Journal of the Korea Academia-Industrial cooperation Society
Vol. 20, No. 3 pp. 562-569, 2019

https://doi.org/10.5762/KAIS.2019.20.3.562
ISSN 1975-4701 / eISSN 2288-4688

계층 클러스터 트리 기반 라만 스펙트럼 식별 고속 검색 알고리즘

김순금[1,2], 고대영[1,2], 박준규[1], 박아론[1], 백성준[1*]
[1]전남대학교 전자컴퓨터공학부, [2]씨에스에너지(주)

A Hierarchical Cluster Tree Based Fast Searching Algorithm for Raman Spectroscopic Identification

Sun-Keum Kim[1,2], Dae-Young Ko[1,2], Jun-Kyu Park[1], Aa-Ron Park[1], Sung-June Baek[1*]
[1]School of Electronics and Computer Engineering, Chonnam National University
[2]CS Energy Corporation

요 약 최근에 원 거리에서 폭발 물질의 감지를 위해 라만 분광 기기의 관심이 점차 증가하고 있다. 더불어 측정된 화학물질에 대한 라만 스펙트럼을 대용량 데이터베이스의 알려진 라만 스펙트라와 비교하여 식별할 수 있는 고속 검색 방법에 대한 요구도 커지고 있다. 지금까지 가장 간단하고 널리 사용되는 방법은 주어진 스펙트럼과 데이터베이스 스펙트라 사이의 유클리드 거리를 계산하고 비교하는 방법이다. 하지만 고차원 데이터의 속성으로 검색의 문제는 그리 간단하지 않다. 가장 큰 문제점중의 하나는 검색 방법에 있어서 연산량이 많아 계산 시간이 너무 오래 걸린다는 것이다. 이러한 문제점을 극복하기 위해, 우리는 정렬된 분산에 따른 MPS Sort+PDS 방법을 제안하였다. 이 방법은 벡터의 두 개의 주요한 특징으로 평균과 분산을 사용하여 후보가 될 수 없는 많은 코드워드를 계산하지 않으므로 연산량을 줄이고 계산 시간을 줄여준다. 본 논문에서 우리는 기존의 방법보다 더욱 더 향상된 2가지 새로운 방법의 고속 검색 알고리즘을 제안한다. PCA+PDS 방법은 전체 데이터를 사용하는 거리 계산과 똑같은 결과를 가지면서 PCA 변환을 통해 데이터의 차수를 감소시켜 계산량을 줄여준다. Hierarchical Cluster Tree 알고리즘은 PCA 변환된 스펙트라 데이터를 사용하여 이진 계층 클러스터 트리를 만든다. 그런 후 입력 스펙트럼과 가장 가까운 클러스터부터 검색을 시작하여 후보가 될 수 없는 많은 스펙트라를 계산하지 않으므로 연산량을 줄이고 계산 시간을 줄여준다. 실험은 정렬된 분산에 따른 MPS Sort+PDS와 비교하여 PCA+PDS는 60.06%의 성능 향상을 보였다. Hierarchical Cluster Tree는 PCA+PDS와 비교하여 17.74%의 성능향상을 보였다. 실험결과는 제안된 알고리즘이 고속 검색에 적합함을 확인시켜 준다.

Abstract Raman spectroscopy has been receiving increased attention as a standoff explosive detection technique. In addition, there is a growing need for a fast search method that can identify raman spectrum for measured chemical substances compared to known raman spectra in large database. By far the most simple and widely used method is to calculate and compare the Euclidean distance between the given spectrum and the spectra in a database. But it is non-trivial problem because of the inherent high dimensionality of the data. One of the most serious problems is the high computational complexity of searching for the closet spectra. To overcome this problem, we presented the MPS Sort with Sorted Variance+PDS method for the fast algorithm to search for the closet spectra in the last paper. the proposed algorithm uses two significant features of a vector, mean values and variance, to reject many unlikely spectra and save a great deal of computation time. In this paper, we present two new methods for the fast algorithm to search for the closet spectra. the PCA+PDS algorithm reduces the amount of computation by reducing the dimension of the data through PCA transformation with the same result as the distance calculation using the whole data. the Hierarchical Cluster Tree algorithm makes a binary hierarchical tree using PCA transformed spectra data. then it start searching from the clusters closest to the input spectrum and do not calculate many spectra that can not be candidates, which save a great deal of computation time. As the Experiment results, PCA+PDS shows about 60.06% performance improvement for the MPS Sort with Sorted Variance+PDS. also, Hierarchical Tree shows about 17.74% performance improvement for the PCA+PDS. The results obtained confirm the effectiveness of the proposed algorithm.

Keywords : Raman Spectrum, Fast Search Algorithm, PCA(Principal Component Analysis), PDS(Partial Distortion Search), Hierarchical Cluster Tree

이 논문은 전남대학교 학술연구비(과제번호:2016-2517) 지원에 의하여 연구되었음.

*Corresponding Author : Sung-June Baek(Chonnam National Univ.)
Tel: +82-62-530-1795 email: tozero@chonnam.ac.kr

Received December 26, 2018 Revised January 30, 2019
Accepted March 8, 2019 Published March 31, 2019

1. 서론

라만 분광법(Raman Spectrocopy)은 레이저를 이용하여 물질 내 분자들의 진동 에너지 레벨을 측정하기 위한 강력한 비접촉 기술이다[1]. 라만 스펙트럼(Raman Spectrum)에 의해 제공되는 진동 값은 분자의 화학 물질 구성에 대한 특징적인 정보를 갖는다. 아래의 Fig. 1은 레퍼런스 스펙트라의 DNT(2,6-Denitrotoluene)와 HMX(1,3,5,7-Tetranitro-1,3,5,7-tetrazocane) 화학물질의 라만 스펙트럼을 나타낸다. Fig. 1에서 보이는 것처럼 라만 스펙트럼은 각 화학 물질에 따라 각기 다른 진동 주파수 성분을 갖는 피크들로 구성된다.

특히 최근에 원 거리에서 폭발 물질의 감지를 위해 라만 분광 기기의 관심이 점차 증가하고 있다[2-6]. 더불어 측정된 화약물질에 대한 라만 스펙트럼을 대용량 데이터베이스의 알려진 라만 스펙트라(Spectra)와 비교하여 식별할 수 있는 고속 검색 방법에 대한 요구도 커지고 있다.

지금까지 가장 간단하고 널리 사용되는 방법은 주어진 스펙트럼과 데이터베이스 스펙트라 사이의 유클리드 거리(Euclidean distance)를 계산하고 비교하는 전체 검색(Full Search) 방법이다. 이러한 방법은 데이터베이스 내의 모든 스펙트라와 주어진 스펙트럼과의 유클리드 거리를 계산한 후 최소거리 값을 갖는 하나의 스펙트럼을 찾는다. 따라서 전체 검색 방법은 비교될 스펙트라의 개수가 증가하고 데이터베이스의 크기가 대용량화됨에 따라 스펙트럼 식별에 많은 계산 시간이 필요하다.

이러한 전체 검색의 문제점을 극복하기 위해 벡터 양자화(Vector Quantization) 분야에서 PDS(Partial Distortion Search or Elimination), MPS(Mean Pyramids Structure or Search) 알고리즘들이 널리 사용되고 있다[7-12]. 우리는 지난 논문에서 MPS방법과 MPS에 PDS를 결합한 MPS+PDS(Partial Distortion Search or Elimination) 방법이 1차원 라만 스펙트럼 데이터의 고속 검색에 적합함을 확인하였다. 또한 기존의 검색 방법보다 성능이 개선된 방법으로 정렬된 분산에 따른 MPS Sort(이하, MPS Sort)+PDS 방법을 제안하였다[13]. 이러한 방법은 분산 신호의 피크가 신호를 구별하는데 주요한 특징으로 작용하므로 라만 스펙트라 전체에 대한 분산 값을 사용하는 검색방법이다. 레퍼런스 스펙트라 전체 분산 신호의 최대 피크 값에서 최소 피크 값으로 내림차순 정렬한 후 PDS를 결합한 MPS Sort+PDS 방법이 MPS+PDS와 비교하여 55.2%의 가장 좋은 성능 향상 결과를 보였다.

본 논문에서는 MPS Sort+PDS 방법보다 더욱 더 향상된 라만 스펙트럼 데이터의 고속 검색에 적합한 새로운 방법을 제안하고자 한다. 이를 위해 우리는 라만 스펙트라의 PCA(Principal Component Analysis) 선형변환을 통해 데이터 차수를 줄이고 이를 바탕으로 이진 계층 클러스터 트리(Hierarchical Cluster Tree)를 만든 후 계층 클러스터 트리를 검색하는 방법을 제안하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 라만 스펙트라의 PCA 선형 변환과 PCA에 PDS를 결합한 PCA+PDS 방법에 대해 살펴본다. 3절에서는 계층 클러스터링과 계층 클러스터 트리 방법을 사용한 고속 검색 알고리즘을 제안하고 설명한다. 4절에서는 실험결과를 통해 제안된 방법이 라만 스펙트럼의 고속 검색에 적합함을 확인한다. 마지막으로 5절에서 간단한 결론으로 본 논문을 마무리 한다.

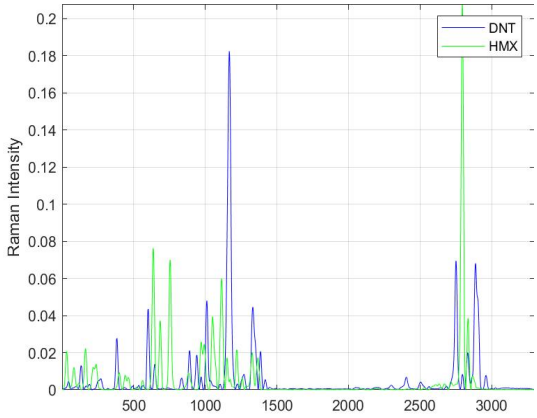

Fig. 1. DNT and HMX of Reference Spectra

2. 기존 검색 알고리즘

2.1 PCA(Principlal Component Analysis)

전체 검색을 기반으로 하는 기존의 검색 방법들은 거리 계산에 N(=3300)개의 데이터를 갖는 스펙트럼 간의 거리를 계산한다. 거리 계산에 전제 데이터를 사용하는 것은 계산량 증가로 인해 고속 검색에 적합하지 않다. 우리는 라만 스펙트럼 데이터의 차수를 감소시켜 계산량을 줄이기 위해 PCA 변환을 수행한다.

아래의 Fig. 2는 PCA 선형 변환의 전체 과정을 나타낸다. 먼저 $N\times M$ (=3300×14085) 레퍼런스 스펙트라 행렬 Y에 대해 다음의 식 (1)과 같이 $N\times N$ 상관 관계 행렬(Correlation Matrix) R을 구성한다.

$$R = YY^T \tag{1}$$

다음으로 상관 관계 행렬 R의 고유값(EigenValue)과 고유벡터(EigenVector)를 구하여 큰 값에서부터 작은 값의 순서대로 내림차순 정렬한다. 그 후에 아래의 식 (2)와 같이 내림차순 된 고유값 행렬 Λ와 고유벡터 행렬 V의 곱으로 상관 관계 행렬 R을 재구성한다.

$$R = V\Lambda V^T \tag{2}$$

위 식(2)에서 행렬 Λ는 대각 성분이 $(\lambda_1,\lambda_2,...,\lambda_N)$이고 $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$을 만족하는 $N\times N$ 고유값 행렬을 나타낸다. 그리고 행렬 V는 고유벡터 계수 v_{ij} $(i \text{ and } j = 1,...,N)$를 갖는 $N\times N$ 고유벡터 행렬이다.

N개의 데이터 포인트를 갖는 각 스펙트럼은 N차원 공간의 점으로 생각할 수 있다. 다음 단계에서 원시 데이터를 새로운 N개의 좌표축을 갖는 공간의 점으로 바꾸기 위해 아래의 식(3)을 이용하여 각 스펙트럼에 대한 PCA 선형 변환을 수행한다.

$$W = V^T Y \tag{3}$$

위 식(3)에서 행렬 Y는 원래 좌표 $(y_1,y_2,...,y_N)$, 행렬 W는 새로운 좌표 $(w_1,w_2,...,w_N)$로 이루어진다. 새로운 좌표 w_1는 첫 번째 주성분(PC) 축 방향으로 모든 스펙트럼의 가장 큰 분산 값을 갖도록 선택한다. 두 번째 주성분은 첫 번째 축에 직교(Orthogonal)하는 축을 따라 가장 큰 분산 값을 갖는다. 따라서 다음의 주성분에 의해 표현되는 분산은 계속해서 감소된다. 선형변환에 의해 만들어진 주성분들(PCs)은 서로 비상관 관계에 있고 서로 직교한다. 또한 차수가 증가함에 따라 주성분에 포함된 데이터의 정보량은 감소한다.

다음으로 새로운 좌표 $(w_1,w_2,...,w_N)$에서 정보량의 집중도를 알아보고 데이터 차수를 줄이기 위해 아래의 식(4)를 이용한다.

$$Energy(K) = \frac{\sum_{i=1}^{K}\lambda_i}{\sum_{i=1}^{N}\lambda_i} \tag{4}$$

위 식(4)에서 $Energy(K)$는 모든 고유값 성분 $(\lambda_1,\lambda_2,...,\lambda_N)$의 합에서 처음부터 K개의 고유값 $(\lambda_1,\lambda_2,...,\lambda_K)$ 성분의 합이 차지하는 비율로 표시된다. 선형변환의 마지막 단계에서 데이터 차수를 감소 시켜 계산량을 줄이기 위해 N(=3300)개의 주성분으로부터 $Energy(K)$의 값이 거의 1에 근사하는 K를 찾아 K (=250)개의 주성분을 선택한다.

PCA 변환 후 K가 250인 라만 스펙트럼 사이의 거리 계산은 PCA 변환 전 N이 3300인 거리 계산과 똑같은 결과를 가지면서 계산량은 더 줄기 때문에 빠른 검색이 가능하다.

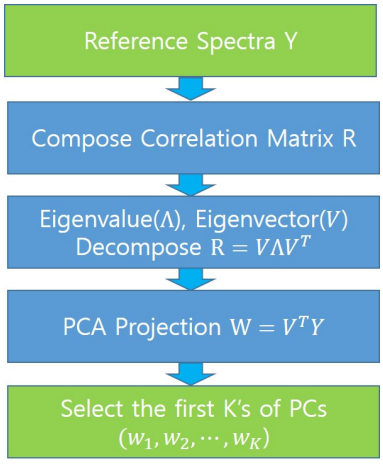

Fig. 2. Process of PCA Linear Transformation

2.2 PCA+PDS(Partial Distortion Search)

PDS 알고리즘은 입력 벡터 x와 레퍼런스 벡터 y_i 사이의 거리의 제곱 $d^2(x,y_i)$ 계산에 있어서 종료 조건을 두어 조기에 종료 시키는 방법이다. 이 알고리즘은 먼저 입력 벡터 $x=(x_1,x_2,...,x_N)$와 임의의 레퍼런스 벡터 y_i 사이의 거리의 제곱 $d^2(x,y_j)$값을 계산하여 $d^2_{\min}$으로 놓는다. 그 다음 레퍼런스 벡터 $y_j=(y_{j1},y_{j2},...,y_{jN})$에 대해, 처음부터 k까지 거리의 제곱 값의 누적 합이 $d^2_{\min}$보다 큰 다음 식(5)의 조건을 만족하는 $k<N$가 존재하면, 레퍼런스 벡터 y_j의 거리의 제곱 계산을 멈춘다.

$$\sum_{n=1}^{k}(x_n - y_{jn})^2 \geq d^2_{\min} \tag{5}$$

이 방법은 $(N-k)$ 번의 곱셈과 $2(N-k)$ 번의 덧셈 계산을 줄여준다.

PDS 방법은 유클리드 거리의 제곱의 누적 합이 가장 가까운 레퍼런스 후보의 거리보다 크면 거리 계산을 종

료하여 계산량을 줄이는 방법이다. 그러므로 PCA에 PDS를 결합하는 방법을 생각할 수 있다. PCA+PDS 방법은 입력 벡터와 임의의 레퍼런스에 대한 유클리드 거리의 제곱 부분 누적 합이 $d^2_{\min}$ 값보다 크면 거리 계산이 종료되므로 PCA 방법보다 계산량을 더 줄임으로써 더욱 빠른 검색이 가능하다.

하지만 K(=250)개의 PCA 변환 데이터를 갖는 스펙트럼 간의 거리 계산도 여전히 오버헤드가 크므로 계산량을 더욱 더 줄이기 위해 계층 클러스터 트리 검색에 기반 한 고속 검색 방법을 제안한다.

3. 제안된 고속 검색 알고리즘

3.1 계층 클러스터링(Hierarchical Clustering)

계층 클러스터링 방법은 PCA 변환을 수행한 후 얻은 $K\times M$ (=250×14085)개의 PCA 변환 데이터에 대해 이진 클러스터 트리(Cluster Tree)를 만들어 각각의 변환 데이터를 N(=70)개의 클러스터에 할당한다. Fig. 3은 계층 클러스터링 전체 과정을 나타낸다. 계층 클러스터 분석을 수행하기 위해 먼저 개체 사이의 유사성을 측정 한다. 유사성 측정(Similarity Measures)은 라만 스펙트라 데이터 집합의 모든 개체 쌍 사이의 거리를 계산한다. m개의 개체로 구성된 데이터 세트의 경우, 데이터 세트에 m * (m-1)/2 쌍이 있다.

데이터 집합 개체 사이의 근접 거리가 계산되면 다음으로 Matlab의 linkage 함수를 사용하여, 개체를 클러스터로 그룹화 할 수 있다. linkage 함수는 생성된 거리 정보를 가져와서 서로 가까이 있는 개체 쌍을 두 개체로 구성된 이진 클러스터로 링크한다. 그런 다음 원본 데이터 집합의 모든 개체가 계층 트리(Hierarchical Tree)에 함께 링크 될 때까지 새로 형성된 클러스터를 다른 개체에 연결하여 더 큰 클러스터를 작성한다. 이진 클러스터 계층 트리를 작성한 후에는 matlab의 cluster 함수를 사용하여 데이터를 N개의 클러스터로 나누어 할당한다.

마지막으로 각 클러스터의 중심이 되는 개체를 찾아 클러스터의 중심과 클러스터 내의 모든 개체 사이의 거리를 구한다. 그런 후에 각 클러스터의 중심에서 가장 가까운 거리에 있는 개체와 가장 먼 거리에 있는 개체에 대한 인덱스와 거리 값 정보를 저장한다. 여기서 각 클러스터의 중심은 모든 개체의 평균에서 가장 가까운 위치에 있는 실제 개체가 존재하는 위치가 각 클러스터의 실제 중심 위치가 된다.

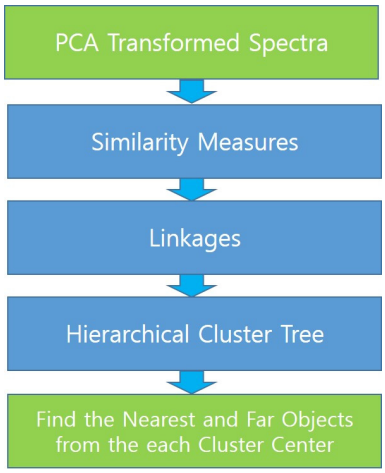

Fig. 3. Process of the Hierarchical Clustering

3.2 계층 클러스터 트리 검색

본 논문에서 제안하는 계층 클러스터 트리(Hierarchical Cluster Tree) 검색 알고리즘은 Fig. 4와 같다. 제안된 알고리즘은 먼저 입력 라만 스펙트럭 x와 레퍼런스 스펙트라 중 가장 가까운 후보 y_j를 찾기 위해 입력 스펙트럼으로부터 N(=70)개의 클러스터 중심과의 거리를 계산한다. 입력 스펙트럼으로부터 가장 가까운 클러스터 중심까지의 거리의 제곱 값을 $d^2_{\min}$ 으로 놓는다. 가장 가까운 거리에 있는 클러스터 내부에 입력 스펙트럼과 가장 유사한 후보가 있을 가능성이 높으므로 가장 가까운 클러스터에서부터 검색을 시작한다. 또한 입력 스펙트럼과 클러스터 중심의 거리 d_{center}와 클러스터 중심에서 가장 먼 거리에 있는 개체의 거리 $d_{\max}$의 차가 $d^2_{\min}$ 보다 크다면 다음 식(6)의 삼각 부등식에 의하여 후보에서 제외한다.

$$d_{center} - d_{\max} > d^2_{\min} \tag{6}$$

다음으로 입력벡터와 임의의 스펙트라의 거리 제곱 누적 합이 $d^2_{\min}$ 값 보다 크면 거리 계산을 종료한다. 마지막으로 $d^2(x,y_j) < d^2_{\min}$ 이면 $d^2(x,y_j)$을 $d^2_{\min}$으로 놓고 N개의 클러스터에 대한 반복루프를 마쳐 최종적으로 입력벡터와 가장 가까운 스펙트라를 찾는다.

```
while all n∈N Clusters
    calculate the each d_center from N clusters
    sort the N clusters from near to far
    find the cluster with minimum distance
end
let the minimum distance as d²_min

while all n∈N Clusters
    from the nearest cluster
    if d_center − d_max > d²_min
      continue;
    end
    for i=2 : Number of each cluster elements
      for j=1:Number of PCA
          d²(x,y_j) > d²_min
           break;
      end
      if d²(x,y_j) < d²_min
        d²_min = d²(x,y_j)
        closest spectra Index=i
      end
    end
end
```

Fig. 4. Hierarchical Tree based Search Algorithm

본 논문에서 제안하는 계층 클러스터 트리 기반 검색 알고리즘은 위 (식) 6의 삼각 부등식에 의해 $d^2_{\min}$보다 큰 값을 갖는 클러스터의 많은 개체 들이 거리 계산 없이 후보에서 제외되므로 계산량을 줄여 빠른 검색이 가능하다. 또한 입력벡터와 임의의 스펙트라의 거리 제곱 누적 합이 $d^2_{\min}$값 보다 크면 거리 계산이 종료되므로 더욱 빠른 검색이 가능하다.

4. 실험결과

4.1 실험방법

레퍼런스 데이터는 화학물질 총 14,085개의 라만 스펙트라 데이터로 구성된다. 각 스페트라 데이터는 크기가 3,300인 1차원 벡터이다. 알고리즘의 성능 비교 실험을 위해 레퍼런스 데이터로부터 균일한 간격으로 샘플링하고 약 25 dB(SNR)의 잡음을 추가하여 2,817개의 라만 스펙트럼 테스트 데이터를 생성하였다.

알고리즘에 대한 성능 평가 기준으로 실행시간을 사용하면 프로세서의 속도에 따라 정확한 비교가 불가능하다. 따라서 본 논문에서는 알고리즘의 성능 평가 기준으로 곱셈연산, 덧셈연산 및 총 연산 횟수를 사용하여 객관적인 성능 평가가 이루어지도록 한다. 각 연산에 대한 계산량은 테스트 데이터 2,817개에 대한 평균 연산 횟수를 사용한다.

각 알고리즘의 성능평가를 위해 지난 논문에서 제안했던 Full Search+PDS, MPS+PDS, MPS Sort+PDS 검색 알고리즘과 비교를 위해 위에 기술한 실험방법에 의해 다시 실험을 수행하였다[13]. 또한 본 논문에서 제안하는 PCA+PDS 알고리즘에 대한 성능 평가 실험과 더불어 3절에서 제안하는 Hierarchical Cluster Tree 알고리즘에 대한 성능 평가를 수행하였다.

4.2 실험결과

각 알고리즘들의 성능 실험결과는 곱셈연산과 덧셈연산을 더한 연산 횟수로 나타낸다. Fig. 5와 Table 1은 Full Search+PDS, MPS+PDS, MPS Sort+PDS 기존 검색 알고리즘들과 본 논문에서 제안한 PCA, PCA+PDS, Hierarchical Tree 알고리즘들의 성능 실험 결과를 총 연산 횟수로 나타낸다. 실험결과는 평균 총 연산 횟수 중 최적의 성능을 나타내는 연산 횟수를 표시하였다.

Fig. 5와 Table 1에서 보이는 실험결과와 같이 Full Search+PDS는 Full Search와 비교하여 80.97%의 성능 향상 결과를 보였다. 이를 통해 제안하는 알고리즘에 PDS를 결합하는 방법이 유효함을 확인하였다. 또한 Full Search+PDS와 비교하여 MPS+PDS는 53.9%의 성능이 향상 되었다. 그리고 MPS+PDS와 비교하여 MPS Sort+PDS는 56.03%의 성능향상 결과를 보였다. 지난 논문에서 살펴본 바와 같이 MPS Sort+PDS 알고리즘이 가장 좋은 성능을 나타내었다. 이 실험을 통해 레퍼런스 스펙트라 전체에 대한 평균 또는 분산 값을 사용하여 분산 신호의 최대 피크 값에서 최소 피크 값으로 내림차순 정렬한 후 PDS를 결합한 방법이 유효한 방법임을 알 수 있다.

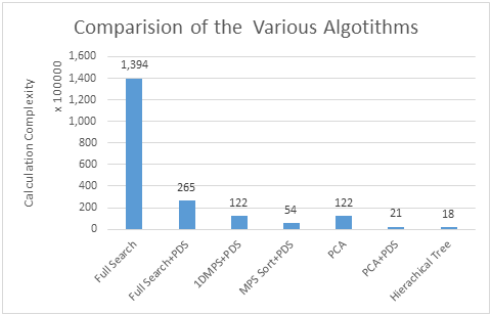

Fig. 5. Comparison of the previous Algorithms and the Proposed Fast Searching Algorithms

Fig. 5와 Table 1에서 보이는 실험결과와 같이 PCA+PDS는 PCA와 비교하여 82.4%의 성능향상 결과를 보였다. 또한 MPS Sort+PDS과 비교하여 60.06%의 성능향상 결과를 보였다. 이를 통해 본 논문 2절에서 제안하는 N이 3300인 라만 스펙트럼에 PCA 변환을 수행하여 K가 250인 라만 스펙트럼으로 데이터의 차수를 감소시킨 후 PDS를 결합한 PCA+PDS 방법이 계산량을 줄이고 고속 검색에 적합함을 확인할 수 있다. 본 논문의 3절에서 제안하는 Hierarchical Cluster Tree 알고리즘은 MPS Sort+PDS과 비교하여 67.15%의 성능향상 결과를 보였다. 또한 최고의 성능을 보인 PCA+PDS 방법과 비교하여 17.74%의 성능 향상 결과를 보였으며 가장 우수한 성능 결과를 보였다. 실험 결과 본 논문에서 제안된 Hierarchical Cluster Tree 기반 검색 방법은 입력 벡터와 가까운 후보가 될 수 없는 더 많은 라만 스펙트라를 제외하여 많은 양의 계산을 줄이므로 고속 검색에 가장 우수한 성능을 보여주고 있다.

Table 1. Multiplication, Addition, Total Calculation Complexity with the best Performance for the Previous and the Proposed Fast Searching Algorithms

Method	Multiplication	Addition	Total
Full Search	46,480,500	92,946,915	139,427,415
Full Search+PDS	8,846,894	17,679,704	26,526,598
1DMPS+PDS	1,210,602	11,012,972	12,223,574
MPS Sort+PDS	853,773	4,520,642	5,374,415
PCA	4,346,250	7,853,165	12,199,415
PCA+PDS	995,283	1,151,231	2,146,514
Hierachical Tree	864,074	901,576	1,765,650

5. 결론

Full Search 방법은 주어진 스펙트럼과 데이터베이스 라만 스펙트라 사이의 유클리드 거리를 계산하고 비교하는 방법이다. 이 방법은 레퍼런스 라만 스펙트라 전체를 검색하므로 계산량이 너무 많다는 단점이 있다. MPS+PDS 방법은 입력 라만 스펙트럼과 임의의 라만 스펙트라의 평균값을 유클리드 거리 계산에 사용하고 유클리드 거리의 제곱의 누적 합이 가장 가까운 라만 스펙트라 후보의 거리보다 크면 거리 계산을 종료하여 계산량을 줄이는 방법이다.

우리는 지난 논문에서 최고의 성능 결과를 보인 MPS Sort+PDS 고속 검색 방법을 제안하였다. 이 방법은 벡터의 두 개의 주요한 특징으로 평균과 분산을 사용하고 신호 피크의 최대에서 최소값 순서로 내림차순 정렬한 후 PDS 방법을 적용하였다.

본 논문에서는 입력 라만 스펙트럼을 데이터베이스의 알려진 라만 스펙트라와 비교하여 식별할 수 있는 새로운 고속 검색 방법으로 PCA+PDS, Hierachical Cluster Tree 알고리즘을 제안하였다. PCA+PDS 방법은 먼저 라만 스펙트럼 데이터에 PCA 변환을 수행하여 데이터의 차원을 감소 시켜 유클리드 거리 계산을 수행한다. Hierarchical Cluster Tree 알고리즘은 PCA 변환 된 라만 스펙트라의 클러스터 트리를 만들고 입력 스펙트럼과 가장 가까운 클러스터에서부터 검색을 시작하여 후보가 될 수 없는 많은 스펙트라를 계산에서 제외하여 연산량을 줄이고 계산 시간을 줄여준다. 실험결과 기존의 검색 방법보다 현저하게 성능이 개선되었고 제안된 Hierarchical Cluster Tree 알고리즘이 고속 검색에 적합함을 확인하였다.

References

[1] D. J. Gardiner, “Practical Raman Spectroscopy”, New York:Springer-Verlag 1989.
DOI: http://dx.doi.org/10.1007/978-3-642-74040-4

[2] R. L. McCreery, “Raman Spectroscopy for Chemical Analysis”, NJ Hoboken:Wiley 2000.
DOI: http://dx.doi.org/10.1002/0471721646

[3] M. Gaft L. Nagli "Standoff laser-based spectroscopy for explosives detection" Proc. SPIE Conf. Ser. vol. 6739 pp. 673903 2007-Oct.
DOI: http://dx.doi.org/10.1117/12.736631

[4] S. Wolf P. J. Wrzesinski M. Dantus "Standoff chemical detection using single-beam CARS" Proc. Conf. Lasers Electro-Opt./Int. Quantum Electron. Conf. pp. 1-2 2009. DOI: http://dx.doi.org/10.1364/CLEO.2009.CFU1

[5] S. Wallin A. Pettersson H. Östmark A. Hobro "Laser-based standoff detection of explosives: A critical review" Anal. Bioanal. Chem. vol. 395 no. 2 pp. 259-274 Sep. 2009. DOI: http://dx.doi.org/10.1007/s00216-009-2844-3

[6] N.R. Butt, M. Nilsson A, Jakobsson M. Nordberg A. Pettersson S. Wallin and H. Ostmark "Classification of raman spectra to detect hidden explosives" Geoscience and Remote Sensing Letters IEEE vol. 8 no. 3 pp. 517-521 may 2011. DOI: http://dx.doi.org/10.1109/LGRS.2010.2089970

[7] BEI, C.D, GRAY, R.M, "An improvement of the minimum distortion encoding algorithm for vector quantisation", IEEE Trans., pp. 1132-1133, 1985. DOI: http://dx.doi.org/10.1109/TCOM.1985.1096214

[8] HSIEH, C.H, LU,P,C, CHANG,J,C, "Fast codebook generation algorithms for vector quantisation of images", Pattern Recognition Lett, pp. 605-609, 1991. DOI: http://dx.doi.org/10.1016/0167-8655(91)90014-D

[9] ORCHARD, M. D, "A fast nearest-neighbour search algorithm", IEEE ICASSP, pp. 2297-2300, 1991. DOI: http://dx.doi.org/10.1109/ICASSP.1991.150755

[10] GUAN, L, KAMEL, M, "Equal-average hyperplane partitioning method for vector quantisation of image data", Pattern Recognition Lett, pp. 693-699, 1992. DOI: http://dx.doi.org/10.1016/0167-8655(92)90098-K

[11] Lee, C.H, Chen, L,H, "Fast closet codeword search algorithm for vector quantization", IEEE Porc., pp. 143-148, 1994. DOI: http://dx.doi.org/10.1049/ip-vis:19941140

[12] Lee, C.H, Chen, L,H, "A Fast Search Algorithm for Vector Quantization Using Mean Pyramids of Codewords", IEEE Trans., pp. 1697-1702, 1995. DOI: http://dx.doi.org/10.1109/26.380218

[13] Ko, Dae-Young, Baek, Sung-June, Park, Jun-Kyu, Seo, Yu-Gyeong, Seo, Sung-Il, "The Fast Search Algorithm for Raman Spectrum", Journal of the Korea Academia-Industrial cooperation Society, Vol. 16, pp. 3378-3384, 2015. DOI: http://dx.doi.org/10.5762/KAIS.2015.16.5.3378

김 순 금(Sun-Geum Kim) **[정회원]**

- 2012년 2월 : 동신대학교 수소에너지학과 (공학학사)
- 2018년 8월 : 전남대학교 산업대학원 전기전자컴퓨터공학과 전자공학(공학석사)
- 2018년 9월 ~ 현재 : 전남대학교 일반대학원 전자컴퓨터공학과(박사과정)
- 2012년 12월 ~ 현재 : 씨에스에너지(주) 대표이사

<관심분야>
디지털 신호처리,패턴인식,태양광시스템,ESS저장장치시스템

고 대 영(Dae-Young Ko) **[준회원]**

- 1999년 2월 : 전남대학교 전자공학과 (공학학사)
- 2002년 2월 : 전남대학교 전자공학과 (공학석사)
- 2016년 8월 : 전남대학교 전자공학과 (공학박사)
- 2018년 4월 ~ 현재 : 씨에스에너지(주) 연구소장

<관심분야>
디지털 신호처리, 임베디드 시스템

박 준 규(Jun-Kyu Park) **[정회원]**

- 2009년 2월 : 전남대학교 전자컴퓨터공학부(공학사)
- 2017년 2월 : 전남대학교 전자공학과(공학박사)
- 2018년 10월 ~ 현재 : 한국생산기술연구원 Postdoc

<관심분야>
디지털 신호처리, 패턴인식, 머신러닝

박 아 론(Aa-Ron Park) **[정회원]**

- 2006년 2월 : 전남대학교 전자컴퓨터정보통신공학부 (공학학사)
- 2008년 2월 : 전남대학교 전자공학과 (공학석사)
- 2012년 2월 : 전남대학교 전자공학과 (공학박사)

<관심분야>
디지털 신호처리, 패턴 인식, 바이오 응용 패턴 인식, 특징 추출/선택

백 성 준(Sung-June Baek) **[정회원]**

- 1989년 2월 : 서울대학교 전자공학과 (공학학사)
- 1992년 2월 : 서울대학교 전자공학과 (공학석사)
- 1999년 2월 : 서울대학교 전자공학과 (공학박사)
- 2002년 3월 ~ 현재 : 전남대학교 전자컴퓨터공학부 교수

<관심분야>
의료 통신 음성 관련 디지털 신호처리