비디오 분류를 이용한 가상 피아노

강성재 O 김재윤 윤성의

한국과학기술원 전산학과

tjdwo2744@kaist.ac.kr, jaeyoon1603@gmail.com, sungeui@kaist.ac.kr

Virtual Piano using Video Classification

SeongJae Kang O Jaeyoon Kim sungeui Yun

School of Computing, Korea Advanced Institute of Science and Technology

요 약

기계학습, 특히 딥러닝 기법이 컴퓨터 비전 분야에 적용되면서 기술적으로 많은 발전을 가져오고 있다. 또한 디지털 음악 분야에서는 시각적으로 음악을 분석하려는 시도들이 이루어지고 있다. 본 논문에서는 압력센서를 이용해 피아노 건반을 직접 제작하고, 이를 통한 연주 동영상에서 연주자의 손 동작을 검출하여 연주하는 피아노 건반의 위치, 길이, 세기 등을 분석해 낼 수 있는 방법을 제시하며 이후 연구의 방향과 가능성을 제시한다.

1. 서 론

기계학습, 특히 딥러닝 기법은 인공지능 분야에 큰 발전을 가져다 주었다. 특히 컴퓨터 비전은 딥러닝의 영향으로 크게 받은 분야로서, 컨볼루션 인공 신경망을 활용한 이미지 분류, 물체 인식 등의 가시적인 성과를 보여왔다 [5, 6, 7]. 뿐만 아니라, 컨볼루션 인공 신경망을 이용한 딥러닝 기술은 동영상을 분석하는 데에도 두각을 드러내고 있으며, 동영상에서의 동작 인식, 물체 추적, 분류 등에 있어서 딥러닝 모델을 활용한 방법은 기존의 고전적인 모델들보다 높은 정확도를 보이고 있다 [8].

음악은 소리뿐 아니라 시각적인 요소와 함께 소비될 때 청중들에게 큰 영감을 준다. 음악 연주회가 악기와 함께 연주자의 퍼포먼스가 더해지고, 현대 음악들이 비디오와 함께 소비되는 것을 통해 이를 알 수 있다. 음악을 소리와 함께 시각적인 요소들과 결합하여 연구하려는 시도 또한 이루어졌는데, 대표적으로 기타연주 영상에서 줄의 흔들림을 분석하거나 [9], 피아노 연주 영상에서 건반의 움직임을 분석하는 등의 연구가 진행되어 왔다 [1, 2, 3, 4].

한편, 기존의 피아노의 건반 움직임을 분석하는 연구들은 공통적으로 건반을 누르는 세기, 즉 음의 세기를 고려하지 않았다는 한계점이 있다[1, 2, 3, 4]. 피아노 연주자는 음의 위치, 길이와 함께 음의 강약을 통해서 감정을 표현하기 때문에 이러한 세기 요소를 분석하는 것은 중요하기 때문이다.

본 연구는 음악의 시각적인 요소만을 컴퓨터 비전과 딥러닝을 이용해 분석하려는 연구이다. 구체적으로는, 피아노 연주 동영상 내에서 연주자의 손동작을 분석하여 연주자가 어떤 피아노 건반을 어떤 세기로 누르는지에 대해 파악하는 데에 목적이 있다.

2. 관련 연구

2.1 피아노 건반 검출

피아노 건반을 검출해내는 다양한 방법[1, 2, 3, 4, 10]들이 연구되었으며, 기존의 컴퓨터 비전 알고리즘들을 통해서 구현되었다. 피아노는 건반들이 직사각형을 이루고, 흰 건반과 검은 건반이 같은 패턴으로 반복된다. 따라서 선행 연구들은 허프 변환을 통해서 피아노 건반 영역의 가장자리를 이루는 직선을 찾아 건반 영역을 검출하였다. 또한 이후에 이진화 알고리즘을 통해서 흰 건반과 검은 건반을 구분해 내고, 연결 성분 알고리즘 등을 통해서 건반을 검출하는 방법을 사용하였다.

2.2 Convolutional Neural Network(CNN)

컨볼루션 인공 신경망의 활용은 이미 이미지 분류에 있어서 인간의 정확도를 뛰어 넘는 등 큰 두각을 드러내었다 [5, 6, 7]. 한편 한 장의 이미지분류가 아닌 동영상을 분류하는 연구는 기존의 이미지에서 시간에 대한 축이 더해지기 때문에 좀 더 많은 데이터와 새로운 모델을 필요로 한다. 특히 짧은 동영상에 대해서는, 논문[8]에서 목적에 따른 컨볼루션 인공 신경망을 적용하는 방법을 제시하였으며, 도식화하면 그림 1과 같다.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 사업의 연구결과로 수행되었음"(2016-0-00018)
“본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임”(2019R1A2C3002833)

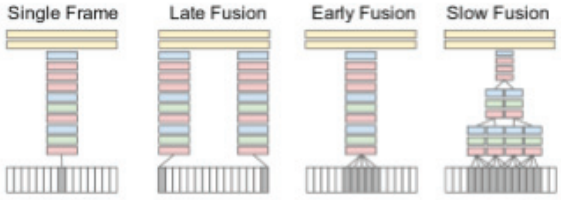

그림 1 컨볼루션 인공 신경망을 짧은 동영상에서 적용하는 방법을 제시한다 [8].

2.3 피아노 연주 분석

컴퓨터 비전을 이용해서 피아노 연주를 분석하여는 시도 또한 연구되어 왔다. 빛을 피아노 위 건반으로 비추고 사람의 손에 의해 생기는 그림자를 이용한 방법[4], 건반 가장자리의 밝기 값 변화를 이용한 방법[1, 2], 그리고 건반 가장자리와 함께 사람의 손 위치를 동시에 학습시키는 방법[3] 등이 연구되었다. 첫번째 방법[4]의 경우 손이 건반을 눌렀을 때 손과 바닥에 생기는 그림자의 면적이 최소화 된다는 사실에 기반한 알고리즘이며, 이 때문에 빛의 위치에 따라서 정확도가 달아질 수 있는 한계점이 있다. 두번째인 건반의 가장자리를 이용하는 방법[1, 2]은, 피아노 건반이 눌렸을 때 상대적으로 주변 건반의 밝기가 두드러지는 현상을 이용한 것이다. 하지만 이러한 방법 또한 손이 건반을 누르는 속도, 즉 음의 세기를 분석하기 어렵다는 한계점이 있다. 마지막으로 사람의 손의 위치를 동시에 학습시키는 방법[3]은 건반의 영역을 특징 기술자로 사용하여 학습을 시키는 방법으로, 이 역시 동영상내에서 하나의 이미지만을 이용해서 학습시켰기 때문에 음의 세기를 분석하기 어렵다는 한계점이 있다.

3. 피아노 건반 검출

3.1 피아노 건반 영역 검출

본 연구는 사람의 손의 움직임을 통해서 음의 위치와 세기를 연구해내는 데에 목적이 있었다. 따라서 기존의 모델들과 달리, 빛의 위치가 항상 건반 위를 바라보지 않을 수 있고, 건반의 가장자리 정보를 이용하지 않아야 한다. 따라서 이러한 특징을 반영할 수 있는 건반을 직접 제작하였으며 그림 2와 같다.

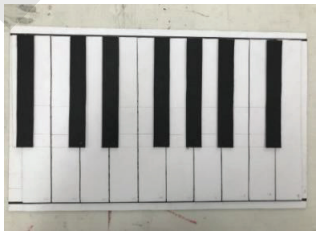

그림 2 제작한 피아노 건반 모형이다.

본 연구는 건반을 직접 제작하였기 때문에 기존 피아노 건반 검출 알고리즘과 다른 방법을 사용하여야 한다. 따라서 기본적인 아이디어는 기존 알고리즘을 참고하되 경우에 따라서 새로운 알고리즘을 고안하였다. 기본적인 방법은 먼저 건반이 있는 이미지에서 허프 변환을 통해서 건반 영역의 가장자리를 이루는 4개의 선을 찾고, 이를 통해서 건반 영역을 검출하였다. 그리고 검출된 영역에서 이진화 알고리즘을 통해서 검은 건반과 흰 건반을 분리한 뒤, 연결 성분 알고리즘을 통해서 검은 건반을 검출하였다. 마지막으로 검은 건반과 흰 건반이 동일한 패턴으로 반복되기 때문에 이러한 패턴을 이용해서 흰 건반도 검출해낼 수 있었다. 이렇게 검출된 건반은 그림 3과 같이 시각화 하였다.

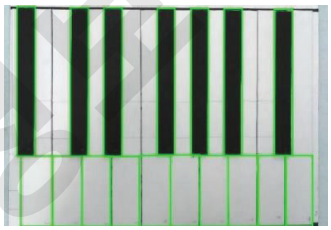

그림 3 건반 검출 알고리즘의 결과를 시각화 하였다.

3.2 특징 기술자 추출

기계 학습을 이용해서 연주를 분석할 것이기 때문에 학습과 예측에 쓰일 특징 기술자를 추출할 필요가 있다. 특히 본 연구에서는 컨볼루션 인공 신경망을 이용할 것이기 때문에, 특징 기술자를 추출할 이미지에서 배경 이미지의 차 영상을 구한 뒤, 이 차 영상에서 건반이 있는 영역을 임의의 직사각형 사이즈로 변환을 시키고 이를 특징 기술자로 사용하기로 하였다. 피아노 건반이 이미 검출된 상태이므로 제작한 건반의 실제 크기를 고려하여 그림 4와 같이 흰 건반을 20×40, 검은 건반을 10×75의 크기로 변환 후 이를 특징 기술자로 사용하기로 하였다

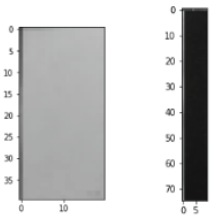

그림 4 손이 올려져 있지 않은 건반의 특징 기술자이다. 시각적으로 이해를 돕기 위해서 차 영상이 아닌 기술자를 추출할 이미지에서 바로 변환을 한 그림이다.

4. CNN을 이용한 연주 분석

검출된 건반을 바탕으로 연주자가 피아노를 연주했을 때에 어떤 건반을 어떤 세기로 누르는 지에 대해 분석하기 위해서 컨볼루션 인공 신경망을 이용하기로 하였다. 이는 [3]에서 제시된 방법을 발전시킨 것으로, 기존에는 하나의 이미지만을 이용해서 건반의 눌림을 분석하는 모델이었다면 이 방법은 동영상에서의 인접한 시간의 여러 이미지들을 함께 학습시키는 방법을 사용하여야 한다. 이때 연주자의 손이 순간적으로 움직이는 동작을 분석하여야 하므로 [8]에서 제시된 것처럼 Early Fusion 방법을 적절히 변형해서 사용한다면 좋은 결과를 얻을 것이다.

기존에 행해진 연구가 아니므로 학습 데이터를 직접 제작하였다. 이때 데이터의 수집은 그림 5와 같이 제작한 건반에 압력센서를 부착하여 건반을 눌렀을 때의 압력 값을 얻을 수 있도록 하였으며, 웹 카메라가 건반의 전체적인 모습을 담을 수 있도록 지지대에 카메라를 연결하여 동영상과 압력 정보를 동시에 수집할 수 있는 환경을 만들었다. 이를 바탕으로 1차 데이터 집합을 만들었으며, 이후 이 압력 값을 적절히 해석하여 건반의 눌림, 그리고 건반을 누르는 세기에 대한 라벨로 변형시켜 학습한다면 의미 있는 결과를 얻을 수 있을 것이다.

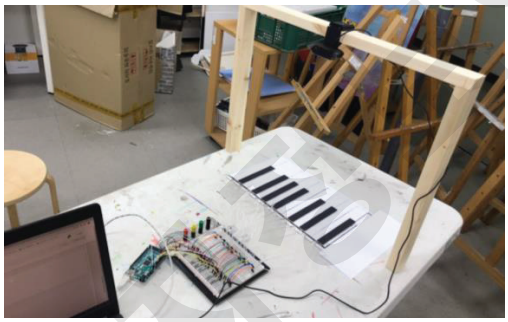

그림 5 학습 데이터 집합 제작

5. 결론

본 논문에서는 소리가 없는 피아노 연주 동영상에서 음을 분석해내는 방법에 대한 관련 논문을 소개하고 기존의 모델의 한계를 개선할 수 있는 방법을 제시하였다. 또한 실제로 피아노 건반을 검출하는 알고리즘을 만들고 컨볼루션 인공 신경망에 쓰일 1차 데이터 집합을 만드는 일련의 과정을 통해 컴퓨터 비전을 통해서 가상의 피아노를 구현할 수 있는 연구 방향과 가능성을 제시하였다.

6. 참고문헌

[1] Suteparuk P: Detection of piano keys pressed in video. Tech. rep., Department of Computer Science, Stanford University (2014)

[2] Mohammad Akbari, Howard Cheng: Real-Time Piano Music Transcription Based on Computer Vision. IEEE Trans. Multimedia 17(12): 2113-2121 (2015)

[3] Mohammad Akbari, Jie Liang, Howard Cheng: A real-time system for online learning-based visual transcription of piano music. Multimedia Tools Appl. 77(19): 25513-25535 (2018)

[4] Boga Vishal, K Deepak Lawrence Paper piano - Shadow analysis based touch interaction. 2nd International Conference on Man and Machine Interfacing (MAMI) (2017)

[5] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012: 1106-1114

[6] Karen Simonyan, Andrew Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition. CoRR abs/1512.03385 (2015)

[8] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Fei-Fei Li: Large-Scale Video Classification with Convolutional Neural Networks. CVPR 2014: 1725-1732

[9] Shir Goldstein, Yael Moses: Guitar Music Transcription from Silent Video. BMVC 2018: 309

[10] Adam Goodwin, Richard D. Green: Key detection for a virtual piano teacher. IVCNZ 2013: 282-287