영상의 특정 의미를 반영하는 Key Frame의 추출 방법

하종우O, 노정담**, 윤성웅*, 김민수***, 안창원*
O(주)바이브컴퍼니,
*(주)바이브컴퍼니,
**(주)아프리카TV,
***부경대학교 산업및데이터공학과
e-mail: jongwoo0.ha@gmail.comO, {swyoon, ahn}@vaiv.kr*,
wjdeka93@gmail.com**, minsky@pknu.ac.kr***

Finding focused key frames of a given meaning on video data

Jong-Woo HaO, Jung-Dam Noh**, Soungwoong Yoon*, Min-Soo Kim***, Chang-Won Ahn*
OVAIV Company inc.,
*VAIV Company inc.,
**AfreecaTV Corp.,
***Dept. of Industrial and Data Engineering, Pukyong National University

● 요 약 ●

영상을 구성하는 프레임 중에 키프레임은 일반적으로 영상 정보를 효과적으로 요약하거나 용이한 분석을 위해 선정된다. 화상이 가진 의미는 인물/사물 등의 객체탐지를 통해 추출되는데, 기존의 키프레임 관련 연구는 영상이 가지는 의미를 반영하는 키프레임을 찾아내기 어렵다. 본 논문에서는 영상이 가지는 특정 의미가 있다고 할 때 이를 반영하는 키프레임을 효과적으로 추출하는 방법을 실험적으로 탐구하였다. 구체적으로 영상을 통할하는 의미를 피로라고 가정하고 영상의 졸음 인식 관련 연구에 사용되는 DDD 데이터셋을 이용하여 효과적인 키프레임 추출 기법을 적용해 보았으며, 실험 결과 졸음이라는 특정 정보에 대한 해석을 도울 수 있는 의미 있는 요약을 제공하는 키프레임들을 효과적으로 추출하는 분석 기법을 찾아낼 수 있었다.

키워드: 키프레임(keyframe), 영상분석(video analysis), 키프레임 추출(keyframe detection)

I. Introduction

현대 사회에는 많은 형태의 정보가 있는데, 이 중 매우 중요한 정보가 바로 영상정보로서 다양한 정보가 영상으로 생산되고 있으며 그 가치 또한 높아지고 있다. 일반적으로 영상의 분석을 위해 먼저 영상을 프레임(frame)으로 분해하는데, 특정 개수의 프레임으로 분리 시 무작위 또는 수동으로 처리하면 정확도가 떨어질 수 있다. 영상을 대표할 수 있는 프레임을 키프레임이라고 하는데, 영상분석상 키프레임 관련 연구는 주로 영상정보를 요약하거나 분석 대상의 범위를 줄이는 방향으로 전개되어 왔다.

이러한 분석방법을 통해 영상이 가진 의미를 분석하는 것은 매우 제한적인데, 만일 모델을 이용하여 자동적으로 영상의 의미를 따르는 키프레임을 추출할 수 있다면 영상의 의미적 요약에 중요한 기점을 마련할 수 있으며 보다 효율적인 영상 분석의 기반을 마련할 수 있을 것이다.

본 연구에서는 프레임 중에서 영상의 의미를 반영하는 키프레임을 추출 기법을 기반으로 효과적으로 찾아내는 방법을 탐구하였다. 이 경우 키프레임 추출 기법은 영상의 의미 요약에 활용될 수 있으며, 분석을 수행할 경우 적절한 프레임을 자동으로 추출할 수 있게 한다. 영상이 가진 많은 의미 중에서 특히 피로와 관련된 프레임의 추출을 위하여 사람의 안면 영상을 이용한 실험셋인 DDD(Driver Drowsiness Detection)[1] 데이터셋을 활용하여 실험적 검증을 시도하였다.

II. Related Works

1. 영상 압축을 위한 영상분석 방법

영상 데이터를 압축하기 위한 키프레임 검출에 관한 기존 연구들은 지도학습과 비지도학습을 기반으로 한 방법론들이 주로 연구되었다.

지도학습 기반 방법은 모델의 구현이 비교적 간단하고 특정 프레임만 모델을 통과시켜 검출하기 때문에 속도가 빠르지만, 지도학습을 위한 데이터셋 구축에 많은 시간이 소모되고 데이터 셋에 대한 의존성이 매우 큰 단점도 있다. 대표적으로 BI-LSTM[2], CNN-LSTM-GAN[3]을 이용한 방법이 있다. 비지도학습 기반 방법은 데이터셋 구축에 대한 필요성이 필수적이지는 않으나 비교적 속도가 느리다는 단점이 있으며, 대표적으로 k-means clustering을 활용하는 방법[4]이 있다. 많은 연구들이 영상 용량 혹은 내용의 압축을 위해 사용되어 왔으나, 본 연구에서는 영상에서 특정 의미를 나타내는 키프레임을 추출하고자 하였으며, 특히 졸음을 의미하는 키프레임을 추출하고자 하였다.

2. 영상의 의미탐지를 위한 영상분석 방법

영상에서 특정 의미에 맞는 프레임을 찾아내기 위한 연구들은 영상기반의 감성 인식을 주로 다루고 있다. 초기에는 얼굴의 특징을 분류하는 연구가 주로 진행되었는데 얼굴의 정면 이미지를 활용하는 기존 HOG 방법[5]으로 시작하여 자연스러운 영상을 이용한 분석을 추구하고 있다. AFEW 데이터셋[6]은 얼굴의 이차원적 특징을 추출하는데 활용되어 현재는 화남, 역겨움, 행복함, 중립, 슬픔, 놀라움 등 7가지 감성으로 분류하는 방법이 적용되고 있으며, 이를 통해 흥분 정도, 감정 변화를 탐지하는 기술이 연구되고 있다[7]. 대표적으로 CNN-RNN [8]과 Convolution 3D를 활용한 기법[9]이 있다.

III. Keyframe Detection Methodology

영상 속에서 특정 의미를 가진 키프레임을 검출하기 위하여 사람의 안면을 포함하는 영상에서 키프레임을 검출하는 방법론들을 검토하였다. 이때 고려사항으로 키프레임 검출시 속도가 빠르고 정보의 손실이 없어야 하며, 영상 전체에 대한 사전분석이 없이도 프레임 단위로 적용이 가능한 방법론을 채택하였는데, 이는 많은 영상 속에서 수동으로 특정 의미를 표기(tagging)하고 이를 이용하여 다른 영상을 평가하는 방법은 평가자에 따라 특정 의미가 다르게 적용될 수 있으므로, 이러한 상황을 포괄하여 의미적으로 적합한 화상을 올바르게 찾아낼 수 있는 방법론을 탐구하기 위함이다. 최종적으로 두가지 방법이 채택되었다.

1. Method 1: Color Histogram

첫째로 색상분포도를 활용하여 프레임들을 검사하고, 이전 프레임과의 유사도 차이가 크게 나타나는 프레임을 키프레임으로 선정하는 방법이다. 색상분포도는 화상 내부의 픽셀이 가질 수 있는 모든 픽셀값들의 출현 빈도를 나타내는데 사용한다. 이러한 색상분포도는 화상 내 픽셀들의 대략적인 분포를 확인할 수 있고 다른 화상과의 비교 시 활용될 수 있다.

먼저 영상의 프레임에 해당하는 색상분포도를 생성하는데, 영상 데이터는 프레임의 흐름 (색상분포도의 흐름)으로 확인할 수 있다. 이 흐름에서 임계값(threshold) 이상의 변화가 발생하는 시점을 키프레임으로 추출하게 되는데, 이때 cosine 유사도를 활용하여 전후 프레임의 색상분포도의 유사도를 계산하였다. 이 방법은 프레임 단위의 색상분포도를 활용하므로 처리 속도가 빠르고 사전에 특정 데이터에 의존되는 함수나 전처리 방법을 사용하지 않으므로 여러 데이터셋의 활용을 위한 구현이 용이하다.

Fig 1에 색상분포도의 생성과 이를 활용한 키프레임의 추출 방법을 보였다. 우선 입력 영상을 n개의 프레임으로 분리한 후 각 프레임 이미지에 대한 색상분포도를 생성한다. 이후 생성된 색상분포도를 영상의 시간 순서에 따라 나열하게 되고, 각 색상분포도를 이전의 것과 비교하여 유사도를 측정하는데, 측정된 유사도가 설정된 임계값보다 작은 경우 해당 프레임은 키프레임으로 추출된다. 하나의 영상에서 n개의 프레임에 대해 같은 동작을 수행하게 되고 마지막 프레임의 검사가 끝난 후 추출된 모든 프레임을 해당 입력영상에 대한 키프레임으로 정의할 수 있다.

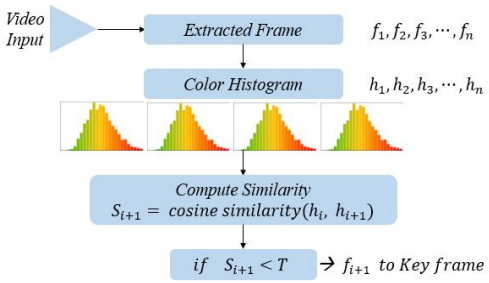


Fig. 1. Key frame detection framework using color histogram

2. Method 2: GMM[10]

둘째로 Gaussian Mixture Model (GMM)을 활용하여 영상의 프레임 흐름을 분석한 후, 프레임 간의 전후 차이가 급격하게 나타나는 부분을 발견하여 해당 프레임을 키프레임으로 추출하는 방법이다. 구체적으로 영상을 프레임으로 분할, GMM을 이용하여 각 프레임에서 개체(object)를 제거하고 배경을 추출한다. 추출된 배경의 픽셀값만큼 원본 프레임에서 차감하여 배경이 제거된 개체 중심의 프레임 정보를 추출할 수 있으며, 개체만 존재하는 각 프레임에서 모든 픽셀을 합연산한 것을 해당 프레임의 행동정보(motion information)로 정의한다. 프레임의 묶음(Window)을 변화시키며 순차적으로 정해진 수의 크기만큼 프레임들이 가진 행동정보를 확인하고 해당 묶음에서 가장 큰 지역최대값(local maxima)을 선정, 각 묶음당 최대값을 가지는 프레임이 키프레임으로 추출된다.

색상분포도를 사용하는 경우 프레임의 흐름에서 큰 변화가 나타나는 것을 감지하기 힘든 경우가 있는데, 두 번째 방법에서는 움직임의 정보를 활용하여 이를 보완하는 효과를 확인할 수 있다. Fig 2는 GMM을 활용한 키프레임 검출 방법을 나타낸다.

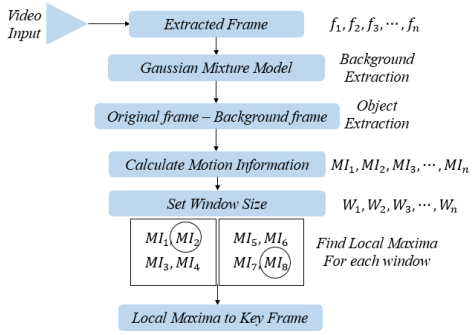


Fig. 2. Key frame detection framework using GMM

IV. Experiments

전장의 두 방법을 특정 의미를 가지는 영상데이터에 적용하여 실험을 진행하였다. 영상 전체에 존재하는 특정 의미는 피로로 규정하였고, 입력 영상데이터로는 DDD 데이터셋을 활용하였다.

1. Driver Drowsiness Detection (DDD) Dataset

DDD 데이터셋은 운전자의 졸음 검출 연구를 위해 제작되었다. 총 18명의 피실험자 (3개 인종, 남/여 구분)에 대한 영상 데이터를 수집하였으며, 피실험자의 안면을 포함하는 운전 상태의 영상을 수집하였으며, 각 영상의 프레임 들은 졸음에 해당될 경우 1, 졸음 상태가 아닐 경우 0의 값이 부여되어 있다. 본 연구에서는 데이터셋에서 무작위로 356개의 영상데이터를 추출하여 실험에 사용하였다.

전장의 방법들을 활용한 키프레임 추출시 목표는 영상속에서 졸음에 해당하는 부분을 모두 찾아내는 것이 아니라 졸음에 해당하는 대표적인 프레임들을 최대한 많이 찾아낼 수 있는지 실험하였다.

2. Method 1 적용 결과

임계값을 변경해가며 경험적인 방법으로 데이터셋에 적절한 값을 찾아내는 방식으로 실험을 진행하였다. 실험 결과는 Fig 3에 실험 결과 하나의 영상에 대해 도출된 키프레임을 예로 보였다.

Fig. 3. Keyframe detection results (threshold = 0.9)

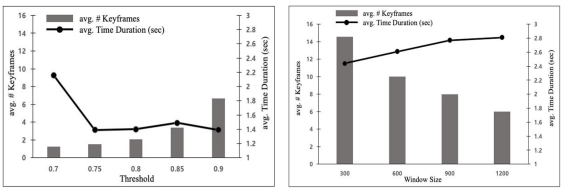
Method 1의 실험 결과 추출된 모든 프레임이 졸음으로 분류된 프레임이었으며, 임계값별 추출 프레임수 및 소요시간은 Fig 5(a)와 같다. Method 1은 모델의 구현 편의성이 뛰어나고 성능이 어느 정도 확보되지만, 임계값의 설정에 따라 모델 성능의 편차가 커서 모델의 해석이 다소 어려울 수 있는 점이 관찰되었다. 또 연속된 모든 프레임을 검사하는 과정에서 비슷한 프레임들이 키프레임으로 추출되는 결과를 보인다. 이러한 점은 영상데이터의 특성에 따라 추출된 키프레임의 품질이 낮아질 수 있음을 나타내며, Method 1의 성능의 영상데이터에 강하게 의존한다고 결론지을 수 있다.

3. Method 2 적용 결과

Method 2의 경우 경험적 방법을 통해 여러 Window 크기를 적용하여 실험을 진행하였다. Fig 4에 추출된 키프레임의 예를 보였다.

Fig. 4. Keyframe detection results (window size = 300)

Method 2 실험 결과도 추출된 모든 프레임이 졸음에 해당하는 프레임이었으며, 묶음별 추출 프레임수 및 소요시간은 Fig 5(b)와 같다. 묶음의 크기가 증가할수록 추출되는 키프레임의 수는 줄어들면서 처리시간이 증가하는 현상을 보인다. 또한 Method 1에 비해 추출된 키프레임간의 유사성이 낮아 보다 감증적인 키프레임의 추출이 가능하다. 이는 묶음을 활용함으로써 연속되는 프레임들이 모두 키프레임으로 추출되는 것을 방지할 수 있는 Method 2의 특성이 잘 나타난 것이다. Method 2의 묶음 300에 해당하는 결과가 소요시간 및 추출 프레임의 성질상 최적의 결과로 판단되었다.



(a) Method 1 (b) Method 2

Fig. 5. Experimental results

V. Conclusions

본 연구에서는 영상에 피로라는 특정 의미를 부여하여 이를 나타낼 가능성이 높은 졸음에 해당하는 프레임을 찾아내기 위하여 키프레임 추출기법을 비교 분석하였다. 실험 결과 키프레임 추출 방법의 적용성에 대해 파악할 수 있었으며, 졸음이라는 요소의 특성상 배경이 제거된 피실험자의 형상에서 보다 높은 확률로 의미 있는 키프레임을 추출할 수 있음을 보였다.

영상의 의미는 피로 뿐만 아니라 다양하게 존재할 수 있다. 본 연구에서는 피로라는 구체적 상황에서 키프레임 추출에 대한 실험을

진행해 보았으나, 다양한 다른 의미에 상응하는 프레임의 추출에 적용할 수 있음을 알 수 있었다. 향후에는 키프레임 추출 방법을 보다 발전시켜 영상에서 도출하고자 하는 의미에 맞도록 적절한 키프레임을 도출하는 방법을 연구하고, 제시된 방법 이외에 강화학습 등 딥러닝 기법을 이용한 새로운 키프레임 추출 방법을 연구하여 영상에서 의미 있는 키프레임을 추출하도록 해야 하겠다.

ACKNOWLEDGMENT

본 논문은 대한민국 정부 (산업통상자원부 및 방위사업청) 재원으로 민군협력진흥원에서 수행하는 민군겸용기술개발사업의 연구비 지원으로 수행되었습니다. (과제번호 20-CM-BD-13)

REFERENCES

[1] Weng, C. H., Lai, Y. H., & Lai, S. H. (2016, November). Driver drowsiness detection via a hierarchical temporal deep belief network. In Asian Conference on Computer Vision (pp. 117-133). Springer, Cham.

[2] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo.Unsupervisedextraction of video highlights via robust recurrent auto-encoders. In Proceedings of the IEEE International Conference on Computer Vision, pages 4633–4641, 2015.

[3] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstmnetworks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[4] De Avila, Sandra EF, et al. "VSUMM: An approach for automatic video summarization and quantitative evaluation." 2008 XXI Brazilian Symposium on Computer Graphics and Image Processing. IEEE, 2008.

[5] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE.

[6] Yu, Z., & Zhang, C. (2015, November). Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 435-442).

[7] Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., & Gedeon, T. (2017, November). From individual to group-level emotion recognition: Emotiw 5.0. In Proceedings of the 19th ACM international conference on multimodal interaction (pp. 524-528).

[8] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2285-2294).

[9] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

[10] Zivkovic, Z. (2004, August). Improved adaptive Gaussian mixture model for background subtraction. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 2, pp. 28-31). IEEE.