http://dx.doi.org/10.14400/JDC.2014.12.3.255

스마트 TV 환경에서 키넥트 센서를 이용한 사진 검색 시스템

최주철
경희대학교 창업보육센터

Photo Retrieval System using Kinect Sensor in Smart TV Environment

Choi Ju Choel
Dept. of Business Incubator Kyung Hee University

요 약 디지털 카메라, 스마트폰, 타블렛과 같은 스마트 기기의 대중화와 소셜 네트워크 서비스를 통해서 사진과 같은 멀티미디어 데이터의 양이 빠르고, 급격하게 확산되고 있다. 사진 검색 방법은 키워드 기반의 검색 방법, 예제 기반의 검색 방법, 시각화 질의 기반의 검색 방법의 세 가지 분류될 수 있다. 이전에 연구된 사진 검색 기법은 일반 PC 환경에 최적화되었기 때문에 최근에 등장한 스마트 TV 환경에서 사진 검색하기 위한 방법으로 사용하는 것은 적합하지 않은 상황이다. 본 논문에서는 스마트 TV 환경에서 키넥트를 이용한 소셜 네트워크에 존재하는 사진 검색 시스템을 제안하였다. 이를 위해서 키넥트 센서를 사용하여 마우스의 컨트롤을 제어할 수 있도록 구현하였으며, 제안하는 시스템의 검색 결과는 임계값이 0.7일 때, 평균 재현율과 평균 정확도는 각각 81%, 80%의 성능을 보였다.

주제어 : 멀티미디어 검색, 소셜 네트워크 서비스, 시각화 질의, 스마트 TV, 키넥트 센서

Abstract Advances of digital device technology such as digital cameras, smart phones and tablets, provide convenience way for people to take pictures during his/her life. Photo data is being spread rapidly throughout the social network, causing the excessive amount of data available on the internet. Photo retrieval is categorized into three types, which are: keyword-based search, example-based search, visualize query-based search. The commonly used multimedia search methods which are implemented on Smart TV are adapting the previous methods that were optimized for PC environment. That causes some features of the method becoming irrelevant to be implemented on Smart TV. This paper proposes a novel Visual Query-based Photo Retrieval Method in Smart TV Environment using a motion sensing input device known as Kinect Sensor. We detected hand gestures using kinect sensor and used the information to mimic the control function of a mouse. The average precision and recall of the proposed system are 81% and 80%, respectively, with threshold value was set to 0.7

Key Words : Multimedia Search, Social Networking Service, Visual Query, Smart TV, Kinect Sensor

1. 서론

최근 디지털 카메라, 스마트 폰, 태블릿과 같은 디지털 기기 대중화로 인해서 사진, 비디오 등의 멀티미디어 데이터의 양이 급격하게 증가하고 있으며, 이러한 현상은 페이스북(facebook)[1], 인스타그램(instagram)[2], 플러커(flickr)[3]와 같은 소셜 네트워크 서비스(social network service)를 통해 보다 빠르게 확산되고 있다. 소

Received 4 February 2014, Revised 6 March 2014
Accepted 20 March 2014
Corresponding Author: Choi Ju Choel
(Dept. of Business Incubator Kyung Hee University)
Email: choijc@khu.ac.kr
ISSN: 1738-1916

셜 네트워크를 이용하는 사용자들은 자신의 사소한 일상부터 특별한 기념일까지 이르는 모든 일상을 사진 또는 동영상의 형태로 촬영하고, 소셜 네트워크를 통해서 지인들과 공유하고 있다. 또한 소셜 네트워크를 이용하는 사용자들은 자신의 멀티미디어 데이터를 공유하는 것뿐만 아니라 지인들이 공유한 멀티미디어 데이터에 자신의 의견 또는 태그 추가하는 행위를 일상적으로 하고 있다 [4, 5].

소셜 네트워크에서 멀티미디어 정보를 공유하거나 의견을 또는 태그를 추가하는 일들로 인해서 개인이 멀티미디어를 접하는 양과 시간이 폭발적으로 증가하기 때문에 자신이 찾고자 하는 이전에 봤던 사진 검색을 위해서 해당 사이트에 방문하여 찾는 것은 점점 더 어려워지고 있다 [6, 5, 7].

이러한 검색의 문제를 해결하는 방법에 대한 많은 연구들이 진행되었으며[6, 5, 7, 8, 9], 특히 사진과 같은 멀티미디어 데이터의 검색의 성능을 향상시키기 위해서는 질의 표현 (Query Formulation) 측면과 질의 매칭 (Query Matching) 측면을 고려한 연구들이 진행되었다 [8]. 질의 표현은 사용의 의도를 정확하게 해석하는 것이 검색 성능을 향상할 수 있다는 것이고, 질의 매칭은 사용자의 질의에 맞는 내용을 데이터베이스에서 정확하게 찾는 것이 검색의 성능을 향상 시킬 수 있다는 것이다. 그러나 질의 매칭은 질의를 어떻게 표현하는 지에 따라서 성능이 달라질 수 있기 때문에 보다 정확한 질의를 사용자가 시스템에게 표현할 수 있도록 제공하는 것이 검색의 성능을 향상시키는 중요한 요소라고 볼 수 있다.

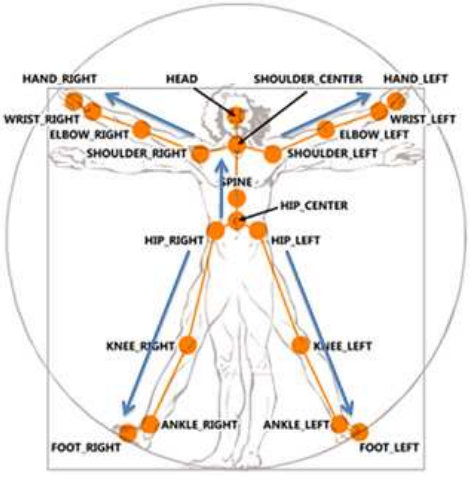
이러한 검색의 방법을 최근에 등장한 스마트 TV 환경에 직접적으로 적용하기에는 어려운 점이 존재한다. 지금까지의 연구는 일반 PC 환경에서 멀티미디어 데이터를 검색하는 방법에 대한 연구가 진행되었으나, 스마트 TV 환경에서는 새로운 방법의 질의 생성 표현 방법이 제공되어야 할 것이다. 예를 들어, 일반 스마트 TV 리모트 컨트롤(remote control)을 이용하여 사진을 검색하는 것은 어려운 일이다 [10, 11]. 즉, 자신이 어디선가 봤던 사진들을 단순한 마우스 기능과 키보드 입력 기능을 제공하는 환경에서 검색하는 것은 사용자들로 하여금 매우 많은 시간을 소비하게 만든다.

따라서, 본 논문에서는 스마트 TV에서 사진 검색을 위해 사용자가 자신이 찾고자 하는 질의를 시스템에게 정확하게 표현할 수 있는 시각화 질의 방법 및 시스템을 제안하고자 한다. 이를 위해서 마이크로소프트의 키넥트 센서(kinect sensor)를 이용하여 사용자가 질의를 생성할 수 있는 인터페이스를 제공하여, 사용자는 자신의 페이스북 사진 첩과 지인들의 사진첩에서 사진 검색할 수 있다.

본 논문의 구성은 2장에서는 관련연구에 대해서 소개를 하고, 3장에서는 제안하는 시스템 구조에 대해서, 그리고 4장에서는 제안하는 시스템의 프로토타입에 대해서설명하고 마지막으로 결론 및 향후 연구로 구성되어 있다.

2. 관련연구

2.1 모션 인식



[Fig. 1] Skeletal Information in Kinect Sensor

현재의 제스처 인식(gesture recognition) 기술은 사용자 신체의 움직임을 인식하여 컴퓨터와 상호작용하는 기술이다. 마이크로소프트사의 키넥트 센서는 신체 골격(skeleton)을 인식하여 각 관절(joint)의 정보를 이용하여 상호 작용을 할 수 있는 SDK 라이브러리[12]를 제공하고 있다. 키넥트 센서는 검출 된 사용자의 골격의 관절에 해당하는 [Fig. 1]와 같은 20개의 거리 정보와 위치 좌표를 제공 해준다.

사람의 골격 정보를 기반으로 제스처 기반 상호작용을 제공하기 때문에 정밀한 인식 성능은 제공하지 못하지만, 다양한 포즈에 대한 인식과 손에 대한 추적에 대

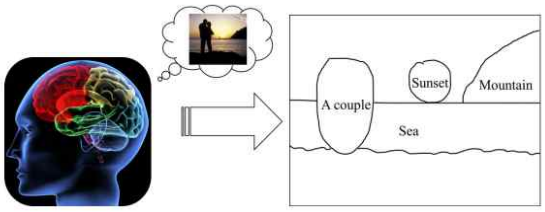
한 기능은 1.7 이후 버전부터 지원하고 있다. 정교한 손 동작 인식과 같은 기능의 구현을 위해서는 키넥트에서 인식한 손 위치에 대한 정보를 바탕으로 OpenCV[13]와 같은 비전 라이브러리를 이용하여 손동작 인식을 수행할 수 있다. 키넥트 센서는 총 2개의 깊이 감지 센서와 1개의 640 X 480의 해상도를 갖는 컬러(RGB) 카메라, 4개의 마이크로폰 어레이로 구성 되어 있다. 동작 깊이 인식 범위는 1.2m - 3.5m 이다. 실제 구동을 하면 2차원 평면이 아닌 X, Y좌표와 거리를 알 수 있는 Z좌표 전부를 인식하여 3차원으로 인식 하게 된다.

2.2 이미지 검색

이미지 검색 기술은 세 가지 형태로 분류하여 연구되어져 왔다. 첫 번째는 텍스트 기반의 이미지 검색(text-based image retrieval)이다. 구글이나 빙(bing)과 같은 포털 사이트에서 키워드를 이용하여 자신이 찾고자 하는 이미지를 검색하는 방법이다. 이 방법은 구현하기 쉬운 장점이 있으나, 실제 이미지 콘텐츠의 내용을 사용자가 직접 설명할 수 없는 단점이 존재하며, 이러한 단점을 해결하기 위해서는 사용자들이 콘텐츠의 내용에 대한 협업적인 주석 처리를 통해서 극복될 수 있다.

두 번째는, 예제 기반의 검색(example-based image retrieval) 방법이다. 사용자는 자신이 찾기를 희망하는 이미지를 질의로 사용하게 되며, 사용자가 질의한 유사한 이미지의 사진들을 검색 할 수 있게 된다. 이러한 검색의 방법은 질의에 사용된 이미지와 동일하거나 유사한 이미지를 검색 하는 데에 있어서 매우 유용하지만, 질의를 하기위해서 이미지를 찾는 어려움이 존재한다.

세 번째는, 시각화 질의 기반의 이미지 검색(visual query-based image retrieval) 방법이다. 이 방법은 위에서 설명했던 텍스트 기반의 검색과 예제 기반의 검색의 장점을 모두 활용한 방법으로, 사용자는 자신이 찾고자 하는 질의에 해당하는 예제를 직접 만들 수 있게 된다. 이러한 시각화 질의는 키워드 기반의 이미지 검색에서 처리될 수 있는 형태로 변환되고, 각 객체들의 위치도 함께 표현된다. 이렇게 표현된 시각화 질의는 사용자가 이미지 안에 있는 객체들의 위치 관계에 대한 설명을 정확하고 쉽게 표현 할 수 있기 때문에 키워드 검색의 성능보다 향상된 결과를 얻을 수 있게 된다. [Fig. 2]는 시각화 질의에 대한 개념과 예시이다.

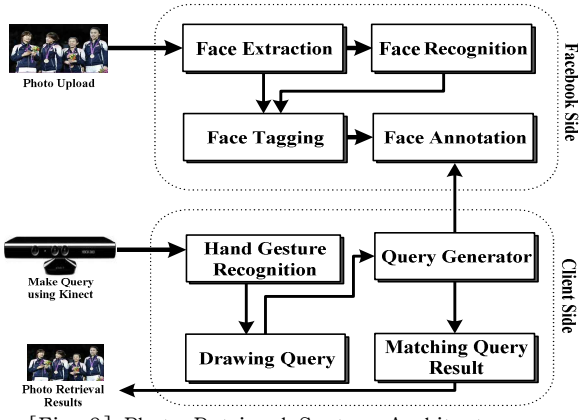


(a) Example of Visual Query

(b) Comparison result

[Fig. 2] Example of Visual Query and Results[8]

3. 시스템 구조도

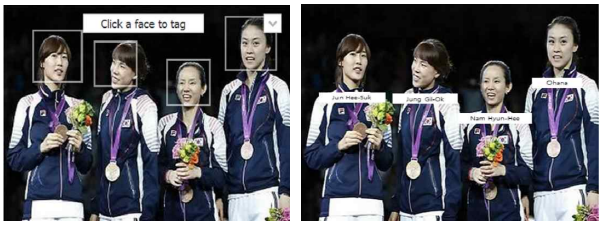


[Fig. 3] Photo Retrieval System Architecture

[Fig. 3]은 본 논문에서 제안하는 스마트 TV 환경에서 시각화 질의 기반의 사진 검색을 위한 시스템이다. 스마트 TV 환경에서 사용자는 자신이 찾고자 사진 속에 포함된 객체들에 대한 정보를 단순한 키워드가 아닌, 시각화된 질의(사용자의 얼굴 위치)를 통해서 키워드(이름)와 위치관계(얼굴 위치)에 대한 질의를 생성하여 시스템에게 정확하게 표현할 수 있게 된다. 제안하는 시스템은 Facebook side와 Client side의 부분으로 구성되었다.

Facebook side에서는 사용자가 사진을 업로드(upload)하는 과정에 대한 것으로, 페이스북에서 사용자

가 사진을 업로드하게 되면, [Fig. 4]의 (a)와 같이 얼굴 영역이 추출되며, 추출된 해당 얼굴에 대한 태그를 추가할 수 있으며, [Fig. 4]의 (b)는 태그가 추가된 결과 화면이다. 본 논문에서의 가정은 페이스북의 사용자들은 자신들의 페이스북 페이지 뿐 만 아니라 자신들의 지인들의 페이지에 방문하여 공유된 사진에서 자신의 얼굴 또는 지인들의 얼굴에 태그를 추가하는 행위를 할 수 있다는 것을 전제로 하고 있다.



(a) Face Extraction (b) Tagged Name

[Fig. 4] Photo Tagging Information on Facebook

이렇게 추가된 얼굴 태그 정보는 페이스북의 API를 이용하여 [Fig. 5]과 같은 JASON 형태로 표현되는 것을 확인 할 수 있다.


```
{
  "id": "606913969338016",
  "tags": {
    "data": [
      {
        "name": "Ohana",
        "x": 85.843373493976,
        "y": 18.345368715084
      },
      {
        "name": "Nam Hyun-Hee",
        "x": 63.855421686747,
        "y": 36.435754189944,
        "id": "1038102463241 75"
      },
      {
        "name": "Jung Gil-Ok",
        "x": 38.453815261045,
        "y": 27.586592178771,
        "id": "107502599279168"
      },
      {
        "name": "Jun Hee-Suk",
        "x": 18.273092369478,
        "y": 24.782122905028
      }
    ],
    "paging": {
      "next": "https://graph.facebook.com/606913969338016/tags?f
    }
  }
}
```


[Fig. 5] Tagging information presented by Jason

Client side는 스마트 TV에 연결된 키넥트 센서를 이용하여 제스추어를 인식하게 된다. 등록된 제스추어를 인식하게되면 드로잉을 할 수 있게 된다. 이렇게 만들어진 얼굴의 위치를 이용하여 비주얼 쿼리를 작성하게 된다. 이렇게 작성된 쿼리를 이용하여 페이스북의 데이터와 비교하여 사용자의 쿼리에 만족하는 결과를 사용자에게 제공한다.

사용자가 작성한 시각화된 질의를 통해서 페이스북에 있는 사진들을 검색하기 위해 매칭 전략과 랭킹 전략이 필요하다 [5]. 우선, 사용자가 작성한 질의를 Q라고 하고 페이스북에 있는 i번째 사진을 Pi라고 하면, Q와 Pi의 유사도를 구하기 위한 수식은 식(1)과 같다. Q, Pi의 벡터모델은 얼굴의 영역, 위치, 이름에 대한 정보가 있다.

$$Sim(Q,P_j) = w(qa,fa) \times \frac{Q \bullet P_j}{|Q||P_j|} \quad (1)$$

*w(qa, fa)*를 계산하기 위해서는 식(2)와 같이 두 가지 경우를 고려해야한다.

$$w(qa,fa) = \begin{cases} p(qa,fa) & if\ ia > 0 \\ \dfrac{1}{d(qa,fa)} & otherwise \end{cases} \quad (2)$$

첫째 경우는, *w(qa, fa)*에 인터섹션이 존재하는 경우에는 인터섹션이된 영역의 크기가 클수록 우선순위가 높아야되며, 이를 위해 식(3)을 이용해서 계산을 한다.

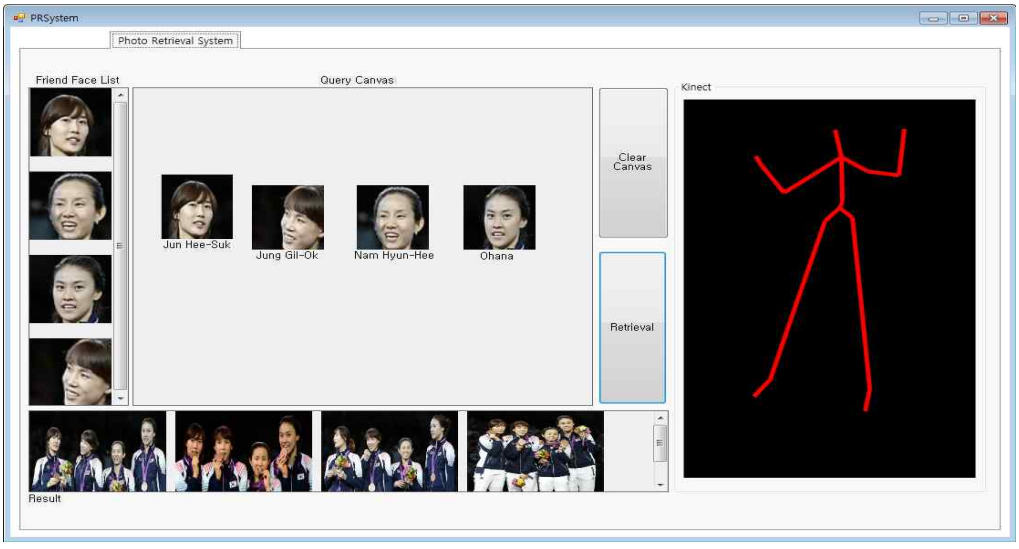
$$p(qa,fa) = \frac{ia}{fa} \times 100 \quad (3)$$

여기서, *w(qa,fa)*는 사진에 존재하는 얼굴의 영역의 위치(*qa*)와 사용자가 작성한 질의 영역(*fa*)과의 유사도를 계산하는 것이다. *p(qa,fa)*는 두 개의 영역에 인터섹션의 영역(*ia*)을 의미한다.

두 번째 경우는, 인터섹션이 발생하지 않은 경우로써, 이때에는 유클리디안 거리 공식을 이용하여 2개의 영역에 대한 유사도를 식(4)를 이용하여 계산할 수 있다.

$$d(qa,fa) = \sqrt{(fc_x \quad qc_x)^2 + (fc_y \quad qc_y)^2} \quad (4)$$

여기서 *fc*는 *fa* 영역에 대한 중심점을 의미하고, *qc*는 *qa*의 영역의 중심점을 의미한다.



[Fig. 6] Photo Retrieval System Interface

4. 실험 및 구현

본 논문에서 사용자의 의도를 정확하게 표현할 수 있는 시각화 질의 시스템을 위해 MS의 키넥트 센서를 이용하였으며, [Fig. 6]과 같은 인터페이스를 위해서 C#을 이용하여 구현하였다. 또한, 사용자의 입력을 보다 효과적으로 처리하기 위해서 [Fig. 7]과 같은 손동작을 인식하여 마우스 다운 이벤트와 마우스 업 이벤트를 구현하였다. 손동작 인식을 위한 단계로는 손영역 추출 및 이진 영상의 단계와 손 중심 모멘트와 손끝좌표 검출의 단계를 OpenCV를 이용하여 구현하였으며, 약 95% 이상의 인식률을 보였다.

Type		
Mouse Event	Mouse Down	Mouse Up

[Fig. 7] Hand Gestures Type for Mouse Control

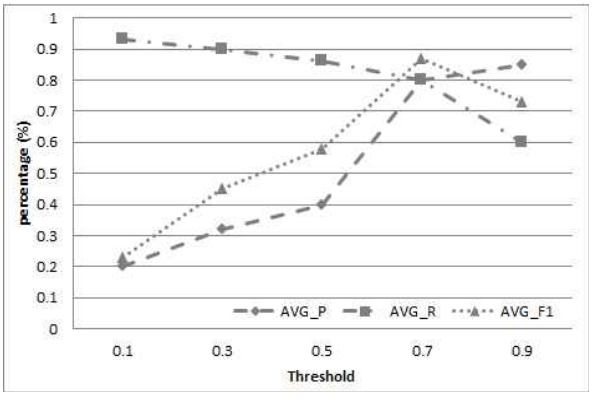
- 손영역 추출 및 이진 영상: 본 논문의 실험을 위해서는 조명과 배경은 변하지 않는 것을 가정한다. 배경은 손 영역을 쉽게 추출하기 위해서 프로그램이 시작되는 시점에 캡쳐하여 등록한다. 키넥트 센서를 통해서 오른쪽 손의 위치를 인식하고, 해당 영역에서 100 X 100의 영역을 추출한다. 추출된 영역과 등록된 배경의 차영상을 구한다. 구분된 차영상은 마스크로 활용한다. 원 RGB 영상에서 손의 영역만 남기고, 나머지는 흰색으로 처리하여 손 영역만 추출 한다. RGB 공간에서 이진화는 피부영역을 추출하는데 매우 유용한 방법이다.
- 손끝 점 검출을 위해서는 곡률(curvature)정보를 이용한 방법으로 손가락 끝에 해당하는 부분은 손의 여러 부분에 비해 특별한 각도를 이루고 있기 때문에 정확성이 뛰어나다. 손가락 각도 0~200도 -0.4와 1의 사이 값으로 설정한다. 곡률기반 알고리즘은 검출된 손 영역의 윤곽선을 일정한 간격으로 나눈 후 이어진 세 점의 사잇각을 계산하여 손 끝의 특징 점을 검출한다.

마우스의 더블 클릭은 마우스 다운과 마우스 업의 이벤트의 조합, 마우스 드래그 기능은 키넥트 센서와 마우스 다운 이벤트의 조합으로 구현 하였다.

검색 성능을 평가하기 위해서 평균 100명의 지인들이 있는 20명의 참가자를 대상으로 제안한 시스템을 이용하여 5개의 사진을 검색하는 실험을 진행하였다.

[Fig. 8]은 유사도의 임계값 변화에 따른 검색 결과의 평균 정확도(AVG_P), 평균 재현율(AVG_R), 그리고 평균 F1-measure의 값(AVG_F1)의 변화를 그래프로 나타낸 것이다. 일반적으로 유사도의 임계값이 높아질수

록 평균 정확도는 올라갔으나, 평균 재현율은 낮아지게 되므로, 실험 결과 유사도의 임계값이 0.7일때 F1-measure에 의한 결과가 가장 우수했다. 따라서 본 논문에서 유사한 사진의 검색을 위해 유사도의 임계값을 0.7로 설정을 하였다.



[Fig. 8] Recall, Precision and F1-measure results per threshold

5. 결론

본 논문에서는 스마트 TV 환경에서 키넥트를 이용한 소셜 네트워크에 존재하는 사진 검색 시스템을 제안하였다. 이를 위해서 키넥트 센서를 사용하여 마우스의 컨트롤을 제어할 수 있도록 구현하였으며, 마우스 컨트롤은 마우스 다운(mouse down), 마우스 업(mouse up), 이를 조합한 마우스 클릭과 더블클릭에 대한 기능을 제공하였다. 손 모양의 인식은 약 95% 이상의 정확도를 보였으며, 제안하는 시스템의 검색 결과는 임계값이 0.7일 때, 평균 재현율과 평균 정확도는 각각 81%, 80%의 성능을 보였다.

향후 연구로는 보다 다양한 형태의 검색이 가능한 UI/UX에 대한 연구와 사용자 맞춤형 질의 학습 방법에 대한 연구가 필요하다.

REFERENCES

[1] Facebook - http://www.facebook.com/
[2] Instagram - http://instagram.com/
[3] flickr - http://www.flickr.com/
[4] M. Crampes, J. Oliveira-Kumar, S. Ranwez, J. Villerd, Visualizing social photos on a Hasse Diagram for eliciting relations and indexing new photos, IEEE Trans. Visualization and Computer Graphics, pp. 985-992, 2009.
[5] K.-S. Lee, J.-G. Jung, K.-J. Oh, G.-S. Jo, U2Mind: visual semantic relationships query for retrieving photos in social network, Proceedings of the Third international conference on Intelligent information and database systems, pp. 20-22, 2011.
[6] Y.-H. Lei, Y-Y. Chen, B-C.Chen L.Iida, W.H. Hsu, Where is who: Large-scale photo retrieval by facial attributes and canvas layout. In ACM SIGIR, pp. 701-710. 2012.
[7] H.-N. Kim, A. E. Saddik, K.-S. Lee, Y.-H. Lee, G.-S. Jo, Photo search in a personal photo diary by drawing face position with people tagging, Proceedings of the 16th international conference on Intelligent user interfaces, pp. 13-16, 2011.
[8] C. Wang, Z. Li, L. Zhang, MindFinder: Image search by interactive sketching and tagging. 19th international conference on World wide web, pp. 1309-1312, 2010.
[9] H. Xu, J. Wang, X. S. Hua, S. Li, Image search by concept map, 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 275-282, 2010.
[10] C. J. Choel, S.-B. Park, Social Photo Retrieval and Its Application in Smart TV, International Conference on Information Science and Applications, pp. 1-2, 2013.
[11] W.-P. Lee, C. Kaoli, J.Y. Huang, A smart TV system with body-gesture control, tag-based rating and context-aware recommendation, Knowledge-Based Systems, vol.56, pp. 167-178, 2014
[12] kinect sensor - http://www.microsoft.com/en-us/kinectforwindows/discover/features.aspx
[13] openCV - http://opencv.org/

최 주 철(Choi Ju Choel)

- 1998년 2월 : 경희대학교 기계공학과(공학사)
- 2005년 2월 : 경희대학교 경영학과(경영학석사)
- 2009년 2월 : 경희대학교 경영학과(경영학박사)
- 2009년 12월 : 롯데정보통신(주) 수석
- 2012년 3월 ~ 현재 : 경희대학교 부교수
- 관심분야 : 추천시스템, 임베디드소프트웨어,데이터마이닝
- choijc@khu.ac.kr