JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 10, pp. 85-92, October 2023
https://doi.org/10.9708/jksci.2023.28.10.085

Reproducing Summarized Video Contents based on Camera Framing and Focus

Hyung Lee*, E-Jung Choi**

*Professor, Dept. of Broadcasting Contents, Daejeon Health Institute of Technology, Daejeon, Korea
**Professor, Dept. of Media and Visual Communications, Hannam University, Daejeon, Korea

[Abstract]

In this paper, we propose a method for automatically generating story-based abbreviated summaries from long-form dramas and movies. From the shooting stage, the basic premise was to compose a frame with illusion of depth considering the golden division as well as focus on the object of interest to focus the viewer's attention in terms of content delivery. To consider how to extract the appropriate frames for this purpose, we utilized elemental techniques that have been utilized in previous work on scene and shot detection, as well as work on identifying focus-related blur. After converting the videos shared on YouTube to frame-by-frame, we divided them into a entire frame and three partial regions for feature extraction, and calculated the results of applying Laplacian operator and FFT to each region to choose the FFT with relative consistency and robustness. By comparing the calculated values for the entire frame with the calculated values for the three regions, the target frames were selected based on the condition that relatively sharp regions could be identified. Based on the selected results, the final frames were extracted by combining the results of an offline change point detection method to ensure the continuity of the frames within the shot, and an edit decision list was constructed to produce an abbreviated summary of 62.77% of the footage with F1-Score of 75.9%

▸ **Key words**: Illusion Depth, Golden Section, Medium Shot, Blur Detection, Video Editing

[요 약]

본 논문에서는 장편의 드라마나 영화에서 스토리 기반의 축약된 요약본을 자동으로 제작하기 위한 방법을 제안한다. 촬영 단계에서 황금분할을 고려한 공간감 있는 프레임 구성과 내용 전달 차원에서 시청자들의 시선을 집중시키기 위한 관심 대상에 대한 초점을 기본 전제로 했다. 이에 적정한 프레임들을 추출하기 위한 방법을 고려하기 위해서 기존의 씬(scene) 및 숏(shot) 검출에 대한 연구, 초점과 관련된 블러 정도를 파악하는 연구들에서 활용되었던 요소 기술들을 활용했다. 유튜브에서 공유되는 영상을 프레임 단위로 변환한 후 프레임별로 특징을 추출하기 위한 영역으로 프레임 전체 영역과 3개의 부분 영역으로 구분했고, 해당 영역별로 각각 라플라시안 연산자와 FFT를 적용한 결과들을 비교하여 상대적으로 일관성 있고 강건한 FFT를 선택했다. 프레임 전체에 대한 계산값과 3개 영역의 계산값들을 비교하여 상대적으로 선명한 영역을 확인할 수 있는 조건을 기반으로 대상 프레임을 선별했다. 이렇게 선별된 결과를 토대로 숏 내에서 프레임들의 연속성을 확보하기 위해 오프라인 변화점 탐지기법을 적용한 결과와 접목시켜 최종 프레임들을 추출했고, 이를 기반으로 편집결정리스트를 구성하였으며, F1-스코어 75.9%를 갖는 62.77%로 축약된 요약본을 제작했다.

▸ **주제어**: 깊이 환영, 황금 분할, 미디엄 숏, 블러 탐지, 영상 편집

• First Author: Hyung Lee, Corresponding Author: E-Jung Choi
*Hyung Lee (hyung@hit.ac.kr), Dept. of Broadcasting Contents, Daejeon Health Institute of Technology
**E-Jung Choi (ejchoi@hnu.kr), Dept. of Media and Visual Communications, Hannam University
• Received: 2023. 09. 22, Revised: 2023. 10. 18, Accepted: 2023. 10. 20.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

쏟아지는 데이터를 기반으로 정보를 생산하고 이를 가공하여 공유함으로써 정보의 가치는 어느 정도 공평하게 영위될 수 있겠다. 이러한 배경에는 첨단 기술을 기반으로 텍스트 위주의 정보 전달 보다는 시청각 기반의 정보 전달의 파급력은 이용 대상을 고려해 본다면 상대적으로 쉽게 수용될 수 있기 때문이다. 아울러 소수의 관심사라도 상호간의 다양성을 인정하여 이를 공유하고 참여할 수 있는 시대적인 흐름에 기인한 것이라고 볼 수 있겠다. 이러한 과정에서 정보 생산자들 역시 특정 전문인이 아닌 일반들도 손쉽게 생산할 수 있는, 특히 다양한 미디어를 전문 제작자가 아닌 일반인 누구나 제작 및 가공하여 공유할 수 있게 되었다. 이러한 측면에서 정보 공유의 매체인 영상물도 누구나 쉽게 제작하여 유튜브와 같은 플랫폼에서 공유함으로써 영상물 제작은 더 이상 전문가의 영역이 아니게 되었고 경제적 소득과 이어지는 생산적인 활동이 되었다.

어떤 주제의 또는 어떤 장르의 영상물을 제작하기 위해서는 기본적으로 고려해야 할 두 가지 기본 원칙이 있는데, '무엇을 볼 것인가?'와 '어떻게 볼 것인가?'이다. 첫 번째는 장면을 구성하는 방법과 관련이 있으며, 두 번째는 관객의 시점과 관련이 있다. 결과적으로 무엇을 어떻게 보여줄 것이냐의 문제인데, 이야기 전달 측면에서 관객의 시선을 제작자의 의도에 맞게 어떻게 이끌 것이냐의 문제이기도 하다[1].

우리는 3차원 공간에 살고 있지만 이를 반영한 영상물은 2차원 평면으로 표현되기에 시청자가 생각하는 것과 실제 보이는 영상 사이에서 차이점이 생긴다. 영상은 폭, 높이, 깊이 등 3가지 차원으로 구성된 하나의 입체적인 상자로 볼 수 있는데, 프레임을 구성하는 폭과 넓이는 사실적이지만 영상의 깊이는 환영으로만 인식될 수 있다[2]. 그렇기에 제작자의 관점에서는 시청자에게 무엇을 보게 할 것인가라는 측면에서는 영상의 깊이를 충분히 고려하여 이들의 시선을 유도할 수 있어야 한다. 이를 위해서 제작 단계에서는 촬영 현장의 설정을 고려해서 장면을 근경, 중경, 원경으로 구성하던가 렌즈의 심도를 조정하여 배경에서 관심 대상을 분리할 수 있겠다. 제작 단계에서 이를 고려하지 못했다면 후반제작에서 영상의 깊이를 영상처리 효과 등을 적용하여 이를 구현할 수 있다. 즉, 장면을 입체적인 프레임으로 구성하기 위해 2차원 평면은 프레임의 넓이와 높이로 구성되고 관심의 대상인 피사체를 그 평면에 위치시키는 것이지만 깊이 관점에서는 원근을 고려해 위치시킴으로써 영상의 깊이를 표현할 수 있다[2]. 전문적인 영상 제작자들은 스토리 전달 측면에서 시청자의 시선을 이끌기 위해 입체감 있는 장면 구성을 고려하여 영상물을 제작한다.

영상 제작은 일반적으로 기획, 제작, 후반작업 등 3단계를 거치게 되는데, 입체감 있는 장면 구성은 모든 단계에서 고려될 수 있겠지만, 특히 두 번째 단계인 제작의 촬영 과정에서 고려해야 될 구도와 관련성이 높다. 경우에 따라서는 후반작업의 편집 과정에서 인위적으로, 특히 VFX의 경우에는 합성의 마무리 단계에서 렌즈 블러 등의 효과를 적용하여 장면 내 원근감을 구현할 수 있다. 촬영의 가장 작은 단위인 숏은 영상제작의 구성에서 동기, 정보, 구성, 사운드, 앵글, 연속성 등 6가지의 중요한 요소들을 갖추어야 하며, 숏의 크기에 따른 용도 측면에서는 대략 9가지 정도로 구분할 수 있다[1-3]. 이들 중 미디엄 숏은 관심 대상의 가상 시선 및 이의 변화를 확실히 확인할 수 있고 배경보다 더 주의를 끌고 피사체의 시선으로 시청자의 시선을 이끌게 한다. 아울러 근경, 중경, 원경에 관심 대상들을 적정하게 배치하여 입체감을 구성함과 동시에 화면 내 관심 대상의 위치와 크기를 고려하여 촬영 시 초점과 심도를 조정하게 된다. 즉, 시청자의 시선을 관심 대상으로 이끌기 위해 아웃 포커싱 등의 방식으로 배경을 분리시킬 수 있고 입체감을 위한 촬영 환경을 구성할 수 있다[1,2].

이렇게 촬영된 숏들을 편집하여 최종 영상물을 제작하는데, 영상 편집의 주요 원칙 중의 하나는 스토리의 연속성을 고려하여 숏들을 선택하고 배열함에 있어서 연속성, 복합성, 내용 등을 충분히 고려해야 한다. 이러한 원칙은 절대적이라기 보다는 관례적이긴 하지만 편집의 연속성 확보는 명확한 스토리 전달 측면에서 가장 기본이라고 할 수 있으며 이를 확보하기 위해서는 등장인물의 동일성 및 배치, 가상선을 구성하는 벡터, 피사체의 움직임, 색상, 음향 등 다양한 요소들이 고려된다[1-3]. 아울러 편집의 유형 중 연속편집은 주요 사건의 설명에 목적을 둔 편집방식으로 많은 정보가 의도적으로 누락 되더라도 스토리가 유연하게 진행되도록 어떤 사건의 세부 사항에서 다음으로 자연스럽게 전환이 되게 하는, 앞뒤로 이어지는 화면 내에서 시간, 장소, 내용적인 연속성이 중요 시 되는 편집방법이다. 즉, 관람하고 시청하는 대부분의 영상물들은 이러한 연속편집이 충분히 반영되었다고 볼 수 있겠다[1-3].

이러한 영상문법에 의거하여 제작된 영상물들인 장편의 드라마와 영화 등을 유튜브 등에서 '몰아보기' 형태의 요약 편집본으로 쉽게 접할 수 있으며, 영향력 있는 채널의 경우에는 상당한 구독자를 보유하고 있다. 이러한 요약본은 재편집본으로써 비록 시청 시간을 단축시켰지만 원본

의 전달 내용을 충분히 전달할 수 있어야 한다. 결국, 원본의 내용을 충분히 반영한 압축본을 제작하기 위해서 편집자들이 의미있는 숏들을 추출하여 재구성 또는 재배열해야 한다. 분명 서사적 내용의 숏들과 서정적 내용의 숏들, 그리고 시청자의 지루함을 달래기 위해 내용의 연속성을 기반으로한 숏 크기 변화 등도 고려할 수 있겠지만, 관심대상에 집중된 장면만을 기반으로한 숏들을 재배열함으로써 내용만을 전달하기에 적절할 수 있을 것이다. 그래서 촬영 시 카메라의 초점이 시청자의 관심을 유도하기 위함이라는 전제하에 미디엄 숏 정도의 크기만으로도 충분히 스토리의 뼈대를 전달할 수 있는 요약 편집본을 제작할 수 있을 것으로 예상된다.

이를 위해 본 논문에서는 방송국에서 방영한 드라마들 중에서 해당 방송사가 직접 운영하는 유튜브 채널에서 스트리밍하는 '몰아보기' 요약본을 대상으로 선정하여 영상제작에 대한 경험기반으로 미디엄 숏 크기 정도의 숏들을 추출하여 요약본 영상물을 제작 방법을 제안한다. 본 논문의 구성으로 2장에서 영상물을 기반으로 구성별 분리 및 추출을 위한 기존의 연구 분야 및 방법 등을 살펴보고, 3장에서는 프레임 구성(구도)을 고려하여 관심 대상에 초점을 맞춘 숏들과 미디엄 숏 크기 정도의 숏들을 추출하여 편집결정 리스트를 구성하는 방법을 제안한다. 마지막 4장에서는 연구 방법 및 활용의 한계와 개선 방안 및 추가적인 연구 방향을 기술한다.

II. Preliminaries

1. Related works in Producing Environment

1.1 Illumination of Depth

공간감이 부여된 장면을 구성하기 위해서는 2차원 장면 내에서 3차원의 깊이 환영을 인식하기 위한 몇 가지 방법들이 촬영 및 후반작업에서 고려 되었다. 사람은 일반적으로 시선이 고정되면 자동으로 초점을 맞추고 상대적인 거리를 인지하여 피사체의 크기로 원근을 파악할 수 있는 데 이는 평균적인 크기를 가늠하여 상대적인 크기로 거리를 파악하는 것이다. 아울러 소실점과 연결된 선들과 물체들 간의 상대적인 거리를 파악한다. 2차원에서 가장 강력하게 깊이를 느낄 수 있는 방법은 한 물체가 다른 물체를 가림으로서 이는 깊이와 운동 지각의 단서가 된다. 또한 빛으로 인해 물체에 만들어진 음영으로 3차원적인 입체감을 전달할 수 있다. 그래서 촬영 시 입체감을 위해 조명이 필요기도 하지만 장면 내의 밝기 차이를 둠으로써 시청자들의 시선을 이끌 수 있다.

자연현상에서는 산란현상으로 인해 원경은 채도가 낮고 흐릿해 보이며, 전경의 물체는 채도가 높고 선명하기에 이를 기반으로 전경에 시청자들의 시선이 머물게 할 수 있지만 이는 일상생활에서 접할 수 있는 것이기 때문에 굳이 의도적인 만들어진 장면이 아닌 이상 자연스럽게 수용될 수 있다. 파사체의 위치(근경, 중경, 원경) 및 피사체의 방향과 속도 등의 움직임을 파악하여 공간에 대한 정보를 구성할 수 있기에 이를 기반으로 장면 내 공간감을 구성할 수 있도록 피사체들을 배치하고 이동시킬 수 있다.

추가적으로 입체영상제작에서는 초점 외에 영점이 추가되기 때문에 이를 기반으로한 깊이 콘티를 별도로 고려하여 시청자들의 시선을 이끌 수 있다.

1.2 Framing for Medium Shot Family

공간감을 구현할 수 있는 촬영 환경을 구성하기 위한 방안들 외에도 주피사체의 프레임 내 위치 역시 중요한 요소가 된다. 이는 황금비율에 기인한 것이지만 이를 매 촬영 시 제대로 적용하기에는 여건상 어려운 점이 있다. 하지만 대부분의 촬영 감독들은 이를 기반으로 화면을 3분할하여 관심대상을 적정한 위치에 배치시켜고 시청자들의 시선을 유도하기 위해 초점을 맞춘다[4].

상황에 따라 1숏, 2숏, 그리고 간헐적으로 3숏, 특히 2숏에서 오버 숄더 숏의 경우에는 주피사체 이외는 모두 아웃 포커싱 기법을 적용하여 입체감과 다른 피사체들 사이의 거리감을 부여하여 시청자들의 시선을 묶어 놓는다. 내용 전달 차원에서 이러한 숏들은 대부분 미디움 숏 정도의 크기를 활용하며, 그 외에는 숏들은 연기자들의 감정 표현이나 전반적인 상황을 설명하기 위한 상황에 더 적절하게 활용되는 것이 제작 환경에서 일반적이다[2].

2. Related works in Research Area

2.1 Scene Detections

제작자 관점에서 영상은 스토리의 내용과 이와 관련된 시간 및 장소 등을 기반으로 프로그램(program), 시퀀스(sequence), 씬(scene), 숏(shot) 등의 계층으로 구성되는데, 예를 들어, 프로그램을 완성된 책이라고 본다면 숏은 문장, 씬은 문단, 시퀀스는 장에 대응된다고 볼 수 있겠다. 여기서 씬은 동일 장소와 시간대에 관련된 숏들의 모음이라고 할 수 있고, 하나의 숏은 연속적은 프레임들의 모음이라고 볼 수 있으며, 이러한 비디오의 구조적인 계층도에서는 그림 1과 같이 프레임(frame)이 추가된다. 이와 관련하여 일반적으로 비디오는 초당 25 혹은 30프레임, 영화는 24프레임으로 구성된다.

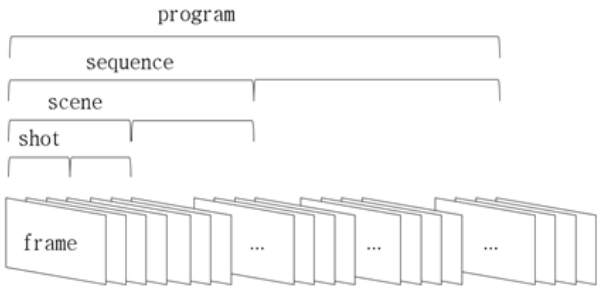

Fig. 1. Video Hierarchy

대부분의 씬 탐지법들은 프레임을 기반하고 있는데, 활용도로는 비디오를 개별 씬으로 분리하던가, 상업광고를 분리 혹은 감시 비디오를 처리 및 분리, 통계적 분석으로 반복되는 씬 검색, 평균적인 숏의 길이에 대한 연구 등 효율적인 비디오 관리 및 검색을 위해 연구되었다[5].

제작자 관점에서 씬은 동일 장소에서 동일 시간 내에 일어난 일련의 숏들을 배열한 것이기에 촬영 환경이 제대로 잘 구성되었다면 숏들은 어느 정도 일관성 있는 색공간 및 밝기를 갖게 되고 이야기 전개에 따라 씬들 사이에 이유 있는 장면전환 효과들이 적용될 가능성이 높다. 이러한 정보를 기반으로한 씬 탐지법들은 오래전부터 후반작업에 활용되는 선형 혹은 비선형 편집기에서는 제공된 기능들이다.

인접 프레임들을 대상으로 동일한 색공간 포함 유무, 경계선 검출 결과를 상호 비교, 밝기를 기반으로 설정된 임계값을 비교하는 등의 방법들을 활용함에 있어 성능 및 정확도는 활용 대상에 따라 달라질 것으로 파악된다[5]. 그리고 시각적 특징으로 색상 히스토그램, 코너 에지, 객체 색상 히스토그램을 기반으로 상향식 계층 클러스터링 방법으로 유사도를 평가하여 연관된 숏들을 군집화하는 방법[6]에서는 숏 경계 식별 성능의 정확도는 평균 93.3%, 비디오 씬 탐지 성능의 정확도는 평균 83.3%의 성능을 보였다. 그 외에 지능망을 활용하여 촬영 환경의 다양성과 이로 인한 잡음관계를 특징의 쌍으로 설정하여 씬을 탐지하는 방법[7], 기존의 지능망에 의미추출과 변화탐지를 분리하여 좀 더 강건하고 효율적인 방안[8] 등 해당 분야에 대한 연구들이 꾸준히 진행되고 있다.

2.2 Shot Detections

숏 탐지와 관련된 다양한 연구들도 진행 되었는데, 이들은 일반적으로 숏 경계(장면전환 포함) 추출에 초점을 두고 있다. 이와 관련된 연구들을 체계적으로 정리한 [9]에서는 물체와 카메라의 움직임 여부 및 장면전환 적용 여부를 구분하기 위해 불변 특징과 민감 특징을 고려해서 프레임별로 구분하는 방법, 연속적인 프레임들 사이의 유사성으로 Minkowski distance를 특징으로 하는 방법, 임계값을 기반으로 장면전환을 구분하는 방법 등이 비교 분석되었다. 이러한 방법들을 적용하기 위한 근간 기술들로 인접한 프레임들을 기반으로 화소값 차이, 색상 분포, 경계선 검출 후 이의 변화, DFT와 DCT를 적용한 후 신호의 변화, 블록매칭 방법으로 블록의 모션 벡터, 그리고 마지막으로 성능에서 좋은 결과를 보이지 못했다는는 산술 통계값들을 특징으로 하는 방법 등이 소개되었다.

제작자 관점에서의 씬 내의 숏들은 일정한 범위 내에서의 화면의 크기 변화만을 고려할 수 있겠다. 단, 씬 변화에 따른 인접 숏들에 장면전환 효과들이 추가될 수 있기에 씬 탐지보다는 좀 더 고려해야 할 요소들이 추가될 수 있겠지만 접근 방법은 대동소이하다고 볼 수 있다.

이상에서 소개한 씬 및 숏 탐지 방법들은 결과적으로 인접한 프레임들을 대상으로 프레임별 특징(들)의 유사성과 연속성을 분석하여 씬과 숏의 경계를 찾는 것이 일반적이었다고 볼 수 있겠다.

2.3 Focus Measure Operators

영상 내에서 초점과 관련된 논문들이 많이 소개 되었는데 이들 중 [10]에서는 컴퓨터비전에서 수동적 깊이 복구 및 3차원 재구성을 위한 방법에서 활용된 기존의 초점 측정 연산자들을 비교 분석하였다. 여기서는 6가지 수학적 원리를 기반으로 기울기, 라플라시안, 웨이블릿, 통계, DCT, 그리고 이들 5가지 분류에 포함되지 않는 원리를 토대로 36가지의 변형된 연산자들을 선정하여 비교 분석하였다. 그리고 이미지의 잡음 수준, 대비, 채도 및 커널 크기와 같은 다양한 조건 하에서 36가지 초점 측정 연산자들을 이미지의 모든 화소에 대한 초점 수준을 계산하여 성능을 비교 제시하였다. 해당 연구를 통해 6가지 수학적 원리 내의 연산자들은 대동소이한 결과를 보였으나, 특히 라플라시안 기반 연산자들이 앞서 언급했던 이미징 조건 하에서 가장 우수한 성능을 보였다고 한다. 하지만 이는 촬영 환경에 종속적이기에 그 중에서도 촬영장비에 의해 크게 좌우 되기 때문에 적정하게 상호 매칭하여 연구될 필요가 있다고 했다. 이와 관련된 연구로 상대적으로 아웃 포커싱된 블러(카메라 블러와 모션 블러 포함)를 탐지하는 연구[11-14]도 활발하게 진행 되었는데 이는 [10]에서 언급된 연산자들의 변형이라고 볼 수 있다.

2.4 Shot Type Detection

화면 내 인물의 얼굴 크기를 기반으로 숏을 구분하는 연구[15]에서는 클로즈업 숏, 미디엄 숏, 롱 숏의 감지 성능

으로 각각 95%, 90%, 53.3%을 보였다. 전달하려는 내용 측면에서 관심대상이 인물인 경우로 제한되며 숏의 크기 구분을 위한 판단 기준이 특정값으로 한정 되었다.

III. The Proposed Scheme

1. Basic Idea and Assumption

본 연구는 서론에서 언급했듯이 유튜브 등의 동영상 공유 플랫폼에서 활용 가능한 요약본 영상을 제작하는 것으로 장르는 드라마로 한정하였다. 아울러 전문적인 영상 제작자가 상황에 맞는 적정한 구도를 고려하여 입체적인 공간 구성으로 프레이밍 하였고, 드라마를 대상으로 했기에 인물들의 대화만으로도 충분히 프로그램의 스토리를 전달할 수 있다고 가정하였다. 이러한 가정을 기반으로 본 논문의 핵심은 촬영 시 충분히 화면분할을 고려하여 관심 대상을 위치 시켜 초점을 맞추고 인물들의 대화에 집중할 수 있는 미디엄 숏 크기 정도의 숏들을 추출하는 것이다. 결국, 이러한 인물 중심의 숏들은 1숏 혹은 2숏 정도가 적정하며, 입체감을 확보하고 시청자의 시선을 이끌기 위해서 장면 내 특정 피사체에 포커싱이 되었을 가능성이 높다는 일반적인 환경으로 한정하였다.

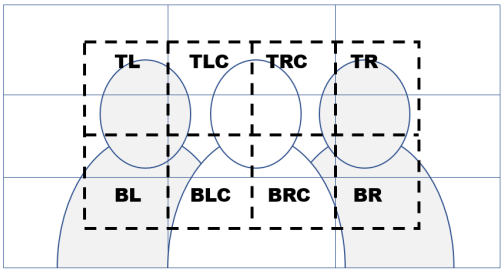

Fig. 2. Eight RoIs: Based on Golden ratio, a frame is divided into 3 regions, and then 8 small regions which named ‘TL’, ‘TLC’, ‘TRC’, ‘BL’, BLC’, ‘BRC’, and ‘BR’ are choose. Where ‘T’ means top, ‘L’ as left, ‘C’ as center, ‘R’ as right, and ‘B’ as bottom.

구도 측면에서는 1숏과 2숏은 3분할 화면 내에서 그림 2와 같이 중앙 혹은 좌측과 우측에 포커싱된 관심 대상이 위치할 가능성이 높으며, 해당 영역 내의 선명도가 다른 영역보다 상대적으로 높을 것이라고 예측 되었다. 이러한 실제작 환경에서의 경험적 가정과 예측을 기반으로 초점과 관련하여 선명도를 측정하기 위해서 앞선 여러 연구들을 기반으로 샘플 프레임에 라플라시안 연산자를 적용하여 확인한 결과는 그림 3과 같았다.

Fig. 3. some results with Laplacian operator

그림 3의 경우는 적정한 샘플 프레임을 선정하여 라플라시안 연산자들의 변형 중에서 분산을 적용한 것으로, 적용 영역들을 그림 2를 기반으로 한 프레임을 24개의 기준선을 설정한 후 이를 4x3 비율로 세분화하여 6,912개의 영역에 개별적으로 적용한 결과이다. 좌우측 상단과 우측 하단의 짙은 부분은 프로그램명과 방송사 로고 등이 수퍼임포우즈 되었기 때문이며, 그림 3의 결과를 토대로 해당 연산자는 선명한 영역과 선명하지 않은 영역을 구분하는데 적절하다고 판단할 수 있었다.

2. Blur Measure Test

그림 3의 경우에는 자막으로 인하여 미치는 영향이 크기 때문에 이를 제거한 후 그림 2에서의 8개 영역에 관심대상의 위치 및 미디엄 숏 크기의 프레임들의 구성을 고려하여 6개의 영역으로 재구성하였다. 자막이 선명하기 때문에 이것이 미치는 영향을 확인하기 위해서 그림 2에서 분할 했던 8개의 영역들을 6개의 영역(TL = TL + TLC, TC = TLC + TRC, TR = TRC + TR, BL = BL + BLC, BC = BLC + BRC, BR = BRC + BR)으로 병합하고 라플라시안 연산자를 적용해서 그림 4와 같은 결과를 얻을 수 있었고, 이를 통해 TR 영역과 BR영역이 배제됨을 확인할 수 있었다.

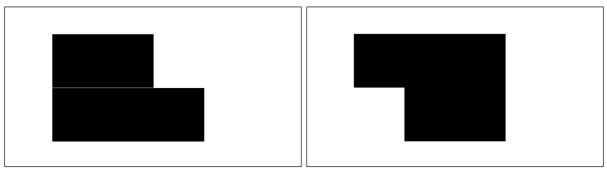

Fig. 4. Six regions with Laplacian operator; the left is the result of the above image of Fig. 3, and the right to the below.

Table 1. Correlations of regions with Laplacian operator at the total sample frames.

	TL	TC	TR	BL	BC	BR	L	C	R
TL	1.00	0.79	0.68	0.50	0.41	0.34	0.82	0.64	0.58
TC	0.79	1.00	0.88	0.46	0.45	0.42	0.53	0.79	0.77
TR	0.68	0.88	1.00	0.34	0.35	0.42	0.55	0.69	0.81
BL	0.50	0.46	0.34	1.00	0.86	0.67	0.63	0.78	0.62
BC	0.41	0.45	0.35	0.86	1.00	0.88	0.56	0.83	0.72
BR	0.34	0.42	0.42	0.67	0.88	1.00	0.45	0.76	0.80

Table 2. Correlations of regions with Laplacian operator at the chosen frames that may be of medium shot family.

	TL	TC	TR	BL	BC	BR	L	C	R
TL	1.00	0.12	0.00	0.09	0.04	0.00	0.92	0.08	0.00
TC	0.12	1.00	0.64	0.20	0.27	0.33	0.00	0.52	0.51
TR	0.0	0.64	1.00	0.02	0.12	0.34	0.03	0.35	0.71
BL	0.09	0.20	0.02	1.00	0.87	0.57	0.29	0.81	0.46
BC	0.04	0.27	0.12	0.87	1.00	0.85	0.24	0.90	0.68
BR	0.00	0.33	0.34	0.34	0.85	1.00	0.13	0.78	0.84

해당 실험을 통해서 포커싱된 영역은 그렇지 않은 영역보다 선명도가 높음을 파악할 수 있었지만, 영역별로 선명도를 구분하기 위한 임계값 설정이 강건하지 못하다는 결론에 도달했다. 분명 개별 영역들의 선명도는 명확히 수치상으로 구분이 가능했지만 이의 임계값들을 책정하기에 적정하지 않았다. 즉, 그림 3에서 상단 그림 프레임에 대한 측정 평균값은 94정도인데 반해 하단 그림은 10정도였다. 충분히 납득 가능한 결과이지만, 전체 프레임의 평균값을 기반으로 6개 영역들을 각각 비교하여 어떤 영역이 상대적으로 '선명하다'와 6개 영역들끼리 상호 비교하여 어떤 영역이 상대적으로 '선명하다'는 가정을 일반화할 수 없었다. 이에 대한 추가적인 연구가 필요하겠지만, [9]에서 언급했던 연산자들 중에서 실험을 통해 공간 도메인을 신호 도메인으로 변경하여 계산할 수 있는 FFT를 적용했을 경우에는 상대적으로 좀 더 일관성 있는 결과를 얻을 수 있었다. 이에 대한 근거가 될 수 있는 실험 데이터는 표1 ~ 표4에서 제시한다.

Table 3. Correlations of regions with FFT at the same frames in Table 1.

	TL	TC	TR	BL	BC	BR	L	C	R
TL	1.00	0.78	0.53	0.76	0.54	0.40	0.89	0.74	0.54
TC	0.78	1.00	0.77	0.62	0.67	0.60	0.72	0.87	0.74
TR	0.53	0.77	1.00	0.38	0.52	0.70	0.50	0.72	0.87
BL	0.76	0.62	0.38	1.00	0.72	0.45	0.87	0.72	0.48
BC	0.54	0.67	0.52	0.72	1.00	0.76	0.66	0.85	0.69
BR	0.40	0.60	0.70	0.45	0.76	1.00	0.46	0.73	0.87

Table 4. Correlations of regions with FFT at the same chosen frames in Table 2.

	TL	TC	TR	BL	BC	BR	L	C	R
TL	1.00	0.59	0.13	0.72	0.34	0.09	**0.83**	0.53	0.17
TC	0.59	1.00	0.57	0.49	0.56	0.43	0.50	**0.79**	0.55
TR	0.13	0.57	1.00	0.05	0.29	0.60	0.07	0.49	**0.80**
BL	0.72	0.49	0.04	1.00	0.57	0.18	**0.86**	0.60	0.19
BC	0.34	0.56	0.29	0.57	1.00	0.66	0.48	**0.83**	0.55
BR	0.09	0.43	0.60	0.18	0.66	1.00	0.14	0.62	**0.85**

연산자 및 영역 선택을 위한 실험은 30여 분 분량의 47,779개의 프레임들을 대상으로 8개의 영역들과 3개의 영역들 (L = TL + BL, C = TC + BC, R = TR + BR)에 라플라시안 연산자와 FFT를 각각 적용하여 영역별 상관성을 비교하기 위해 표1과 표3에서 제시하였고, 이들 프레임들 중 해당 연구의 취지에 적정한 숏 기반의 30,003개를 선택한 후 영역들간의 상관성을 표2와 표4에 각각 제시하였다. 표1과 표3을 비교해 본다면 라플라시안 연산자도 8개 영역에 적용한 결과가 적정해 보일 수 있겠으나, 반면에 표2와 표4와 비교해 본다면 FFT를 적용함으로써 3개의 영역만으로도 충분하다고 판단할 수 있다. 또한, 실험을 통해 FFT를 적용했을 때 선명도를 구분하기 위한 임계값 설정이 라플라시안 연산자를 적용했을 때 보다 영역별의 선명도를 일반화하기에 상대적으로 적합하다고 판단되었다. 즉, 실험을 통해 선명한 영역은 상대적으로 전체 영역의 선명도 보다 최소한 같거나 높을 것이라는 가설검정을 기반으로 선명도를 구분하기 위한 임계값으로 프레임 전체를 대상으로 계산한 FFT의 결과를 적용할 수 있었다.

3. Experiments

실험에 사용된 동영상의 정보는 H.264 - MPEG-4 AVC (part 10) (avc1), 1280 x 720, 29.97fps, Planer 4:2:0 YUV, ITU-R BT.709이고 대략 러닝타임은 26분 35초이다. 가능한 최소한의 자막 삽입 등을 고려한 원본 드라마의 내용을 유지하는 동영상을 선정하였고 개별 프레임들을 JPEG 형태로 분리 저장한 대략 47,779개의 프레임을 기반으로 실험 하였다. MPEG 기반 4:2:0이기에 해상도를 보존하기 위해 JPEG 화질을 3가지 정도로 구분하였으나 결과적으로 JPEG 화질이 미치는 영향이 미미하였다.

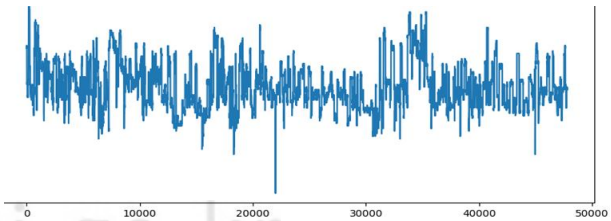

Fig. 5. Signal yielded by FFT to total frames

3장에서 언급했던 라플라시안 연산자와 FFT를 개별적으로 적용한 결과를 기반으로 영역들을 비교하여 최종적으로 화면을 수직으로 3분할 (Left, Center, Right) 영역으로 구분하였고, 프레임별 특징은 프레임 전체 영역과 3개의 영역들, 총 4개의 FFT 계산값으로 설정하였다. 프레임 전체 영역에 대한 FFT 계산값 파형은 그림 5와 같으며, 앞 부분인 1,500개 프레임들에 대응하는 확장된 파형은 그림 6과 같다.

프레임 전체 영역에 대한 연산 결과의 평균값과 3개 영역별 연산 평균값을 비교하여 같거나 작다면 이는 분명히 구도 기반의 상대적으로 초점이 맞는 영역이라고 판단하였다. 이러한 결과는 그림 7의 상단 이미지이며 높은 레벨('1')은 조건에 맞는, 낮은 레벨인 '0'은 조건에 맞지 않는 경우를 의미한다. 즉, '1'인 프레임만 선별하여 추출하는 것이다. 초기의 예상으로는 3개 영역들 중 전체 평균과 비교를 하되 상대적으로 선명한 영역이 존재한다면 어느 정도의 편차를 고려해야하고 이를 기반으로 분포가 어느 정도 이상인 영역을 검색하여 이를 기반으로 적정 프레임들을 선정하려 했으나 실험 결과로는 첫 번째 비교기준인 프레임 전체의 평균치를 기반으로 구분하는 것과 별다른 차이가 크지 않았다.

그림 7의 상단 그림은 조건에 적합한 결정 유무를 그래프로 나타낸 것으로 74.8% 정도 단축시킬 수 있었으나, 이를 기반으로 편집결정리스트를 추출하여 비선형편집기에서 온라인 렌더링으로 영상을 출력한다고 하더라도 연속적인 프레임 내에서 프레임의 결손(예를 들어 동일 숏 내에서 1초 동안 30프레임들 중 몇 개의 프레임들이 사라지는)은 시청자들로 하여금 불편함을 유발할 수 있기에 오프라인 변화점 탐지기법[16]을 적용하여 682개의 변화점을 검출했으며 이를 기반으로 수정 보완하여 그림 7의 하단 그림과 같은 결과를 얻었다.

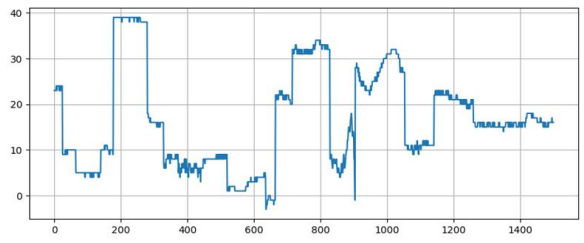

Fig. 6. Zoomed signal from 1st to 1,500th frame

결과적으로 47,799개의 입력 프레임들로부터 연속된 프레임들의 234개 모음으로 구분하였고 원본 대비 62.77%의 요약본을 제작할 수 있었다.

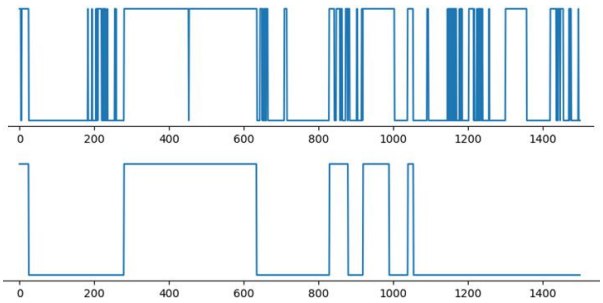

Fig. 7. the above figure by the first condition and the below is final results after applying change point detection in the same range of Fig. 6

이렇게 선정된 프레임들을 수작업으로 분류한 프레임들과 비교하여 혼돈행렬 기반의 평가지표로 정확도는 66.85%, 정밀도는 69.80%, 재현율은 83.18%, F1-스코어는 75.9%를 보였다.

IV. Conclusions and Future Work

본 논문은 영상이 내포한 의도를 전달하기 위해서 시청자들의 시선을 끌고 다닐 수 있어야 함을 전제로, 프레임 내에 공간감을 확보하고 시청자들의 시선을 유도 및 집중시키기 위해 특정 피사체에 포커싱된 프레임들과 대사 위주의 숏으로 간주될 수 있는 미디엄 숏 크기의 프레임들을 추출하기 위해, 기존에 연구되어 발표된 요소 기술들로 FFT를 기반으로 프레임을, 그리고 오프라인 변화점 탐지기법 등을 적용하여 숏들을 추출했고, 이를 기반으로 편집결정 리스트를 구성하여 75.9%의 F1-스코어를 갖는 62.77%의 요약본 영상을 제작했다.

본 연구는 몇 가지 문제점들을 내포하고 있었다. 첫 번째는 입력 영상의 화질이다. 앞서 입력 영상 정보에 대해서 기술했듯이 MPEG기반의 4:2:0 압축이기에 예측 프레임들에 대한 낮은 화질과 촬영 원본이 아닌 몇 번의 렌더링으로 인한 화질의 열화, 두 번째는 촬영 과정에서 조명의 변화 및 렌즈 블러 외 모션 블러 등과 카메라의 급격한 움직임, 세 번째는 후반편집에서 적용되었을 수도 있는 여러 효과들, 네 번째는 자막 삽입이다. 비록 렌즈 및 카메라 자체의 결함(초점이 맞지 않는)을 무시한다 하더라도 상기 상황들이 해당 연구에 있어서 상당한 걸림돌이 되었고 실험을 통해 고려해야할 상황의 경우 수가 너무 다양했다.

비록 실험 대상 영상이 원본이 아닌 편집된 요약본이었기에 어느 정도의 제약이 있었지만, 해당 연구가 스토리의 연속성을 유지하고 단시간 내에 내용 파악을 위한 요약본 영상 제작에 어느 정도 도움이 될 수 있을 것으로 보이며,

추가적인 연구로는 오디오(다이얼로그) 분석을 추가하여 추출한 숏들의 명확한 바운딩과 이들의 특징점들을 기반으로 비지도학습 모형들을 활용하여 본 연구에서 특정하는 상황별 프레임들을 군집하는 방법을 검토하고, 이의 결과를 기반으로 자동분류할 수 있는 방법을 모색할 예정이다.

ACKNOWLEDGEMENT

This work was supported by 2022 Hannam University Research Grant.

REFERENCES

[1] E. J. Choi, “Theory of Film Making”, Communication Books, 2020.
[2] Roy Thompson, “Grammar of the Shot,” Focal Press, pp. 26-27, 66-97, 2021.
[3] Hervert Zettl, “Television Production Handbook,” Thomson, pp. 401-408, 2004.
[4] Hwon, Hyuk Min, Kim, Hyen Ki, “A Study on the Ratio of Video Screen Using Modulor,” Korea Design Forum, no. 15, pp.41-48. 2007
[5] V. T. Chasanis, A. C. Likas and N. P. Galatsanos, "Scene Detection in Videos Using Shot Clustering and Sequence Alignment," in IEEE Trans. Multimedia, vol. 11, no. 1, pp. 89-100, Jan. 2009, DOI:10.1109/TMM.2008.2008924.
[6] Dongwook Shin, et al., “Video Scene Detection using Shot Clustering based on Visual Features,” Journal of Korea Intelligent Information Systems Society, vol. 18, no.2, pp.47-60, 2012.
[7] K. Sakurada, M. Shibuya and W. Wang, "Weakly Supervised Silhouette-based Semantic Scene Change Detection," 2020 IEEE ICRA, Paris, France, 2020, pp. 6861-6867, DOI: 10.1109/ICRA40945.2020.9196985.
[8] Enqiang Guo, et. al., “Learning to Measure Change: Fully Convolutional Siamese Metric Networks for Scene Change Detection,” DOI:10.48550/arXiv.1810.09111
[9] S. H. Adbulhussain, et al, “Methods and challenges in Shot Boundary Detection: A Review,” Entropy(Basel), 2018 Arp; 20(4):214. DOI:10.3390/e20040214
[10] Pertuz, S., et. al., (2013). Analysis of focus measure operators for shape-from-focus. Pattern Recognition, 46(5), 1415-1432. DOI:10.1016/j.patcog.2012.11.011
[11] Rupali Yashwant Landge, Rakesh Sharma, “Blur Detection Methods for Digital Images-A Survey.” Int. Journal of Computer Applications Technology and Research, Vol.2-4, pp. 495-498, 2013, DOI:10.7753/IJCATR0204.1019
[12] Renting Liu, Zhaorong Li and Jiaya Jia, "Image partial blur detection and classification," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8, DOI:10.1109/CVPR.2008.4587465.
[13] P. A. Pagaduan, et al, “iBlurDetect: Image Blur Detection Techniques Assessment and Evaluation Study,” In Proceedings of the International Conference on CESIT 2020, pp.286-291, DOI:10.5220/0010307700003051
[14] S. Alireza Golestaneh, Lina J. Karam, “Spatially-Varying Blur Detection Based on Multiscale Fused and Sorted Transform Coefficients of Gradient Magnitudes,“ Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5800-5809
[15] Y.T Baek, S.B Park, “Shot Type Detecting System using Face Detection,” Journal of The KSCI, Vol. 17, No.9, pp. 49-56, Sep. 2012. DOI:10.9708/jksci/2012.17.9.049
[16] Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. Signal Processing, 167, 107299. DOI:10.1016/j.sigpro.2019.107299

Authors

Hyung Lee received the B.S. degree in Computer Science, M.S. and Ph.D. degrees in Computer Engineering from Chungnam National University, Korea, in 1995, 1997 and 2015, respectively.

Dr. Lee has been a professor in the department of Broadcasting Contents at Daejeon Health Institute of Technology in Daejeon, Korea since 2001. He is interested in interactive media based on information technology.

E-Jung Choi received the M.S. and Ph.D. degrees in Media Communication from Hankuk University of Foreign Studies, Korea, in 1989 and 2004, respectively. Dr. Choi joined the faculty of the department of Multimedia, Hannam University, Dajeon, Korea, in 1999. He is currently a professor in the department of Media & Visual Communications, Hannam University. He is interested in TV program production.