사운드 분류기를 이용한 영상검색에 관한 연구

A Study on Image Retrieval Using Sound Classifier

김승한, 이명순, 노승용
(KIM SEUNG HAN, LEE MYEONG SUN, ROH SEUNG YONG)

Abstract - The importance of automatic discrimination image data has evolved as a research topic over recent years. We have used forward neural network as a classifier using sound data features within image data, our initial tests have shown encouraging results that indicate the viability of our approach.

Key Words : music speech classification, image retrieval , audio signal processing, neural network

1. 장 서론

1.1절 개 요

최근 디지털 영상과 비디오 데이터의 양이 늘어감에 따라 이를 색인하고 검색할 수 있는 방법이 필수적으로 요구되고 있다. 특히 디지털 방송 관리, 주문형 비디오,디지털 라이브러리, 의료진단 시스템, 홈쇼핑등 멀티미디어 데이터를 다루는 대부분의 서비스뿐 아니라 최근에는 인터넷의 검색 기술의 발전과 함께 영상 검색시스템의 개발에 대한 시도와 함께 영상 정보 검색에 대한 수요가 나날이 증가하고 있다. 최근의 영상 정보검색 기술은 영상을 분석하여 특징을 추출한 다음, 이를 색인화 하고 유사한 특징을 가지는 영상을 찾아내는 기술이다.

기본적으로 영상간의 유사도 판단을 위해서는 영상이 가진 내용을 파악하여 그 정보들을 서로 비교 가능한 데이터로 변환하여야 한다. 따라서 영상정보의 검색 기술은 궁극적으로 영상 분석(image analysis), 영상 이해(image under standi -ngs), 또는 객체 인식(object recognition) 기술과 함께 발전해 나가야 한다. 그러나 현재로서는 이러한 기술들의 수준이 영상 검색에 사용할 만큼 좋은 성능을 보이지 못하고 있다. [5][6]

이에 본 논문에서는 일반적으로 영상 데이터에 있는 음성신호 정보를 활용하여 영상의 검색, 분류를 하는 방법을 제안함으로써 영상검색 시스템의 성능의 특징요소로서 활용할 뿐 아니라 그 차제로도 좋은 영상 분류 시스템을 제안한다.

1.2절 연구방법

본 논문은 크게 두 가지의 분류과정으로 구성되어 진다. 첫째로 그림 1과 같이 동영상 데이터로부터 사운드 데이터를 추출한 후 추출한 사운드 데이터에서 1000ms 단위로 사운드 데이터를 나누어서 2장에서 살펴볼게 될 특성에 대한 특징벡터를 추출한다. 여기서 해당 구간에 해당하는 특징 벡터의 값을 이용해서 크게 사운드를 음성, 비음성, 환경음으로 분류하는 사운드 분류과정을 거치게 된다. 비음성으로는 효과음, 음악등 유성음 계통의 사운드를 의미한다.

이와 같은 방법으로 전구간의 사운드 데이터 샘플에 대한 분류작업을 완료하면 전 사운드 데이터 구간에 대해서 음성, 비음성, 환경음의 시간과 발생 빈도에 대한 데이터를 구할 수 있다. 그런 다음 음의 발생시간과 발생패턴에 대한 자료를 기반으로 2차 동영상 분류작업을 통해서 사운드 데이터의 비중과 발생 빈도를 입력으로 동영상을 분류와 검색에 도움이 되도록 한다.

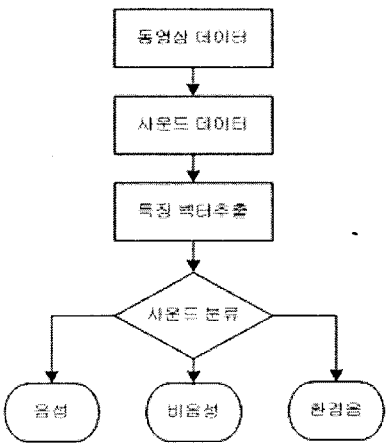

그림 1. 1차 사운드 분류를 위한 흐름도

2. 장 사운드 분석 기법 및 결과

우선 1차 사운드 분류와 세그먼테이션의 정확성을 얻기 위해서 적절한 특징벡터를 선택해야 한다. 그 동안의 연구 결과에 비추어서 zero-crossing rate ratio(ZCR), 단구간 에너지함수(STE, short time energy)[1], and spectrum flux(SF)을 사용하였다. 분류방법으로는 음성과 비음성의 분류에는 KNN(K-nearest neighbor) 분류기를 사용하였으며 비음성과 환경음의 분류에는 규칙기반(rule-based) 기법을 이용하였다.

2.1절 Zero-Crossing Rate Ratio(ZCR)

ZCR은 이산시간신호에서 연속된 샘플의 부호가 다를 경우를 나타내며 다음과 같이 정의 된다.

$$Z_n = \frac{1}{2}\sum_m |sgn[x(m)] - sgn[x(m-1)]|w(n-m)$$
$$sgn[x(n)] = \begin{cases} -1, x(n) \geq 0 \\ 1, x(n) < 0 \end{cases} \quad (1)$$

여기서 x(n)은 이산시간신호이며 w(n)은 길이 N인 윈도우(window) 함수이다. ZCR 시간에 대한(sample index)에 대한 변화값은 그림 2와 같다. 평균ZCR은 무성음(unvoiced)이 유성음(voiced)보다 일반적으로 높은 ZCR값을 가지기 때문에 유성음과 무성음(unvoiced)을 특성을 구별하는 데에 좋은 특징 벡터가 된다. 특히 음성신호(speech signal)의 경우는 무성음과 유성음의 반복으로 구성되는 반면 음악(music) 데이터일 경우는 그렇지 않기 때문에 음성신호에서 일반적으로 높은 ZCR 값을 가지게 된다.

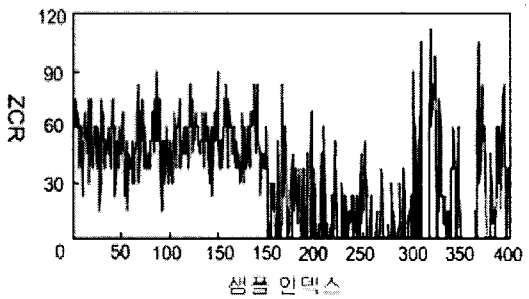

그림 2. 사운드 데이터의 구간별 ZCR값의 변화

그림 2. 에서 음성 세그먼트는 0~200샘플 구간이며 30에서 90사이의 ZCR값을 가진다. 음악세그먼트는 200~300 샘플 구간이며 낮은 ZCR의 값을 나타낸다. 그리고 마지막으로는 환경음에 대한 구간으로 잡음과 다양한 오디오 특성을 나타내는 신호들의 조합으로 구성되어있다. 따라서 변동이 큰 구간으로 ZCR 값을 가지게 된다.

2.2절 단구간 에너지 함수(STE)

단구간 에너지 함수는 다음식과 같이 정의된다.

$$E_n = \frac{1}{N}\sum_m [x(m)w(n-m)]^2$$
$$w(n) = \begin{cases} 1, 0 \leq n \leq N-1 \\ 0, otherwise \end{cases} \quad (2)$$

여기서 x(n)은 이산시간신호이며 n은 샘플값, w(m)은 윈도우 함수이다. 단구간 에너지는 다음과 같이 크게 다음의 두 가지 목적으로 사용된다. 첫째, 음성신호의 경우 ZCR과 같이 무성음과 유성음을 구분하는 특징벡터로서 사용된다. 무성음 구간에서는 STE 값이 유성음구간에 비해서 현저히 낮아지기 때문이다. 둘째, SNR(signal to noise, 신호대잡음비)가 높을 경우에 STE는 묵음구간을 판단하는 근거로서 사용된다.[4]

2.3절 Spectrum Flux(SF)

SF은 신호의 스펙트럼의 변화량을 평균한 값으로 다음과 같이 정의된다.[2]

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}[\log(A(n,k)+\delta) - \log(A(n-1,k)+\delta)]^2$$
$$A(n,k) = \left|\sum_{m=-\infty}^{\infty} x(m)w(nL-m)e^{j\frac{2\pi}{L}km}\right| \quad (3)$$

여기서 A(n,k)는 입력신호의 n번째 구간의 이산 푸리에 변환(Discrete Fourier Transform, DFT)이면 x(m)은 원(origin) 오디오 데이터, w(m)은 윈도우 함수이다. L은 윈도우 길이, K는 DFT의 차수, N은 총 구간의 합을 의미한다.SF값을 이용한 실험결과 음성신호가 비 음성신호에 비해서 높은 SF값을 보였고 특히 환경음의 경우는 ZCR과 유사하게 SF값의 급격한 급등락이 나타나는 특징을 나타냈다.

2.4절 분류 방법 및 결과

사운드 분류시스템은 다음과 같다. 동영상으로부터 얻어진 사운드 신호는 샘플링(16KHz)과 특징 추출 단계를 거치게 된다. 특징 추출 단계에서는 2장에서 살펴본 바와 같이 ZCR, STE, SF값을 계산하게 된다. 계산된 ZCR, STE, SF는 KNN분류기를 통해서 음성과 비음성으로 판별되며 여기서 비음성으로 판별시에는 규칙기반 기법으로 주어진 구간의 오디오 신호가 비음성인지, 환경음 인지를 판별하게 된다.[3][4] 다음은 음성, 비음성, 환경음에 대한 분류기의 실험결과이다. 실험데이터는 44.1KHz와 모노 채널 그리고 16bit/sample의 사운드 데이터이며 3시간 분량의 학습구간과 5시간 분량의 실험구간으로 나누어서 분류기 테스트를 하였다.

저자 소개

김승한 : 서울시립대학교 전자전기컴퓨터공학부 博士課程

이병순 : 서울시립대학교 전자전기컴퓨터공학부 博士課程

노승용 : 서울시립대학교 전자전기컴퓨터공학부 教授

음성	비음성	환경음
3.25±1.55	5.98±3.54	11.98±2.51

표 1. 사운드 분류 시스템의 에러율과 편차

3. 장 동영상 분류기법과 결과

동영상으로부터 추출된 사운드 신호에서 2-3장에서 분류된 사운드 분류기법을 적용하여 동영상은 음성, 비음성, 환경음의 발생 빈도와 시간으로 표시가 가능하였다. 예를 들면, 강의 동영상의 경우는 음성의 비중이 크고 비음성이나 환경음의 비중이 상대적으로 작게 나타난다. 따라서 본 논문에서는 위와 같은 패턴과 다음의 전방향 신경망(feed-forwad NN)을 통해 분류기를 학습하는 방법을 사용하였다. 분류기의 다이어그램은 다음과 같다.

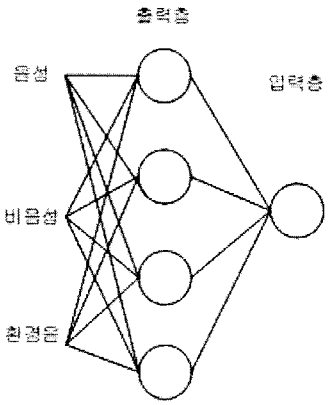

그림 4. 전방향 신경망 다이어그램

다음은 전방향 신경망을 이용한 동영상 분류 시스템의 실험결과이다. 실험은 방송프로그램 사에서 강의, 영화와 드라마, 음악, 토론으로 구분해 놓은 동영상을 1시간 분량으로 각각 100개씩 임의로 선택하였고 각 동영상의 샘플링 데이터는 32KHz, 16bit 모노로 샘플링 되었다. 이와 같이 선택된 영상에서 추출된 사운드 데이터 중 60% 신경망 학습, 40% 실험의 데이터셋(data set) 과 70% 신경망 학습, 30% 실험의 데이터 셋으로 나누어서 테스트 하였다.

뉴런의수	학습세트	정확도
5	60%	82.50%
10	60%	88%
5	70%	93%
10	70%	98%

표 2. 신경망 분류기의 정확도 (단위: %)

4. 장 결론

이번 논문에서는 사운드 데이터의 특성(ZCR, STE, SF)을 이용하여 동영상에서 사운드 데이터를 추출하여 동영상의 분류와 검색을 자동적이고 효율적으로 전개하는 방법에 대해서 제안 하였다. 표2의 실험결과에서와 같이 문서의 분류의 정확도가 70% 학습 데이터 셋에서는 90%이상의 정확성을 보여주었다. 앞으로의 연구는 사운드 분류기의 정확도와 함께 음성인식과 사운드 인식 등의 알고리즘을 접목하여 영상 분류기의 성능을 향상 시킬 수 있는 방법에 대한 연구가 필요함과 동시에 기존의 영상인식, 검색, 분류기법에서의 하나의 특징 벡터로 분류기법을 사용하여 더 나은 영상인식 방법을 만들어 내는 연구도 필요할 것으로 보인다. 또한 실시간 스트림의 경우 전체 오디오 데이터를 받을 수 없는 상황에서의 분류가 이루어져야 하므로 실시간 상황에 적절한 오디오 분류기의 연구도 병행되어야 할 것이다.

참 고 문 헌

[1] Scheirer, E. and Slaney, M., "Construction and Evaluation of A Robust Multifeatures Speech / Music Discriminator", ICASSP 97, Vol.2, 1997

[2] L.Lu, H.Jiang, and H.J. Zhang, "Content Analysis for Audio Classification and Segmentation", IEEE Trans. Speech and Audio Processing, Vol. 10, October 2002

[3] Foote, J "Content-based retrieval of music and audio", Proceedings of SPIE 97, Dallas, 1997

[4] T.Zhang, C.-C.Jay Kuo, "Heuristic Approach for Generic Audio Data Segmentation and Annotation"ACM Multimedia, Vol 1. pp.67-76, November,1999

[5] 노현기, 황본우, 문종섭, 이성환, "내용기반 영상검색 정보기술의 현황", 대한 전자공학회 논문지, 8호, 제 25권, pp. 798-806, 1998. 8.

[6] ISO/IEC FDIS 15938-4, "Part4: Audio," MPEG-7, June 2001

[7] C.Panagiotakis, G.Tziritas, "A Speech/Music Discriminator Based on RMS and Zeros-Crossings", IEEE Trans. Multimedia, February, 2005