멀티카메라 기반 동영상 요약 시스템

임승빈○ 박한샘 민준기 황금성 조성배
연세대학교 컴퓨터과학과
{envymask○, sammy, loomlike, yellowg}@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Video Summarization System Based on Multi-Camera

Seung-Bin Im0, Han-Saem Park, Junki Min, Keum-Sung Hwang and Sung-Bae Cho
Department of Computer Science, Yonsei University

요 약

디지털 카메라 기술의 발전과 보급으로 공공건물의 보안 카메라부터 개인 휴대 단말기의 카메라까지 동영상 데이터를 수집할 수 있는 수단이 크게 늘었으며, 그 활용 또한 매우 일반화되었다. 동영상 데이터는 문서나 음성 등의 다른 데이터보다 훨씬 구체적이고 사실적인 정보를 포함하므로 과거의 기억을 정리하고 복원하기 위한 유용한 방법이 될 수 있다. 동영상 데이터의 증가와 함께 동영상 요약에 대한 연구가 최근에 활발히 진행되고 있는데, 이들 연구의 대부분은 하나의 동영상을 요약하고 분석하기 위한 것이다. 본 논문에서는 사무실에 여러 대의 카메라를 설치하여 데이터를 저장하며, 이렇게 수집된 동영상 데이터를 효과적으로 요약하고 검색하는 시스템을 구축한다. 동일한 이벤트를 여러 방향에서 바라보고, 그 상황을 가장 잘 설명한 카메라를 선택 할 수 있다는 점에서 멀티 카메라의 사용은 장점을 갖는다. 사전에 정의된 이벤트에 따라 전문가가 어노테이션을 부여하도록 하였으며, 전문가가 설정한 유틸리티에 따라 카메라 선택 및 요약이 이루어진다. 다양한 옵션에 따라 요약된 결과로 사용자 평가를 수행하였다.

1. 서 론

카메라 기술의 발전과 보급화로 주요 공공건물뿐만 아니라 일상생활 환경이나 길거리, 개인 휴대 단말기에서까지 많은 동영상 정보를 수집할 수 있게 되었다. 특히, 방송의 대상이 되는 무대나 스포츠 경기에서는 넓은 범위와 다양한 시점을 커버하기 위해서 많은 수의 카메라가 사용된다. 따라서 각각의 동영상을 어떻게 분석하고 결합하느냐에 따라서 하나의 사건에 대해서도 다양한 형태의 결과를 얻을 수 있는 가능성을 가지게 되었다. 또한 사용자가 일상생활에서 다양한 영상 매체에 노출되고 수집되기 때문에 굳이 자신이 기록을 남기지 않더라도 자신의 삶을 요약해서 저장할 수만 있다면 되돌아보거나 검색할 수 있는 보조 기억 매체로서의 활용가능성을 얻게 되었다. 본 논문에서는 이러한 개발에 필요한 기반 기술로써 사무 환경에서 수집된 멀티 카메라 영상을 분석하고 요약하기 위한 시스템 구축하였다.

최근 폭발적으로 늘어나고 있는 동영상 정보들은 멀티미디어 연구 분야에서 많은 과제들을 만들고 있는데, 특히 동시에 다양한 카메라에서 수집되는 영상의 경우 영상에 따라 담고 있는 정보의 질과 양이 다르기 때문에 사용자가 원하는 바에 가장 근접하는 정보를 추출하는 것뿐만 아니라 가장 적절한 대상을 선택해야 한다. 이때 필요한 것이 각각의 영상이 어떠한 정보를 얼마나 정확하게 담고 있는지를 분석하고 어노테이션을 부여하는 것이다. 본 논문에서는 사무실 환경에서 발생 가능한 여러 가지 상황을 사전에 정의하였고, 전문가가 이에 맞추어 어노테이션을 수행하도록 하였다. 동시에 여러 카메라를 사용하는 환경에서는 많은 동영상이 한꺼번에 수집되기 때문에 사용자가 원하는 영상만을 검색하거나 요약하기 위한 기술의 연구가 필요하다. 본 논문에서는 일상적인 사무 환경에서 수집된 다른 카메라 동영상을 분석하여 사용자가 원하는 영상만을 효과적으로 검색·요약하기 위한 방법에 대해 다룬다.

2. 관련 연구

앞에서 언급했듯이 동영상 정보는 이제 매우 일반적인 데이터가 되었으며, 따라서 이를 요약하거나 원하는 정보를 찾아내는 연구 또한 많은 연구자들에 의해 진행되고 있다. 동영상 요약 방법은 결과의 형식에 따라 두 가지로 나뉘는데[1], 첫째는 중요한 여러 장의 정지 영상을 선택해, 시간 순으로 나열하는 방법(pictorial summary)이며[2], 둘째는 여러 개의 짧은 동영상 클립을 합한 형태의 요약 방법(video skimming)이다[3]. 후자가 더 자연스러운 형태의 결과를 제공하며, 더 많은 연구가 진행되고 있다.

표 1은 동영상 요약 관련연구를 정리한 것이다. Y. Li 등은 시청각 데이터의 길이와 속도 등을 분석하여 동영상을 요약하는 방법을 제안하였으며[3], Y.-F. Ma 등은 사용자가 관심을 갖는 부분에 초점을 맞춘 요약 프레임워크인 사용자 주의 모델(user attention model)을 제안하였다[4]. 계층적 접근방식은 동영상 요약을 위해 자주 사용되는 방식 중 한가지인데, 하위 레벨에서 영상처리 등의 분석을 이용한 어노테이션을 거친 후 내용 정보를 바탕으로 요약을 수행한다[1, 5]. X. Zhu 등은 내용을 효과적으로 기술하기 위해 온톨로지를 활용하였으며, D. Chen 등은 영상분석 과정에서 온톨로지를 이용한 체계적인 어노테이션을 수행하여 행동 인식의 성능을 높였다[6].

위와 같이 대부분의 동영상 요약 연구는 하나의 비디오를 분석하고 요약하기 위한 것이었다. 아직 초기 단계이긴 하지만 최근에는 멀티카메라를 이용한 연구 또한 일부에서 연구되고 있다. G. C. de Silva 등은 유비쿼터스 가정환경에서의 정보 요약 및 검색을 위해 멀티 카메라를 사용하였으며[7], Y. Sumi 등은 학회장에 다양한 센서를 설치하여 데이터를 수집한 후 기본 행동 간의 관계를 정의한 템플릿을 이용하여 참가자들의 행동을 요약하였다[8]. 이러한 연구들은 멀티 카메라를 넓은 공간을 커버하기 위해 사용하였는데, 이와 달리 본 논문에서는 유사한 영역을 다양한 각도에서 커버하기 위한 목적으로 멀티 카메라를 사용하였다.

표 1. 동영상 요약 관련연구

연도	저자	내용
2003	X. Zhu, et al.	온톨로지를 활용한 계층적 내용 기술 및 요약
2004	D. Chen, et al.	영상 분석 방법에 온톨로지를 기반으로 한 체계적인 접근 방법을 사용, 카메라에서 관찰된 상호작용을 탐지
2005	Y.-F. Ma et al.	사용자 주의(attention) 모델의 이용
2006	Ying Li, et al.	시청각 데이터의 길이와 속도를 분석하여 동영상을 요약

3. 멀티카메라 기반 동영상 요약 시스템

본 논문에서는 무선 네트워크로 연결된 다방향의 멀티 카메라를 이용하여 실내 사무 환경에서의 일상생활을 영상으로 저장하고, 사용자의 요구에 맞게 요약, 검색하는 시스템을 구현하였다. 영상 요약을 위한 과정은 다음과 같다. 먼저 다양한 높이와 각도에서 동영상 데이터를 수집하고, 수집한 동영상의 이벤트 어노테이션이 수행된다. 마지막으로 어노테이션된 영상의 최적 카메라 영상 선택 및 요약이 이루어진다.

3.1. 카메라 설치 및 데이터 수집

카메라의 촬영영역에서 이벤트의 발생 영역이 카메라의 촬영 영역과 일치하더라도, 사람과 물체의 방향에 따라 영상이 가려지거나 잘 보이지 않게 되는 경우가 있다. 즉 동시에 다양한 카메라에서 수집되는 영상의 경우 영상에 따라 담고 있는 정보의 질과 양이 다르기 때문에 성공적인 요약을 위해서는 각 상황에 맞는 가장 적절한 정보를 포함한 영상의 선택과 영상의 중요한 의미를 잃지 않는 범위에서의 데이터의 압축이 매우 중요하다.

실내 사무 환경에서 일어나는 다양한 이벤트를 멀티카메라를 이용하여 수집하기 위하여 본 논문에서는 사무실 내의 3.8m x 3.2m 영역을 목적 영역으로 설정하고 세 대의 카메라를 설치하였다. 그림 1은 사무실의 구조도와 설치된 카메라들의 위치를 보여준다. 위쪽 코너와 아래쪽 가운데 위치한 세 개의 노란색 원은 카메라를 나타내며 각 카메라가 커버할 수 있는 각의 범위도 함께 표현되었다. 이 그림에서 우리의 목적 영역은 책상 다섯 개가 모여 있는 사무실의 안쪽 영역이다.

영상 수집은 소니 네트워크 카메라 SNC-P5를 이용하여 320x240, 30fps로 저장하였다. 그림 2는 사용한 카메라의 모습을 보여준다. 설치된 카메라는 무선 네트워크를 이용하여 연결되었으며 획득된 데이터는 MPEG 형식의 avi 동영상 파일로 저장되며 주기적으로 각 무선 네트워크 카메라와 연동된 개별 서버에 저장된다. 설치된 세 대의 카메라의 촬영 범위는 큰 차이가 없지만 카메라의 위치와 방향이 다르기 때문에 수집한 영상은 담고 있는 정보의 질과 양에서 큰 차이를 보이게 된다.

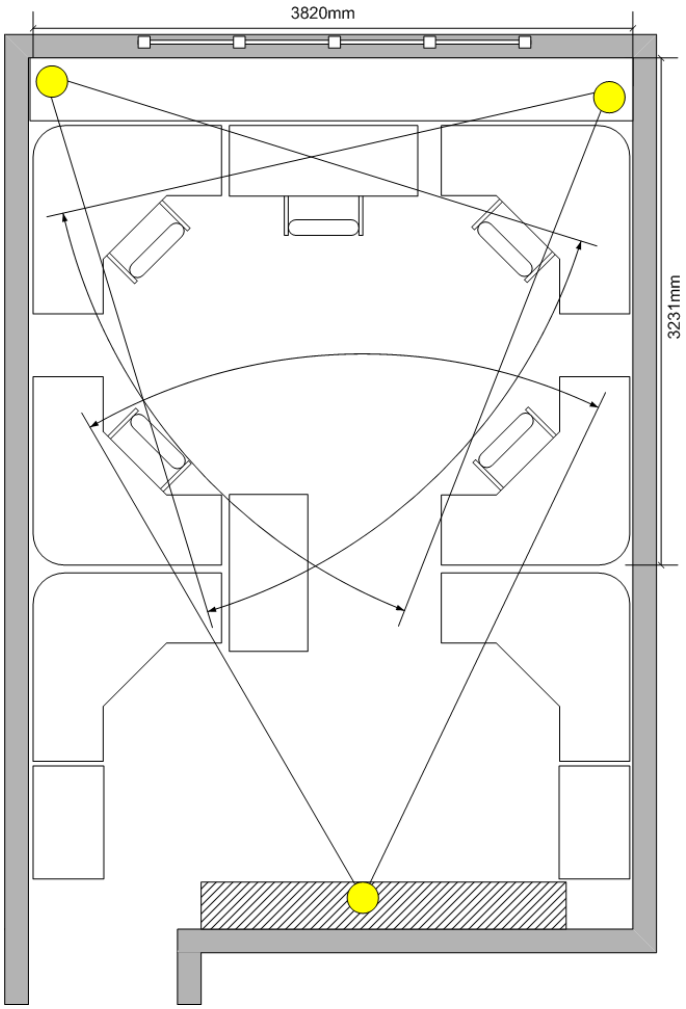

그림 1. 사무실 내부의 자리 배치와 카메라의 위치

그림 2. 소니 네트워크 카메라 SNC-P5

3.2. 이벤트 어노테이션

비디오 요약을 위해서는 영상에서 사람의 행위나 이벤트가 랜드마크 정보로 필요하며, 화면 선택의 기준으로 사용할 수 있는 신뢰도 혹은 정확도 값 등이 레이블링 되어있어야 한다[10]. 본 논문에서는 이벤트를 판단하는 기준이 그 구성요소에 있다고 보고, 이를 사람의 행위, 참여 수, 사람이 다루고 있는 물체에 따라 결정하였다. 이를 위해 사무실 환경에서 발생할 수 있는 이벤트, 사람의 행위, 사용가능한 물체를 다음과 같이 정의한다. 표 2는 이를 이용하여 각 이벤트를 결정하는 규칙을 보여주고 있다.

- 이벤트: Work, Call, Eat, Cleaning, Presentation, Meeting, Conversation, In, Out, Sleep
- 행위: Sit, Talk, Gaze, Move, Object hold, Object move
- 물체: Computer, Book, Telephone, Food, Cleaning kit

표 2. 사무실 환경에서 발생하는 이벤트의 종류와 결정 규칙

이벤트	행위	개체	사람수	지속시간 (초)	비고
Work	Sit, Gaze	Computer or Book	1	>=60	업무
Call	Talk, Hold	Telephone	1	>=60	전화통화
Eat	Hold, Obj_Move	Food	1	>=60	음식섭취
Cleaning	Move, Hold	Tool	1	>=60	청소
Presentation	Talk, Gaze	Computer	>=2	>=60	한사람이 다른 사람의 컴퓨터 화면을 보며 설명을 들음
Meeting	Talk, Sit, Gaze		>=3	>=60	3인 이상의 사람이 모여서 회의를 함
Conversation	Talk, Gaze		>=2	>=60	2인이 대화
In					사무실 들어옴
Out					사무실 나감
Sleep					자리에서 수면

이벤트의 레이블링을 위해서는 영상분석을 통한 자동 행동인식기술이 필요하다. 뉴스와 같이 비디오영상 내에 자막이 있는 경우는 문자인식을 통해 비교적 쉽게 이벤트의 레이블링이 가능하다[9]. 그러나 대부분의 영상은 자막이 없기 때문에 영상처리를 통한 행동인식이나 영상 변화량 분석 등의 기술이 필요하며, 미리 정의해놓은 규칙에 따라 사람이 직접 어노테이션 하기도 한다[10]. 본 논문에서는 앞에서 정의한 규칙에 따라 이벤트의 레이블을 사람이 직접 어노테이션 하였으며, 그림 3은 이를 위해 구현한 어노테이션 툴을 보여준다.

각 이벤트는 촬영한 카메라의 위치나 다른 사람, 혹은 환경 내의 다른 물체에 의해 정확하게 보이거나 가려지기도 한다. 또한 사람의 시선(Gaze)이나 쥐고 있는 물체 등 이벤트를 구성하는 요인들이 화면에 나타나지 않는 경우도 있기 때문에 툴에서 간단한 확률 값(50, 75, 100%)을 이벤트 도중에 변경할 수 있도록 하였다.

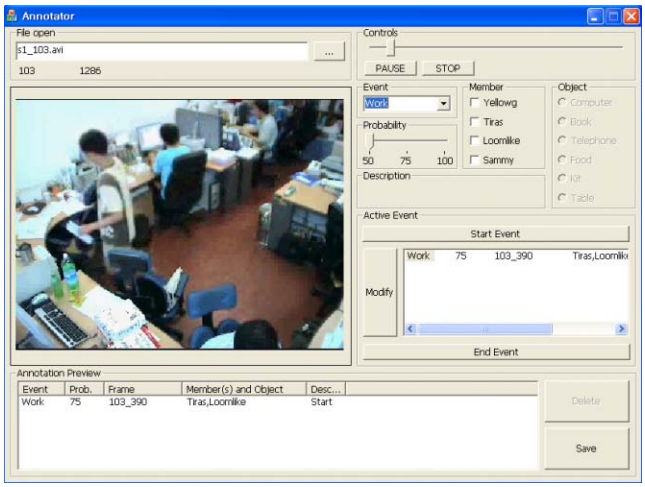

그림 3. 이벤트 어노테이션 툴

3.3. 카메라 선택 및 요약

이벤트 어노테이션이 종료되면, 어노테이션을 바탕으로 카메라 선택 및 요약을 수행한다. 동일한 이벤트를 여러 카메라의 동영상이 모두 포함한다고 하면, 어노테이션시에 전문가가 부여한 유틸리티 값에 의해 시스템은 그 중 하나의 카메라를 선택한다. 전문가는 동영상이 포함하는 사람의 방향, 그 사람이 갖고 있는 물체, 그리고 특정 행동을 얼마나 잘 표현하고 있는지를 고려하여 유틸리티를 부여한다. 동일한 이벤트의 오랜 지속 혹은 큰 변화가 없는 영상의 유틸리티 값은 줄어들게 된다. 이렇게 주어진 유틸리티 값과 이벤트 인식에 가장 적절한 카메라의 선택으로 요약이 수행된다.

$$\arg\max_{1\le i\le N}\{Utility_i(direction, object, reprsentativity)\} \quad (1)$$

이벤트를 요약할 때, 각 이벤트는 시간적으로 겹치거나 동시에 일어나는 경우가 있다. 이런 경우 각 이벤트를 모두 보여주려면 필연적으로 영상의 반복 재생이 필요하다. 예를 들어 meeting 이벤트의 진행 도중 어떤 사람은 음식을 먹는 eating 이벤트를 수행할 경우, 두 가지 이벤트를 강조해서 요약하려면 meeting을 보여주고, 다시 eating을 보여주는 등의 반복 재생이 필요하다. 따라서 이벤트 요약방식은 영상의 반복재생을 허용하더라도 모든 이벤트를 보여주는 이벤트 기반 방식과, 시간의 흐름을 중요시하여 겹치는 이벤트는 결합하여 영상의 반복재생을 허용하지 않고 요약하는 시퀀스 기반 요약방식으로 나뉠 수 있다. 이벤트 기반 방식은 반복재생이 수행될 경우 어떠한 이벤트를 보여주고 있는지 알아내기 어려운 경우가 많으므로 요약된 이벤트 결과를 텍스트정보로 제공해 준다.

요약된 영상을 표현할 때, 요약 알고리즘에 의하여 줄여진 영상이라 하더라도 이동이나 변화가 적은 영상은 사용자가 지루함을 느끼기 쉽다. 따라서 행동의 변화가 적은 이벤트가 오랫동안 지속될 경우 재생속도의 변화를 주어 지루함을 덜어줄 수 있도록 하였다.

3.4. 어플리케이션

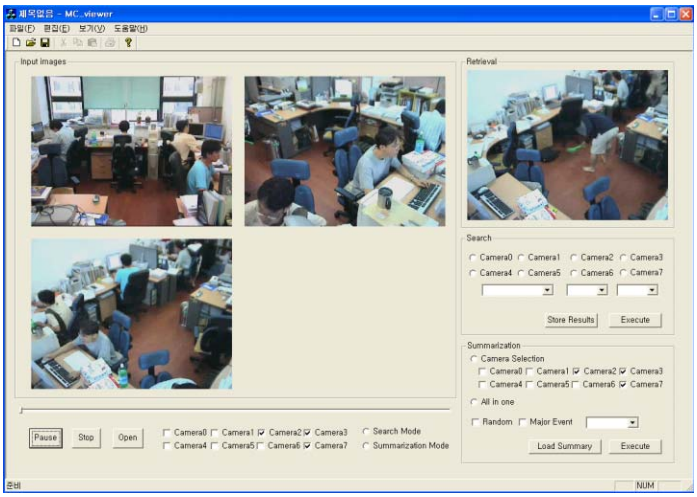

그림 4. 영상 요약 및 검색 어플리케이션

본 논문에서는 제안하는 요약방법을 실제 동영상 요약에 적용하기 위하여 DirectShow기반의 동영상 뷰어 및 영상 요약 어플리케이션을 구현하였다. 어플리케이션은 수집된 멀티카메라 영상의 동시 재생이 가능하며 원본 영상을 요약하여 하나의 결과 영상으로 만들어준다. 또한 일반적인 요약이 아닌, 특정 사람의 하루 일과나 특정 이벤트의 요약 같은 사용자 질의문에 따른 요약이 가능하며 어노테이션된 결과를 이용한 영상 검색 또한 가능하다. 그림 3은 실제 구현된 영상 요약 및 검색 어플리케이션의 모습을 보여준다.

4. 사용자 평가

본 논문에서는 실제 실내 사무 환경에서의 일상생활 영상을 실험으로 사용하기 위하여 3개의 시나리오를 따라 영상을 수집하였다. 수집된 영상들은 전문가의 어노테이션 과정을 거친 후 제안하는 요약 시스템을 통하여 요약된다. 각각의 요약된 영상의 사용자 평가를 통하여 제안하는 시스템의 유용성을 보였다.

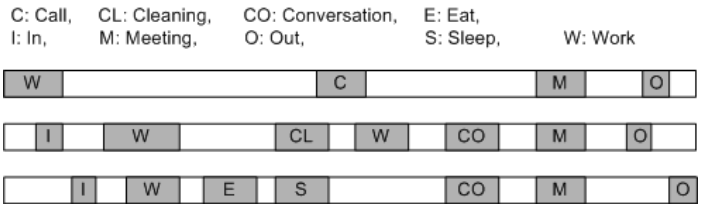

그림 5. 실제 실험에서 사용한 시나리오의 예

제안하는 방법 영상요약이 사용자에게 얼마나 만족도를 줄 수 있는지를 판단하기 위하여 사용자 평가를 수행하였다. 데이터는 각 시나리오당 약 30분간 촬영하였고, 요약결과 약 4분에서 6분정도의 영상으로 영상의 길이를 줄여줄 수 있었다. 10명의 사용자에게 원본 영상과 요약된 영상을 제공하고 각 요약된 영상에 대하여 1부터 5까지의 만족도를 측정하였다. 그림 5는 실제 실험에서 사용한 시나리오의 한 예이다.

첫 번째로, 요약된 영상을 재생할 때 어떻게 영상을 재생하면 사용자가 지루함을 느끼지 않고 볼 수 있는가를 조사하기 위하여 다음과 같은 실험을 수행하였다. 기본적으로 움직임이 적은 computer work 같은 경우에는 동일한 시간으로 요약된 다른 이벤트에 비하여 영상의 변화가 적으므로 지루함을 느끼기 쉽다. 또한 presentation 이벤트도 computer work에 비해서는 변화가 크지만 비슷한 영상이 오래 지속될 수 있으므로 미리 정의된 룰에 따라 각 이벤트별 선택 재생속도 조절을 통하여 요약 영상을 재생해주면 영상의 지루함을 줄여줄 수 있다. 표 3은 실험에서 사용한 각 이벤트별 재생속도를 보여주고 있다. 표 4는 속도가 조절되지 않은 영상과 재생속도 조절이 적용된 영상의 실제 사용자 평가 결과이다. N은 보통 속도 재생을 나타내고 SC는 속도 조절 재생 방식이다. 세 가지 시나리오 모두에서 선택적 속도 조절이 더 좋은 만족도를 보이는 것을 확인할 수 있다.

표 3. 이벤트별 재생속도

	Computer work	Presentation	Cleaning	Sleep
재생속도	x4	x2	x2	x3

표 4. 선택적 배속 재생 결과.

	S1		S2		S3	
	N	SC	N	SC	N	SC
만족도	2.5	3	2	4	2	4.2

두 번째 실험으로 이벤트 기반 요약방식과 시퀀스 기반 요약 방식의 비교실험을 수행하였다. 표 5는 요약방식의 비교실험 결과를 보여주고 있다. E는 이벤트 기반 요약방식, C는 이벤트 기반방식에 캡션이 추가된 요약방식, S는 시퀀스 기반 요약방식을 나타낸다. 결과를 살펴보면, 이벤트 기반 요약방식은 재생의 반복을 허용하므로 지루해지기 쉽다는 단점 때문에 만족도가 떨어졌으며 이해를 돕기 위한 텍스트 정보를 추가로 주었을 때 만족도가 상승하는 것을 볼 수 있다. 그러나 반복을 허용하지 않고 반복영상의 경우 한번만 보여주는 시퀀스 기반 요약방식의 경우에는 텍스트 정보가 추가된 이벤트 방식보다 더 좋은 만족도를 내는 것을 확인할 수 있다. 시나리오 2는 각 이벤트들이 연속적으로 일어나는 경우가 많아 재생의 반복이 다른 시나리오에 비하여 더 많이 일어나 만족도가 저하되었으며 세 번째 시나리오는 1과 2보다 사람들의 동적인 이동이 많아 만족도에서 더 좋은 결과를 얻었다.

마지막 실험으로 단순 요약이 아닌 사용자 질의에 따른 요약에 대한 사용자의 만족도를 측정하였다. 해당 영상에서 대상 사람이 어떤 행동을 하였는지에 대한 요약 질의에서 표 6과 같은 결과를 확인할 수 있었는데 대체적으로 일반적 요약에 비하여 높은 만족도를 보이는 것

을 확인할 수 있었다.

표 5. 요약방식의 비교실험 결과.

	S1			S2			S3		
	E	C	S	E	C	S	E	C	S
만족도	2	3.5	4	1	3.2	3	2	3	4

표 6. 질의요약방식의 실험 결과.

	S1		S2		S3	
	A	B	A	B	A	B
만족도	4	5	3	3	5	3.5

5. 결론 및 향후 계획

본 논문에서는 사무실 환경에 설치된 멀티 카메라에서 수집된 동영상을 여러 가지 형태로 요약하여 간단하게 보여주는 시스템을 구축하였다. 동영상에서의 컨텍스트 정보 및 상황을 부여하기 위해 어노테이션 규칙을 정의하고 체계적인 어노테이션을 수행하였다. 이때, 어노테이션 도구를 개발하여 사용자가 쉽게 어노테이션을 할 수 있도록 하였다. 동시에 여러 카메라에서 수집된 동영상은 어노테이션의 비교·분석을 통해 사용자의 요구에 따라 일부가 선택되고 요약되었다.

실제 실내 사무환경에서 여러 카메라를 이용하여 수집된 동영상에 어노테이션을 수행하고 요약을 수행해 본 결과, 사용자들은 요약 방식에 따라 다양한 평가를 내렸으며, 사용자의 의도에 따른 요약이 잘 수행되고 있음을 알 수 있었다. 본 연구는 현재 진행 중인 연구로써 현재는 사용자에 의한 수동 어노테이션을 통해 데이터를 수집하고 있으나, 향후에는 다양한 영상 처리 기법을 접목하여 자동 어노테이션이 가능할 것이다. 또한 도메인을 확장하여 더 많은 카메라에서 더 오랫동안 수집된 데이터를 이용한 실험이 수행될 것이다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 지원사업의 연구결과로 수행되었음, IITA-2005-(C1090-0501-0019).

참고문헌

[1] X. Zhu, J. Fan, A. K. Elmagarmid and X. Wu, “Hierarchical video content description and summarization using unified semantic and visual similarity,” *Multimedia Systems*, vol. 9, no. 1, pp. 31-53, Jul. 2003.

[2] M. Yeung and B. Yeo, "Video visual ization for compact presentation and fast browsing of pictorial content," *IEEE Trans CSVT*, vol. 7, pp. 771-785, 1997.

[3] Y. Li, S.-H. Lee, C.-H. Yeh and C.-C. Jay Kuo, “Techniques for movie content analysis and skimming,” *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 79-89, Mar. 2006.

[4] Y. Ma, X. Hua, L. Lu and H. Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE Trans. on Multimedia*, vol. 7, no. 5. pp. 907-919, 2005.

[5] B. L. Tseng, C.-Y. Lin and J. R. Smith, “Video personalization and summarization system for usage environment,” *Journal of Visual Communications and Image Representation*, vol. 15, no. 3, pp. 370-392, Sep. 2004.

[6] D. Chen, J. Yang, and H. D. Wactlar, "Towards Automatic Analysis of Social Interaction Patterns in a Nursing Home Environment from Video," *Proc. of ACM Multimedia, 6th ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, pp. 283-290, New York, Oct. 15-16, 2004.

[7] G.C. De Silva, T.Yamasaki and K. Aizawa, "Experience retrieval in a Ubiquitous Home," *ACM Multimedia Workshop on Continuous Archival of Personal Experience*, pp. 35-44, Singapore, Nov. 11, 2005.

[8] Y. Sumi, S. Ito, T. Matsuguchi,, S. Fels, and K. Mase, "Collaborative capturing and interpretation of interactions," *Pervasive Workshop on Memory and Sharing of Experiences*, pp. 1-7, Apr. 2004.

[9] M. Bertini, A. D. Bimbo, and P. Pala, "Content-based indexing and retrival of TV news", *Pattern Recognition Letters*, Vol. 22, pp. 503-516, 2001.

[10] A. Amir, S. Basu, G. Iyengar, C.-Y. Lin, M. Naphade, J. R. Smith, S. Srinivasan, and B. Tseng, "A Multi-modal system for the retrieval of semantic video events," *Computer Vision and Image Understanding*, vol. 96, pp. 216-236, 2004.