OpenCV를 활용한 k-means clustering 기반의 포스터 색감 분석 기법 및 추천 시스템

김태홍*, 오수진*, 김응모**
*성균관대학교 정보통신대학
**성균관대학교 소프트웨어대학
e-mail : s009255@skku.edu, bgbanana4@gmail.com, ukim@skku.edu

k-means clustering analysis of a movie poster colors using OpenCV, and recommendation system.

Tae Hong Kim*, Sujin OH*, Ung-Mo Kim**
*College of Information and Communication Engineering, Sungkyunkwan University
**College of Software, Sungkyunkwan University

요 약

본 연구는 영화 포스터를 대상으로 OpenCV를 활용하여 k-means clustering 기반의 색감을 분석하는 기법을 제안한다. 또한 이를 활용하여 영화 포스터 간의 유사도를 구하고, 특정 영화와 유사한 대표색을 가지는 영화를 추천하는 시스템을 제안한다. 이를 위해, 본 연구에서는 다음과 같은 가정을 기반으로 한다. 첫 번째, 포스터는 해당 영화를 가장 잘 나타내는 이미지로, 포스터의 색감은 영화의 전반적인 분위기를 가진다. 두 번째, 영화 사이에 유사한 색감을 가진다면, 해당 영화들은 유사한 분위기를 가진다. 본 연구에서는 2단계로 나누어 연구를 진행한다. 우선 k-means clustering 기법을 통하여 데이터를 전 처리하여 영화별 대표색을 선정한다. 이 때, 선정된 대표색을 이용하여 각 영화간 색감 유사도를 분석한 결과를 통해, 같은 장르의 영화는 유사도가 높음을 확인할 수 있었다. 다음으로 앞의 색감 유사도 분석을 통하여, 특정 영화와 높은 유사도를 가지는 영화를 추천한다. 본 연구에서 추천된 영화는 기존의 영화 선택 기준에 비하여 사용자 본인의 취향을 반영한다. 본 연구 내용이 영화를 추천하는 과정에서 반영된다면 추천 시스템의 정확도와 사용자 만족도 향상에 기여할 것으로 기대된다.

1. 서론

영화는 대중적인 문화생활로, 현재 어떤 문화 콘텐츠보다도 강력한 영향력을 가진다. 영화 선택에 있어서, 영화의 줄거리, 장르, 분위기, 주변의 관람평, 포털사이트에서 제공하는 영화 평점 등 다양한 요소가 중요한 기준으로 작용한다. 하지만, 주변의 관람 평 혹은 평점과 같은 요소들은 지극히 주관적인 성향을 띄기에, 이들 요소를 기반으로 선택한 영화가 생각과 다른 경우가 다수 발생한다.

영화 선택에 있어서 중요한 역할을 하는 또 다른 요소로 포스터를 들 수 있다. 영화 포스터는 영화 매체를 이용 시 처음 관심을 증폭시키는 것에 최소의 시공간으로 접할 수 있는 수단이다. 한 컷으로 예상 관객들로 하여금 핵심내용을 상상하는 수단이 되며, 감독의 전달 의도를 최대한 잘 반영한 이미지라 할 수 있다[1, 2]. 따라서 본 연구에서는 영화 포스터로부터 색감을 추출하여 영화를 선택하는 새로운 기준을 제안하고자 한다. 이 기준은 다른 사람의 주관적인 평을 배제하고 영화를 보는 당사자만의 취향을 반영한다.

최근 영상 처리와 관련된 관심과 더불어 이를 처리하기 위해 여러 알고리즘이 제안되고 있다. 특히, 본 연구에서 사용한 OpenCV 라이브러리의 경우, 영상 처리를 위해 여러 알고리즘을 제공한다. 본 연구에서는 OpenCV 라이브러리의 k-means clustering 알고리즘을 이용하여 영화 포스터의 이미지 처리를 진행한다. 각 영화 포스터로부터 해당 영화를 대표하는 색과 각 색의 비중에 대한 정보를 추출한다. 이를 분석하여 영화간의 색감 유사도를 구하고, 이 때, 얻어진 유사도를 기반으로 사용자의 취향을 반영한 영화 추천 시스템을 제안한다.

본 연구에선 영화 포스터를 데이터 기반으로 하여 대표색을 추출하고 이를 통해 색상 차를 구하는 방법을 제안하며 또한, 구해진 색상 차를 활용하여 유사한 분위기의 영화를 추천하는 시스템은 제안한다. 2장에선 본 연구의 선행 연구 및 기술 현황에 대하여 소개하며, 3장에선 데이터 수집 및 전처리 과정을 소개한다. 4장에서는 실제 구현과 분석된 데이터를 통한 결과를 제공하며, 이어 5장에선 본 연구의 결론을 제시한다.

2. 관련 연구

본 절에서는 선행 연구로 OpenCV 라이브러리와 k-means Clustering 기법에 대하여 서술한다.

2.1. OpenCV 라이브러리

OpenCV는 BSD 라이센스하에 배포되어 학술 및 상업적 용도로 이용가능하다[3]. C++, python 및 Java로 이용이 가능하고 현재 활용되는 대부분 OS에서 사용이 가능하다. OpenCV는 계산 효율성과 실시간 응용 프로그램 제작에 중점을 두고 설계되었으며 현재는 간단한 아키텍쳐와 함께 사용하는 경우부터 광산 검사와 로봇 공학에 까지 다양한 분야에 이용되고 있다[4]. 전문적 지식 없이 코드 몇 줄로 구현할 수 있어 영상처리 분야를 대중화 시켰다.

2.2. k-means clustering

머신러닝은 어떤 문제가 주어지고 이를 풀기위한 일반적인 알고리즘 개발의 한 형태로 볼 수 있으며, 성능 측면에서는 프로그램 자체가 학습을 하여 개선이 되도록 하는 알고리즘의 개발을 목적으로 한다. Clustering은 머신러닝의 일종으로, 카테고리화 되지 않은 데이터들을 클러스터로 분류한다. 클러스터를 정의하는 방법은 여러 가지가 있지만, 본 연구에서는 사용하는 k-means 방식을 사용한다. 같은 클러스터 내부의 데이터는 가깝다고 가정되며, 각각의 클러스터마다 중심이 존재한다. 클러스터 내의 데이터가 중심과 얼마나 가까운지에 따라 비용이 책정되고, 그 비용을 최소화 하는 클러스터를 찾는 알고리즘이다. 각 클러스터를 찾을 때 최적의 중심을 찾기 위해 클러스터의 *mean*을 계산하고, 총 k개의 클러스터를 찾는다. k-means clustering은 이미지 또는 비디오 시퀀스의 픽셀 그룹화에 사용되고 있다.

3. 데이터 수집 및 전처리

본 연구에서는 아래와 같은 2가지 가정을 기반으로 한다.

- 포스터는 영화를 대표하는 요소이며 영화의 핵심 내용과 감독의 의도를 반영한다[1, 2]. 그러므로 대표 이미지로서의 포스터를 분석하여 얻은 색감은 영화의 전반적인 색감으로 판단할 수 있다.
- 같은 장르의 영화는 다른 장르의 영화에 비하여 전반적인 분위기가 유사함을 가정한다. 본 연구에서는 이를 활용하여 포스터 속의 색감 유사도를 기반으로 분위기 및 장르를 구분한다.

3.1. 데이터 수집

본 연구에서는 장르별로 전반적인 분위기에 차이가 있는 것을 이용하기 위하여, 장르를 기준으로 데이터를 수집하였다. 하지만 하나의 영화에 다양한 장르를 포함하는 경우도 많고, 다른 장르라도 느껴지는 분위기가 비슷한 경우가 있다. 예를 들어 스릴러와 공포는 확실히 다른 장르로 구분하지만 전반적인 분위기는 유사성을 가지고 있다. 그러므로 수집 장르로 드라마와 공포 두 가지로 한정하여 데이터를 수집한다. 드라마와 공포 장르는 한 영화에서 공존하기 힘들며 전반적인 분위기에서 큰 차이를 보인다. 드라마 장르에서 12가지 영화에 대한 포스터 이미지를, 공포 장르에서 10가지 영화에 대한 포스터 이미지를 데이터로 수집하여 사용한다.

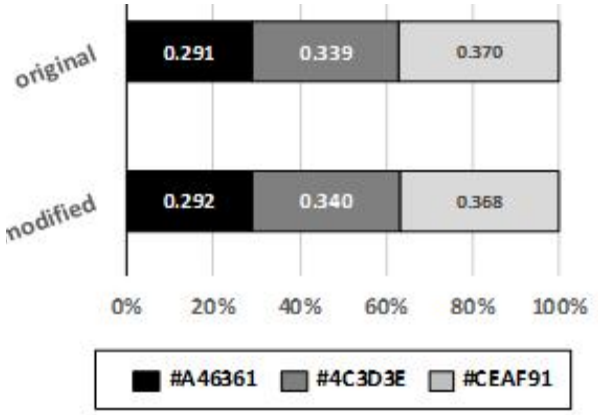
3.2. 데이터 전처리

3.2.1. 흰색 배경 처리

흰색 배경의 경우 색상 자체에 의미가 있기 보다는 여백으로서 의미를 가진다. 그러므로 이 비중이 충분히 크다면 결과에 의미 없는 흰색이 나타난다. 이를 해결하기 위해, 클러스터의 개수 k 대신 $k+1$을 파라미터로 사용한다. 그리고 흰색의 비중이 높은 경우, 이를 제거하고 나머지 색으로 데이터를 구성한다. 흰색 배경이 아닌 경우, k개의 결과만을 얻기 위하여, 가장 낮은 비중의 색을 제거한다.

3.2.2. 속도 개선

입력 이미지에 따라도 상이한 처리 속도를 가진다. 본 연구에서는 이미지의 전반적인 색감만을 필요하기 때문에, 이미지를 구성하는 픽셀 수를 줄여 속도를 개선한다. python의 OpenCV 라이브러리에서 제공하는 resize 메소드를 사용하여 x, y축을 기준으로 pixel수를 반으로 감소시킨다.



(그림 1) 영화 *써니*의 clustering 결과(k=3)

(그림 1)은 원본 파일과 축소된 파일의 clustering 결과를 시각화한 그래프이다. 두 결과를 비교하였을 때, 선택된 색상의 RGB값의 차이는 거의 없었으며, 그 비중 또한 매우 유사함을 확인하였다.

4. 색감 유사도 Similarity 분석

본 연구에서는 전처리 후 데이터를 이용하여 포스터 대표 색감 유사도 분석한다. 유사도는 Lab색상체계에서 구할 수 있는 색상 차이 값 $\delta(A_i, B_j)$를 이용한다.

$$Sim(A,B) = \sum_{i=1}^{k}\sum_{j=1}^{k}\delta(A_i,B_j)*w_{Ai}*w_{Bj} \quad (1)$$

수식 (1)은 두 이미지 A, B 데이터 사이의 색상 유사도 $Sim(A,B)$를 얻는 식이다. 이 때, A_i는 이미지 A에서 i번 째 clustering 데이터를 의미한다. 또한, w_{Ai}는 그 비중을 나타내며, 그 값은 폐구간 [0,1]을 가진다.

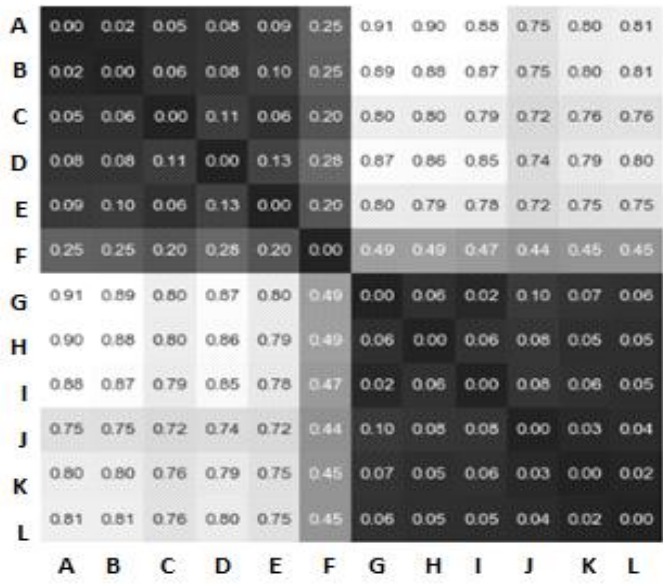
증명. $Sim(A,B)$의 최솟값과 최댓값을 구하기 위해, 다음과 같이 clustering 파라미터 k값을 1로 가정한다. 우선 이미지의 색이 동일한 상황을 가정한다. 두 이미지가 동일하기에 $\delta(A_1, B_1) = 0$이며, $Sim(A,B)$는 최솟값으로 0을 가진다. 다음으로는 이미지 색상이 상반되는 경우, 즉 $RGB(0,0,0)$와 $RGB(255,255,255)$를 가정한다 (즉, $\delta(A_1, B_1) = 1$, $w_{A1} = w_{B1} = 1$). 따라서 $Sim(A,B)$는 최댓값으로 1을 가진다. 이를 통해, $Sim(A,B)$는 폐구간 [0,1]임을 알 수 있다. 다음의 증명에서는 k값을 1로 가정하였지만, 실제 1보다 큰 k값을 사용한다면, 동일한 이미지를 비교하여도 $Sim(A,B)$는 0이 될 수 없으며, 최댓값 또한 1이 될 수 없다.

4.1. 색감 유사도 분석

본 절에서는 3장에서 제시한 본 연구의 가정 중 두 번째 가정을 증명한다.

포스터의 대표 색감 유사도 분석을 위해, python colormath 라이브러리를 이용하여, RGB 색상을 Lab 색상으로 변환한다. Lab 색상체계에서는 두 색상간의 차이를 폐구간 [0,1]값으로 반환할 수 있다.

<표 1>은 각 장르별로 6가지씩 영화를 선정하여, 색감 유사도 분석까지의 과정을 마친 결과이다. 가독성을 높이기 위해 영화 제목을 알파벳으로 대체하였으며, <표 1>의 우측에 관련 정보를 표기한다.



(그림 2) 색감 유사도 분석 히트맵

(그림 2)은 <표 1>를 이용하여 $Sim(A,B)$을 히트맵으로 시각화 한 것이다. 같은 장르의 영화에서는 낮은 $Sim(A,B)$, 즉 높은 유사율을 가짐을 확인할 수 있다.

<표 1> 수집 데이터 간의 대표 색감 유사도 분석 결과

	A	B	C	D	E	F	G	H	I	J	K	L
A	0	0.02	0.05	0.08	0.09	0.25	0.91	0.9	0.88	0.75	0.8	0.81
B	0.02	0	0.06	0.08	0.1	0.25	0.89	0.88	0.87	0.75	0.8	0.81
C	0.05	0.06	0	0.11	0.06	0.2	0.8	0.8	0.79	0.72	0.76	0.76
D	0.08	0.08	0.11	0	0.13	0.28	0.87	0.86	0.85	0.74	0.79	0.8
E	0.09	0.1	0.06	0.13	0	0.2	0.8	0.79	0.78	0.72	0.75	0.75
F	0.25	0.25	0.2	0.28	0.2	0	0.49	0.5	0.48	0.44	0.45	0.45
G	0.91	0.89	0.8	0.87	0.8	0.49	0	0.06	0.02	0.1	0.07	0.06
H	0.9	0.88	0.8	0.86	0.79	0.5	0.06	0	0.06	0.08	0.05	0.05
I	0.88	0.87	0.79	0.85	0.78	0.48	0.02	0.06	0	0.08	0.06	0.05
J	0.75	0.75	0.72	0.74	0.72	0.44	0.1	0.08	0.08	0	0.03	0.04
K	0.8	0.8	0.76	0.79	0.75	0.45	0.07	0.05	0.06	0.03	0	0.02
L	0.81	0.81	0.76	0.8	0.75	0.45	0.06	0.05	0.05	0.04	0.02	0

Brief	Movie
A	cat
B	closeknit
C	life
D	ride
E	wonder
F	timetraveler
G	annabelle
H	counjuring2
I	jigsaw
J	gonziam
K	hereditary
L	getout

5. 결과 분석

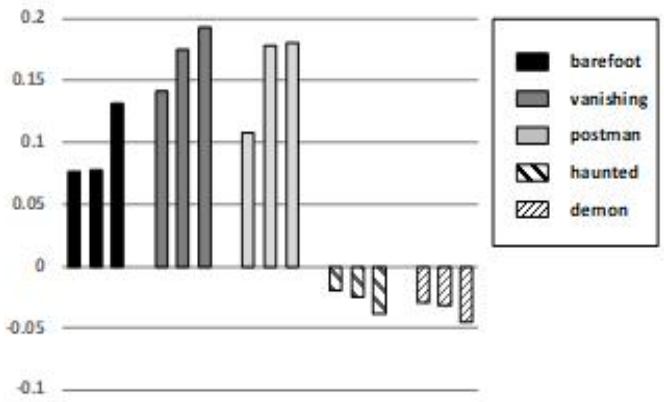
본 연구에서는 OpenCV 3.4.1을 사용하며, N값으로 수식 (2)와 같이 사용한다.

$$N = 3 \quad (2)$$

본 연구에서는 영화간의 $Sim(A,B)$을 구하고, 이들 중 가장 낮은 값, 즉 가장 유사율이 높은 영화 N개를 추천한다. 이를 위하여, 드라마 장르의 영화 3개(barefoot, vanishing, postman)과 공포 장르의 영화 2개(haunted, demon) 총 5개의 영화를 선정하여 영화 추천 결과를 확인한다.

<표 2> 영화별 추천 결과

test movie			1	2	3
1	barefoot	movie	wonder	life	cat
		value	0.077	0.078	0.131
2	vanishing	movie	sunny	timetraveler	dangsin
		value	0.142	0.175	0.193
3	postman	movie	ride	cat	closeknit
		value	0.108	0.178	0.181
4	haunted	movie	jigsaw	annabelle	getout
		value	0.019	0.025	0.038
5	demon	movie	counjuring2	annabelle	jigsaw
		value	0.029	0.032	0.045



(그림 3) 영화별 추천 결과 그래프

<표 2>는 새롭게 추가된 5가지 영화에 대하여 결과를 정리한 표이다. 각 영화별로 색감 유사도 순위에 따라, 영화 제목과 색감 유사도 값(value)을 나타낸다. (그림 3)은 장르의 구분을 위해 드라마 장르의 값은 양수로 공포 장르의 값은 음수로 하여, <표 2>를 시각화 한 그래프이다. <표 2>와 (그림 3)을 통해 새롭게 추가된 영화의 추천된 영화가 동일한 장르의 영화임을 확인할 수 있다.

6. 결론

본 연구는 영화 포스터 속 색감을 통하여 영화 분위기를 분류하고, 이를 기반으로 영화를 추천하는 시스템을 제안한다. 이를 위하여, OpenCV 라이브러리의 k-means clustering을 활용하여 포스터에서 대표 색감 k개를 추출한다. 추출된 색감으로부터 색감 유사도 분석을 하여 영화간의 유사도를 수치화한다. 마지막으로 유사도를 기반으로 특정 영화와 유사한 분위기를 지니는 영화 N개를 추천한다. 특정 영화와 선정된 영화가 동일한 장르를 가짐을 확인하였으며, 본 연구에서 제안한 추천 시스템의 활용 가능성이 높음을 알 수 있다.

본 연구에서는 2가지의 장르로만 결과 분석하였지만, 다양한 장르에 대하여 확장될 가능성이 있다. 또한, 데이터로 수집된 영화의 수가 더욱 많을수록 다양한 풀에서의 비교가 가능해지기 때문에 향상된 성능을 보일 것으로 예상된다. 본 연구를 통해 포스터의 색감이 영화를 추천하는 새로운 기준이 될 수 있음을 확인할 수 있었기에, 추천 시스템에 하나의 척도로 활용될 수 있다.

참고문헌

[1] 김형석, 김성훈, “국내 다양성영화 포스터의 시각적 상징성에 관한 연구” 한국디자인문화학회, 3월 2015.

[2] 조성근, 김종근, “계획행동 이론을 적용한 영화관람 의도의 결정요인에 관한 연구: 영화포스터 표현형식의 조절역할을 중심으로” 한국벤처창업학회, 12월 2015.

[3] OpenCV, https://opencv.org/

[4] Dr. S. Syed Ameer Abbas, Dr. P. Oliver Jayaprakash, M. Anitha, X. Vinitha Jaini, “Crowd Detection and Management using Cascade classifier on ARMv8 and OpenCV-Python”, 2017 International Conference on Innovations in Information, Embedded and Communication Systems, 3월 2017.