ETRI

2014
Electronics and
Telecommunications
Trends

소프트웨어 기술동향 특집

영상 빅데이터 분석기술 동향

Technologies Trends in Image Big Data Analysis

고종국 (J.G. Ko) 분석소프트웨어연구실 선임연구원
배유석 (Y.S. Bae) 분석소프트웨어연구실 책임연구원
박종열 (J.Y. Park) 분석소프트웨어연구실 실장
박경 (K. Park) 빅데이터 SW 플랫폼연구부 부장

최근에 스마트폰, CCTV, 블랙박스, 고화질 카메라 등으로부터 수집되는 영상 데이터의 양이 급격히 증가하고 있어 이에 따른 비정형 영상 빅데이터를 기반으로 인물이나 사물 등을 인식하여 의미있는 정보를 추출하고 내용을 시각적으로 분석하고 활용하기 위한 요구사항이 증대되고 있다. 영상 빅데이터 분석기술은 이러한 대규모 영상들에 대해 학습 및 분석을 수행하여 원하는 영상을 검색하거나 이벤트 발생 등의 상황인식을 위한 제반 기술들을 말한다. 본고에서는 영상인식을 위한 학습기술 및 영상 빅데이터 분석기술의 현황 및 관련 이슈들에 관하여 살펴보고자 한다.

Ⅰ. 서론

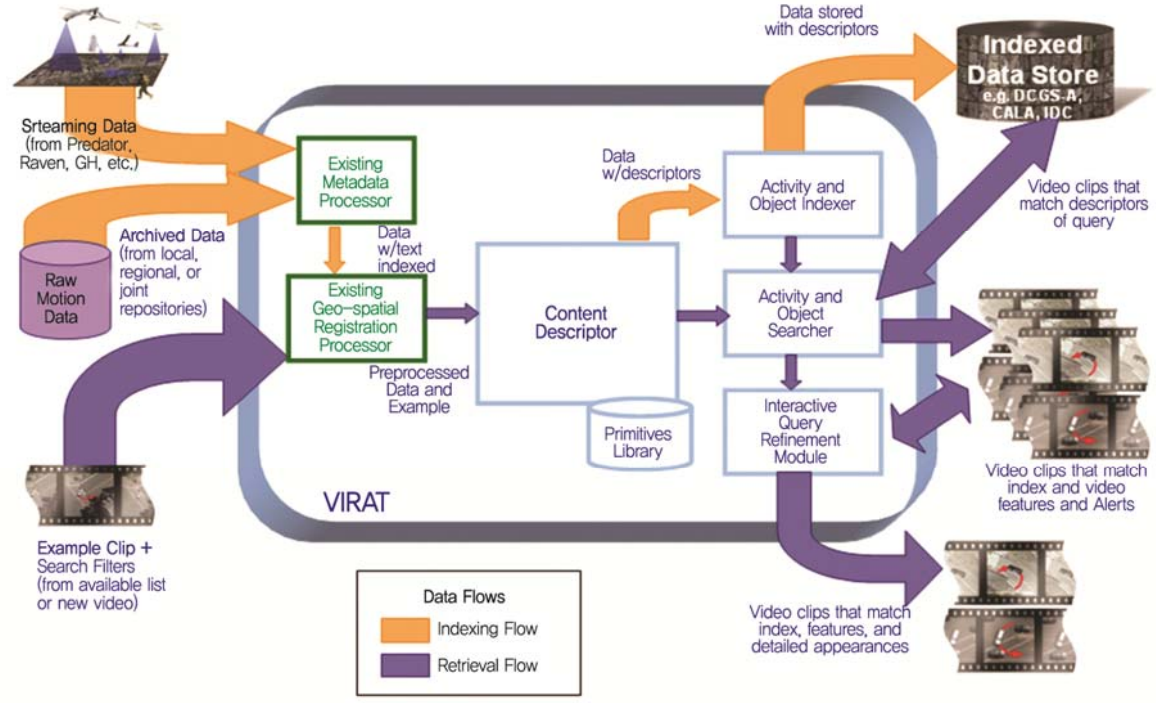
많은 전문가들이 예견하고 있듯이 스마트폰, CCTV, 블랙박스, 드론, 인공위성, 디지털 카메라 등에서 수집되는 영상 데이터의 양은 기하급수적으로 증가하고 있으며, 이에 따른 비정형 영상 데이터를 인식하고 내용을 분석하여 활용할 수 있는 기술 요구가 점차 증대되고 있다. 이미지, 비디오, 오디오와 같은 멀티미디어 데이터는 인터넷 트래픽의 60%, 모바일 폰 트래픽의 70%, 이용 가능한 비정형 데이터의 70% 이상을 차지하고 있을 정도로 급증하고 있으며, 웹 사용자는 분당 72시간 분량의 비디오를 YouTube에 업로드하고 있고, 소셜 미디어 사용자는 평균적으로 하루에 3억개의 사진을 페이스북에 포스팅하고 있다.

또한, 최근 들어 하드웨어 및 소프트웨어 기술의 급격한 발전으로 인해 대용량의 데이터를 수집하고 빠른 시간 내에 분석 및 처리할 수 있는 기반 환경이 구축되고 있으며, 다양한 빅데이터 관련 솔루션들이 점차 실생활에 접목되고 있다.

대규모 영상 데이터와 영상 분석기술의 만남은 기존의 영상 자체에 대한 인식의 범위를 뛰어넘어 의미 있는 정보추출과 내용분석 등을 통하여 더욱 발전된 기능을 제공함으로써 새로운 가치를 창출하고, 보다 나아가 미래 변화를 예측하며 능동적으로 대처할 수 있는 좋은 기회를 제공할 것으로 예상된다.

영상의 내용을 이해하는 기술은 미국을 중심으로 많은 연구가 진행되고 있으며, (그림 1)과 같이 상황의 이해와 행동을 분석하여 다음에 발생할 상황을 예측하는 기술로까지 이어지고 있다. 현재 가장 널리 알려진 기술은 DARPA 주도의 프로젝트로 VIRAT(Video and Image Retrieval and Analysis Tool) 과제[1]가 대표적이다.

본고에서는 최근 이슈화된 영상 빅데이터 들의 영상분석기술 기술동향 및 관련 이슈들에 대해 살펴보고자 한다.



(그림 1) DARPA 영상분석 프로젝트 개요도

II. 기술현황

1. 영상인식을 위한 학습기술

가. Labeled data 기반 감독 학습기술

감독 학습 기반의 영상인식기술은 labeled data를 가지고 특징을 추출하여 영상을 분류하는 학습을 수행하는 방식으로 영상을 인식/분류 하는 기능을 수행한다((그림 2) 참조).

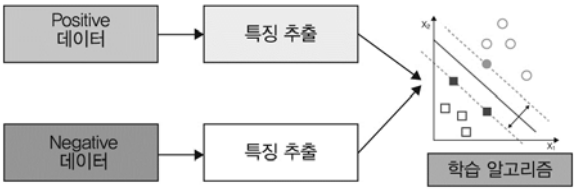
대표적인 학습 알고리즘은 SVM(Support Vector Machine)으로 분류하려는 데이터들을 구분하는 마진을 최대로 하여 분류하는 기능을 수행한다.

대표적인 예로 IBM IMARS(IBM Multimedia Analysis and Retrieval System) 시스템[2]은 레이블된 이미지들에 대해 Color 히스토그램, SIFT, HOG 등의 다양한 특징들을 추출한다. 이 특징들을 SVM, 신경망, GMM 등의 감독학습 알고리즘들에 이용하여 영상을 인식/분류하는 기능을 제공한다.

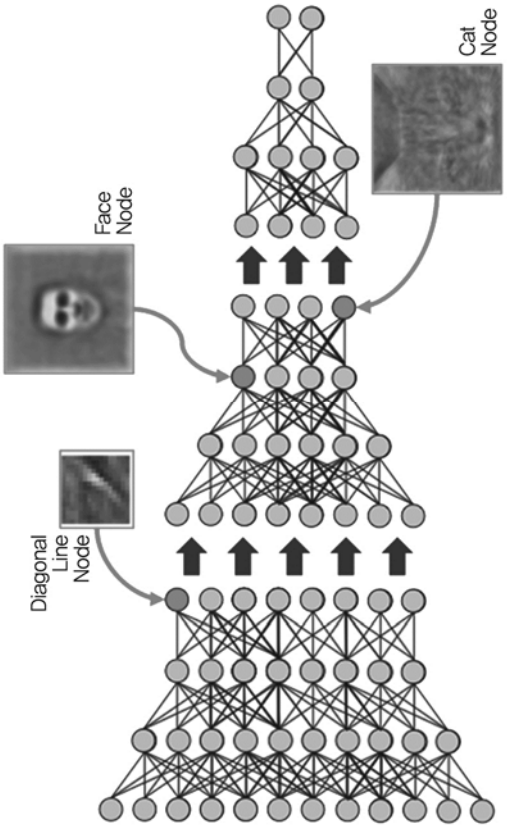
감독학습 기술에 있어서 특징 추출 및 학습 알고리즘과 함께 중요한 것이 레이블된 학습 데이터들의 수집에 있다. 많은 레이블된 학습 데이터들이 존재하면 학습은 효과적으로 될 수 있다. 하지만 현실적으로 많은 레이블 데이터들을 수집하는 데는 한계가 있다.

나. Unlabeled data 기반 비감독 학습기술

최근에 Deep learning 기술로 객체를 인식/분류하는 기술이 이슈화가 되고 있다. Deep learning은 여러 개



(그림 2) Labeled 데이터 기반 감독학습



(그림 3) Unlabeled 데이터 기반 비감독 학습

의 레이어들로 구성된 multi-layer 네트워크이다. 각 레이어들은 이전 레이어들의 출력을 입력으로 받고 상위 레벨의 특징들을 생성한다. 이 기술은 토론토 대학의 Hinton 교수[3]에 의해 주목 받기 시작했는데 이 기술은 기존의 신경망과 달리 unlabeled 데이터들을 입력값으로 하고 비감독 학습방식으로 상위 레벨의 특징들을 추출하는 학습을 수행 된다((그림 3) 참조). 이렇게 여러 단계의 학습을 하나로 통합하여 전체 학습 네트워크를 구성한다. 대표적인 예로 구글 시스템[4]은 천만장의 unlabeled 이미지들에 대해 지역적 신경망을 구성하여 비감독 학습을 수행하고 이러한 학습과정을 한 단계씩 쌓아 올려 전체 multi-layer 네트워크를 구성하였다. 이러한 deep learning 기반의 영상인식기술은 최근에 객체인식 및 음성인식 등에 좋은 성능을 보여주고 있다.

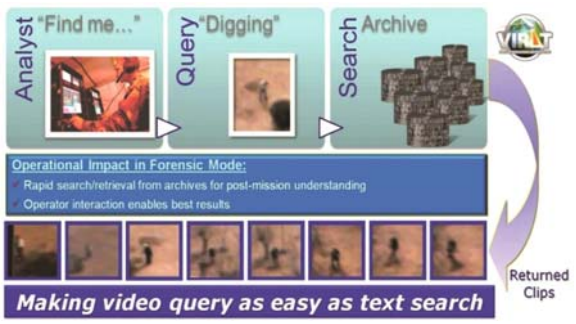
2. 영상검색 및 내용분석 기술

미국 DARPA에서는 VIRAT 프로젝트[1]를 수행하고 있다((그림 4) 참조). 대규모의 영상정보 데이터베이스를 구축하고 content-based searching 기능을 통해 빠른 속도로 대용량 영상 콘텐츠를 검색하는 기능을 수행한다. VIRAT의 목적은 수천 시간의 동영상 데이터베이스에서 다음과 같은 형태들이 발생하는 것을 검색하는 도구를 제공하는 것이다.

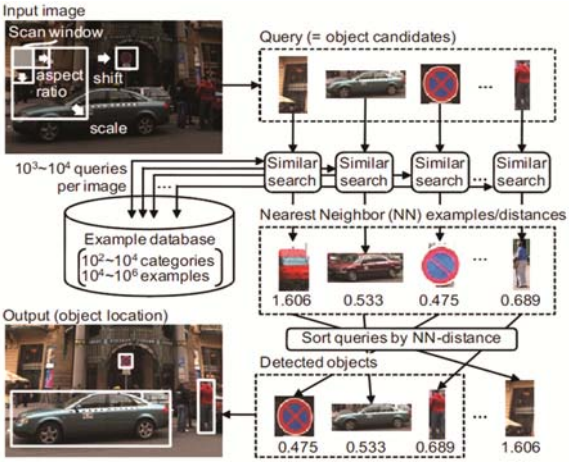
- Single Person: 배회, 투기, 걷기, 달리기 등
- Person-to-Person: 미팅, 악수, 물건교환, 군집, 해산 등
- Person-to-Vehicle: 운전, 승차, 하차, 태우기 등
- Person-to-Facility: 들어가기, 나오기, 서있기 등
- Vehicle: 턴, 정차, 차량 군집이동, 차량화재 등

일본 히타치 연구소[5]에서는 영상 빅데이터 기술과 관련하여 이미지가 포함되어 있는 유사한 장면을 자동으로 검출하고 해당 장면으로 이동하고, 영상으로부터 특정 객체를 검출하며, 추출한 영상에 대하여 주석(annotation)을 자동으로 태깅하는 3가지의 기술 요소를 제시하였다.

- 유사 이미지 검색기술: 쿼리 이미지에 비슷한 이미지를 데이터베이스에서 찾아오는 기술로, 이 기술을 이용하며 대량의 영상 데이터 중에서 원하는 이미지를 추출할 수 있음. 히타치에서는 특징량 벡터 클러스터링을 기반으로 고속 유사 벡터 검색기법을 사용하여 유사 이미지 검색기능을 제공하고 있음((그림 5) 참조).
- 객체 탐지 기술: 이미지 중에서 사람의 얼굴이나 자동차 등의 객체 영역을 식별하는 기술로, 입력 이미지의 부분 영역과 검출 대상의 사례 이미지를 유사 이미지 검색의 특징량 기준으로 일치하는 객체 영역을 검출함. 이 기술을 사용해서 점포 내에서 인원수를 세거나 이상 행동의 탐지, 대량의 감시 영상에서 특정 장면을 찾을 수 있음.
- 이미지 주석 기술: 이미지가 나타내는 내용에 해당하는 메타 데이터를 자동으로 부여하는 기술로 주어진 이미지 쿼리에 대하여 유사 이미지 검색을 하고 검색결과의 이미지에 나오는 텍스트의 단어를 확률적 지표에 의해 평가하여 특별한 사전학습 없이 이미지에 의미를 부여하는 키워드를 추정할 수 있음.

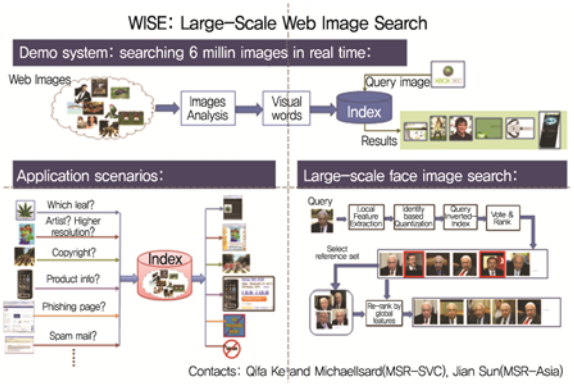


(그림 4) VIART 개요도

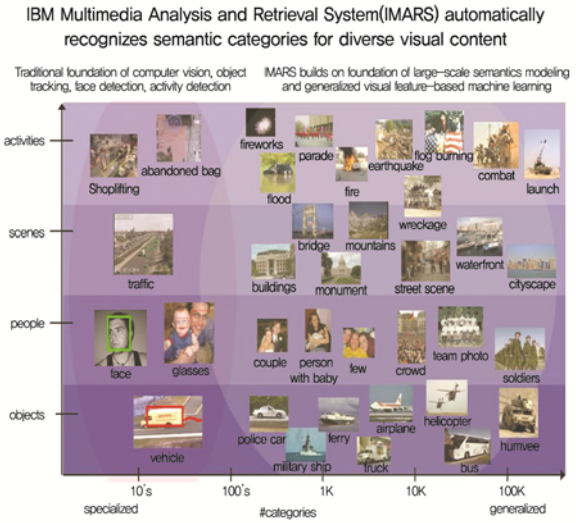


(그림 5) 유사 이미지 검색 기반 객체 검출

Microsoft에서는 대규모 웹 이미지 검색과 탐색을 위한 WISE(Web Image Search and Exploration) 프로젝트[6]를 통하여 이미지 재현을 위한 대규모 기계학습 및 효율적인 이미지 인덱싱과 질의방법을 개발하고 있으며, 프로젝트 내에서 컨텐츠 기반 이미지 검색을 위해 인덱싱과 스케일러블 이미지 재현 및 알고리즘을 개발하고 Bing 검색엔진에 활용하여 10억개 이상의 이미지에 대한 인덱싱과 검색기능을 지원하고 있다((그림 6)



(그림 6) 마이크로소프트 WISE 시스템 구성도



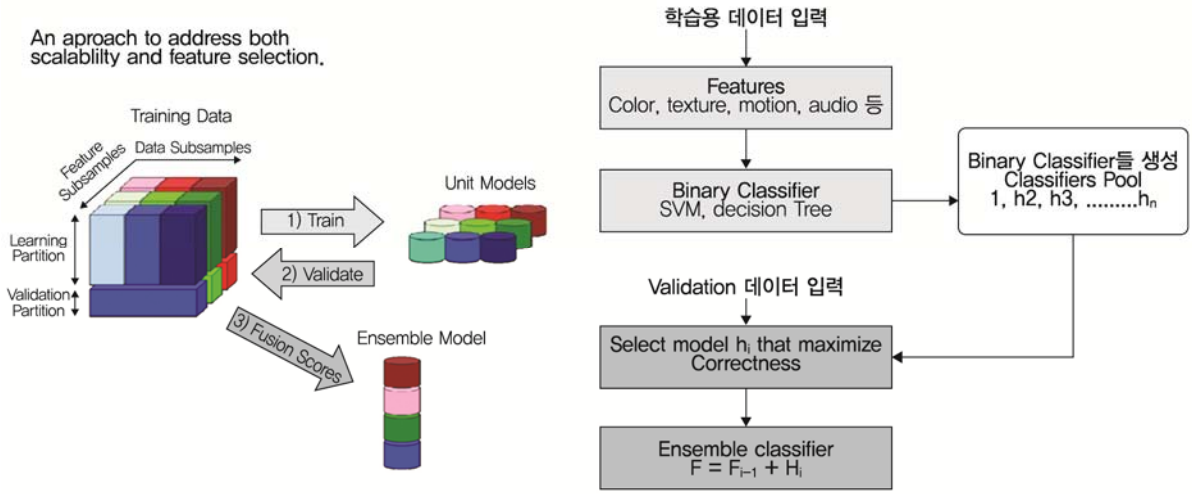
(그림 7) IBM IMARS 개요도

참조). 또한, 웹 스케일 얼굴 이미지 인식과 검색기능을 제공하는 얼굴 특징과 인덱싱을 위한 파이프라인을 개발하였다. 추가로, 대규모 웹 이미지를 클러스터링 하기 위해 부분 복제 웹 이미지를 찾는 효율적인 해싱 알고리즘을 개발하였다.

IBM은 2000년 초반부터 현재까지 이미지 검색 및 이벤트 탐지를 위한 IMARS 시스템[2]을 개발하고 있으며 매년 영상검색 및 이벤트 탐지 평가를 위한 TRECVID (TREC Video Retrieval Evaluation) 학회에 개발제품을 평가해오고 있다((그림 7) 참조).

좁은 범위에서는 영상에서 사람, 자동차 등의 객체를 탐지하는 것으로부터 넓은 범위에서는 사람의 이상행동, 화재, 전쟁 등의 이벤트/액션 등을 탐지하는 형태로 진행되고 있다. IBM 시스템은 레이블된 입력 데이터들에 대해 감독학습방식으로 데이터들에 대한 학습 알고리즘을 수행한다. 학습을 위한 특징들을 다름과 같은 특징들로 여러 가지 특징들을 사용하여 학습을 수행한다.

- 전역적 특징: Color Histogram, Color Moments, Color Correlogram 등
- 지역적 특징: Scale Invariant Feature Transform, Local Binary Patten, Histogram of Oriented



(그림 8) IBM IMARS 시스템 학습 알고리즘 구성도

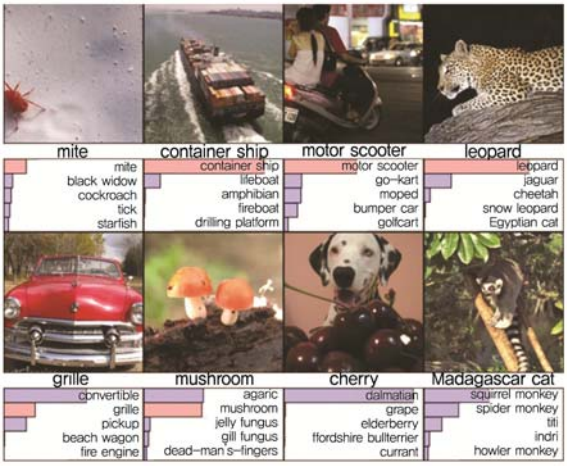
Gradients 등

추출된 각각의 특징들에 대한 학습 분류기로는 SVM(Support Vector Machine), GMM(Gaussian Mixture Model) 등의 여러 학습방법을 사용하여 각각의 유닛 모델들을 생성한다. 이렇게 생성된 유닛 모델들은 검증과정을 거쳐 최적의 앙상블 분류기로 결합된다. (그림 8)은 이 과정을 설명한다.

구글 시스템[4]은 deep learning 방식으로 unlabeled 데이터를 입력으로 받고 비감독 학습방법으로 학습을 수행하여 멀티 레이어 네트워크를 구성한다. 시스템은 3단계로 구성되었고 각 단계는 각각 3개의 레이어들로 구성되어 총 9개 레이어들로 이루어졌다. 전체 시스템은 총 1B개의 파라미터들로 네트워크를 구성하였다. 구글 시스템은 이동, 회전, 스케일링 등의 지역적 왜곡을 극복하기 위해 L2 Pooling과 Local Contrast Normalization 방법을 사용하였다.

구글은 학습을 위해 유튜브에서 천만개의 동영상을 수집하고 이들에서 각각 한장씩 이미지를 랜덤하게 추출하여 학습에 사용하였다((그림 9) 참조). 구글 시스템은 단지 unlabeled된 데이터로만 학습을 수행하였음에도 불구하고 테스트 영상에 대해 81.7%의 얼굴 인식 성능을 보였다.

최근 토론토 대학[7]에서는 대규모 이미지 검색을 위해 deep learning 방식인 Deep Convolutional Neural Networks 기술을 개발하여 탁월한 객체 검색기술 성능을 보였다. 이 시스템은 7개의 hidden 레이어로 구성되



(그림 10) 토론토 대학의 ImageNet 시험결과

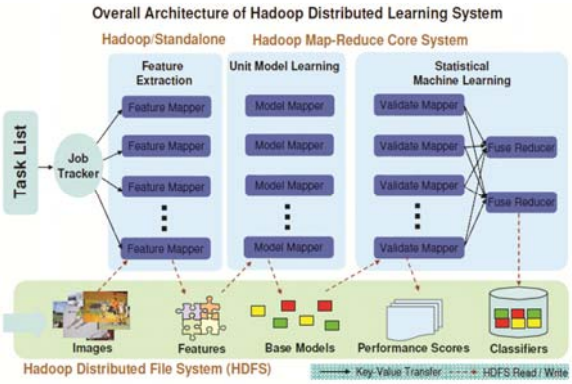
고 60M개의 파라미터들로 구성된 멀티 네트워크를 생성 하였다. 이 시스템은 1,000개의 클래스들로 구성된 ImageNet 시험 데이터에 대해 5순위 내 검색결과가 83%의 성능을 보였다. (그림 10)은 검색결과를 보여주는데 입력된 쿼리 영상에 대해 5순위 검색 결과를 보여준다. 막대그래프의 크기는 검색된 확률 크기값을 표시한다.

3. 영상 빅데이터 분산/병렬처리

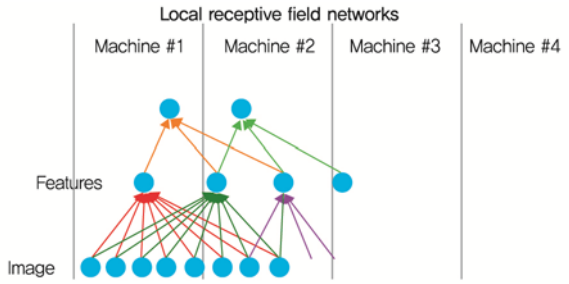
IBM에서는 대규모 영상 검색을 위한 시스템 개발에 있어서 하둡(Hadoop) 기반의 분산/병렬처리 시스템을 적용하여 개발하였다((그림 11) 참조).

버지니아 대학[8]에서는 영상 빅데이터에 대한 분산 컴퓨팅 처리를 위한 API(Application Program Interface)

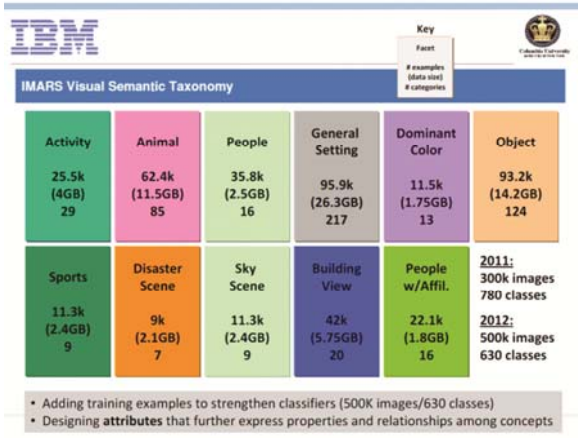
(그림 9) 구글 시스템 학습 알고리즘 구성도



(그림 11) 하둡 기반 IBM 시스템 구성도



(그림 13) 구글 시스템 학습 병렬처리 구성도



(그림 14) IBM 영상검색 데이터베이스

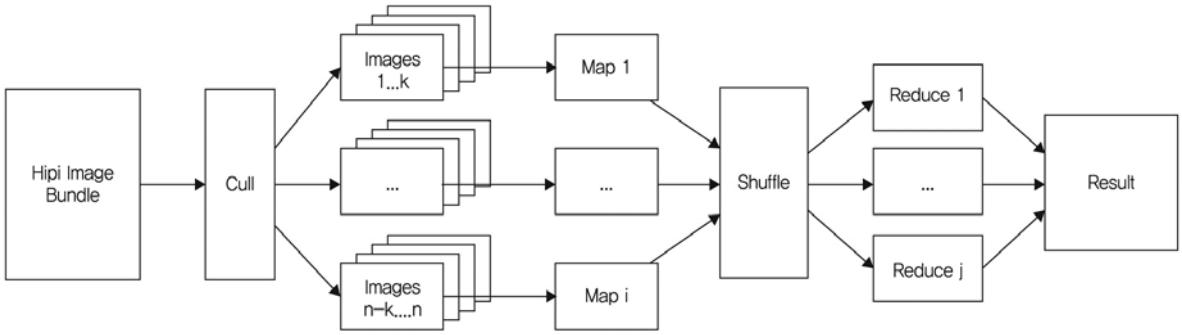
를 제공하는 하둡 MapReduce 라이브러리인 HIPI(Hadoop Image Processing Interface for image-based map-reduce Tasks) 프레임워크를 개발하고 있다((그림 12) 참조). HIPI는 MapReduce 프레임워크 기반으로 영상처리 및 비전 응용프로그램에 개방적이고 확장 가능한 라이브러리를 제공한다. 사용자가 MapReduce 프레임워크의 자세한 내용을 파악할 필요없이 영상 기반 분산/병렬처리가 가능하도록 지원한다.

구글 시스템은 천만개의 200x200 이미지들에 대해 1B의 파라미터들을 학습하기 위해 16개의 코어들을 가지고 있는 1,000개의 머신을 이용하여 병렬처리를 수행하였다((그림 13) 참조).

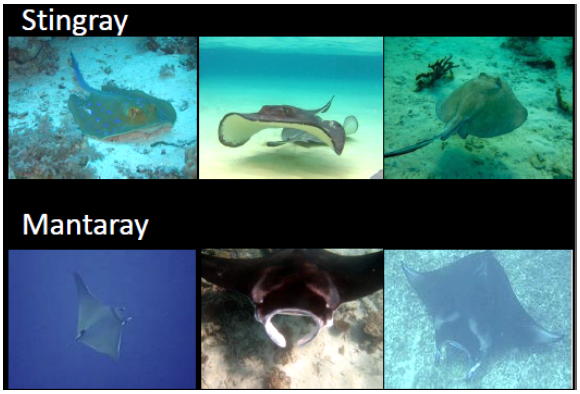
4. 영상 빅데이터 데이터베이스

IBM[9]은 영상검색을 위한 데이터베이스를 구성하였는데 2012년에는 500K개의 이미지들로 구성된 총 630개의 클래스들로 이루어졌다((그림 14) 참조). 각 사각형에서 맨 위는 각 클래스들을 의미하고 아래 숫자들은 순서대로 샘플의 수, 데이터 사이즈, 그리고 카테고리 개수를 의미한다.

스탠포드 대학의 ImageNet[10]은 이미지 검색을 위



(그림 12) 버지니아 대학의 HIPI 라이브러리 구성도



(그림 15) ImageNet 데이터베이스 예제

(그림 16) MIT 데이터베이스

한대규모 영상 데이터베이스로 2012년에는 총 16M images들로 구성된 22,000개의 범주로 나누어져 있다. (그림 15)는 이미지 샘플들을 보여준다.

MIT는 대규모 장면 인식과 분류를 위한 SUN(Scene Understanding) 데이터베이스[11]를 구축하고 벤치마크 자료를 공유하고 있다. 웹으로부터 장면 관련 이미지를 유형별로 수집하여 데이터베이스로 구축하고, 정제된 397개 카테고리를 사용하여 최대 908개 카테고리까지 확장이 가능한 특징이 있다. 최대 확장 가능한 카테고리는 분류가 가능하다는 것을 의미하며 인식의 정확도를 고려하여 908개 중에서 상위 397개 카테고리를 사용하고 있다((그림 16) 참조).

Ⅲ. 결론

본고에서는 영상 빅데이터 분석을 위한 관련 학습기술들과 관련 기술 현황 및 주요 이슈들에 대해 살펴보았다. 최근 다양한 소스로부터 다양한 형태의 비정형 영상 데이터들의 증가에 대한 영상 분석기술의 접목을 통해 여러 다양한 발전기능을 제공하고 있음을 알 수 있었다. 또한, 영상 빅데이터 처리를 위한 분산/병렬처리 등의 기술의 요구사항이 증가하고 있음을 알 수 있었다.

추후, 대규모 영상 데이터들의 증가는 기하급수적으로 늘어날 것으로 예상된다. 따라서 이를 지원하기 위한 좀더 일반화된 학습 모델의 개발 및 대용량 처리를 위한 하드웨어 및 플랫폼 개발이 많이 이루어져야겠다.

용어해설

영상검색 저장된 데이터 셋으로부터 입력된 쿼리 이미지와 유사한 이미지를 찾아주는 기술

하둡(Hadoop) 대용량 데이터를 분산처리하게 해주는 오픈 소스 프로젝트

약어 정리

API	Application Program Interface
GMM	Gaussian Mixture Model
HIPI	Hadoop Image Processing Interface for image-based map-reduce Tasks
IMARS	IBM Multimedia Analysis and Retrieval System
SUN	Scene Understanding
SVM	Support Vector Machine
TRECVID	TREC Video Retrieval Evaluation
VIRAT	Video and Image Retrieval and Analysis Tool
WISE	Web Image Search and Exploration

참고문헌

[1] DARPA, “BAA-08-20: Video and Image Retrieval

and Analysis Tool (VIRAT)," Mar. 3th, 2008.

[2] IBM Multimedia Analysis and Retrieval System, http://mp7.watson.ibm.com/imars

[3] G. E. Hinton "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, 2006.

[4] G. E. Hinton, S. Osindero, and Y.-W. Teh "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, July 2006, pp. 1527-1554.

[5] IT Pro, http://itpro.nikkeibp.co.jp/article/COLUMN/20121012/429404

[6] Microsoft, Web Image Search and Exploration (WISE), http://research.microsoft.com/en-us/projects/WISE

[7] A Krizhevsky, "ImageNet Classification with Deep Convolutional Neural Networks," *NIPS*, 2012.

[8] HIPI, University of Virginia, "HIPI : Hadoop Image Processing Interface," http://hipi.cs.virginia.edu

[9] IMB Research-Columbia University, "Semantic Indexing Task," 2012.

[10] Stanford, ImageNet, http://www.image-net.org

[11] MIT, SUN database, http://groups.csail.mit.edu/vision/SUN/