 여기서 추출한 특징은 의상의 색깔, 무늬나 배경 등 컨텍스트 정보가 아닌 특징점으로 표현하는 골격 정보이다. 보고된 바에 의하면 골격 정보를 찾는 특징점 기반의 동작 분류 방법은 다른 컨텍스트 찾는 분류 방법에 비해 강건하고 분류 성능이 우수하다[4-5]. 본 논문에서 제안된 방법은 안정적인 소리꾼 프레임과 소리꾼 영역 검출을 제공하며, 본 논문의 실험 결과 7개의 부류로 분류한 발림 동작은 정확도가 84.8%에 달하였다.

본 연구는 인공지능 기법으로 판소리 영상으로부터 발림 동작을 자동으로 분류하는 최초의 논문으로 의미가 있으며 향후 국악 연구에서 빠른 속도의 판소리 소리꾼 발림 동작 분석과 판소리 교육 등에 활용될 것으로 기대된다.

Ⅱ. 제안된 방법

제안된 방법의 파이프라인은 소리꾼 검출과 발림 동작 분류로 구성된다. 우선 소리꾼만 아니라 고수 등 다른 사람도 포함한 판소리 공연 전경 영상에서 소리꾼을 검출 후 그림 1에 보이는 소리꾼의 관절과 부채 위치 등 골격 정보를 의미하는 특징점들을 검출하여 발림 동작 분류한다. 이러한 소리꾼 검출과 특징점 검출을 통한 발림 동작 인식 과정은 동영상를 구성하는 매 프레임에 반복 적용해서 자동으로 분석결과를 제공한다.

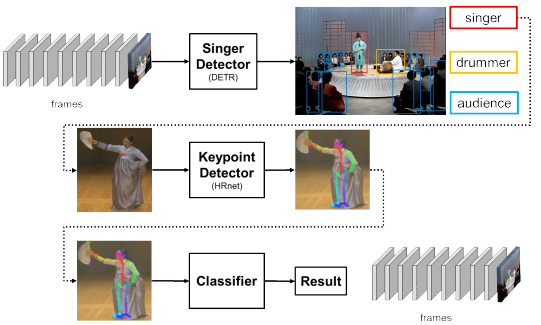

그림 1. 제안된 방법의 파이프라인
Fig. 1. The pipeline of the method

2.1 DETR를 이용한 소리꾼 검출

2.1.1 학습 데이터의 구성

동작 분류를 수행하기 전에 먼저 동작을 취하는 소리꾼을 찾아야 한다. 판소리 동영상에 등장하는 인물은 소리꾼, 고수, 청중(사회자, 스태프들도 청중으로 간주함)들이 있다. 본 논문에서는 소리꾼만 찾은 것이 목적이지만 소리꾼을 잘 찾기 위해서 소리꾼 검출을 위해 소리꾼, 고수, 청중 3개의 클래스로 나누어 객체 검출을 수행하였다.

소리꾼과 고수 클래스의 데이터는 인터넷에서 수집하여 총 2,013장으로 구성되었다. 소리꾼만 포함한 영상 739장이었으며 고수까지 포함한 영상은 1,274장이었다. 청중 클래스의 학습 데이터 부족을 보충하기 위해 MS COCO 데이터셋에서 person 클래스 영상들을 청중 클래스로 추가하여 총 3013장을 7대3으로 학습 데이터와 테스트 데이터로 나누었다.

2.1.2 소리꾼 검출 방법

앞서 언급한 바와 같이 소리꾼 동작을 분류하기 위해서는 동영상을 구성하는 매 프레임에서 소리꾼을

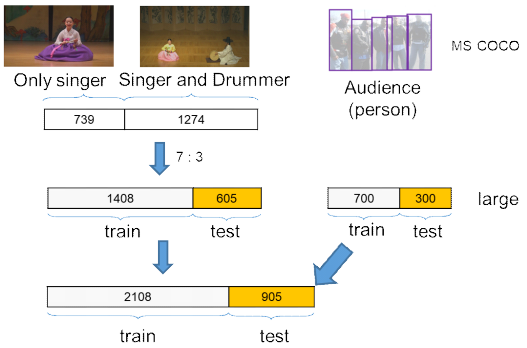

그림 2. 소리꾼 검출 데이터의 구성
Fig. 2. The structure of the data for the singer detection

검출해야 한다. 객체 검출에 관한 연구는 2-stage의 R-CNN 시리즈[6]에서 1stage의 YOLO(You Only Look Once) 시리즈[7] 까지 수많은 연구가 있어 왔으며, 최근 몇 년간 자연어 처리 분야에 많이 적용된 트랜스포머[8] 모델이 컴퓨터 비전 분야에서도 활발하게 적용되고 있다. 본 연구에서는 객체의 크기 변화가 심하지 않을 경우 성능이 우수하며 NMS(Non-Maximum Suppression)연산이 제외되어 검출객체 수가 작은 경우 속도 면에서도 유리한 DETR[9]라는 모델을 사용한다.

대다수의 검출 모델이 proposal, anchor 혹은 windows 등을 이용해서 검출 문제를 분류 문제와 회귀 문제로 구성하여 간접적으로 해결하는 것과 다르게 DETR 모델은 그림 3과 같이 검출 문제를 직접 end-to-end방식인 집합 예측(set prediction) 문제로 간주하여 트랜스포머의 encoder-decoder 구조로 N 개의 객체 바운딩 박스를 예측하며, 여기서 N은 사전에 설정한 영상 중에 존재한 객체의 개수보다 훨씬 큰 정수이다. 학습은 예측한 바운딩 박스와 ground truth의 비교는 이분 그래프(bipartitegraph)의 이분 매칭을 수행하여 예측한 좌표와 클래스를 ground truth에 점점 접근하게 손실함수를 정의하고 이를 감소시키는 방식으로 진행된다. 본 논문에서는 소리꾼을 검출하기 위해 객체 크기 변화가 크지 않고 검출할 객체수가 한정된 경우 Faster R-CNN 모델의 성능과 비슷하며 실행 속도도 더욱 빠른 DETR를 사용하였다. 즉 DETR는 작은 객체에 대한 식별력이 Faster R-CNN보다 못하지만 본 논문에서 검출해야 할 소리꾼 객체는 일반적인 판소리 동영상에서 작지 않으며 비슷한 크기를 가진다.

2.2 골격 정보를 이용한 발림 동작 분류

본 논문의 동작 분류는 특징점들 검출하여 소리꾼의 관절 위치 정보와 부채 위치 정보를 이용하여 분류하는 방식을 사용한다. 백본으로 사용된 HRNet[10]이 이러한 특징점들을 찾는 능력을 갖추기 위해서는 ground truth인 특징점의 위치에 있는 화소를 중심으로 가우시안분포에 의해 그림 4 같이 생성된 headmap을 표적으로 학습하고 학습된 백본 다음 단에 FC Layer를 붙여서 학습하여 동작 분류한다.

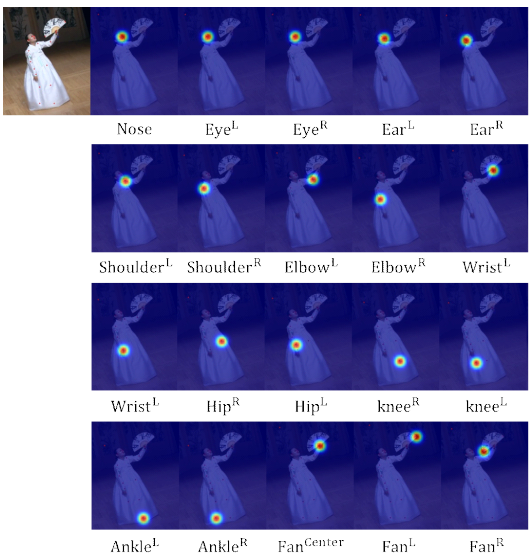

그림 4. 특징점 ground truth의 예
Fig. 4. An example of ground truth for the keypoints

2.2.1 특징점 학습 데이터 구성

먼저 골격 정보를 이용하기 위해서는 백본으로 사

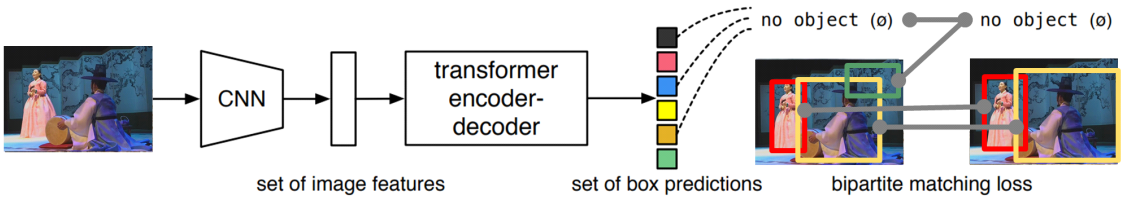

그림 3. DETR 모델의 파이프라인
Fig. 3. The pipeline of DETR

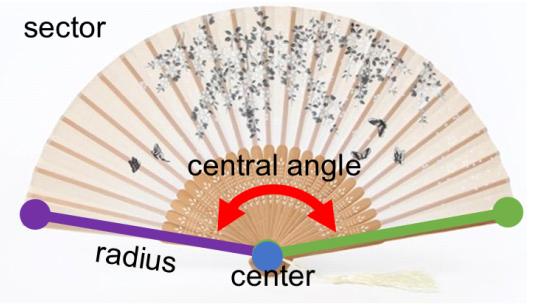

그림 5. 부채 특징점의 선정
Fig. 5. Selection of the keypints on the fan.

용된 HRNet을 특징점 데이터셋에서 초기 학습을 수행한다. 초기 학습에 사용된 특징점 데이터는 인터넷에서 영상을 수집하였고 특징점 주석을 부가하였다. 수집한 데이터는 총 575장이고 7대3의 비율로 학습 데이터와 테스트 데이터로 나누었다. 여기서 판소리의 발림에 부채는 중요한 역할을 하는 존재이기 때문에 소리꾼의 골격의 특징점뿐만 아니라 부채를 표현하는 특징점도 추가하였다. 기본적으로 부채는 부채꼴 모양이므로 중심, 반경과 중심각이 확정되면 모양이 확정된다. 이런 이유로 그림 5와 같이 부채 표현할 수 있는 특징점 3개가 추가로 선정되었다. 그러므로 특징점이 그림 4와 같이 코, 좌 우 양쪽의 눈, 귀, 어깨, 팔꿈치, 손목, 둔부, 무릎, 발목 등 17개 소리꾼의 골격 특징점과 선정된 부채의 특징점 3개로 총 20개로 구성되어 있다.

2.2.2 동작 및 학습 데이터 구성

발림 동작은 판소리 마당, 바디, 소리꾼에 따라 다른 형태로 진행된다. 따라서 본 논문에서는 판소리 발림과 관련된 논문[2-3]에서 가장 정형화된 7개의 클래스를 그림 7과 같이 정의하였다. 즉 본 연구에서는 동영상 동작 인식을 위해서는 너무 다양한 판소리 동작을 구분하기 어렵고, 동영상을 구성하는 정지 영상인 프레임 단위에서 인식할 수 있는 주요 부류를 범주로 설정하였다.

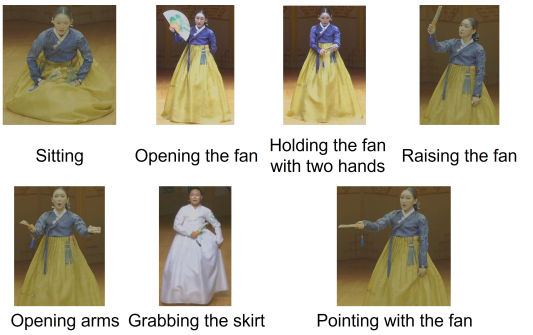

그림 7. 정의된 발림 동작의 카테고리
Fig. 7. Categories of motions

표 1. 각 동작별 데이터 수
Table 1. Number of data for each motion.

	학습	테스트
앉음(sitting)	140	60
부채 펼침(opening the fan)	140	60
부채를 두 손으로 듦 (holding the fan with two hands)	140	60
부채 듦(raising the fan)	140	60
팔 벌림(opening arms)	140	60
치마 잡기(grabbing the skirt)	140	60
부채로 지시 (pointing with the fan)	140	60

인터넷 동영상에서 클래스별 200장씩 영상을 수집했으며, 표 1같이 7대 3의 비율로 학습 데이터와 테스트 데이터로 나누었다.

2.2.3 골격 정보 찾는 백본(HRNet)

본 논문에서는 골격 정보가 포함된 특징점을 찾기 위해서 HRNet 모델을 사용했다. 특징점 검출은 마지막에 특징점 heatmap을 출력하여 ground truth와 평균 제곱 오차를 계산하는 방식으로 훈련을 수행하기 때문에 해상도가 중요하다. HRNet 발표되기 전에는 특징점 검출에 high-to-low와 low-to-high의 다해상도 프레임워크를 사용했었다[12-15]. 그러나 그림 6과 같이

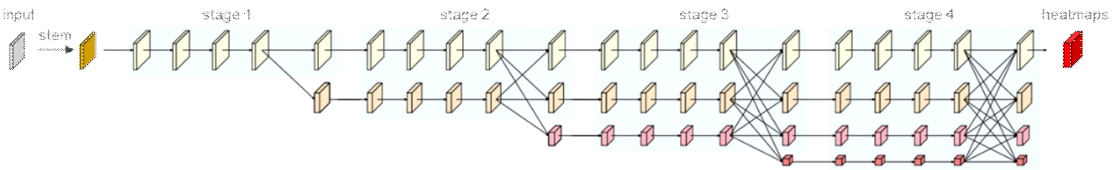

그림 6. HRNet의 아키텍처
Fig. 6. The architecture of HRNet

HRNet는 이전의 네트워크와 달리 처음부터 끝까지 고해상도를 계속 유지하며 스테이지가 계속되며 high-to-low 해상도 서브 네트워크를 점차적으로 추가하고 업 샘플링과 다운 샘플링을 통해서 다중 해상도 융합을 실현한다. 본 논문에서 다중 해상도 융합을 편리하게 구현하기 위해 HRNet을 사용하였다.

Ⅲ. 실험 및 결과

데이터 증강(augmentation) 방법[23]을 사용함으로써 학습데이터를 늘릴 수 있지만 판소리 동영상에서 소리꾼 검출과 발림 동작 분류라는 특별한 도메인을 대상으로 하는 본 연구에 적절한 방법을 마련하기 어려웠다. 따라서 과적합 없이 전체 네트워크를 학습시키기에는 부족하기 때문에 본 연구에서는 많은 데이터를 사용하여 다양한 클래스에 대해 미리 학습된 가중치를 재사용하는 전이 학습(transfer learning) 방법[17]을 유용하게 사용하였다. 즉 보편적인 분류 문제에서 다양한 분포를 가지는 데이터를 통해 학습된 가중치를 사용하는 방식을 채택하였다.

본 연구에서는 현재 수행하고자 하는 작업과 관련된 객체 검출과 특징점 검출 데이터셋인 전체 MS COCO 데이터셋으로 학습된 가중치를 이용하여 전이 학습을 수행한 뒤, 작은 학습률을 적용하여 전체 네트워크를 미세 조정하는 식으로 학습을 수행하였다. 실험은 크게 프레임 중에 소리꾼 영역을 자동으로 찾는 객체 검출 실험과 찾은 소리꾼 영역에서 소리꾼의 관절과 부채 위치인 특징점들 검출하여 동작을 분류하는 실험으로 나눌 수 있다.

3.1 소리꾼 검출

3.1.1 학습 방법

최적화 도구로는 많이 사용되는 Adam(Adaptive moment estimation)에서 Decoupled weight decay를 더해서 일반화 능력 향상시킨 AdamW[24]를 사용하였고, 전이 학습에 대한 초기 학습률 파라미터는 0.0001로, 미세 조정에 대한 초기 학습률 파라미터는 가중치가 거의 수렴하였다는 가정 하에 이보다 작은 0.00001로 설정하여 학습하였다.

3.1.2 실험결과

학습된 트랜스포머 DETR를 활용 테스트 데이터의 추론을 진행하여 표 2 같은 결과를 얻을 수 있었다. 그림 8 같이 객체 검출의 평가 지표로 사용된

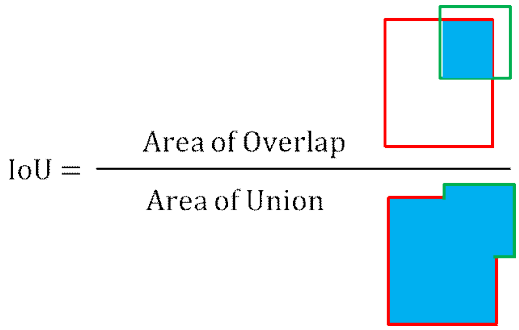

그림 8. IoU(Intersection over Union): 객체 검출의 평가 지표
Fig. 8. IoU(Intersection over Union): A evaluation metric of object detection

표 2. 소리꾼 검출의 실험 결과
Table 2. Experimental results of singer detection

	AP (IoU=0.95)	AP (IoU=0.75)	AP (IoU=0.50)
소리꾼 검출	0.74	0.91	0.935

IoU(Intersection over Union)의 문턱치를 95%, 75%, 50%로 설정할 때 AP(Average Precision)가 각각 74%, 91%, 93.5%로 나왔다. 여기서 주의해야 할 것은 본 시험에서 나온 객체는 큰 사이즈의 객체로 크기 변화가 심하지 않는 것이며 소리꾼 발림 동작 분류를 얻기 위해서는 IoU 75% 이상이면 약 90%의 소리꾼 검출이 가능하고 발림동작 분류에 큰 영향을 줄지 않는다.

만약 더 많은 객체를 검출하여 발림 동작을 판단하기 위해서는 IoU 문턱치를 줄일 수 있는데 이 경우는 발림 동작 분류를 수행하기 전에 후 처리를 통해 객체를 확장해 소리꾼 영역을 크게 해 반드시 소리꾼이 발림동작 분류에 포함되어야 한다. 그러나 매 프레임 발림 동작의 부류가 변하지는 않기 때문에 후처리 과정 없이도 약 90%의 프레임에서 소리꾼 객체를 찾으므로 빠진 프레임이 있더라도 분석에 큰 영향을 주지 않을 것으로 예상된다.

3.2 발림 동작 분류

본 연구의 파이프라인에서는 소리꾼을 검출한 후 특징을 추출해서 동작을 분류해야 한다. 여기서 주의해야 할 것은 영상 중의 컨텍스트가 동작을 인식하는 데에 있어 오히려 잘못된 결과로 이어질 수도 있다는 점이다. 실제로[4,5] 등 연구에서 컨텍스트가 동작 인식을 방해하는 예를 제공한다. 본 연구에서는 골격 정보 추출을 통해 근본적으로 컨텍스트를 제거하는 방법을 이용하고 있다. 이는 컨텍스트를 찾는 Resnet[11] 실험

결과와도 비교하였다.

3.2.1 학습 방법

우선 HRNet 백본을 MS COCO 특징점 데이터셋으로 학습을 수행한 뒤, 앞서 언급한 전이 학습 방법으로 그림 5 같이 부채의 의미 있는 특징점을 추가한 데이터셋에서 미세조정을 수행하였다. 최적화 도구는 adam[25]을 사용하였고 학습률은 초기 학습에 0.001로 미세 조정에 0.0001로 백본 학습을 수행하였다.

백본을 학습 수행 한 다음에 백본의 학습된 가중치를 고정시키고 FC Layer 분류기를 붙여서 학습하며, 최적화 도구는 adam으로 학습률은 0.01로 분류기 학습을 수행하였다.

또한 컨텍스트 정보 이용하는 모델과 골격 정보 이용하는 모델을 비교하기 위해 Resnet152를 Imagenet 데이터셋에 학습한 결과를 본 연구의 데이터셋에 대해 전이 학습하였다. 학습에 사용된 최적화 도구는 adam으로 학습률은 0.01로 분류기 미세조정을 수행하였다.

3.2.2 실험결과 및 검토

테스트 데이터에 대한 실험결과는 표 3과 표 4와 같다. 우선 HRNet 백본 전이 학습 실험에 사용하는 특징점 평가 척도는 식 (1) 의OKS(Object Keypoint Similarity)[1]다. 식 (1) 중에 d_i는 ground truth 특징점과 검출된 특징점 사이의 유클리드 거리고 v_i는 ground truth 특징점의 가시성 표시고 s는 객체 영역의 제곱근이고 k_i는 특징점마다 저하를 제어하는 상수다. 이 값은 표준 편차인 sk_i와 함께 정규화 되지 않은 가우시안 함수에 d_i를 통과시킨 값을 의미하며 각 특징점에 대해 0과1사이의 특징점 유사성을 계산한 값이 된다. 이 값들을 주석이 있는 전체 특징점 ($v_i > 0$)에 대해 평균을 구하여 특징점 평가한다. OKS 문턱치를 95%, 75%, 50%로 설정할 때 AP가 각각 81.6%, 99%, 99%의 분류 결과를 보였다.

$$OKS = \frac{\sum_i \exp\left(-d_i^2/2s^2k_i^2\right)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \quad (1)$$

또한 보통 데이터간의 불균형이 심각할 시, 정확도가 모델의 성능을 제대로 평가할 수 없기 때문에 정밀도와 재현율의 조화평균을 구하는 식 (2)의 F 스코어[26]의 β값을 1으로 (즉, 동일한 가중치로) 부여한 F_1 스코어를 사용하며 본 연구에서 FC층의 분류기를 부가하여 발림 동작 분류 학습의 결과는 F 스코어의 β 값을 재현율에 더 편중된 2로 부여한 F_2 스코어를 사용하여 측정하였다.

$$F = \frac{(1+\beta^2)(precision \times recall)}{\beta^2 precision + recall} \quad (2)$$

실험결과 표 4와 같이84.8%의 높은 값을 얻을 수 있었으며, 비교하기 위해 컨텍스트를 활용하는 Resnet 실험 결과는 64.0%에 불과하였다. 컨텍스트가 동작을 인식하는 데 부적합 수도 있는 것을 보여주는 결과로 직관적으로 어떤 동작을 취하고 있는지는 의상의 색깔, 무늬나 배경과의 관계보다 골격 정보와의 관계가 훨씬 중요하기 때문으로 풀이된다.

3.3 발림 동작 분류

앞서 소리꾼 검출 모델과 동작 분류 모델의 학습을 끝난 뒤 그림 1과 같이 동영상의 정지된 프레임에서 자동으로 소리꾼을 찾고 발림 동작 분석할 수 있게 프로그램을 작성하였고 길이 12 분 29초이고 프레임 수 총 21,721인 동영상에 적용하여 결과를 출력했다.

실행한 결과 그림 9 같이 소리꾼이 있는 영역에 바운딩 박스를 그렸고 구체적인 발림 동작으로 표기되어 있다. 통합 실행의 결과에 클래스별 각각 프레임 수는 표 5와 같다.

실행한 결과, 소리꾼과 함께 등장한 고수, 청중, 사회자, 스태프등 인물이 있음에도 불구하고 소리꾼을 찾아서 해당한 발림 동작 클래스에 분류되는 능력이 보였으며, 동작의 최고조 상태가 아닌 프레임에는 성능이 약간 떨어지는 현상이 보였다. 이와 같은 현상은

표 3. HRNet 백본 초기 학습의 실험 결과
Table 3. Experimental results of pretraining HRNet for keypoint detection

	AP (OKS=0.95)	AP (OKS=0.75)	AP (OKS=0.50)
특징점 검출	0.816	0.99	0.99

표 4. 특징점 기반 실험 결과 및 컨텍스트 기반 분류 비교 실험 결과
Table 4. Comparision of experimental results using context information and skeletal information

	F2-score
콘텍스트	0.64
골격	0.848

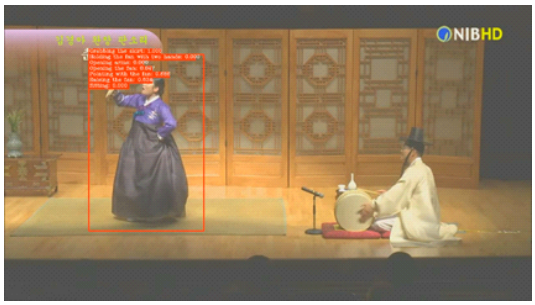

그림 9. 통합 실행의 결과
Fig. 9. The result of overall experiment

표 5. 통합 실행 결과의 통계
Table 5. the statistic result of overall experiment

	프레임 수
총	21721
소리꾼 존재	19378
앉음(sitting)	145
부채 펼침(opening the fan)	948
부채를 두 손으로 듦 (holding the fan with two hands)	6271
부채 듦(raising the fan)	145
팔 벌림(opening arms)	248
치마 잡기(grabbing the skirt)	7475
부채로 지시(pointing with the fan)	6731

학습 데이터에 있는 발림 동작 영상이 모두 소리꾼이 동작 취하는 최고조 상태였기 때문이라고 여긴다. 이를 개선하는 방법으로 최고조 상태가 아닌 프레임 영상도 학습 데이터에 추가하는 방식을 고려할 수도 있다.

또한 부채의 특징점 정보가 중요한 "부채 펼침", "부채를 두 손으로 듦", "부채 듦", "부채로 지시" 등의 경우에는 HRNet 백본이 부채의 특징점을 충실하게 검출할 수 있도록 초기 학습에 데이터를 충분하게 준비하면 발림 동작 분류의 성능도 향상될 것이 예상된다.

그림 10. 부채를 두 손으로 듦의 혼동 (좌: 정확함, 우: 혼동함)
Fig. 10. The confusion of holding the fan with two hands (Left: the example correctly classified, Right: the example confusedly classified).

특별히 그림 10의 오른쪽과 같이 부채를 한 손으로 들고 있고 다른 한 손은 허리 뒤에 숨어 있지만, 그 숨어 있는 손이 소리꾼 앞에서 볼 때 부채와 겹쳤으면 그림 10의 왼쪽같이 "부채를 두 손으로 듦"으로 혼동되는 현상이 있었다. 이러한 혼동은 본 논문에서 골격 정보를 얻는 방법이 상, 하, 좌, 우를 구분할 수 있으나 전, 후를 구분할 수 없는 2D 방법을 이용해서 생기는 것으로 분석된다.

Ⅳ. 결 론

판소리는 감정표현 방법의 하나로서 노래 내용에 따라 다른 발림 동작을 취하여 감정을 표현하고, 이러한 발림 동작은 판소리에서 중요한 역할을 한다. 본 논문에서는 동영상의 정지된 프레임에서 소리꾼을 찾고 컨텍스트 정보에 비교적 영향을 받지 않는 골격 정보를 이용하여 자동적으로 발림 동작을 분석하는 방법을 제안하였다.

제안된 방법은 소리꾼 검출은 약 90%의 정확도를 보였으며 7개 부류로 분류한 동작인식에서 스코어가 84.8%에 달하였다. 향후 문화 콘텐츠 분석 및 보존, 교육 등 분야의 활용을 기대되며, 설장구[16]와 같이 무용극이나 뮤지컬 등의 연구에도 확장할 수 있을 것으로 예상된다.

본 연구에서 제시한 방법은 정지된 프레임 영상에서 동작을 분류하여 인식하는 방법이었고 향후에는 다양한 동영상 발림의 부류가 정의되면 동작의 시작과 완료까지 다 포함하는 동영상 클립을 활용하는 동작 인식이 예상되고, 판소리에서 소리꾼의 얼굴이 역시 또 가락이나 노래 내용에 따라 희로애락을 표현하기 때문에 이런 방향의 연구도 기대된다.

References

[1] T. Y. Lin, et al., "Microsoft coco: Common objects in context," *in Proc. ECCV 2014*, pp. 740-755, Zurich, Switzerland, Sep. 2014.

[2] Jeonggeum Seo, A study on Balim in Pansori -Focusing on shim chong ga(2008), Retrieved Aug., 11, 2008, from https://academic.naver.com/article.naver?doc_id=15292732.

[3] Hyeonju Lee, The Analysis of Balim as

'Pansori Dance' and the Study on the Choreographer's Role in Changgeuk Opera : on the topic <Sugungga>(2014), Retrieved Aug., 01, 2014, http://www.riss.kr/link?id=T13525109.

[4] Y. Li, and N. Vasconcelos, "Resound: Towards action recognition without representation bias," *in Proc. ECCV 2018*, pp. 513- 528, Munich, Germany, Sep. 2018.

[5] P. Weinzaepfel and G. Rogez "Mimetics: Towards understanding human actions out of context," *Int. J. Computer Vision*, vol. 129, no. 5, pp. 1675-1690, March 2021.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in NIPS*, pp. 91-98, Montreal, Canada, Dec. 2016.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE CVPR*, pp. 779-788, Las Vegas, USA, 2016.

[8] A. Vaswani, et al., "Attention is all you need," *Advances in NIPS*, pp 6000 - 6010, Long Beach, USA, Dec. 2017.

[9] N. Carion, et al., "End-to-end object detection with transformers," *Eur. Conf. Comput. Vision*, pp. 213-229, Glasgow, UK, Aug. 2020.

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. CVPR*, pp. 5693-5703, Long Beach, USA, June 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, pp. 770-778, Las Vegas, USA, June 2016.

[12] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *Eur. Conf. Computer Vision*, pp. 483-499, Amsterdam, The Netherlands, Oct. 2016.

[13] Y. Chen, et al., "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE Conf. CVPR*, pp. 7103-7112, Salt Lake City, USA, June 2018.

[14] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. ECCV 2018*, pp. 466-481, Munich, Germany, Sep. 2018.

[15] E. Insafutdinov, et al., "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," *Eur. Conf. Comput. Vision*, pp. 34-50, Amsterdam, The Netherlands, Oct. 2016.

[16] E. H. Park, "The study on dance performance of the Seoljanggu: Focused on the style of Yang, Do Il in Chungchong Province," *The Korean J. Dance Stud.*, vol. 8, no. 8, pp. 139-161, Dec. 2001.

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Eng.*, vol. 22, no. 10, pp. 1345-1359, Oct. 2009.

[18] A. Karpathy, et al., "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. CVPR*, pp. 1725-1732, Columbus, USA, June 2014.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, pp. 568-576, Montreal, Canada, Dec. 2014.

[20] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 35, no. 1, pp. 221-231, Jan. 2013.

[21] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 40, no. 6, pp. 1510-1517, June 2018.

[22] E. Mansimov, N. Srivastava, and R. Salakhutdinov, Initialization strategies of spatio-temporal convolutional neural networks(2015), Retrieved Mar., 25, 2015, from https://doi.org/10.48550/arXiv.1503.07274.

[23] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *J. Big Data*, vol. 6, no. 1, pp. 1-48, July 2019.

[24] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *International Conference on Learning Representations*, New Orleans, USA, May 2019.

[25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, San Diego, USA, May 2015.

[26] Y. Sasaki, *The truth of the f-measure(2007)*, Retrieved Oct., 26, 2007, from https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf[accessed 2021-05-26].

오 문 흠 (Wenqin Wu)

2020년 : Zhejiang Gongshang University Hangzhou College of Commerce 공학사

2020년 9월~현재 : 전북대학교 전자정보공학과(컴퓨터공학) 석사과정

<관심분야> 영상처리, 딥러닝

[ORCID:0000-0002-0322-7142]

이 혜 정 (Hyejeong Lee)

2021년 2월 : 전북대학교 한국음악학과 졸업

2021년 3월~현재 : 전북대학교 한국음악학과 석사과정

<관심분야> 한국음악, 국악, 국악이론, 판소리

[ORCID:0000-0001-7667-3234]

이 준 환 (Joonwhoan Lee)

1980년 2월 : 한양대학교 전자공학과 공학사

1982년 2월 : 한국과학기술원 전자공학과 공학석사

1990년 8월 : 미국 미주리대학 전기 및 컴퓨터공학과 공학박사

1990년 10월~현재 : 전북대학교 컴퓨터공학부 교수

<관심분야> 영상처리, 감성 분석, 인공지능

[ORCID:0000-0003-1854-9643]