특집논문 (Special Paper)
방송공학회논문지 제24권 제6호, 2019년 11월 (JBE Vol. 24, No. 6, November 2019)
https://doi.org/10.5909/JBE.2019.24.6.983
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

인공신경망을 이용한 샷 사이즈 분류를 위한 ROI 탐지 기반의 익스트림 클로즈업 샷 데이터 셋 생성

강 동 완[a)], 임 양 미[b)‡]

Generating Extreme Close-up Shot Dataset Based On ROI Detection For Classifying Shots Using Artificial Neural Network

Dongwann Kang[a)] and Yang-mi Lim[b)‡]

요 약

본 연구는 영상 샷의 크기에 따라 다양한 스토리를 갖고 있는 영상들을 분석하는 것을 목표로 한다. 따라서 영상 분석에 앞서, 익스트림 클로즈업 샷, 클로즈업 샷, 미디엄 샷, 풀 샷, 롱 샷 등 샷 사이즈에 따라 데이터셋을 분류하는 것이 선행되어야 한다. 하지만 일반적인 비디오 스토리 내의 샷 분포는 클로즈업 샷, 미들 샷, 풀 샷, 롱 샷 위주로 구성되어 있기 때문에 충분한 양의 익스트림 클로즈업 샷 데이터를 얻는 것이 상대적으로 쉽지 않다. 이를 해결하기 위해 본 연구에서는 관심 영역 (Region Of Interest: ROI) 탐지 기반의 이미지 크롭핑을 통해 익스트림 클로즈업 샷을 생성함으로써 영상 분석을 위한 데이터셋을 확보 방법을 제안한다. 제안 방법은 얼굴 인식과 세일리언시(Saliency)를 활용하여 이미지로부터 얼굴 영역 위주의 관심 영역을 탐지한다. 이를 통해 확보된 데이터셋은 인공신경망의 학습 데이터로 사용되어 샷 분류 모델 구축에 활용된다. 이러한 연구는 비디오 스토리에서 캐릭터들의 감정적 변화를 분석하고 시간이 지남에 따라 이야기의 구성이 어떻게 변화하는지 예측 가능하도록 도움을 줄 수 있다. 향후의 엔터테인먼트 분야에 AI 활용이 적극적으로 활용되어질 때 캐릭터, 대화, 이미지 편집 등의 자동 조정, 생성 등에 영향을 줄 것이라 예상한다.

Abstract

This study aims to analyze movies which contain various stories according to the size of their shots. To achieve this, it is needed to classify dataset according to the shot size, such as extreme close-up shots, close-up shots, medium shots, full shots, and long shots. However, a typical video storytelling is mainly composed of close-up shots, medium shots, full shots, and long shots, it is not an easy task to construct an appropriate dataset for extreme close-up shots. To solve this, we propose an image cropping method based on the region of interest (ROI) detection. In this paper, we use the face detection and saliency detection to estimate the ROI. By cropping the ROI of close-up images, we generate extreme close-up images. The dataset which is enriched by proposed method is utilized to construct a model for classifying shots based on its size. The study can help to analyze the emotional changes of characters in video stories and to predict how the composition of the story changes over time. If AI is used more actively in the future in entertainment fields, it is expected to affect the automatic adjustment and creation of characters, dialogue, and image editing.

Keyword : video storytelling, shot sizes types, artificial neural network, region of interest, image saliency

Copyright Ⓒ 2016 Korean Institute of Broadcast and Media Engineers. All rights reserved.
"This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered."

Ⅰ. 서 론

MIT 미디어랩은 최근 비디오 스토리텔링(video story-telling)에서의 기계학습과 인간의 협업 가능성에 대한 연구를 진행했다[1]. 이 연구는 오디오 정보, 이미지 정보, 대사 정보를 활용하여 기쁨, 슬픔, 분노 등 캐릭터들의 감정적 흐름을 시간의 흐름에 따라 2차원 곡선 그래프에서 긍정적 정점과 부정적 정점으로 정의하였는데, 이를 바탕으로 인기 있는 비디오 스토리와 그렇지 않은 비디오 스토리에 대한 이론정립을 시도하였다. 이와 같이, 비디오 스토리텔링의 정량적 분석 및 평가 가능성을 시사하는 연구들이 등장하고 있는데, 다양한 분야에서 괄목할만한 성과를 보여주고 있는 딥러닝(deep learning) 기술을 활용한 연구들이며, 이러한 연구가 점차 늘어나는 추세에 있다.

본 연구는 비디오 스토리텔링 흐름을 분석하기 위해 영화 장면연출에 따른 샷 사이즈(shot size) 변화 추이를 활용하고자 한다. 따라서 영상의 샷 사이즈를 탐지하는 기술이 필요한데, 본 연구에서는 인공 신경망(Artificial Neural Network)을 이용하여 샷 사이즈에 따른 영상 분류기를 생성한다. 이를 위해, 익스트림 클로즈업 샷(extreme close-up shot), 클로즈업 샷(close-up shot), 미디엄 샷(medium shot), 풀 샷(full shot), 롱 샷(long shot)등으로 분류된 그림 1의 최종 단계인 샷 사이즈 이미지 (shot size images classification DB) 훈련 데이터셋이 필요하다. 하지만 일반적으로 영화의 샷 사이즈 분포는 클로즈업 샷, 미디엄 샷, 풀 샷, 롱 샷에 치중되어 있는 반면, 익스트림 클로즈업 샷은 등장 빈도가 상대적으로 적은 편이기 때문에 데이터셋 확보에 어려움이 있다. 더욱이 익스트림 샷은 대체로 영화 내의 스토리 흐름에서 중요한 의미를 갖는 경우가 많기 때문에 익스트림 샷 데이터셋 확보의 필요성이 더욱 증대된다. 본 연구에서는 기(旣) 확보한 클로즈업 샷 영상으로부터 관심 영역을 탐지한 뒤 크롭핑하여 익스트림 클로즈업 샷을 생성함으로써 익스트림 클로즈업 샷 데이터셋을 확보(그림 1. Data Augmentation)하는 방안을 제안한다.

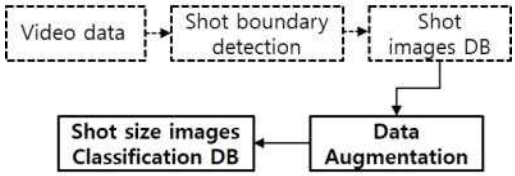

그림 1. 본 연구의 전체 흐름도
Fig. 1. Flowchart of the study

일반적으로 클로즈업샷에는 사람의 얼굴이 포함되어 있다. 따라서 본 논문에서는 얼굴이 포함된 영역을 관심 영역으로 정의하고, 얼굴 검출을 통해 관심 영역을 찾아 크롭핑하는 전략을 사용한다. 하지만 얼굴 검출에 실패하거나 얼굴이 포함되지 않은 샷에 대해서는 얼굴 검출 방법만으로 관심 영역을 찾을 수 없다. 이를 해결하기 위해, 본 연구에서는 세일리언시 검출을 보조적으로 활용하는 방법을 제안한다. 세일리언시 검출[2]은 영상으로부터 시각적으로 주의를 끄는 영역을 찾는 것을 의미하는데, 컴퓨터 비전 및 인지 과학 등의 분야에서 관심 영역을 탐지하기 위한 용도로 널리 사용되어 왔다[3]. 본 연구에서는 먼저 얼굴 검출을 통해 관심 영역 탐지를 시도한 뒤, 그것이 실패한 경우 세일리언시 검출을 수행하여 관심 영역을 추정하는 방법을 사용한다.

본 연구에서는 이를 통해 확보한 데이터셋을 인공신경망을 통해 학습 및 샷 사이즈 분류에 효과적으로 사용됨을 보임으로써 제안하는 방법이 적절히 익스트림 클로즈업 데이터를 생성함을 검증한다. 제안하는 방법은 스토리를 갖고 있는 영상들의 전체적인 샷 사이즈 변화 양상 파악에 활용될 수 있어 향후 샷 사이즈 변화 패턴을 통한 비디오 스토리텔링 분석[4]이 가능할 것으로 전망된다. 본 논문의 구성은 2장에서 세일리언시 기반의 관련연구를 설명하고 3장에서 부족한 데이터셋 확장을 위한 세일리언시 검출 알고리즘과 샷 사이즈 유형별 분류 모델 구축에 관해 설명한

a) 서울과학기술대학교 컴퓨터공학과(Department of Computer Science and Engineering, Seoul National University of Science and Technology)
b) 덕성여자대학교 IT 미디어공학과(Department of IT Media Engineering, Duksung Women's University)
‡ Corresponding Author : 임양미(Yangmi Lim)
E-mail: yosimi@duksung.ac.kr
Tel: +82-2-901-8350
ORCID: https://orcid.org/0000-0002-3725-0025
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.NRF-2017R1D1A1B03028804).
※ 이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.NRF-2017R1D1A1B03028804).
· Manuscript received September 16, 2019; Revised November 11, 2019; Accepted November 11, 2019.

다. 4장에서 실험결과를 보여주고 결론을 맺는다.

Ⅱ. 관련연구

이미지 내에서 얼굴을 탐지하는 것은 얼굴 인식[5], 얼굴 추적[6], 얼굴 표정 분석[7], 얼굴 속성 인식[8], 얼굴 형태 재구성[9], 영상 검색[10] 등 다양한 컴퓨터 비전 관련 분야에서 오랫동안 연구되어 왔다. 초기의 얼굴 인식 연구들은 주로 지식 기반 접근법과 이미지 기반 접근법으로 구분될 수 있다. 지식 기반 얼굴 탐지 접근법을 사용한 연구들에서는 에지, 명도, 색상 등 저수준 특징을 분석[11-13]하거나 ASM(Active Shape Model)[14-16]과 같은 얼굴 템플릿을 이용하여 얼굴을 탐지하는 방법을 제안하였다. 반면, 이미지 기반 얼굴 탐지 접근법의 연구들에서는 얼굴 이미지에 대해 주성분 분석 등 선형 부분 공간[17,18]을 적용하거나 인공신경망[19,20] 등을 이용하여 얼굴을 탐지하였다. 하지만, 이러한 초기 연구들은 주로 조명, 자세 등 특정 조건 하의 얼굴 이미지에서만 좋은 성능을 보임으로써, 실생활 응용에서 실용적으로 사용되기 어렵다는 한계를 지녔다.

이후, SIFT(Scale Invariant Feature Transform), SURF (Speeded Up Robust Features) 등 강인한 특징 추출 방법들[21,22]과 SVM(Support Vector Machines), 딥러닝(인공신경망) 등의 강력한 식별 방법론[23,24]을 적용한 얼굴 탐지 방법들이 등장하면서 얼굴 인식의 성능이 향상되었다. 특히, 신경망을 다층(layers) 구조로 구성하여 입력층과 출력층 사이에 숨겨진 여러 층을 거치면서 대량의 데이터 분류와 학습을 통해 데이터 기반의 예측이 가능하도록 하는 기술인 딥러닝[25]이 얼굴 탐지에 사용되면서 탐지 성능의 급격한 향상을 가져왔다. 본 연구에서는 합성곱 신경망(Convolutional Neural Network)을 사용하여 얼굴 탐지와 얼굴 정렬을 결합하여 해결함으로써 얼굴 탐지 성능을 높인 방법[26]을 사용하여 효과적으로 얼굴을 탐지하였다.

본 연구에서는 이미지에서 관심영역을 찾기 위해 세일리언시 탐지 방법을 활용한다. 세일리언시는 인간의 시각적 주의 메커니즘을 모델링하여 이미지에서의 관심도를 픽셀별로 계산한 것으로써, 컴퓨터 비전의 다양한 분야에서 활용되어져 왔다[27]. 세일리언시 탐지 알고리즘은 크게, 상향식(bottom-up), 하향식(top-down)으로 구분할 수 있다. 상향식 세일리언시 탐지 모델[28,29]은 색상, 밝기, 방향 등과 같은 이미지 정보 특징을 활용하여 지형적으로 배열한 맵 모델을 만들고 시선을 추정하여 관심영역을 찾는다. 이와 같은 접근 방법은 구현이 용이하고, 장면에 대한 사전 지식을 요구하지 않는다는 장점이 있다. 반면, 하향식 세일리언시 탐지 모델[30,31]은 관심 영역에 대한 사전 지식을 이용하는 방법으로써, 이미지 문맥에 대한 높은 이해를 요구하기 때문에 계산 비용이 높다.

이후, 딥러닝 기술의 발전으로 합성곱 신경망을 이용한 연구들이 등장하게 되면서 세일리언시 탐지의 정확도가 비약적으로 상승하였다. 본 연구에서는 얼굴 검출이 실패한 경우, 효과적으로 관심 영역을 검출하기 위하여 문맥 탐지 기반의 PiCANet(Pixel-wise Contextual Attention)[32]을 세일리언시 탐지 알고리즘으로 사용하였다. 문맥에 대한 고려 없이 영상처리 알고리즘에 기반한 기존의 세일리언시 탐지 알고리즘들과 달리, 픽셀 기반의 문맥적 세일리언시를 탐지함으로써 더 나은 성능을 보여주는 PiCANet은 국지적 문맥과 전역적 문맥을 합성곱 신경망을 이용해 조인트 학습한 모델을 만들고 U-Net 아키텍처에 결합함으로써 문맥적으로 눈에 띄는 대상을 탐지할 수 있는 모델을 생성하여 세일리언시를 탐지한다.

Ⅲ. 시스템 설계 및 구현

1. 시스템 개요 및 데이터 셋 구축

비디오 스토리텔링과 같은 긴 영상들은 시청자들에게 감정 전달을 위해 장면전환을 전제로 한 샷 사이즈 변화를 통해 극적 효과를 나타낸다. 예를 들어 클로즈업 샷은 캐릭터의 상반신을 근접 촬영하여 감정 변화를 자세히 보여주고, 미들 샷은 인물 간의 대화 장면을 통해 객관적 상황을 나타낸다. 풀 샷은 인물 전체를 표현하거나 배경과 함께 상황을 보여주는 경우에 사용한다[4]. 이러한 비디오 스토리텔링의 흐름을 분석하기 위해 본 연구에서는 입력 영상의 샷 사이즈 유형을 자동으로 분류하는 방법을 제안한다. 샷 사이즈 유형의 자동 분류를 위해 인공신경망을 이용한 기계학습을 사용하는데,

그림 2. 샷 사이즈 5개의 유형 데이터
Fig. 2. Type data of 5 shot sizes

이를 위해서는 샷 사이즈 유형별 훈련 데이터 셋 구축이 요구된다. ImageNet이나 MNIST 등, 널리 검증된 기계학습을 위한 영상 데이터 셋이 관련 연구들에서 사용되고 있지만, 분류의 유형이 샷 사이즈의 유형이라는 본 연구의 특성상 5가지 샷 사이즈 유형별 데이터 셋(그림 2와 같이 익스트림 클로즈업, 클로즈업, 미듐, 풀, 롱)을 구축하여 사용한다. 다음은 데이터 셋 구축을 포함한 본 시스템의 개요를 나타낸다.

1) 영화의 클라이맥스 부분의 2~3분 분량의 영상을 샷 경계를 탐지하여 샷 단위의 영상으로 자르고, 영상 길이의 중간에 해당하는 프레임을 추출한다. (씬의 단위인 장소 중심의 작은 이야기 기준으로 2~3분 분량의 비디오 선별 함)[4]
2) 추출한 프레임들을 5가지 샷 사이즈 유형으로 수작업으로 분류하여 이미지(프레임)들로 구성된 5개의 데이터 셋을 구축한다.
3) 데이터 셋의 크기가 상대적으로 작은 롱 샷 데이터 셋을 보충하기 위해 온라인 이미지 검색을 통한 크롤링을 수행하여 보완한다.
4) 마찬가지로 익스트림 클로즈업 데이터 셋을 보충하기 위해 얼굴 검출을 통한 크롭핑을 시도한다. 클로즈업 샷 데이터 셋의 이미지들에 대해 얼굴 검출을 시도하여 얼굴이 검출될 경우 얼굴을 위주로 이미지를 크롭핑한다. 만약, 얼굴 검출에 실패할 경우, PiCANet을 이용해 세일리언시 검출을 수행하여 관심 영역을 탐지하고 해당 영역을 위주로 이미지를 크롭핑한다. 크롭핑이 수행된 이미지는 익스트림 클로즈업 샷을 위한 데이터 셋으로 사용한다.
5) 합성곱 신경망을 이용해 5개의 데이터 셋을 학습시켜 샷 사이즈 유형별 이미지 분류 모델을 만든다.

2. 얼굴 및 세일리언시 검출 기반 익스트림 클로즈업 샷 생성

그림 3은 클로즈업 샷 이미지로부터 익스트림 클로즈업 샷 이미지를 생성하기 위해 제안하는 알고리즘의 개요를 보여준다. 먼저, 클로즈업 샷 이미지에서 얼굴 검출을 시도한다. 얼굴 검출 알고리즘으로는 MTCNN을 사용한다[26]. 여러 개의 얼굴이 검출되는 경우, 본 연구에서는 검출 된 얼굴 바운딩 박스들의 크기를 구하고 그 중 가장 큰 박스의 얼굴을 대표 얼굴로 사용한다. 다음으로, 대표 얼굴을 포함하는 크롭핑 박스를 찾는다. 얼굴을 포함하는 크롭핑 박스는 그 크기 및 위치를 달리하는 다양한 조합으로 만들 수 있다. 본 연구에서는 입력 클로즈업 샷 크기 대비 출력 익스트림 클로즈업 샷 크기를 미리 지정하여 크롭핑 박스의 크기를 고정하였으며, 검출된 얼굴이 크롭핑 박스의 중앙에 위치할수록 좋다고 가정하였다. 따라서 검출된 얼굴의 바운딩 박스의 중점과 크롭핑 박스의 중점을 가장 가깝게 만듦으로써 최적의 크롭핑 박스를 구할 수 있다. 본 논문에서는 그림 4의 우측 하단 노란 점들과 같이, 크롭핑 박스 중점

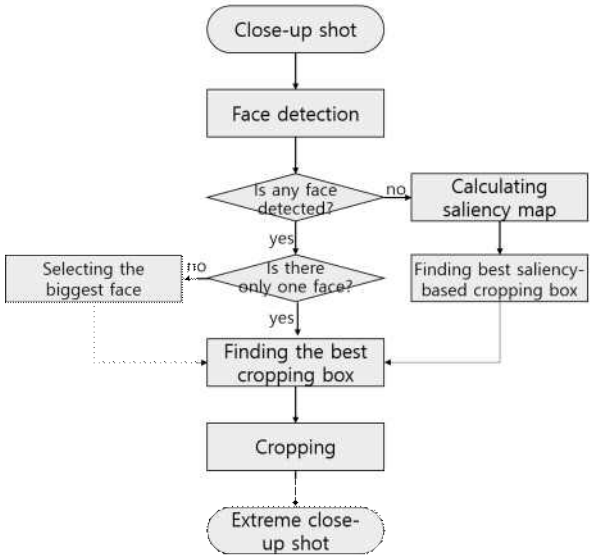

그림 3. 익스트림 클로즈업 샷 생성 알고리즘
Fig. 3. Extreme close-up shot generation algorithm

이 위치할 후보들을 격자 단위로 사전에 정의하고, 그 후보들 중 대표 얼굴로 검출된 얼굴의 바운딩 박스의 중점과 거리가 가장 가까운 것을 크롭핑 박스의 중점으로 사용하는 방법을 제안한다.

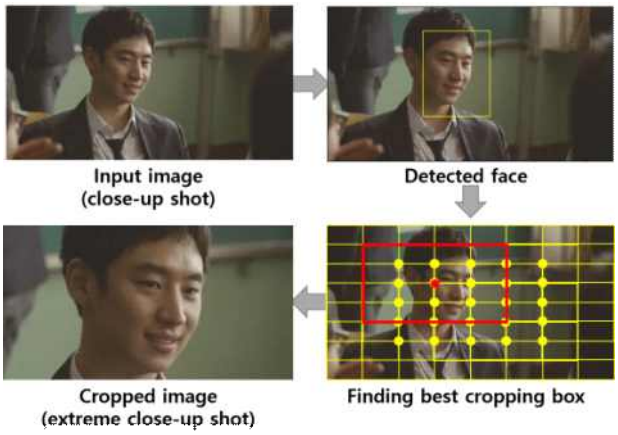

그림 4. 얼굴 검출 기반의 익스트림 클로즈업 샷 생성
Fig. 4. Extreme close-up shot generation based on face detection

만약, 그림 5와 같이 얼굴이 가려져 있어 얼굴 검출에 실패하거나 이미지 내에 얼굴이 포함되지 않는 경우 앞서 설명한 얼굴 검출 기반의 크롭핑을 수행하는 것이 불가능하다. 이를 해결하기 위해, 본 연구에서는 세일리언시 기반의 관심 영역을 검출하고 이를 크롭핑에 활용하여 익스트림 클로즈업 샷을 생성하는 방법을 제안한다. 먼저, PiCANet[32]을 이용하여 그림 4의 우측 상단과 같이 세일리언시 맵(Saliency Map)을 구한다. 세일리언시 맵은 이미지의 각 픽셀마다 이미지 내에서 눈에 띄는 정도를 수치화하여 그에 대응하는 맵으로 만든 것이다. 본 연구에서는 수식 (1)과 같이, 세일리언시 맵의 값을 가중치로 사용하여 세일리언시 맵 픽셀들의 가중치 무게 중심 c를 구하는 방법을 제안한다.

$$c = \frac{\sum_{p \in I} s(p) \times p}{\sum_{p \in I} s(p)} \tag{1}$$

수식 (1)에서 p는 입력 이미지 I의 픽셀 좌표를 나타내고 s(p)는 픽셀 p의 세일리언시 값을 나타내며 c는 세일리언시 맵의 가중치 무게 중심에 해당한다. 얼굴 검출 기반의 크롭핑과 동일한 방법으로, 크롭핑 박스의 중점 후보들 중 c와의 거리가 가장 가까운 후보를 찾아 크롭핑을 수행함으로써 관심 영역을 포함하는 익스트림 클로즈업 샷을 구한다.

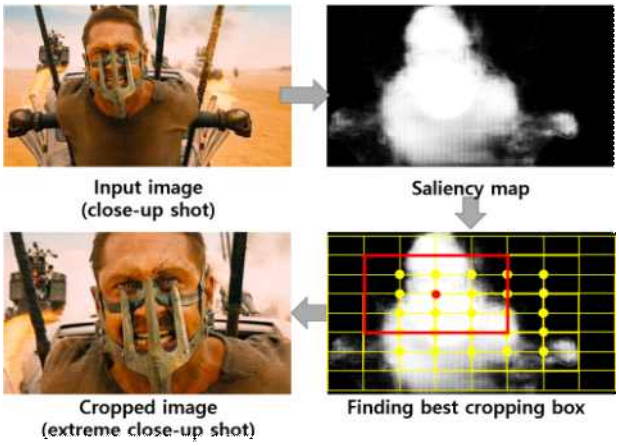

그림 5. 세일리언시 기반의 익스트림 클로즈업 샷 생성
Fig. 5. Extreme close-up shot generation based on saliency detection

3. 샷 사이즈 유형별 분류 모델 구축

본 연구의 최종 목표는 샷 사이즈 유형별 샷 분류 모델 구축이다. 그림 6은 본 연구의 선행연구에서 수행한 샷 경계 검출[33]과 샷 사이즈 분류 시스템[34]의 개요를 보여주며 아래와 같이 크게 두 단계로 나누어 설명할 수 있다.

1단계) 비디오 스토리텔링에서 숏전환 부분 경계 이미지 검출
2단계) a) 검출된 이미지에서 얼굴 탐색 바운딩 박스 처리(세일리언시 적용)
b) 익스트림 숏이미지의 가공과 숏사이즈 추정, 학습 및 분류

앞서 얼굴 및 세일리언시 검출을 통해 생성한 익스트림 클로즈업 샷은 크롭핑 과정을 통해 이미지의 크기가 줄어들었기 때문에, 샷 사이즈 유형별 분류를 위한 모델을 생성하기 위해서는 다른 샷 사이즈 데이터 셋의 이미지들과 그

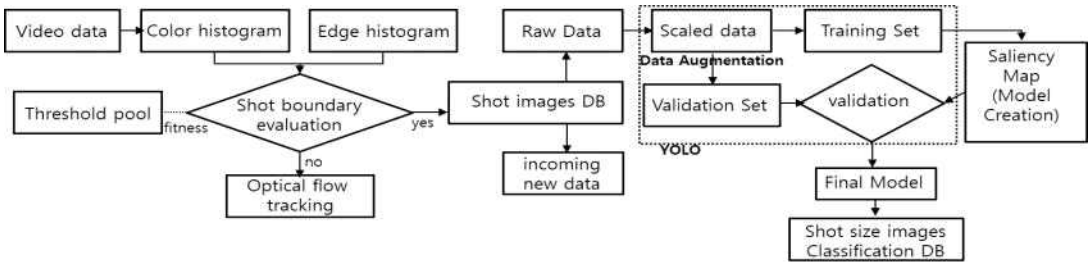

그림 6. 샷 경계 검출, 데이터 확장, 샷 사이즈 유형 분류 흐름도
Fig. 6. Shot boundary detection, data expansion, shot size type classification flowchart

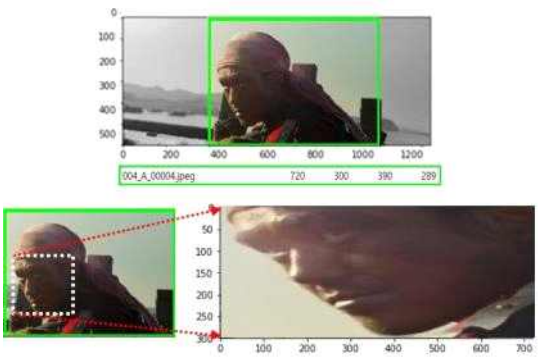

그림 7. 입력이미지 해상도 조정하기
Fig. 7. Adjusting input image resolution

표 1. ImageDataGenerator 파라메터
Table 1. ImageDataGenerator parameter

```
from keras.preprocessing.image import ImageDataGenerator
datagen = ImageDataGenerator(
        rotation_range= 0.8,
        width_shift_range= 0.8,
        height_shift_range= 0.8,
        rescale=1./(0~255),
        shear_range= 0.8, zoom_range= 0.8,
        horizontal_flip= True, fill_mode='nearest'
```

크기를 일치시키는 과정이 필요하다. 본 연구에서는 데이터를 회전하거나 확대를 하는 등에 사용되는 Keras API의 ImageDataGenerator를 사용하여 이를 수행하였다. 표 1은 해당 함수의 사용 예를 보여주며, 이를 이용해 그림 7과 같이 720*480으로 데이터 셋을 구성하는 이미지들의 크기를 일괄 조절하였다. 3.1과 3.2절에서 생성한 익스트림 클로즈업 샷 이미지가 추가된 데이터 셋을 활용하여 분류 모델을 만들기 위해 CNN을 이용한 학습이 필요하다. 본 연구에서는 인공신경망의 구현에서 널리 사용되고 있는 API인 Keras를 사용하였다. Keras API에서는 직접 데이터를 가공, 연결을 위해 Lambda layer와 Concatenate layer를 사용한다.

1) Lambda layer를 사용하면 신경망에 새로운 계층을 추가 수정할 수 있어 입력값과 출력값의 형태 변경이 가능하다[35].
2) Concatenate layer를 사용하면 여러 개의 레이어를 하나의 레이어로 만들 수 있어 분리된 두 모델을 연결할 수 있다[36].

Ⅳ. 실험 결과

아래의 표 2와 표 3은 Keras에서 제공하는 Lambda와 Concatenate의 파라메타 사용 방법을 설명하며 표 4는 Keras를 이용한 구현 예제를 보여준다.

표 2. Lambda layer 파라메터
Table 2. Lambda layer parameter

```
keras.layers.Lambda(function, output_shape=None,
mask=None, arguments=None)
```

표 3. Concatenate layer 파라메터
Table 3. Concatenate layer parameter

```
keras.layers.Concatenate(axis=-1)
```

표 4. Lambda layer 와 Concatenate layer의 프로그램 구현
Table 4. Program implementation of Lambda layer and Concatenate layer

```
def slice_4(t):
        return t(:, 1, :, :]

        slice_layer1 = Lambda(slice_1)
        slice_layer2 = Lambda(slice_2)
        slice_layer3 = Lambda(slice_3)
        slice_layer4 = Lambda(slice_4)

        features = []
        for k1 in range(w):
          features1 = slice_layer1(cnn_features)
          cnn_features = slice_layer2(cnn_features)
          for k2 in range(h):
            features2 = slice_layer3(features1)
            features1 = slice_layer4(features2)
            features.append(features2)

          relations = []
            concat = Concatenate()
                for features1 in features:
```

본 연구의 실험에서는 샷 사이즈 유형별 5개 데이터 셋을 학습에 사용하였다. 이때, 각 데이터 셋은 500장의 샷 대표 이미지들로 구성되었다. 본 연구에서는 5개의 샷 유형에 대한 분류를 목표로 하였으므로 최종 계층의 출력수를 5로 설정하였고, 오차를 최소화하기 위해 Adam 최적화를 사용하였다. 네트워크 가중치 학습율에 사용된 값은 0.001, 세대 수의 값은 20이 사용되었다. 이와 같은 학습을 통해 생성된 샷 사이즈 유형별 분류기에 대해서 10겹의 교차 검증을 통해 분류 정확도를 측정하였고 약 86.8%의 정확도를 얻을 수 있었다. 이를 통해, 본 논문에서 제안한 클로즈업 샷의 크롭핑을 통해 생성한 익스트림 클로즈업 샷 데이터 셋이 인공신경망의 학습 데이터로써 효과적으로 사용될 수 있음을 검증하였다.

V. 결 론

본 연구는 영화의 영상 데이터로부터 샷 사이즈 유형별 분류를 위해 부족한 데이터 셋을 확장하는 방안의 연구를 진행하였다. 하지만 샷 사이즈 분류는 최소 익스트림 클로즈업, 클로즈업, 미들, 풀, 롱으로 분류해야 본 연구의 궁극적 목적인 샷 사이즈 분류를 통한 클라이맥스 패턴 연구가 가능하다. 이를 위해 영화의 샷 이미지 데이터를 구성을 위한 클로즈업 샷에서 얼굴 위주의 관심영역을 잘라내어 익스트림 클로즈업 샷에 대한 데이터 셋을 얻는 방법을 제안하였다. 특히 샷 이미지 내에서 얼굴 검출에 실패한 경우 세일리언시 기반의 크롭핑을 사용함으로써 클로즈업 샷 내에서의 바운딩 박스 검출을 효과적으로 보완할 수 있는 결과를 얻었다.

향후 연구 중 하나는 익스트림 롱 샷 이미지 데이터 셋을 획득하는 것이다. 익스트림 클로즈업 샷과 마찬가지로 익스트림 롱 샷 이미지도 비디오 스토리에서 거의 나타나지 않으므로 분류된 샷을 기반으로 비디오 스토리를 분석할 때 데이터 셋을 보강하는 것은 중요하다. 이 문제를 해결하기 위해, 본 연구에서 롱 샷 이미지를 보충하기 위해 사용했던 웹기반 이미지 크롤링을 활용할 예정이며, 롱 샷과 익스트림 롱 샷을 구분하여 수집할 수 있는 방안을 연구할 계획이다.

본 연구에서 사용한 모델의 분류 정확도를 높이는 것 역시 향후 연구에서 수행되어져야 한다. 영상 내 오브젝트들의 종류, 크기, 위치 등을 분석함으로써 샷 사이즈 유형별 분류의 정확도를 높일 수 있을 것으로 기대된다. 이를 위해 YOLO[37] 등의 연구에서 사용되어진 방법들을 본 연구에 접목하는 것이 필요할 것으로 예측된다.

참 고 문 헌 (References)

[1] E. Chu and D. Roy, “Audio-Visual Sentiment Analysis for Learning Emotional Arcs in Movies,” in published 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, Nov. 2017, https://doi.org/10.1109/ICDM.2017.100

[2] L. Itti, and C. Koch, “Computational modelling of visual attention,” Nature reviews neuroscience, Vol. 2, No. 3, pp. 194-203, 2001, https://doi.org/10.1038/35058500

[3] K. Duncan, and S. Sarkar, “Saliency in images and video: a brief survey,” IET Computer Vision, Vol. 6, No.6 pp. 514-523, 2012, https://doi.org/10.1049/iet-cvi.2012.0032

[4] Y.M. Lim, “The Climax Expression Analysis Based on the Shot-list Data of Movies,” Journal of The Korean Society Of Broad Engineers, Vol. 21, No. 6, pp. 965-976, November 2016, https://doi.org/10.5909/JBE.2016.21.6.965

[5] W. Zhao, R. Chellappa and P.J. Phillips, "A. Rosenfeld Face recognition: a literature survey," ACM Comput. Surv. (CSUR), Vol.35, No.4, pp. 399-458, 2003, https://doi.org/10.1145/954339.954342

[6] Z. Kalal, K. Mikolajczyk and J. Matas, "Face-tld: tracking-learning-detection applied to faces," 17th IEEE International Conference on Image Processing (ICIP), pp. 3789-3792, 2010, https://doi.org/10.1109/ICIP.2010.5653525

[7] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol.22 No.12, pp. 1424-1445, 2000, https://doi.ieeecomputersociety.org/10.1109/34.895976

[8] N. Kumar and A.C. Berg, P.N. Belhumeur and S.K. Nayar, "Attribute and simile classifiers for face verification," IEEE 12th International Conference on Computer Vision, pp. 365-372, 2009, https://doi.org/10.1109/ICCV.2009.5459250

[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187-194, 1999, https://doi.org/10.1145/311535.311556

[10] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg and S.M. Seitz, "Exploring photobios," ACM Transactions on Graphics (TOG), Vol. 30, No. 61, 2011, https://doi.org/10.1145/2010324.1964956

[11] S. J. McKenna, S. Gong, and Y. Raja, " Modelling facial colour and identity with Gaussian mixtures," Pattern Recognition, April, 1998, https://doi.org/10.1016/S0031-3203(98)00066-1

[12] X.-G. Lv, J. Zhou, and C.-S. Zhang, "A novel algorithm for rotated human face detection," in IEEE Conferenceon Computer Vision and Pattern Recognition, 2000, https://doi.org/10.1109/CVPR.2000.855897

[13] J. Wang and T. Tan, "A new face detection method based on shape information," Journal of Pattern Recognition Letters, Vol.21, Issue 6-7, pp. 463 – 471, 2000, https://doi.org/10.1016/S0167-8655(00)00008-8

[14] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," in Proceedings of 1st International Conference on Computer Vision, London, 1987, https://doi.org/10.1007/BF00133570

[15] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic tracking, coding and reconstruction of human faces, using flexible appearance models," IEEE Electron. Letters, Vol.30, pp.1578 – 1579, 1994, https://doi.org/10.1049/el:19941110

[16] K. C. Yow and R. Cipolla, "Feature-based human face detection," Image Vision Comput. Vol.15, No.9, 1997, https://doi.org/10.1016/S0262-8856(97)00003-6

[17] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991, https://doi.org/10.1162/jocn.1991.3.1.71

[18] M. Tanaka, K. Hotta, T. Kurita, and T. Mishima, "Dynamic attention map by ising model for human face detection," in Proceedings of International Conference on Pattern Recognition, 1998, https://doi.org/10.1109/ICPR.1998.711870

[19] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp.:203-208, 1998, https://doi.org/10.1109/34.655647

[20] H. A. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in Proceedings of .IEEE International Conference on Computer Vision and Pattern Recognition, pp. 38 – 44, 1998, https://doi.org/10.1109/CVPR.1998.698585

[21] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3666-3673, 2013, https://doi.org/10.1109/CVPR.2013.470

[22] J. Li, Y. Zhang, "Learning surf cascade for fast and accurate object detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3468 – 3475, 2013, https://doi.org/10.1109/CVPR.2013.445

[23] E. Osuna, R. Freund and F. Girosi, "Training support vector machines: An application to face detection," in Proceedings of CVPR, 1997, https://doi.org/10.1109/CVPR.1997.609310

[24] C. Zhang, Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1036-1041. 2014, https://doi.org/10.1109/WACV.2014.6835990

[25] Y. Bengio, Y. LeCun, and G. Hinton, "Deep Learning," Nature, 521, (7553): 436 – 444, 2015, https://doi.org/10.1038/nature14539

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks,". IEEE Signal Processing Letters, Vol. 23, No. 10, pp. 1499-1503, 2016, https://doi.org/10.1109/LSP.2016.2603342

[27] A. Borji, M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," IEEE Transactions on Image Processing, Vol. 24, No. 12, pp. 5706-5722, 2015, https://doi.org/10.1109/TIP.2015.2487833

[28] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11, pp. 1254-1259, 1998, https://doi.ieeecomputersociety.org/10.1109/34.730558

[29] S. Mahamud, L.R. Williams, K.K. Thornber, K. Xu, "Segmentation of multiple salient closed contours from real images," IEEE Trans. Pattern Anal. Mach. Intell., Vol.25, No.4, pp. 433 – 444. 2003, https://doi.org/10.1109/TPAMI.2003.1190570

[30] D. Gao, S. Han and N. Vasconcelos "Discriminant saliency, the detection of suspicious coincidences and applications to visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., Vol. 31, pp. 989 – 1005, 2009, https://doi.org/10.1109/TPAMI.2009.27

[31] V. Gopalakrishnan, Y. Hu and D. Rajan,"Salient region detection by modeling distributions of color and orientation," IEEE Trans. Multimedia, Vol. 11, pp. 892 – 905, 2009, https://doi.org/10.1109/TMM.2009.2021726

[32] N. Liu, J. Han, and M. H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3089-3098, 2018, https://doi.org/10.1109/CVPR.2018.00326

[33] S.R. Park and Y.M. Lim, "The Implementing a Color, Edge, Optical Flow based on Mixed Algorithm for Shot Boundary Improvement," Journal of Korea Multimedia Society, Vol. 21, No. 8, pp. 829-836, August 2018, https://doi.org/10.9717/kmms.2018.21.8.829

[34] S.R. Park, J.E. Eom and Y.M. Lim, "The System Design and Implementation for Detecting the Types of Shot Size," Proceeding of Conference Korea Multimedia Society, Vol. 21, No. 1, pp. 968-996,

Seoul, Korea, May 2018.
[35] Keras Documentaion. https://keras.io/layers/core/#lambda. (accessed Sept. 07, 2019)
[36] Keras Documentaion. https://keras.io/layers/merge/. (accessed Sept. 07, 2019)
[37] J. Redmon and S. Divvala et al., "You only look once: Unified, real-time object detection." IEEE CVPR, pp. 779-788, 2016, https://doi.org/10.1109/CVPR.2016.91

저 자 소 개

강 동 완

- 2006년 2월 : 중앙대학교 컴퓨터공학과 공학사
- 2013년 2월 : 중앙대학교 컴퓨터공학과 공학박사
- 2015년 7월 ~ 2017년 8월 : 영국 Bournemouth University Visiting Researcher
- 2018년 2월 ~ 2018년 8월 : 영국 Bournemouth University Marie Curie Fellow
- 2018년 9월 ~ 현재 : 서울과학기술대학교 컴퓨터공학과 조교수
- ORCID : https://orcid.org/0000-0001-7210-4595
- 주관심분야 : 컴퓨터그래픽스, 영상처리, 감성컴퓨팅

임 양 미

- 1998년 3월 : 큐슈대학교 정보전달학과 석사
- 2009년 2월 : 중앙대학교 첨단영상대학원 박사
- 2010년 ~ 현재 : 덕성여자대학교 IT미디어학과 부교수
- 2016년 ~ 현재 : 한국방송·미디어공학회 논문지편집위원
- 2015년 ~ 현재 : 한국멀티미디어학회 이사
- ORCID : https://orcid.org/0000-0002-3725-0025
- 주관심분야 : 멀티미디어, 인터랙티브아트, UX/UI