거리 학습과 재서열화를 이용한 방송 콘텐츠에 대한 블로그 포스팅 태깅

손정우, 김선중, 김화숙, 조기성
한국전자통신연구원 방송통신미디어연구소 스마트미디어플랫폼연구실
e-mail : {jwson, sjkim, wskim, chokis}@etri.re.kr

Distance Learning and Re-Ranking based Broadcasting Contents Tagging with Blog Postings

Jeong-Woo Son, Sun-Joong Kim, Hwa-Suk Kim, Keeseong Cho
Smart Media Platform Research Section
Broadcasting & Telecommunications Media Research Laboratory
Electronics and Telecommunications Research Institute

요 약

이미지 혹은 영상에 대한 자동 태깅은 해당 콘텐츠에 대한 추가적인 정보를 자동으로 시스템에 제공하는 기술로써 영상 인식, 콘텐츠 매시업, 정보 검색 등 다양한 기술/서비스 분야에서 여러 목적으로 활용되고 있다. 특히, 방송 콘텐츠는 많은 양의 정보를 제한된 영역 및 시간에 축약하여 담고 있기 때문에 영상 처리 기술을 통한 객체 인식이나, 콘텐츠 매시업, 추천 서비스 등의 성능 향상을 위해 자동 혹은 수동 태깅을 통한 정보 제공이 요구된다. 본 논문에서는 블로그를 이용한 프레임 단위의 방송 콘텐츠 태깅 기술을 제안한다. 제안하는 기술은 기존의 콘텐츠 단위의 정보 제공이나, 수동 태깅 된 정보를 제공하는 기술들과 달리, 영상의 각 프레임에 대한 자동 태깅을 목표로 한다. 제안하는 방법은 거리 학습을 통해 영상의 각 프레임이 가지는 특성을 고려한 모델을 구축한 후, 이를 토대로 영상의 프레임들과 블로그의 이미지를 매칭한다. 매칭된 결과를 기반으로 특정 블로그는 영상 내 특정 프레임 구간에 태깅 된다. 제안한 방법은 이미지 매칭 성능을 측정하여 평가하였다. 블로그 이미지에 대해 Top 1 매칭 프레임을 살펴본 결과, 70%의 정확률을 보였다. 소프트 매칭(Top n)의 경우, 최대 90%의 성능을 얻을 수 있음을 실험을 통해 알 수 있었다.

1. 서론

방송 콘텐츠는 일반 사용자가 가장 손 쉽게 접근할 수 있는 형태의 콘텐츠 소비 창구이다. 최근 사용자의 콘텐츠 접근 경로가 SNS, 포털 등 다양해지면서, 방송 콘텐츠 역시 변화하는 콘텐츠 소비 환경에 맞춰 변신을 꾀하고 있다. 차세대 방송 서비스의 대표적인 예로 콘텐츠 내의 객체를 인식하고, 인식된 결과를 토대로 콘텐츠 외부의 정보를 매시업하여 사용자에게 전달 혹은 추천하는 서비스를 들 수 있다.

차세대 방송 서비스를 실현하기 위해서는 다양한 기술들이 요구된다. 먼저, 콘텐츠에 등장하는 객체를 인식하는 영상 인식 기술을 들 수 있다. 콘텐츠에 등장하는 인물, 자동차 등을 검출하거나, 좀 더 세부적으로 식별하는 작업은 콘텐츠를 시청하고 있는 사용자가 요구하는 정보가 무엇인지 판별하기 위한 초기 과정이다. 객체가 식별된 후에는 연관된 정보를 얻기 위한 검색 및 추천 기술이 요구되며, 다양한 콘텐츠를 효과적으로 구성할 수 있는 매시업 기술 또한 필요하다. 위와 같은 기술들은 최근 다양한 연구가 이루어지고 있으나, 아쉽게도 현재 영상만을 이용하여 만족할 만한 성능을 얻기는 힘들다. 대표적인 예로 영상 처리에 기반한 객체 식별의 경우, AP(Average Presicion) 0.5 를 넘기기 힘든 실정이다.

서비스를 구성하는 기술들의 성능을 높이는 효과적인 방법 중 하나는 추가적인 정보를 제공하는 것이다. 영상에 더해, 해당 영상이 어떤 객체를 담고 있는 지, 출연하는 배우는 누군지 등의 정보만으로도 쉽게 성능을 올릴 수 있다. 현재 서비스하거나, 서비스 준비 중인 많은 시스템에서도 위와 같은 정보를 수동 혹은 일부 자동으로 제공하고 있다. 다만, 완전히 자동으로 동작하지 않는 기술의 특성 상, 콘텐츠 단위의 정보 제공에 머물고 있는 실정이다. 이러한 형태의 정보 제공은 장면에 나타난 객체 단위의 서비스와 같이, 장면(scene), 샷(shot), 프레임(frame)과 같은 콘텐츠 단위보다 좀 더 세부적인 영상 단위에서 이루어지는 서비스에는 분명한 한계점을 가지고 있다.

본 논문에서는 방송 콘텐츠의 각 프레임에 대한 자동 블로그 태깅 기술을 제안한다. 제안한 방법은 특

정 프레임에 대해 연관된 블로그를 태깅함으로써 방송 콘텐츠에 대한 기존의 정보 제공 기술의 단점을 해소하고자 한다. 제안한 방법은 먼저, 콘텐츠와 연관된 블로그 포스팅(blog posting)을 수집한다. 이때, 콘텐츠 방영일과 블로그 생성 시기와의 관계를 고려하여, 대상 콘텐츠와 연관된 블로그만을 수집한다. 다음으로, 수집된 블로그에서 대상 콘텐츠에 등장하는 장면만을 모으기 위해 제품 이미지와 같은 노이즈들을 제거 한다. 정제된 블로그 이미지는 대상 콘텐츠의 프레임들과 비교하여 특정 프레임에 태깅한다. 마지막으로 블로그 이미지가 태깅된 결과를 바탕으로 해당 블로그를 프레임에 태깅한다.

본 논문에서 제안하는 태깅 기술의 핵심은 블로그 이미지와 콘텐츠의 프레임 간 매칭에 있다. 일반적인 1 시간 분량의 방송 콘텐츠의 경우, 10 만여 프레임으로 이루어져 있다. 따라서 블로그 이미지를 모든 프레임과 단순 비교하는 것은 많은 계산량을 요구한다. 본 논문에서는 이를 해소 하기 위해 임의 선택된 특정 프레임에 대해서 모델을 학습 한 후, 학습된 모델을 기반으로 매칭을 수행한다. 이와 같은 방법은 1) 매칭에 드는 시간을 줄이고, 2) 각 프레임의 특성을 모델에 반영하여 성능을 높일 수 있다. 프레임에 대한 모델은 거리 학습(distance learning)을 통해 구축하였으며, 더 정확한 태깅을 위해 재서열화(re-ranking)를 사용하였다.

재안한 방법의 성능은 블로그 이미지와 프레임 간의 매칭 정확도를 사용하여 측정하였다. 성능 검증을 위해 “별에서 온 그대” 1 화를 방송 콘텐츠로 하여, 블로그 24 개를 수집하였다. 실험의 결과 제안한 방법에서는 Top 1 매칭에 대해 최대 70%의 정확률을 보였으며, Top n 의 경우, 최대 90%의 성능을 얻을 수 있었다.

본 논문의 구성은 아래와 같다. 먼저, 2 장에서는 방송 콘텐츠와 블로그의 특성을 토대로 블로그 포스팅이 방송 콘텐츠에 대한 정보 제공에 쓸모가 있음을 보인다. 3 장에서는 본 논문에서 제안하는 거리 학습과 재서열화를 이용한 방송 콘텐츠 태깅 방법을 설명한 후, 4 장에서 실험을 통해 성능을 검증한다. 마지막으로 5 장에서 결론을 맺는다.

2. 방송 콘텐츠와 블로그

웹을 통한 사용자의 참여가 확대되는 사회 전반에 걸친 현상은 방송 콘텐츠에 대해서도 예외 없이 나타나고 있다. 특정 콘텐츠가 방영되면, 블로그, SNS 등을 통해 다양한 사용자의 생각이 전파된다. 타 콘텐츠와 달리 방송 콘텐츠는 해당 콘텐츠에 대한 설명을 토대로 시청 의견이나 상품, 촬영지에 대한 정보들이 웹을 통해 전파되는 특성이 있다. 방송 콘텐츠에 대한 태깅 대상을 선정하기 위해 다양한 웹 데이터들을 검토하였다. SNS 의 경우, 작성의 편의성으로 인해 많은 양의 메시지가 전달되나, 트위터(Twitter), 페이스북(Facebook) 등을 통해 전파되는 메시지는 길이가 짧고 많은 축약이 일어남을 알 수 있었다. 반면, 블로그 포스팅은 그 수가 SNS 와 비교할 경우 적으나, 1) 특정 객체나 장소 등의 정보를 자세히 기술하고 있는 점, 2) 콘텐츠의 방영에 따라 충분히 많은 양의 포스팅이 생성되는 점 등을 확인할 수 있었다. 그림 1 은 “별에서 온 그대”가 방영된 후, 해당 콘텐츠에 대해 언급하는 블로그 포스팅의 수를 보여준다. 그림에서 보여주듯이, 콘텐츠 방영일을 시작으로 종영일까지 꾸준히 증가함을 알 수 있다.

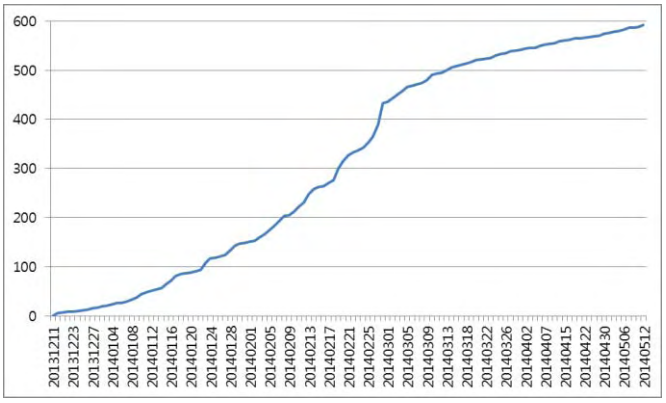

그림 1. 콘텐츠 방영 후, 블로그 포스팅 추이

블로그 포스팅을 영상에 태깅하기 위해서는 영상의 프레임과 포스팅을 연결할 수 있는 매개체가 필요하다. 100 여개의 블로그 포스팅을 검토한 결과, 방송 콘텐츠에 대한 블로그 포스팅은 해당 콘텐츠의 특정 장면을 설명하기 위한 이미지와 이에 대한 사용자의 설명으로 이루어져 있었다. 사용자의 설명은 해당 장면의 촬영지, 등장 인물 및 객체, 그리고 감상평이 주를 이루었다. 주목할 점은 블로그 포스팅을 작성할 때, 콘텐츠의 특정 장면에 대해 논한다는 것과 이 때, 특정 장면은 이미지를 이용해서 설명한다는 것이다. 사용되는 이미지는 방송 콘텐츠의 특정 프레임에 대응되나, 색조, 밝기 등을 조정한 수정본인 경우도 빈번히 보였다. 본 논문에서는 이와 같이 콘텐츠의 특정 장면을 설명하는 이미지를 추출하여, 프레임과 효과적으로 매칭함으로써 방송 콘텐츠의 특정 프레임에 대해 블로그를 태깅하고자 한다.

3. 블로그 태깅을 위한 거리 학습 및 재서열화

1 회차분의 방송 콘텐츠는 60 분을 기준으로 10 만여 프레임으로 이루어져 있다. 웹에서 수집된 블로그 포스팅의 경우, 8~10 여개의 이미지를 포함하고 있으며, 이중 약 60%인 5~6 개의 이미지가 태깅을 위해 매칭되어야 할 대상이다. 모든 프레임에 대해 모든 이미지를 매칭할 경우, 간단한 색상 히스토그램 조차도 많은 시간을 소요하기 때문에, 실제 시스템에 적용하기는 힘들다. 따라서 본 논문에서 제안하는 방법은 3 단계에 걸쳐, 1)노이즈 제거, 2) 매칭, 3) 재서열화를 통해 속도와 정확률을 모두 고려하고자 한다.

3.1. 노이즈 제거

블로그에는 다양한 이미지가 포함되어 있다. 이 중, 제품 사진, 직접 촬영한 풍경 등 해당 콘텐츠에서 나

타나지 않는 모든 이미지를 제안한 방법에서는 노이즈로 간주하고 필터링한다. 노이즈를 제외한 이미지는 실제로 콘텐츠에 등장하는 특정 프레임의 전체 혹은 일부이다. 따라서 노이즈가 아닌 경우, 이미지는 콘텐츠의 프레임 중, 일부와 매우 높은 유사도를 보인다. 특히, 장소, 등장 객체, 조명 등이 같기 때문에 색상 공간에서도 쉽게 의미 있는 유사도를 측정할 수 있다. 제안한 방법에서는 노이즈 제거를 위해 HS(Hue Saturation) 히스토그램을 기반으로 샘플링된 프레임과 비교하여 유사도를 얻는 함수 $f_t(x_1, x_2)$를 코사인 유사도를 바탕으로 정의하였다. $f_t(x_1, x_2)$를 통해 얻은 유사도 중, 최대값 n 개의 평균이 임계값 Θ_t 보다 낮은 경우, 노이즈로 보고 제거하였다.

3.2. 거리학습 기반 이미지 매칭

노이즈 제거에 사용된 HS 히스토그램은 블로그이미지를 특정 프레임에 태깅하기에는 적합하지 않다. 이는 서로 동떨어진 프레임 간에도, 배경, 인물, 극중 시간 등의 일치로 인해 높은 유사도를 보일 수 있기 때문이다. 이 경우, 배경이나 카메라 각도가 동일하기 때문에 SIFT[1]나 SURF[2] 등 경계선 정보에 기반한 특징을 사용하더라도, 쉽게 구분 지을 수 없다. 같은 장소, 같은 인물, 같은 조명이 사용된 사로 상이한 두 프레임을 구분 짓는 것은 매우 작은 크기의 소품일 수도 있기 때문이다. 이와 같이 프레임 간의 구분이 가능한 특징의 추출을 위해 본 논문에서는 거리학습(distance learning)[3]을 사용한다. 거리 학습이란, 두 벡터 x_1 과 x_2 가 주어졌을 때, 이들의 유사도가 특정 조건이 만족하도록 유사도 함수 $f_s(x_1, x_2)$를 학습하는 것을 의미한다. 이때, $f_s(x_1, x_2)$는 두 벡터를 변환하는 행렬 $\mathbf{A}$ 를 학습을 통해 추정해야 할 파라미터로 가진다. 본 논문에서는 1 부터 m 까지의 프레임이 주어졌을 때, k 개의 프레임을 샘플링하여 각 프레임에 적합한 k 개의 행렬 $\{\mathbf{A}_i\}_{i=1,\dots,k}$를 학습한다. 이 때, i 번째 프레임에 대한 행렬 $\mathbf{A}_i$ 를 학습하기 위한 제약 조건은 두가지로 주어진다.

1) i 프레임의 주변 $\pm s$ 개의 프레임 $j \in J$ 중, $f_t(i, j)$ 가 Θ_s 보다 클 경우, i 프레임과 같은 것으로 간주 한다.
2) $i \pm s$ 프레임 외, 나머지 프레임 $l \in L$ 의 경우, i 프레임과의 유사도는 $|\#i - \#l|$에 비례하여 작아지도록 한다. 이때, $\#i$ 는 i 프레임의 프레임 넘버를 의미한다.

위의 제약 조건을 토대로 $\{\mathbf{A}_i\}_{i=1,\dots,k}$ 는 SVMRank[4]를 이용하여 학습되었다.

이와 같은 거리 학습 기반의 매칭 함수가 가지는 장점은 다음과 같다. 1) 프레임을 구분하는 매우 작은 특징이라 하더라도, 자동으로 추출될 수 있다. 2) 특정 콘텐츠에 대한 모델 학습에 시간이 걸리나, 블로그 이미지가 매칭되는 시점에서 비교 횟수를 줄일 수 있다.

3.3. ORB 기반 재서열화

콘텐츠의 특정 프레임이 블로그에 사용될 때, 많은 경우, 수정을 거친다. 이들 변화를 고려하기 위해 행렬 A 는 어느 정도의 오류를 허용하여 학습이 된다. 따라서, $f_s(x_1, x_2)$에 기반한 태깅 결과는 정확성에 한계를 가진다. 물론, 색상 정보와 별개로, 경계선을 활용한 SIFT 와 같은 특징을 함께 고려하여 학습할 경우, 정확률은 높일 수 있다. 하지만, 이 경우, 속도가 매우 느려지는 단점이 있다. 본 논문에서는 속도를 최대한 유지하며 SIFT 나 SURF, ORB[5]와 같이 색상과는 다른 정보를 가지는 복잡한 특징을 사용하기 위해 재서열화 (re-ranking)[6]를 사용한다.

즉, $f_s(x_1, x_2)$를 통해 얻어진 블로그 이미지 x_1 에 대한 n 개의 프레임 $\{x_c\}_{c=1\dots n}$ 에 대해 상위 r $(r < n)$개의 이미지에 대해 ORB 특징을 기반으로 재서열화를 수행한다. 재서열화 함수 f_o 는 두 ORB 벡터 간의 유사도 $S_{ORB}(x_1, x_2)$와 더불어, f_s 를 함께 사용하여

$$f_o(x_1, x_2) = \alpha S_{ORB}(x_1, x_2) + (1-\alpha) f_s(x_1, x_2),$$

로 정의되며, α $(0<\alpha<1)$는 사용자 파라메터이다. 재서열화는 정확성이 높으나 속도가 느린 특징 추출 기법을 소량의 데이터에만 적용함으로써 속도를 유지할 수 있는 장점이 있다.

4. 방송 콘텐츠를 이용한 성능 검증

제안한 방법의 성능은 블로그 이미지에 대한 방송 콘텐츠의 프레임 단위의 매칭 정확률을 측정하여 검증하였다. 실험을 위해, 방송 콘텐츠로 “별에서 온 그대” 1 회차를 선정하였으며, 해당 콘텐츠의 방영일인 2013 년 12 월 18 일부터 12 월 24 일까지 생성된 블로그 중, “별에서 온 그대” 혹은 “별그대” 중 하나 이상을 포함하고, 극중 여자 주인공이 착용한 자켓인 “LM2181775228”을 포함하는 포스팅을 수집하였다. 수집에는 NaverAPI 를 사용하였다.

수집된 총 블로그 포스팅 수는 24 개이며, 이들 포스팅에 포함된 이미지의 수는 총 232 개로 평균 10 여개의 이미지를 포함하고 있다. 수집에 사용된 질의어 중, 자켓의 제품명은 일반적으로 알기 어려운 형태이며, 극 중, 하나의 장면에 등장한다. 따라서, 수집된 블로그 데이터의 수는 방송 콘텐츠의 특정 조건을 만족하는 일부에 대한 정보도 블로그 포스팅을 통해 얻을 수 있음을 보여준다. 수집된 블로그 이미지 232 개 중, 53 개인 약 23%가 노이즈였으며, 노이즈를 제외한 이미지의 경우, 해당 이미지와 매칭되는 장면의 시작 프레임과 끝 프레임을 태깅하여 실험 데이터를 구축했다.

먼저, 노이즈 제거 함수의 경우, 이미지를 4x5 의 영역으로 나눈 후, 18x32 크기의 HS 벡터를 추출하였다. 즉, 이미지 하나에 대해 4x5x18x32 사이즈의 벡터가 추출되었으며, Θ_t 는 0.8 로 설정했다. 프레임은 초당 1 장의 이미지를 임의 선택하여 사용하였다. 실험 결과 노이즈 제거 성능은 precision, recall 모두 0.94 로 우수한 성능을 보였다. 모두 노이즈가 아닌 것으로 판단할 경우, precision 0.77, recall 1.00 임을 고려해 볼 때, 좋은 성능임을 알 수 있다.

매칭의 경우, 총 592 개 프레임을 2 초 단위로 임의 선택하여 사용하였다. 즉, 592 개의 모델을 학습하였으며, 하나의 모델을 학습하는 데, 총 20 여초가 소요되었다. 매칭 성능은 Top n 개의 프레임 중, 정답 프레임

이 하나라도 포함되어 있는 경우의 비율을 계산하였으며, 성능은 그림 2 와 같다. 그림에서 알 수 있듯이, 매칭 함수는 n=1 부터 n=30 까지 0.66 에서 0.90 의 성능을 보였다. 592 개 모델에 대한 비교에는 0.04 초가 소요되었음을 고려 해볼 때, 짧은 계산 시간에도 불구하고 높은 성능을 보임을 알 수 있다.

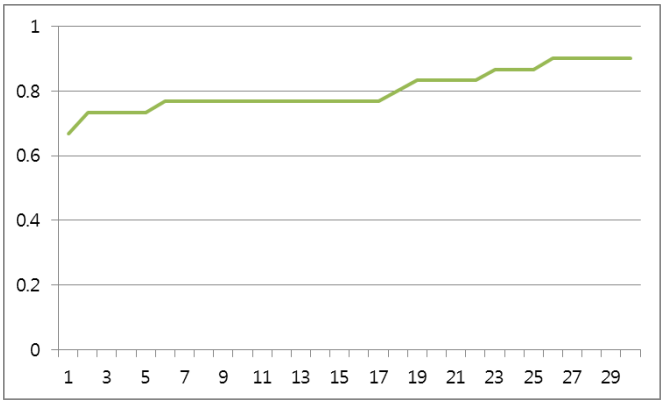

그림 2. 거리학습 기반 매칭 성능

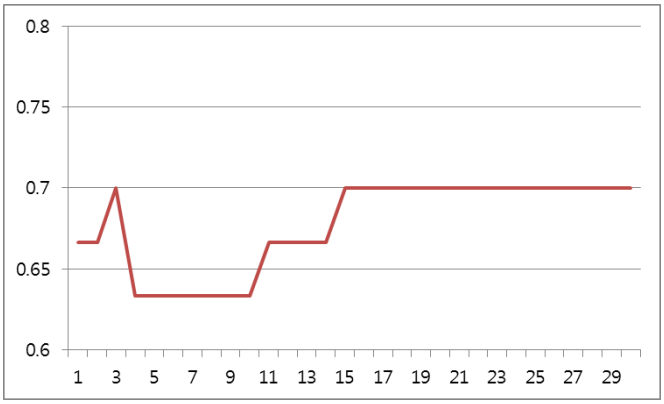

그림 3. 재서열화 성능

그림 3 은 재서열화의 성능을 보여준다. 재서열화는 Top 1 에 대해 정확하게 매칭 되었는가를 판단하여 성능을 측정하였다. 그림에서 알 수 있듯이, 매칭을 통해 2 개 이상의 후보를 추출할 경우, 재서열화 함수의 성능은 0.70 으로 재서열화를 하지 않을 경우와 비교해볼 때, 4.5%의 성능 향상을 볼 수 있었다.

5. 결론

본 논문에서는 방송 콘텐츠에 대한 프레임 혹은 장면 단위의 지식 태깅의 사전 작업으로써, 이미지 매칭을 통한 콘텐츠 프레임 단위의 블로그 태깅 기술을 제안하였다. 제안한 방법은 프레임 간의 특징을 매칭에 반영하고, 비교 횟수를 줄이기 위해 거리 학습 기반의 매칭 기술을 도입하였다. 거리 학습에서 프레임 간의 거리를 제약으로 정의함으로써 전체적으로 유사하나 같은 장면에 속하지 않는 프레임 간에 유사도를 낮추도록 했다. 뿐만 아니라, 매칭의 성능을 높이기 위해 재서열화를 통해 경계선 정보를 매칭에 반영하였다.

두 단계 매칭 기법의 성능은 실제 방송 콘텐츠를 대상으로 한 실험을 통해 검증하였다. 실험에서 거리 학습 기반 매칭을 사용한 경우, Top n 개의 후보 내에 정답 프레임이 속하는 비율이 90%에 이르렀으며, Top 1 을 고려할 경우에도 66%의 성능을 보였다. 거리 매칭 결과를 기반으로 한 재서열화의 경우에도 Top 1 에 대해 70%의 정확률을 보였으며, 이는 거리 매칭의 성능을 4% 향상 시키는 결과로 해석된다.

본 논문에서 제안한 방법은 프레임 단위 태깅의 초기 단계에 활용될 수 있다. 따라서, 추후, 태깅된 블로그의 텍스트를 분석하여, 정제된 지식 정보를 태깅하는 방법과 이를 기반으로 객체 인식, 매시업 서비스 등에 적용하는 방법에 대한 연구가 필요하다. 뿐만 아니라, 더 많은 데이터를 기반으로 한 검증이 필요할 것으로 보이며, 이들 둘은 추후 연구에 반영할 예정이다.

Acknowledgement

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 수행하였음. [14-000-11-002, 방송용 영상 인식 기반 객체 중심 지식융합 미디어 서비스 플랫폼 개발]

참고문헌

[1] D. Lowe. “Object Recognition from Local Scale-Invariant Features,” In *Proceedings of the International Conference on Computer Vision*, pp. 1150-1157, 1999.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Gool. “SURF: Speeded Up Robust Features,” Computer Vision and Image Understanding, Vol. 100, No. 3, pp. 346-359, 2008.

[3] K. Weinberger, J. Blitzer, and L. Saul. “Distance metric learning for large margin nearest neighbor classification,” In *Advances in Neural Information Processing Systems*, pp. 1473-1480, 2005.

[4] M. Schultz and T. Joachims. “Learning a distance metric from relative comparisons,” In *Advances in Neural Information Processing Systems: 41*, 2004.

[5] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: an efficient alternative to SIFT or SURF,” In *Proceedings of IEEE International Conference on Computer Vision*, 2011.

[6] J. Carbonell and J. Goldstein. “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” In *Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335-336, 1998.