시각 단어 가중치 부여 방법을 이용한 장면 영상 분류

황천섭[O,1] 박운상[3] 낭종호[3] 이성환[1,2]
[1]고려대학교 컴퓨터·전파통신공학과, [2]고려대학교 뇌공학과, [3]서강대학교 컴퓨터공학과
hcs64648@korea.ac.kr, {unsangpark, jhnang}@sogang.ac.kr, sw.lee@korea.ac.kr

Scene Classification using Visual Word Weighting

Cheon-Seob Hwang[O1] Unsang Park[3] Jongho Nang[3] Seong-Whan Lee[1, 2]
[1]Department of Computer and Radio Communications Engineering, Korea University
[2]Department of Brain and Cognitive Engineering, Korea University
[3]Department of Computer Science and Engineering, Sogang University

요 약

장면 영상 분류는 멀티미디어 자산 관리와 같은 여러 응용 분야를 위해 컴퓨터 비전 분야에서 활발히 연구되고 있다. 대부분의 장면 영상 분류 방법에서 이미지의 표현을 위해 BoVW(Bag-of-Visual-Words) 모델을 이용해 왔다. BoVW은 간편하고 효율적인 이미지 모델링 방법으로 널리 이용되고 있으나 공간 정보의 손실이나 분류 성능의 부족과 같은 한계를 가지고 있다. 본 논문에서는 BoVW의 단점을 개선하여 영상 분류의 성능을 높이기 위해 공간 피라미드 매칭(Spatial Pyramid Mathching) 방법에 기반한 로컬 특징 추출과 특징 단계 융합법(Feature-Level Fusion)을 이용하고 각 시각 단어의 중요도에 따라 차등적 가중치를 부여하는 시각 단어 가중치 부여 방법(Visual Word Weighting)을 제안한다. 시각 단어의 중요도에 따라 산출된 가중치는 BoVW 모델을 이용해 얻은 히스토그램(Histogram)에 적용되어 장면 영상 분류 시스템의 성능을 높이는 데 이용된다.

1. 서론

최근 자동 이미지 분류 기술은 컴퓨터 비전 분야에서 많은 관심을 얻고 있다. 특히 웹의 발전에 따라 기하급수적으로 증가하는 다량의 미디어 자산을 효율적으로 관리하기 위해 장면 영상의 자동 분류 기술이 다양하게 연구되고 있다. 대부분의 장면 영상 분류 방법들에서 효율적인 이미지 모델링을 위해 BoVW(Bag-of-Visual-Word) 모델[1]을 이용하고 있다. 하지만 BoVW 방법은 모델링의 과정에서 공간적 정보를 대부분 손실한다는 점과 모델 자체가 가지는 분류 능력에 한계가 있다는 점과 같은 단점을 가지고 있다.

이러한 단점을 극복하기 위해 다양한 연구가 시도되고 있는데, BoVW가 가지는 공간적 정보의 손실을 극복하기 위해 공간 피라미드 매칭(Spatial Pyramid Matching) 기술이 개발 되어 이용되고 있다[2]. 공간 피라미드 매칭 방법은 이미지를 단계별로 여러 개의 부분 공간으로 분할하여 부분 히스토그램을 생성하고 각 부분 히스토그램들을 결합하여 최종 히스토그램을 생성하는 방법으로 BoVW의 공간 정보 손실을 부분적으로 해결하여 장면 영상 분류의 성능을 크게 향상시켰다. BoVW가 가지는 분류 능력을 증진시키기 위한 또 다른 시도로 시각 단어에 가중치를 부여하는 방법이 있다. 문서의 모델링에서 널리 이용되는 TF-IDF(Term Frequency-Inverse Term Frequency)를 이용한 방법[1], 분류 성능을 최대화하는 최적 가중치 학습 방법[3] 등이 그 예이다.

본 논문에서는 BoVW의 분류 능력의 개선을 위해 각 시각 단어에 대응하는 클러스터가 가지는 레이블 복잡도(Label Complexity)[4]와 구성 성분들의 집약도(Compactness)를 이용한 새로운 시각 단어 가중치 부여 방법을 제안한다. 각 단어가 가지는 고유의 중요도에 따라 차등적인 가중치를 부여함으로 BoVW 히스토그램의 분류 능력을 증진시킬 수 있게 된다. 장면 영상 분류 시스템의 성능을 향상시키기 위해 시각 단어 가중치 부여 방법 외에도 영상 분류를 위한 로컬 특징들의 구성과 이들의 특징 단계 융합(Feature-Level Fusion)을 이용한다. 로컬 특징들의 특징 단계 융합은 커널 단계 융합(Kernel-Level Fusion) 방법에 비해 매우 간단하게 구현되어 사용될 수 있음에도 그와 유사한 성능을 얻을 수 있는 장점을 가진다. 단, 전체 벡터의 차원이 증가하게 되어 학습과 분류에 소요되는 시간을 증가시킬 수 있으나 히스토그램 인터섹션(Histogram Intersection) 커널을 이용한 SVM(Support Vector Machine) 학습을 통해 이를 완화할 수 있다.

2. 제안하는 방법

제안하는 방법의 전체 시스템 구성은 그림 1과 같다. 먼저 각각의 이미지로부터 로컬 특징들을 추출하고 공간 피라미드 방법에 기반한 BoVW 모델링을 통해 히스토그램 벡터 형태로 표현한다. 다음으로 BoVW 모델링으로 얻어진 시각 단어들을 이용하여 그들의 중요도에 따른 가중치를 계산하고 히스토그램 벡터에

곱하여 적용한다. 마지막으로 얻어진 로컬 특징 별 히스토그램 벡터들을 특징 단계의 융합법을 이용하여 최종 벡터 형태로 변경하고 이를 히스토그램 인터섹션 커널 SVM을 이용하여 학습한다.

2.1 특징 추출 및 융합

본 논문에서는 특징 추출을 위해 3가지의 외형 특징, SIFT(Scale Invariant Feature Transform), ORB(Oriented FAST and Rotated BRIEF), BRISK(Binary Robust Invariant Scalable Keypoints)와 1가지의 텍스처 특징, LBP(Local Binary Patterns)를 이용한다.

장면 영상 분류에서는 특징 점(Interest Point)을 이용한 방법에 비해 밀집 샘플링(Dense Sampling)을 기반으로 한 방법이 대부분의 경우 우수한 성능을 보이기 때문에, 3가지의 외형 특징, SIFT, ORB, BRISK에 대해 16x16 픽셀 크기의 패치와 4 픽셀 단위의 기준선을 이용한 밀집 샘플링을 이용하여 특징을 추출한다. 얻어진 각각의 특징들에 대하여 총 3단계로 이루어진 공간 피라미드 매칭과 300개의 시각 단어들을 구성하는 BoVW 모델링을 적용하여 히스토그램 벡터들을 생성한다. LBP의 경우 이미지 전체에서 나타나는 텍스처 특징으로부터 얻어진 공간적인 정보를 손실 없이 모두 포함하기 위하여 BoVW 모델링 없이 직접적으로 사용한다.

위의 과정으로 얻어진 총 4개의 로컬 특징 들은 각각의 시각 단어 가중치를 적용한 후 간단한 특징 연결(Feature Concatenation) 방법을 통해 하나의 융합된 특징 벡터로 표현된다. 융합된 특징 벡터는 분류기 학습을 위한 최종 학습 벡터 형태로 이용된다.

2.2 시각 단어 가중치 부여 방법

본 논문에서는 각 시각 단어의 중요도에 따라 차등적인 가중치를 적용함으로써 BoVW 모델링의 분류 능력을 증진시키는 방법을 제안한다. 시각 단어의 중요도를 측정하기 위한 방법으로 클러스터를 이루는 구성 성분들의 레이블 복잡도와 집약도를 이용한다.

레이블 복잡도는 클러스터가 얼마나 다양한 클래스 레이블을 가지는 성분들로 구성되어 있는지를 측정하기 위한 목적으로 엔트로피 모델[4]을 이용하여 계산된다. 각 클러스터의 엔트로피는 총 시각 단어의 수가 V, 전체 클래스의 수를 C, 클러스터 i ($i=1,...,V$)의 성분의 수를 N_i 그리고 클래스 j를 레이블로 가지는 클러스터 i의 성분의 수를 n_{ij}라고 할 때, 아래의 식 (1)과 같이 계산된다.

$$E_i = -\sum_{j=1}^{C} \frac{n_{ij}}{N_i} \log \frac{n_{ij}}{N_i} \quad (1)$$

엔트로피 E_i는 클러스터의 모든 성분들이 하나의 클래스에만 속해 있을 경우 최저값인 0을 가지게 된다. 클러스터의 레이블 복잡도가 낮을수록 분류에 유리하므로 해당 시각 단어에 더 높은 가중치를 부여한다.

두 번째 중요도 측정법인 클러스터의 집약도는 성분들이 클러스터의 중심에 얼마나 집약되어 있는지를 측정한다. 언어적 단어와 달리 시각 단어는 클러스터링의 결과로 추정되어 만들어지므로 단어가 가지는 신뢰성이 일정하지 않게 된다. 본 논문에서는 클러스터링이 얼마나 잘 되었는지에 대한 신뢰성의 측정을 집약도를 이용하여 계산하고 높은 집약도를 가지는 클러스터가 단어로서의 높은 명확성과 표현성을 가질 것으로 기대하여 높은 가중치를 부여한다. 클러스터 i의 j번째 성분을 x_{ij}, 중심점을 w_i라고 할 때, 집약도 CP_i는 유클리드 거리를 이용하여 아래의 식 (2)와 같이 계산된다.

$$CP_i = \frac{1}{1+\frac{1}{N_i}\sum_{j=1}^{N_i}\left\|x_{ij}-w_i\right\|^2} \quad (2)$$

클러스터의 모든 성분들이 중점과 같은 위치에 존재할 때 집약도는 1로 최대값을 가진다. 최종적으로 클러스터 i의 가중치를 얻기 위해 엔트로피와 집약도의

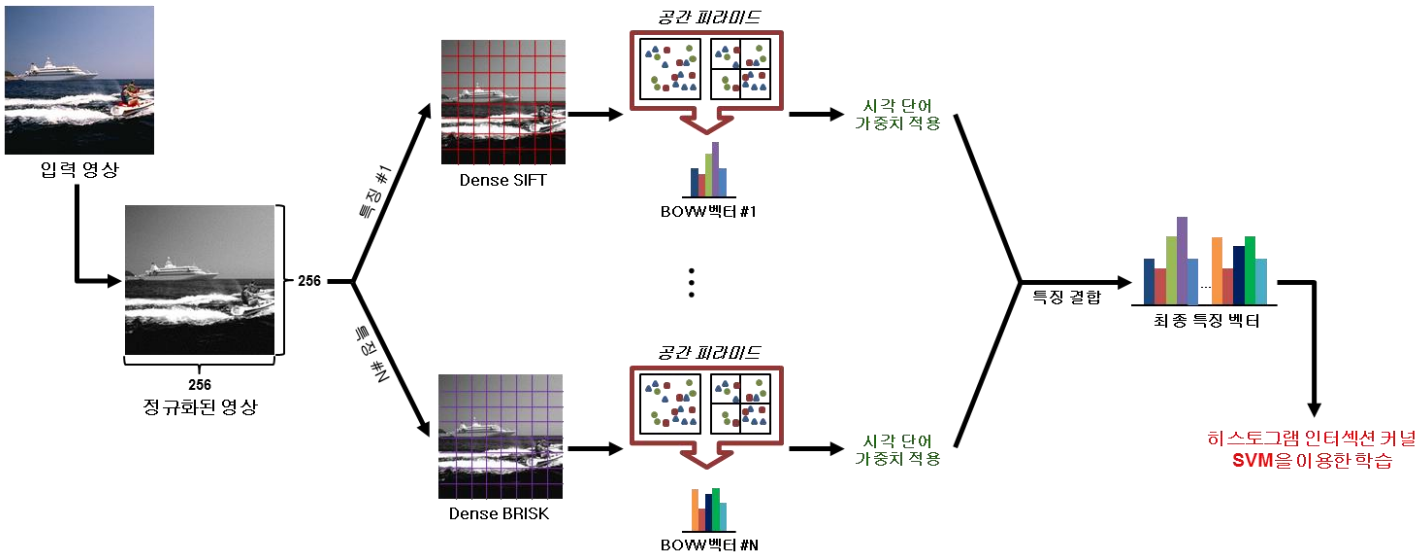

그림 1 장면 분류 시스템의 전체 학습 과정 개관도

결과 값을 아래의 식 (3)과 같이 결합한다.

$$wp_i = \alpha\left(1-\frac{E_i}{\log C}\right)+\beta\frac{CP_i}{\max CP_j},\quad j=1,...,V \quad (3)$$

α 와 β 는 엔트로피와 집약도의 영향력을 조절하는 파라미터이고 $\log C$ 와 $\max CP_j$는 정규화를 위해 사용된다. 계산된 가중치는 각 시각 단어에 대응되는 히스토그램 빈(bin)의 값에 곱하여 이용된다.

3. 실험 및 결과

본 논문에서 제안하는 장면 영상 분류 방법을 검증하기 위해 15개 클래스와 4,485개의 이미지로 구성된 Scene-15 공용 데이터셋[2]을 이용하여 실험을 진행하였다. 각 클래스당 100개의 이미지를 임의로 선택하여 학습에 이용하였고 남은 이미지를 테스트에 사용하였다. 실험은 총 5회 반복하여 진행하였고 각 클래스별 분류율의 평균을 이용하여 성능을 측정하였다.

표 1은 각 특징 추출 방법들에 대한 분류 실험의 결과이다. 3가지의 외형 정보를 추출하는 로컬 특징, SIFT, ORB, BRISK는 공간 피라미드 방법을 기반한 BoVW 방법을 적용하였을 때 각각 기존보다 10% 가량 분류 성능이 상승하는 것을 확인하였고 텍스처 정보를 추출하는 LBP 특징은 BoVW 모델링 없이 전체적 특징으로 이용하였을 때 다른 특징들보다 우수한 성능을 보이는 것을 통해 장면 영상 분류에 적합한 특징 추출 방법임을 확인할 수 있었다.

표 1 로컬 특징에 따른 분류 성능 비교(%)

특징 종류	분류율(평균±표준편차)
SIFT	64.33±0.41
ORB	63.66±0.68
BRISK	66.47±0.49
SIFT+공간 피라미드	74.92±0.51
ORB+공간 피라미드	73.80±0.41
BRISK+공간 피라미드	74.90±0.52
LBP	77.44±0.43

다음으로 시스템의 분류 성능을 높이기 위해 제안된 특징 단계 융합 방법과 시각 단어 가중치 부여 방법을 적용한 실험을 진행하였다. 표 2는 그 실험 결과를 나타낸다. 공간 피라미드를 이용한 SIFT, ORB, BRISK와 LBP 특징을 하나의 벡터 형태로 융합하여 분류에 이용하였을 때 각각의 특징 별 분류 능력이 상호 보완되어 전체적인 성능을 증진시키며 80% 이상의 높은 분류율을 기록하였다. 또한 각 시각 단어의 중요도를 제안한 방법으로 계산하여 히스토그램 벡터에 곱하여 사용하였을 때 특징 융합을 사용한 분류 성능을 추가적으로 증진시킬 수 있었다.

표 2 특징 융합과 시각 단어 가중치 실험 결과(%)

방법	분류율(평균±표준편차)
특징 단계 융합	81.75±0.32
시각 단어 가중치	82.87±0.31

4. 결론 및 향후 연구

본 논문에서는 장면 영상 분류 시스템의 성능을 향상시키기 위해 분류에 적합한 로컬 특징 추출 및 특징 단계의 융합 방법을 적용하였고 시각 단어들의 중요도를 레이블 복잡도와 집약도의 두 가지 측정 방법으로 산출하여 BoVW 히스토그램 벡터의 가중치로서 사용하는 방법을 제안하였다. 제안된 방법은 공용 데이터셋을 이용하여 성능을 측정하였고 각 방법의 적용이 분류 성능을 점진적으로 증가시켜 최종적으로 높은 분류율을 얻었다. 향후에는 여러 분류기를 상호보완적으로 혼합하여 사용하는 부스팅 방법을 적용하여 분류 시스템의 성능을 개선하는 연구를 진행할 계획이다.

감사의 말씀

본 연구는 미래창조과학부/한국산업기술평가관리원 산업융합원천기술개발사업의 지원을 받아 수행된 연구 결과임(10047058, 멀티미디어 데이터 소비/유통 활성화를 위한 서비스 컴포넌트 기반 스마트 미디어 자산 관리 기술 개발).

참고 문헌

[1] J. Sivic and A. Zisserman, “Video Google: A Text Retrieval Approach to Object Matching in Videos,” *IEEE International Conference on Computer Vision*, pp. 1470-1477, 2003.

[2] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169-2178, 2006.

[3] H. Cai, F. Yan, and K. Mikolajczyk, “Learning Weights for Codebook in Image Classification and Retrieval,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2320-2327, 2010.

[4] J. Hou and M. Pelillo, “A Simple Feature Combination Method based on Dominant Sets,” *Pattern Recognition*, Vol. 46, No. 11, pp. 3129-3139, 2013.