정규논문 (Regular Paper)

방송공학회논문지 제18권 제2호, 2013년 3월 (JBE Vol. 18, No. 2, March 2013)

<http://dx.doi.org/10.5909/JBE.2013.18.2.185>

ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

스마트폰에서의 시공간적 중요도 기반의 비디오 요약

이 원 범^{a)}, 윌 리 엄^{a)}, 박 인 규^{a)†}

Spatiotemporal Saliency-Based Video Summarization on a Smartphone

Won Beom Lee^{a)}, Williem^{a)}, and In Kyu Park^{a)†}

요 약

본 논문에서는 스마트폰의 플랫폼으로 하여 시공간적 중요도 기반으로 비디오를 요약하는 효율적인 기법을 제안한다. 제안하는 기법은 주어진 비디오에서 카메라 및 물체의 움직임에 강건한 색상 히스토그램의 차분으로 장면 전환을 검출하고 연속적인 프레임간의 유사성, 얼굴의 영역, 개별 프레임(frame)의 중요도를 통해 시공간적 중요도를 분석한다. 그리고 검출된 장면 전환을 이용하여 과분할된 계층적 트리를 생성하고 비디오 분석 과정에서 계산한 병합 및 유지 에너지를 이용하여 반복적으로 갱신한다. 또한 갱신된 계층적 트리에서 사용자가 요구하는 재생 길이와 최소 구간 길이를 충족하고 동시에 높은 중요도를 가진 노드들로부터 탐욕 알고리즘(greedy algorithm)을 통해 프레임을 추출한다. 실험 결과 상용 스마트폰에서 2분길이 분량의 입력 비디오를 10초 내외의 수행시간으로 요약할 수 있었으며, 그 결과는 상용 비디오 편집 소프트웨어인 Muvce보다 우수함을 보였다.

Abstract

In this paper, we propose a video summarization technique on a smartphone, based on spatiotemporal saliency. The proposed technique detects scene changes by computing the difference of the color histogram, which is robust to camera and object motion. Then the similarity between adjacent frames, face region, and frame saliency are computed to analyze the spatiotemporal saliency in a video clip. Over-segmented hierarchical tree is created using scene changes and is updated iteratively using merge and maintenance energies computed during the analysis procedure. In the updated hierarchical tree, segmented frames are extracted by applying a greedy algorithm on the node with high saliency when it satisfies the reduction ratio and the minimum interval requested by the user. Experimental result shows that the proposed method summarizes a 2 minute-length video in about 10 seconds on a commercial smartphone. The summarization quality is superior to the commercial video editing software, Muvce.

Keyword : video summary, video segmentation, scene change detection, saliency, smartphone

a) 인하대학교 정보통신공학부 (School of Information and Communication Engineering, Inha University)

† Corresponding Author : 박인규(In Kyu Park)

E-mail: pik@inha.ac.kr

Tel: +82-32-860-9190

※ 본 연구는 (주)삼성전자와 인하대학교의 지원으로 수행되었음.

· Manuscript received August 20, 2012 Revised January 7, 2013

Accepted March 4, 2013

I. 서 론

최근 다양한 멀티미디어 기기의 발전과 광대역 네트워크를 통해 UCC (user created contents)의 저작 및 공유가 활성화되고 있다. 특히 기업 및 개인에 의해 생성된 비디오 콘텐츠는 카메라가 장착된 모바일 장치의 증가로 손쉽게

게 취득되고 있으며, UCC를 공유하는 문화현상에 따라 YouTube 등에 업로드 건수가 폭발적으로 증가하고 있는 추세이다. 이와 같은 현상으로 인해 실제적인 비디오의 내용을 탐색 및 분석하기 위해 많은 시간이 소요되며, 비디오 압축 기술의 발전에도 불구하고 여전히 많은 저장 공간을 요구하고 있다. 따라서 넘쳐나는 비디오를 효과적으로 탐색하고 사용자가 쉽게 이해하도록 하기 위한 비디오 요약 기술의 수요가 급증하고 있다.

한편 최근 널리 보급된 스마트폰은 고해상도 카메라와 GPS 수신기와 같은 다양한 센서를 장착하고 있으며, 안정적인 고속 무선 네트워크를 통해 자료를 활발히 전송할 수 있다. 또한 하드웨어 기술의 발전과 함께 프로세서의 성능 및 디스플레이의 품질도 향상되고 있다. 이와 같은 환경에서 사용자들은 스마트폰을 이용하여 수많은 영상과 비디오를 취득하고 있으며 이를 효율적으로 관리하기 위한 도구를 필요로 하게 되었다. 비디오 요약은 스마트폰 사용자가 자신의 스마트폰에 저장된 비디오를 빠른 시간에 효율적으로 탐색하고 선택할 수 있는 적절한 도구로서 사용될 수 있다.

비디오 요약 기법은 크게 샘플링(sampling) 기법과 세그먼테이션(segmentation) 기법으로 분류할 수 있다. 샘플링 기법은 특정 규칙에 따라 비디오에서 프레임(frame)을 추출하여 재구성하는 것으로 비디오 특성에 따라 중요한 구간에서 보다 많은 샘플링을 수행하는 비균등 샘플링

(nonuniform sampling) 기법이 많은 연구에서 적용되고 있다. Bennett 등은 움직임(motion)과 장면 변화에 따른 비균등 샘플링을 적용하였다^[1]. DeMenthon 등은 특징 공간에서 표현된 곡선을 이진 곡선 분할 기법(binary curve splitting)을 이용하여 반복적으로 단순화함으로써 키 프레임을 선택하였다^[2]. Zhuang과 Hanjalic 등은 비감독적 클러스터링 기반의 키 프레임 추출 기법을 제안하였다^{[3][4]}. Chang 등은 프레임간의 semi-Hausdorff 거리기반의 그래프를 정의하고 Quine-MacCluskey 알고리즘을 이용하여 최소가 되는 프레임의 집합을 구하였다^[5]. 그리고 다양한 방법을 통해 추출한 프레임 간의 합성을 통해서 보다 축약된 형태로 비디오 요약을 수행하는 연구도 진행되고 있다. Kang 등은 모든 프레임의 중요도 지도를 이용하여 최대의 중요도를 가질 수 있도록 3차원 시공간 볼륨(volume)간의 패킹(packing) 및 병합 알고리즘을 제안하였다^[6]. Pritch 등은 활성 에너지와 불연속성 에너지로 정의한 비용 함수(cost function)를 최소화하는 기법을 제안하여 하나의 프레임에서도 많은 사건(event)을 보여줄 수 있었다^[7]. Mei 등은 비디오에서 추출한 키 프레임의 합성을 통해 한 장의 합성 영상을 생성하는 기법을 제안하였다^[8].

세그먼테이션 기법은 먼저 비디오 내에서 의미단위로 분리하고 중요한 구간을 추출하여 재구성한다. Ngo 등은 모든 프레임에 대해 무향그래프를 생성하고 정규화된 컷(normalized cut) 알고리즘을 이용하여 클러스터로 분할하

그림 1. 제안하는 시스템의 전반적인 개요도
Fig. 1. Overview of the proposed system

였다^[9]. 이 클러스터는 다시 유향그래프로 표현되며 최단 경로 알고리즘(shortest path)을 적용하여 효율적으로 비디오 장면을 검출하였다. Gong 등은 SVD (singular value decomposition)통해 시각적인 특성에 대한 행렬을 유도하고 이를 이용한 비디오 요약 기법을 제안하였다^[10]. Yu 등은 비디오에서 검출된 샷(shot)에 대하여 사용자의 의한 방문 링크(visiting links)와 샷 간의 유사성 링크로 우선순위를 정하는 방법을 제안하였다^[11]. Herranz 등은 반복적인 랭킹 기법(iterative ranking)을 제안하여 입력된 비디오에 대해서 단일한 해를 구하는 것이 아닌 요약 수준에 따른 다양한 해들의 집합을 구하였으며, 단 한 번의 비디오 분석과정을 통해 다양한 길이의 요약된 비디오를 생성할 수 있다^[12]. Fu 등은 다시점에서 취득한 비디오에 대한 요약 기법을 제안하였다^[13]. 먼저 각 시점에서의 비디오의 샷 간의 관계를 통해 그래프 모델로 정의하고 가우시안 엔트로피 융합 모델(Gaussian entropy fusion model)을 통해 샷의 중요도를 분석한다. 다양한 시점의 샷에 대하여 최대의 중요도와 샷 수의 최소화를 만족하는 해를 찾기 위해 정수 계획법(integer programming)을 도입하여 해결하였다. Tseng 등은 Palm OS PDA에서 비디오 요약 시스템에서 구축하였으나, 실질적으로 비디오 요약을 수행하는 미들웨어(middleware)는 PC 상에 존재하며 모바일 기기와의 통신으로 결과를 생성하고 전송한다^[14].

본 논문에서는 원본 비디오에서 추출한 시공간적 중요도를 기반으로 비디오 요약기법을 제안한다. 제안하는 알고리즘은 스마트폰에서의 구동을 전제로 설계되었다. 전체 프레임에서 중요한 프레임에 대한 척도를 제시하고 최대한 중요한 프레임을 보존하면서 사용자의 요구조건에 맞춰 다

양한 재생 길이의 비디오 요약을 수행한다. 이러한 비디오 요약 기법은 마치 비디오의 하이라이트를 자동으로 생성해주는 것과 유사하다. 또한 같은 주제의 여러 비디오를 통합할 때 핵심적인 부분만 추출하여 비교적 짧은 재생길이의 비디오로 재구성이 가능하다. 본 논문에서는 실제 상용 스마트폰을 플랫폼으로 하여 비디오 요약 시스템을 구현하고 실험하였으며 알고리즘의 복잡도와 비디오 요약 성능 측면에서 검증되었다. 그림 1에 제안하는 시스템의 전반적인 개요도를 도시하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 비디오 프레임의 중요도를 평가하는 방법에 대하여 서술한다. 제3장에서는 비디오 요약을 위한 계층적 트리의 생성 및 재구성 기법에 대하여 설명한다. 제4장에서는 스마트폰에서의 구현 및 실험 결과를 제시하고 분석하며, 제5장에서 결론 및 향후 연구에 대해 서술한다.

II. 시공간적 중요도 추정

일반적으로 비디오 요약을 위해서는 어떤 프레임을 보존하고 어떤 프레임을 제거할 지를 명확히 판단해야 한다. 본 장에서는 프레임의 본질적인 중요도를 측정하는 공간적 중요도와 인접하는 프레임과의 상관관계에 따라 정해지는 시간적 중요도의 추정 기법을 제안한다.

프레임은 윤곽선 및 색상, 밝기, 히스토그램 등과 같은 저차원 정보와 물체, 얼굴, 문자 등의 고차원 정보를 가지고 있다. 일반적으로 비디오에서 의미 있는 객체가 등장하는 프레임은 보존하는 것이 합리적이며, 뉴스 등의 비디오의

그림 2. 공간적 중요도 지도 비교. (a) 원본 영상, (b) 지역적 엔트로피, (c) 중요도 지도

Fig 2. Comparison of the spatial saliency maps (a) original image, (b) local entropy, (c) saliency map

경우 문자를 포함하는 프레임이 보다 중요도가 높다고 판단할 수 있다. 비디오 요약에서는 객체에 대한 의미를 해석하는 것보다 의미 있는 객체의 존재여부 판단에 초점을 둔다.

1. 공간적 중요도 분석

비디오 요약을 수행하기 위하여 비디오를 구성하는 많은 프레임들 중에서 어떤 프레임이 중요한 지 판단해야 한다. 예를 들어 많은 윤곽선과 다양한 색상을 포함하고 있는 프레임과 다소 균일한 화소값을 가지는 프레임과의 비교시 전자의 경우가 정보량이 많으므로 더 중요하다. 이와 같은 단일 프레임의 중요도를 평가하는 연구가 활발하게 진행되고 있다. 본 논문에서는 지역적 엔트로피(local entropy) 또는 Itti 등의 방법^[15]을 이용하여 프레임 고유의 공간적 중요도를 계산한다. 지역적 엔트로피는 지역적 히스토그램을 기반으로 화소 단위의 엔트로피이며, 이를 통하여 프레임의 전체 엔트로피를 구한다. 본 논문에서는 Lab 색상 체계에서 a, b 채널에 대한 지역적 히스토그램을 이용하여 각 화소의 엔트로피를 계산하고 프레임 내의 모든 2차원 좌표에서 정규화된 지역적 엔트로피의 총합으로 영상 전체의 엔트로피를 구한다. Itti 등의 방법^[15]은 주변 영역과의 차이가 뚜렷하게 드러나는 중심 영역이 시각적으로 보다 민감하다는 인간의 시각적 특성을 모델링하여 밝기, 색상, 방향 성분에 대한 통합된 프레임의 중요도 지도를 구한다. 이와 같이 구해진 중요도 지도 또는 영상 전체의 엔트로피를 통해서 한 영상의 공간적 중요도 W_s 를 계산한다. 본 논문에서는 하드웨어의 성능에 맞춰 엔트로피를 이용한 방법과

중요도 지도를 이용한 방법을 선택할 수 있게 하였다. 그림 2에서는 두 가지 방법으로 생성한 중요도 지도를 도시하였다.

일반적으로 비디오에서는 얼굴을 포함한 프레임이 그렇지 않은 프레임에 비해 상대적으로 높은 중요도를 지닌다고 말할 수 있다. 따라서 본 논문에서는 얼굴의 개수와 차지하는 면적을 구하여 얼굴에 대한 중요도 W_f 를 계산한다. 이 때 얼굴 검출 알고리즘으로는 Viola 등의 방법^[16]을 채택하여 이용한다.

2. 시간적 중요도 분석

비디오에서는 일반적인 영상의 성질과 달리 짧은 시간에 많은 수의 영상을 취득하기 때문에 인접한 프레임은 거의 유사한 장면들로 구성되어 있다. 따라서 시간의 흐름에 따라 점진적으로 변화하게 되고 객체들의 움직임 또한 여러 장의 프레임을 통해 관찰할 수 있다. 이러한 시간적 중요도는 객체의 움직임이나 장면간의 급격한 변화 및 점진적 변화에 의해 정의된다. 따라서 객체의 움직임이 활발하거나 유사한 장면이 오랫동안 지속되면 중요도가 높다고 판단한다. 이는 비디오를 생성할 때 상대적으로 중요하다고 생각하는 부분의 취득 시간이 길기 때문이다. 따라서 연속된 영상 사이에서의 장면 전환을 검출하기 위해서 유사성을 측정할 필요가 있다. 본 논문에서는 색상 히스토그램의 차분을 기준으로 인접한 프레임간의 유사성을 측정하였다. 이미 기술한 바와 같이 색상 히스토그램의 차분은 연산량이 상대적으로 적고 비교적 정확하게 급격한 장면 전환을 검출할 수 있다. 또한 카메라의 떨림으로 인해 발생하는 오차

그림 3. 비디오 프레임간의 히스토그램 차분 처리 과정 (a) 초기 차분값, (b) 평활화된 차분값, (c) 검출된 국부 최대값

Fig 3. Processing histogram difference of video frames (a) raw difference, (b) smoothed difference, (c) detected local maxima

는 히스토그램에서 크게 나타나지 않으므로 추가적으로 카메라 보정이 필요하지 않아 효율적이다. 하지만 실제 프레임간의 차분 값은 카메라 및 물체의 움직임으로 인해 균일하지 않아 유사한 장면임에도 불구하고 다소 차이가 발생하게 된다. 그림 3(a)에서는 차분 히스토그램에 대한 그래프를 도시하였다. 기존의 연구에서는 장면 전환을 판별하는 문턱치를 적용하였지만 정확한 문턱치에 대한 정의가 어렵다. 따라서 본 논문에서는 경계선을 보존하는 평활화(smoothing) 기법^[17]을 적용하여 차분 히스토그램의 장면 전환에 대한 신호를 보다 명확하게 표현한다. 그림 3(b)는 Xu 등의 방법^[17]을 이용하여 평활화된 신호를 도시한다. 마지막으로 앞서 구한 평활화된 신호를 보다 명확하게 만들기 위해서 지역적으로 최대값만을 보존한다. 그림 3(c)는 이와 같이 검출된 최대값을 의미하며 이는 장면 전환 검출을 위해서 이용된다.

시간적 중요도 추정에 쓰이는 객체간의 움직임 추정하기 위해서 본 논문에서는 먼저 영상을 매크로 블록으로 나눈다. 그리고 참조 프레임과 추정 프레임 사이에서의 움직임(motion) 벡터를 구한 후 벡터크기의 총합으로 움직임에 대한 중요도 W_m 을 정의한다. 최종적으로 비디오의 한 프레임에 부여되는 중요도는 앞서 정의한 공간적 중요도와 시간적 중요도를 통합하여 정의한다. 따라서 지역적 엔트로피의 양이 많고 검출된 얼굴의 수와 면적이 넓을수록 공간적 중요도는 높아지고 움직임이 많을수록 시간적 중요도가 높아진다. 본 논문에서는 식 (1)과 같이 통합된 시공간

적 중요도 W 를 계산한다.

$$W = \alpha W_s + \beta W_f + \gamma W_m \quad (1)$$

III. 비디오 요약

본 논문에서는 본래의 내용을 최대한 상실하지 않는 범위 내에서 원하는 길이의 비디오를 생성하는 비디오 요약 기법을 제안한다. 입력 비디오를 다양한 재생 길이를 가지는 비디오로 재구성하기 위해서 매번 분석과정이 반복되는 것은 많은 계산량이 필요하다. 이는 비디오 분석을 기반으로 생성한 계층적 트리를 도입하여 해결한다. 계층적 트리는 과분할 구조를 가지며 요구된 재생 길이를 만족하는 노드들로 조합하기 위해서 하위 노드부터 병합을 수행하여 한다. 또한 사용자로부터 입력 받은 노드의 최소 재생 길이를 통해 장면 전환간의 간격이 매우 짧은 노드의 선택을 방지하고 다수의 노드에서 균형적으로 프레임을 추출함으로써 시각적 결함을 제거할 수 있다.

본 논문에서는 검출된 장면 전환을 기준으로 생성한 초기의 과분할 노드들을 반복적으로 병합하여 최적의 노드집합을 구하는 기법을 제안한다. 따라서 비디오 분석을 통하여 얻은 프레임의 중요도 및 장면 전환 검출을 통해 계층적 트리를 생성하는 것이 선행되어야 한다. 계층적 트리의 각 노드는 비디오에서 특정 구간을 의미하며 프레임의 중요도에 의해 노드의 중요도가 결정된다. 그림 4와 같이 계층적 트리의 최상위 노드는 비디오의 전체 프레임을 포함하는 구간을 가진다. 또한 계층적 트리의 다른 특징은 한 노드의 구간은 모든 자식 노드들의 구간의 합집합이고 형제 노드 간에는 중첩되는 프레임 존재하지 않는다.

1. 계층적 트리의 생성

계층적 트리를 생성하기 위해서는 먼저 장면 전환 리스트를 구한다. 장면전환 리스트는 비디오 분석과정에서 장면 전환으로 검출된 프레임들의 번호와 순위, 인접한 프레임과의 유사도로 구성한다. 순위는 장면 전환을 검출된 전

그림 4. 비디오 분석 결과 생성되는 계층적 트리의 사례
Fig 4. An example of hierarchical tree based on the result of video analysis

프레임들 사이에서 유사도를 기준으로 오름차순으로 정한다. 한편 계층적 트리의 노드는 하나의 장면 전환 리스트가 할당된다. 계층적 트리의 노드는 부여 받은 장면 전환 리스트에서 유사도가 가장 낮은 값을 가지는 프레임을 기준으로 분할되어 자식 노드를 생성한다. 만약 동일한 유사도를 가지는 프레임들이 존재한다면 모두 분할의 대상이다. 그리고 장면 전환 리스트도 분할하여 자식 노드에 다시 할당한다. 이와 같은 분할과정은 최상위 노드로부터 선행되어 적용되며 장면 전환 리스트가 존재하지 않는 노드가 발생할 때까지 반복적으로 수행한다. 계층적 트리의 노드는 서로 다른 깊이를 가지며 장면 전환 리스트에서 분할 기준 프레임의 순위를 통해 깊이가 정해진다. 예를 들면, 깊이가 0인 최상위 노드가 분할되면 자식 노드들의 깊이는 1이다.

2. 계층적 트리의 단계적 병합

계층적 트리의 특성은 한 노드의 구간은 부모 노드의 구간에 포함된다. 따라서 사용자의 요구에 부합하는 비디오 요약을 위해서 노드를 선택해야 한다. 본 논문에서는 통합 에너지와 유지 에너지를 통해 단계적으로 병합하는 기법을 제안한다. 병합은 한 노드의 모든 자식 노드들이 통합되는 것을 의미하지만 모든 자식 노드가 제거되는 것과 동일하며 자식 노드가 최하위 노드만으로 구성되는 노드에서만 수행된다. 그리고 조건에 만족하는 노드의 통합 에너지가 유지 에너지보다 높을 경우 병합이 수행되고 해당 노드의 모든 자식 노드가 제거되어 최하위 노드로 지정된다. 따라서 통합 에너지와 유지 에너지를 정교하게 정의해야 한다. 통합 에너지는 자식 노드들 간의 구간 길이 비율이 균일하고 중요도 곡선이 복잡할수록 높다. 식 (2) 및 식 (3)과 같이 통합 에너지 E_m 을 계산한다.

$$\alpha = \frac{\sum_{i=1}^n (1 - nC_l(i))^2}{n(n-1)} \quad (0 < \alpha < 1) \quad (2)$$

$$E_m = (1 + \alpha)(N_l^2 + \beta N_\sigma) \quad (3)$$

여기서 n 은 자식 노드의 수, C_l 은 자식노드의 구간길이의 비율, N_l 은 현재 노드의 구간길이, N_σ 는 중요도 곡선에서의 해당 구간의 분산을 의미한다. α 는 자식 노드의 구간길이의 비율에 따라 정해지는 가중치이다.

유지 에너지는 인접한 구간을 가지는 형제 노드간의 유사도가 낮고 자식 노드와의 깊이 차이가 클수록 높다. 식 (4)와 같이 유지 에너지 E_s 를 계산한다.

$$E_s = D_L D_H \sum_{i=1}^n (C_l^2 + \beta C_\sigma(i)) \quad (4)$$

여기서, D_L 은 자식 노드와의 깊이에 대한 차이 값이고 병합 수행 시 자식 노드의 깊이로 현재 노드의 깊이를 갱신한다. 이는 장면 전환 리스트에서 유사도 기준으로 정의한 순위 차이가 큰 노드들의 병합을 방지한다. D_H 는 자식노드를 분리하는 히스토그램의 차이 값을 의미한다.

3. 프레임 추출

본 논문에서는 탐욕적 접근 방법을 적용하여 계층적 트리로부터 순차적으로 프레임을 추출하여 최종적으로 결과물을 생성한다. 먼저 추출되는 프레임의 연속성을 보장하기 위해서 중요도 곡선을 개선한다. 동일한 장면으로 판정된 구간 내에서의 중요도 곡선을 볼록한 곡선(convex curve)으로 변경한다. 그리고 최종 프레임을 추출하기 위하여 계층적 트리의 최하위 노드 중 사용자가 요구한 최소 구간 길이를 충족하는 노드들을 선택한다. 그리고 선택된 노드들을 순차적으로 방문하여 한 프레임씩 추출하며, 각 노드는 지역적 최대값을 가지는 프레임으로부터 탐욕적 접근을 통해 높은 중요도를 가지는 방향으로 확장한다. 만약 노드의 구간경계를 초과하는 확장이 요구되면 확장 방향을 전환한다. 이는 다수의 세그먼트에서 균등하게 추출되는 것을 보장한다. 그러나 선택된 노드들의 수가 많다면 균등하지만 노드마다 매우 짧은 구간을 추출할 수 있다. 식 (5)와 같이 비디오 요약을 위해 사용할 수 있는 최대 노드의 수 S_{max} 를 계산한다.

$$S_{\max} = \text{round}\left(\frac{R_l}{R_s}\right) \quad (5)$$

여기서 R_l 는 사용자가 요구하는 전체 재생 길이이고 R_s 는 요구된 최소 구간 길이이다. 최소 구간 길이를 만족하는 최하위 노드를 중요도 기준으로 우선순위를 부여하여 $S_{\max}$ 를 초과하지 않는 수만큼 상위 노드를 선택한다. 그림 5는 비디오 요약에 사용되는 노드를 선택하는 일련의 과정을 도시하였다.

그림 5. 노드 선택 과정
Fig 5. Step on selecting nodes

IV. 실험 결과

본 논문에서는 10개의 다양한 비디오에 대하여 실험하였

으며, 표 1에서는 실험에 사용된 비디오의 특징을 제시하였다. 본 논문에서의 알고리즘의 구현 및 실험에 사용된 스마트폰은 ARM Cortex A9 1.2GHz CPU를 장착한 삼성전자의 SHW-M250S(갤럭시 SII)이며 FFmpeg 라이브러리를 이용하여 비디오 디코딩을 수행하였다. FFmpeg은 다양한 코덱에 대해서 디코더(decoder) 및 인코더(encoder)를 제공하고 있으며 NEON엔진에 대한 최적화 옵션을 두어 FFmpeg 내의 멀티미디어와 신호처리 알고리즘의 성능을 향상시킬 수 있다. 이러한 FFmpeg 라이브러리는 크로스 컴파일러(cross compiler)를 통해 컴퓨터뿐만 아니라 스마트폰의 많은 멀티미디어 플레이어에서 널리 사용되고 있다.

한편, 스마트폰에서의 구현은 제한된 CPU 성능과 메모리를 이용하기 때문에 수많은 프레임에서 비디오 분석과 같은 계산량이 많은 알고리즘의 경우 최적화가 필수적이다. 본 논문에서는 직접 계산하는 연산을 최소화하기 위하여 참조 테이블의 사용을 최대화하였다. 특히 지역적 엔트로피 계산 수행시 반복적으로 호출되는 로그함수를 제거하여 미리 계산하여 참조테이블에 저장한다. 이러한 연산은 수행되는 프레임의 해상도와 수에 따라 계산량이 결정되는데 비디오에서는 수많은 영상에 대해서 처리하기 때문에 이점이 있다. 또한 본 논문의 구현은 다양한 LOD (level of detail) 설정에 대한 성능 분석을 수행하여 영상 처리에 사용되는 해상도, 지역적 윈도우의 크기, 분석에 사용되는 프레임의 샘플링 주기 등을 조절한다. 이는 인접한 프레임이 비교적 짧은 시간 간격으로 취득이 되어 유사하고 결과 품질에 영향을 주지 않는 범위 내에서 영상 처리의 해상도를 낮추면 연산량을 상당히 줄일 수 있기 때문이다. 본 논문에서는 비교적 많은 연산량을 요구하는 얼굴 검출 주기와 영

표 1. 실험에 사용된 입력 비디오의 특징
Table 1. Characteristics of the input videos for the experiment

번호	1	2	3	4	5	6	7	8	9	10
프레임 수	3,000	4,224	2,544	5,640	1,320	1,680	3,567	4,584	870	960
프레임 넓이	640	640	640	640	640	640	640	320	640	480
프레임 높이	360	360	360	360	524	360	360	240	360	272
초당 프레임 수 (FPS)	24	24	24	24	24	24	29	24	29	24

표 2. 스마트폰(SHW-M250S)에서의 비디오 분석 수행시간 (단위: 초)

Table 2. Processing time to analyze videos on a smartphone (sec)

번호	1	2	3	4	5	6	7	8	9	10
유사도	1.74	3.04	1.96	4.34	1.31	1.24	2.78	3.99	478	566
엔트로피	2.28	3.13	1.96	4.20	1.42	1.26	2.16	4.55	557	728
얼굴검출	5.28	7.57	4.53	10.43	3.67	2.90	5.36	10.66	1.28	1.86
합계	9.30	13.74	8.44	18.97	6.39	5.40	10.30	19.20	2.32	3.15

상 처리의 해상도와 지역적 엔트로피를 구할 때의 윈도우 크기를 실험적으로 정하였다. 표 2는 스마트폰에서의 각 실험 비디오의 분석에 소요된 수행시간이다.

비디오 요약 을 위해서 사용자의 요구사항(요약 수준, 최

소 구간 길이, 장면 전환 효과, 얼굴 가중치의 유무, 움직임 가중치의 유무)을 입력 받아 처리한다. 그림 6에서는 다양한 요약 수준에서의 비디오 요약 결과를 도시하였다. 제시된 영상은 요약된 비디오의 키프레임을 의미한다. 실험 결

그림 6. 다양한 요약 비율에 대한 제안하는 알고리즘의 비디오 요약의 결과

Fig 6. Result of video summarization of the proposed algorithm for various reduction rate

그림 7. 상용 비디오 편집 도구(Muvee)와의 요약 성능 비교 (50% 축약) (a) 테스트 비디오 6의 요약 결과, (b) 테스트 비디오 10의 요약 결과
Fig 7. Comparison of video summarization (50% reduction) performance with a commercial video editing tool (Muvee) (a) Summarization result of 6th test video, (b) Summarization result of 10th test video

과, 요약 수준이 높아질수록 많은 사건을 포함하고 적은 장면 전환을 가지는 비디오일수록 한 사건에서 많은 프레임을 추출한다.

제안하는 알고리즘의 성능 비교를 위해 모바일 상용 비디오 편집 도구인 Muvee와의 성능 비교를 수행하였으며, 실험 비디오 6번과 10번에 대해 50% 요약을 수행한 결과를 그림 7의 좌우에 각각 제시하였다. 그림 7의 좌측에 제시한 비교 결과에 나타난 바와 같이 Muvee의 비디오 요약은 전체 비디오에 대해 균형있게 생성되지 못하였다. 즉, 요약된 비디오는 전체 비디오의 약 50%가 경과한 지점에서 시작되는데, 입력 비디오의 전반부에도 중요한 장면들이 등장하지만 요약된 비디오에는 이들이 전혀 포함되지 않았다. 반면 제안하는 기법은 입력 비디오의 시작부의 두 샷의 요약이 포함되어 있어 전체적으로 균형있는 요약이 되었다고 볼 수 있다. 또한, 그림 7의 우측에 제시한 요약 결과에 나타난 바와 같이 Muvee의 비디오 요약은 매우 짧은 부분을 다수 포함하여 사용자의 시각적 불편함을 초래하였다. 반면 제안하는 기법은 이러한 부분은 요약된 비디오에 포함하지 않고, 어느 정도 길이가 유지되는 비디오 샷만을 요약된 비디오에 포함하여 균형 잡힌 결과를 생성하였다. 이와 같이 제안하는 알고리즘은 기존의 상용 비디오 편집 도구가 제공하는 비디오 요약에 비해 우수한 성능을 보인다고 할 수 있다.

V. 결 론

본 논문에서는 스마트폰의 플랫폼으로 하여 각 프레임에서의 시공간적으로 중요도를 분석하여 비디오 요약을 수행하는 기법을 제안하였다. 다양한 환경에서 취득한 비디오를 입력으로 하여 단 한 번의 분석을 통해 반복적이고 불필요한 구간에 대해서 정의하고 이를 바탕으로 사용자로부터 요구된 재생 길이를 가지는 비디오를 반복적으로 생성할 수 있다. 이러한 요약 기술을 통해 사용자들은 수많은 비디오에 대해 요약된 정보들을 손쉽게 얻을 수 있고 한정적인 저장 공간과 비디오 공유를 위한 네트워크를 통한 전송에 있어서 많은 이점을 지닌다. 또한 비디오 분석 제약이 많은

저장 공간에서 많은 이득을 볼 수 있다. 본 논문에서는 스마트폰을 플랫폼으로 하여 제안하는 비디오 요약 기법을 구현하였다. 스마트폰은 비디오 관리, 취득, 공유가 가능하고 하드웨어의 발전에 힘입어 다양한 멀티미디어 처리가 가능하므로 비디오 요약에 적합한 기기이다.

본 논문에서는 검출된 장면 전환에 의해 나뉜 각 장면마다 한 장면에 대해서 시공간적 중요도를 계산한다. 이와 같은 특성을 이용하여 다수의 비디오를 하나의 비디오로 저작할 수 있고, 비디오의 세그먼트에 대한 태깅(tagging)과 접목하면 많은 비디오로부터 유사한 세그먼트만 취합할 수 있을 것으로 예상된다. 또한 계층적 트리로부터 노드를 선택함에 있어 추정된 중요도 곡선은 최종적인 결과의 품질에 영향을 끼친다. 따라서 중요도 곡선의 정확도를 향상시키기 위한 향후 연구가 필요하다.

참 고 문 헌

- [1] E. P. Bennett and L. McMillan, "Computational time-lapse video," ACM Trans. on Graphics, vol. 26, no. 3, Article No. 102, July 2007.
- [2] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," Proc. of 6th ACM Conference on Multimedia, September 1998.
- [3] 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, October 1998.
- [4] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," IEEE Trans. on Circuits and Systems for Video Technology, vol. 9, no. 8, pp. 1280-1289, December 1999.
- [5] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," IEEE Trans. on Circuits and Systems for Video Technology, vol. 9, no. 8, pp. 1269-1279, December 1999.
- [6] H. W. Kang, X. Q. Chen, Y. Matsushita, and X. Tang, "Space-time video montage," Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1331-1338, June 2006.
- [7] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1971-1984, November 2008.
- [8] T. Mei, B. Yang, S. Q. Yang, and X. S. Hua, "Video collage: presenting a video sequence using a single image," The Visual Computer, vol. 25, no. 1, pp. 39-51, December 2008.
- [9] 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, pp. 296-305, February

- 2005.
- [10] Y. Gong and X. Liu, "Video summarization using singular value decomposition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 174-180, June 2000.
- [11] B. Yu, W. Y. Ma, K. Nahrstedt, and H. J. Zhang, "Video summarization based on user log enhanced link analysis," Proc. of 11th ACM Conference on Multimedia, pp. 382-391, November 2003.
- [12] L. Herranz and J. M. Martinez, "A framework for scalable summarization of video," IEEE Trans. on Circuits and Systems for Video Technology, vol. 20, no. 9, pp. 1265-1270, September 2010.
- [13] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z. H. Zhou, "Multi-view video summarization," IEEE Trans. on Multimedia, vol. 12, no. 7, pp. 717-729, November 2010.
- [14] B. L. Tseng, C. Y. Lin, and J. R. Smith, "Video summarization and personalization for pervasive mobile devices," Proc. SPIE : Storage and Retrieval for Media Databases, vol. 4676, pp. 359-784, December 2001.
- [15] L. Itti, C. Koch, and E. Niebur, "A model of saliency based visual attention for rapid scene analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, November 1998.
- [16] P. Viola and M. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, May 2004.
- [17] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via L_0 gradient minimization," ACM Trans. on Graphics, vol. 30, no. 6, Article No. 174, December 2011.
- [18] Muvee, <http://www.muvee.com/en/>

저 자 소 개

이 원 범

- 2010년 2월 : 인하대학교 컴퓨터정보공학부 학사
- 2012년 8월 : 인하대학교 정보통신공학과 석사
- 2012년 8월 ~ 현재 : LG전자 CTO 부문 연구원
- 주관심분야 : 스마트폰에서의 영상처리, 비디오 요약, 영상기반 3차원 형상복원

월 리 엄

- 2011년 7월 : 인도네시아 Bina Nusantara University 컴퓨터과학과 학사
- 2011년 8월 ~ 현재 : 인하대학교 정보통신공학과 석박사통합과정
- 주관심분야 : 3차원 비디오, computational photography, GPGPU

박 인 규

- 1995년 2월 : 서울대학교 제어계측공학과 학사
- 1997년 2월 : 서울대학교 제어계측공학과 석사
- 2001년 8월 : 서울대학교 전기컴퓨터공학부 박사
- 2001년 9월 ~ 2004년 3월 : 삼성종합기술원 멀티미디어랩 전문연구원
- 2007년 1월 ~ 2008년 2월 : Mitsubishi Electric Research Laboratories (MERL) 방문연구원
- 2004년 3월 ~ 현재 : 인하대학교 정보통신공학부 부교수
- 주관심분야 : 컴퓨터 그래픽스 및 비전 (영상기반 3차원 형상 복원, 3차원 카메라, computational photography), GPGPU