SIFT 특성 분포를 이용한 비디오 스트림의 장소 변화 예측

유준희 ○, 석호식, 장병탁

서울대학교 컴퓨터공학부

{jhyoo, hsseok, btzhang}@bi.snu.ac.kr

Location Change Estimation in a Video Stream based on SIFT Feature Distributions

Jun Hee Yoo ○ , Ho-Sik Seok, Byoung-Tak Zhang

Seoul National University

School of Computer Science & Engineering

요 약

비디오 데이터의 지능적인 처리를 위해서는 사전에 작성한 메타데이터에 제한 받지 않는 유연한 접근방법이 필요하다. 본 논문에서는 엔트로피를 이용하여 적절한 특징을 추출한 후 비디오를 처리하는 방법을 소개한다. 이미지 인식이 잘 될 경우 일정한 이미지 조합으로 비디오의 배경을 설명할 수 있지만, 이미지 인식이 어렵기 때문에 동일한 배경일지라도 등장 인물의 움직임, 촬영 각도의 변화 등 사소한 변화가 발생하면 컴퓨터는 다른 이미지인 것으로 간주하게 된다. 우리가 제안하는 방법은 비디오를 구성하는 이미지 프레임에서 추출한 SIFT(Scale Invariant Feature Transform) 특성의 분포를 엔트로피에 기반하여 재구성한 후 분포 변화를 통해 장소 변화를 추정하는 방법이다. 제안 방법은 비디오 데이터의 이미지를 특징 짓는 비주얼 워드의 분포를 활용하기 때문에 사소한 변화 정도의 영향을 받지 않으면서 동시에 배경의 확연한 변화를 나타낼 수 있다. 우리는 실제 TV 드라마 데이터에 적용하여 제안 방법의 유용성을 확인하였다.

1. 서 론

사회가 고도로 정보화되어 가면서 이전에는 상상도 할 수 없을 만큼 빠른 속도로 정보가 쌓이고 있다. 쏟아지는 정보들 속에서 사용자가 원하는 정보를 찾는 검색 방법 또한 발전에 발전을 거듭하여 각종 문서와 그림, 비디오 찾아 전달한다. 그러나 대부분이 텍스트인 문서와는 달리, 그림이나 비디오를 검색하는 경우는 해당 이미지나 비디오가 있는 페이지의 단어들을 이용하여 검색하거나 데이터가 무엇을 담고 있는지를 색인 해둔 메타데이터를 이용하여 검색을 하는 것이 일반적이다. 이 경우 컴퓨터가 스스로 비디오나 사진의 내용을 파악하여 메타데이터를 입력할 수 없기 때문에 사람이 정보를 주지 않으면 컨텐츠의 내용이 무엇인지 컴퓨터는 전혀 알 수가 없다. 현재 널리 쓰이고 있는 방법인 작성자가 직접 메타데이터를 입력하는 경우와 다수의 사용자로 하여금 메타데이터를 입력하는 방식은 각각 키워드가 적거나 적합하지 못한 키워드를 적는 문제가 생길 수 있기 때문에 컴퓨터가 메타데이터의 도움을 받지 않고 비디오의 내용을 파악할 수 있다면 비디오 데이터를 지능적으로 처리하는데 큰 도움이 될 것이다.

본 논문에서는 비디오 데이터의 지능적인 처리를 위하여 비디오의 이미지 데이터에 기반하여 장소의 변화를 추정하는 방법을 소개한다. 제안 방법은 SIFT (Scale Invariant Feature Transform) 특징[1]과 엔트로피를 이용하는 것으로, SIFT 특징은 회전이나 확대에 의한 변화가 발생해도 유사한 속성을 유지할 수 있다는 장점이 있다. 또한 엔트로피를 이용한 비주얼 워드 인덱스 재배치로 보다 효율적으로 데이터를 변환할 수 있다.

우리는 이 같은 장점을 이용하여 비디오의 배경 변화를 추정하였다. 제안 방법에서는 비디오로부터 1 초 간격으로 이미지를 추출하고, 추출된 이미지들로부터 SIFT 특징들을 수집한다. 수집된 SIFT 특징에 K-means 클러스팅을 적용하여 비주얼 워드를 정의한 후, 각 워드속성의 엔트로피를 계산해서 비주얼 워드로 구성된 표현공

간을 변환하였다. 우리는 TV 드라마[1] 이미지 데이터의 비주얼 워드 분포를 추정하여 제안 표현 방법이 배경 변화를 얼마나 잘 표현할 수 있는지 확인하였다.

2. 관련 연구

Nicholas Morsillo 등은 컴퓨터를 이용하여 자동으로 YouTube 의 비디오에서 이용해 비슷한 영상을 찾아서 태그를 자동으로 생성하는 방법을 소개하였다[2]. 이 연구에서는 영상에서 비주얼 특징(Visual Feature)들을 추출, 서로 다른 영상을 비교하여 비슷한 정도를 이용하여 영상에 대한 태그를 수정하거나 추가하였다. 본 논문에서 제안하는 방법은 하나의 비디오 안에서 발생하는 동적인 변화를 추정한다는 점에서 Nicholas Morsillo 등의 연구와는 차이가 있다.

비디오 컨텐츠의 장소를 인식하는 연구는 아니지만 맥락이 유사한 연구로는 카메라로부터 입력되어 들어오는 비디오 데이터에서 주요 개체들을 인식하고, 이를 이용하여 현재의 장소가 어디인지 구분하는 방법에 대한 연구[3,4]등이 있다. 이들 연구는 장소 구분이라는 목적에서 본 연구와 유사하지만 본 논문에서 제안한 방법은 등장인물의 클로즈업 등 각종 화면 전환 효과가 있는 데이터를 다루기 때문에 [3,4]의 연구와 큰 차이가 있다. 그 뿐만 아니라 [3,4]의 연구는 인식된 물체가 어떤 사물인지를 알아내기 위해 다수의 사람들이 직접 개체를 구분하고 레이블을 부여해 놓은 'LabelMe' 프로젝트[5]을 이용하여 물체를 구분하기 때문에 비디오의 데이터만을 활용하는 본 연구와 차이가 있다.

3. 실험 방법

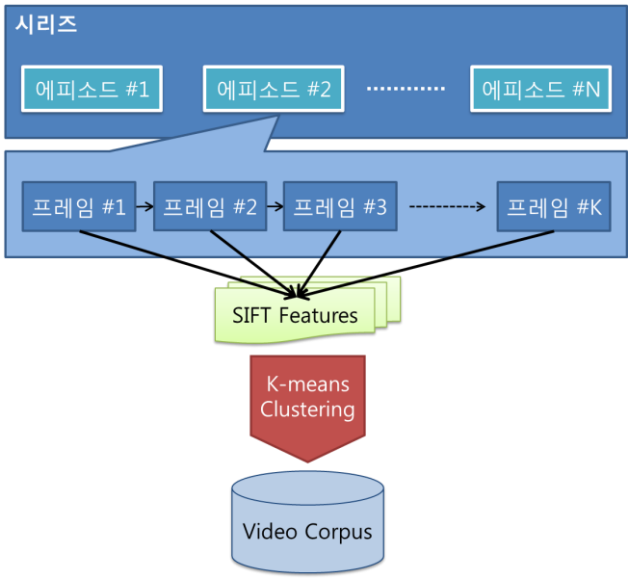

그림 1 데이터 전처리 과정

[1] 본 논문에서는 미국 NBC 방송국의 유명 시트콤인 '프렌즈'를 사용하였다.

3.1. 데이터 전처리 과정

드라마의 한 시리즈에서 하나의 에피소드를 선택하고 에피소드에서 초당 한 장의 이미지를 추출해 낸다. 추출한 각각의 이미지에서부터 SIFT 알고리즘을 이용하여 SIFT 특징들을 찾아내고 찾아진 모든 SIFT 특징들을 k-means 클러스터링을 적용하여 각 클러스터를 대표하는 k 개의 중심점을 찾는다. 찾아진 k 개의 중심점을 비주얼 워드(visual word)로 정의하여 비디오 코퍼스를 구성한다.

3.2. 실험 과정

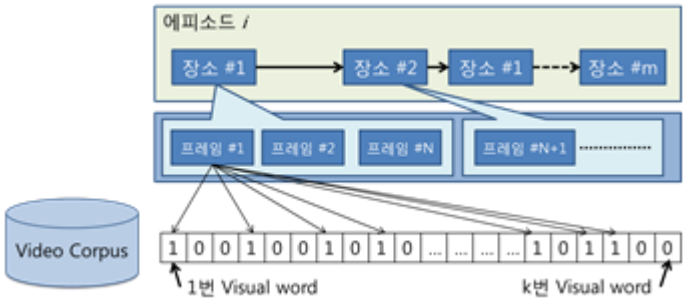

그림 2 하나의 프레임에서 비트스트링을 만드는 방법

각 프레임에서 나온 SIFT 특징들을 비주얼 코퍼스 크기만큼의 길이를 가지는 이진수 배열(2^k)로 표현한다. 이 때, 반복 출현은 기록하지 않고, 해당 비주얼 워드의 등장 여부만을 1 과 0 을 이용하여 기록한다(그림 2).

그림 3 비트스트링의 인덱스 재배열

각각의 프레임에서 출현한 비주얼 워드들 중에서 장소를 구분하는데 효과적인 속성을 찾기 위해 비주얼 워드의 속성[2] 별로 엔트로피를 계산한다(그림 3). 이 단계를 위해 학습 데이터에 장소를 기록하고 공개 소프트웨어 툴인 WEKA 3.6[6]을 이용하여 엔트로피 계산을 하였다. 계산된 엔트로피를 기반으로 가장 효과적인 속성을 1 번 인덱스로, 가장 효과적이지 못한 속성을 K 번 인덱스로 비주얼 워드의 속성을

[2] SIFT 특징은 128 개의 속성을 가진다.

재배치하고, 이진수 배열을 정수로 바꾸어 장소별 평균과 분산을 구해 장소 별 분포를 구한다.

본 논문에서는 배경 예측을 위한 예비 연구로 배경 변화를 잘 나타낼 수 있는 비주얼 워드 표현 방법을 찾고자 한다. 이를 위해 사람이 판단한 배경 변화 구간에 속한 이미지 프레임들의 비주얼 워드 분포를 그려 보았다.

4. 실험결과

데이터에는 모두 7 개 의 서로 다른 장소가 있었다. 장소 변화에 집중하기 위해 실험 시, 드라마의 인트로 부분과 중간중간의 화면 전환 장면들은 장소가 없는 것으로 취급하였다.

그림 4 장소 변화의 예

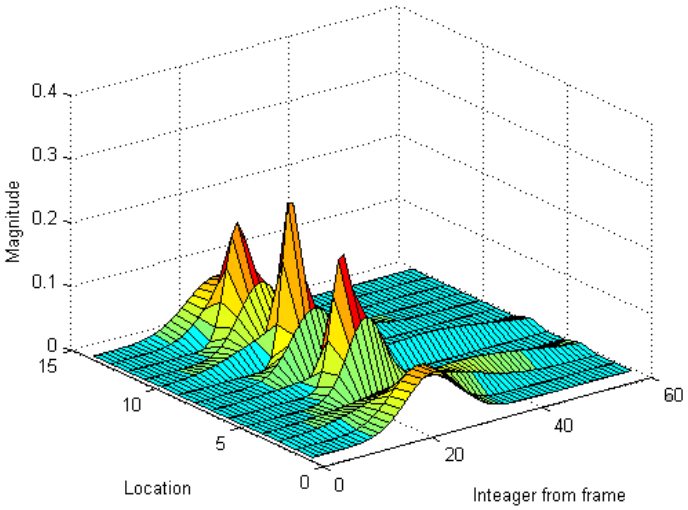

그림 5 각 장소의 분포도

시간에 따른 장소변화와 장소의 분포도 그래프를 눈으로 비교하면 상당히 유사한 형태를 보이고 있으나, 유의 수준 0.05 를 기준으로 양측 z 검정을 해보면 [표 1]과 같이 이전 장소와 다음 장소간의 차이가 나는 구간 많이 있었다.

장소 1 과 장소 4 만을 추려서 비교해 보면(그림 4, 그림 6), 장소 1 은 거의 비슷한 분포로 나타나는데, 장소 4 는 한 분포가 다른 두 분포와는 차이를 보이는 것을 알 수 있다. 이것은 카메라가 그 장소의 한 부분만을 찍었기 때문에 나타난 결과이다.

표 1 양측 z 검정 결과 ($Z_{0.025}$=1.96)[3]

장소 변화 시점	z 값	장소 변화 시점	z 값
1	-1.17	8	**2.20**
2	**-2.11**	9	**-2.78**
3	**-4.07**	10	**2.27**
4	**8.52**	11	**2.70**
5	**2.23**	12	**-3.67**
6	**-12.9**	13	-0.09
7		14	
8	**11.6**		

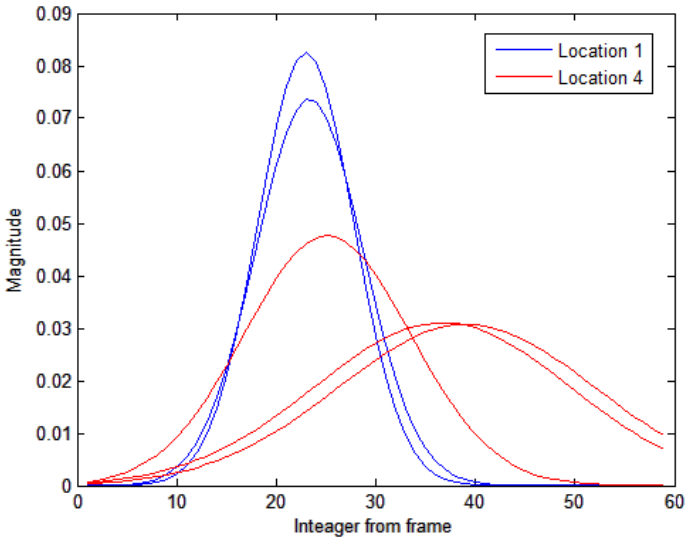

그림 6 장소 1 과 장소 4 의 분포도

5. 결론 및 향후 연구

본 논문에서는 SIFT 특성과 엔트로피를 이용하여 비디오의 장소 변화를 나타낼 수 있는 표현 방법을 제안하였다. 우리가 제안한 표현 방법은 컴퓨터에 의한 배경 변화 예측의 예비 연구로 향후 우리는 각 배경의 비주얼 워드 분포를 설명할 수 있는 은닉 변수 모델 개발에 제안 표현 방법을 활용할 것이다.

[3] 에피소드에 나오는 장소는 7 개의 장소이지만, 각 장소가 여러 번 나타나기도 하기 때문에 장소 변화는 모두 13 번 일어난다.

감사의 글

이 논문은 교육과학기술부의 재원으로 국가연구재단의 지원을 받아 수행된 연구(0421-20110032, 지능형 추천 서비스를 위한 인지기반 기계학습 및 추론기술, Videome)이며, 지식경제부 산업원천기술개발사업(10035348, 모바일 플랫폼 기반 계획 및 학습 인지 모델 프레임워크 기술 개발) 및 교육과학기술부의 BK21-IT 사업에 의해 일부 지원되었음.

[참고 문헌]

[1] David G. Lowe, "Object recognition from local scale-invariant features," Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol.2, no., pp.1150-1157 vol.2, 1999 doi: 10.1109/ICCV.1999.790410

[2] Nicholas Morsillo, Gideon Mann and Christopher Pal, “YouTube Scale, Large Vocabulary Video Annotation “VIDEO SEARCH AND MINING Studies in Computational Intelligence, Volume 287/2010, pp.357-386, DOI: 10.1007/978-3-642-12900-1_14, 2010

[3] Kai Ni, Anitha Kannan, Antonio Criminisi, John Winn, "Epitomic Location Recognition", Pattern Analysis and Machine Intelligence, IEEE Transactions on, On page(s): 2158 - 2167, Volume: 31 Issue: 12, Dec. 2009

[4] Antonio. Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin, “Context-Based Vision System for Place and Object Recognition,” Proc. Int’l Conf. Computer Vision, vol. 1, pp. 273-280, 2003.

[5] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, William T. Freeman, “LabelMe: a database and web-based tool for image annotation,“ International Journal of Computer Vision, pages 157-173, Volume 77, Numbers 1-3, May, 2008.

[6] Holmes, G.; Donkin, A.; Witten, I.H.; , "WEKA: a machine learning workbench," Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on , vol., no., pp.357-361, 29 Nov-2 Dec 1994