

영상 구성 파라미터 추출을 위한 융합 분석 알고리즘 연구

맹 채 정 *

하 동 환 **

* 중앙대학교 창의ICT공과대학 연구원

** 중앙대학교 소프트웨어대학 교수

Convergence Analysis Algorithm Study for Extracting Image Configuration Parameters

Maeng, Chae Jung *

Har, Dong-Hwan **

* Researcher, College of ICT ,Chung-Ang University

** Professor, College of Software ,Chung-Ang University

이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A5A2A01031624)

** Corresponding Author : Har, Dong Hwan, dhhar@cau.ac.kr

THE KOREAN SOCIETY OF SCIENCE & ART

한국과학예술융합학회

THE KOREAN SOCIETY OF SCIENCE & ART Vol.37(3)_Regular article or full paper

* Contribution : 2019.05.13_Examination : 2019.06.10_Revision : 2019.06.25_Publication decision : 2019.06.30

목차

Abstract

국문초록

I. 서론

1.1 연구배경 및 목적

2.1 연구방법

II. 본론

2.1 프로그램 개발

2.2 실험결과

III. 결론

Reference

Endnote

Abstract

This study was conducted to organize a program to classify and analyze the characteristics of images for the automation of background music selection in the video content production process. The results and contents of the study are as follows: video characteristics are selected as subject category, emotion, pixel motion speed, color, and character material. Subject categories and feelings were extracted using Microsoft's Azure Video Indexer, Pixel Movement Speed was an Optional flow, Color was an Image Histogram for Image, and character materials was CNN(Convolutional Neural Network). The results of this study are significant in that video analysis was conducted to match background music in the recent content production process of 'Internet One-person Broadcasting Creators'.

국문초록

본 연구는 영상콘텐츠 제작과정에서 배경음악 선정의 자동화를 위하여 영상의 특성을 분류, 분석할 수 있는 프로그램을 구성하였다. 연구 결과 및 내용은 다음과 같다. 영상의 특성은 '주제 범주', '감정', '픽셀 움직임 속도', '색상', '등장인물'로 선정하며, '주제 범주'와 '감정'은 Microsoft사의 Azure Video Indexer를, '픽셀 움직임 속도'는 Optical flow, '색상'은 Image Histogram, '등장인물'은 CNN (Convolutional Neural Network)을 활용하여 데이터를 추출하였다. 이러한 본 연구의 결과는 최근 주목을 받고있는 '인터넷 1인 방송 크리에이터'들의 콘텐츠 제작과정에서 배경음악 매칭을 위한 영상 특성 분석이 이루어졌다는 점에서

의의가 있다.

Key Words

Internet One-person Broadcasting Creators(인터넷 1인 방송 크리에이터), Optical Flow(옵티컬 플로우), Image Histogram(이미지 히스토그램), Convolutional Neural Network(컨벌루션 뉴럴 네트워크), Convergence Content(융합콘텐츠)

I. 서론

1.1 연구배경 및 목적

최근 영상 공유 플랫폼 발달과 영상 제작 기술의 대중화로 어느 때보다 미디어 분야에서 영상이 미치는 영향력이 높아졌다. 전 세계 13억 인구가 사용하는 유튜브(YouTube)에서부터 젊은이들에게 큰 인기를 얻고 있는 뮤직비디오 플랫폼 틱톡(TikTok)까지, 영상을 직접 만들고 공유하는 것은 이미 대중들의 익숙한 소통 수단이 된지 오래이다.

이처럼 영상 제작 경험이 보편화되고 있는 요즘, 영상에 알맞은 배경 음악을 찾아 넣는 것은 영상 제작에 있어 매우 중요한 요소이다. 영상의 내용과 적절히 어우러지는 배경 음악은 영상이 전하고자 하는 메시지를 더욱 분명하게 하거나 혹은 강조할 수 있기 때문이다. 하지만 전문적으로 음악이나 음향관련 학습을 하지 않은 대부분의 일반 영상 제작자들은 본인의 주관적 느낌에 따라 배경음악을 선정하는 경우가 많고, 결과적으로 적절치 못한 영상과 배경 음악의 조화로 인해 영상제작물의 수준 저하가 야기되는 경우가 많다. 이러한 문제에 근거하여 일반 사용자들이 본인이 제작한 영상에 맞는 적절한 배경음악(OST)을 자동적으로 추천 또는 매칭 시켜주는 프로그램을 개발하는 것이 본 연구의 출발점이다.

영상과 음악을 매칭 시키기 위해서는 영상 콘텐츠의 요소별 특징을 규정할 수 있어야 하며 이 규정들은 음악의 구성 요소들 각각과 관계 파악이 가능하도록 가공될 수 있어야 한다. 본 연구는 음악의 구성 요소들과 매칭 시킬 수 있는 영상 요소 다섯 가지를 선정하고 영상 분석 및 데이터를 추출하여 인덱싱할 수 있는 프로그램 구성을 목적으로 한다. 머신러닝 기술과 영상처리 기술의 발달은 영상 콘텐츠의 요소별 특징 분석, 인덱싱을 위하여 필수적인 도구가 되며 다음과 같은 다섯 가지 기술적 화두가 있다.

첫째, 의미기반 비디오 온톨로지 데이터(semantic video ontology data)를 활용한 검색 시스템에 대한

연구이다. 비디오 장면의 내용에 대한 키워드를 트리 구조로 저장한 장면이름 온톨로지와 저수준 정보인 색상, 모양, 재질과 고수준 정보인 객체, 이벤트의 의미적 차이를 극복하며, 장면을 정의하는 단어의 의미를 분석하여 모양은 다르지만 의미적으로 근접한 장면들을 검색하는 방식이다¹⁾. 이는 온톨로지 데이터를 다량으로 구축해야 자동분석이 가능하다는 점에서 일반 1인 크리에이터가 사용하기에 한계가 있다.

두 번째는 키프레임 추출 및 색인으로 영상을 검색하는 시스템이다. 배경차분법, Canny Edge, Optical Flow, 사람 의상 색상 정보 추출 등의 객체 인식 기법으로 키프레임을 추출하고, 사람과 관련된 주요 정보를 자동으로 색인 및 저장하여 시간, 목적, 장소, 사건 종류별 검색이 가능하며, 그 결과를 XML 문서로 제공한다²⁾. 이러한 결과는 영상 콘텐츠의 맥락 분석으로 세부 주제를 확인할 수 있으며, 음악 요소들과 매칭하기 위한 ‘감정’, ‘색상’, ‘픽셀 움직임 속도’, 등의 요소 확인에 유용하게 활용할 수 있다.

세 번째는 질의와 RGB 히스토그램을 활용한 기법들을 통한 영상 인덱싱 에이전트 설계이다. 자동 주석 처리 기법은 사용자 질의로 키워드를 추출하고 키워드들에 대한 의미 가중치 계산으로 키프레임 주석 정보를 수집한다. 키프레임 특징처리 기법은 컬러 히스토그램을 사용하여 키프레임 검출 후 픽셀 그룹에 대한 RGB 평균값 계산을 통해 키프레임 유사도를 구한다³⁾. 이는 질의 횟수가 높아질수록 검색 정확도가 높아지지만 주석 정보 수집을 사용자 질의에 의존한다는 점에서 대용량 인덱싱 데이터 수집에 어려움이 있다. 또한 사용자 질의 정보의 주관적 측면으로 높은 정확도와 신뢰도를 얻기에 한계가 있다.

네 번째는 비디오 클립과 스크립트를 동시에 이용한 멀티모달(multimodal) 방법과 텍스트 마이닝(text mining) 기반의 뉴스 비디오 마이닝 시스템(news video mining system)이다. 텍스트 마이닝을 통한 군집 분석 기법은 뉴스 비디오를 주제별로 분류하고 동향과 군집의 성장패턴, 군집 간 상호 연관성을 분석하여 뉴스기사 내용 변화와 같은 잠재적 지식을 도출한다⁴⁾. 정보 전달이 목적인 뉴스 기사와 달리 정해진 형식이 없는 1인 크리에이터들의 영상은 ‘주제 범주’ 외로 다양한 요소로 해석될 수 있다. 이것은 음악 요소들과 매칭을 위하여 ‘주제 범주’, ‘픽셀 움직임 속도’, ‘색상’, ‘감정’, ‘등장인물(성별과 나이)’ 등의 데이터를 수집함으로써 분석 요소를 다각화 할 수 있다.

다섯 번째는 주석기반 및 내용기반 메타데이터 통합비디오 인덱싱 기법이다. 인덱싱 정보에는 촬영기법, 사진 제작지식과 비디오 몽타주로 추출된 주석기

반 및 영상소재에 의한 내용기반 특징 등이 있다. 그 예로 패턴의 변화, 색상, 질감, 개체 모양, 개체간 공간상 위치관계, 의미 정보가 있다. 다수의 특징들을 분석하여 비디오 인덱싱을 하는 것은 유사하지만 촬영기법, 사진 제작지식과 같은 객관적인 요소를 추가할 수 있으므로 분류의 정확성을 높일 수 있다. 그러나 질의 프로그램을 이용하여 데이터를 수집하는 방식을 사용함으로써 대용량 데이터베이스가 구축되어야 하는 단점이 있다⁵⁾.

이와 같은 연구를 바탕으로 영상의 주요 분석 요소를 ‘주제 범주’, ‘감정’, ‘픽셀 움직임 속도’, ‘색상’, ‘등장인물’ 5가지로 선정하며, ‘주제 범주’와 ‘감정’은 Microsoft사의 Azure Video Indexer를, ‘픽셀 움직임 속도’는 Optical flow, ‘색상’은 Image Histogram, ‘등장인물’은 CNN(Convolutional Neural Network)을 가공 및 활용하여 데이터를 추출하고자 한다.

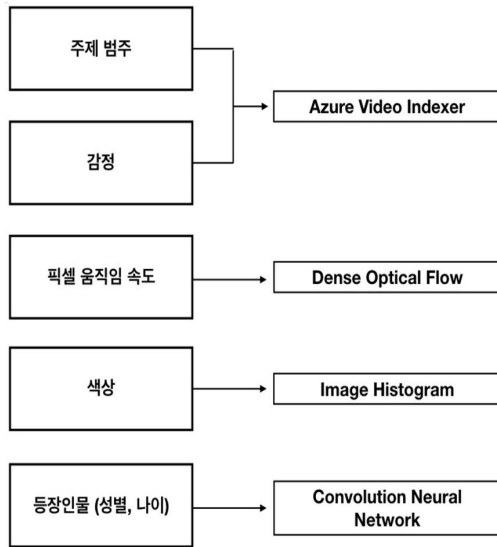
1.2 연구방법

II. 본론

2.1 영상 구성 파라미터 추출 프로그램 개발

(1) Azure Video Indexer를 이용한 영상 구성 파라미터 추출

본 연구에서는 영상 평가의 구성요소를 ‘주제 범주’, ‘감정’, ‘픽셀 움직임 속도’, ‘색상’, ‘등장인물’로 정하였다. <그림-01>은 영상 요소 데이터 추출 프로그램의 전체적인 블록 다이어그램을 보여준다. 본 연구에서 ‘주제 범주’, ‘감정’ 분석을 위해 사용한 Microsoft사의 Azure Video Indexer는 미디어 분석 클라우드 애플리케이션으로 오디오 전사, 화자 통계, OCR, 감정 분석, 얼굴 감지, 키워드 추출, 주제 유추 등 분석하려는 영상에 대해 27가지 분석결과를 제공한다.



<그림-01> 영상 5가지 요소 데이터 추출 과정

‘주제 범주’는 키워드 추출, 주제 유추, 레이블 식별 기능을 이용해 Binary 형태로 추출하였다. 키워드는 음성과 시각적 텍스트를 분석하여 추출하였고, 주제는 1st-level IPTC taxonomy가 포함된 대본에서 유추하였으며 레이블은 시각적 객체와 행동을 인식하여 식별해 내었다.

‘감정’은 Emotion detection과 Sentiment analysis 기능으로 Binary 데이터로 추출하였다. Emotion은 joy, sadness, anger, fear로 나뉘어 음성 및 오디오 신호로부터 인식하였고 Sentiment는 positive, negative, neutral로 나뉘어 음성 및 시각적 텍스트를 통해 식별해 내었다.

본 연구에 맞는 프로그램을 구축하기 위하여 Azure Video Indexer로 추출한 비디오 분석 결과를 JSON(Java Script Object Notation) 형식으로 받아온 후 JavaCC(Java Compiler Compiler)로 파싱하여 CSV(comma-separated values) 파일로 변환하였다. JSON은 ‘속성-값’ 쌍 또는 ‘키-값’ 쌍으로 이루어진 데이터 오브젝트를 전달하기 위한 개방형 표준 포맷이며⁶⁾, JavaCC는 parser generator 기능과 lexical analyzer generator 기능을 수행하며 EBNF 표기법으로 작성된 형식 문법의 Parser를 생성한다⁷⁾. [표-01]은 Azure Video Indexer로 추출한 ‘주제 범주’와 ‘감정’ 데이터 결과를 정리한 것으로 사용한 얼굴 이미지 데이터는 Flickr 웹사이트에서 연령, 성별이 레이블된 LMDB(Lightning Memory-Mapped Database) 형태의 사진 26,580장을 임의로 선정하였다⁸⁾.

[표-01] Azure Video Indexer를 이용한 ‘주제 범주’, ‘감정’ 데이터 추출 결과

| Video 번호 | Topic | Sentiment | | | Emotion | |
|-------------|----------------------------------|-----------|--------|-------------------|---------|-------------------|
| | | type | score | duration ratio | type | duration ratio |
| 1 | Sport | Positive | 0.9844 | 0.0351 | Joy | 0.0351 |
| | | Neutral | 0.5 | 0.9649 | | |
| 2 | Society | Positive | 0.9579 | 0.0744 | Joy | 0.0744 |
| | | Neutral | 0.5 | 0.8189 | Sad | 0.0844 |
| | | Negative | 0.1552 | 0.1066 | Fear | 0.0222 |
| 3 | Lifestyle, Leisure | Neutral | 0.5 | 0.9194 | Joy | 0.068 |
| | | Positive | 0.9518 | 0.068 | | |
| | | Negative | 0.1522 | 0.0125 | Sad | 0.0125 |
| 4 | Economy, Business, Finance | Positive | 0.9569 | 0.0476 | Joy | 0.0476 |
| | | Neutral | 0.5 | 0.9523 | | |

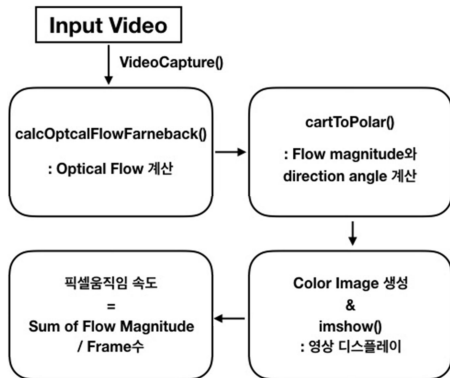
‘픽셀 움직임 속도’와 ‘색상’은 각각 Dense optical flow와 Image Histogram을 이용해 데이터를 추출하였다. Dense optical flow는 Gunner Farneback’s algorithm을 기반으로 하며, 이는 Polynomial Expansion 바탕의 연속된 두 프레임 간 움직임 측정 알고리즘이다. Optical flow는 영상에서 물체나 표면, 엣지의 움직임 패턴을 의미하며 물체 또는 카메라의 움직임 때문에 발생하는데 이것은 이미지 밝기 패턴의 움직임 속도 분포로 해석할 수 있다⁹⁾. Image Histogram은 Gray Level을 갖는 각 픽셀의 수 혹은 총 픽셀 수에 대한 비율을 표시한 함수로 픽셀 값들을 막대그래프 혹은 직선그래프로 표시한다¹⁰⁾.

‘등장인물’ 분석은 CNN(Convolutional Neural Network)을 이용하여 이미지 데이터를 학습시킨 딥러닝 모델을 통해 영상 프레임별로 얼굴 검출 후 성별과 나이를 추정하였다. CNN은 MLP(Multilayer perceptrons)를 사용하여 계산과 특징 학습을 진행하고 이미지를 분류하는 역할을 수행하였다.

(2) Optical Flow를 이용한 영상 픽셀 움직임 속도 분석 알고리즘

<그림-02>는 Dense optical flow 알고리즘을 사용한 ‘픽셀 움직임 속도’ 데이터 추출 알고리즘 블록도이다. 영상을 구성하는 모든 픽셀들에 대한 Optical flow는 calcOpticalFlowFarneback 함수를 이용하여 계산하며 각 이미지에 대한 pyramid들을 생성하였다.

그리고 `cartToPolar()` 함수를 이용하여 2D 벡터들의 flow magnitude와 direction angle을 cartesian에서 polar로 변환하여 계산하였다. flow magnitude는 픽셀이 움직인 거리를 의미하고 direction angle은 픽셀이 움직인 방향을 의미한다. Flow magnitude와 direction angle 값은 각각 value(밝기 정도)와 hue(색깔)로 지정하여 Color Image 영상을 생성하였다.







<그림-02> ‘픽셀 움직임 속도’ 데이터 추출 알고리즘

`cvtColor()` 함수를 이용하여 `COLOR_HSV2BGR` 형식으로 바꾼 영상은 `imshow()` 함수로 Optical flow 결과를 재생한다. Flow magnitude는 픽셀이 움직인 거리이므로 전체 합계는 영상 내 픽셀 움직임 거리를 나타낸다. 각 Frame의 Flow magnitude 배열 평균값을 `numpy.mean()` 함수로 계산하고 그 합을 전체 Frame 수로 나누었다. 도출된 각 Frame의 Flow magnitude 평균이며 ‘픽셀 움직임 속도’에 대한 비교 수치를 구하였다. [표-02]는 실험 영상들에 대한 분석 결과로 Flow Magnitude 평균 및 합계와 Frame 수, Color Image Video의 캡처된 이미지를 보여준다.

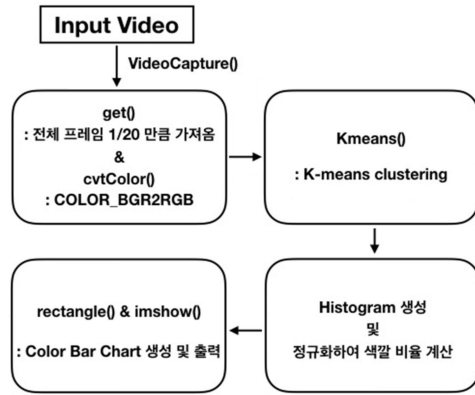
(3) Image Histogram을 통한 영상 ‘색상’ 데이터 추출 알고리즘

<그림-03>은 입력되는 영상에 대하여 알고리즘을 적용하여 Image Histogram의 픽셀값 분포 계산과 영상의 ‘색상’ 데이터를 추출과정을 보여준다. 이 과정은 `VideoCapture()` 함수로 영상을 연속적으로 캡처하는 과정으로 시작하며, `get()` 함수로 전체 프레임의 1/20을 가져와 `cvtColor()` 함수를 적용하여 색상 순서를 BGR에서 RGB로 바꾸어준다. 이 이미지들은 `Kmeans()` 함수로 K-means clustering을 하여 5개 clusters를 만들어 학습시켰다.

[표-02] ‘픽셀 움직임 속도’ 데이터 추출 결과 및 Color Image Video 캡처화면

| Video 번호 | Flow Magnitude | | Frame 수 | Dense Optical Flow Color Image Video |
|-------------|-------------------|---------|------------|---|
| | 평균 | 합계 | | |
| 1 | 1.546 | 18778.7 | 1214 5 |  |
| 2 | 1.850 | 32320.2 | 1746 6 |  |
| 3 | 2.002 | 53628.0 | 2677 7 |  |
| 4 | 0.589 0 | 4302.74 | 7304 |  |

K-means clustering이란 머신러닝의 비지도 학습 모델 중 하나로 k개의 데이터 평균들을 계산하여 data clustering을 수행하는 알고리즘이다¹¹⁾. 이것은 n개 관측치를 좌표에 점들로 나타내었을 때 점 사이 거리 합을 최소로 하는 k개의 중심점(centroid)들을 찾아 관측치를 k개 clusters로 분할한다. 이 결과들은 `fit()` 함수로 K-means clustering 알고리즘을 이미지들에 적용시켜 pixel lists를 clustering한다.



<그림-03> '색상' 데이터 추출 알고리즘 블록도

[표-03] Image Histogram을 이용한 '색상' 데이터 추출 결과

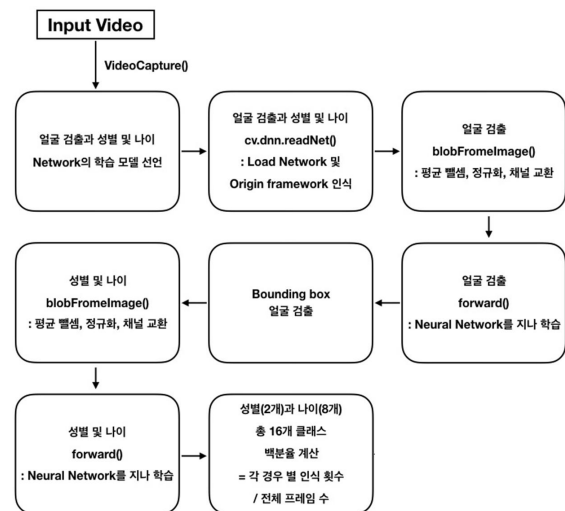
| Video 번호 | Ratio | R(Red) | G(Green) | B(Blue) |
|----------|--------|--------|----------|---------|
| 1 | 0.469 | 22.0 | 27.7 | 45.0 |
| | 0.088 | 141 | 150 | 161 |
| | 0.231 | 44.7 | 52.3 | 71.4 |
| | 0.100 | 86.0 | 95.5 | 93.4 |
| | 0.117 | 168 | 197 | 235 |
| 2 | 0.377 | 187 | 153 | 144 |
| | 0.169 | 229 | 224 | 240 |
| | 0.104 | 154 | 102 | 106 |
| | 0.026 | 89.1 | 24.6 | 37.4 |
| | 0.324 | 207 | 177 | 171 |
| 3 | 0.254 | 122 | 153 | 172 |
| | 0.113 | 26.9 | 32.3 | 31.0 |
| | 0.376 | 164 | 194 | 208 |
| | 0.0934 | 127 | 107 | 99.9 |
| | 0.165 | 68.2 | 80.9 | 54.3 |
| 4 | 0.107 | 161 | 130 | 127 |
| | 0.479 | 17.5 | 85.9 | 115 |
| | 0.186 | 87.1 | 247 | 246 |
| | 0.113 | 228 | 210 | 200 |
| | 0.115 | 110 | 42.8 | 49.6 |

다음은 각 컬러들이 전체에서 차지하는 분포비율을 계산하는 과정으로 이미지 내 pixel들을 각 clusters에 할당된 후 그 값에 대한 히스토그램을 생성하였다. 또한 np.arange() 함수로 clusters 개수에 따라 주어진 간격 내 균일하게 분포된 값들을 계산하여 라벨(label) 값을 계산하였다. 라벨 값들을 이용해 Numpy

의 histogram() 함수로 히스토그램을 생성한 후 astype() 함수로 배열 값의 합이 1이 되도록 정규화하여 비율을 계산하였다. 여기서 반환된 히스토그램 값은 각 clusters 간격에 차지하고 있는 각 컬러들의 비율을 의미한다. [표-03]은 실험영상들에 대한 RGB 값과 각 컬러에 대한 비율 분석결과를 보여준다.

히스토그램과 중심점들을 각 clusters에서 컬러-백분을 분포와 컬러로 지정하여 clusters의 상대적 비율을 cv2.rectangle()로 bar chart 형태로 보여준다. 각 cluster에 할당된 픽셀 개수를 계산하여 히스토그램 값에 할당하고, 그 값을 받아 각 컬러에 레이블한 픽셀 개수에 해당하는 숫자를 표현하는 bar 객체를 생성하였다. 마지막으로 Matplotlib의 imshow() 함수를 이용하여 bar를 시각화하여 [표-03]과 같은 Color Bar Chart 결과를 생성하였다.

(4) CNN을 이용한 얼굴 인식 및 성별과 나이 데이터 추출 알고리즘



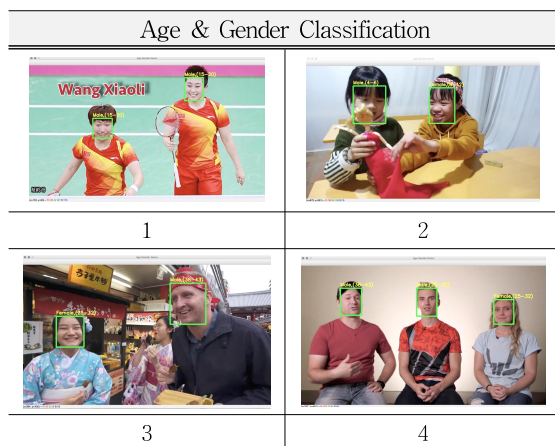
<그림-04> '등장인물' 성별과 나이 데이터 추출 알고리즘

<그림-04>는 성별과 나이 데이터 추출을 위한 CNN 알고리즘과 클래스별 백분율을 계산 과정이다. 성별 예측 모델은 분류 문제로 'Male'과 'Female' 두 클래스로 나타내는 Softmax 형태의 합성곱 신경망(CNN)을 사용하였다. 나이 예측 모델 또한 0에서 100 세까지를 8개 클래스로 나누는 분류하기 위하여 3개의 convolutional layers와 2개의 fully connected layers, 하나의 final output layer를 사용하는 뉴럴 네트워크를 이용하였다. 얼굴 검출 네트워크 학습 모델과 성별과 나이 네트워크 학습 모델에 대한 text 파일

과 훈련된 weight 값들은 binary 파일에 포함시켜 가져왔다. 그리고 cv.dnn.readNet() 함수를 이용하여 두 파일 내부에 origin framework를 인식하고 dnn의 Neral Network 모델을 읽어왔다¹²⁾.

얼굴 검출 네트워크에 입력되는 영상은 blobFromImage() 함수로 평균 땀샘, 정규화, 채널 교환 과정을 거치고 forward() 함수를 통하여 네트워크를 통하여 학습하였다. Output detection들은 [0, 0, i, 3]와 같은 4D matrix 형태이며 3번째 차원의 i는 영상 속에서 탐지된 얼굴들에 반복적으로 적용되는 횟수이고 4번째 차원은 bounding box와 각 얼굴 점수를 의미한다. bounding box의 결과 좌표에 원본 이미지 Width와 Height를 곱함으로써 0에서 1 사이로 정규화하여 해당 영상에 알맞은 좌표를 구하였다. 마찬가지로 성별과 나이 인식 네트워크를 지나는 입력 영상은 blobFromImage() 함수와 forward() 함수를 거쳐 학습된 후 결과 값을 반환하였다. 성별과 나이에 대한 분석 결과는 리스트 형태로 각 2개, 8개 클래스로 나누어져 결과 값을 저장한다. argmax() 함수를 이용하여 가장 큰 confidence 값을 가지는 인덱스를 구한 후 최종적으로 결정된 성별과 나이 범주를 출력하였다¹³⁾.

[표-04] CNN을 이용한 얼굴검출 및 성별, 나이 인식 과정



수집된 데이터를 성별이 'Female'일 경우와 'Male'일 경우로 나누고 나이별로 재차 분류하였다. 성별은 2개, 나이는 8개 클래스로 나누었으므로 총 경우의 수는 16가지가 된다. 프레임을 인식할 때마다 분류하여 경우별 인식 횟수를 계수하고 총 횟수로 나누어 백분율을 계산하였다. [표-04]는 각 영상의 얼굴검출 및 성별, 나이 인식 과정을 보여주며, [표-05]는 각각의 입력영상에 대한 인물들의 성별과 나이 분포 추출 결과를 보여준다.

[표-05] CNN을 이용한 '등장인물' 성별, 나이 데이터 추출 결과

| Video번호 | | 1 | | 2 | | 3 | | 4 | |
|----------|--------|--------|--------|---------|-------|--------|--------|--------|--------|
| Gender | | Female | Male | Female | Male | Female | Male | Female | Male |
| A | 0-2 | 0.3676 | 3.554 | 2.144 | 7.924 | 0.5435 | 1.492 | 0.0 | 0.0100 |
| | 4-6 | 0.2451 | 1.103 | 7.890 | 48.82 | 1.833 | 24.90 | 0.0 | 0.2301 |
| | 8-12 | 10.05 | 24.88 | 1.812 | 10.45 | 3.476 | 28.73 | 0.8403 | 17.42 |
| | 15-20 | 9.926 | 12.75 | 0.4973 | 5.758 | 1.795 | 1.125 | 0.2401 | 5.472 |
| | 25-32 | 3.186 | 29.17 | 4.000 | 6.244 | 3.957 | 11.50 | 32.18 | 40.24 |
| | 38-43 | 0.3676 | 1.225 | 0.06631 | 1.293 | 0.405 | 19.13 | 0.8103 | 2.561 |
| | 48-53 | 0.1225 | 0.6127 | 0.5304 | 2.332 | 0.4930 | 1.277 | 0.0904 | 6.433 |
| | 60-100 | 1.103 | 0.4902 | 0.1547 | 1.392 | 0.1390 | 0.4171 | 0.000 | 0.000 |
| Total(%) | | 25.58 | 74.41 | 16.87 | 83.13 | 12.49 | 87.51 | 32.07 | 67.93 |

2.2 실험결과

Microsoft Video Indexer에서 추출한 JSON 데이터를 CSV로 변환하여 '주제 범주'와 '감정'에 대한 데이터를 수집하였다. 세부 데이터 항목은 Topic, Sentiment type, Sentiment score, Sentiment duration ratio, Emotion type, Emotion duration ratio이며 이를 [표-01]로 정리하였다. Topic은 Sport, Lifestyle, Leisure, Society, Crime, Law, Justice, Economy, Business, Finance 등 해당 영상의 주제 범주 키워드 형태로 추출하였다. Sentiment type은 Positive, Neutral, Negative 세 가지로 분류되었으며 각 type의 score와 duration ratio 수치 데이터를 수집하였다. Emotion type은 Joy, Sad, Fear, Anger 4가지 type으로 나누어지며 해당 type의 duration ratio 값도 얻을 수 있었다.

'픽셀 움직임 속도'는 Dense Optical flow 알고리즘을 통해 얻은 flow magnitude 값의 합계를 Frame 수로 나눈 평균값으로 비교하였다. Dense Optical flow를 시각화한 Color Image 영상은 2D 벡터들의 flow magnitude와 direction angle을 계산하고 각각 value(밝기 정도)와 hue(색깔)로 지정하여 표현하였다. '색상'은 R(Red), G(Green), B(Blue) 수치 값으로 나타내었으며 전체 영상에서 각 색상이 차지하는 분포비율을 합이 1이 되도록 하여 구하였다. 색상과 분포비율에 대한 수치 데이터를 한눈에 볼 수 있게 하기 위하여 Color Bar Chart를 이용해 시각화하였다. '등장인물'에 대한 성별, 나이를 16가지 경우로 나누어 각 영

상에 등장하는 인물의 성별과 나이 인식 결과를 백분율로 계산하여 분포를 나타내었다. 인물 얼굴 위치는 bounding box로 나타내고 그 위에 유추한 성별과 나이를 레이블링하여 Text로 나타내었다.

III. 결론

비디오 스트리밍 기술과 디스플레이, 네트워크 등 영상을 제작하고 공유할 수 있는 기술의 발달로 영상은 하나의 보편적인 소통 수단이 되었다. 영상에서 음악의 역할은 장면의 의도와 전달하고자 하는 메시지를 극대화시킬 수 있는 도구로 사용될 수 있다. 그러나 영상에 맞는 음악을 선정하기 위한 과정은 1인 미디어 크리에이터들에게는 매우 번거롭고 어려운 일이 될 수 있다. 본 연구는 제작된 영상콘텐츠를 자동으로 분석하고 어울리는 음악과 매칭, 추천해 줄 수 있는 방안을 제시하기 위하여 시작하였다. 이를 위하여 첫 번째 세부 연구로, 영상 분석을 위한 5가지 요소를 ‘주제 범주’, ‘등장인물’, ‘감정’, ‘픽셀 움직임 속도’, ‘색상’으로 선정하고 데이터를 추출을 위하여 다음과 같은 분석방법을 제시하였다.

배경음악이 있는 7가지 영상에 대해 Microsoft Video Indexer로 ‘주제 범주’, ‘감정’ 데이터를 선별하여 추출하였다. Dense optical flow 알고리즘을 사용하여 연속된 두 프레임 간 optical flow 값을 구하고 Magnitude 평균값으로 ‘픽셀 움직임 속도’를 비교하였다. ‘색상’ 데이터는 영상을 같은 프레임 간격으로 캡처한 이미지를 K-means clustering한 후 히스토그램을 생성하여 각 색깔에 대한 분포비율로 구하였다. ‘등장인물’의 성별과 나이 데이터는 분류 문제로 접근하여 각각 2개, 8개 노드를 가진 Softmax 형태의 합성곱 신경망(CNN)을 사용하여 추출하였다. 영상의 프레임 인식할 때마다 입력되는 이미지는 네트워크를 통과해 반환된 값 중 가장 큰 confidence 값을 가지는 성별과 나이를 출력하였다.

본 연구를 통해 영상과 음악 매칭 프로그램을 위한 첫 단계인 영상 구성 파라미터들을 추출해 내었다. 음악의 분석 요소로는 음고, 음가, 음색, 화성, 조성, 빠르기 등이 있으며 추후 연구를 통해 영상 분석 요소들과의 연관 정도를 실험하고 자동 매칭 및 음악 추천 프로그램을 제작하고자 한다. 본 연구로 1인 미디어 크리에이터들의 음악 검색 폭을 넓히고 높은 품질의 영상 콘텐츠 생산을 유도하여 미디어 산업의 발전에 이바지하기를 기대하는 바이다.

Reference

- [1] Byeongcheol Kim, Changjin Kim, Seongcheol Yun, Kyungsook Han, “Implement Static Analysis Tool using JavaCC,” Journal of the Korea Society of Computer and Information 23(12), pp.89-94, 2018.
- [2] Byoungjun Kim, Joonwhoan Lee, “A Deep-Learning Based Model for Emotional Evaluation of Video Clips,” INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS 18(4), pp.245-253, 2018.
- [3] Edman Paes dos Anjos, “Experimental Evaluation of Succinct Representations of JSON Documents,” Seoul National University : Graduate School of Computer Science and Engineering, 2016.
- [4] Gil Levi, Tal Hassner, “Age and Gender Classification using Convolutional Neural Networks,” IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.34-42, 2015.
- [5] Hansung Lee, Younghee Im, Jaehak Yu, Seunggeun Oh, Daihee Park, “A News Video Mining based on Multi-modal Approach and Text Mining,” Journal of KIISE : Database 37(3), pp.127-136, 2010.
- [6] Ji-Hyun Cho, Ji-Won Yoon, “Design of Unstructured Big Data Matching System for Unidentified Aircraft Tracking Using Open CV,” Journal of Digital Forensics 10(2), pp.1-20 2016.
- [7] Joohyung Kang, Sooyeong Kwak, “Violent Behavior Detection using Motion Analysis in Surveillance Video,” Journal of broadcast engineering Volume.20 Number.3, pp.430-439, 2015.
- [8] JooSung Kim, KyongYeon Kim, HakIl Kim, Yoo-Sung Kim, “A Video Annotation System with Automatic Human Detection from Video Surveillance Data,” Journal of KIISE : Computing Practices and Letters 18(11), pp.808-812, 2012.
- [9] JongHui Lee, O Hae Seog, “Design of Indexing Agent for Semantic-based Video Retrieval,” KIPS Transactions B Volume.10-B Number.6, pp.687-694, 2003.
- [10] Jong-In Kim, “Automatic classification of classical music based on musical contents,” Seoul National University : Graduate School of Cognitive Science, Master’s thesis, 2018.
- [11] Min Young Jung, Sung Han Park, “Semantic-based Scene Retrieval Using Ontologies for Video

- Server”, Journal of the Institute of Electronics and Information Engineers-CI 45(5), pp.32-37, 2008.
- [12] Tae-Dong Lee, Min-Gu Kim, “Design and Implementation of Content-based Video Database using an Integrated Video Indexing Method,” Journal of KIISE, Korean Institute of Information Scientists and Engineers 7(6), pp.661-683 2001.
- [13] Zubair Khan, Jianjun Ni, Xinnan Fan, Pengfei Shi “AN IMPROVED K-MEANS CLUSTERING ALGORITHM BASED ONAN ADAPTIVE INITIAL PARAMETER ESTIMATION PROCEDUREFOR IMAGE SEGMENTATION,” International Journal of Innovative Computing, Information and Control Volume 13 Number 5, pp.1509 - 1525, 2017.

Endnote

- 1) 정민영, 박성환, “비디오 서버에서 온톨로지를 이용한 의미기반 장면 검색”, 전자공학회논문지-CI 45(5), pp.32-37, 2008.
- 2) 김주성, 김정연, 김학일, 김유성, “비디오 감시 데이터로부터 사람의 자동 인식을 이용하는 비디오 주석 시스템,” 정보과학회논문지 : 컴퓨팅의 실제 및 레터 18(11), pp.808-812, 2012.
- 3) 이종희, 오해석, “의미기반 비디오 검색을 위한 인텍싱 에이전트의 설계,” 정보처리학회논문지 B 제10-B 권 제6호, pp.687-694, 2003.
- 4) 이한성, 임영희, 유재학, 오승근, 박대회, “멀티모달 방법론과 텍스트 마이닝 기반의 뉴스 비디오 마이닝,” 정보과학회논문지 : 데이터베이스 37(3), pp.127-136, 2010.
- 5) 이태동, 김민구 “통합된 비디오 인텍싱 방법을 이용한 내용기반 비디오 데이터베이스의 설계 및 구현,” 한국정보과학회논문지, 한국정보과학회7(6), pp.661-683 2001.
- 6) Edman Paes dos Anjos, “Experimental Evaluation of Succinct Representations of JSON Documents,” 서울대학교 대학원 : 컴퓨터공학부, 2016.
- 7) 김병철, 김창진, 윤성철, 한경숙, “JavaCC를 이용한 정적 분석 도구 구현,” Journal of the Korea Society of Computer and Information 23(12), pp.89-94, 2018.
- 8) Gil Levi, Tal Hassner, “Age and Gender Classification using Convolutional Nerual Networks,” IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.34-42, 2015.
- 9) 강주형, 곽수영, “감시 영상에서 움직임 정보 분석을 통한 폭력행위 검출,” 방송공학회논문지 제20권 제3호, pp.430-439, 2015.
- 10) Byoungjun Kim, Joonwhoan Lee, “A Deep-Learning Based Model for Emotional Evaluation of Video Clips,” INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS 18(4), pp.245-253, 2018.
- 11) Zubair Khan, Jianjun Ni, Xinnan Fan, Pengfei Shi “AN IMPROVED K-MEANS CLUSTERING ALGORITHM BASED ONAN ADAPTIVE INITIAL PARAMETER ESTIMATION PROCEDUREFOR IMAGE SEGMENTATION,” International Journal of Innovative Computing, Information and Control Volume 13 Number 5, pp.1509 - 1525, 2017.
- 12) Gil Levi, Tal Hassner, “Age and Gender Classification using Convolutional Nerual Networks,” IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.34-42, 2015.
- 13) Gil Levi, Tal Hassner, “Age and Gender Classification using Convolutional Nerual Networks,” IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.34-42, 2015.