시니어 라이프 로깅을 위한 심미적 특징 기반의 행동 요약 시스템

김선대, 류일웅, 유재성, 굴람 무즈타바, 박은수, 김승환, 류은석
가천대학교 컴퓨터공학과
{ele7004, dlfdnd96, poopoo96, mujtaba, dmsehf804, whitekomani}@gc.gachon.ac.kr,
esryu@gachon.ac.kr

Aesthetic Feature-based Activity Summarization for Senior Life Logging

Seondae Kim, Il-Woong Ryu, Jaesung Ryu, Ghulam Mujtaba, Eunsoo Park, Seunghwan Kim and Eun-Seok Ryu
Department of Computer Engineering, Gachon University

요 약

본 논문은 시니어 라이프 로깅을 위한 데이터베이스를 효과적으로 구축하기 위해 영상의 심미적 특징을 통한 행동 별 영상 요약을 소개한다. 실내의 TV 앞에서 오랜 시간을 보내는 시니어의 상태를 체크하기 위해 일반 카메라 또는 360 카메라를 통해 HD 급 화질 이상의 영상을 주기적으로 수집하고, 이를 머신러닝 또는 딥러닝 기반의 행동인식 시스템에 이용하기 위한 전처리 단계에 응용할 수 있는 방법을 서술한다. 이 연구에서는 영상 데이터에서 얻을 수 있는 색상을 이용한 HSV 히스토그램, 영상신호의 Jitter 를 줄이는 고정도, 움직임 에너지 등을 이용하여 짧은 시간 내에 행동별로 구분된 영상(샷, shot)을 자르고 요약하는 방법을 서술한다.

1. 서론

최근 들어 정보통신 기술의 급격한 발전에 힘입어, 음성 신호와 영상 신호를 포함하는 멀티미디어 데이터의 활용이 급속도로 증가하고 있다. 이에 수많은 연구가 음성, 영상신호 및 기타 정보를 이용하여 사람의 상태를 체크하고, 병을 조기 진단하는 시스템, 그리고 원격 진료 시스템 등과 같은 사람들을 위한 헬스케어 시스템에 활용되고 있다[1].

이러한 흐름에 맞추어 본 연구에서는 시니어(노인)라는 특정한 대상으로 한 라이프 로깅 시스템을 서술한다. 본 연구의 라이프 로깅 시스템은 크게 1) 카메라를 통한 영상/음성 데이터 수집; 2) 수집한 데이터를 요약하여 일간/주간/연간 등의 특정 시간대 별로 저장; 3) 축적된 데이터를 바탕으로 인공지능 기반의 분석/판단을 수행하는 시스템으로, 시니어의 상태 및 행동과 그에 대한 패턴을 파악하고, 조기 진단할 수 있는 병이나 급박한 위급상황 등에 대비하는 것이 목표이다.

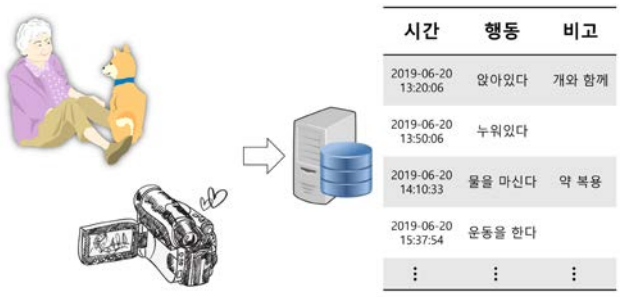

그림 1. 본 연구에서 수행하는 라이프 로깅 시스템의 예.

이전 연구[2]에서는 시니어 라이프 로깅 전 과정에 대한 간단한 실험 및 검증, 활용가능성을 측정하였다. 그 결과로 추가적인 연구로써 시니어로부터 얻은 데이터를 (1), (2), (3)과 같은 절차에 따라 본격적으로 데이터를 수집하고 분석하여 평가를 내린다. 본 논문에서는 앞서 언급한 (1) 영상/음성 데이터 수집 단계에 대한 내용을 주로 서술한다. 본 논문의 구성은 다음과 같다. 2 절에서는 본 연구를 수행하는데 있어서 참고한 관련 연구들을 서술한다. 3 절에서는 본 연구에서 수행하고 있는 영상 캡처 및 요약 시스템에 대해 서술한다. 4 절에서는 제안하는 기법을 통해 실험한 내용을 설명하고, 논의할 점을 서술한다. 마지막으로, 5 절에서는 본 논문에 대한 결론을 맺는다.

2. 관련 연구

이 장에서는 본 연구를 수행하기 위한 두 가지 관련연구를 소개한다. 첫번째로, 연구 [3]은 최근 YouTube, Flickr, Instagram, Facebook 등 동영상을 이용할 수 있는 다양한 플랫폼에서 사용자가 영상을 보기 전, 영상의 대표 이미지(썸네일)을 뽑아 자동으로 추천해주는 기법이다. 또한 연구 [5]는 영화와 같은 시공간적인 데이터를 담고 있는 영상에 대해 장르를 추출하는 연구를 수행하였다. 이는 시각적인 데이터와 함께 음성 데이터를 통해 특징을 추출하고 영화의 장르를 예측하는 시스템을 나타내고 있으며, 추후 시니어를 위한 라이프 로깅을 위해 다음 단계의 연구로써 비슷한 맥락으로 수행할 예정이다.

2.1. 영상 요약 및 썸네일 추출 기법

[3]에서는 영상의 대표적 키프레임인 썸네일(Thumb-nail)을 총 세 가지의 단계를 거쳐 출력한다. 그림 2 와 같이 영상이 입력되면 1) 프레임 필터링 2) 키프레임 추출 3) 썸네일 선택을 순차적으로 수행하여 키프레임 후보군 중 영상을 대표하는 최고의 키프레임을 출력한다. 이 랭킹은 특히 (3)에서 색상, 텍스처, 이미지 품질, 이미지 조화도 등을 고려하여 키프레임마다 랭크가 정해진다.

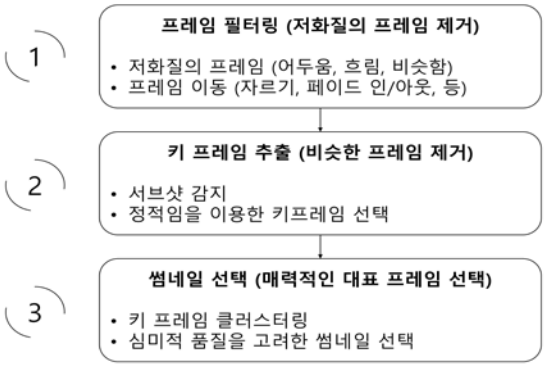

그림 2. 자동 썸네일 추출 시스템의 동작 순서[3].

(1)에서는 저화질의 프레임을 필터링한다. 색상 값이 어둡거나, 흐리거나, 비슷한 프레임들의 경우 굳이 다음 단계를 거칠 필요가 없는 일종의 쓸모 없는 프레임들을 제외하는 역할을 한다. 웹에 올려지는 동영상 중 대부분은 영상을 여러 영상편집 툴로 편집되어 업로딩된다. 예를 들면, 페이드 인/아웃, 영상 자르기, 디졸브 (영상이 흐려지거나 분해되면서 다음 샷으로 넘어가는 효과), 등 다양한 영상효과들이 들어가게 된다. 이처럼, 키프레임 추출에는 쓸모 없는 프레임을 최대한 제거하게 되어 썸네일 후보의 선택 확률을 높인다.

(2)에서는 (1)에서 필터링 된 프레임들을 K-means 알고리즘을 통하여 클러스터링한다. 영상에서 비슷한 프레임끼리 클러스터링하게 되면, 시각적으로도 비슷한 프레임끼리 묶이게 되고, 샷의 경계를 쉽게 구분할 수 있고, 샷과 같은 연속적인 프레임들에 대한 중심값 (Centroid)을 얻을 수 있다. 이 중심값을 통하여 그 중심값과 제일 가깝고 대표적인 키프레임을 추출할 수 있게 된다. 클러스터링을 위한 특징은 OpenCV 를 활용하여 얻어낸 HSV 히스토그램과, 엣지 히스토그램을 이용한다. 또한 움직임 에너지를 연속적인 두개의 프레임 간에 계산하여, 움직임 편차를 추출해 정적인 부분을 필터링한다. 움직임 고정도가 강하면 움직임이 적고 낮은 품질의 불필요한 프레임을 제외시킬 수 있다.

이 연구에서는 컬러와 텍스처, 그리고 이미지 품질과 같은 심미적인 특성들로 구성한 특징 벡터를 통해 프레임의 질을 따져 필요한 프레임을 선택하여 추출한다. 이 연구에서는 썸네일과 함께 GIF (Graphics interchange format) 파일 그리고 키샷을 추출해낼 수가 있다. GIF 파일과 키샷은 추출하는 영상의 길이(초), 샷의 범위, 필터링, 프레임, 해상도 등 정해진 파라미터 내에서 영상을 추출해낼 수 있다. 입력 영상에서 측정된 여러 샷들의 범위에 따라 임시 파일로 분할된 영상을 추출하고, 추출된 샷들을 합치게 된다. 영상 및 GIF, 키프레임 추출 및 출력의 전 과정은 FFmpeg 을 통해 이미지 파일 포맷 및 영상 포맷의 파일로 추출한다.

2.2. Spatio-temporal 데이터를 위한 딥러닝 네트워크

본 연구에서 영상을 요약하여 딥러닝 네트워크에 응용하려는 아이디어는, 그림 3 과 같이 영상 데이터를 멀티 모달 형식의 네트워크를 추후에 적용하기 위함이다. [5]는 영화 트레일러를 시간 순서대로 특징을 추출하여 장르 예측을 수행한다. 딥 러닝 네트워크로는 당시 이미지 인식 챌린지인 ImageNet 에서 뛰어난 성능을 보인 ResNet 을 이용하였다. 영화 트레일러와 같은 짧고 중요한 영상을 입력소스로 받아들여 각 프레임의 특징을 추출하는데 사용한 후 Convolution-Through-Time Module(CTT)로 명명한 모듈을 이용하여 시간 순서대로 연속적인 프레임들을 m 차원의 특징 벡터로 각각 만들어 풀링하고, 이를 특징으로 이용한다.

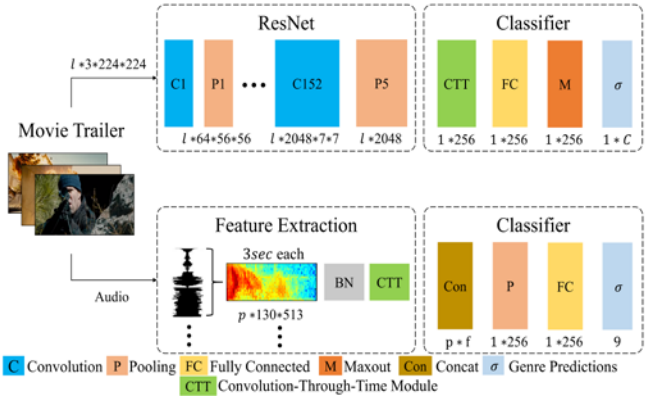

그림. 3 CTT 모듈을 활용한 Convolution Network 를 적용한 영화 장르 분류 시스템[5].

CTT 모듈은 입력된 영상에서 16 프레임씩 특징을 뽑아 이를 시간 순서대로 Convolution 및 Max pooling 을 이용해 특징을 순서대로 정리한 후 결과값으로 장르를 예측한다. 또한 시각적 데이터(프레임)과 별개로 오디오의 특징을 Spectrogram 을 통하여 오디오로만 독립적으로 영화 트레일러의 장르를 예측하고 추후 이 결과를 합산한 확률을 통해 장르를 예측한다. 본 연구에서는 추후 이러한 멀티 모달리티 네트워크로 시니어의 인지능력을 체크할 것이다.

3. 시니어를 위한 행동 별 라이프 로깅

시니어는 보통 몸이 편치 않거나 여가시간이 많아, 대부분의 시간을 집 안에서 보내게 된다[4]. 본 연구에서는 집에서 시간을 보내는 시니어를 대상으로 한정하여 연구한다. 시니어들은 또한 TV 를 매우 자주, 많은 시간동안 시청하기 때문에, TV 주변의 일반 카메라 또는 360 카메라를 이용하여 HD (1280 * 720)급 이상의 영상과 음성 데이터를 수집하여 분석하는 상황을 가정한 후 연구를 수행한다.

3.1. 행동 인식을 위한 행동 별 영상 데이터의 저장

현재 본 논문의 시니어 라이프 로깅을 위한 영상 요약은 시니어가 약한 움직임이나, 특정 행동들을 했을 때를 포착하여 이를 딥 러닝 기반의 네트워크에 넣을 전처리 부분이다. 입력 받는 영상은 사용자가 원하는 범위의 시간, 또는 자동차

블랙박스와 비슷하게 24 시간을 모두 촬영한 영상이다. 분할된 영상에서 시니어가 행동하는 부분의 샷을 추출하여 영상으로 수집하여, 각 영상이 하나의 행동으로, 시간이 측정되어 정리되는 시니어 라이프 로깅 데이터베이스를 만든다.

이 데이터베이스는 추후 현재 개발중인 인공지능 모델의 입력 소스가 되며, 행동별로 나누어진 데이터는 행동 인식 모델의 학습을 위한 UCF-101[6] 등과 같은 데이터셋에 호환될 수 있으며, 추가적으로 이는 본 연구에서 개발 중인 행동인식 모델의 Fine-Tuning 에 유리하다. 추후 시니어에 대한 자료들을 주기적으로 업데이트하여, 시니어를 위한 i)행동인식 및 시니어 라이프 로깅을 위한 행동 별 데이터셋 제작, ii)개발중인 인공지능 모델 개발에 적용할 수 있도록 한다.

3.2. 행동 별 파일 출력을 위한 파라미터 조정

현재 오픈소스[7]에서 공식적으로 지원하는 파일 출력의 범위는 키프레임 그리고 GIF 를 개수 n 을 정해서 커맨드 라인으로 원하는 개수를 입력하여 출력할 수 있다. 다만, 영상을 출력하는 점에 있어서는 현재 기본적으로 제공하는 15 초에 맞추어, 요약되고 병합된 영상만을 출력할 수 있다. 현재의 코드는 사용자 파라미터로 15 초 이외의 영상 요약 제한 시간을 입력하게 되면 오류로 인해 출력되지 않는다. 따라서 이 점을 출력부분에서 필수적으로 고쳐야 한다.

그림 4. 본 연구에 응용하기 위해 변경한 부분(굵은 글씨).

기존 연구는 샷 후보군을 여러 개 임시적으로 뽑은 다음 그 이후에 제한된 동영상의 길이와 샷 후보군에 심미적으로 부여된 점수에 따라, 샷을 일괄적으로 결합하는 방식이어서, 시니어의 행동들이 많이 생략되고 잘리는 경우가 있다. 따라서 본 연구에서 응용하기 위해 수정한 점은 다음 그림 4 에서 표시한 바와 같다. Parser 부분에서는 영상의 초당 프레임과 상관없이 휴리스틱하게 앞, 뒤 부분의 영상을 0.5 초 구간을 삭제하는 코드를 제외한다. 기존 연구는 웹 영상의 대표 썸네일 선택을 위한 것이기 때문에, 평균적으로 행동이 느린 시니어를 위해서는 최소 샷의 길이를 늘려주어 적당히 샷을 분포시키기 위함이다.

또한, Highlighter 에서는 너무 짧거나 긴 샷을 제외한다. 이 부분에서 중요한 점은 노인의 행동이 수행되는 시간과 밀접한 관련이 있다. 너무 짧은 샷은 크게 의미가 많지 않은 불필요한 결과를 낳는다. 반대로 너무 긴 샷은 행동이 길지 않은 이상 행동 별로 나누려는 샷의 경계에서 변화가 일어나서 그 범위 안의 샷 경계를 자동적으로 만들게 되므로 크게 관련이 없다. 따라서 휴리스틱하게 최소 샷 길이보다 2~3 배 이상 최대 샷의 길이를 지정한다.

추가적으로, 정적인 샷을 제외할 수 있다. 이는 2.1 장에서 잠시 설명하였듯이, t 프레임과 그 다음의 t+1 프레임 간의 픽셀 평균 변화량(움직임 에너지)을 구해 변화량이 정해진 파라미터를 넘지 않으면 정적인 프레임으로 계산하여 제외하게 된다. 하지만, 앞서 언급하였듯이, 기존 연구[3] 에서는 웹 영상의 기준으로 파라미터를 적용하였다. 본 연구에서는 집과 같은 곳에서 고정된 카메라로 촬영하는 환경이므로, 정적도를 기존보다 조금 더 낮추어야 한다. 추가적으로, Parser 부분에서 영상의 전체 프레임에 휴리스틱하게 적용되던 알고리즘을, Highlighter 에서는 샷 후보군의 단위로 적용시켜 1 초 전후의 앞 뒤 영상을 샷 구간에서 제외시킨다. 이는 각 행동구간별 샷을 추출할 시에 앞 뒤 행동이 다를 경우, 앞 행동이 뒤 행동구간에 샷의 범위가 걸쳐서 생기는 오차를 줄이기 위함이다.

마지막으로, Generator 에서는 나누어진 샷 구간을 프레임의 인덱스로 추출하는 과정이다. 기존에는 정해진 길이(15 초)를 벗어나는 파라미터를 입력하면 출력이 되지 않는다. 이를 FFmpeg 을 이용하여 버퍼에 입력된 샷을 삭제하지 않고 그대로 추출한다. 추출한 샷에 대한 영상들은 각각의 행동을 담고 있는 대로 편집되어 출력된다. 사용자가 원하는 경우 키프레임 및 GIF, 키 샷으로만 병합된 영상 또한 얻을 수 있다.

4. 실험 결과

영상	초당 프레임	입력된 영상의 길이	평균 프레임 편차	촬영된 장소	처리 시간 (초)	출력 결과 (샷 개수/길이)
*남성1	29.98	61.2초	0.0467	실내 TV앞 (멀리)	5.01	9/41.2초
남성2	25.0	132.2초	0.0813	실내 TV앞 (가까이)	8.02	24/94.32초
여성1	29.97	273.6초	0.0528	실내 (운동)	20.07	42/196.83초
여성2	30.00	133.8초	0.0807	실내 TV앞 (가까이)	9.09	23/103.76초
사람들	23.98	292.8초	0.0813	실내 TV앞 (가까이)	18.05	45/198.95초

표 1. 본 연구를 적용한 실내 영상 실험 데이터의 정보.

(* : 직접 피실험자를 촬영한 영상)

기존 연구[3]의 코드를 이용할 경우, 수행한 행동의 많은 부분이 잘리고 제대로 기록되지 않는다. 본 논문에서 소개한 대로 Parser, Highlighter 그리고 Generator 를 수정하여 실험을 진행한다. 본 연구에서 수행한 실험은 여러 행동이 담긴 영상을 얼마나 구간을 잘 나누어 분류하는지 검증한다. 1) TV 를 보면서 다양한 활동을 하는 성인 남성을 시니어라고 가정하고 촬영한 영상과 2) 웹 영상 중 TV 앞을 가정하고 촬영한 영상들을 통해 실험하였다. 특히, (1)의 경우 다양한 행동을 수행하는 피실험자를 촬영하여 녹화하고, 이후 서버급의 Intel Zeon * 2ea, 64GB RAM 급 서버에서 실험을 진행하였다. 약 4~5 스레드 정도를 사용하였으며, 연속적으로 촬영된 영상을 입력 데이터로 이용한다. 프레임 간의 변화량 편차의 제한은 0.0025 로 줄이고 실험하였다.

그림 5. 제안하는 방법으로 추출한 행동 별로 추출된 샷의 키프레임 (표 1 에 표기된 영상의 순서와 동일).

표 1 에 사용된 영상파일은 그림 5 에서 볼 수 있듯이 일반적으로 실내에서 생활하는 사람들의 모습을 담고 있다. 샷의 길이조정은 평균 2~8 초로 수정하였다. 남성 1 영상은 직접 피험자가 여러 행동들을 수행한 모습을 촬영하였다. 걷기, 먹기, 자기, 일어서기, 앉기, 넘어지기 등 TV 와 약 4~5M 떨어진 곳에서 다양한 행동을 수행하였고, 각각의 행동에 맞게 샷이 구간별로 알맞게 추출되었다.

이와 같이 남성 2 영상에서도, 샷 별로 행동하는 모습이 제각각 다르고, 해당하는 행동들도 다르게 나타난다. 다만, 지정한 샷 범위에 따른 의미 있는 샷을 추출하는 것이기 때문에, 연속적인 여러 개의 샷에 있어서는 비슷한 행동이 더 추출될 수 있다. 여자 1 영상은 본 연구의 샷 경계 성능을 웹 비디오에서도 효과가 있는지 검증하기 위해 실험하였다. 여기서 하나의 아닌 여러 대의 카메라로 촬영된 영상은 카메라마다 모습이 프레임의 차이가 극명하게 달라지기 때문에, 샷의 길이가 행동별로 일정하지 않음을 확인하였다.

여자 2 영상의 경우, 샷 경계의 차이가 일반적인 영상보다 차이가 크기 때문에 결과는 확실히 남자 1,2 영상보다 뚜렷한 샷 구간을 얻을 수 있다. 하지만, 이렇게 샷 경계의 차이가 있더라도, 사람들 1 과 같은 영상에서는 사람들이 많기 때문에 픽셀의 변화량, 움직임 에너지가 급변하는 경우가 많아 샷 경계가 모호해지는 경우가 생겨 행동 별 구간의 샷을 추출하는 성능이, 객체가 많을 경우 다소 떨어지는 한계를 보인다.

5. 결론

본 논문은 시니어 라이프 로깅을 위한 효과적인 데이터베이스를 구축하기 위해 영상의 심미적 스코어링을 통한 행동 별 영상 요약을 수행하였다. 실내의 TV 앞에서 오랜 시간을 보내는 시니어의 상태를 체크하기 위해 일반 카메라 또는 360 카메라를 통해 HD 급 화질의 원본 영상을 얻어, 이를 머신러닝 또는 딥러닝 기반의 행동인식에 이용하기 위한 전처리 단계에 응용할 수 있는 방법을 서술하였다. 영상의 심미적 요소, 움직임 에너지 등을 이용하여 짧은 시간 내에 행동별로 영상을 자르고 요약한다. 요약된 영상 데이터는 평소 시니어의 상태확인 및 이상징후 판단에 사용되어, 시니어의 삶의 질을 향상시킬 수 있다.

Acknowledgement

본 연구는 경기도의 경기도 지역협력연구센터 사업의 일환으로 수행하였음. [GRRC-가천 2017(B01), 시니어 라이프 로그 기반 행동 분석]

참고문헌

[1] Miotto, Riccardo, et al. Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics, 2017, 19.6: 1236-1246.

[2] 김선대, 박은수, 정종범, 구자성, 류은석, 딥 러닝 기반의 API 와 멀티미디어 요소를 활용한 시니어 라이프 데이터 수집 및 상태 분석. 한국방송미디어공학회 하계학술발표대회 논문집, 2018, 244-247.

[3] Song, Yale, et al. To click or not to click: Automatic selection of beautiful thumbnails from videos. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016. p. 659-668.

[4] 황남희, 한국 노년층의 여가활동 유형화 및 영향요인 분석. 보건사회연구, 2014, 34.2: 37-69.

[5] Werhrmann, Jônatas; BARROS, Rodrigo C. Movie genre classification: A multi-label approach based on convolutions through time. Applied Soft Computing, 2017, 61: 973-982.

[6] Soomro, Khurram; ZAMIR, Amir Roshan; SHAH, Mubarak. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[7] Song, Yale, Yahoo Hecate, https://github.com/yahoo/hecate, 2016. [Accessed on May 15th. 2019.]