# 스마트폰에서의 시공간적 중요도기반 비디오 요약

이원범, 박인규
인하대학교 정보통신공학부
luckybeom@gmail.com, pik@inha.ac.kr

# Spatiotemporal Saliency-Based Video Abstract on a Smartphone

Won Beom Lee, In Kyu Park
School of Information and Communication Engineering, Inha University

## 요 약

본 논문에서는 동영상을 시공간적 중요도 기반으로 요약하는 기법을 제안한다. 동영상 내에서 각 프레임의 중요도를 평가하여 높은 우선순위를 가지는 구간들의 집합으로 요약을 수행한다. 화면내의 얼굴면적의 비율, 영상의 복잡도를 통하여 각 프레임이 가지는 공간적 중요도를 분석하고 인접한 프레임간의 비교를 통해 밝기 히스토그램과 움직임(motion)의 양을 추정함으로써 시간적 중요도를 구한다. 에지 보존 스무딩 필터를 밝기 히스토그램에 적용하여 장면 전환을 검출한다. 분리된 장면들로 과분할 구조를 가지는 계층적 트리를 생성하여 사용자가 요구한 재생길이를 가지는 동영상을 자동으로 저작한다. 본 논문에서는 동영상 분석 및 저작을 제한적인 환경인 스마트폰에서 효과적으로 작동하도록 구현 및 최적화를 수행하였다.

## 1. 서론

최근 멀티미디어 기기의 발전과 UCC(user created contents)의 저작 및 공유가 활성화되어 수많은 동영상 콘텐츠가 온라인상에 존재한다. 이러한 동영상 콘텐츠는 제목, 키워드 등으로 검색을 하지만 실제적으로 동영상의 내용을 분석하기 위해서는 반복적이고 불필요한 구간까지 모두 확인하게 되어 시간이 낭비된다. 또한 압축기술의 발전에도 불구하고 동영상은 여전히 많은 저장공간을 요구한다. 따라서 내용측면에서 동영상 요약기술이 요구되며, 군집화 기법[1], 그래프 모델[2] 등을 도입하여 많은 연구가 진행되고 있다.

본 논문에서는 원본 동영상에서 추출한 시공간적 중요도를 기반으로 동영상 요약기법을 제안한다. 전체 프레임에서 중요한 프레임에 대한 척도를 제시하고 최대한 중요한 프레임을 보존하면서 사용자의 요구조건에 맞춰 비디오 요약을 수행한다. 이러한 비디오 요약기술은 마치 동영상의 하이라이트를 자동으로 생성해주는 것과 유사하다. 또한 같은 주제의 여러 동영상을 통합할 때에도 핵심적인 부분만 추출하여 비교적 짧은 재생길이의 비디오로 재구성하게 만든다. 그림 1 은 비디오 요약의 전체과정을 도시하였다.

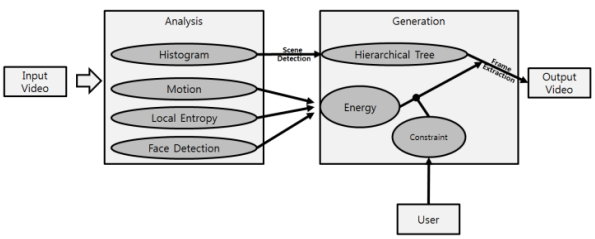



그림 1. 비디오 요약의 전체과정

## 2. 시공간적 중요도 평가

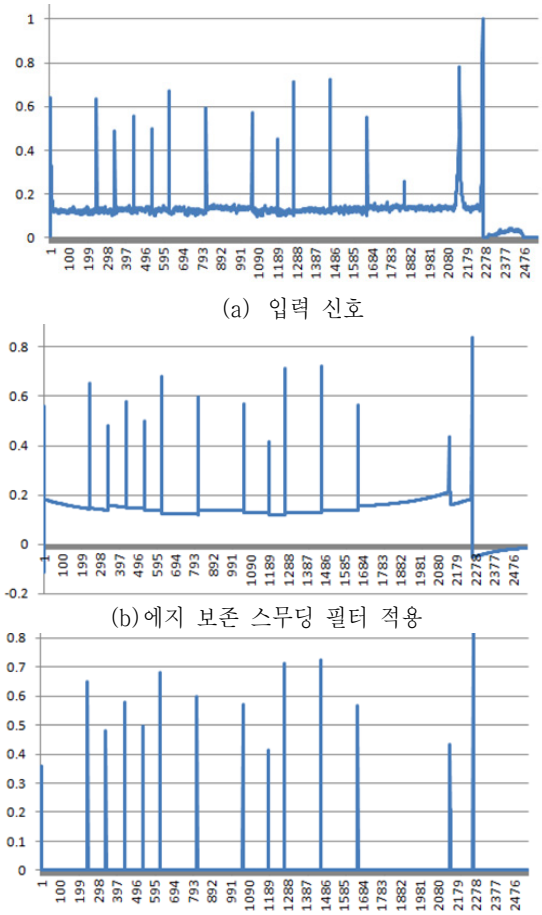
동영상을 요약하기 위해서는 중요한 프레임으로 지정되는 기준에 대하여 명확히 정의해야 한다. 따라서 단일 영상에서 추정 가능한 공간적 중요도와 인접한 프레임간의 상관관계를 통해 추정하는 시간적 중요도를 통하여 프레임을 평가한다.

본 논문에서는 영상 자체의 정보의 양을 의미하는 지역적 엔트로피와 얼굴 검출을 통해 공간적 중요도를 정의한다. 지역적 엔트로피는 지역적 히스토그램을 기반으로 화소 단위의 엔트로피를 구하고 영상의 전체의 엔트로피를 구한다. 그리고 일반적으로 동영상에서는 상대적으로 얼굴을 포함한 프레임이 중요할 가능성이 높으므로 [3]의 방법을 채택하여 얼굴의 개수와 차지하는 면적을 구하여 얼굴에 대한 중요도를 평가한다.

한편, 시간적 중요도는 인접한 프레임에서의 움직임의 양과 장면 전환 검출을 통해 유사한 장면의 지속성으로 정의한다. 인접한 프레임 사이에서 움직임 벡터를 추정하고 밝기 히스토그램의 차이를 통해 장면 전환을 검출하였다. 이는 비교적으로 연산량이 적고 카메라의 떨림으로 발생하는 오차를 줄일 수 있어 효율적이다.

장면 전환 검출을 위해 상대적으로 히스토그램의 차이가 많은 프레임들을 선택해야 한다. 하지만 기준이 모호하기 때문에 본 논문에서는 입력 신호와의 유사성과 단조로운 신호간의 균형을 유지하는 에지 보존 스무딩 필터[4]를 히스토그램 차이 그래프에 적용하였다. 그리고 지역적 최대값을 가지는 신호만을 보존한다. 그림 2 는 히스토그램 차이 그래프의 개선 과정을 도시하였다.

최종적으로 시공간적 중요도는 각 가중치를 적용한 움직임, 엔트로피, 얼굴의 영역의 에너지의 합으로 정의한다.



(a) 입력 신호

(b)에지 보존 스무딩 필터 적용

(c) 장면 전환 검출

그림 2. 히스토그램 차이 그래프의 개선 과정

## 3. 계층적 트리 생성 및 비디오 요약

본 논문에서 제안하는 비디오 요약 기법은 본래의 내용을 최대한 상실하지 않는 범위 내에서 원하는 길이의 동영상을 생성하는 것이다. 따라서 미리 생성한 계층적 트리를 이용하면 분석과정을 생략하고 다양한 재생길이를 가지는 동영상을 생성할 수 있다.

계층적 트리를 생성하기 위해 장면 전환으로 검출된 프레임들로 구성된 리스트를 생성한다. 계층적 트리의 최상위 노드는 전체프레임이고 밝기 히스토그램의 차이를 기준으로 분리하여 하위노드를 생성한다. 또한 장면 전환 리스트도 시간을 기준으로 분리하여 하위노드에 할당한다. 이 과정은 최하위 노드를 더 이상 분할할 수 없을 때까지 반복한다.

과 분할된 계층적 트리에서 요구 재생 길이에 만족하는 노드를 선택하기 위해서 자식 노드를 통합하려는 에너지와 유지하려는 에너지를 정의하여 서로 비교하여 높은 에너지의 성질을 따른다. 통합에너지는 자식노드의 재생길이 및 중요도 곡선이 복잡할수록 높다. 유지에너지는 인접한 형제노드와의 장면 전환이 크고 분할 순서에 따른 레벨차이가 클수록 높다. 통합 및 유지에너지를 바탕으로 요구 재생 길이를 충족시킬 때까지 최하위 노드부터 단계적으로 병합을 수행한다. 또한 시각적인 결함을 줄이기 위해 최소 재생길이를 만족하도록 추출하고 노드의 최대 개수를 제한하여 비교적 짧은 구간의 프레임만 선택되는 것을 방지한다.

## 4. 실험결과

지역적 엔트로피 연산은 상당히 많은 수의 로그 연산이 필요하다. 윈도우의 크기에 따라 한정적인 로그 값이 필요하여 참조테이블을 통해 해결하였다. 또한 고정 소수점 연산으로 부동소수점 데이터의 처리속도 향상하였으며, 영상 처리의 해상도, 윈도우의 크기, 프레임 추출 주기 등의 다양한 LOD(level of degree)설정에 대한 성능분석을 통해 최적의 파라미터를 구하였다.

표 1 에 ARM Cortex A9 1.2GHz Dual Core CPU 를 장착한 삼성전자의 Galaxy II 에서 640×360 해상도의 동영상 분석에 대한 수행시간을 제시하였다.

표 1. 수행시간 분석 (단위: 프레임, 초)

| | 동영상 1 | 동영상 2 | 동영상 3 |
|---|---|---|---|
| 총 프레임 수 | 870 | 3000 | 5640 |
| 디코딩 | 19.27 | 61.29 | 99.04 |
| 밝기 히스토그램 | 0.56 | 2.07 | 4.40 |
| 엔트로피 | 3.20 | 12.93 | 23.35 |
| 얼굴검출 | 1.99 | 8.02 | 15.14 |
| 합계 | 25.02 | 84.31 | 141.93 |

## 5. 결론

본 논문에서는 동영상에 대해 시공간적 중요도를 분석하여 대표적인 부분을 최대한 보존하는 요약 기술을 제안하였다. 다양한 환경에서 취득한 동영상을 스마트폰에서 사용자의 입력에 따라 자유자재로 재생길이를 조절하여 요약된 동영상을 생성할 수 있다. 또한 여러 개의 동영상을 통합하여 저작하는 방식으로 응용할 수 있다. 이러한 요약 기술을 통해 사용자들은 수많은 비디오에 대해 요약된 정보를 쉽게 얻을 수 있고 제약이 많은 저장공간에서 이득을 볼 수 있다.

## 감사의 글

본 연구는 삼성전자(주)의 지원을 받아 수행되었음.

## 참고문헌

[1] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, “Adaptive key frame extraction using unsupervised clustering,” *Proc. of International Conference on Image Processing*, vol. 1, pp. 866-870, Oct 1998.

[2] C. W. Ngo, Y. F. Ma, and H. J. Zhang, “Video summarization and scene detection by graph modeling,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 2, Feb 2005.

[3] P. Viola and M. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, May 2004.

[4] L. Xu, C. Lu, Y. Xu, and J. Jia, “Image smoothing via $L_0$ gradient minimization,” *ACM Trans. on Graphics*, vol. 30, no. 6, Dec 2011.