사용자의 요구를 반영하는 동영상 요약 알고리즘

(A Video Abstraction Algorithm Reflecting Various Users Requirement)

정진국[†] 홍승욱[†] 낭종호[††] 하명환[†††] 정병희[†††] 김경수[†††]

(Jin-Guk Jeong) (Seung-Wook Hong) (Jong-Ho Nang) (Myung-hwan Ha) (Byung-Hee Jung) (Gyung-Su Kim)

요 약 자동으로 동영상을 요약하는 알고리즘은 다양한 방향으로 진행되어 왔다하지만 이러한 알고리즘들은 한가지 저급수준 내용정보만을 이용하여 동영상을 요약하였기 때문에 요약하는 사용자의 주관을 반영할 수 없다는 단점이 있다. 즉, 동영상 요약이라는 것은 내용에 대한 전반적인 이해에 바탕을 두고 중요한 샷을 선택하는 것이라고 정의할 수 있는데 이 경우에 중요한 샷이라고 결정하는 것은 요약자의 주관에 따라 달라질 수 있기 때문에 사용자의 주관을 반영할 수 없다는 것은 큰 단점으로 대두될 수 있다.본 논문에서는 사용자의 요구를 반영하는 동영상 요약 알고리즘을 제시한다.알고리즘에서는 일반적으로 많이 사용하는 동영상 요약에 대한 목적함수와 이들에 대한 가중치를 이용한다.본 논문에서는 동영상 요약을 목적함수를 극대화 시킬 수 있는 샷들의 집합으로 정의하는데 이 경우 문제점으로 제시될 수 있는 것이 계산량이 많다는 것이다. 즉, n개의 샷을 가진 동영상에 대하여2^n번의 계산량이 필요하기 때문에 전체 알고리즘의 수행 시간이 많아지는 것이다.본 논문에서는 이러한 문제점을 해결하기 위하여 빠른 시간에 근사 샷들의 집합을 구할 수 있는 Simulated Annealing 알고리즘을 이용하였다.실험 결과에 의하면 본 논문의 알고리즘은 요약자의 주관을 반영시킬 수 있고 Simulated Annealing 을 이용하여 빠른 시간에 원하는 요약을 할 수 있음을 확인하였다.본 논문의 알고리즘은 동영상을 대상으로 하는 디지털 비디오 라이브러리와 같은 응용 분야에 이용할 수 있을 것이다.

키워드 : 동영상 요약, 멀티미디어, 목적 함수, 디지털 비디오 라이브러리

Abstract Video abstraction is a process to pick up some important shots in a video, while the important shots might vary on the persons subjectivity. Previous works on video abstraction use only one low level feature to choose an important shot. This thesis proposes an abstraction scheme that selects a set of shots which simultaneously satisfies the desired features(or objective functions) of a good abstraction. Since the complexity of the computation to find a set of shots which maximizes the sum of object function values is $O(2^n)$, the proposed scheme uses a simulated annealing based searching method to find the suboptimal value within a short period of time. Upon the experimental results on various videos, we could argue that the proposed abstraction scheme could produce a reasonable video abstraction. The proposed abstraction scheme used to build a digital video library.

Key words : Video abstraction, Multimedia, Object function, Digital Video Library

.본 연구는 서강대학교 산업기술연구소 지원으로 이루어졌음

† 비 회 원 : 서강대학교 컴퓨터학과
jjguk@mlneptune.sogang.ac.kr
plvesta@hotmail.com
†† 종신회원 : 서강대학교 컴퓨터학과 교수
jhnang@ccs.sogang.ac.kr
††† 비 회 원 : KBS 기술연구소 연구원
mhha@kbs.co.kr
bhjung@kbs.co.kr
odyssey@kbs.co.kr

논문접수 : 2002년 1월 8일
심사완료 : 2003년 3월 17일

1. 서 론

인터넷 및 컴퓨터 네트워크가 발달함에 따라 멀티미디어 데이타의 사용이 일반화되었다. 특히 비디오 데이타는 사람이 쉽게 인지할 수 있기 때문에 다른 미디어 데이타에 비해 점점 더 중요한 역할을 하고 있다. 하지만 이러한 비디오 데이타를 사용하는 데 있어서 비디오 데이타의 큰 용량은 문제점으로 대두 될 수 있다. 예를

들어 사용자가 인터넷을 통해 방송 프로그램을 보는 경우에 내용을 알기 위해서는 전체 비디오를 수신하면서 상영하여 보아야 하는 것이다. 이러한 문제점을 해결하기 위해 많이 사용되는 방법 중에 하나가 동영상 요약 방법이다.

동영상 요약 방법[1,2,3,4]에 대해서는 기존에 많은 연구들이 진행이 되어 왔다. 샷의 길이를 이용하는 동영상 요약 방법[2]이 있었고, 칼라 히스토그램[3]을 이용하는 동영상 요약 방법이 있었다. 이외에도 움직임 에너지[4], 전체적인 색 구성[1]을 이용하는 방법 등 다양한 형태로 많은 연구가 진행되어 왔다. 하지만 기존의 방법에서는 단순히 하나의 저급수준 내용정보만을 이용하여 동영상을 요약하였기 때문에 요약하는 사용자의 주관을 반영할 수 없다는 단점이 있다. 동영상 요약이라는 것은 내용에 대한 전반적인 이해에 바탕을 두고 중요한 샷을 선택하는 것이라고 정의할 수 있는데 이 경우에 중요한 샷이라는 것은 요약자의 주관에 따라 달라질 수 있기 때문에 사용자의 주관을 반영할 수 없다는 것은 큰 단점으로 대두될 수 있다.

본 논문에서는 사용자의 요구를 반영하는 동영상 요약 알고리즘을 제시한다. 일반적으로 많이 사용되는 동영상 요약에 대한 목적함수와 이들에 대한 가중치를 사용하여 사용자가 원하는 목적함수에 가중치를 많이 줄 수 있도록 한다. 본 논문에서는 동영상 요약을 목적함수를 극대화 시킬 수 있는 샷들의 집합으로 정의하는데 이 경우 문제점으로 제시될 수 있는 것이 계산량이 많다는 것이다. 즉, n개의 샷을 가진 동영상에 대하여 알고리즘을 적용하는 경우 2^n번의 계산이 필요하기 때문에 계산량이 많아지는 것이다. 이러한 문제점을 해결하기 위하여 본 논문에서는 빠른 시간에 근사 샷들의 집합을 구할 수 있는 Simulated Annealing 알고리즘을 이용하였다. 실험을 통해 본 논문의 알고리즘을 이용하면 사용자의 주관을 반영할 수 있음을 알 수 있었고 Simulated Annealing을 이용하여 시간에서의 문제점을 해결할 수 있음을 알 수 있었다. 본 논문의 알고리즘은 동영상을 대상으로 하는 VOD(Video On Demand)나 DVL(Digital Video Library)와 같은 응용 분야에 유용하게 이용할 수 있을 것이다.

2. 연구 배경

동영상 요약이라는 것은 본래의 동영상이 가지는 핵심적인 내용은 그대로 유지하는 상태에서 동영상 전체 길이를 줄인 후에 새로운 동영상 혹은 이미지의 집합을 만드는 것을 얘기한다. 일반적으로 동영상은 여러 개의 샷과 이러한 샷들의 집합인 신으로 이루어진다. 이러한 구조를 고려하는 상태에서 동영상을 요약한다는 것은 그림 1과 같이 동영상의 여러 샷/신 들 중에서 중요한 샷/신을 선택하여 짧은 길이의 동영상을 만드는 것을 의미한다.

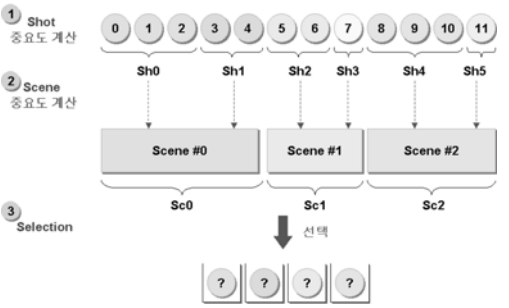


그림 1 동영상 요약의 전체적인 구성

Moca 시스템[1]에서는 전체적인 색 구성을 이용하여 동영상을 요약하고 있다. 우선 동영상 데이타를 샷으로 묶고, 이러한 샷을 바탕으로 신으로 그룹화한다. 그룹화 과정이 끝나면 신으로부터 색 정보와 움직임 정보를 추출하고 오브젝트를 인식한다. 이러한 작업들을 수행한 후 각각의 데이타를 미리 정해진 특성을 가지고 있는 여러 가지 이벤트와 비교하여 어떤 이벤트인지 결정하는 것이다. 예를 들면 대화 장면이라는 것은 규칙적인 인물의 등장과 짧은 샷들의 반복적인 특성이 있다는 특징을 이용하여 인식하게 되는 것이다. 이외에도 오디오 정보의 크기, 주파수, 음의 고저 등을 고려하여 총 소리, 폭발 소리 등을 인식하게 되고 이러한 특징을 바탕으로 하나의 요약된 동영상을 만들게 된다.

Lienhart의 방법[2]에서는 샷의 길이를 이용한다. 이 방법에서는 디지털 비디오 카메라의 특징을 이용한다. 즉, 각 프레임 안에 포함된 시간 정보를 이용하여 샷의 거리를 구하는 것이다. 하나의 샷은 샷의 처음 프레임의 시간 t_b^i와 마지막 프레임의 시간 t_e^i로 구성된 $S_i = [\, t_b^i, t_e^i]$로 나타낼 수 있다. 그 후 이렇게 표시된 샷과 샷의 거리 $\Delta t(S_i, S_j)$는 다음과 같은 식으로 구하게 된다.

$$\Delta t(S_i, S_j) = \begin{cases} 0 & if \quad [t_b^i, t_e^i] \cap [t_b^j, t_e^j] \neq \varnothing \\ t_b^j - t_e^i & if \quad (t_b^j > t_e^i) \end{cases}$$

즉, 샷 S_i,와 S_j가 시간적으로 중첩되지 않으면 샷의 거리 $\Delta t(S_i, S_j)$는 그 샷 S_j의 시작시간과 S_i의 종료시간과의 차이로 구하게 된다. 만일 중첩이 되면 거리가 0인 같은 샷으로 본다. 이와 같이 각 샷의 거리를 구한

이후에 인접한 샷은 비슷하다고 가정한 후 인접한 샷들을 클러스터링하는 과정을 거치고 클러스터링된 각 집합내의 샷들은 동일한 중요도를 갖게 되므로 각각의 클러스터에서 무작위로 클립을 선택하여 요약을 하게 되는 것이다.

칼라 히스토그램을 이용하는 방법에는Sull의 방법[3]이 있다. 이 방법은 그림 2와 같이 칼라 히스토그램을 이용하여 프레임과 프레임 사이의 Distance를 구한 후, 칼라 히스토그램의 차이가 가장 적은 집합을 Semi-Hausdorff Distance를 이용하여 추출하게 된다. 이렇게 추출된 프레임은 그 프레임이 속한 집합을 대표하게 되고, 다시 이러한 프레임들끼리 Semi-Hausdorff Distance를 되풀이하는 것이다. 이러한 과정을 하나의 프레임이 나올 때까지 되풀이하면 마지막 하나의 프레임이 전체 스트림을 대표하는 프레임이 되는 것이고 이 과정에서 어느 레벨의 프레임들을 모으면 전체 동영상의 요약이 되는 것이다.

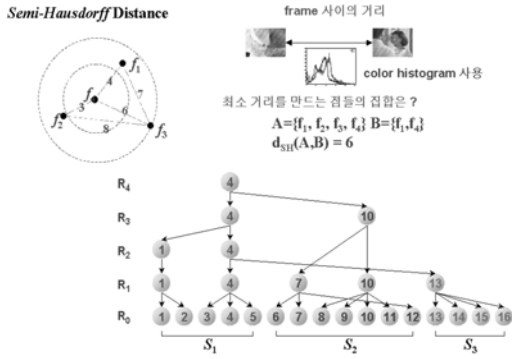


그림 2 칼라 히스토그램을 이용한 동영상 요약

샷의 움직임 에너지를 이용하여 동영상을 요약하는 방법에는 Nam의 연구[4]가 있다. 이 방법은 하나의 동영상을 sub-shot으로 나눈 후에 각각의 sub-shot k에 포함된 L개의 프레임 i에 대하여 1차원 Wavelet 변환을 이용해서 움직임 에너지 $m_i^k(m,n)$을 구한다. 이러한 프레임 에너지 절대값의 평균이 sub-shot의 움직임 에너지가 되고 이를 수식으로 나타내면 다음과 같다.

$$\text{Motion intensity index} = \frac{1}{L}\sum_{b}^{e}\sum_{m,n}\left|m_i^k(m,n)\right|$$

이러한 sub-shot의 움직임 에너지를 기준으로 움직임이 많은 sub-shot에서는 많은 프레임을 추출하고, 움직임이 적은 sub-shot에서는 적은 수의 프레임을 추출하여 동영상을 요약하는 것이다.

이와 같이 동영상 요약에 대한 기존의 연구에서는 샷의 길이[2], 칼라 히스토그램[3], 움직임 에너지[4], 전체적인 색 구성[1] 등을 이용하였다. 기존의 연구의 공통점은 프레임이나 샷에 대해 한가지 저급 수준 내용정보만으로 평가하여 일률적으로 적용하였다는 것이다. 하지만 요약이라는 것은 그 내용에 대한 전반적인 이해에 바탕을 두고, 내용을 인식하여 수행하는 것이기 때문에 요약하는 사람의 주관과 주변에 있는 샷에 따라 바뀔 수 있는 특성이 있다. 그렇기 때문에 단순히 한가지 저급 수준 내용 정보만으로 동영상을 요약하는 방법에는 문제가 있을 수 있고 또, 이러한 요약자의 주관과 저급 수준 내용 정보와의 관계를 선형함수로 표시하기에는 어려움이 따른다는 문제점이 있다.

3. 목적 함수를 이용한 비디오 요약

본 논문에서는 요약자의 주관을 반영하기 위해 여러 개의 목적함수(Object Function)와 이에 대한 가중치를 조절하여 요약자에게 가장 맞는 요약을 만들 수 있는 방법을 제안한다. 본 논문에서는 동영상 요약을 그림3에서와 같이 목적 함수들의 가중치를 합한 값을 극대화 시키는 동영상 내의 샷들의 집합으로 정의하고 있다.그림 6에서 O(X)는 동영상 요약에 사용되는 목적 함수들을 나타내는 것이고, W는 이러한 목적 함수들에 대한 가중치를 나타내는 것이다. 이 장에서는 이러한 목적 함수들의 종류에 대해서 설명하도록 한다.

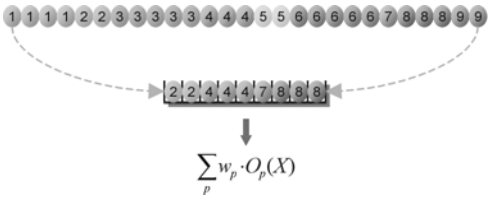


그림 3 동영상 요약

3.1 목적 함수를 이용한 비디오 요약

본 논문에서 제안하는 목적 함수를 이용하여 동영상 요약을 하기 위해서는 몇 가지 정의할 사항들이 있는데 그림 4는 이를 위해 정의된 용어들을 보여주고 있다. 비디오 요약은 n개의 샷으로 이루어진 동영상 V={v_i | 0<= i <=n}에서 k개의 샷을 선택하여 이루어진 집합 $X = \{x_i \mid x_i \in V, 0 \le i \le k, 0 \le k \le n\}$로 정의할 수 있다. 또, 각 샷들은 m개의 특성(feature)를 가질 수 있는데, i번째 샷이 갖는 j번째 특성을 f_j^i라 정의한다. 각각의 샷 x_i는 시작 프레임을 나타내는 S_i, 끝 프레임을 나타내는

E_i, 샷의 길이를 나타내는 L_i로 정의한다. 그리고 원하는 요약의 길이는 T로 정의한다.

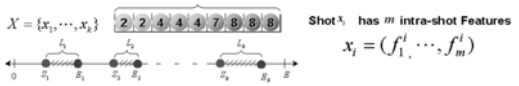


그림 4 동영상 요약을 위해 사용된 용어

이러한 정의를 바탕으로 본 논문에서 사용하는 목적함수는 7가지이다. 목적함수는 크게 두 종류로 나눌 수 있다. 하나는 샷내의 특성을 고려하여 계산하는 목적함수이고, 다른 하나는 샷끼리의 특성을 고려하는 목적함수들이다.

3.1.1 샷내의 특성을 고려하는 목적함수

- Event(O_4): 영화 예고편을 보면 움직임이 많은 샷들로 구성되는 것이 대부분이다. 기존의 연구[4]에 의하면 동영상에서 이러한 움직임들은 움직임 에너지로 표현될 수 있다. 본 논문에서 사용된 목적함수에서는 요약된 비디오 안에 움직임 에너지가 큰 샷들이 많을수록 높은 값을 가질 수 있게 하기 위하여 각각의 샷 x의 움직임 에너지 f의 절대값의 합인 $\frac{1}{k}\sum_{i=1}^{k}|f_j^i|$를 이용하고 있다. 즉, $O_4(X)=\frac{1}{k}\sum_{i=1}^{k}|f_j^i|$로 정의한다.
- Not too short(O_3): 기존의 연구[5]에 의하면 사람이 동영상의 내용을 이해할 수 있는 최소한의 샷의 길이는 3.5초이다. 이 목적함수는 최소한의 샷의 길이보다 작은 샷들을 제거하기 위한 함수이다. 3.5초에 해당하는 상수를 C라고 정의한 이후에 C보다 작은 정도에 따라 더 작은 값을 가질 수 있도록 하기 위하여 $\frac{x-|x|}{2}$를 C에 대해 평행이동하여 구한 식 $\frac{1}{k}\sum_{i=1}^{k}\frac{L_i+C-|L_i-C|}{2C}$을 목적함수로 사용한다. 즉, $O_3(X)=\frac{1}{k}\sum_{i=1}^{k}\frac{L_i+C-|L_i-C|}{2C}$로 정의한다.
- Shot Exclusion(O_6): 영화 예고편 같은 경우에는 일반적으로 결론을 포함시키지 않는다. 이 목적함수는 이와 같은 사용자의 요구를 받아들이기 위한 함수이다. 이 함수는 결론에 해당하는 부분의 시간값(프레임)을 C라고 정의하는 경우 C를 넘게 되면 작은 값을 가질 수 있어야 한다. 이를 위해 Not too short와 같이 $\frac{x-|x|}{2}$를 C를 중심으로 평행이동 하여 구한 식

$$\frac{1}{k}\sum_{i=1}^{k}\left(\frac{E-E_i-|E_i-C|}{2(E-C)}+\frac{1}{2}\right)$$

을 사용한다. 즉,

$$O_6(X)=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{E-E_i-|E_i-C|}{2(E-C)}+\frac{1}{2}\right)$$

로 정의한다.

3.1.2 샷사이의 특성을 고려하는 목적함수

- Well Distributed(O_1): 동영상을 요약하는 경우에는 동영상 내에서 한쪽 부분에 치우치지 않고 동영상 전체에 걸쳐서 고루 선택될 수 있으면 좋다고 말할 수 있을 것이다. 이 목적함수는 이러한 동영상 요약의 특성을 반영하기 위한 목적 함수이다. 이를 반영하기 위하여 사용하는 값은 샷 사이의 거리값과 편차이다. 집합 X내에 속한 샷들 사이의 거리 평균은 $\mu=\frac{E-\sum_{i=1}^{k}L_i}{k+1}$와 같이 구할 수 있게 되고, 샷간 거리 편차는 $|S_1-\mu|+|E-E_k-\mu|+\sum_{i=2}^{k}|(S_i-E_{i-1})-\mu|$로서 나타낼 수 있다. 이러한 편차가 작을수록 큰 값을 나타나야 하기 때문에 거리 편차의 역수인

$$\frac{1}{|S_1-\mu|+|E-E_k-\mu|+\sum_{i=2}^{k}|(S_i-E_{i-1})-\mu|}$$

를 목적함수로 정한다. 즉,

$$O_1(X)=\frac{1}{|S_1-\mu|+|E-E_k-\mu|+\sum_{i=2}^{k}|(S_i-E_{i-1})-\mu|}$$

가 된다.
- Well Fitting(O_2): 요약된 동영상이 사용자의 요구로 들어오는 요약 길이와 같으면 잘된 요약이라고 할 수 있을 것이다. 이 목적 함수는 이러한 동영상 요약의 요구 사항을 위한 목적 함수이다. 이를 위하여 선택된 샷들 길이의 합인 $L=\sum_{i=1}^{k}L_i$이 주어진 요약 길이(T)와의 차이가 작으면 높은 값을 가질 수 있게 하기 위하여 sech(x)를 평행이동 한 후에 축소한

$$\frac{2}{e^{C(L-T)}+e^{-C(L-T)}}$$

을 사용한다. 즉,

$$O_2(X)=\frac{2}{e^{C(L-T)}+e^{-C(L-T)}}$$

가 된다.
- Concise(O_5): 요약된 동영상 안에서는 비슷한 샷이 많

은 것보다는 다양한 샷이 존재하는 것이 요약하는 사용자를 더 만족시킬 수 있을 것이다. 이 목적 함수는 이러한 요구사항을 만족시키기 위한 함수로 요약된 동영상의 샷안에 다양한 칼라 히스토그램을 가질수록 높은 값을 가질 수 있도록 하였다. 한 샷내 키 프레임의 칼라 히스토그램의 Bin인 f_z^i와 뽑힌 다른 샷내 키 프레임의 칼라 히스토그램의 Bin인 f_z^{i+1}의 차이를 내적(inner product)을 구할 때 사용하는 각의 크기인 $\frac{1}{k}\sum_{i=0}^{k-1}\frac{\vec{f}_z^{\,i}\cdot\vec{f}_z^{\,i+1}}{\left|\vec{f}_z^{\,i}\right|\cdot\left|\vec{f}_z^{\,i+1}\right|}$ 를 이용하여 구한다. 즉,

$$O_5(X)=\frac{1}{2}-\frac{1}{2}\cdot\frac{1}{k}\sum_{i=0}^{k-1}\frac{\vec{f}_z^{\,i}\cdot\vec{f}_z^{\,i+1}}{\left|\vec{f}_z^{\,i}\right|\cdot\left|\vec{f}_z^{\,i+1}\right|}$$ 가 된다.

- Non Bias(O_7): 기존 연구[2]에 따르면 동영상 요약에 포함된 샷들 중 한 샷이 너무 많은 부분을 차지하게 되면 좋은 요약이라고 하기 어렵다. 이 목적 함수는 이와 같이 너무 큰 샷은 제거하기 위하여 이용되는 목적 함수로 뽑힌 샷 중에 최대 샷의 길이의 비가 적을 때 큰 값을 가질 수 있도록 하였다. 앞서서 뽑힌 샷들의 평균 길이인 $\mu=\frac{E-\sum_{i=1}^{k}L_i}{k+1}$와 제일 큰 샷의 길이인 $\max(L_i)$와의 차의 비를 목적 함수로 사용한다. 즉, 목적함수는

$$O_7(X)=\frac{1}{\left|\mu-\max(L_i)\right|},\ \forall i\in\{1,\cdots,k\}$$ 가 된다.

이러한 목적 함수를 이용하여 사용자가 요구하는 가중치를 적용하게 되면 다음과 같은 식을 얻을 수 있게 된다.

즉, 좋은 요약이란 함수 G(X)의 값을 극대화 시킬 수 있는 샷의 집합 X라고 할 수 있다.

앞서 설명한 여러 가지 목적 함수들에게 가중치를 적용하기 위해서는 목적 함수들이 같은 값의 범위를 갖고 있어야 하기 때문에 정규화 과정이 필요하게 된다. 목적 함수 중 O_1과 O_7의 경우에는 값의 범위가 0부터 ∞까지이다. 무한대의 값을 가지는 경우에는 1의 값을 갖게 하고, 0의 값을 가지는 경우에는 0으로 나타내기 위하여 [0:1]의 범위로 정규화 하게 된다. 이를 위하여 그림 5에서와 같이 정규화하는 함수인 $\frac{x}{x+1}$로 다시 사상(mapping)해야 한다. 이는 $[0:\infty]$의 범위를 갖는 $\frac{1}{x}$을 단조 증가 함수인 $-\frac{1}{x}$로 변환한 후 x가 [0:1]사이의 범위에 있게 하기 위해 x축으로 -1만큼 이동시킨 $-\frac{1}{x+1}$로 변환한 후 다시 y도 [0:1]의 범위에 있게 하기 위해 y축으로 1만큼 이동시킨 $-\frac{1}{x+1}+1$을 사용함으로서 x,y 모두 [0:1]사이의 범위를 갖는 정규화된 함수를 사용하게 된다.

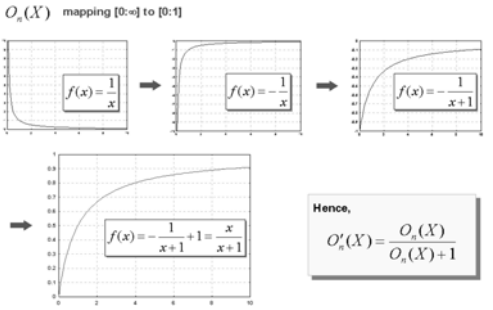


그림 5 목적 함수를 정규화하기 위한 변환 함수

본 논문에서 제안하는 목적 함수에 기반하는 동영상 요약 알고리즘을 이용하면 기존의 여러 방법 또한 적용할 수 있다. 그림 6에서와 같이 우선 시간 축을 바탕으로 한 요약은 전체 동영상에서 골고루 샷이 선택될 수 있게 O_1과 O_2에만 가중치를 두어 표현할 수 있고, 칼라 히스토그램을 이용한 요약 방법[3]은 O_2와 O_5에만, 움직임 에너지를 이용한 요약 방법[4]은 O_2와 O_4에만, 마지막으로 일반적인 예고편의 경우에는 O_2와 O_6에만 가중치를 부여하여 표현할 수 있다.

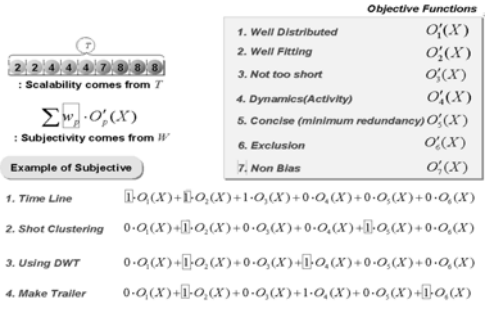


그림 6 목적함수로서 표현한 기존의 요약 방법

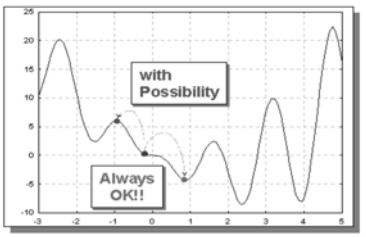
본 논문에서 제안한 목적 함수를 이용한 요약 방법의 장점은 확장(Scalability)과 주관(Subjectivity)적 요약이

가능하다는 데 있다. 여기서 확장이라는 것은 요약 길이(T)를 변화시킴에 따라 목적 함수(O)가 변화될 수 있다는 점이고, 주관은 본 논문에서 제안하는 목적 함수에 대한 가중치(W)를 변경함으로 해서 요약자의 주관을 반영시킬 수 있다는 것이다. 반면 단점으로는 계산의 복잡도가 크기 때문에 요약하는 데 시간이 오래 걸린다는 점이다. 즉, 샷들 n개로 구성된 동영상에서 목적 함수의 최대값을 구하기 위해서는 약 2^n번의 계산이 필요하다는 것이다. 이러한 단점을 극복하기 위하여 본 논문에서는 목적 함수를 최대화 할 수 있는 근사 샷들의 집합을 빠른 시간에 구할 수 있는 Simulated Annealing을 이용한다.

3.2 Simulated Annealing을 이용한 탐색

Simulated Annealing은, 기본적으로 국소탐색(Local Search) 방법을 개선한 방법으로 조합 최적화(Combinatorial Optimization) 문제에 많이 쓰이지만, 유전자 알고리즘이나 신경망 학습 등 다른 분야에도 폭 넓게 쓰이는 방법[6]이다.

그림 7은 Simulated Annealing알고리즘을 나타내는 그림이다. 주어진 구간에서 최소값을 찾기 위해서는 모든 점을 탐색해야만 한다. 하지만 이런 경우에는 너무 시간이 걸리기 때문에 이러한 단점을 보완하기 위해 Simulated Annealing은 그림 7에서와 같은 알고리즘을 사용한다. 알고리즘에 대해 보다 자세히 살펴보면 주어진 절대 온도 T가 있고, 임의의 점 C_i를 선택한다. 그 후 점 C_i의 함수값 $E_i = F(C_i)$을 구한다. 또 다시 임의의 점 C_{i+1}을 선택하고 함수값 $E_{i+1} = F(C_{i+1})$을 구한다.



1. An initial configuration, C_i is chosen, at random or by some other, usually heuristic method. This configuration has an associated $E_i = F(C_i)$.
2. An initial temperate T is determined.
3. A random change in the configuration is generated this change produces a possible new configuration C_{i+1}.
4. If $E_{i+1} \le E_i$ then the change is a always accepted.
5. If $E_{i+1} \ge E_i$ then the change is accepted with probability
6. Update temperature T if required. Go back to 2.

Repeat loop until done.

그림 7 Simulated Annealing 알고리즘

만일 E_{i+1}이 E_i 보다 작으면 언제나 최소값으로 받아들이지만, 그렇지 않은 경우에는 $P = e^{-\frac{(E_i - E_{i+1})}{kT}}$ 확률로 받아들인다. 항상 작은 값을 받아들이게 되면 국소 최소값(Local Minimum)만을 찾게 되기 때문에 이와 같은 방법을 이용하게 된다. 이 때 확률 P는 절대 온도 T가 클수록, $E_{i+1} - E_i$의 차이가 클수록 커지게 된다. 이와 같은 과정을 거친 후에는 T에 $\rho(0 \le \rho \le 1)$값을 곱한 후, 다시 반복하게 되는데, 이 ρ가 클수록 빨리 근사값을 찾게 된다.

그림 8은 본 논문에서 사용하는 목적 함수를 이용한 동영상 요약 알고리즘을 나타내는 그림이다. 알고리즘을 설명하면 우선 하나의 샷 집합을 임의로 뽑고 가중치를 적용하여 이 샷 집합에 대한 목적 함수들의 합(e1)을 구한다. 또 다시 하나의 샷 집합을 뽑고 가중치를 적용한 목적 함수들의 합(e2)을 구한다. 만약 e2가 e1보다 크게 되면 e2와 이에 해당하는 집합을 현재의 샷 집합으로 결정한다. 그렇지 않은 경우에는 e1, e2와 T로 Boltzman 확률(ρ)을 구하고 이를 만족하는 지 계산하게 된다. 만족하게 되면 e2와 이에 해당하는 집합을 샷의 집합으로 채택하고, 아니면 T의 값에 ρ를 곱하여 줄이게 된다. 이렇게 하여 T가 임계값 ε보다 작게 되면 현재의 샷 집합을 채택함으로 해서 전체 알고리즘이 종료하게 된다.

```
Procedure SimAbstraction(ρ, T,ε, W[])
Real ρ, T; // SA의 파라미터
Real ε;    // 종료 조건
Real W[];  // 목적함수의 가중치

Real e1,e2; // 목적함수의 결과 값
Real p      // e1, e2로부터의 Boltzman 확률

Integer shotset[]   // 뽑힌 샷 들
Integer nshotset    // 뽑힌 샷의 개수

Begin
shotset ←ChooseShotSet(&nshotset); //임의의 한 개 샷 집합선택
e1 ←ObjectFunction(shotset, nshotset, W); //샷 집합의 목적함수 값

while (T > ε) do
shotset ←ChooseShotSet(&nshotset); //임의의 한 개 샷 집합선택
e2 ←ObjectFunction(shotset, nshotset); //샷 집합의 목적함수 값

if (e2≥ e1) // 현재의 샷 집합이 더 큰 값을 가지면 이 집합을 채택
e1 ← e2;
else
ρ ←Prob(e1, e2, T); // 현재의 샷 집합이 더 작은 값을 가지면
if (Condition(p))   // (e1, e2, T)로 Boltzman 확률을 구하여
e1 ← e2;          // 확률을 만족하면 현재 샷 집합을 채택한다.
else
T←ρ* T;           // 현재 shot집합을 채택하지 않고 T를 줄인다.
endif
endif

return e1, shotset, nshotset;
end.
```

그림 8 목적 함수를 이용한 요약 알고리즘

4. 실험 결과 및 분석

4.1 Simulated Annealing 을 이용한 실험 결과

제안한 알고리즘의 유용성을 검증하기 위하여 계산 가능한 크기의 샷으로 구성된 동영상에 대하여 목적 함수를 최대로 만족시키는 값을 구하고 Simulated Annealing 을 이용한 경우와 그렇지 않은 경우에 대하여 설명하도록 한다. 실험을 위해서 사용된 동영상은 그림 9와 같이 16개의 샷으로 구성되어 있다. 이러한 동영상을 1,500 프레임으로 요약한다고 했을 때 30 프레임 이하의 샷과 전체 동영상의 80%에 해당하는 2,778 프레임 이상의 샷들은 제외하게 된다.

그림 10은 모든 샷 집합에 대한 목적 함수의 값과 이에 대한 히스토그램을 나타낸 그림이다. 그래프에서 x축은 샷의 집합을 나타내고, y축은 목적 함수에 대한 값을 나타낸다. 오른쪽 그림은 이러한 샷들의 집합이 갖는 목적 함수의 히스토그램을 정규화하여 나타낸 그림이다.

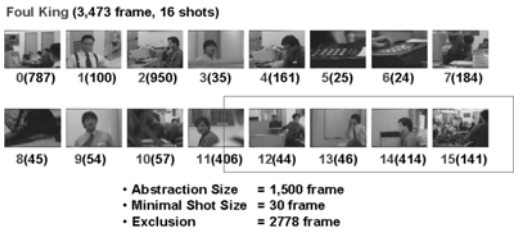


그림 9 실험에 사용된 동영상의 구성

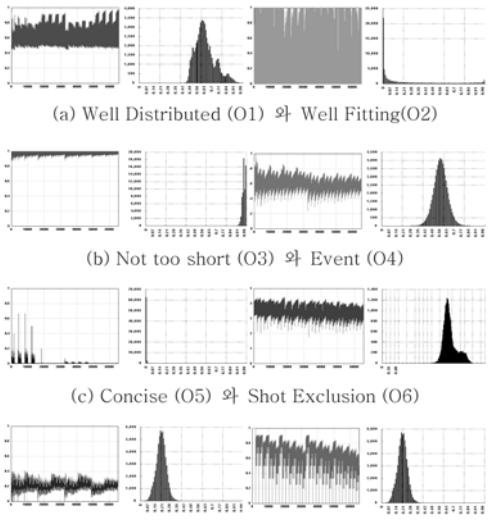


그림 10 각 샷집합에 대한 목적 함수의 값과 분포

이 그림을 통하여 볼 때 평균값 주위로 값들이 분포되어 있음을 알 수 있고 이를 통해 목적 함수의 값을 최대화하는 샷의 집합과 그렇지 않은 샷의 집합을 구분할 수 있음을 알 수 있다.

다음은 Simulated Annealing을 사용하는 경우와 그렇지 않은 경우의 차이를 알아보기 위한 실험 결과들이다. 여기서 숫자로 표현된 것은 각 샷의 번호가 되는 것이고 이는 그림 9에서 확인할 수 있다. 또한 E는 정규화 된 목적 함수값을 나타내고 있다.

- Well Distributed(O_1): 그림 11은 목적 함수 O_1에 대한 실험 결과를 나타내는 그림이다. 목적 함수 O_1을 최대로 하는 샷의 집합은 그림 11에서와 같이 모든 샷이 다 포함되는 경우이다. 그 이유는 각 샷들의 간격 편차가 0이 되기 때문이다. 그림 11의 아래쪽에 보이는 그림은 Simulated Annealing을 사용하여 최적 샷 집합을 구한 결과를 보여주고 있는데 이를 통해 ρ 값이 증가할수록 점차 많은 샷들이 포함되는 것을 알 수 있고, 그에 따라 목적 함수 O_1값이 커지는 것을 알 수 있다.

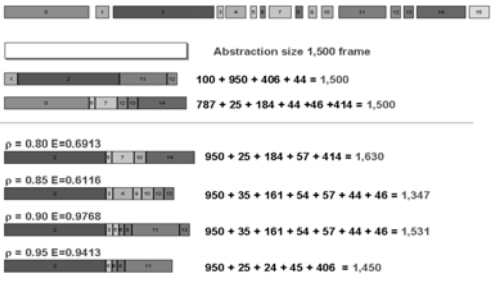


그림 11 목적 함수 Well Distributed 에 대한 결과

- Well Fitting(O_2): 그림 12는 목적 함수 O_2에 대한 실험 결과를 나타내는 그림이다. 목적 함수 O_2를 최대로 하는 샷의 집합은 그림 12에서 볼 수 있듯이 결과로 나오는 요약 파일의 크기가 1,500 프레임이 되는 경우

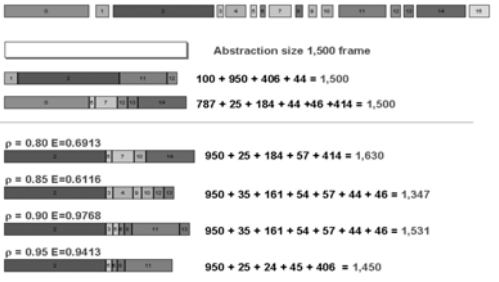


그림 12 목적 함수 Well Fitting에 대한 결과

이다. 그림 15의 아래쪽에 보이는 그림은 Simulated Annealing을 사용하여 목적 함수 O_2를 만족시키는 샷의 집합을 구한 결과를 보여주는 그림인데 이를 통해 ρ값이 증가할수록 요약 파일의 크기가 1,500프레임에 가까워짐을 알 수 있고, 그에 따라 목적 함수 O_2 값이 커지는 것을 알 수 있다.

- Event(O_4): 그림 13은 목적 함수 O_4에 대한 실험 결과를 나타내는 그림이다. 그림에서 왼쪽에 보이는 샷들은 움직임 에너지가 큰 순으로 나열되어진 샷들을 보여주는 것이다. 움직임이 큰 샷들의 집합을 계산한 결과 10번 샷만 포함된 집합이 목적 함수의 값이 가장 크게 나옴을 알 수 있었고, 그 외에 두 번째, 세번째로 움직임이 큰 샷들이 포함된 경우 목적 함수의 값이 크게 나옴을 알 수 있었다. 또한 Simulated Annealing을 사용하여 최적 샷들의 집합을 구한 결과 ρ값이 커짐에 따라 점차 움직임 에너지가 큰 샷들이 집합에 포함되는 것을 알 수 있다.

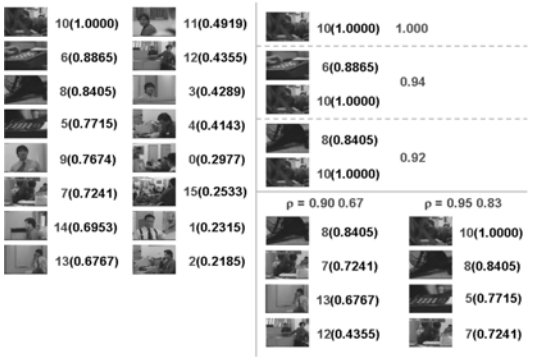


그림 13 목적 함수 Event에 대한 결과

- Concise(O_5): 그림 14는 목적 함수 O_5에 대한 실험 결과를 나타내는 그림이다. 그림에서 위쪽에 보이는 그림은 모든 샷들에 대한 히스토그램을 보여주는 그림이고 아래쪽에 보여주는 것은 최적 샷 집합과, Simulated Annealing에 의한 샷 집합을 보여주고 있다. 목적 함수 O_5를 최대화하기 위해서는 선택된 샷들이 서로 다른 히스토그램을 가져야 한다. 계산 결과 3번, 5번 샷이 선택된 경우가 목적 함수 O_5값이 가장 크게 나옴을 알 수 있었다. 또한 Simulated Annealing을 사용한 결과 다른 실험 결과와 마찬가지로 ρ값이 클수록 목적 함수 O_5의 값 또한 크게 나옴을 알 수 있었다.

위의 여러 가지 실험에서 알 수 있듯이 Simulated

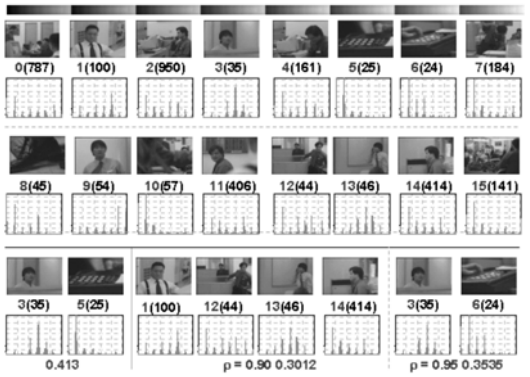


그림 14 목적 함수 Concise에 대한 결과

Annealing을 사용하는 경우에 비교적 최적값에 가까운 근사값을 찾는다는 것을 알 수 있고, 특히 ρ값이 클수록 더욱 최적값에 가깝게 근사값을 찾는다는 것을 알 수 있다.

그림 15는 본 논문에서 사용되는 목적 함수들이 종속(dependent)적인 특성을 갖는 지를 알아보기 위한 실험 결과를 나타내는 그림이다. 그림은 앞서 이용된 216개 샷의 집합을 가질 수 있는 동영상에 대한 목적 함수들의 값과 비교에 사용되는 y=x의 함수를 통해 목적 함수의 특징을 이용하여 목적 함수의 상관도를 나타내고 있다. 각각의 목적 함수들이 종속적이기 위해서는 각 값들의 분포가 양의 상관 관계를 나타내던지 음의 상관 관계를 나타내어야 하는데, 그림 15를 보면 알 수 있듯이

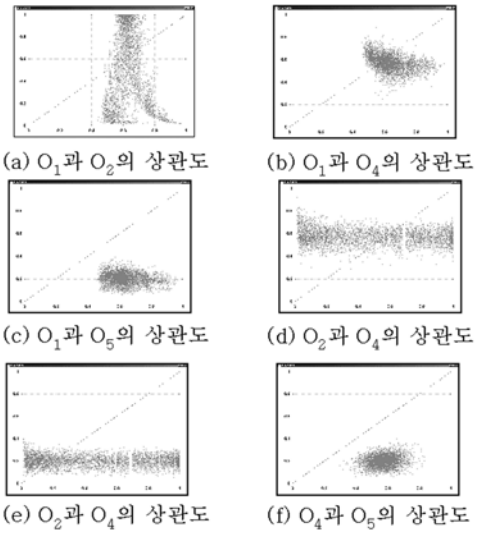


그림 15 목적 함수들의 상관도

목적 함수들의 값이 고르게 분포되어 있음을 알 수 있다. 그러므로 본 논문에서 사용되는 목적 함수들은 서로 종속적이라고 하기 어렵다는 것을 알 수 있다.

4.2 가중치 적용 방법

그림 16은 시트콤을 대상으로 Simulated Annealing을 이용하여 3,600 프레임으로 요약한 결과이다. 각 그림에서 위쪽에 위치한 그림은 요약된 비디오의 키 프레임들을 표시한 그림이고 아래 위치한 그림은 전체 동영상에서 각 샷들의 위치를 나타내는 그림이다. 그림을 통해 ρ값이 클수록 3,600 프레임에 가깝게 요약되는 것을 알 수 있고, 뽑힌 샷들의 분포도 점차 일정하게 된다는 것을 알 수 있다. 또한 중복되는 샷이 줄어드는 것도 키 프레임들을 통해 확인할 수 있다. 하지만 실제 결과 동영상을 살펴보면 산술적인 수치와는 다르게 줄거리가 제대로 요약이 되지 않았다는 것을 알 수 있다. 그 이유는 우선 동영상의 장르가 시트콤이기 때문에 여러 가지 부수적인 이야기가 있게 된다. 이러한 동영상의 특성을 반영하지 못했기 때문이고 또 다른 문제는 이야기의 중심이나 초점이 맞추어지는 지점이 장르마다 다르기 때문이다. 이러한 이유로 인해 장르에 따라 각 목적 함수의 가중치 설정을 달리해야 하는 필요성이 생기게 된다. 이를 위해 우선 이야기의 구성이 비교적 단조로운 다큐멘터리에 대하여 요약하는 방법과 이에 맞는 가중치 설정 방법에 대하여 설명하도록 한다.

그림 16 Simulated Annealing을 이용한 실험 결과

그림 17은 ρ가 0.95이고 목적 함수(1,2,4,5)의 가중치를 모두 동일하게 1.0으로 주었으며 1,800 프레임으로 요약하였을 때의 결과를 나타내는 그림이다. 결과를 보면 알 수 있듯이 요약된 동영상의 길이도 비슷하고 뽑힌 샷들의 칼라 히스토그램도 다르며 움직임이 많은 샷들을 포함하고 있어 이론적으로는 이상적인 요약 결과라고 할 수 있다. 실제로도 시트콤을 요약한 결과보다는 나은 요약임을 알 수 있었다. 하지만 이 또한 시트콤과 마찬가지로 사용자가 내용을 이해하는 데에는 부족함이 있었다. 그 이유는 모든 목적 함수의 가중치가 일정하였기 때문이다.

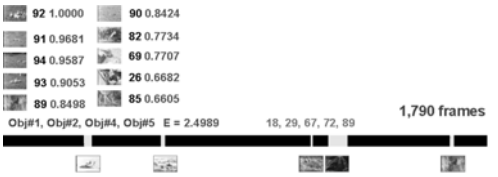


그림 17 다큐멘타리에 대한 실험 결과

그림 18은 요약자가 요약하였을 때의 결과를 입력으로 얻어낸 가중치를 보여주는 그림이다. 비록 요약 동영상이 움직임이 많은 샷을 포함하는 것도 아니고 칼라 히스토그램이 비슷한 샷도 많은 것을 알 수 있기는 하지만 내용을 아는 상태에서 요약을 한 것이기 때문에 보다 나은 요약 파일이라고 할 수 있다. 이와 같이 다른 결과를 나타내는 이유는 목적 함수의 가중치가 다르기 때문이다. 즉, O2에는 0.9999, O1에는 0.7873, O4는 0.2737, O5에는 0.1976이다. 이러한 가중치가 같은 장르의 다른 동영상에서도 비슷하게 나오는 지 확인하기 위해 다른 다큐멘터리에도 적용한 결과를 보여주는 그림이 그림 19이다. 그림을 보면 조금씩 다르기는 하지만 비슷하게 나오는 것을 알 수 있다. 이렇게 하여 장르에 따라 사용할 수 있는 가중치를 얻게 되는 것이다.

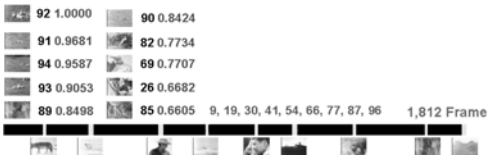


그림 18 요약자에 의해 얻은 가중치(다큐멘터리 1)

그림 19 요약자에 의해 얻은 가중치(다큐멘터리 2)

시간적인 측면에서 알고리즘의 성능을 살펴보면 모든 동영상의 특성 자체를 알고리즘이 수행되면서 바로 구하는 것이 아니라 미리 구할 수 있기 때문에 즉, 움직임 에너지(Event), 컬러 히스토그램(Concise)의 경우에는 미리 구해 놓을 수 있기 때문에 본 논문의 알고리즘은 빠른 시간에 요약을 미리 만들어 낼 수 있다.

표 1은 관련 연구와의 비교를 나타내고 있다 동영상 요약 방법은 동영상의 장르나 요약자의 주관에 따라 원하는 요약이 달라지는 특성이 있기 때문에 비교를 하는데 어려움이 있다. 하지만 기능적인 측면에서 살펴보는 경우 기존의 연구[1-4]에서는 동영상 요약의 방법만 제시하였고, 요약자의 주관에 대해서는 고려하지 않는 반면 본 논문의 알고리즘을 이용하면 요약자의 주관(Subjectivity)과 더불어 요약자가 원하는 요약 길이에 맞출 수 있다는 장점(Scalability)을 가지고 있다.

표 1 관련 연구와의 비교

	사용한 내용정보	기본 단위	Scalability	Subjectivity
Moca System[1]	Mood	shot	×	×
Lienhart[2]	shot 길이	shot	O	×
Sull[3]	히스토그램	Frame	O	×
Nam[4]	움직임 에너지	sub-shot	×	×
목적함수를 이용한 요약	위의 방법 모두	shot	O	O

5. 결 론

여러 가지 기술의 발달로 인해 동영상의 사용이 일반화되고 있다. 하지만 큰 파일 크기로 인해 여러 가지 문제점이 생겨났는데 동영상 요약 방법은 이를 해결할 수 있는 방법 중에 하나이다 기존에 여러 가지로 동영상 요약 알고리즘이 제시되어지고 있기는 하지만 하나의 저급 수준 내용 정보만을 이용하고 있어서 요약자의 주관을 반영할 수 없다는 단점이 있다 이를 해결하기 위하여 본 논문에서는 목적 함수를 이용하여 동영상을 요약하는 방법을 제시하고 있다 실험 결과 본 논문의 알고리즘을 적용하면 요약자의 주관을 반영할 수 있고다양한 크기의 요약 동영상을 생성시킬 수 있음을 알 수 있었다. 너무 큰 탐색 영역이 있어서 시간이 오래 걸린다는 단점이 있었지만 Simulated Annealing알고리즘을 통해 빠른 시간 내에 요약할 수 있음을 알 수 있었다 본 논문의 알고리즘은 디지털 비디오 라이브러리 및 편집 시스템에 유용하게 사용될 수 있을 것이다.

참 고 문 헌

[1] S. Pfeiffer, R. Lienhart, S. Fischer and Wolfgand Effelsberg, "Abstracting Digital Movies Automatically," *Journal of Visual Communication and Image*, Vol. 7, No. 4, pp. 345-353, Dec. 1996.

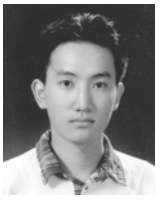
[2] R. Lienhart, "Dynamic Video Summarization of Home Video," *Proceedings of SPIE on Storage and Retrieval for Media Databases 2000*, Vol. 3972, pp. 378-389, Jan. 2000.

[3] H.S. Chang, S.H. Sull and S.U. Lee, "Efficient Video Indexing Scheme for Content-Based Retrieval," *IEEE Transaction on Circuits and Systems for Video Technology*, Vol.9, No. 8, pp. 1269-1279, Dec. 1999.

[4] J. Nam and A. H. Tewfik, "Video Abstract of Video," *Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP '99)*, pp. 117-122, Sep. 1999.

[5] J. Saarela and B. Merialdo, "Using Content Models to Builds Audio-Video Summaries," *Proceedings of SPIE on Electronic Imaging*, pp. 318-321, Jan. 1999.

[6] R.H.J.M. Ottern and L.P.P.P van Ginneken, *The Annealing Algorithm*, Kluwer Academic Publishers, 1989.



정 진 국

1998년 서강대 전자계산학과 졸업 2000년 서강대 컴퓨터학과 석사 2000년~현재 서강대 컴퓨터학과 박사 과정

홍 승 욱

1994년 서강대 수학과 졸업 1996년 서강대 컴퓨터학과 석사 2001년 서강대 컴퓨터학과 박사. 2002년~현재 Nokia SW Design Engineer

낭 종 호

1986년 서강대 전자계산학과 졸업 1988년 한국과학기술원 석사. 1992년 한국과학기술원 박사. 1992년~1993년 Fujitsu 연구소 연구원. 1993년~현재 서강대학교 컴퓨터학과 교수. 현재 서강대학교 컴퓨터학과 부교수

하 명 환

1993년 2월 경북대학교 공과대학 전자공학과 졸업(학사). 1995년 2월 한국과학기술원 전기 및 전자공학과 졸업(석사) 1995년 2월~현재 KBS 기술연구소 연구원. 관심분야는 멀티미디어 방송 제작 시스템, 비디오 인덱싱, 영상 처리

정 병 희

1994년 2월 이화여자대학교 전자계산학과 졸업(학사). 1996년 2월 한국과학기술원 전산학과 졸업(석사). 1996년 1월~현재 KBS 기술연구소 연구원. 2000년 9월~현재 한국과학기술원 전산학과 박사과정 재학중. 관심분야는 멀티미디어 방송 시스템, 미디어 아카이브, 초고속 네트워크 시스템

김 경 수

1983년 2월 서울대학교 공과대학 제어계측공학과 졸업(학사). 1985년 2월 서울대학원 제어계측공학과 졸업(석사). 1985년 3월~현재 KBS 기술연구소 차장. 관심분야는 멀티미디어 방송 제작 시스템, 스트리밍 미디어 및 서비스, 콘텐츠 보호 및 관리 기술