사용자 로그분석을 이용한 멀티 카메라 사무실 이벤트 요약

박한샘O 조성배
연세대학교 컴퓨터과학과
sammyO@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

A Summarization of Multi-Camera Office Event Using User Log Analysis

Han-Saem Park0 Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

최근 카메라를 비롯한 다양한 센서 기술 및 디지털 저장장치의 발달로 사용자의 일상생활의 기록인 라이프 로그를 수집하고 분석하는 연구가 활발히 이루어지고 있다. 라이프 로그는 모바일 디바이스에 포함된 다양한 센서를 통해 실외에서 수집되는 경우와 실내에 카메라를 중심으로 한 센서를 설치하여 수집되는 경우로 나누어 볼 수 있으며, 수집된 로그는 다양한 방법을 통해 분석하여 사용자에게 요약이나 검색과 같은 서비스 제공에 활용될 수 있다. 본 논문은 오피스 환경에 다수의 카메라를 설치하여 수집한 실내 비디오 로그 데이터를 대상으로 하며, 사용자의 어플리케이션 로그를 분석하여 요약을 위해 활용한다. 다수의 카메라는 오피스의 가운데 부분을 비추도록 하여, 발생한 하나의 이벤트에 대한 다양한 시점의 영상을 얻을 수 있도록 하였다. 전체 요약 과정은 크게 데이터 어노테이션, 사용자 로그분석을 이용한 이벤트 시퀀스 요약, 도메인 지식을 이용한 카메라 뷰의 선택으로 나뉘어 수행된다. 최종적으로 실험을 통해 제안하는 요약 방법이 좋은 결과를 보임을 확인하였다.

1. 서 론

최근 카메라, GPS를 비롯한 다양한 센서 기술의 발달과 초고속 네트워크 서비스의 보급, 디지털 저장장치의 가격하락 등으로 인해 사용자의 일상생활의 기록이라고 할 수 있는 라이프 로그를 수집하고 분석하는 연구가 활발히 이루어지고 있다[1]. 이러한 연구는 개인의 생활을 기록으로 남김으로써 원하는 기억을 나중에 검색하고자 할 때에 유용하다.

라이프 로그는 실외에서 수집되었을 때와 실내에서 수집되었을 때, 데이터의 성격이 많이 다르다. 실외에서는 주로 모바일 디바이스에 포함된 다양한 센서를 활용하여 데이터가 수집되며, 수집된 정보는 모바일 컨텍스트의 추론, 주요 사건 요약 등의 서비스 제공을 위해 활용된다[2]. 실내에서는 카메라를 통해 수집된 비디오 데이터를 중심으로 다양한 로그가 수집될 수 있으며, 이를 분석하여 사용자에게 요약이나 검색과 같은 서비스를 제공해 줄 수 있다[3].

본 논문은 오피스 환경에 여러 대의 카메라를 설치하여 수집한 실내 비디오 로그 데이터를 대상으로 하며, 사용자의 어플리케이션 사용 로그를 분석하여 이벤트의 요약을 위해 활용하였다. 제안하는 방법은 또한 실내 오피스 이벤트를 멀티 카메라를 통해 잡아냄으로써 하나의 이벤트에 대한 다양한 뷰를 제공할 수 있는 장점을 갖는다.

2. 관련연구

2.1. 실내 라이프 로그 데이터의 활용

실내 환경에서 수집된 라이프 로그는 카메라를 통해 수집된 비디오 데이터가 주가 되며, 그 외에 필요에 따라 소리 센서, 압력 센서 등이 적절히 로그 데이터 수집을 위해 활용된다.

Y. Sumi 등은 학회장에서 학회 참석자들의 데이터를 수집하기 위해 천장의 유비쿼터스 센서와 안내 로봇의 카메라, 일부 학회 참석자들의 카메라를 이용하였고, 참석자들에게 간단한 요약 서비스를 제공하였다[4]. G. C. Silva 등은 유비쿼터스 홈에서 사용자들의 로그를 수집하여 요약 및 검색이 가능한 멀티미디어 다이어리 서비스를 제공하였다[3]. 유비쿼터스 홈 내에서의 위치 정보를 수집하기 위해 바닥에 압력센서를 사용하였으며, 대부분의 영역을 커버할 수 있도록 방마다 카메라와 마이크를 설치하여 데이터를 수집하였다. C. Zhang 등은 실내 이벤트 데이터를 다양한 시점에서 비추기 위해 네 대의 카메라를 사무실 구석에 설치하였으며, 정확한 위치를 잡아내기 위해 사용자에게 적외선 센서를 부착하고 적외선 카메라를 네 대의 카메라 옆에 설치하였으며, 이를 이용하여 사용자에게 검색 서비스를 제공하였다[5].

2.2. 비디오 요약

본 논문에서 요약의 대상으로 다루어지게 될 비디오 데이터는 영화, 뉴스, 스포츠 등 전문가에 의해 촬영된 것에서 미리 셋업 된 환경에 카메라를 설치하여 연구자가 직접 수집한 것까지 그 범위가 다양하다. 영화, 뉴스와 같이 장면(scene)과 샷(shot)의 구분이 명확하고 배경의 변화가 동적인 비디오의 요약 및 검색을 위한 절차는 크게 내용 분석(content analysis), 구조 분석(structure parsing), 요약(summarization) 및 검색(indexing & retrieval)의 네 단계로 나누어진다[6]. 내용 분석은 하위수준의 특징으로부터 의미 수준의 내용을

연결하는 단계로서, 영상 처리 기법과 인식 기술 등을 이용해서 자동으로 수행되기도 하고[7], 사람에 의해 수동으로 수행되기도 한다[8]. 구조 분석은 내용 분석 결과를 바탕으로 비디오 데이터를 개별 장면이나 샷으로 나누는 단계이다. 요약은 이렇게 분석된 정보를 이용하여 중요한 내용을 간결하게 보여주는 단계이다.

직접 환경을 구축하고 카메라를 설치하여 수집된 데이터의 경우 2.1의 실내 라이프 로그 데이터와 같이, 주로 실내에서 여러 대의 카메라를 함께 사용하는 멀티 카메라 환경을 대상으로 한다[3-5]. 이러한 데이터는 실내가 배경이 되므로 배경 변화가 정적이고, 따라서 장면보다는 사용자의 위치나[3], 해당 도메인에서 발생하는 행동(activity)이나 이벤트(event) 단위로 비디오 데이터를 나누는 방법을 주로 사용한다[4].

3. 멀티 카메라 사무실 환경

3.1. 실험 환경

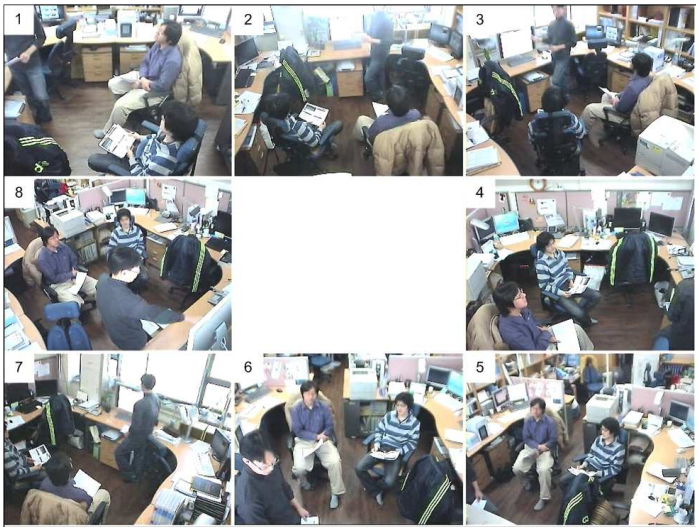

그림 1. 멀티 카메라 사무실 환경에서 수집된 영상 예

실험 환경은 Park 등이 수행했던 기존 연구와 같다[9]. 간단히 설명하면, 연구실 내 4m×4m 영역을 타겟으로 8대의 카메라가 둘러서 중앙을 바라보도록 설치하여 하나의 이벤트를 다양한 시점에서 바라볼 수 있도록 환경을 설정하였다. 설치된 카메라에 의해 수집된 영상은 그림 1과 같다. 왼쪽 위에서부터 시계방향으로 카메라의 설치순서와 동일하게 그림을 배치하였으며, 제시된 예는 미팅장면에 대한 것이다. 실험을 위해 사용된 카메라는 소니 네트워크 카메라 SNC-P5이며, 30 fps의 MPEG 파일로 저장하였다.

3.2. 이벤트 어노테이션

본 논문에서는 사무실 환경에서 발생할 수 있는 일반적인 이벤트를 사전에 정의하였으며, 이를 바탕으로 동영상의 어노테이션이 이루어졌고, 그 상위에서 요약이 수행되었다. 사무실 이벤트 정의는 다음과 같다.

- Entry (*A*),
  if stand (*A*, *entrance-area*) and face (*A*, *in*)
- Leaving (A),
  if stand (*A*, *entrance-area*) and face (*A*, out)
- Calling (*A*),
  if hold (*A*, *phone*) and speak (*A*)
- Vacuuming (*A*),
  if hold (*A*, *vacuum cleaner*) and stand (*A*, *center-area*)
- Eating (*A*),
  if hold (*A*, *food*)
- Nap (*A*),
  if rest (*A*, *corner-area*)
- Work (*A*),
  if sit (*A*, *corner-area*) and
  {use (*A*, *computer*) or hold (*A*, *document*)}
- Printing (*A*),
  if exist (*A*, *printer-area*) and hold (*A*, *printout*)
- Conversation (*A*, *B*),
  if {exist (*A*, *x*) and exist (*B*, *y*) and close (*x*, *y*)} and
  {speak (*A*) or speak (*B*)}
- Meeting (*A*, *B*, *C*),
  if {exist (*A*, *x*) and exist (*B*, *y*) and exist (*C*, *z*) and close (*x*, *y*, *z*)} and
  {speak (*A*) or speak (*B*) or speak (*C*)} and
  {hold (*A*, *document*) or hold (*B*, *document*) or hold (*C*, *document*)}
- Seminar (*A*, *B*),
  if {stand (*A*, *screen-area*) and sit (*B*, *center-area*)} and speak (*A*)

영상처리 및 패턴인식 기법을 통한 물체 및 이벤트 인식 연구는 그 자체로 중요한 연구 이슈로써, 기존의 비디오 요약 연구에서는 대부분 자동 혹은 반자동으로 물체나 이벤트의 어노테이션을 수행하였다[7, 10]. 본 논문에서는 사람의 어노테이션으로 대체하였으며, 이 부분은 향후 자동화 할 예정이다.

4. 제안하는 방법

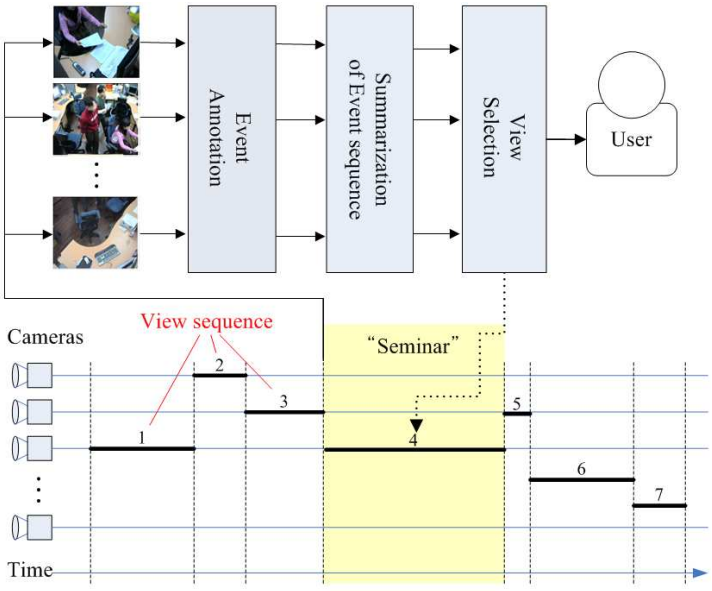

그림 2. 제안하는 방법의 개요

제안하는 방법의 전체적인 개요는 그림 2와 같다. 8대의 카메라를 이용해 수집된 이벤트 시퀀스를 입력으로 받아 요약이 수행되며 전체 과정은 이벤트 어노테이션,

이벤트 시퀀스의 평가 및 요약, 그리고 카메라 뷰 선택의 세 부분으로 나뉜다. 이벤트 어노테이션은 3.2에 기술된 바와 같다.

4.1. 사용자 로그 분석을 이용한 샷 평가 및 요약

어노테이션 된 이벤트 시퀀스는 샷으로 나뉜다. 샷은 "이벤트, 사람, 물체 정보가 동일하게 유지되는 연속된 프레임의 집합" 으로 정의하였다. 나뉘어진 샷은 사용자 로그 분석을 바탕으로 다음의 식 (1)~(5)에 의해 평가된다. 이 때, 사용자 로그는 사용자가 요약을 위해 선택한 키워드의 로그를 의미한다.

$$SV(S_i) = \sum_{\forall k \in S_i} KV(k) + w \cdot \left(\sum_{\forall k \in S_{i-1}} KV(k) + \sum_{\forall k \in S_{i+1}} KV(k) \right) \quad (1)$$

$$KV(k_i) = KV'(k_i) \cdot \sum_{\forall b \in keyword\ set} C(k_i, k_b) \quad (2)$$

$$KV'(k_i) = \sum_{j=1}^{N_{UserLog}} P(1 - dp \cdot j) \quad (3)$$

$$P = \begin{cases} 1 & if\ k_i\ was\ used\ in\ UserLog_j \\ 0 & otherwise \end{cases} \quad (4)$$

$$C(k_a, k_b) = C(A, B) = \frac{\sum AB - \frac{\sum A \sum B}{N}}{\sqrt{\left(\sum A^2 - \frac{(\sum A)^2}{N} \right) \left(\sum B^2 - \frac{(\sum B)^2}{N} \right)}} \quad (5)$$

위 수식에서 SV는 샷의 평가 결과인 샷 점수(shot value)를, KV는 키워드 점수(keyword value)를, 그리고 C는 두 키워드 간의 상관관계(correlation)를 의미한다. P는 특정 키워드가 사용자에게 선택 되었는지 아닌지를 의미하는 predicate이며, dp(decay parameter)는 시간이 지날수록 키워드의 중요도를 떨어뜨리는 역할을 하는 변수이다.

수식의 의미는 다음과 같다. 사용자가 과거에 많이 선택한 키워드는 이후에도 키워드로 선택할 가능성이 높으므로 높은 키워드 점수를 받게 된다. 샷 점수는 샷이 포함한 키워드 점수의 합이며, 컨텍스트의 고려를 위해 근접한 샷의 점수 또한 w(weight)만큼의 비율로 반영하도록 하였다. 사용자가 직접 선택한 키워드 이외의 다른 키워드는 선택된 키워드와의 상관관계에 따라 점수가 주어진다. 사용자 로그를 통해 선택 키워드와 사용 패턴이 유사하면 높은 상관관계를 갖게 되며, 키워드 점수 또한 높아지게 된다. 식 (5)에서 A, B는 두 키워드 k_a와 k_b가 로그에서 사용되었는지 여부에 따라 값이 결정되는 (사용되면 1, 사용되지 않으면 0) predicate의 벡터이다.

위와 같이 샷의 점수가 계산되면 높은 점수를 받은 샷을 선택하여, 요약이 가능하다. 샷의 길이가 긴 경우, 하나의 샷을 모두 요약에 사용할 수 없으므로 일부분의 프레임을 선택하는 과정이 필요하다. 본 논문에서는 샷의 시작이나 끝 보다는 가운데 부분이 특정 이벤트를 잘 반영할 가능성이 높다고 판단하여 가운데 부분의 프레임 집합을 선택하였으며, 이 부분은 향후 개선될 여지가 있을 것으로 판단된다.

4.2. 도메인 지식을 이용한 뷰 선택

이벤트 평가와 그에 따른 요약이 수행되면, 동일한 이벤트를 비추는 여러 대의 카메라 뷰 가운데 하나를 선택하는 뷰 선택 과정이 수행된다. 본 논문에서는 도메인 지식을 이용하여 뷰 선택을 수행하였다.

뷰 선택을 위한 도메인 지식은 사용자 서베이를 통해 얻은 지식을 포함한다. 도메인 지식을 바탕으로 여러 규칙을 구성하였으며, 규칙은 그림 3과 같이 표현된다. 이 규칙은 "사람 A가 2부분에 위치하며 Work이벤트를 수행하고 있을 때는 3번 카메라 뷰를 선택한다" 는 규칙을 표현한 것이다.

```
<owlr:Rule>
        <owlr:antecedent>
                <owlr:individualPropertyAtom owlr:property="locate">
                        <owlr:variable owlr:name="A">
                        <owlr:variable owlr:name="2">
                </owlr: individualPropertyAtom>
                <owlr:individualPropertyAtom owlr:property="happen">
                        <owlr:variable owlr:name="Work">
                </owlr: individualPropertyAtom>
        </owlr:antecedent>
        <owlr:consequent>
                <owlr:individualPropertyAtom owlr:property="view">
                        <owlr:variable owlr:name="3">
                </owlr: individualPropertyAtom>
        </owlr:consequent>
</owlr:Rule>
```

그림 3. 뷰 선택을 위한 규칙 예

5. 실험결과

5.1. 시나리오 및 데이터 수집

3.2에서 정의된 이벤트를 이용하여, 3명의 사용자가 세미나를 수행하는 상황을 중심으로 약 20분 분량의 시나리오를 설계하였고 그에 따라 데이터를 수집하였다. 그림 4는 A, B, C 세 사람이 시간에 따라 이벤트를 수행하는 시나리오를 보여주며, CA, CONV, EN, LE는 calling, conversation, entry, leaving의 약자이다.

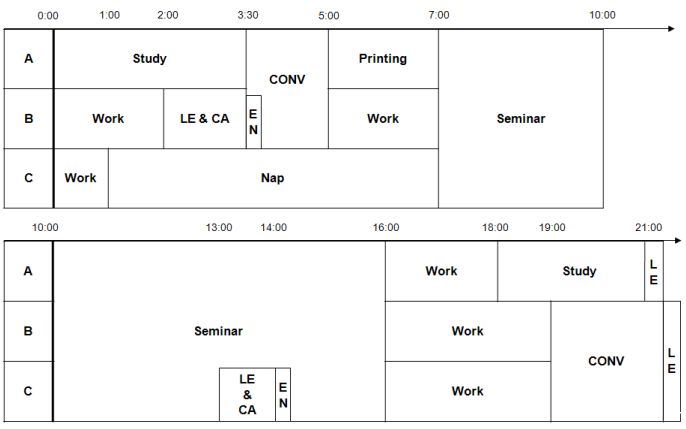

그림 4. 세미나 중심의 시나리오

위의 시나리오를 바탕으로 수집된 데이터를 어노테이션 한 결과 총 32개의 샷으로 나뉘어졌다.

5.2. 샷 평가 및 요약 결과

그림 5는 실험에 사용된 사용자 로그를 보여준다. 사용자가 키워드를 선택하고 요약을 수행하면 그 때마다 함께 사용한 키워드가 로그에 누적 기록된다. 실험을 위해 선택된 키워드는 "Conversation, Seminar"이며 dp는 0.05, w는 0.2가 사용되었다.

```
* 2008.04.01
Calling,Entry
Calling,Meeting,Seminar
Conversation,Meeting,Seminar
Calling,Conversation
Entry,Leaving,Study,Work
Conversation,Meeting
Calling,Entry,Leaving,Nap,Printing,Study,Vacuuming,Work
Meeting,Seminar,Study,Work
Calling,Conversation,Nap
Seminar
Meeting,Seminar
Calling,Seminar
* 2008.04.12
Calling,Entry,Meeting
Calling,Conversation,Nap,Vacuuming,Work
Calling,Leaving
```

그림 5. 사용자 로그

그림 5의 로그를 바탕으로 제안하는 방법을 통해 32개 샷을 평가한 결과 그림 6의 그래프를 얻었으며, 이 그래프의 가로, 세로축은 샷 번호와 샷 점수를 의미한다. 샷의 정의에 따라, 모든 샷의 길이가 동일하지 않으므로 시나리오에서 이벤트의 비율과 해당 샷의 비율은 다를 수 있다.

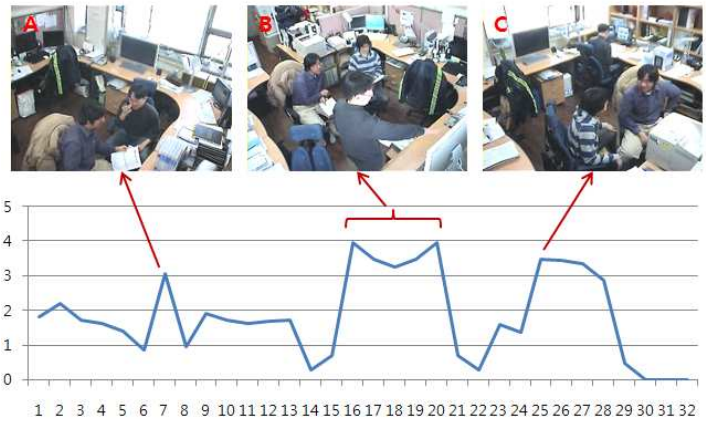

그림 6. 샷 평가 결과 및 높은 점수를 받은 샷

그림 6의 그래프에서 샷 점수가 높은 샷은 7번, 16~20번, 25~29번이며, 세 부분을 각각 A, B, C라고 하면 A, C는 Conversation, B는 Seminar에 해당하는 부분이다. 사용자가 선택한 키워드에 해당하는 부분이므로 높은 점수가 할당되는 것이 당연한 결과라고 할 수 있다. 선택 키워드에 해당되지 않는 나머지 이벤트에 대한 점수는 사용자 로그에서의 선택 키워드와 상관관계에 따른 키워드 점수를 바탕으로 계산된다. Conversation과 Seminar다음으로 키워드 점수가 높은 이벤트는 로그에서 두 키워드와의 상관관계가 높은 Meeting, Calling, Nap, Work 정도이다. Nap 이벤트는 2~13번 샷까지 연속해서 포함되어 있으며, Work 이벤트는 그림 4의 시나리오에서 보이듯이 전체적으로 많은 부분에 포함되어있다. Meeting이벤트는 수집한 데이터에 해당되는 부분이 없으며, Calling이벤트는 대부분 Leaving이벤트와 함께 발생해 밖으로 나가서 통화를 하는 내용이라 어노테이션에서 제외되었다. 14번이나 22번 샷과 같이 점수가 가장 낮은 부분은 서로 다른 이벤트가 연결되는 부분으로 이벤트가 정의되지 않아 Unknown으로 어노테이션 된 부분에 해당한다.

요약 결과는 평가된 샷 가운데 사용자가 선택한 기준 점수보다 높은 점수를 가진 샷으로 구성하되, Conversation이나 Seminar의 경우처럼 샷의 길이가 길어지는 경우 가운데 부분을 선택하도록 하였다.

5.3. 어플리케이션

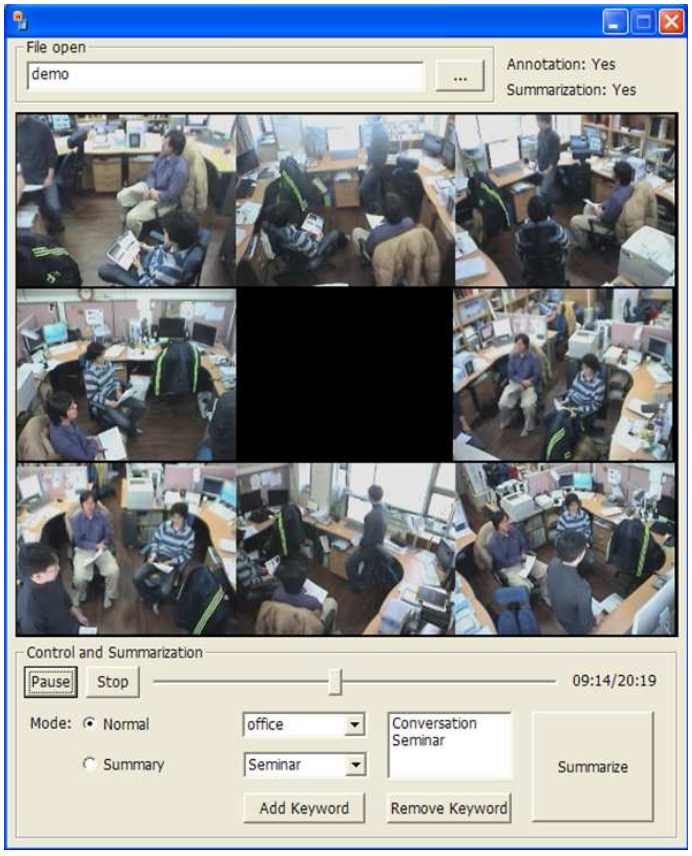

그림 7. 어플리케이션 스크린 샷 (재생모드)

수집된 비디오 데이터를 살펴보고, 원하는 이벤트 키워드를 추가하여 요약을 수행하기 위한 어플리케이션을 구현하였다. 그림 7은 8대의 카메라에 의해 수집된 데이터를 한 번에 보여주는 재생모드의 스크린 샷이며, 그림

8은 선택된 키워드에 따라 요약된 결과를 보여주는 스크린 샷이다. 요약 모드에서 요약을 수행하려면 먼저 파일을 연 후 해당 도메인을 선택하고 (현재의 도메인은 office), 현재 도메인 내에서 선택 가능한 키워드 가운데 관심이 있는 키워드를 선택하여 리스트에 추가하고 요약 버튼을 통해 요약을 수행하면 된다. 요약 버튼이 눌리면 현재 선택된 키워드가 사용자 로그에 추가되며, 제안하는 방법에 따른 샷 평가 및 요약 과정이 수행되게 된다. 그림 7은 Seminar이벤트를 그림 8은 Conversation이벤트에 해당하는 부분을 각각 보여준다.

그림 8. 어플리케이션 스크린 샷 (요약모드)

6. 결론 및 향후연구

본 논문은 동일한 이벤트에 대한 다양한 시점을 얻기 위해 오피스 환경에 다수의 카메라를 같은 장소를 향하도록 멀티 카메라 시스템을 구성하였고, 이로부터 수집한 비디오 로그 데이터를 대상으로 요약 및 뷰 선택을 수행하였다. 요약 과정에서의 샷 평가를 위해 사용자가 과거에 선택한 키워드 로그를 분석하여 활용하였으며, 도메인 지식을 이용하여 뷰 선택을 수행하였다. 실험 결과는 사용자 로그 분석을 통한 요약 결과가 잘 동작함을 보여주었다.

본 논문은 비디오 데이터에 대해 수동 어노테이션을 수행하고, 그 위에서 요약 및 뷰 선택을 수행하였다. 수동 어노테이션 부분은 장기적으로 영상처리 및 패턴인식 기법을 활용한 자동 어노테이션으로 대체되어야 하며, 어플리케이션 또한 키워드나 샷 점수 등 의 분석 결과까지 보여줄 수 있도록 개선될 필요가 있다. 이후 제안하는 방법과 어플리케이션의 검증을 위한 사용성 평가 또한 수행할 계획이다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT 연구센터 지원사업의 연구결과로 수행되었음 (IITA-2008-(C1090-0801-0046))

※ 참고문헌

[1] J. Gemmell, G. Bell and R. Lueder, "MyLifeBits: A personal database for everything," *Communications of the ACM*, vol. 49, no. 1, pp. 88-95, 2006.

[2] S.-B. Cho, K.-J. Kim, K.-S. Hwang and I.-J. Song, "AniDiary: Daily cartoon-style diary exploits Bayesian networks," *IEEE Pervasive Computing*, pp. 66-75, 2007.

[3] G. C. Silva, T. Yamasaki and K. Aizawa, "An interactive multimedia diary for the home," *IEEE Computer*, pp. 52-59, 2007.

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

[5] C. Zhang, S.-B. Cho and S. Fels, "MyView: Personalized event retrieval and video compositing from multi-camera video images," *Lecture Notes in Computer Science*, vol. 4557, pp. 549-558, 2007.

[6] N. Sebe, M. S. Leu and A. W. M. Smeulder, "Editorial introductoin: Video retrieval and summarization," *Computer Vision and Image Understanding*, vol. 92, pp. 141-146, 2003.

[7] B. L. Tseng, C.-Y. Lin, and J. R. Smith, "Using MPEG-7 and MPEG-21 for personalizing video, IEEE Multimedia, vol. 11, no. 1, pp. 42-53, 2004.

[8] A. Ekin, A. M. Tekalp and M. Mehrota, "Automatic soccer video analysis and summarization, *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796-807.

[9] H.-S. Park and S.-B. Cho, "Fuzzy rule-based summarization of event sequences in an indoor multi-camera environment, *Proceedings of the KIISE*, vol. 34, no. 2(C), pp. 288-292, 2007.

[10] 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, pp. 31-53, 2003.