Tag Refinement를 이용한 이미지 어노테이션
(Image Annotation using Tag Refinement)

차 재 성 † 조 선 영 †† 어 영 정 † 김 성 도 † 변 혜 란 †††
(Jaeseong Cha) (Sunyoung Cho) (Youngjung Uh) (Seongdo Kim) (Hyeran Byun)

요 약 최근 Flickr, Facebook과 같은 사진 공유 기반의 소셜 미디어 공유 사이트의 발전으로 인해 이미지의 양이 폭발적으로 증가하면서, 효율적인 이미지 검색을 위한 연구가 활발히 진행되고 있다. 이와 함께 이미지에 자동으로 관련된 태그를 어노테이션하는 이미지 태깅 연구가 진행되고 있으며, 이미지 뿐 아니라 태그와 같은 이미지에 달린 컨텍스트 정보까지도 함께 고려함으로써 태깅 성능을 높이려는 시도를 하고 있다. 본 논문에서는 일관성 있는 태그 추출을 위해 Tag refinement를 이용한 웹이미지 어노테이션 방법을 제안한다. 제안하는 방법은 1) 관심영역 기반의 이미지 특징을 이용하여 쿼리의 이웃 이미지를 검색한다, 2) 검색된 이웃 이미지의 태그로부터 NMF 클러스터링을 기반으로 하여 쿼리와 관련된 태그를 추출한다, 3) 추출된 태그의 순서를 Tag Refinemt를 통해 관련성 순으로 결정하여 태깅한다. Flickr로부터 수집한 태그가 달린 이미지 데이터셋에 대해 실험하였고, 일반적인 이웃 투표 기법을 포함하여 클러스터링 방법과 제안하는 방법을 비교함으로써 태깅 성능을 평가한 결과 제안하는 방법이 기존의 방법보다 더 높은 태깅 정확도를 가지고 있음을 보였다.

키워드 : 이미지 어노테이션, 이미지 태깅, 소셜 태그, 태그 개선

Abstract Recently, with the development of social multimedia tagging through Flickr and Facebook, many researches have studied by using social tag-annotated web image. However, previous methods produced the tag list with low consistency since low correlation among extracted tags. This paper proposes web image annotation through tag refinement for the consistent tag extraction. Our method produces tag result with high relevancy and consistency. We conduct an experiment on annotated web image dataset collected from Flickr. We show that the proposed method gives more high performance in tagging accuracy and consistency than the previous methods.

Key words : Image Annotation, Image Tagging, Social Tag, Tag Refinement

1. 서 론

최근 인터넷이 발달하고 디지털 카메라가 보편화되면서, Flickr, Facebook, YouTube와 같은 소셜 미디어 공유 사이트가 급격히 성장하고 있다. 그 결과, 인터넷과 웹 기술의 발전으로 멀티미디어의 양이 폭발적으로 증가하면서, 대량의 멀티미디어 컨텐츠를 관리하기 위해 데이터에 대한 설명을 추가하는 어노테이션 메커니즘이 많이 사용되고 있다. 특히, 이미지 어노테이션은 폭발적으로 증가하는 웹이미지에 따른 효율적인 이미지 검색의 필요성으로 인해 점점 중요해지고 있다. 대부분의 이미지 검색 연구는 주로 이미지의 내용을 분석하는 내용기반 이미지 검색(CBIR: Content-based Image Retrieval) 방법이 많이 진행되어 왔다. 이러한 초기 이미지 태깅 및 어노테이션 연구는 색상, 텍스쳐 및 형태와 같은 시각적 특징을 이용하여 이미지의 내용을 분석한다. 비록 이러한 방법은 정의하는 태그의 개수가 적을 경우에는

· 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2011-0027450)

† 비 회 원 : 연세대학교 컴퓨터과학과
cjs135@yonsei.ac.kr
sheldon@yonsei.ac.kr
seongdo@yonsei.ac.kr
†† 학생회원 : 연세대학교 컴퓨터과학과
sycho22@yonsei.ac.kr
††† 종신회원 : 연세대학교 컴퓨터과학과 교수
hrbyun@yonsei.ac.kr
(Corresponding author임)

논문접수 : 2012년 2월 8일
심사완료 : 2012년 5월 17일

정보과학회논문지: 소프트웨어 및 응용 제39권 제8호(2012.8)

잘 작동하지만, 데이터셋이 커지고 태그의 종류가 다양해짐에 따라 성능이 떨어지게 된다. 또한, 이러한 모델 기반 방법은 저차원 이미지 특징과 고차원 이미지 시맨틱 간의 차이로 인한 시맨틱 갭 문제(Semantic gap problem)를 갖는다. 따라서 이러한 시맨틱 갭을 줄이고 이미지 태깅 성능을 향상하기 위해 제목, 태그나 설명 같은 이미지에 달린 여러 컨텍스트 정보를 활용하는 연구들이 많아지고 있다. 특히, 최근에는 웹의 발전과 함께 이미지의 양이 폭발적으로 증가하면서 웹에서의 활용성을 고려한 데이터 기반(data-driven) 방식의 연구들이 제안되고 있다. 데이터 기반 방식은 먼저 대용량의 이미지 데이터셋으로부터 이미지 특징을 이용하여 시각적으로 유사한 이웃 이미지 집합을 검색한다. 검색된 이웃 이미지 집합에 대해, 각 이미지의 컨텍스트 정보로부터 다양한 기법을 이용하여 쿼리 이미지와 관련된 태그만을 추출한다. 데이터 기반 방식의 이미지 태깅에서 태깅 성능은 크게 두 가지 단계에 의해 결정된다고 볼 수 있다. 먼저 첫 번째 단계인 쿼리와 유사한 이미지 집합을 검색하는 단계이다. 이 단계는 이미지의 시각적 특징을 이용하여 쿼리와 유사한 이웃 이미지를 찾음으로써, 두 번째 단계에서 분석될 후보 태그들을 필터링하는 역할을 한다. 따라서 쿼리와 얼마나 유사한 이미지들이 잘 검색되었는지에 따라 후보 태그들이 달라지며 결과적으로 추출되는 태그의 정확성이 결정된다. 즉, 만약 쿼리와 관련이 없는 이미지들이 검색된 경우, 그 이미지들의 태그를 이용하여 추출한 태그 역시 관련 없는 태그가 될 수 있다. 태깅 성능을 결정하는 또 다른 단계는 이웃 이미지 집합의 태그를 이용하여 쿼리와 관련된 태그를 추출하는 단계이다. 이는 첫 번째의 이미지 검색 단계는 오랫동안 연구되어온 분야이지만 성능적인 한계가 존재하기 때문에 태그 정보를 함께 사용함으로써 태깅 성능을 향상시키기 위함이다. 그러나 일반적으로 이미지에 달린 태그는 각 개인 사용자에 의해 달린 태그이기 때문에 주관적이고 개인화된 잡음 태그를 포함하고 있다. 따라서 대부분의 이미지 태깅 연구들은 다양한 방식을 통해 잡음 태그를 포함한 태그들로부터 쿼리와 관련된 객관적인 태그만을 추출하는 방법을 제안함으로써 태깅 성능을 높이려는 시도를 하고 있다. 소셜 태그는 여러 일반 사용자들에 의해 달려진 것으로, 애매모호하고 일관성이 떨어지며 주관적이라는 단점이 존재한다. 따라서 태그들을 분석하여 이미지와 관련없는 잡음(noise) 태그들을 필터링하고, 관련있는 태그만을 추출하여 이미지에 달아주는 이미지 어노테이션은 효율적인 이미지 검색을 위해 중요하다. 이에 대한 많은 연구들이 진행되고 있지만, 기존의 방법들은 여전히 추출된 태그들간에 연관성이 적어 일관성이 떨어지는 결과를 준다. 본 논문에서는 tag refinement 기법을 통한 이미지 어노테이션 알고리즘을 통해, 관련 태그를 추출하면서도 태그들 간 연관성이 높은 결과가 추출되는 방법을 제안한다. 특히 기존의 방법에 비해 추출된 태그들 간에 서로 관련성이 높다는 것을 보인다. 본 논문의 구성은 다음과 같다. 2장에서는 기존의 연구를 소개하고 문제점을 분석한다. 3장에서는 제안하는 이미지 어노테이션 방법을 서술하고, 4장에서는 다른 알고리즘과의 비교를 통해 제안하는 알고리즘의 우수성을 입증한다. 마지막으로 5장에서는 결론 및 향후 연구 방향에 대해 기술하겠다.

2. 관련 연구

최근 멀티미디어 양이 폭발적으로 증가하면서 효율적인 멀티미디어 검색 기술은 중요한 이슈가 되고 있다. 이를 위해 이미지에 관련된 태그를 자동으로 달기 위한 이미지 태깅(Image Tagging)이나 태그 추천(Tag Suggestion) 기술이 활발히 연구되고 있다. 대부분의 초기 이미지 태깅 및 어노테이션 연구는 색상, 텍스쳐 및 형태와 같은 시각적 특징을 이용하여 이미지의 내용을 분석한다[1-3]. 이 방법은 다양한 패턴인식 기법을 적용하여 이미지 특징과 태그 간 관계를 사상(mapping)하는 모델을 정의한다. Mori의 알고리즘[1]에서는 이미지를 여러 서브이미지로 나누고, 양자화한 각 서브이미지의 특징 벡터와 태그를 사상한다. Duygulu가 제안한 방법[2]에서는 이미지를 영역단위로 분할한 후, EM(Expectation Maximization) 알고리즘을 이용하여 영역 타입과 키워드간 사상관계를 학습한다. Blei가 제안한 방법[3]은 CORR-LDA(Correspondence Latent Dirichlet Allocation)를 이용하여 이미지 영역과 워드 집합의 latent 변수 표현 간 관계를 찾는 모델을 정의한다. 모델 기반 방식은 일반적으로 이미지 특징과 어노테이션 텍스트 간 대응관계를 모델링하기 위해 확률 모델이나 분류 알고리즘을 이용한다[4,5].

Cusano가 제안한 방법[4]에서는 SVM(Support Vector Machine)을 이용하여 이미지 영역을 7개 클래스 중 하나로 분류함으로써 이미지 어노테이션을 수행한다. Carneiro와 Vasconcelos가 제안한 방법[5] 역시 어노테이션 텍스트를 클래스 레이블로 분류하기 위해 베이즈 결정 규칙(Bayes decision rule)의 응용을 통한 분류 방법을 제안한다. 그러나 이러한 모델 기반 방식은 이미지 특징으로부터 어노테이션 텍스트를 검출하기 위한 학습 과정에서 시맨틱 갭 문제가 발생하고, 제한된 태그 개념만을 다룬다는 단점이 존재한다. 비록 이러한 방법은 정의하는 태그의 개수가 적을 경우에는 잘 작동하지만, 데이터셋이 커지고 태그의 종류가 다양해짐에 따라 성능이 떨어지게 된다. 또한, 이러한 모델기반 방

법은 저차원 이미지 특징과 고차원 이미지 시맨틱 간의 차이로 인한 시맨틱 갭 문제(Semantic gap problem)를 갖는다.

따라서 이러한 시맨틱 갭을 줄이고 이미지 태깅 성능을 향상하기 위해 제목, 태그나 설명 같은 이미지에 달린 여러 컨텍스트 정보를 활용하는 연구들이 많아지고 있다[6-9]. Yeh의 알고리즘[6]에서는 이미지와 키워드를 혼합한 검색 방법을 제안한다. 이 방법은 웹 페이지로부터 관련된 키워드를 추출하여, 쿼리 이미지와 가장 비슷한 이미지를 찾는데 이용한다. Weinberger의 알고리즘[7]에서는 태그분포의 가중치된 KL(Kullback-Leibler) divergence에 기반하여 태그를 분석함으로써 이미지에 대한 태그를 추천한다.

특히, 최근에는 웹의 발전과 함께 이미지의 양이 폭발적으로 증가하면서 웹에서의 활용성을 고려한 데이터 기반(data-driven) 방식의 연구들이 제안되고 있다[8-11]. 이 방식에서는 태그 활용 시, 주관적이고 애매모호한 태그의 특성으로 인해 잡음 태그를 필터링하고 관련있는 태그만을 추출하는 것이 중요하다고 할 수 있다. Wang이 제안한 방법[8]에서는 컨텍스트 정보로부터 마이닝 기법을 적용하여 관련된 용어나 구절을 추출한다. 그리고 이미지와 키워드 검색을 통해 쿼리와 유사한 이웃 이미지들을 찾고, 그 이미지들의 태그들을 SRC (Search Result Clustering)를 통해 마이닝함으로써 관련된 태그만을 추출한다. Li의 알고리즘[9]에서는 이웃 이미지 집합에 공통적으로 존재하는 태그일수록 쿼리와 관련있는 태그라는 가정을 가지고 이웃 투표 기법을 통해 태그를 추출한다. 이웃 투표 기법을 통해 추출된 태그들로부터, Liu의 알고리즘[10]에서는 태그 추천 시 확률 모델 기반 관련성 추정과 랜덤 워크(Random walk) 기반 필터링을 통해 관련성 순으로 재배열 한다. Lu의 알고리즘[11]에서는 웹 이미지 데이터셋으로부터 신뢰도 맵과 컨텐트-컨텍스트 유사도 행렬을 통해 적은 시맨틱 갭을 갖는 태그를 추출하고, 마이닝 과정을 통해 잡음 태그들을 필터링한다.

데이터 기반 방식은 먼저 대용량의 이미지 데이터셋으로부터 이미지 특징을 이용하여 시각적으로 유사한 이웃 이미지 집합을 검색한다. 검색된 이웃 이미지 집합에 대해, 각 이미지의 컨텍스트 정보로부터 다양한 기법을 이용하여 쿼리 이미지와 관련된 태그만을 추출한다. 마지막으로 추출된 태그들의 관련성 순으로 태깅되는 순서를 결정한다. Li의 알고리즘[9]에서는 유사한 이미지 집합에 존재하는 공통적 태그들은 쿼리 이미지와 관련된 태그일 것이라는 가정을 가지고, 이웃 투표(neighbor voting) 기법을 제안함으로써 관련된 태그를 추출한다. 이러한 방법들은 컨텍스트 정보를 활용함으로써 사용자의 태깅 경향을 반영하고 효율적으로 관련된 태그를 추출한다. 본 논문에서도 이미지 내용뿐만 아니라 태그 정보도 함께 이용하며, 데이터 기반 방식에 의해 이미지의 태그를 추출한다.

3. 제안하는 방법

제안하는 이미지 태깅 알고리즘은 데이터 기반 방식으로써 다음의 3가지 과정으로 구성된다. 먼저, 주어진 쿼리 이미지로부터 피사체 영역으로 추정되는 관심영역을 추출한다. 관심영역에 대해 이미지 특징을 추출하고 이를 통한 내용 기반 검색을 통해 쿼리 이미지와 유사한 이웃 이미지 집합을 찾는다. 다음은 이웃 이미지 집

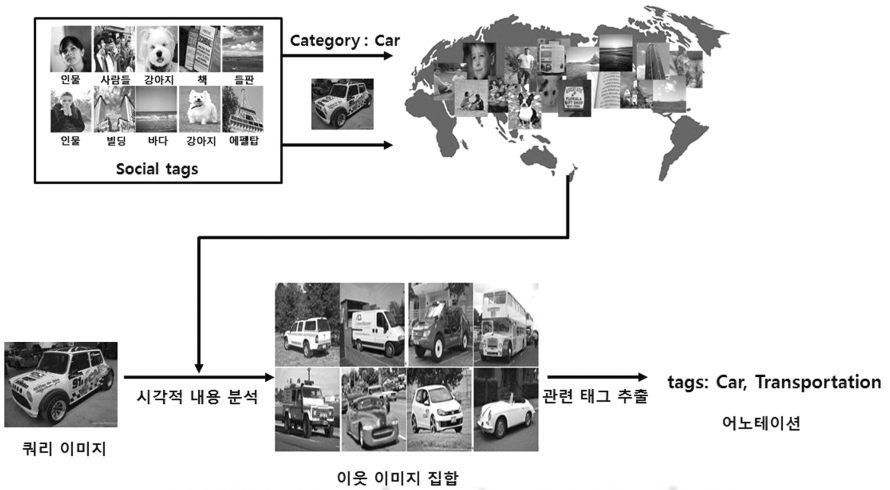

그림 1 제안하는 이미지 태깅 방법의 흐름도

합에 포함된 각 이미지의 태그들에 대해, NMF 기반 클러스터링 알고리즘을 이용하여 쿼리 이미지에 대한 태그를 추출한다. 마지막으로 앞에서 추출된 태그들에 대해 관련성 순으로 태깅되는 순서를 결정한다. 그림 1은 제안하는 알고리즘의 흐름도를 보여준다.

3.1 시각적 내용 분석을 통한 이웃 이미지 검색

주어진 쿼리 이미지에 대해, 이미지의 피사체라고 여겨지는 관심영역을 추출하기 위해 GBVS(Graph-Based Visual Saliency) 모델을 적용한다[12]. 이 모델은 그래프 이론을 이용하여 추출된 특징 맵(Feature map)에서의 활성화 지도(Activation map)를 만들고, 마코브(Markovian) 알고리즘을 이용하여 활성화 지도를 정규화 함으로써 관심영역을 강조한다. 먼저 특징 맵 $M: [n]^2 \rightarrow R$이 주어지면 활성화 맵 A를 계산하는 과정은 $A: [n]^2 \rightarrow R$로 볼 수 있다. $[n]=\{1,2,\cdots,n\}$, R은 실수를 나타낸다. 활성화 맵은 마코브 알고리즘을 통해 구할 수 있다. 일반적으로 두 픽셀간의 차이를 계산하는 데는 식 (1)이 사용된다.

$$d((i,j)\|(p,q)) \triangleq \left|\log \frac{M(i,j)}{M(p,q)}\right| \quad (1)$$

하지만 마코브 알고리즘에서는 $|M(i,j)-M(p,q)|$, 즉 $M(i,j)$와 $M(p,q)$ 간의 차이를 이용한다. 이제 완전 연결 방향 그래프 G_A가 있다고 가정 하면, 노드 (i,j)에서 (p,q)까지의 가중치는 식 (2)와 같다.

$$w_1((i,j),(p,q)) \triangleq d((i,j)\|(p,q)) \cdot F(i-p,j-q), \quad (2)$$
$$where\ F(a,b) \triangleq \exp\left(-\frac{a^2+b^2}{2\sigma^2}\right)$$

식 (2)를 보면 노드 (i,j)에서 (p,q)까지의 가중치는 그것의 차이값에 비례함을 알 수 있다.

위의 구해진 가중치는 활성화 지도를 정량화 시키는 단계를 거치게 되는데 이때 가중치는 식 (3)을 통해 정량화 할 수 있다.

$$w_2((i,j),(p,q)) \triangleq A((i,j)\|(p,q)) \cdot F(i-p,j-q) \quad (3)$$

이 방법은 간단하면서도 효율적으로 인간의 관심과 주의를 끄는 이미지 영역을 추출한다.

추출된 이미지의 관심영역에 대해, 색 현저도(Color saliency)와 지역 형태(Local shape)의 두 가지 이미지 특징을 추출한다. 먼저, 색 현저도 특징을 추출하기 위해 색 현저도 부스팅(Color saliency boosting) 알고리즘을 이용한다[13]. 이 알고리즘은 그림 2와 같이 해리스 검출기(Harris detector)에 의해 검출된 현저한 점(Salient point)에 대해 색 현저도를 추출한다.

영역 전체가 아닌 현저한 점 위치에서의 색 정보만을 특징으로 추출하기 때문에 간단하면서도 효율적으로 특징을 추출한다. 이미지의 지역 형태 특징을 추출하기 위해서는 PHOG(Pyramid Histogram of Oriented Gradients)를 이용한다[14]. PHOG는 지역 형태와 그에 대한 공간적 레이아웃을 이용하여 이미지를 표현한다. 지역 형태는 에지 방향의 히스토그램에 의해 표현되며, 작은 회전에 강인하다는 장점을 갖는다.

위에서 기술한 두 가지 이미지 특징은 이미지 전체 영역이 아닌 관심영역에 대해서만 추출함으로써 피사체가 아닌 다른 영역에 의해 유사하지 않은 이웃 이미지가 검색되는 경우를 줄여준다. 전체 데이터베이스에 있는 이미지로부터 유사한 이웃 이미지 집합을 찾기 위해서는 k-최단 이웃(k-nearest neighbor) 알고리즘을 적용한다. 이때 이미지 특징 간 시각적 유사도를 판단하기 위해서는 유클리디안 거리(Euclidean distance)를 이용한다.

3.2 NMF기반 클러스터링을 통한 태그 추출

CBIR 기반 이미지 검색을 통해 추출된 쿼리 이미지와 유사한 이웃 이미지들에 대해, NMF 기반 클러스터링 알고리즘을 이용하여 쿼리 이미지에 대한 태그를 추출한다. 본 논문에서는 이웃 이미지들의 태그를 분석하여 일정한 투표값 이상의 태그만을 이용하여 데이터 행렬을 구성하고, 그러한 데이터 행렬의 NMF 기반 클러스터링을 통해 쿼리 이미지와 관련된 핵심 태그들을 추출해 낸다.

먼저, 이웃 이미지 집합 Φ내에 존재하는 이미지들의 태그들의 집합을 $W=\{w_1, w_2, \cdots, w_m\}$라고 하자. 이 때, W에 포함된 각 태그들은 일정한 투표 임계값 이상을 가지며, 투표 임계값을 어떻게 설정하느냐에 따라 클러

(a) (b) (c) (d)

그림 2 (a)(c) 질의 영상, (b)(d) 그래프 기반의 현저한 점 검출 결과