論文2002-39SP-4-4

반복 과정을 통한 율-제한 주요 화면 선택 기법

(Rate-Constrained Key Frame Selection Method using Iteration)

李 熏 哲 *, 金 聖 大 **

(Hun Cheol Lee and Seong Dae Kim)

요 약

주요 화면은 보다 적은 양의 데이터를 사용해서 비디오가 갖는 시각적 내용물의 변화량을 효과적으로 표현하기 위해 많이 사용된다. 이와 같은 비디오 표현 방식은 대역폭이나 저장 용량이 제한된 상황에 적합하다. 이 경우 대역폭이나 저장 용량에 따라 주요 화면의 개수를 조절하는 능력은 주요 화면 선택 기법의 중요한 필요 사항 중 하나다. 본 논문에서는 주요 화면의 개수가 제한 조건일 때 순차적인 주요 화면을 찾는 방법을 제안한다. 제안하는 기법은 먼저 원하는 개수의 초기 주요 화면을 미리 선택하고 이들이 대표하는 서로 중복되지 않는 시구간을 정한 후 반복 과정을 통해 주요 화면의 위치와 시구간의 크기를 조절하면서 왜곡값이 최소가 되도록 주요 화면과 시구간을 찾는다. 실험 결과 제안하는 방법이 선택하는 주요 화면들은 율-왜곡 관점에서 기존의 방법보다 우수하고 인간의 시각 인지와도 일치함을 알 수 있었다.

Abstract

Video representation through representative frames (*key frames*) has been addressed frequently as an efficient way of preserving the whole temporal information of sequence with a considerably smaller amount of data. Such compact video representation is suitable for the purpose of video browsing in limited storage or transmission bandwidth environments. In a case like this, the controllability of the total key frame number (i.e. key frame rate) depending on the storage or bandwidth capacity is an important requirement of a key frame selection method. In this paper, we present a sequential key frame selection method when the number of key frames is given as a constraint. It first selects the desired number of initial key frames and determines non-overlapping initial time intervals that are represented by each key frame. Then, it adjusts the positions of key frames and time intervals by iteration, which minimizes the distortion. Experimental result demonstrates the improved performance of our algorithm over the existing approaches.

Key words : 비디오 검색, 주요 화면 선택, 비디오 요약

* 學生會員, 韓國科學技術院 電子電算學科 電氣및電子工學
(Division of Electrical Engineering, Dept. of Electrical Engineering and Computer Science, KAIST)
** 正會員, 韓國科學技術院 電子電算學科 電氣및電子工學
(Division of Electrical Engineering, Dept. of Electrical Engineering and Computer Science, KAIST)
接受日字:2002年1月16日, 수정완료일:2002年6月12日

I. Introduction

최근 들어 멀티미디어 데이터를 효과적으로 표현하고 검색하는 기법에 관한 연구와 표준화 작업이 많이 진행되고 있다.[1~3] 특히 비디오 데이터 베이스에 접근하는 방법을 다루는 연구가 가장 널리 주목을 받고 있으며 여기에는 비디오 데이터의 분석(analysis), 표현

(representation), 훑어보기(browsing) 및 검색(retrieval) 등이 있다.[1,3] 주요 화면이란 비디오의 시각적 내용물을 몇 개의 화면으로 대표하는 화면을 의미하며, 비디오의 내용물을 간략하게 표현하거나[9,10] 시각적 특징을 이용해서 비디오 검색을 하는 분야에 사용할 수 있다[7,8]

일반적인 비디오 검색 엔진들은 비디오에서 샷 경계를 찾은 후[1,4~6] 샷 안에서 선택한 주요 화면들을 사용해서 비디오 검색을 위한 인덱스를 만들거나 샷들 사이의 유사도를 정의한다.[7,8] 그리고 나서 이를 이용해서 비디오를 계층적으로 혹은 장면 전환 그래프(scene transition graph) 등으로 표현한다.[9,10] 이런 비디오 표현 기법을 이용해서 사용자는 관심이 있는 비디오 세그먼트를 찾기 위해 비디오 훑어보기나 탐색을 할 수 있다. 이와 같이, 보다 적은 양의 데이터를 사용해서 효과적으로 비디오가 갖는 시각적 내용물의 변화량을 나타낼 필요가 있을 때 주요 화면을 쓴다. 특별히 주요 화면을 사용하는 비디오 표현 방식은 대역폭이나 저장 용량이 제한된 상황에 적합하다. 예를 들어 휴대폰으로 동영상 서비스를 제공하는 경우, 전송 대역폭이나 단말기의 저장 용량 등에 따라 주요 화면의 개수를 달리하면서 여러 종류의 요약 시퀀스를 만들어서 사용자가 전체 동영상을 서비스 받을지 여부를 결정할 수 있게 할 수 있다. 이 경우 대역폭이나 저장 용량에 따라 주요 화면 개수를 조절하는 능력은 주요 화면 선택 기법의 중요한 필요 사항 중 하나다. 또한, 계층적 비디오 훑어보기 등에서도 계층에 따라 주요 화면의 개수를 달리 하면서 비디오의 내용물을 요약할 수 있다.

주요 화면 선택 기법들은 비용 함수 정의 방법에 따라 *클러스터 기반 방법*[13~15]과 *순차적인 방법*[16~18]으로 분류할 수 있다. 클러스터 기반 방법은 샷이 포함하는 모든 화면들을 한꺼번에 사용해서 비용 함수를 정의하는 방법으로, 각각의 화면들을 n-차원 특징 벡터로 표현하고 n-차원의 공간에 표시한 후 클러스터링을 수행한다. 그리고 나서 각 클러스터의 중심 화면을 주요 화면으로 선택한다. 이 방법에는 시간 개념이 없다. 즉 화면의 등장 순서에 상관없이 화면의 내용에 따라 주요 화면을 선택하는 방법이다. 따라서 이 방법은 비디오의 내용물을 전체적으로 이해하기에는 도움이 되지만 시간에 따라 내용물이 변하는 비디오의 특성을 나타내지 못한다는 단점이 있다.

한편 순차적인 방법은 시간 개념을 사용해서 주요 화면을 선택한다. 이 방법은 화면들의 등장 순서에 따라 순차적으로 주요 화면을 찾는 방법으로, 이전 주요 화면의 위치를 참고하여 현재 주요 화면의 위치를 결정하는 방식이다. 이 방법은 주요 화면은 서로 중복하지 않는 임의의 시구간 안에 있는 모든 화면을 대표하는 화면이라고 간주한다. 그리고 이를 이용해서 국부적으로 각 시구간에서 왜곡을 계산한 후 이 왜곡들을 더해서 전역적인 비용 함수를 정의하고 이 비용 함수를 최소로 만드는 주요 화면을 찾는다. 이 방법을 사용하면 시간에 따른 내용물의 변화량이 큰 부분에서는 조밀하게 주요 화면을 선택할 수 있고 그렇지 않으면 듬성듬성하게 주요 화면을 선택을 할 수 있다. 순차적인 방법에 의해서 선택된 주요 화면은 시간에 따른 비디오의 내용물의 변하는 특성을 잘 나타낼 수 있다는 장점이 있다. 따라서 샷들 사이의 유사도를 정의하기 위해서는 클러스터 기반으로 선택된 주요 화면보다 순차적 방법으로 선택된 주요 화면을 사용하는 것이 더 타당하다.

본 논문에서는 주요 화면 개수가 제한 조건일 때 순차적인 방법을 사용해서 주요 화면들을 찾는 방법을 제안한다. 제안하는 방법은 각 시구간에서 발생하는 국부 왜곡들을 더해서 전역 왜곡을 정의한다. 그리고 주요 화면 개수를 일정하게 유지시킨 채 전역 왜곡을 최소화하는 주요 화면과 주요 화면이 대표하는 시구간을 찾는 것을 목적으로 한다. 이 방법은 시간 축 상에서 행하는 비균일 샘플링과 비슷한 개념이다. 이를 위해 먼저 원하는 개수의 초기 주요 화면을 선택하고 이들이 대표하는 시구간을 결정한 후 반복 과정을 통해 주요 화면의 위치와 시구간의 크기를 조정하면서 왜곡값이 최소가 되도록 주요 화면과 시구간을 찾는다. 이런 반복 과정을 통해 현재의 왜곡은 이전의 왜곡보다 항상 같거나 적음이 보장된다. 이런 방식으로 찾은 원하는 개수의 주요 화면들은 샷들 사이의 유사도를 정의하거나 사용자의 요구에 따라 계층적으로 비디오 내용물을 표현하는데 사용할 수 있다.

본 논문의 구성은 다음과 같다. 2절에서는 기존의 주요 화면 선택 기법을 설명한다. 그리고 3절에서는 제안하는 기법에 관해서 설명한다. 제안하는 기법에는 크게 초기 주요 화면을 선택하는 과정과 반복 과정에 의해 주요 화면의 위치를 조정하는 과정이 있다. 그리고 나서 4절에서는 실험 결과에 관해 설명하고 5절에서는 결론 및 향후 과제를 말함으로써 논문을 맺고자 한다.

Ⅱ. Review of Key Frame Selection Methods

주요 화면을 선택하는 분야에는 여러 종류의 기법들이 있다. 가장 간단한 방법은 샷의 첫 번째 혹은 마지막 화면을 주요 화면으로 선택하거나[11] 일정한 시간 간격을 두고 화면들을 주요화면으로 선택하는 것이다.[12] 하지만 이 방법들은 비디오의 내용물을 전혀 고려하지 않기 때문에 빠른 카메라의 움직임이 발생하거나 내용물의 변화가 심한 경우에는 효과적으로 샷을 표현할 수 없다. 이런 단점을 보완하기 위해 다음과 같이 비디오의 내용물을 고려한 주요 화면 선택 방법들이 제안되었다.

우선 클러스터 기반 방법에는 다음과 같은 것들이 있다. Y. Zhuang *et. al*[13]는 화면을 나타내는 특징 벡터로서 컬러 히스토그램을 사용하였으며 클러스터 탐색 알고리즘을 이용해서 전체 화면들을 몇 개의 클러스터로 나누고 크기가 일정한 값 이상이 되는 클러스터의 중심 화면을 주요 화면으로 선택하는 방법을 제안하였다. Doulamis *et. al*[14]은 컬러와 움직임을 사용해서 화면의 특징 벡터를 정의하고 k-평균 클러스터링과 로그 탐색을 사용해서 주요 씬(key scene)과 주요 화면을 찾는 방법을 제안하였다. Chang *et. al*[15]은 컬러 히스토그램을 이용해서 화면들을 특징 벡터들의 집합으로 표시한 후 준-하우스도르프 거리(semi-Hausdorff distance)를 이용해서 비용 함수를 정의하고 왜곡을 주어진 문턱 값 이하로 유지하면서 가장 적은 개수의 주요 화면을 선택하는 방법을 제안하였다.

순차적 방법으로 주요 화면을 선택하는 기법들은 다음과 같은 것들이 있다. Yeung과 Liu[16]는 샷의 첫 번째 화면을 최초의 주요 화면으로 초기화한 후 뒤따라오는 현재 화면과 이전 주요 화면 사이의 비유사도가 주어진 문턱 값보다 더 크면 현재 화면을 주요 화면으로 선택하는 방법을 제안하였다. 이 방법은 순차적 주요 화면은 시간적으로 인접한 화면들을 대표한다는 성질은 잘 이용했지만 정량적이지 못하고 휴리스틱한 면이 있다. 그리고 원하는 개수의 주요 화면을 선택하기 위해 문턱 값을 바꾸면서 여러 번 시행착오를 반복해야 한다는 단점이 있다. 반면 Hanjalic et. al[17,18]은 N 개의 주요 화면을 선택하고자 할 때, 비감소하는 성질을 띤 누적된 움직임 활성도(accumulated motion activity)를 사용해서 주요 화면과 연속하는 두 주요 화면 사이에 있는 분절점(break-point) 사이의 재귀적인 관계를 아래와 같이 정량적으로 유도하였다.

$$k_i = \frac{1}{2}(t_{i-1} + t_i), \quad i = 1,2,\cdots,N \tag{1}$$

$$f(t_i) = \frac{1}{2}(f(k_i) + f(k_{i+1})), \quad i = 1,2,\cdots,N-1 \tag{2}$$

여기에서 $f(t)$는 시각 t에 위치한 화면에 대한 화면 기술자이며 비감소하는 성질을 갖는다. 그리고 k_i는 i 번째 주요 화면의 위치, t_i는 k_i와 k_{i+1} 사이에 있는 분절점의 위치를 의미하며 t_0는 샷의 처음 화면이 시간 축 상에 놓여진 위치를 의미한다. 이와 같이 정의한 후 Hanjalic은 최초의 분절점(t_1)의 위치를 초기 값으로 정하고 재귀적으로 다음 주요 화면의 위치와 분절점의 위치를 찾는 방법을 제안하였다. 이 방법의 저자들은 마지막 분절점(t_N)의 위치가 샷의 마지막 화면이 되도록 최초의 분절점의 위치를 조정하면 원하는 개수의 주요 화면을 얻을 수 있다고 강조하고 있다. 이 방법 또한, Yeung과 Liu의 방법과 마찬가지로, 원하는 개수의 주요 화면을 선택하기 위해서는 최초의 분절점의 위치를 여러 번 변화시키는 시행 착오의 과정이 필요하다.

이런 단점을 극복하기위해, 본 논문의 저자들은, 문턱 값이나 초기 분절점의 위치를 조정할 필요가 없는 율-제한 주요 화면 선택 기법을 제안하였다 [19]. 여기에서는 이전의 주요 화면 위치가 현재 주요 화면의 위치에 영향을 주지 않는다는 가정 하에, 샷의 내용물의 변화량이 각각의 주요 화면에 동일하게 분배되도록 주요 화면과 분절점의 위치를 계산하였다. 이 방법은 주요 화면의 위치는 각 분절점 사이에서는 국부적으로 최적이지만, 분절점들은 주요 화면 사이에서 국부적으로 최적이 되지 않을 수도 있다.

Ⅲ. Iterative Key Frame Selection Method

[19]에서 제안한 방법을 개선하기 위해, 본 논문에서는 반복 과정을 통해 왜곡을 될 수 있는 한 최소로 만들면서 주요 화면과 분절점을 동시에 국부적으로 최적이 되도록 하는 주요 화면 선택 방법을 제안한다. 이 절에서는 먼저 왜곡을 정량적으로 정의하고 반복 과정

을 통해 왜곡을 최소로 만드는 주요 화면과 분절점을 선택하는 방법을 설명한 후, 반복 과정에 의해 왜곡이 점점 수렴함을 정량적으로 증명한다.

1. Distortion Measures of Selected Key Frames

집합 $\mathbf{S}=\{s_1,s_2,\cdots,s_M\}$를 M개의 화면들의 집합이라 하고 $\mathbf{K}=\{k_1,k_2,\cdots,k_N\}$를 그 샷에서 선택된 N개의 주요 화면들의 집합이라 하자. 여기에서 s_i는 샷의 i번째 화면을 나타내고 k_j는 j번째 주요 화면을 의미하며, $\mathbf{K}\subset\mathbf{S}$, $N<M$의 관계를 만족해야 한다. 그리고 $\mathbf{T}=\{[t_{i-1},t_i]\,|\,i=1,\cdots,N\}$를 각각의 주요 화면이 대표하는 서로 중복하지 않는 N개의 시구간(time interval)들의 집합이라 하자. 즉 $t_{i-1}\le k_i\le t_i$의 관계가 성립해야 하며 이는 시구간 $[t_{i-1},t_i]$에 포함된 모든 화면들은 시각 k_i에 있는 화면으로 대표할 수 있다는 것을 의미한다. 이런 방식으로 M개의 화면을 가지는 샷을 N개의 주요 화면으로 표현하면 그림 1과 같은 왜곡이 발생한다. 이 왜곡을 정량적으로 정의하기 위해 $f(t)$를 시간 축 상의 위치 t에 놓여 있는 화면을 기술하는 화면 기술자, $d(t_i,t_j)$를 시각 t_i와 시각 t_j에 위치한 화면들 사이의 거리 값이라 두자. 이 거리 값은 $f(t)$를 사용해서 계산할 수 있으며 응용 분야에 따라 여러 형태를 가질 수 있다. 그리고 기호 표시와 수식 유도의 편의를 위해 앞으로는 t를 연속 변수라 가정하고 설명을 한다. 그러면 주요 화면 선택 문제는 그림 1과 같은 왜곡을 최소로 만드는 집합 **K**와 집합 **T**를 찾는 문제로 표현할 수 있다. 본 논문에서는 두 가지 관점에서 왜곡을 정의한다. 이 두 종류의 왜곡은 본질적으로는 서로 동일한 의미를 가지며, 단지 논리 전개의 편의를 위해 기호를 다르게 사용했을 따름이다.

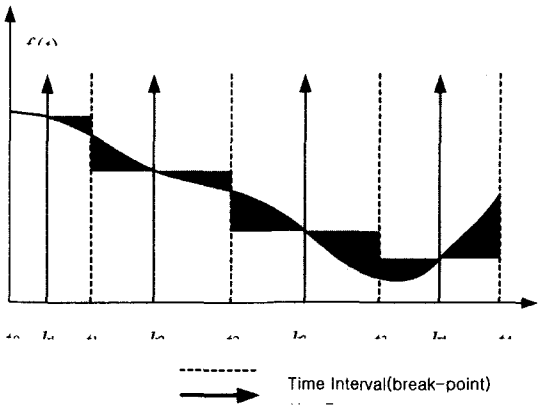


그림 1. 왜곡을 정의하는 방법($N=4$인 경우)
Fig. 1. Definition of distortion (in case of $N=4$).

먼저 첫째 정의에서는 식 (3)과 같이 전체 왜곡을 서로 중복하지 않는 시구간에서 발생하는 국부 왜곡들의 합으로 생각한다. 이 식은 시구간이 미리 주어져 있을 때, 왜곡을 최소로 만드는 주요 화면을 찾는데 유용하게 사용될 수 있다. 여기에서 t_0와 t_N은 각각 샷의 처음 화면과 마지막 화면이 시간 축 상에 놓여진 위치를 의미한다.

$$D_1(\mathbf{K},\mathbf{T})=\sum_{i=1}^{N}\int_{t_{i-1}}^{t_i} d(t,k_i)dt \tag{3}$$

그리고 각각의 시구간에서 국부 왜곡을 최소로 만드는 주요 화면은 다음과 같이 구할 수 있다.

$$k_i=\arg\min_{t_{i-1}\le k_i\le t_i}\int_{t_{i-1}}^{t_i} d(t,k_i)dt,\ i=1,2,\cdots,N \tag{4}$$

둘째 정의에서는 식 (5)와 같이 전체 왜곡을 서로 이웃하는 두 개의 주요 화면 사이에서 발생하는 국부 왜곡들의 합으로 생각한다. 이 식은 주요 화면이 주어져 있을 때 왜곡을 최소로 만드는 시구간을 결정하는데 사용될 수 있다. 여기에서 $k_0=t_0$, $k_{N+1}=t_N$을 의미한다.

$$D_2(\mathbf{K},\mathbf{T})=\sum_{i=0}^{N}\left[\int_{k_i}^{t_i} d(t,k_i)dt+\int_{t_i}^{k_{i+1}} d(t,k_{i+1})dt\right] \tag{5}$$

그리고 두 개의 이웃하는 주요 화면 사이의 국부 왜곡을 최소로 만드는 분절점(break-point)은 다음과 같이 찾을 수 있다.

$$t_i=\arg\min_{k_i\le t_i\le k_{i+1}}\left[\int_{k_i}^{t_i} d(t,k_i)dt+\int_{t_i}^{k_{i+1}} d(t,k_{i+1})dt,\ i=1,2,\cdots,N-1\right] \tag{6}$$

2. Optimal Approaches

위와 같이 왜곡을 정의하는 경우 원하는 개수의 주요 화면을 찾는 최적의 방법에는 그림 2와 같이 두 가지 방법이 있다. 그림 2(a)는 첫째 방법을 설명하고 있다. 우선 원하는 개수의 주요 화면을 임의로 선택한다. 그리고 식 (6)을 이용해서 각각의 주요 화면이 대표하는 서로 중복되지 않는 시구간을 구한다. 그러면 임의의 주요 화면 집합에 의해 발생하는 왜곡을 계산할 수 있다. 만약 샷 안에 있는 화면들의 개수가 M개이고 원하는 주요 화면의 개수가 N개라면 ${}_MC_N$개의 경우가

존재한다. 이런 모든 경우에 해당하는 주요 화면 집합들 가운데 왜곡을 최소로 만드는 집합이 그 샷을 대표하는 최적의 주요 화면들이다. 한편 최적의 주요 화면을 찾는 둘째 방법은 그림 2(b)에서 설명하고 있다. 먼저 서로 중복하지 않는 시구간을 선택하고 식 (5)를 이용해서 각 시구간을 대표하는 주요 화면을 결정한 후 왜곡을 계산한다. 이 과정을 모든 가능한 시구간 집합에서 되풀이한 후 왜곡이 최소가 되는 시구간에 속하는 주요 화면을 찾으면 된다. 이 두 가지 방법으로 찾은 주요 화면은 서로 동일하다.

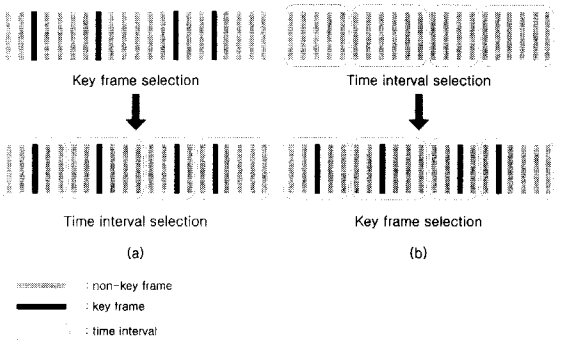


그림 2. 최적의 순차적 주요 화면 선택 기법
Fig. 2. An optimal method of sequential key frame selection.

3. Iterative Key Frame Selection

최적의 순차적 주요 화면 선택 방법은 샷의 길이가 길어질수록 왜곡을 계산해야 할 주요 화면(시구간) 집합의 경우의 수가 많아져서 계산량이 기하급수적으로 증가한다는 단점이 있다. 그리고 정량적인 해가 존재하지 않고 모든 경우의 왜곡을 서로 비교해야 한다는 단점이 있다. 이런 문제를 해결하기 위해 이 절에서는 반복 과정을 통해 주요 화면을 선택하는 방법을 설명한다.

먼저 식 (4)와 식 (6)에서, 왜곡을 최소로 만드는 주요 화면의 집합 **K**와 시구간의 집합 **T**가 서로 의존적인 관계에 있음을 알 수 있다. 즉 하나가 미리 정해지면 다른 하나는 자동으로 결정된다. 예를 들어 **T**가 미리 정해진 경우, 식 (4)를 사용해서 각각의 시구간에서 국부 왜곡이 최소가 되는 주요 화면의 위치를 계산하면 식 (3)을 최소로 만드는 **K**를 쉽게 얻을 수 있다. 마찬가지로 **K**가 미리 정해진 경우에는 식 (6)을 사용해서 연속하는 두 주요 화면 사이의 국부 왜곡이 최소가 되는 분절점들을 계산하면 식 (5)를 최소로 만드는 시구간의 집합 **T**를 찾을 수 있다.

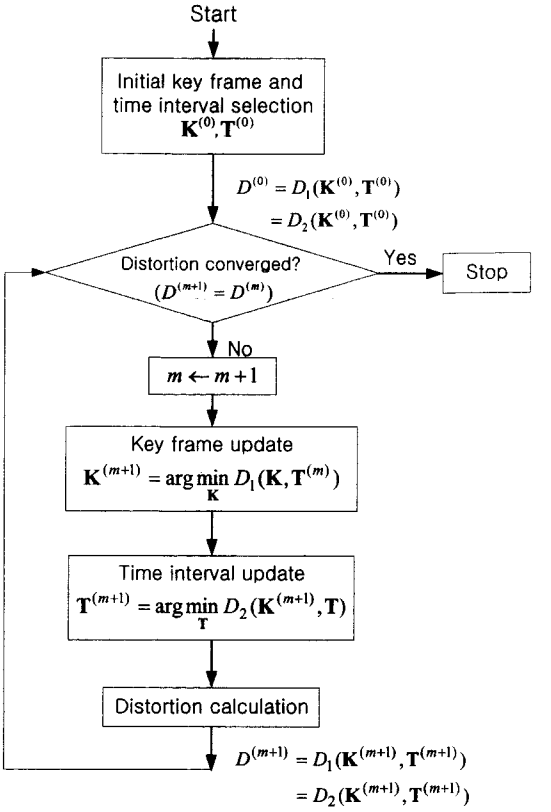


그림 3. 제안하는 기법의 흐름도
Fig. 3. Overview of the proposed method.

임의의 **K**와 **T**가 있을 때, D_1과 D_2의 값은 서로 동일하다. 하지만 D_1의 관점에서 최소가 된다고 해서 반드시 D_2의 관점에서 최소가 되는 것은 아니다. 즉 주요 화면이 분절점 사이에서 국부적으로 최적이 되었다고 해서 반드시 분절점들도 주요 화면 사이에서 국부적으로 최적이 되는 것은 아니다. 반대의 경우도 마찬가지다. 이런 문제를 해결하기 위해 본 논문에서는 그림 3과 같이 반복 과정을 진행함에 따라 주요 화면의 위치와 시구간의 위치가 조정되도록 하였다. 이런 과정을 반복하면 전체 왜곡은 점점 줄어듦을 알 수 있으며 전체 왜곡이 수렴할 때까지 주요 화면과 분절점의 위치를 반복해서 조정하면 식 (3)과 식 (5)의 관점에서 동시에 왜곡이 최소가 되는 주요 화면과 시구간을 찾을 수 있다.

이제 반복 과정을 진행함에 따라 현재 왜곡은 항상 이전 왜곡과 같거나 더 적게 됨을 보이고자 한다. 이를 위해 m번째 반복 과정 후 발생하는 왜곡을 $D^{(m)}$, 주

요 화면의 집합을 $\mathbf{K}^{(m)}=\{k_1^{(m)},k_2^{(m)},\cdots,k_N^{(m)}\}$, 시구간의 집합을 $\mathbf{T}^{(m)}=\{[t_0^{(m)},t_1^{(m)}],[t_1^{(m)},t_2^{(m)}],\cdots,[t_{N-1}^{(m)},t_N^{(m)}]\}$ 이라 두자. 즉 $D^{(m)}=D_1(\mathbf{K}^{(m)},\mathbf{T}^{(m)})=D_2(\mathbf{K}^{(m)},\mathbf{T}^{(m)})$ 의 관계가 성립한다. 먼저 시구간이 고정된 경우, 각각의 시구간에서 식 (4)를 사용해서 주요 화면의 위치를 조정한다. 이것은 식 (3)의 관점에서 왜곡을 최소로 만든다. 따라서 이렇게 위치가 바뀐 주요 화면에 의해 임의의 시구간에서는 항상 다음의 부등식이 성립함을 알 수 있다.

$$\int_{t_{i-1}^{(m)}}^{t_i^{(m)}} d(t,k_i^{(m+1)})dt \le \int_{t_{i-1}^{(m)}}^{t_i^{(m)}} d(t,k_i^{(m)})dt,\ i=1,2,\cdots,N \tag{7}$$

즉 다음의 부등식이 항상 성립한다.

$$D_1(\mathbf{K}^{(m+1)},\mathbf{T}^{(m)}) \le D_1(\mathbf{K}^{(m)},\mathbf{T}^{(m)}) \tag{8}$$

주요 화면의 위치가 이와 같이 조정된 경우 이제 식 (5)를 최소로 만드는 관점에서 식 (6)을 사용해서 시구간의 위치를 조정하면 아래의 식이 항상 성립함을 알 수 있다.

$$\begin{aligned}&\int_{k_i^{(m+1)}}^{t_i^{(m+1)}} d(t,k_i^{(m+1)})dt+\int_{t_i^{(m+1)}}^{k_{i+1}^{(m+1)}} d(t,k_{i+1}^{(m+1)})dt\\ &\le \int_{k_i^{(m+1)}}^{t_i^{(m)}} d(t,k_i^{(m+1)})dt+\int_{t_i^{(m)}}^{k_{i+1}^{(m+1)}} d(t,k_{i+1}^{(m+1)})dt,\quad i=1,2,\cdots,N-1\end{aligned} \tag{9}$$

즉 아래의 부등식이 항상 성립한다.

$$D_2(\mathbf{K}^{(m+1)},\mathbf{T}^{(m+1)}) \le D_2(\mathbf{K}^{(m+1)},\mathbf{T}^{(m)}) \tag{10}$$

한편 $D_1(\mathbf{K}^{(m+1)},\mathbf{T}^{(m)})=D_2(\mathbf{K}^{(m+1)},\mathbf{T}^{(m)})$ 의 관계에 있다. 이 사실과 식 (8), 식 (10)을 이용하면 아래의 부등식이 항상 성립함을 알 수 있다.

$$D_2(\mathbf{K}^{(m+1)},\mathbf{T}^{(m+1)}) \le D_1(\mathbf{K}^{(m)},\mathbf{T}^{(m)}) \tag{11}$$

그리고 $D^{(m)}=D_1(\mathbf{K}^{(m)},\mathbf{T}^{(m)})$, $D^{(m+1)}=D_2(\mathbf{K}^{(m+1)},\mathbf{T}^{(m+1)})$ 의 관계가 성립하므로 아래와 같이 반복 과정을 진행함에 따라 왜곡이 항상 줄어듦이 보장된다.

$$D^{(m+1)} \le D^{(m)} \tag{12}$$

4. Selection of Initial Key Frames and Time Intervals

본 절에서는 초기 주요 화면과 초기 시구간을 결정하는 방법을 설명한다. 순차적 주요 화면 선택 방법은 시간 축 상에서 행하는 비균일 샘플링과 유사한 개념이다. 식 (3)과 식 (5)에서 정의한 왜곡을 최소로 만들기 위해서는 변화량이 큰 부분에서는 조밀하게 주요 화면을 선택하고 변화량이 적은 부분에서는 듬성듬성하게 선택해야 한다. 이런 성질을 고려해서 본 논문에서는 비디오의 내용물의 변화량을 고려해서 초기 시구간과 초기 주요 화면의 위치를 정한다. 먼저 비디오 샷의 전체의 내용물의 변화량을 다음과 같이 정의한다.

$$\Delta_V \equiv \int_0^M \left|\frac{df}{dt}\right| dt \tag{13}$$

그리고 나서 서로 동일한 내용물의 변화량을 갖도록 초기 시구간의 위치를 아래와 같이 정한다.

$$\int_{t_0^{(0)}}^{t_1^{(0)}} \left|\frac{df}{dt}\right| dt = \int_{t_1^{(0)}}^{t_2^{(0)}} \left|\frac{df}{dt}\right| dt = \cdots = \int_{t_{N-1}^{(0)}}^{t_N^{(0)}} \left|\frac{df}{dt}\right| dt = \frac{\Delta_V}{N} \tag{14}$$

그리고 각각의 시구간에서 식 (4)를 만족하는 주요 화면들을 초기 주요 화면으로 선택한다. 그림 4는 초기 주요 화면과 초기 시구간을 선택하는 개념을 보여주고 있다. 이에 대한 자세한 설명은 [19]에 잘 나와 있다.

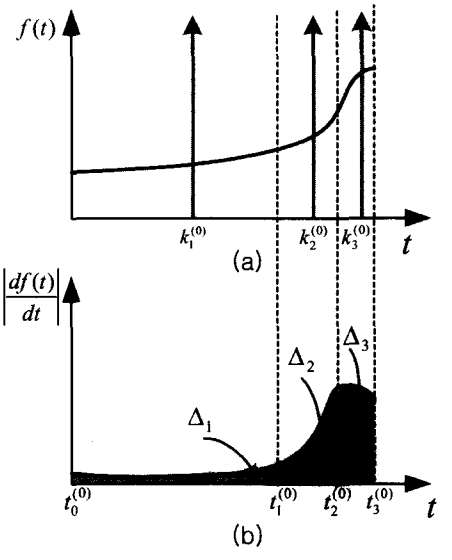


그림 4. 초기 시구간과 초기 주요 화면의 선택 방법 ($\Delta_1=\Delta_2=\Delta_3$)
(a) 초기 주요 화면 (b) 초기 시구간 선택

Fig. 4. Selection of initial time intervals and initial key frames ($\Delta_1=\Delta_2=\Delta_3$).
(a) initial key frames (b) initial time intervals.

Ⅳ. Experimental Results

일반적으로 많이 사용하고 있는 MPEG 영상 열(sequence)인 Foreman 영상 열(0-398)을 사용해서 주요 화면을 선택하는 실험을 수행하였다. Foreman 영상 열은 한 대의 카메라로 단절 없이 촬영한 하나의 샷이며 영상 열 중간에 내용물이 크게 변하는 부분이 있기 때문에 주요 화면 선택의 성능을 평가하기에 좋은 영상 열이다. 이 영상 열은 한 남자가 건물 앞에서 고개를 크게 움직이며 이야기하는 장면이 계속되다가 300번째 화면 부근에서 급격한 카메라 패닝(panning)이 일어난 후 공사장 장면을 잠시 보여주면서 끝난다.

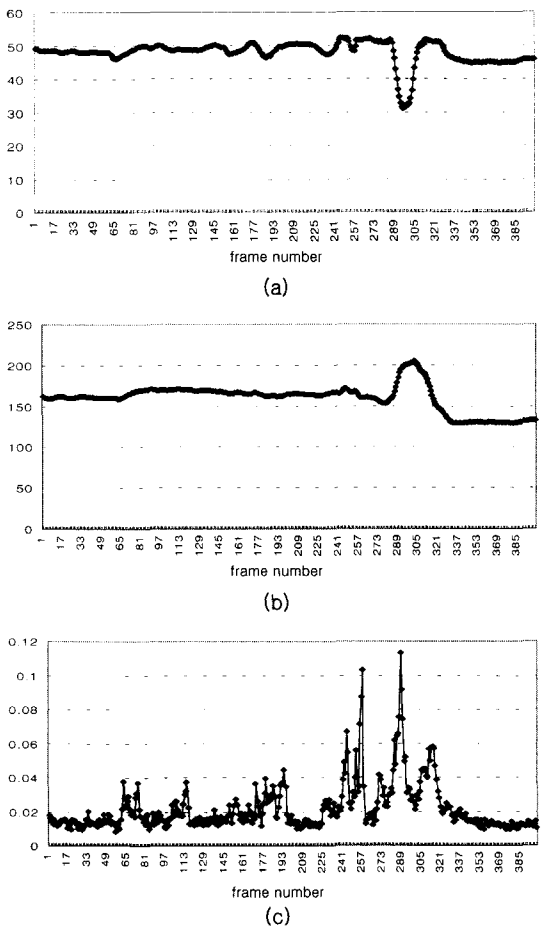


그림 5. Foreman sequence의 화면 기술자 (a) 분산 (b) 평균 밝기 값 (c) 히스토그램 차이

Fig. 5. Frame descriptors of *Foreman* sequence (a) Variance of intensity (b) Average of intensity (c) Frame-by-frame histogram difference.

그림 5는 Foreman 영상 열에서 계산한 여러 종류의 화면 기술자들 가운데 밝기 값을 이용한 화면 기술자의 예를 보여 주고 있다. 그림 5(a)와 그림 5(b)는 각 화면의 밝기 값의 분산과 평균값을 의미하며 그림5(c)는 이웃하는 화면들 사이의 히스토그램 차이를 의미한다. 300번째 화면 부근에서는 카메라 패닝이 갑자기 발생하기 때문에 화면 기술자의 급격한 변화가 일어남을 알 수 있다. 이 그림에서 알 수 있듯이 분산과 평균 밝기는 히스토그램의 차이가 큰 부분에서 크게 변한다. 따라서 밝기를 고려한 화면 기술자를 생각할 때, 밝기의 평균과 분산은 내용물의 변화량을 잘 나타낸다고 할 수 있다. 물론, 앞에서도 언급했듯이, 어떤 종류의 화면 기술자를 사용하느냐에 따라 선택되는 주요 화면이 달라지게 된다. 하지만 제안하는 방법은 주어진 화면 기술자에 대해 왜곡을 최소로 만드는 관점에서 접근하기 때문에, 본 절에서는 동일한 화면 기술자에 대해 기존의 기법과 제안하는 기법의 성능을 서로 비교하였다.

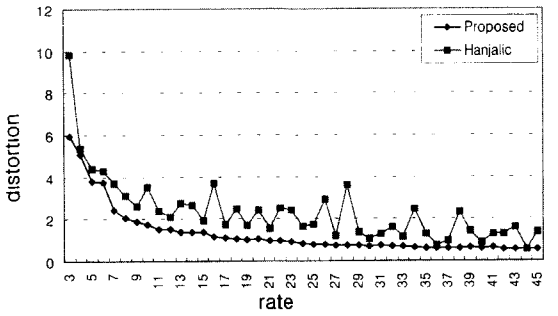
1. Performance comparison

제안하는 방법을 기존의 순차적 주요 화면 선택 방법들 중 정량적인 방법인 Hanjalic의 방법과 비교하였다. Hanjalic의 방법과 공정하게 비교하기 위해, L_1-norm을 사용해서 두 화면 사이의 거리 $d(t_i,t_j)$를 계산하였으며, 식 (15)와 같이 누적된 화면 간 차이 값을 화면 기술자로 사용하였다.여기에서 $v(i)$는 i번째 화면의 밝기 값의 분산을 의미한다.

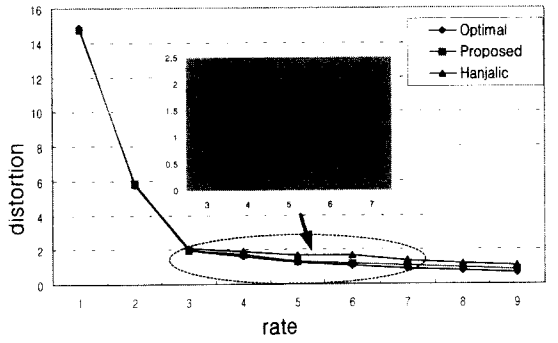
$$f(k)=\sum_{i=1}^{k}|v(i)-v(i-1)| \tag{15}$$

먼저, 객관적인 성능 평가를 위해, 제안된 방법과 Hanjalic의 방법을 율-왜곡의 관점에서 비교하였다. 여기에서 율이란 주요 화면의 개수를 의미하고 왜곡은 식 (3)에서 정의한 왜곡을 의미한다. 그림 6(a)에서 알 수 있듯이, 제안하는 방법의 성능이 율-왜곡의 관점에서 Hanjalic의 방법보다 우수함을 알 수 있었다. Hanjalic의 방법에서는, 주요 화면의 개수가 증가함에 따라 전체적으로는 왜곡이 줄어드는 경향이 있지만 중간 중간에 변동점(fluctuation point)이 발생함을 알 수 있다. Hanjalic의 방법은 전체 왜곡을 주요 화면과 분절점에 관해 1차 미분을 한 값이 0이 되는 위치를 찾음

으로써 식 (1), 식 (2)와 같은 재귀적인 관계를 유도하였다. 하지만, 1차 미분이 0이 된다는 것은 국부 최소인 경우 뿐 아니라, 국부 최대 혹은 전역 최대도 될 수 있음을 의미한다. Hanjalic의 방법에서 발생한 변동점은 이런 원인 때문이다. 반면, 제안하는 방법은 항상 수렴이 보장되는 반복 과정을 사용했기 때문에 이런 변동점은 발견되지 않았다.



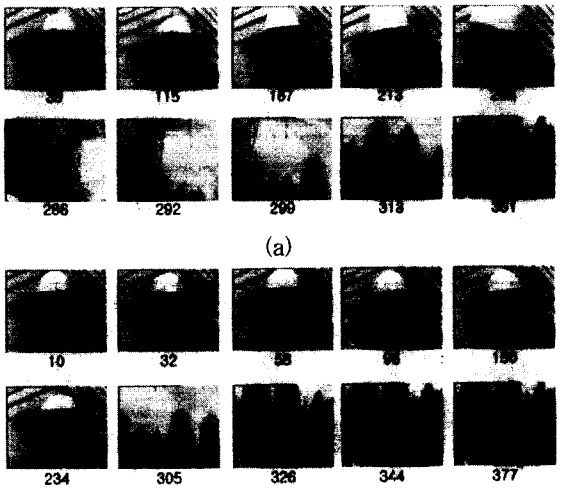
(a)



(b)

그림 6. 실험결과: 율-왜곡 관점 성능 비교
Fig. 6. Experimental result: Performance comparison in the sense of rate-distortion.

다음으로, 제안하는 방법을 최적의 방법과 율-왜곡 관점에서 성능을 비교하였다. 이 경우, foreman 영상열의 모든 화면을 이용하면 계산 시간이 너무 많이 걸린다. 예를 들어, 399개의 화면에서 5개의 최적 주요 화면을 찾기 위해서는 Pentium-III 컴퓨터로 약 95일 정도 걸림이 예상되었다. 따라서 모든 화면을 사용하지 않고 270에서 320까지의 50개의 화면을 사용하여 율-왜곡 성능을 평가하였다. 그림 6(b)에서 알 수 있듯이, 제안하는 방법이 최적의 방법과 거의 비슷한 성능을 보임을 알 수 있다.



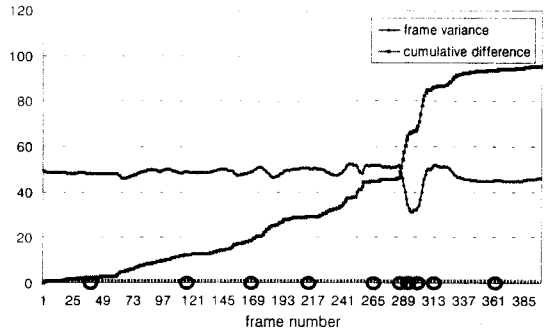
(a)

(b)

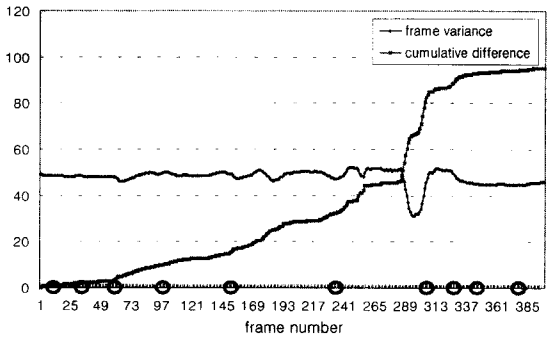
그림 7. 실험결과: 10개의 주요 화면을 선택한 경우 (a) 제안하는 방법 (b) Hanjalic의 방법
Fig. 7. Experimental result: Selected key frames (a) Proposed method (b) Hanjalic's method.

다음으로 주요 화면이 실제로 전체 샷을 얼마나 잘 대표하는지를 정성적으로 비교하였다. 그림 7은 제안하는 방법과 Hanjalic의 방법을 사용해서 10개의 주요 화면을 선택한 결과를 보여 준다. 먼저 그림 7(b)에서 알 수 있듯이 Hanjalic의 방법이 선택하는 주요 화면들은 남자의 얼굴 장면은 잘 표현한다. 하지만 급격하게 카메라 패닝이 일어나는 부분을 잘 표현하지 못함을 알 수 있다. 이 부분에서는 급격하게 내용물이 변화함에도 불구하고 이를 나타내는 주요 화면은 한 장(305번째 화면)밖에 선택되지 않았다. 그리고 공사장 장면에서는 실제 등장하는 시간과 내용물의 변화량에 비해 너무 많은 주요 화면이 선택되었다. 반면 그림 7(a)에서 알 수 있듯이 제안하는 방법이 선택하는 주요 화면들은 이런 부분을 잘 표현하고 있음을 알 수 있다. 즉 남자 얼굴 장면은 전체 샷에서 차지하는 시간이 길기 때문에 5~6장 가량의 주요 화면으로 표현되었으며, 패닝 장면도 실제로는 아주 짧은 시간을 차지하지만 내용물의 변화량이 크기 때문에 3~4장의 주요 화면으로 표현되고 있음을 알 수 있다. 그리고 공사장 장면은 내용물의 변화량은 거의 없고 지속 시간도 짧기 때문에 1~2장 정도의 주요 화면으로 표현되고 있다.

그림 8은 그림 7의 주요 화면들이 시간 축 상에서 어떤 식으로 분포하는지를 보여준다. 그림 8(a)에서 알 수 있듯이 제안하는 방법은 내용물의 변화가 적은 남



(a)



(b)

그림 8. 주요 화면의 분포(x축 위의 동그라미는 주요 화면의 위치를 의미한다.)
(a) 제안하는 방법 (b) Hanjalic의 방법

Fig. 8. Distribution of key frames along a video shot (small circles on the x-axis indicate the temporal positions of key frames)
(a) Proposed method (b) Hanjalic's method

자 얼굴 장면에서는 거의 균일한 간격으로 주요 화면을 선택하였고 카메라 패닝이 일어나는 부분에서는 조밀한 간격으로 주요 화면을 선택함을 알 수 있다. 그리고 장면이 지속되는 시간에도 비례하게 주요 화면을 선택함으로써 전체적으로 시간에 따른 내용물의 변화를 잘 표현하고 있음을 알 수 있다. 반면 그림 8(b)에서 알 수 있듯이 Hanjalic의 방법은 급격한 패닝이 일어나는 부분을 잘 표현하지 못함을 알 수 있다. 그 원인은 다음과 같다. Hanjalic의 방법에서는 현재 주요 화면의 위치는 이전 주요 화면 및 이전 분절점의 위치와 재귀적인 관계로부터 결정된다. 따라서 현재 주요 화면은 내용물이 크게 변하는 부분을 지나서 선택되기도 한다. 7 번째 주요 화면이 그 예이다. 내용물의 변화를 잘 표현하기 위해서는 이보다 앞선 화면이 주요 화면으로 선택되어야 하지만 현재 주요 화면은 이전 주요 화면 및 분절점과 재귀적인 관계가 있기 때문에 그렇게 되지 못하고 있다.

2. Effects of Initial Key Frames

초기 주요 화면이 성능에 미치는 영향을 알아보기 위해, 제안하는 초기 주요 화면 선택 방법을 랜덤하게 초기 주요 화면을 선택하는 방법, 균일한 간격으로 초기 주요 화면을 선택하는 방법과 비교하였다. 그림 9에서 알 수 있듯이, 율-왜곡의 관점에서는 세 방법이 모두 비슷한 성능을 보임을 알 수 있다. 다만, 주요 화면의 개수가 적을 경우, 랜덤하게 선택하는 방법의 왜곡이 다른 두 방법보다 약간 증가하지만, 주요 화면의 개수가 증가할수록, 왜곡에는 큰 차이가 없음을 확인하였다. 균일하게 초기 주요 화면을 선택하는 경우, 한쪽으로 초기 주요 화면이 몰리지 않기 때문에, 수렴 결과가 제안하는 방법과 크게 다르지 않았다. 반면, 랜덤하게 초기 주요 화면을 선택하는 경우에는, 주요 화면의 개수가 적을 때에는 한쪽으로 초기 주요 화면이 몰려버릴 확률이 크기 때문에, 수렴 결과는 더 나쁘게 나올 수도 있다. 하지만 주요 화면의 개수가 증가하면, 그럴 확률이 적기 때문에 성능이 비슷해진다. 이로부터, 초기 주요 화면이 전체 샷에 어느 정도 골고루 분포하면 수렴 결과에는 큰 차이가 없음을 알 수 있다.

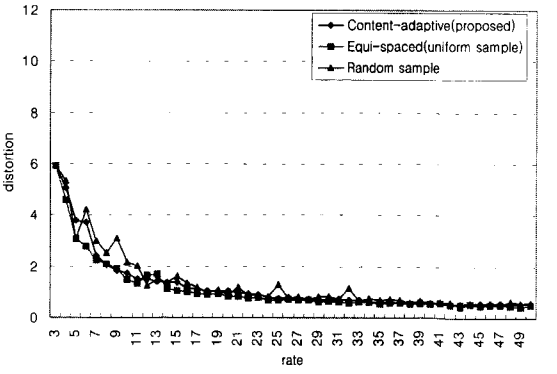


그림 9. 초기 주요 화면의 위치가 율-왜곡 성능에 미치는 영향.

Fig. 9. Effects of initial key frames on the rate-distortion performance.

반면, 아래의 표 1과 같이 수렴에 필요한 평균 반복 회수에서는 제안하는 초기 주요 화면 선택 방법의 반복 회수가 다른 두 방법보다 더 적음을 알 수 있다. 즉 내용물의 변화량을 고려해서 초기 주요 화면을 선택하면 계산 속도가 더 빨라짐을 알 수 있다.

표 1. 초기 주요 화면에 따른 평균 반복 회수 비교

Table 1. Effects of initial key frames on the average iteration numbers.

	Proposed	Uniform sampling	Random sampling
Average Iteration Number	4.98	5.12	7.58

V. Conclusion & Further works

본 논문에서는 주요 화면의 개수가 제한 조건일 때 순차적인 방법을 사용해서 주요 화면을 찾는 방법을 제안하였다. 이를 위해 먼저 원하는 개수의 초기 주요 화면을 미리 선택하고 이들이 대표하는 서로 중복하지 않는 시구간을 정한 후 반복 과정을 통해 주요 화면의 위치와 시구간의 크기를 조정하면서 왜곡이 최소가 되도록 주요 화면과 시구간을 찾는다. 실험 결과 제안하는 방법에 의해서 선택된 주요 화면들은 율-왜곡 관점에서 기존의 방법보다 우수하고 인간의 시각 인지와도 일치함을 알 수 있었다.

한편 제안한 방법은 주요 화면의 개수를 조정함으로써 왜곡-제한 주요 화면 선택 방법으로 쉽게 확장할 수 있다. 그리고 제안한 방법에 의해 선택한 주요 화면들은 비디오 검색을 위해 샷들 사이의 유사도를 정의하는 분야에서 사용될 수 있다. 일반적으로 두 샷들 사이의 유사도를 구하는 방법은 각 샷의 대표적인 화면 하나를 선택해서 이들끼리 비교하는 방법이다.[3] 하지만 이런 방법은 샷 안에 있는 내용물의 시간적인 변화까지는 고려하지 못한다는 단점이 있다. 따라서 시간적인 변화까지 고려해서 두 샷들 사이의 유사도를 정의하는 방법이 제안되었다.[7,8] 본 논문에서 제안하는 주요 화면 추출 방법은 두 번째 경우에 잘 적용될 수 있다. 그리고 계층적으로 비디오를 구성하는 분야에서도 사용될 수 있다. 향후에는 주요 화면을 이용해서 샷들 사이의 유사도를 정의한 후 비디오를 검색하는 방법과 빠른 검색을 위해 비디오 데이터 베이스를 효율적으로 구성하는 방법에 관해 연구를 진행할 계획이다.

References

[1] F. Idris and S. Panchanathan, "Review of Image and Video Indexing Techniques", *Journal of Visual Communication and Image Representation*, Vol. 8, No. 2, June, pp. 146~166, 1997.

[2] MPEG-7: Context, Objectives and Technical Roadmap, V.11, *ISO/IEC JTC1/SC29/WG11/N2729*, March 1999, Seoul.

[3] Hong Jiang Zhang, Jianhua Wu, Di Zhong and Stephen W. Smoliar, "An Integrated System for Content-based Video Retrieval and Browsing", *Pattern Recognition*, Vol. 30, No. 4, pp. 643~658, 1997.

[4] Huncheol Lee, CheongWoo Lee and SeongDae Kim, "Abrupt shot change detection using an unsupervised clustering of multiple features", *Proc. of ICASSP 2000*, Vol. 4, pp. 2015~2018.

[5] R.Brunelli, O.Mich, and C.M.Modena, "A Survey of the Automatic Indexing of Video Data", *Journal of Visual Communication and Image Representation*, Vol. 10, pp. 78~112, 1999.

[6] Ullas Gargi, Rangachar Kasturi, and Susan H. Strayer, "Performance Characterization of Video-Shot-Change Detection Methods", *IEEE Trans. Circuits and Systems for Video Technology*, pp. 1~13, Vol. 10, No. 1, 2000.

[7] Man-Kwan Shan and Suh-Yin Lee, "Content-based Video Retrieval based on Similarity of Frame Sequence", *Proc. Int'l Workshop on Multimedia Database Management Systems*, pp. 90~97, 1998.

[8] Yap-Peng Tan, Sanjeev R.Kulkarni and Peter J. Ramadge, "A Framework for Measuring Video Similarity and Its Application to Video Query by Example", *Proc. Int'l Conference on Image Processing*, 1999.

[9] Minerva M. Yeung and Boon-Lock Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content", *IEEE Trans. Circuits and Systems for Video Technology*, pp. 771~785, Vol. 7, No. 5, 1997.

[10] Minerva M. Yeung and Boon-Lock Yeo, "Segmentation of Video by Clustering and Graph Analysis", *Computer Vision and Image Understanding*, Vol. 71, No. 1, pp. 94~109,

1998.

[11] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," *in Visual Database Systems II*, 1992.

[12] B. Shahrary and D. C. Gibbon, "Automatic generation of pictorial transcript of video programs," *in Proc. IS&T/SPIE Digital Video Compression: Algorithms and Technologies*, San Jose, CA, 1995, pp. 512~519.

[13] Y. Zhuang, Y. Rui, T. S .Huang and S. Mehrotra, "Adaptive Key Frame Extraction using Unsupervised Clustering", *Proc. ICIP*, pp. 866~870, 1998.

[14] Nikolaos D. Doulamis, Anastasios, D. Doulamis, Yannis, S. Avrithis and Stefanos D.Kollias, "Video Content Representation using Optimal Extraction of Frames and Scene", *Proceedings of ICIP 98*, Vol 1, pp. 875~879, 1998.

[15] Hyun Sung Chang, Sanghoon Sulll and Sang Uk Lee, "Efficient Video Indexing Scheme for Content-based Retrieval", *IEEE Trans. Circuits and Systems for Video Technology*, pp. 1269 ~ 1279, Vol.9, No. 8, 1999.

[16] Minerva M. Yeung and Bede Liu, "Efficient Matching and Clustering of Video Shots", *Proceedings of ICIP 95*, 1995, pp. 338~342.

[17] A. Hanjalic, R.L. Lagendijk, J. Biemond, "A New Method for Key Frame based Video Content Representation", *Proceedings of the First International Workshop on Image Databases and Multi-Media Search*, Amsterdam (NL), 1996.

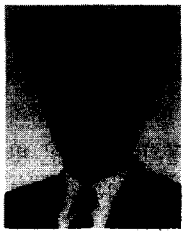
[18] R. L. Lagendijk, A. Hanjalic, M. Ceccarelli, M. Soletic, and E. Persoon "Visual Search in a SMASH System", *Proceedings of ICIP 96*, Vol. 3, pp. 671~674, 1996.

[19] Hun-Cheol Lee and Seong-Dae Kim, "Rate-driven Key Frame Selection using Temporal Variation of Visual Content", *Electronics Letters*, Vol. 38, Issue 5, pp. 217~218, Feb. 2002.

저 자 소 개

李 熏 哲(學生會員)

1995년 KAIST 전기 및 전자공학과 학사. 1997년 KAIST 전기 및 전자공학과 석사. 1997년~현재 KAIST 전기 및 전자공학과 박사 과정, <주관심분야 : 영상 처리, 컴퓨터 비젼, 비디오 검색 등임>



金 聖 大(正會員)

1977년 서울대학교 전자공학과 졸업(공학사), 1979년 한국과학기술원 전기 및 전자공학과 졸업(공학 석사) 1983년 프랑스 INPTENSEEIHT 졸업(공학박사), 1984년~현재 : 한국과학기술원 전자전산학과(전기 및 전자공학전공) 교수, <주관심분야 : 영상처리, 영상 통신, 컴퓨터 비젼, VLSI구현 등>