감시 시스템에서의 비정상 소리 탐지 및 식별

주영민*, 이의종**, 김정식**, 오승근**, 박대희*
*고려대학교 컴퓨터정보학과
**고려대학교 컴퓨터정보학과
e-mail : ymj0704@hanmail.net, {kongjjagae,misia,gmo85,dhpark}@korea.ac.kr

Abnormal Sound Detection and Identification in Surveillance System

Young-min Joo*, Eui-jong Lee**, Jeong-sik Kim**,
Seung-geun Oh**, Dai-hee Park*
*Dept of Database&Mining Lab., Korea University
**Dept of Database&Mining Lab., Korea University

요 약

본 논문에서는 감시카메라 환경에서 취득한 오디오 데이터를 입력으로 하여, 비정상 상황을 인식하는 시스템을 제안한다. 제안된 시스템은 단일클래스 SVM의 대표적인 모델인 SVDD와 최근 얼굴 인식 분야에서 성공적인 업적을 보여주고 있는 신호 처리 분야의 SRC를 계층적으로 결합한 구조로써, 첫 번째 계층에서는 SVDD로 비정상 소리를 신속하게 탐지하여 관리자에게 알람 경고하고, 두 번째 계층의 SRC는 탐지된 비정상 소리를 유형별로 세분화 식별하여 관리자에게 비상 상황을 보고함으로써 관리자의 위기 상황 대처를 돕는다. 제안된 시스템은 실시간 처리가 가능하며, 점증적 갱신의 학습 능력으로 인하여 비정상 오디오 데이터베이스의 변화에도 능동적으로 적응할 수 있다. 실험을 통하여 제안된 시스템의 성능을 검증한다.

1. 서론

최근 공공장소나 주요 시설물에서의 도난과 같은 범죄를 비롯한 위험물 투기 및 방치 등 다양한 특수 범죄들이 빈번히 발생하고 있다. 따라서 유동인구가 많은 공항, 버스정류장, 지하철역 등과 같은 공공장소에서의 응급상황이나 보행자 안전 문제 등이 크게 대두되고 있는 실정이다[1].

최근의 연구동향에 따르면, 보안 감시 분야의 연구는 감시 카메라로부터 획득한 멀티미디어 데이터로부터 원하는 장면 및 정보를 찾는 단순한 보안 검색 시스템에서부터, 객체의 움직임을 적극적으로 검출하고 위험 상황을 미리 인지하여 알람(alert)을 통해 관리자에게 알리거나, 자동으로 객체를 인식 및 식별하고 추적하는 등 고수준의 의미 정보를 찾기 위한 연구들 까지 다양하게 진행되고 있다[2]. 이러한 연구들 중 본 논문에서는 오디오 서베일런스 환경에서의 비정상 상황인식에 관한 연구를 대상으로 한다.

오디오 정보로부터 응급 상황을 탐지하는 연구영역에서는 전통적으로 GMM(Gaussian mixture model)과 같은 확률론적 패턴인식 알고리즘이 주도하고 있으나, 최근에는 패턴 분류 및 함수 근사(function approximation) 등의 문제에서 우수한 성능을 보이는 SVM(support vector machine)을 오디오 기반의 서베일런스 시스템에 적용하고자 하는 연구들이 흥미롭게 발견된다[3-5]. Rouas 등[3]은 철도 환경에서 'shout' 탐지를 위해서 SVM을 사용했으며, Wu 등[4]은 오디오 정보로부터 정상 소리와 'crying', 'groan', 'gun shooting'과 같은 비정상 소리를 이진 분류하는 목적으로 SVM을 사용하였다. 한편, 이의종 등[5]은 해당 클래스만을 독립적으로 표현하는 단일 클래스 SVM의 대표적 모델인 SVDD(support vector data description)를 계층적으로 구조화한 오디오 서베일런스 시스템을 제안하였다. 이는 실시간으로 비정상 소리 탐지가 가능하며 동시에 높은 인식율을 보장해야만 하는 감시 시스템의 요구 사항을 반영하여 설계된 시스템으로, 첫 번째 계층에서는 SVDD로 비정상 소리를 신속하게 탐지하여 관리자에게 알람 경고하고, 두 번째 계층의 다중클래스 SVDD는 탐지된 비정상 소리를 세분화 식별하여 관리자에게 비상 상황을 보고함으로써 관리자의 위기 상황 대처 능력을 돕는 추가 정보를 제공한다.

본 논문에서는 위에서 언급한 SVM 방법론을 계승, 발전시켜 보다 성숙한 모델을 제안하는 차원에서 출발하여, CCTV 등과 같은 감시 카메라 환경에서 오디오 정보를 이용하여 비정상 상황을 탐지 및 식별하는 시스템을 제안하고자 한다. 제안된 시스템은 단일 클래스 SVM의 대표적 모델인 SVDD와 최근 얼굴 인식 분야에서 성공적인 업적을 보여주고 있는 SRC(sparse representation classifier)를 기반으로 한 계층적 구조로써, 첫 번째 계층

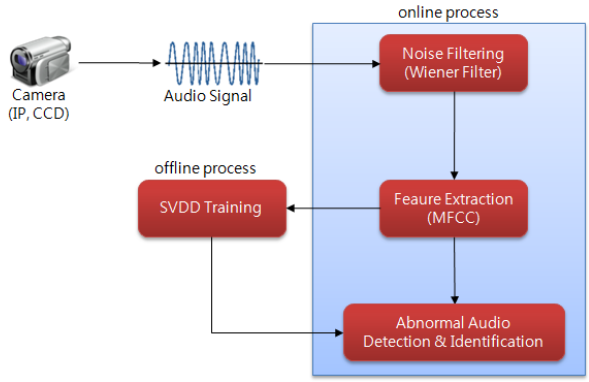
1) 본 연구는 교육과학기술부와 한국산업기술재단의 지역혁신 인력양성사업으로 수행된 연구결과임

에서는 SVDD로 비정상 소리를 신속하게 탐지하여 관리자에게 알람 경고하고, 두 번째 계층의 SRC에서는 탐지된 비정상 소리를 'gun', 'scream', 'siren' 소리 등으로 세분화 식별하여 관리자에게 비상 상황을 보고함으로써 관리자의 위기 상황 대처 능력을 돕는 추가 기능을 제공한다. 제안된 시스템은 실시간 처리를 위하여 1초 단위의 소리 신호 정보를 이용하였으며, 상황에 따라 새로운 비정상 소리 클래스의 추가가 요구되더라도 새로운 비정상 오디오 데이터의 특징만을 데이터베이스에 추가함으로써 시스템의 점증적 갱신(incremental updating) 및 확장을 보장한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 오디오 기반의 비정상 상황 인식 시스템에 대해 기술한다. 3장에서는 실험결과 및 성능 분석을, 마지막으로 4장에서는 결론 및 향후 연구과제에 대해 논한다.

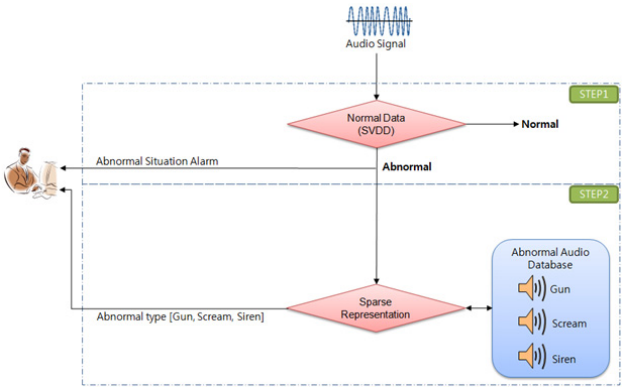
2. 비정상 소리 탐지 및 식별 시스템

본 논문에서 제안하는 비정상 소리 탐지 및 식별 시스템의 구조는 그림 1과 같다.



(그림1) 감시카메라 환경에서의 비정상 소리 탐지 및 식별 시스템의 전체 구성도

본 논문에서 제안하는 비정상 소리 탐지 및 식별 시스템은 총 4개의 모듈로 구성된다. 즉, 1개의 오프라인 처리 모듈인 SVDD 학습 모듈과 3개의 온라인 처리 모듈인 잡음 제거, 특징 추출, 비정상 소리 탐지 및 식별 모듈로 구성된다. 각 모듈의 기능은 다음과 같다. 1) SVDD 학습 모듈에서는 정상 소리 데이터 훈련 집합으로부터 오프라인 상에서 학습을 실시한다. 2) 감시카메라 환경에서의 비정상 소리 탐지 및 식별 문제는 짧은 소리 신호 구간 내에서 순간적으로 발생하는 비정상 혹은 응급 상황을 대변하는 소리를 탐지해야 한다. 따라서 잡음 제거 모듈에서는 짧은 구간에서 일정한 음폭 주파수 스펙트럼(frequency spectrum)의 표현에 장점을 갖고 있는 Wiener 필터 방법[6]을 사용한다. 3) 특징 추출 모듈에서는 인간이 인지할 수 있는 소리 영역 내에서 비정상 소리를 탐지할 수 있도록 MFCC(mel frequency cepstrum coefficient)[7]를 이용하여 특징 벡터를 추출한다. 4) 비정상 소리 탐지 및 식별 모듈에서는 SVDD 학습 모듈에서 학습이 완료된 SVDD와 SRC를 그림 2와 같이 계층적으로 구성하여 실시간으로 유입되는 소리 데이터의 비정상 탐지 및 식별을 수행한다. 첫 번째 계층의 SVDD는 탐지된 소리 데이터가 정상 오디오(normal audio database)에 등록되어 있는지를 빠르게 판단하여, 비정상 오디오로 판단되면 이를 관리자에게 알람 경고한다. 두 번째 계층의 SRC는 최근 얼굴 인식 분야에서 성공적인 업적을 보여주고 있는 일종의 분류 알고리즘으로써, 본 시스템에서는 비정상 소리 데이터베이스에 등록된 비정상 소리 데이터의 종류를 강건하게 식별하여 그 결과를 관리자에게 보고한다.



(그림 2) SVDD와 SRC를 이용한 계층적 비정상 소리 탐지 및 식별 시스템의 구성도

2.1 SVDD기반의 비정상 소리 탐지

오디오 기반 감시 시스템의 가장 중요한 목적은 감시 카메라 환경에서 발생한 소리가 비정상 소리인지 아닌지를 신속하게 판별하는 것이다. 비정상 소리 판별 문제는 비정상 소리 식별 문제와는 달리 대상 소리가 정상 소리 데이터베이스에 등록되었는지 아닌지를 판별하는 이진 클래스 분류(binary class classification) 문제로 볼 수 있다. 그러나 실제 비정상 소리 데이터와 정상 소리 데이터를 구분하기 위해서는 정상 소리 데이터만으로 기계학습을 수행한 후, 입력된 소리의 비정상 여부를 확인하는 테스트를 거치는 것이 실용적이다. 결국 비정상 소리 탐지 문제는 단순히 소리 데이터의 비정상 여부를 확인하는 과정이므로, 이진 클래스 분류 문제가 아닌 단일 클래스 분류(one class classification) 문제로 보는 것이 합리적이다. 따라서 본 논문에서는 단일 클래스 SVM의 가장 대표적인 방법론인 SVDD[14]를 이용하여 본 시스템의 첫 번째 계층에서 비정상 소리 여부를 신속하게 판단하고자 한다.

SVDD를 이용하여 비정상 소리를 인식하는 방법은 다음과 같다. 정상 소리 데이터만을 이용하여 특징 공간에서 정상 소리 데이터 군집만을 포함하는 원형체가 되도록 학습한다. 즉, 테스트 소리 데이터가 원형체 안에 포함되면

정상으로 인식되며 포함되지 않은 데이터는 정상이 아닌 비정상 소리로 판단된다.

2.2 SRC를 이용한 비정상 소리 식별

신호 처리 분야에서 효과적인 압축과 복원을 위해 연구되었던 SR(sparse representation)을 이용한 교사 학습 방법, 즉 SRC(sparse representation classifier)이 최근 얼굴 인식 분야에서 성공적인 업적을 보여주고 있음이 보고되고 있다[8-9]. 본 논문에서는 SRC를 비정상 소리 분류 문제에 적용하고자 한다.

SR 기반의 비정상 소리 분류 문제는 다음과 같이 수식으로 표현된다[10-11]. 일반적으로 n차원을 가지는 고차원 데이터 행렬 A는 데이터 포인트들의 집합으로 $A=a_1,\cdots a_n$으로 표현된다. A에 속하는 한 점 $a_i \in A$은 그 점과 이웃한 점들과의 선형 조합(linear combination)으로 표현된다. 임의의 클래스에 속하는 데이터 포인트들의 집합 $\{a_1,\cdots,a_n\}$이 주어졌다면, 같은 클래스에 속하는 새로운 데이터 포인트 a^*는 $\{a_1,\cdots,a_n\}$을 선형 조합으로 표현된다.

$$a^* = \beta_1 a_1 + \cdots + \beta_n a_n. \quad (1)$$

즉, n개의 학습 샘플 $\{a_1,\cdots,a_n\}$가 주어졌을 때, 선형 조합은 선형 부분 공간(linear subspace) W를 생성(span)하며, 새로운 데이터 포인트 a^*는 포인트가 속하는 클래스에 가장 근사한 부분 공간에 놓이게 된다.

$$W = span\{a_1,\cdots a_n\}. \quad (2)$$

I개의 클래스를 갖는 학습 샘플이 주어졌다면, 패턴 인식의 기본 방법론은 새로운 테스트 샘플을 학습 샘플 클래스를 이용하여 이에 상응하는 클래스로 정확하게 분류하는 것이다. i번째 클래스에 속하는 n_i 학습 샘플들은 행렬 $A=[v_{i,1},v_{i,2},\cdots,v_{i,n_i}] \in R^{m\times n_i}$의 열로 정렬이 되며, 이것은 각각의 학습 샘플 집합 $A=[A_1,A_2,\cdots A_k]$행렬로 표현된다. 선형 표현 가설(assumption)하에서, 테스트 샘플 $y=R^m$는 학습 샘플들에 의해 생성된 선형부분 공간에서 근사 된다. 이것은 다음과 같은 행렬식으로 표현된다.

$$y = Ax \in R^m, \quad (3)$$

위 식에서 x는 계수 벡터(coefficient vector)이다. 클래스 i에 속하는 샘플 y의 계수 벡터 x는 i와 관련된 학습 데이터 값을 제외하고는 0을 갖고, 다음과 같이 표현된다.

$$x = [0,\cdots,0,\beta_{i,1},\beta_{i,2},\cdots,\beta_{i,n_i},0,\cdots,0]^T \in R^n \quad (4)$$

식 (4)는 $y=Ax$ 선형시스템의 방정식을 풀어 x값을 얻을 수 있으며, 다음과 같은 x의 최적해(optimization solution)를 구하는 문제로 변경할 수 있다[10-11].

$$\hat{x_0} = \arg\min_x \|x\|_0 \ subject\ to\ y=Ax. \quad (5)$$

식 (5)를 이용하여 해를 찾는 것은 NP-hard 문제이다. 그러나 만약 x의 해가 충분히 sparse 하다면, l^0-노름(norm) 최소화 문제 (5)는 다음의 convex relaxed optimization 문제를 이용하여, 근사해인 l^1-노름 최소화 문제로 풀 수 있다[10-11].

$$\hat{x_1} = \arg\min_x \|x\|_1 \ subject\ to\ y=Ax. \quad (6)$$

3. 실험 및 결과 분석

본 논문에서 제안한 감시카메라 환경에서의 비정상 소리 탐지 및 식별 시스템을 평가하기 위하여, Wu 등[4]이 실험에 사용하였던 소리 데이터[12]를 샘플링하여 실험하였다. 표 1은 본 실험에서 사용한 정상 소리('people', 'background', 'nautical', 'recreation', 'quotes') 와 비정상 소리('gun', 'scream', 'siren') 데이터 집합을 보여준다. 프레임 단위는 정상 소리와 비정상 소리 모두 특징이 있는 부분을 중심으로 1초 단위로 추출하였으며, 정상 소리는 각 소리 종류별로 40개씩 총 200개의 데이터 집합을 수집하였다. 비정상 소리는 각 소리 종류별로 30개씩 총 90개 데이터 집합으로 구성하였다.

<표 1> 정상 소리와 비정상 소리 데이터 집합

정상소리	people, background, nautical, recreation, quotes
비정상 소리	gun, scream, siren

첫 번째 실험은 비정상 소리를 신속하게 탐지하는 실험으로 200개의 정상 소리 데이터 중 임의로 추출한 정상 소리 데이터 100개만으로 SVDD를 학습하였고, 테스트를 위한 데이터는 학습에 참여 하지 않은 정상 소리 데이터 100개와 비정상 소리 데이터 100개로 테스트 하였다. 이때, filter bank와 fft size 값은 각각 24와 512로 고정하였다. 실험 결과 비정상 소리 탐지 성능은 100%로 완벽하게 비정상 소리를 탐지하였다.

두 번째 실험은 SRC에서 상단 SVDD를 통해서 비정상 소리로 인식된 소리의 종류를 식별하는 실험이다. 비정상 소리 데이터 3개의 클래스에서 식별을 위해서 클래스당 20개의 소리 데이터를 학습 데이터로 사용하고, 학습에 사용되지 않은 나머지 10개의 데이터는 테스트 데이터로

사용하였다. 실험 결과 테스트 데이터로 사용하였던 30개의 데이터 모두 정확하게 식별하여 100%의 완벽한 식별률을 기록하였다. 계층적 SVDD를 이용한 방법론[5] 과 동일한 실험 조건에서 비교했을 때, 본 논문에서 제안한 시스템이 월등히 높은 식별률을 보임을 확인하였다(표 2 참조).

<표 2> 비정상 소리 식별 정확도

	계층적 SVDD [5]	Proposed Method
비정상 소리 식별률	94%	100%

4. 결론

본 논문에서는 CCTV 등과 같은 감시 카메라 환경에서 오디오 정보를 이용하여 비정상 상황을 인식하는 시스템의 프로토타입을 제안하였다. 제안된 시스템의 첫 번째 계층에서는 단일 클래스 SVM인 SVDD로 비정상 소리를 신속하게 탐지하여 관리자에게 알람 경고하고, 두 번째 계층의 SRC는 탐지된 비정상 소리를 'gun', 'scream', 'siren' 소리 등으로 세분화 식별하여 관리자에게 보고함으로써 관리자의 위기 상황 대처 능력을 돕는 추가 정보를 제공한다. 제안된 시스템은 실시간 처리를 위하여 1초 단위의 소리 신호 정보를 이용하였으며, 상황에 따라 새로운 비정상 소리 클래스의 추가가 요구되더라도 전체 시스템을 재학습시킬 필요 없이 새로운 비정상 소리 클래스만을 행렬 A의 열벡터로 추가함으로써 시스템의 점증적 갱신 및 확장이 가능하다. 실험을 통하여 제안된 시스템의 성능을 검증하였다.

향후 연구과제로는 본 연구에서 제안된 프로토타입의 비정상 소리 탐지 및 식별 시스템을 실세계에서 구현·운용하고자 한다.

참고문헌

[1] 전지혜, 박종화, 정철준, 강인구, 안태기, 박구만, "실시간 지능형 감시 시스템을 위한 방치, 제거된 객체 검출에 관한 연구", 한국통신학회논문지, vol. 35, no. 1, pp. 24-32, 2010.

[2] T. Zho, R. Nevatia, and B. Wu, "Segmentation and Tracking of Multiple Human in Crowded Environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 7, pp. 1198-1211, 2008.

[3] J. Rouas, J. Louradour, and S. Ambellouis, "Audio Events Detection in Public Transport Vehicle," 2006 IEEE Intelligent Transportation Systems Conference (ITSC '06), pp. 733-738, 2006.

[4] Wu. H. Gong, P. Che, Z. Zhong, and Y. Xu, "Surveillance Robot Utilizing Video and Audio Information," Journal of Intelligent Robot System, vol. 55, no. 4-5, pp. 403-421, 2009.

[5] 이의종, 주영민, 김정식, 오승근, 유재학, 박대희, "오디오 서베일런스 시스템에서의 비정상 상황인식", 2010한국정보과학회 추계학술대회, 2010. (게재예정)

[6] S. Dolcol, D. Rong, T. Klasen, J. Wouters, S. Haykin, and M. Moonen, "Extension of the Multi Channel Winer Filter with ITD Cues for Noise Reduction in Binaural Hearing Aids," Application of Signal Processing to Audio and Acoustics, vol. 16, no. 16, pp. 70-73, 2005.

[7] K. Murtry, B. Yegnanarayana, "Combining Evidence From Residual Phase and MFCC Features for Speaker Recognition," IEEE Signal Processing Letters, vol. 13, no. 1, pp. 52-55, 2006.

[8] L. Qiao, S. Chen, and X. Tan, "Sparsity Preseving Projections with Applications to Face Recognition," Journal of Pattern Recognition, vol. 43, no. 1, pp. 331-341, 2010.

[9] P. Yannis, K. Constantine, and A. Gonzolo R., "Music Genre Classification Via Sparse Representations of Auditory Temporal Modulations," 17th European Signal Processing Conference (EUSIPCO 2009), pp. 1-5, 2009.

[10] Y. Ji, T. Lin, and X. Tan, "Mahalanobis Distance Based Non-negative Sparse Representation for Face Recognition," International Conference on Machine Learning and Applications, pp. 41-46, 2009.

[11] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and M. Yi, "Robust Face Recognition via Sparse Representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210-227, 2009.

[12] http://www.grsites.com