KIPS Tr. Software and Data Eng.
Vol.3, No.12 pp.547~552 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.12.547

Image Classification Using Bag of Visual Words and Visual Saliency Model

Jang, Hyunwoong† · Cho, Soosun††

ABSTRACT

As social multimedia sites are getting popular such as Flickr and Facebook, the amount of image information has been increasing very fast. So there have been many studies for accurate social image retrieval. Some of them were web image classification using semantic relations of image tags and BoVW(Bag of Visual Words). In this paper, we propose a method to detect salient region in images using GBVS(Graph Based Visual Saliency) model which can eliminate less important region like a background. First, We construct BoVW based on SIFT algorithm from the database of the preliminary retrieved images with semantically related tags. Second, detect salient region in test images using GBVS model. The result of image classification showed higher accuracy than the previous research. Therefore we expect that our method can classify a variety of images more accurately.

Keywords : Bag of Visual Words, SIFT, Visual Saliency Model, GBVS, Image Classification

이미지 단어집과 관심영역 자동추출을 사용한 이미지 분류

장 현 웅† · 조 수 선††

요 약

플리커, 페이스북과 같은 대용량 소셜 미디어 공유 사이트의 발전으로 이미지 정보가 매우 빠르게 증가하고 있다. 이에 따라 소셜 이미지를 정확하게 검색하기 위한 다양한 연구가 활발히 진행되고 있다. 이미지 태그들의 의미적 연관성을 이용하여 태그기반의 이미지 검색의 정확도를 높이고자 하는 연구를 비롯하여 이미지 단어집(Bag of Visual Words)을 기반으로 웹 이미지를 분류하는 연구도 다양하게 진행되고 있다. 본 논문에서는 이미지에서 배경과 같은 중요도가 떨어지는 정보를 제거하여 중요부분을 찾는 GBVS(Graph Based Visual Saliency)모델을 기존 연구에 사용할 것을 제안한다. 제안하는 방법은 첫 번째, 이미지 태그들의 의미적 연관성을 이용해 1차 분류된 데이터베이스에 SIFT알고리즘을 사용하여 이미지 단어집(BoVW)을 만든다. 두 번째, 테스트할 이미지에 GBVS를 통해서 이미지의 관심영역을 선택하여 테스트한다. 의미연관성 태그와 SIFT기반의 이미지 단어집을 사용한 기존의 방법에 GBVS를 적용한 결과 더 높은 정확도를 보임을 확인하였다.

키워드 : 이미지 단어집, SIFT, 관심영역 자동추출, GBVS, 이미지 분류

1. 서 론

플리커, 페이스북과 같은 소셜 미디어 공유 사이트가 인기를 끌면서 이미지 정보의 양이 급격하게 늘어나고 있다. 많은 양의 이미지 데이터가 웹 공간에 저장됨에 따라 사용자들은 직관적이고 정확한 정보를 얻기를 바라게 되었다.

초기에 웹 이미지 검색은 이미지에 달린 태그를 기반으로 하였다. 하지만 폭소노미 기반의 웹 이미지에는 그 이미지의 내용과 관련 없는 여러 개의 태그가 붙는 경우가 많기 때문에 부정확한 이미지가 검색될 수밖에 없었다. 이에 착안하여 태그기반의 이미지 검색의 정확도를 향상시키기 위해 태그들의 의미적 중요도를 분석하여 이미지 검색에 활용하는 연구가 있었다[1].

하지만 태그의 정보가 부족할 수도 있고 사용자의 주관적인 판단으로 추가되는 태그들이 포함되기 때문에 태그기반 이미지 검색에는 분명한 한계가 있었다. 동시에 컴퓨팅 파워의 빠른 발전으로 이미지 내용기반 검색 방법(content based image retrieval)이 웹 이미지에도 적용되고 있다. 이

※ 이 논문은 2010년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2010-0013307)
† 준 회 원 : 한국교통대학교 컴퓨터정보공학과 석사과정
†† 종신회원 : 한국교통대학교 컴퓨터정보공학과 교수
Manuscript Received : September 17, 2014
First Revision : November 18, 2014
Accepted : November 18, 2014
* Corresponding Author : Cho, Soosun(sscho@ut.ac.kr)

미지 내용기반 검색에서 사용하는 색상비교, 객체의 외곽선 추출, 객체 변화에 강인한 특징점 추출 등 여러 가지 방법들 중에서 BoVW(Bag of Visual Words), 즉 이미지 단어집을 기반으로 분류하는 방법이 주목할 만한 성능을 보이고 있기 때문에 자주 사용된다[3].

BoVW 기반의 이미지 내용 추출 과정은 먼저 이미지에서 특징점들을 추출한다. SURF나 SIFT와 같은 특징 추출하는 알고리즘을 통해 이미지에서 변화에 강인한 특징을 추출한다. 추출된 특징들로 K-means 군집화 과정을 수행함으로써 BoVW를 구성할 수 있다. BoVW는 일종의 이미지 단어 사전으로 볼 수 있는데, 이 BoVW를 이용하여 이미지를 표현한다. 이는 전통적인 텍스트 검색에서 사용하는 BoW (Bag of Word) 방법을 이미지에 적용한 것이다. 군집화에서 클러스터의 개수인 K에 따라 이미지 단어의 수가 결정된다. 이미지를 이 단어들로 표현한 후, SVM과 같은 학습 머신에 훈련 데이터를 학습시키고 테스트 이미지를 분류한다.

그러나 소셜 미디어 공유 사이트에 등록되는 일반적인 이미지에는 정형화된 객체 이미지뿐만 아니라 배경을 포함하는 이미지가 대부분이다. 그렇기 때문에 이미지의 특징적인 부분만 추출하는 연구가 많이 있었다[4]. 본 연구에서는 관심영역을 추출하기 위해 GBVS 모델을 사용하기로 한다. GBVS 모델은 간단하면서도 효율적인 방법으로 관심영역을 추출하는 데 좋은 성능을 보이고 있다[5]. 관심영역이란 불필요한 정보를 갖고 있는 영역이 제거된 영역이다. 따라서 GBVS 모델을 적용해서 이미지를 분류할 때 방해가 되는 영역을 제거할 수 있었고 선택된 관심영역을 테스트하여 정확도를 증가시켰다.

본 논문의 구성은 다음과 같다. 2절에서는 관련된 연구에 대해 소개하고 3절에서는 본 논문에서 제안하는 방법을 사용한 실험을 소개한다. 4절에서는 결과를 분석하고 평가한다. 5절에서는 결론을 맺는다.

2. 관련 연구

2.1 이미지 태그와 내용기반의 이미지 분류

이미지에 관련된 태그를 자동으로 추천하는 기술이나 태그들과 검색어의 의미 연관성을 이용한 검색 방법 등이 활발히 연구되어 왔다[6]. 그 중에서도 위키피디아 기반의 의미 연관성을 활용하여 태그들과의 연관성을 판단하고 이를 이용하여 검색순위를 조정하는 방법이 좋은 성능을 보였다[1]. 이 방법은 위키피디아에 기반하여 검색어와 검색 대상 이미지 각 태그들 사이의 연관성을 코사인 유사도로 계산한 후 그 값이 높은 순으로 대표 태그를 찾는 방법이다.

먼저 위키피디아의 링크된 문서들을 하나의 벡터로 보고 링크들 사이의 가중치를 계산한다. 그렇게 되면 두 벡터 사이의 코사인 유사도 값을 이용해서 유사성을 구할 수 있다. 어떤 문서 s로부터 타깃 문서 t로 아웃링크가 있을 때 이 링크의 가중치 w는 다음 식 (1)과 같다.

$$w(s \rightarrow t) = \log\left(\frac{|W|}{|T|}\right) \text{if } s \in T, 0 \; otherwise \qquad (1)$$

여기서 W는 위키피디아 전체 문서 집합이고, T는 타깃 문서 t를 링크하는 모든 문서들의 집합이다. 위의 식에서 보이는 것과 같이 링크 가중치는 타깃 문서를 링크하는 문서들을 전체 문서집합과 나눈 비율이다. 타깃 문서를 링크하는 비율이 높을수록 해당 소스 문서 내에서 그 링크의 중요도는 떨어진다.

두 문서 사이의 연관성을 계산하기 위해서 아래와 같은 코사인 유사도 식 (2)를 이용한다.

$$\text{similarity} = \cos\theta \frac{A \cdot B}{\|A\|\|B\|} \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A)^2} \times \sqrt{\sum_{i=1}^{n} (B)^2}} \qquad (2)$$

검색어와 검색 대상 이미지 각 태그들 사이의 연관성을 코사인 유사도로 계산한 후 그 값이 높은 순서대로 대표 태그를 선택한다. 의미 연관성을 이용해 플리커 이미지를 검색한 결과 효과적인 검색결과를 보임을 입증했다.

한편, 이미지 내용기반의 분류에서는 BoVW(Bag of Visual Words) 방법이 좋은 성능을 보이고 있다[3]. 이미지에서 추출된 특징점들을 군집화하여 BoVW를 생성한다. BoVW를 기반으로 이미지들의 히스토그램을 생성하고, 각 히스토그램 값을 특징 벡터 값으로 해석하여 SVM(Support Vector Machine)과 같은 분류기에 학습을 시켜 이미지를 분류하는 것이 기본적인 과정이다.

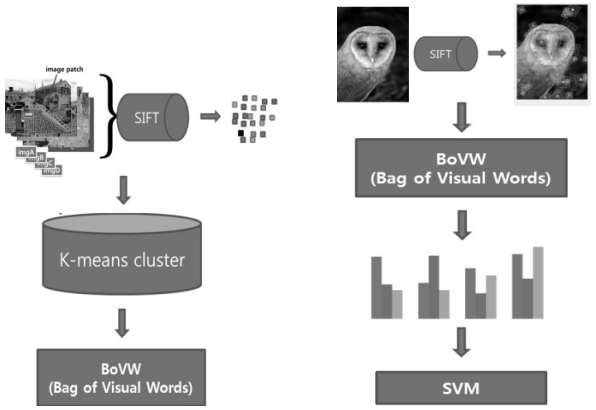

Fig. 1A. BoVW Construction Fig. 1B. Image Classification

Fig. 1. Image Classification based on BoVW

어떤 특징을 추출하느냐와 추출하는 기법에 따라서 이미지 분류의 정확도나 성능이 달라진다. 최근에는 색상, 위치, 크기, 회전 등의 변화에 강인한 이미지 특징점 추출 알고리즘인 SIFT를 사용하는 것이 좋은 성능을 나타내는 것으로 밝혀졌다[7].

SIFT알고리즘을 이용한 BoVW기법으로 효과적으로 이미지의 카테고리를 분류하는 연구가 있었다[2]. 이 방법은 이미지에 적당한 태그가 없거나 태그들이 주관적일 때 성능이 떨어지는 태그기반 검색의 단점을 극복할 수 있다. 하지만 배경과 같은 불필요한 부분도 검색에 사용되기 때문에 검색의 정확도를 떨어트리는 요인이 되기도 한다.

2.2 이미지의 관심영역 자동선택(GBVS)

관심맵(Saliency map)은 이미지 내에서의 특징적인 영역을 구별하여 표현하는 것이다. 관심맵의 설정에 따라 특징적인 것이 적은 영역과 많은 영역으로 나누어진다. 보통 특징적인 것이 많은 영역은 우리가 찾고 싶은 물체나 장소가 된다. 반면에 특징적인 것이 적은 영역은 우리가 관심 없는 배경으로 간주할 수 있다[8].

GBVS(Graph Based Visual Saliency) 모델의 기본은 인간이 시각적으로 이미지를 인식할 때 중요도가 높은 일부의 특징만을 먼저 선별한다는 것이다. GBVS 모델은 시각적으로 중요도가 높은 관심영역을 추출한다. 관심영역을 선택하기 위한 모델은 추출, 활성화, 정규화/조합의 총 3단계로 구성된다.

먼저 추출된 이미지의 특징 벡터들을 사용해서 활성화 지도(Activation maps)들을 형성한다. 그리고 활성화 지도들 중에서 현저함(conspicuity)이 높은 부분들의 조합을 정규화한다. 이때, 마르코프 체인(Markov chains) 알고리즘에 활성화 지도를 정규화시킴으로써 관심영역을 표현할 수 있다[9].

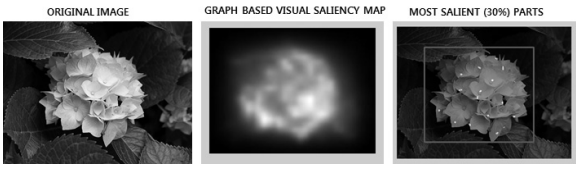

Fig. 2. Salient Area Selection using GBVS

위의 Fig. 2는 이미지에 GBVS 모델을 적용함으로써 30%이상의 불필요한 배경부분을 제거할 수 있음을 보여준다. GBVS 모델로 각 이미지에서 관심영역을 검출하는 작업을 통해 객체 인식의 정확도와 연산시간을 단축시키는 좋은 결과를 얻을 수 있었다[8].

3. 구현 및 실험

구현에서는 SIFT 알고리즘을 사용해 이미지의 특징점(keypoints)을 추출하고 K-means 군집화 알고리즘으로 BoVW(Bag of Visual Words)를 구성하였다. BoVW로 이미지의 히스토그램을 생성하고 SVM(support vector machine) 분류기에 학습을 시켜 테스트 이미지를 분류하였다. 이때, 테스트 이미지에 GBVS 모델을 적용함으로써 불필요한 배경부분을 제거하여 이미지 분류의 정확도를 높였다.

실험에서는 위키피디아 기반의 태그 의미연관성을 이용한 검색으로 수집된 데이터 셋을 이용하였다. 태그들 간의 의미연관성을 비교하여 우선순위에 따라 태그들의 순서를 재배치하는 방법은 웹사이트에 이미지를 업로드하는 시점에 이루어질 수 있으므로 검색에서는 계산량이 현격히 줄어든다.

3.1 SIFT를 이용한 BoVW 기법

구현의 첫 단계는 SIFT 알고리즘을 사용해 이미지의 frame과의 descriptor를 추출하는 것이다. frame은 특징점을 말하는데 (x, y)위치, 특징점의 scale(σ), 각도(Θ) 이렇게 ① x ② y ③ σ ④ Θ 총 4가지의 정보를 가지고 있다.

특징점 주변에 gradients(기울기)를 검출할 수 있는데 특징점 주변 4×4 배열의 히스토그램이 8개의 방향을 가진다. 그렇기 때문에 보통 128bit(4×4×8=128)의 dimension을 사용한다. 8개의 방향으로 나눠진 벡터들을 합하면 keypoint descriptor가 나오게 된다. 방향이 8개인 이유는 SIFT 알고리즘을 최초로 제안한 Lowe의 연구[10]에 의하면 8개일 때가 Correct nearest descriptor의 수치가 가장 높기 때문이다.

특징점을 BoVW로 만들기 위해 k-means를 이용한 클러스터링을 구현한다. k-means 클러스터링은 k개의 센터를 정하고 센터에서 가까운 데이터들을 하나의 집합으로 만들어 가는 것이다. 센터 값을 가진 하나의 집합이 vocabulary (word)가 되는 것이다. 단어의 수를 정하는 것은 휴리스틱한 방법으로 실험을 통해 정할 수 있다.

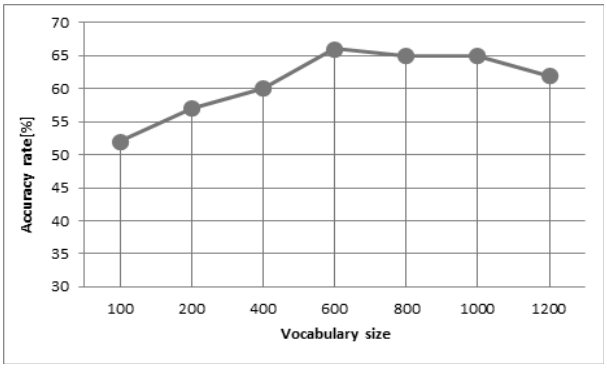

Fig. 3. The evaluation of vocabularies of varying sizes

Vocabulary size를 다양하게 평가해본 결과 600일 때, 좋은 성능을 볼 수 있었다. 그래서 본 연구에서는 단어 수를 600으로 정했다. 이렇게 만들어진 words 집합이 BoVW가 되는 것이다. 이미지 특징 사이에 유사도를 판단하기 위해 유클리드 거리 계산을 이용한다.

BoVW를 가지고 각 이미지에 해당하는 히스토그램을 만들고 히스토그램 값을 특징 벡터값으로 해석하여 SVM 분류기에 넣고 학습을 시킨다. 테스트 이미지의 히스토그램을 만들어 SVM에서 비교함으로써 이미지를 판단했다.

3.2 GBVS 모델 기법

이미지의 특징적인 정보를 지역 형태로 추출하여 관심영역으로 선택할 수 있다. 특징 맵($M: [n]^2 \rightarrow R$)이 주어질 때, 이미지의 지역정보(i, j)에서 활성화 맵($A: [n]^2 \rightarrow R$)을 계산한다. 두 M(i, j)과 M(p, q)의 차이를 정의하기 위해 |M(i, j)-M(p, q)|를 이용한다.

노드(i, j)로부터 노드(p, q)로 가는 엣지는 가중치로 구할 수 있다.

$$w_1((i,j),(p,q)) \doteq d((i,j)|(p,q)) \cdot F(i-p, j-q), where$$
$$F(a,b) \doteq \exp(-\frac{a^2+b^2}{2\sigma^2}) \quad (3)$$

노드(i, j)로부터 노드(p, q)로 가는 엣지 가중치는 차이값에 비례한다. 그래프 모서리 가중치를 정의하고 마르코브 체인 알고리즘을 사용해서 노드들의 가중치를 활성화 지도로 정규화한다.

$$w_2((i,j),(p,q)) \doteq A(p,q) \cdot F(i-p, j-q) \quad (4)$$

이러한 그래프 알고리즘을 통해 간단하고 효율적으로 이미지의 관심영역을 선택할 수 있다.

본 연구에서는 테스트 이미지에서 관심영역을 선택한 후 BoVW로 히스토그램을 구성하고 SVM을 통해 이미지를 분류하였다.

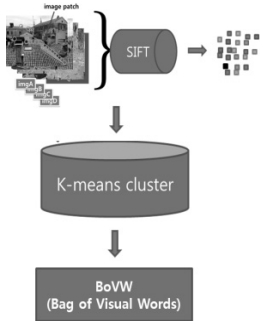

Fig. 4A. BoVW Construction

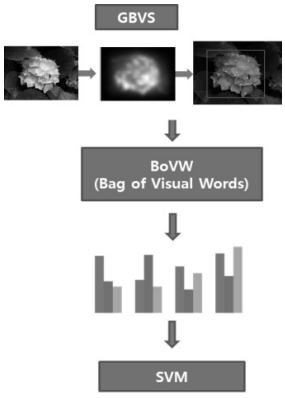

Fig. 4B. Image Classification using GBVS

Fig. 4. Image Classification based on Proposed Method

실험에서는 먼저, 전통적인 이미지 분류의 성능을 테스트할 때 사용되는 데이터 집합인 Caltech101 dataset[11]을 이용하여 제안하는 알고리즘의 정확도를 평가했다. Caltech101 dataset에는 101개의 카테고리가 있고 총 9,146개의 이미지가 있다. Caltech101 dataset은 California Institute of Technology에서 2003년 9월에 Computer Vision 기술개발에 기여하기 위해 만들어진 정제된 이미지 집합이다[11].

Table 1에서 보이는 것처럼 이미 정제된 Caltech101 dataset이미지를 분류해본 결과 94.00%의 높은 정확도를 보이고 있다.

Table 1. Classification using Caltech101 dataset (correctness: 94.00%)

Category	correct images	rates
accordion	10	100%
airplanes	10	100%
anchor	7	70%
ant	10	100%
barrel	10	100%
합계	47	94.00%

하지만 본 연구는 플리커 등 대용량 이미지 공유 사이트에서 볼 수 있는 정제되지 않은 이미지들을 대상으로 하였으므로 본격적인 실험에서는 위키피디아 기반의 의미정보를 이용하여 이미지 태그들의 우선순위를 찾아 이미지를 수집했던 기존 연구를 활용하였다[1]. 기존 연구에서는 실제 인터넷 상에서 검색된 이미지를 분류 대상으로 하기 위해서 플리커 이미지들을 사용했다. 기존 연구의 태그 의미 연관성으로 얻어진 이미지 데이터를 사용해서 총 5,005개의 이미지를 수집하였고, bird(새), car(자동차), cup(컵), house(집), sea(바다) 다섯 개의 카테고리별로 각 700개씩 총 3,500장을 임의로 뽑아 훈련 이미지로 사용하였다. 테스트 이미지로는 각 카테고리에서 임의로 30개씩 뽑아 총 150개의 이미지를 대상으로 분류하는 실험을 하였다.

Fig. 5. 1,001 bird images retrieved from Tag's semantic relations using Wikipedia

4. 결과분석 및 평가

Table 2. Compare the classification performance BoVW with BoVW+GBVS

Category	correct images		rates	
	BoVW	BoVW+ GBVS	BoVW	BoVW+ GBVS
bird	16	17	53.33%	56.66%
car	20	21	66.66%	70.00%
cup	28	29	93.33%	96.66%
house	17	18	56.66%	60.00%
sea	15	14	50.00%	46.66%
합계	96	99	64.00%	66.00%

Table 2는 각 카테고리의 1,001개의 이미지 중 700개를 훈련 데이터로 사용하고 30개씩 임의의 이미지를 합쳐 150개의 이미지를 대상으로 분류 실험을 비교한 결과이다. 'bird' 카테고리는 30개 중 17개를 분류하여 56.66%, 'car'는 30개 중 21개를 분류하여 70.00%, 'cup'은 30개 중 29개를 분류하여 96.66%, 'house'는 30개 중 18개를 분류하여 60.00%, 'sea'는 30개 중 14개를 분류하여 46.66%의 정확도를 나타냈다. 그 결과 평균 66.00%의 정확도를 나타냈다.

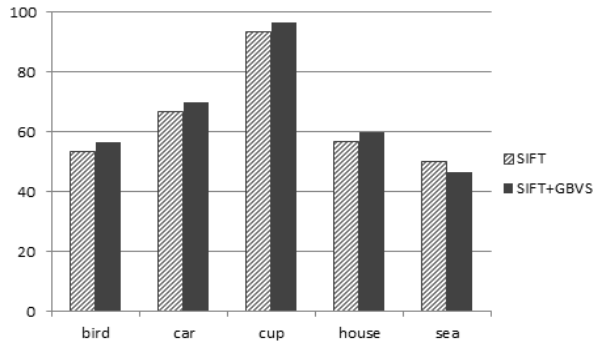

Fig. 6. BoVW vs. BoVW+GBVS

Fig. 6은 SIFT 기반 BoVW를 사용한 분류 결과와 본 연구에서 제시한 GBVS를 추가로 사용한 이미지 분류 결과를 비교한 것이다. Fig. 6에서 빗금 친 그래프인 BoVW보다 BoVW+GBVS를 사용한 것이 전반적으로 좋은 성능을 나타낸다. 하지만 'sea'와 같이 배경이 오브젝트인 경우 BoVW만을 사용한 것보다 성능이 낮아짐을 알 수 있었다. 그 이유는 GBVS는 배경을 제거하기 때문에 sea 이미지에서 충분한 특징점을 얻지 못했기 때문이다. BoVW+GBVS를 사용한 결과 이전 연구보다 전반적으로 좋은 성능을 나타냄을 알 수 있었고 특히 이미지 검색에서 검색어가 장면(scene)을 나타내지 않고 특정 객체(object)를 나타내는 경우에는 효과적일 것임을 알 수 있었다.

5. 결 론

본 연구에서는 위키피디아 기반의 태그 의미 연관성을 이용하여 검색된 이미지들을 훈련 데이터로 사용하였다. 이는 일반 플리커 이미지를 사용할 때보다 더 정확하게 분류 머신을 학습시키기 위한 것이다. 또한 제한된 훈련 데이터셋이 아닌 실제 대용량 이미지 공유 사이트에서 얻을 수 있는 확장 가능한 이미지 데이터셋을 이용하기 위해서이다.

이전 연구를 통해 SIFT 기반의 BoVW를 사용해 이미지 분류를 효과적으로 할 수 있다는 것이 입증되었는데, 본 연구에서는 BoVW+GBVS를 같이 사용해서 정확도를 더 높이는 결과를 보였다. 또한 카테고리의 종류가 객체인지 장면인지에 따라 다른 결과가 나타남을 알 수 있었다.

References

[1] S. J. Lee and S. Cho, "Tagged Web Image Retrieval Re-ranking with Wikipedia-based Semantic Relatedness," Journal of Korea Multimedia Society, Vol.14, No.11, pp.1491-1499, 2011.

[2] H. J. Jeong, J. M. Lee, and J. H Nang, "Image Categoriztion Using SIFT Bag of Word," Korea Computer Congress, pp.1277-1279, 2013.

[3] H. W. Jang and S. Cho, "Flickr Image Classification using SIFT Algorism", the KIPS Spring Conference, Vol.20, No.2, 2013.

[4] R. Bharath, L. Zhi, J. Nicholas and X. Cheng, "Scalable scene understanding using saliency-guided object localization," in *Proceedings of ICCA*, pp.1503-1058, 2013.

[5] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *Proceedings of NIPS*, pp.545-552, 2006.

[6] D. H. Kweon, J. H Hong, and S. Cho, "Web Image Retrieval using Prior Tags based on WordNet Semantic Information," Journal of Korea Multimedia Society, Vol.12, No.7, pp. 1032-1042, 2009.

[7] A. Vedaldi and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms", *Proceedings of the international conference on Multimedia*, New York, pp.1469-1472, 2010.

[8] G. E. Kalliatakis and G. A. Triantafyllidis, "Image based Monument Recognition using Graph based Visual Saliency", Electronic Letters on Computer Vision and Image Analysis, Vol.12, No.2, pp.88-97, 2013.

[9] Z. W. Tu and S. C. Zhu, "Image Segmentation by Data-Driven Markov Chain Monte Carlo", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.24, No.5, pp.657-673, 2002.

[10] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", Journal International Journal of Computer Vision, Vol.60, No.2, pp.91-110, 2004.

[11] Caltech 101 Dataset [Internet], http://www.vision.caltech.edu/Image_Datasets/Caltech101/ (검색일: 2013. 08. 16)

장 현 웅

e-mail : jhwskorg@gmail.com

2013년 한국교통대학교 컴퓨터정보공학과 (학사)

2013년~현 재 한국교통대학교 컴퓨터정보공학과 석사과정

관심분야 : Web Image Mining

조 수 선

e-mail : sscho@ut.ac.kr

1987년 서울대학교 계산통계학과(학사)

1989년 서울대학교 계산통계학과(석사)

2004년 충남대학교 컴퓨터과학과(박사)

1994년~2004년 한국전자통신연구원 소프트웨어연구소 선임연구원

2004년~현 재 한국교통대학교 컴퓨터정보공학과 교수

관심분야 : Data Mining & Information Retrieval