유아용 애니메이션을 위한 R-CNN 기반 캐릭터 감정 학습 기법

허민오O 장병탁
서울대학교 컴퓨터공학부
{moheo, btzhang}@bi.snu.ac.kr

Character Facial Sentiment Learning Methods based on R-CNNs for Kids' Animations

Min-Oh HeoO Byoung-Tak Zhang
School of Computer Science & Engineering, Seoul National University

요 약

영상 내 인물의 표정을 자동적으로 인식하는 기술은 주로 실사 영상 안에서 사람의 얼굴을 검출하고, 얼굴에 드러난 표정을 구분하는 기술에 초점이 맞추어져 있다. 하지만, 다수의 유아용 애니메이션은 실사가 아니며, 사람이 아닌 의인화 된 캐릭터가 등장하므로 유사 기법을 적용하기 어렵다. 본 고에서는 애니메이션 캐릭터 및 캐릭터의 감정을 함께 학습할 수 있는 기계학습 기법을 제안한다. 애니메이션 내 캐릭터의 감정은 주로 눈매와 입모양을 통해 전달되므로, 이 영역을 주로 선택하여 감정을 레이블링(labeling) 하도록 한 후, 객체 검출 딥러닝 기법인 R-CNN (Regions with Convolutional Neural Networks features)을 기반으로 캐릭터의 표정에 대한 지도학습을 수행하였다. 실제 적용 예로서, 유명 유아용 애니메이션인 '뽀롱뽀롱 뽀로로'에 대해 학습을 시도하였다. 실험 결과, 테스트 데이터 상 캐릭터 및 감정 인식 성능으로 mean Average Precision (mAP) 0.75의 인식율을 보였다. 흥미롭게도 스토리 상 얼굴이 완전히 달라진 예와 낙서형식의 그림에서도 일부분 인식에 성공하였다. 또한, 전이학습 측면에서 다른 애니메이션 및 실제 사진에 적용한 결과, 얼굴 검출 및 감정인식을 일부 수행할 수 있었다.

1. 서 론

최근 컴퓨터 비전 기술이 크게 발달함에 따라 다양한 응용들이 나타나고 있다. 그 중 한 가지는 사진 또는 동영상에 나타난 사람의 얼굴을 검출하고 (Face Detection), 얼굴에 드러난 표정(Emotion)을 구분하는 기술이다[1,2]. 이러한 기술은 주로 얼굴의 키포인트를 찾아 분석을 하는 방식을 취하며, 동일인 여부의 판별은 수행하지 않는다.

반면, 유아용 애니메이션에서 캐릭터의 표정으로부터 감정을 읽는 문제를 고려한다면, 위의 기술은 크게 도움이 되지 못한다. 유아용 애니메이션은 주로 실사가 아니며, 사람이 아닌 동물이나 무생물이 의인화 된 캐릭터가 등장하는 경우가 많기 때문에, 유사 기법을 적용하기 어렵다. (예: 뽀롱뽀롱 뽀로로, 꼬마버스 타요, 로보카 폴리 외 다수)

본 고에서는 애니메이션 캐릭터의 표정으로부터 감정(sentiment)을 학습할 수 있는 방법을 제안한다. 키포인트를 잡거나 규칙을 만드는 접근은 사실상 불가능하므로, 데이터 기반의 기계학습 방법이 적절하다. 의인화 된 캐릭터라 하더라도 애니메이션 내 캐릭터의 감정은 주로 눈매와 입모양을 통해 전달되므로[3,4], 이 영역의 정보를 추출하여야 한다. 이를 위해 눈, 코, 입이 포함된 영역만을 선택하여 감정을 레이블링 하고, 객체 검출 딥러닝 방법인 R-CNN (Regions with Convolutional Neural Networks features) [5,6]을 기반으로 캐릭터의 표정에 대한 지도학습을 수행하였다. 실제 적용한 예로서, 유명 유아용 애니메이션인 '뽀롱뽀롱 뽀로로'의 출시된 DVD 내에서 주로 등장하는 9명의 캐릭터에 대해 학습을 시도하였다. 실험 결과, 테스트 데이터 상 캐릭터 및 감정 인식 성능으로 mean Average Precision (mAP) 0.75의 인식율을 보였다. 흥미롭게도 스토리 상 변신하여 얼굴이 완전히 달라진 예에 대해서도 상당 부분 성공적인 인식하였으며, 낙서형식의 그림에서도 인식하는 경우를 관찰하였다. 또한, 전이학습 측면에서 기 학습모델로 다른 애니메이션 및 실제 사진에 적용한 결과, 얼굴 검출 및 감정인식을 일부 수행할 수 있었다.

이후, 본 논문은 다음과 같이 구성된다. . 2 장에서는 제안하는 방법론을 소개하고, 3 장에서는 실험 세팅 및 결과를 보인 후, 4 장에서 결론을 맺는다.

2. 애니메이션 캐릭터의 표정 감정 학습 방법

2.1 애니메이션 캐릭터의 표정 특성

사람이 표정으로 내적 상태를 추론할 때, 눈썹, 눈, 코, 입 등의 특징 요소가 기타 요소에 비해 중요한 요인임이 알려져 있다[3]. 반대로, 애니메이션에서 추상화 된 그림으로 정보를 전달해야 하므로 눈에

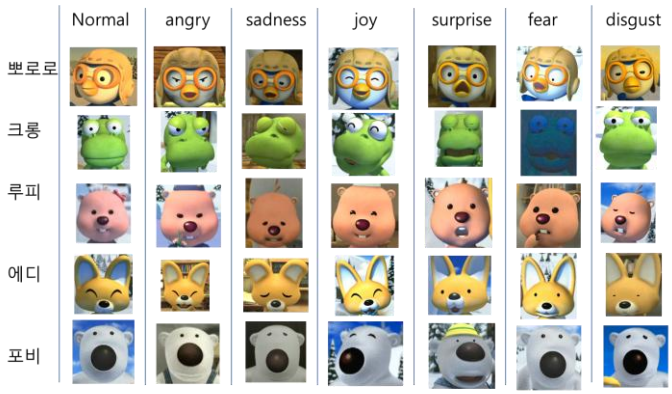

그림 1. '뽀롱뽀롱 뽀로로' 캐릭터 별 정서 표현의 예

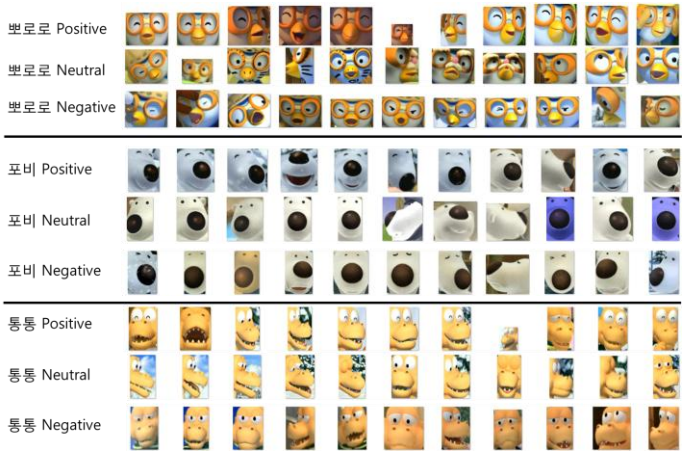

그림 2. 표정 레이블링 된 영상조각의 예

명확하게 보이는 특징인 상기 요소를 중점적으로 다룬다[4]. 그에 따라, 애니메이션의 특성과 캐릭터의 설정에 따라 표현 요소를 명확히 드러내지 않는 경우도 있다. 그림 1은 뽀롱뽀롱 뽀로로의 주요 캐릭터의 감정별 영상의 예이다. 일부 감정 – joy, surprise는 명확히 드러나지만, 기타 감정들은 모호한 경우가 대부분이다.

2.2 표정 레이블링 방법

이러한 특성들을 반영하여 표정 레이블링에 다음과 같은 가이드라인을 부여하였다.

1) 눈, 코, 입 부분을 포함하되 가능하면 작게 영역을 택하고 감정을 레이블링
2) 옆 얼굴 또는 일부가 가려진 경우 1)의 일부라도 보이면 그 부분 영역을 택하고 감정을 레이블링

그림 1과 같이 캐릭터 정서가 사람이 보기에도 차이를 느낄 수 없는 수준이라고 판단되어, 본 논문에서는 그림 2와 같이 상기 기준에 맞추어 캐릭터별로 Positive, Neutral, Negative로 레이블링 하기로 하였다

2.3 객체 검출으로서의 캐릭터 표정 인식기

상기 설명한 방법으로 얻은 훈련 데이터를 이용하여 캐릭터 표정인식 문제를 객체 검출 (object detection) 문제로 바꾸어 캐릭터 및 감정 인식문제를 해결하였다. 본 논문에서는 faster R-CNN[6]을 객체 검출 모델로 이용하였다 (그림 3). R-CNN은 크게 3 가지 구성요소로 설명된다. 입력 영상으로부터 영상인자를 추출하는 CNN 부분 – ImageNet 데이터로 사전 학습한 VGG-16

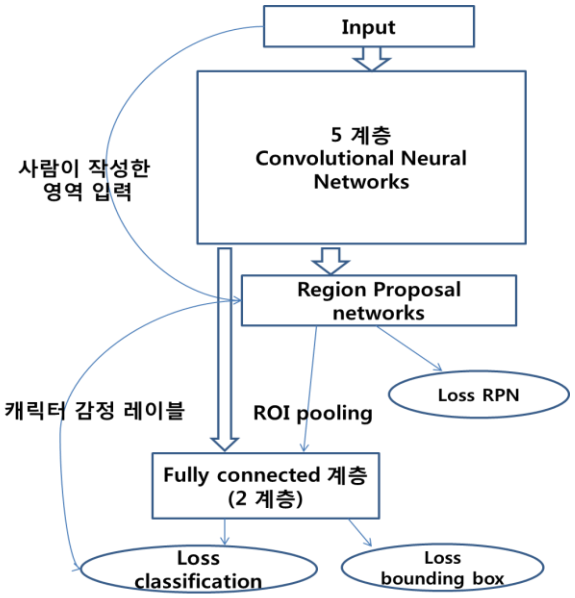

그림 3. 적용한 R-CNN의 기능별 간략 구조도

네트워크의 인자를 이용하였다. 나머지는 입력 영상에 따라 적합한 영역을 학습하는 네트워크 (region proposal network), 제안된 영역과 영상 인자로부터 분류성능 오차와 영역에 대한 오차를 계산하는 부분이다. 딥러닝에서와 마찬가지로 학습에 stochastic gradient descent로 네트워크 전체(end-to-end)를 업데이트한다.

3. 실험 결과

제안한 방법에 대한 작동 여부 및 성능 평가를 위해 유아용 애니메이션 '뽀롱뽀롱 뽀로로'를 이용하였다. 그림 4의 (a)와 같이 DVD로 출시된 에피소드들을 이용하되, 자막이 나오는 시점을 기준으로 영상을 캡처하고, 여기에 2기 이상 등장하는 캐릭터 9 명에 대해 영역과 캐릭터별 감정을 레이블링 하였다. 16 × 16 보다 작은 영역을 제외하고 남은 레이블링 된 영역들의 개수는 그림 4

	Series				Movie		Total
	Season 1	Season 2	Season 3	Season 4	Porong Porong Rescue Mission	Racing adventure	
Number of episodes	39 (3 DVD)	52 (4 DVD)	52 (4 DVD)	26 (4 DVD)	1	1	171
Number of Image-caption Pair	2,012	3,479	4,792	4,590	388	805	16,066
Total Running Time (min)	195 (5 per ep.)	260 (5 per ep.)	260 (5 per ep.)	286 (11 per ep.)	30	77	1,108

(a)

	Positive	Neutral	Negative	Total
Pororo	726	3831	837	5394
Crong	546	3581	571	4698
Eddy	915	2610	504	4029
Poby	635	2574	266	3475
Loopy	665	2397	384	3446
Petty	433	1934	318	2685
Harry	256	1837	264	2357
Rody	186	1379	82	1647
Tongtong	121	645	92	858

(b)

그림 4. 학습 데이터 크기. (a) 영상-자막 쌍 기준 분량 (b) 영역별 레이블링 영상조각 분량

(b)와 같다. 이를 에피소드를 기준으로 7:3의 비율로 무작위로 나누고, 이에 속한 영상조각들을 훈련 데이터와 테스트 데이터로 이용하였다.

그림 5와 같이 정면 모습, 가려진 모습에서도 캐릭터-감정을 잘 검출하는 것을 볼 수 있으며, 그림 6과 같이 잘못 추론한 예의 경우도 순간 포착에 의해 사람도 헷갈려 할만한 경우들이 다수 존재한다. 그림 8에 캐릭터-감정 별 성능이 정리되어 있다. mAP 값이며, Precision-Recall 곡선의 하단 면적 (Average Precision)을 질의(query)별로 구하여 평균을 낸 값이다. 클래스 별 균형이 맞지 않아도 성능이 떨어지지 않는다. (그림4, 그림7) 반면 포비 Negative는 성능이 좋지 않은데

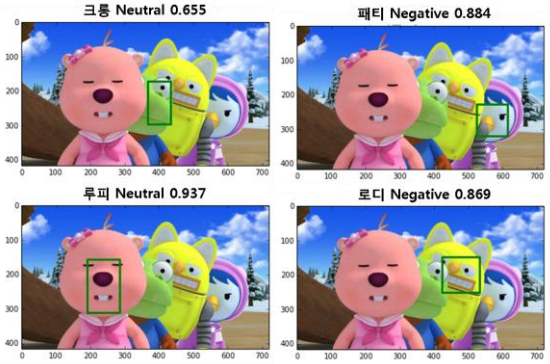

그림 5. 테스트데이터에 대한 캐릭터-감정 검출 예 (영역은 녹색 사각형으로 표시, 확신도는 그림 위에 표기)

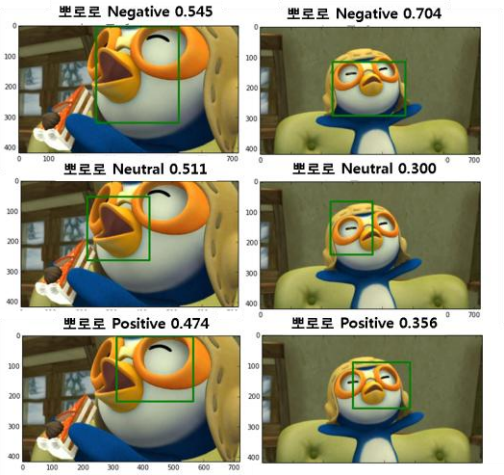

그림 6. 잘못 추론한 예 (우: 기지개 켜는 모습)

그림 7. 에피소드 상 얼굴이 달라졌으나 감정인식 됨

	Neutral	Positive	Negative	
뽀로로	0.903	0.829	0.665	0.799
크롱	0.895	0.735	0.562	0.731
포비	0.88	0.84	0.252	0.657
루피	0.9	0.807	0.746	0.818
에디	0.827	0.817	0.598	0.747
해리	0.805	0.707	0.373	0.628
패티	0.921	0.958	0.825	0.901
로디	0.895	0.866	0.564	0.775
통통	0.826	0.861	0.415	0.701
	0.872	0.824	0.556	

그림 8. 캐릭터-감정 별 검출 성능 (값: mAP)

실제 영상도 구분이 쉽지 않다. (그림 2)

그림 7은 그림 5의 로봇 로디가 친구들처럼 생명체가 되는 소원이 이루어져 얼굴이 달라지는 에피소드를 다루었으나 캐릭터-감정을 대부분 옳게 인식해냈다. (48샷 중에 43은 인식성공, 3은 캐릭터 인식오류, 2는 감정인식 오류. 그림 7의 중앙이 오류의 예). 또한, 낙서형식의 그림에서도 인식하는 경우를 관찰하였다. 전이학습 측면에서 학습된 모델로 다른 애니메이션 및 실제 사진에 적용한 결과, 얼굴 검출 및 감정인식을 일부 수행할 수 있었다.

4. 결론

본 고에서는 딥러닝 기반 객체 검출 기술을 기반으로 애니메니션 캐릭터의 표정을 읽어 감정을 인식하는 기법을 소개하였다. 자동화된 유아용 비디오 분석 및 스토리 학습에 기여할 것으로 기대한다[7]

감사의 글

이 논문은 2016년도 정부(R0126-16-1072-SW스타랩, 10044009-HRI.MESSI, 10060086-RISF)의 재원으로 지원을 받아 수행된 연구이며, 네이버랩스의 지원을 일부 받았음.

참고문헌

[1] K. Sikka, et al., Multiple kernel learning for emotion recognition in the wild, *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 517-524, ACM, 2013.

[2] Noldus, Face Reader, http://www.noldus.com/human-behavior-research/products/facereader, 2008.

[3] G. Rhodoes, Looking at Faces: First-order and Second-order Features as Determinants of Facial Appearance, *Perception*, Vol. 17, pp. 43-64, 1988.

[4] 고혜영, 애니메이션에 나타난 캐릭터 표정 분석을 통한 기본정서 표현, 한국엔터테인먼트산업학회 논문지, Vol.1(1), pp.11-14, 2007.12.

[5] R. Girshick, et al., Rich feature hierarchies for accurate object detection and semantic segmentation, In CVPR 2014, pp. 580-587, 2014.

[6] S. Ren, et al., Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in Neural Information Processing Systems (NIPS)*, pp. 91-99, 2015.

[7] 허민오, 한동식, 김경민, 장병탁, 유아용 비디오를 위한 딥러닝 기반 스토리 학습 프레임워크, 한국정보과학회 동계학술발표회 논문집, pp. 725-727, 2015.12.