비접촉 동작 인식 기기를 활용한 동영상 콘텐츠 검색 시스템

이승재 정다운 이근동 제성관 오원근

한국전자통신연구원 SW·콘텐츠연구소

{seungjee, crisia, zacurr, skj, owg}@etri.re.kr

Video Content Searching System Using Touchless Motion Sensing Device

Lee, Seungjae Jung, Da-Un Lee, Keundong Je, Sungkwan Oh, Weon Geun

SW·Contents Research Lab., ETRI

요약

본 논문에서는 비접촉 동작 인식 기기를 활용한 동영상 콘텐츠 검색 시스템을 소개한다. 제안하는 시스템은 스마트 TV와 같은 인터넷이 가능한 디스플레이의 동영상 콘텐츠를 대상으로 하며, 콘텐츠 자체 또는 콘텐츠 내의 객체에 대한 정보를 검색 결과로 제공한다. 시스템 구현에 앞서 영상 콘텐츠의 검색 요구 사항에 따라 시나리오 및 기능을 수립하고, 각각의 기능은 비접촉 동작 인식 기기인 립모션을 기반으로 손 동작을 지정하였다. 따라서, 제안하는 시스템은 사용자의 손가락 동작에 의해 선택된 영역을 모바일 비주얼 검색 서버에 전송하게 되고, 검색 결과는 모바일 기기를 통해 최종적으로 전달된다. 본 논문에서는 시스템을 실제로 구현하고 다양한 콘텐츠에 대하여 실험하였다. 개발된 시스템을 통해서 사용자는 손을 이용한 간단한 동작에 의해 콘텐츠 정보, 콘텐츠 내 객체의 정보를 실시간으로 모바일을 통해 제공받을 수 있다.

1. 서론

모바일 비주얼 검색(Mobile Visual Search) 기술은 무선 네트워크 기술의 발달과 함께 카메라 디바이스 및 스마트 폰이 보편화되면서 다양한 응용 분야로 확대되고 있다 [1]. 모바일 비주얼 검색은 영상에서 환경 및 시점 변화에 강인한 특징점, 특징을 추출하고 검색하는 기술로 증강 현실, 로봇 비전, 영상 감시 및 자동차, TV 등의 다양한 분야에 서비스를 확장하고 있다.

모바일 폰이 TV와 함께 세컨드 스크린을 위한 멀티 디스플레이로 이용됨에 따라 검색 환경에 대한 관심이 높아지기 시작했다. 많은 개발자들은 자유 동작 센서를 탑재하거나 인터랙션 디자인을 개선함으로써 좀 더 자유로운 사용자 인터페이스를 위한 연구를 진행하고 있다 [2-4]. 그러나, TV 플랫폼의 가격 대비 낮은 성능과 사용자와 디스플레이 사이의 거리에 의해 발생되는 제약된 인터페이스는 사용자에게 한정된 콘텐츠를 제공할 수 밖에 없는 주요한 결점이다. 그래서, 사용자 동작 중심의 리모트 컨트롤이 연구되었지만, 여전히 사용자의 요구를 충족하지 못하고 있다.

최근 비접촉 센서들이 연달아 출시되면서 높은 인식률은 물론 다양한 프로그램 및 앱을 제공하고 있다. 비접촉 동작 인식 기기는 대표적으로 Microsoft의 카넥트(KINECT) [5]와 립모션(Leap Motion) [6]이 있다. 특히, 립모션은 손, 손가락 혹은 손가락과 같은 도구에 대해 위치와 제스처, 움직임을 인식하는 기기로 적외선 카메라를 기반으로 디바이스 위에서 25 ~ 600mm 범위 내에서 동작한다. 립모션은 이러한 정보를 프레임(Frame)이라는 객체를 통해서 손에 대한 동작 정보를 id, hands, fingers, tools gesture 등으로 제공하며, 이를 기반으로 상호 작용이 가능한 다양한 서비스 시나리오 및 프로그램 앱을 구성할 수 있다.

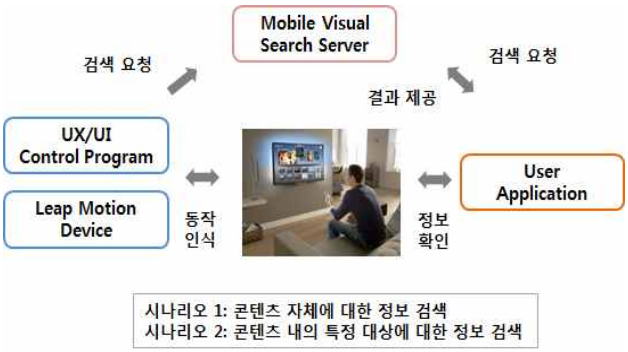

그림 1. 비접촉 동작인식 기기를 활용한 동영상 콘텐츠 검색 및 시나리오

본 논문에서는 립모션을 활용한 동영상 콘텐츠 검색 시스템을 구현하였다. 개발된 시스템에서 사용자는 손을 이용한 간단한 동작에 의해 동영상의 정보 및 동영상 내의 상품에 대한 정보를 검색하고, 검색 결과는 모바일 디스플레이에서 실시간으로 제공받을 수 있다.

2. 시스템 구성 및 동작 시나리오

제안하는 시스템은 그림 1에서 보는 바와 같이 'UX/UI 제어 프로그램', '모바일 비주얼 검색 서버', 그리고 정보 확인을 위한 '안드로이드 앱' 세 가지로 크게 구성된다. UX/UI 제어 프로그램이 동작을 인식하면, 모바일 검색 서버에 검색을 요청을 하고 DB로부터 찾아낸 검색 결과는 모바일 앱으로 즉시 전송하게 된다. 요청된 콘텐츠 검색을 위해 사용되는 특징은 LoG(Laplacian of Gaussian)에 기반한 SIFT(Scale Invariant Feature Transform) [7]를 활용하였으며, 검색 구조는 VLAD [8]를 변형한 자체 구조를 활용하였다.

검색의 과정은 사용자가 거실에서 TV 혹은 동영상을 보고 있는 상황에서 스마트폰 혹은 태블릿을 가지고 직접 검색 하여 정보를 확인하거나, 동작 인식 프로그램의 제어를 통해서 콘텐츠를 검색하고 모바일 앱을 통해서 검색 결과를 확인하는 과정으로 나누어질 수 있다. 이 때, 검색의 대상에 따라 시나리오는 두 가지로 구성된다. 첫 번째 시나리오(시나리오 1)는 콘텐츠 자체의 정보를 검색하는 것으로 주로 Automatic Content Recognition (ACR)라는 콘텐츠 식별 기술을 활용하며, 모바일 비주얼 검색을 통해서도 검색이 가능하다. 이 경우, 화면을 스마트폰 혹은 태블릿으로 직접 촬영하거나 현재 재생 중인 영상의 화면을 동작 인식 프로그램의 제어를 통해서 검색을 위해 전송할 수 있다. 두 번째 시나리오(시나리오 2)는 콘텐츠 내의 특정 대상을 검색하는 경우로 사용자가 원하는 특정 대상에 대해서 특징을 추출하기 위해 영역을 선택하는 과정이 필요하다. 이를 위해 동작 인식 프로그램을 통해서 화면 내의 특정 객체를 선택하고, 이를 검색을 위해 서버로 전송 한다.

이러한 두 가지 시나리오를 만족시키기 위한 사용자 인터페이스는 립모션 기기에서 제공되는 다양한 정보 중에서 식별이 쉽고, 기능과의 연관성을 가지도록 손가락의 개수를 기반으로 기능을 정의하였다. 표 1은 검색을 위해 구분된 단계 및 각 단계에서의 명칭과 기능을 설명하고 있다. 시나리오 1은 '검색 대기'를 통해 검색 준비가 된 상태에서 '검색 전송'으로 전환하여 현재 화면을 전송하여 영상 콘텐츠를 검색한다. 시나리오 2는 '검색 대기'에서 검색하고자 하는 대상에 대해서 '영역 선택' 모드로 전환 후 영역을 선택 후에 '영역 선택 종료' 후에 해당 영역을 서버로 전송한다.

표 1. 검색 단계 및 기능 정의

검색 단계	기능 정의
검색 대기	조건: 손가락 개수가 0에서 1로 변화 시 역할: 검색을 위한 대기 모드
검색 전송	조건: 손가락 개수가 1에서 4개 이상으로 변화 시 역할: 선택된 영역의 이미지 전송 선택된 이미지가 없을 경우 전체 이미지 전송
영역 선택	조건: 손가락이 1초 이상 머무르는 경우 역할: 영역 선택 모드로(원형 혹은 네모) 전환
영역 선택 종료	조건: 영역 선택 후 손가락을 1초 이상 머무르는 경우 역할: 영역 선택을 종료하고 전송 여부를 확인
검색 종료	조건: 손가락 개수가 0인 경우 역할: 현재 선택된 것 모두 취소

3. 시스템 구현 및 결과

제안하는 시스템을 구성하는 UX/UI 제어 프로그램은 립모션 SDK를 기반으로 개발되었고, 검색을 위한 '모바일 비주얼 검색 서버' 와 '검색 결과 확인용 안드로이드 앱'은 각각 C++ 언어와 Java를 통해 개발 되었다.

실험을 위해서 의약품, 침구 청소기, 자동차 등 각종 생활 물품 28종과 상품 광고 2개 및 TV 프로그램 1개를 실험 동영상으로 선정하고 각 항목에 대해서 영상 DB를 구축하였다. 이를 기반으로 앞 절에서 설명한 시나리오를 위한 검색 시스템을 구성하였다.

그림 2. 제안 방법으로 개발된 시스템의 실험 결과 (a) 립모션을 통한 검색 대상 영역 선택 (b) 손가락 개수 변화를 통한 검색 영역 전송 (c) 안드로이드 앱을 통한 검색 결과 확인

그림 3. 개발된 시스템을 통한 영상 콘텐츠 검색 서비스 방법: (a) 화면 전체 검색(시나리오 1) (b) 관심 영역 선택 및 검색(시나리오 2)

그림 2는 최종적으로 구현된 시스템에 대한 실험 결과를 시스템을 구성하는 요소 별로 보여준다. 개발된 시스템을 통해서 사용자는 손 가락을 이용한 간단한 동작에 의해 콘텐츠 자체 또는 콘텐츠 내 객체를 직접 선택하여 전송할 수 있으며, 검색 결과는 안드로이드 앱을 통해 실시간으로 확인할 수 있다.

그림 3은 두 개의 시나리오에 대한 실험 결과를 보여준다. 그림 3 (a)는 시나리오 1에 대한 실험 결과이며, 콘텐츠 자체에 대한 정보 검색을 시도하는 결과를 보여준다. 사용자는 화면에 보여지고 있는 전체 영상에 대해 손가락 하나를 펼친 "검색 대기"상태에서 "검색전송"을

위해 연속으로 남은 손가락을 동시에 펼쳐 이미지를 전송한다. 그림3 (b)는 시나리오 2에 대한 실험 결과이며 원하는 상품의 영역을 선택해 검색하는 결과이다. 사용자는 화면에서 관심 있는 특정 상품에 대해 “검색 대기”에서 1초 이상이 지난 후 “영역 선택” 상태로 전환되어 타원 또는 직사각형 형태의 경계선으로 상품을 선택한다. 선택 도중 손가락을 1초 이상 멈추어 “영역 선택 종료”상태로 전환되고, 시나리오 1과 같이 손 가락 하나를 펼친 “검색 대기”상태에서 연속으로 남은 손가락 동시에 펼쳐 “검색 전송”으로 전환시키면, 선택된 관심 영역 이미지를 검색 서버로 전송하게 된다.

결과적으로, 제안된 시스템은 손가락을 이용한 간단한 동작에 의해 실시간으로 검색할 수 있음을 확인할 수 있었다. 하지만, 실험 과정에서 사용자별로 립모션 기기에 대한 트레이닝이 필요함에 따라 차후 다양한 사용자가 이용할 수 있도록 추가적인 연구가 필요하다.

4. 결론

본 논문에서는 비접촉 센서 기반의 손 동작 인식 기기와 모바일 비주얼 검색 기술을 응용한 동영상 콘텐츠 검색 시스템을 구현하고 사용자 시나리오에 대해서 확인하였다. 제안된 시스템은 스마트 디바이스의 동영상 콘텐츠를 대상으로 하며 콘텐츠 자체 또는 콘텐츠 내의 객체에 대한 부가적인 정보를 사용자가 모바일을 통해 즉시 검색결과를 확인하는 것을 목적으로 하고 있다. 본 논문에서는 실제 구현을 통해 사용자의 검색 시나리오에 따라 동작이 가능함을 확인하였다. 그러나, 실험 과정에서 사용자에 따라 립모션 기기에 대한 별도의 적응 훈련이 필요함에 따라 실제 서비스를 위해서는 본 논문에서 접근한 방식 외에 좀 더 사용자 친화적이고 직관적인 동작 인식 방법이 필요할 것으로 예상된다.

감사의 글

본 연구는 미래창조과학부 첨단융복합콘텐츠기술개발의 일환으로 수행하였음. [R2012030111, UVD(Unified Visual Descriptor) 기반 Smart Mobile Search 기술 개발]

참고 문헌

1. 이승재, 이근동, 나상일, 제성관, 정다운, 오원근, 서영호, 손욱호, “모바일 비주얼 검색: 기술과 표준화 동향”, 전자통신동향분석, vol. 39, no. 1, 2014년 2월.
2. W. T. Freeman and C. D. Weissman, "Television control by hand gestures," in *Proceeding of IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 179-183, Jun 1995.
3. Soonmook Jeong, Jungdong Jin, Taehoun Song, Keyho Kwon, and Jae Wook Jeon,“Single-camera dedicated television control system using gesture drawing,” *IEEE Transactions on Consumer Electronics*, vol. 58, no. 4, pp. 1129-1137, Nov 2012.
4. S. Lian, W. Hu, and K. Wang, “Automatic User State Recognition for Hand Gesture Based Low-Cost Television Control System,” *IEEE Transactions on Consumer Electronics*, vol.60, no. 1, pp 107-115, Feb 2014.
5. KINECT http://www.microsoft.com/en-us/kinectforwindows/
6. Leap Motion https://www.leapmotion.com/
7. D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International J. Comput. Vision*, vol. 60, 2004, pp. 91-110.
8. H. Jegou et al., “Aggregating local descriptors into a compact image representation,” *Proc. IEEE Conf. Comput. Vision Pattern Recognition(CVPR),* San Francisco, CA, June 2010.