JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 24 No. 10, pp. 91-99, October 2019
https://doi.org/10.9708/jksci.2019.24.10.091

Index-based Boundary Matching Supporting Partial Denoising for Large Image Databases

Bum-Soo Kim*

*Researcher, Dept. of Future Technology and Convergence Research,
Korea Institute of Civil Engineering and Building Technology, Gyeonggi, Korea

[Abstract]

In this paper, we propose partial denoising boundary matching based on an index for faster matching in very large image databases. Attempts have recently been made to convert boundary images to time-series with the objective of solving the partial denoising problem in boundary matching. In this paper, we deal with the disk I/O overhead problem of boundary matching to support partial denoising in a large image database. Although the solution to the problem superficially appears trivial as it only applies indexing techniques to boundary matching, it is not trivial since multiple indexes are required for every possible denoising parameters. Our solution is an efficient index-based approach to partial denoising using R*-tree in boundary matching. The results of experiments conducted show that our index-based matching methods improve search performance by orders of magnitude.

▸ **Key words**: Boundary matching, Partial denoising, Time-series data, Time-series matching, Indexing

[요 약]

본 논문에서는 대용량 이미지 데이터베이스에서 보다 빠른 매칭을 위한 색인 기반의 부분 노이즈 제거 윤곽선 매칭을 제안한다. 최근에는 윤곽선 매칭에서 부분 노이즈제거 문제를 해결하기 위해 윤곽선 이미지를 시계열로 변환하는 시도가 있어 왔다. 본 논문에서는 대용량 이미지 데이터베이스에서 부분 노이즈제거를 지원하기 위해 윤곽선 매칭의 디스크 I/O 오버헤드 문제를 다룬다. 이는 색인 기술을 윤곽선 매칭에 단순히 적용하면 되기 때문에 단순해 보이지만 가능한 모든 노이즈제거 매개변수에 대해 여러 개의 색인이 필요하기 때문에 어려운 문제이다. 이 문제를 해결하기 위해 본 논문에서는 윤곽선 매칭에서 R*-tree를 사용하여 부분 노이즈제거에 대한 효율적인 색인 기반 접근 방식을 제안한다. 수행 된 실험 결과, 제안한 색인 기반 매칭 방법은 검색 성능을 수백 배 향상시킨다.

▸ **주제어**: 윤곽선 매칭, 부분 노이즈제거, 시계열 데이터, 시계열 매칭, 색인

• First Author: Bum-Soo Kim, Corresponding Author: Bum-Soo Kim
*Bum-Soo Kim (bumsookim@kict.re.kr), Dept. of Future Technology and Convergence Research, Korea Institute of Civil Engineering and Building Technology
• Received: 2019. 09. 18, Revised: 2019. 10. 07, Accepted: 2019. 10. 07.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

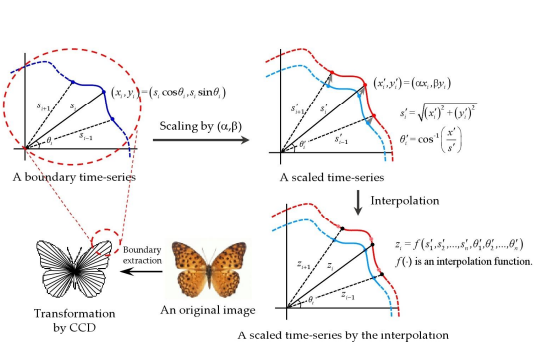
I. Introduction

최근 데이터베이스 크기가 기하급수적으로 증가함에 따라 이러한 데이터베이스에서 유용한 정보를 빠르게 찾을 수 있는 방법들이 절실히 요구되어지고 있다. 특히, 대용량 시계열 데이터베이스를 활용하기 위해 특정 시점의 값을 나타내는 실수 시퀀스인 시계열의 마이닝에 대한 수많은 연구가 수행되고 있다[1, 2, 3, 4, 5]. 예를 들어, 이미지 매칭[6, 7], 생물학적 서열 매칭[8], 필기 인식[9]과 같은 다양한 응용문제를 해결하기 위해 색인, 유사성 모델, 데이터 전처리 같은 시계열 매칭 기법들을 활용하는 몇 가지 방법들이 제안되었다. 본 논문에서는 시계열 매칭 기술들을 사용하는 윤곽선 (이미지) 매칭을 다룬다. 윤곽선 매칭은 먼저 윤곽선 이미지를 윤곽선 시계열로 변환한 다음, 주어진 질의 시계열과 유사한 데이터 시계열을 판별하는 것을 말한다[3, 6, 7, 10, 11].

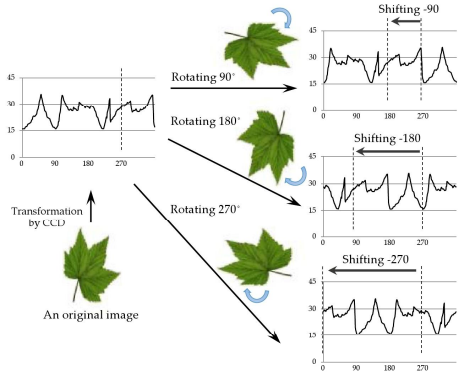
그림 1은 이미지 전처리 작업을 위해 활용된 윤곽선 시계열의 예제들이다. 좀 더 자세히 살펴보면, 그림 1(a)는 보간 함수를 사용함으로써 윤곽선 시계열의 스케일을 변환하는 과정을 나타낸다. 다음으로, 그림 1(b)에서 윤곽선 시계열은 길이 축을 단순히 이동시킴으로써 윤곽선 이미지의 회전된 효과를 얻을 수 있다. 그림 1(c)와 1(d)는 윤곽선 시계열에서 노이즈를 각각 전체적으로나 부분적으로 제거하는 과정을 각각 보여준다. 이때, 그림 1(d)에서 부분 노이즈가 제거된 시계열을 부분 노이즈제거 시계열이라 한다[12]. 이와 같이 윤곽선 이미지들을 윤곽선 시계열로 변환하는 것은 이미지 도메인에서의 윤곽선 전처리 문제를 시계열 도메인에서의 문제로 변환할 수 있음을 의미한다. 따라서, 윤곽선 이미지들의 시계열 변환은 신속한 전처리를 기반으로 빠른 검색에 장점이 있는 시계열 매칭 기술을 활용하여 대용량 이미지 데이터베이스라도 빠른 윤곽선 매칭이 가능하다.

Kim et al. [6]의 최근 연구에서는 부분 노이즈제거를 지원하는 윤곽선 매칭 문제를 해결하기 위해 노력했다. 그러나, 대용량 이미지 데이터베이스에서 이러한 매칭에는 모든 데이터 이미지가 최소한 한 번 이상 접근해야 하므로 디스크 I/O 오버헤드가 발생한다. 만약 데이터 이미지 수가 충분히 많게 되면 이 오버헤드 또한 상당히 크게 발생한다. 따라서, 본 논문에서는 대용량 이미지 데이터베이스에서 부분 노이즈제거 윤곽선 매칭의 디스크 I/O 오버헤드 문제를 해결하는 효율적인 방법을 제안한다.

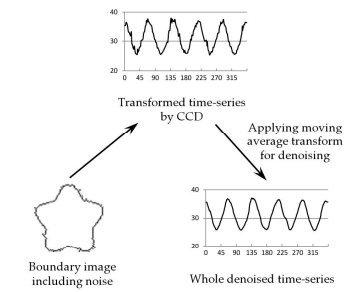
색인 기반의 접근법은 I/O 오버헤드 문제를 해결하기 위한 일반적인 방법이다[1, 12, 13, 14]. 이러한 해결책은 부분 노이즈제거 윤곽선 매칭에 색인 기술들을 단순히 적



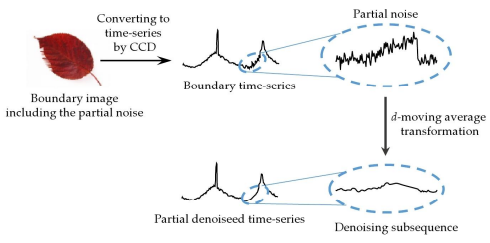
(a) A scale-converted boundary time-series [11].



(b) A rotation-converted boundary time-series [7].



(c) A whole denoised boundary time-series [10].



(d) A partial denoised boundary time-series [6].

Fig. 1. Examples of boundary matching utilized for image preprocessing.

용하기만 하면 된다. 그러나, 각각 노이즈 제거 레벨과 노이즈 길이가 달라질 때마다 새로운 색인들이 필요하므로 단순한 문제가 아니다. 즉, 기존 색인 방법으로는 여러 개의 색인들을 사용해야 하므로 심각한 유지관리 오버헤드와 디스크 용량 부족 문제가 발생한다.

부분 노이즈제거 윤곽선 매칭의 디스크 I/O 오버헤드 문제를 해결하기 위해 본 논문에서는 먼저 다차원 색인을 사용한 색인 기반 접근법을 제안한다. 색인을 구축하기 위해서는 먼저 각각의 데이터 시계열마다 부분 노이즈제거 시계열들을 저차원 점으로 변환한 후에 이 점들을 포함하는 저차원 MBR(minimum bounding rectangle)을 구성한다. 이때, 이 저차원 MBR를 R*-tree[15]와 같은 다차원 색인에 저장한다. 이는 부분 노이즈제거 시계열들을 이루는 값들이 서로 매우 유사하기 때문에 하나의 MBR을 구성할 수 있다. 즉, 부분 노이즈제거 위치가 다른 경우에도 부분 노이즈제거 시계열의 원본에 해당하는 시계열 항목 값들이 다른 부분 노이즈제거 시계열들에 자주 사용되기 때문에 하나의 MBR로 구성해도 검색에 사용할 수 있을 것이라는 직관에 기인한다. 따라서, 본 논문에서는 하나의 데이터 시계열에서 발생하는 모든 부분 노이즈제거 시계열들의 저차원 변환된 값들로 구성한 저차원 MBR인 *dl*-MBR을 정형적으로 정의한다. 또한, 착오 기각[1, 16]이 발생하지 않음을 보이고 색인 기반 방법의 정확성을 증명한다.

실험을 통해 본 논문에서는 제안한 색인 접근법이 우수함을 보이기 위해 기존의 방법과의 걸린 시간을 비교한다. 실험 결과, 제안한 색인 접근법은 기존 방법에 비해 걸린 시간을 수백에서 수천 배까지 줄인다. 이는 제안한 색인 기반 접근법이 색인 단계에서 검색 중 불필요한 많은 윤곽선 이미지들을 전지하기 때문에 당연한 결과이다. 이러한 결과를 바탕으로, 제안한 색인 기반 접근법은 부분 노이즈제거 윤곽선 매칭을 실현하는데 상호대화방식의 실용적인 방법이라 사료된다.

II. Related works

2.1 Time-Series Matching

시계열 매칭은 시계열 데이터베이스에서 주어진 시계열과 유사한 시계열들을 찾는 문제이다[1, 7, 16]. 시계열 매칭에서는 DTW(dynamic time warping)[17], LCSS(longest common subspace)[18], EDR(edit distance on real sequence)[19]과 같이 많은 유사 모델들(similarity models)이 연구되고 있다. 이들 유사 모델들은 제안한 방법과 직교적(orthogonal)이므로 함께 적용이 가능하다. 본 논문에서는 가장 폭넓게 사용하는 유사 모델 중 하나인 유클리디안 거리 모델[1, 16, 20]을 사용한다.

색인 구축 알고리즘에서는 길이 n인 데이터 시계열을 각각 저차원 변환이라고 하는 함수 $F(\cdot)$에 의해 f-차원 점($f \ll n$)으로 변환한 후, 다차원 색인에 저장한다. 저차원 변환은 고차원 문제[21]를 피하기 위해 널리 사용되며, 지나치게 큰 색인 공간을 줄여준다. 저차원 변환을 위해 DFT(discrete Fourier transform)[1], DWT(discrete wavelet transform)[22], SVD(singular value decomposition)[12], PCA(principal component analysis)[23], PAA(piecewise aggregate approximation)[24] 등과 같은 다양한 특성 추출 함수들이 제안되었다. 본 논문에서는 다차원 색인에서 저차원 변환으로 단순하고 직관적인 PAA를 사용한다. PAA는 주어진 길이 n의 시계열 $X(=\{x_0, x_1, \cdots, x_{n-1}\})$과 PAA $F(\cdot)$ 함수에 의해 변환된 f-차원 점 $F(X)=\{\breve{x}_0, \breve{x}_1, \cdots, \breve{x}_{f-1}\}$는 아래와 같이 식 (1)에 의해 정의된다.

$$F(X)=\{\breve{x}_0, \breve{x}_1, \cdots, \breve{x}_{f-1}\}, \quad (1)$$

$$\breve{x}_0=\frac{1}{\omega}\sum_{j=\omega i}^{\omega(i+1)-1} x_j, \text{ where } \omega=\frac{n}{f}.$$

시계열 매칭 알고리즘에서는 우선 질의 시계열을 f-차원 점으로 변환한 다음, 변환된 점과 허용치 ε으로 범위 질의를 구성한다. 이 범위 질의는 색인에서 질의 시계열과 잠재적으로 유사한 후보들(candidates)을 판별할 때 사용한다. 저차원 변환은 착오 기각이 발생하지 않음을 보장하지만 착오 알람의 원인이 된다[1, 16]. 그러므로, 시계열 매칭 알고리즘에서는 범위 질의를 통해 얻어진 후보들의 실제 데이터 시계열을 디스크에 접근하여 실제 유사한지를 판별하는 후처리 작업을 수행해야 한다. 마침내, 질의 시계열과 후보들의 실제 시계열들과 유사 거리를 계산하여 착오 알람의 시계열들을 후보에서 제거하여 유사한 시계열들을 얻게 된다.

2.2 Image Matching

흔히 CBIR(content-based image retrieval)라고 알려진 이미지 매칭은 이미지의 특성들을 사용함으로써 주어진 질의 이미지와 유사한 데이터 이미지들을 찾아내는 문제이다[13, 25, 26]. 이미지 매칭의 활용 분야들은 매칭에

Algorithm 1 *BuildNaiveIndex* ($\mathcal{T}$, d, l)

Input: Boundary time-series database $\mathcal{T}$, Denoising level d, Denoising length l

1: **for each** data time-series $T \in \mathcal{T}$ **do**
2: **for each** denoising position $p \in [0, n-1]$ of T **do**
3: Make a partial denoising time-series $\widetilde{T}_p^{d,l}$ from T;
4: Transform $\widetilde{T}_p^{d,l}$ to an f-dimensional point $F\left(\widetilde{T}_p^{d,l}\right)$ by using the low-dimensional transformation $F(\cdot)$;
5: Make a record $< T\text{-}ID, F\left(\widetilde{T}_p^{d,l}\right) >$, and store it into the index;
6: **end-for**
7: **end-for**

Fig. 2. The naive index-building algorithm for partial denoising boundary matching.

사용하는 특성에 따라 다르다. 본 논문에서는 색상과 질감이 유사한 이미지 매칭에 유용한 모양(shape) 특성에 초점을 둔다[27]. 그동안 모양 기반 이미지 매칭에서 윤곽선 시계열을 이용한 몇 가지 연구들이 수행되어 왔다[3, 6, 10, 11].

노이즈제거를 지원하는 윤곽선 매칭 문제는 Kim et al.[6, 10]에 의해 다뤄졌다. 처음 연구에서는 대용량 이미지 데이터베이스에서 윤곽선 전체 노이즈제거를 지원하는 전체 노이즈제거 윤곽선 매칭을 제안하였다[10]. 다음으로 전체 노이즈제거를 확장하여 부분 노이즈제거를 지원하는 윤곽선 매칭 즉, 부분 노이즈제거 윤곽선 매칭을 역시 제안하였다[6]. 전체 노이즈제거 윤곽선 매칭[10]은 대용량 이미지 데이터베이스를 위해 색인 기반 매칭 방법을 활용한다. 한편으로, 처음 제안한 부분 노이즈제거 윤곽선 매칭[6]은 시계열 매칭의 하한 기술들을 적용하여 효과적인 매칭 방법을 제시하였으나, 대용량 이미지 데이터베이스를 위해 색인 방법은 고려하지 않았다. 즉, 초기 부분 노이즈제거 윤곽선 매칭에서는 디스크 I/O 오버헤드 문제보다는 매칭 기술들에 초점이 맞추었다. 따라서, 본 논문에서는 시계열 매칭에서 널리 사용하는 R*-tree[1, 4, 10, 16, 28]를 이용하여 색인 기반 부분 노이즈제거 윤곽선 매칭을 제안한다.

III. Proposed Index-based Algorithms

일반적으로, 전형적인 색인 방법들은 먼저 고차원 점들을 저차원 점들로 변환한 다음, 색인에서 저차원 점들마다 저장한다[13, 29]. 그림 2는 부분 노이즈제거 윤곽선 매칭의 단순한 색인 구축 알고리즘을 나타낸다. 이 알고리즘은 색인 구축을 위한 기존의 전형적인 방법이다. 비록 노이즈제거 레벨 d와 노이즈제거 길이 l은 사용자에 의해 주어질지라도, 단순한 색인 구축 알고리즘은 모든 부분 노이즈제거 시계열로 부터 저차원 점들을 저장하는데 상당히 심각한 오버헤드가 발생하므로 실용적이지 못하다. 대용량 이미지 데이터베이스인 경우에는 매우 심각한 문제를 초래하게 된다. 결과적으로, 색인 기반 부분 노이즈제거 윤곽선 매칭을 위해서는 부분 노이즈제거 시계열의 저차원 점을 효율적으로 저장할 수 있는 색인 방법이 필요하다.

부분 노이즈제거 윤곽선 매칭에서는 유사 윤곽선 시계열을 얻기 위해 윤곽선 시계열로부터 부분 노이즈를 이미 안다는 것을 전제하고 부분 노이즈제거 윤곽선 시계열을 정형적으로 정의한다. 더 상세히 말하자면, 주어진 윤곽선 시계열 X, 노이즈제거 레벨 d, 노이즈제거 길이 l, 노이즈제거 위치 p가 있을 때, *부분 노이즈제거 윤곽선 시계열* $\widetilde{X}_p^{d,l}$은 k-계수 이동평균변환[28]을 사용하여 서브시퀀스 $X[p:p+l-1]$에서 노이즈를 제거한 서브시퀀스로 교체한 시계열이며, 식 (2)로 정의한다[6].

$$\tilde{X}_p^{d,l} = \left\{\tilde{x}_{p,0}^{d,l}, \tilde{x}_{p,1}^{d,l}, \ldots, \tilde{x}_{p,n-1}^{d,l}\right\}, \tag{2}$$

$$\tilde{x}_{p,i}^{d,l} = \begin{cases} x_i^{d,l} = \dfrac{1}{d}\displaystyle\sum_{j=i}^{i+d-1} x_{j\%n} & \text{if } i \in \{p\%n, \ldots, (p+l-1)\%n\}; \\ x_i & \text{otherwise} \end{cases}$$

여기서, $0 \le p \le n-1$, $2 \le d \le n-1$.

그러나, 부분 노이즈제거 시계열은 가능한 모든 부분 노이즈제거의 레벨, 길이, 위치들에 의해 만들어질 수 있다. 따라서, 부분 노이즈제거를 지원하는 윤곽선 매칭 문제를 단순하게 하기 위해, 본 논문에서는 부분 노이즈제거를 위해 사용자에 의해 노이즈제거 레벨 d와 노이즈제거 길이 l은 주어진다고 가정한다. 이때, 질의 시계열과 모든 가능한 부분 노이즈제거 시계열들과의 최소 거리를 유사 척도로 정의할 수 있다. 즉, X와 Y의 거리 $PDD(X, Y, d, l)$는 X와 Y의 모든 가능한 부분 노이즈제거 시계열들 간의 최소 거리로 정의한다. 그때, 부분 노이즈제거 윤곽선 매칭에서는 식 (3)에 의해 계산되는 $PDD(X, Y, d, l)$를 *부분 노이즈제거 거리*라 부른다[6].

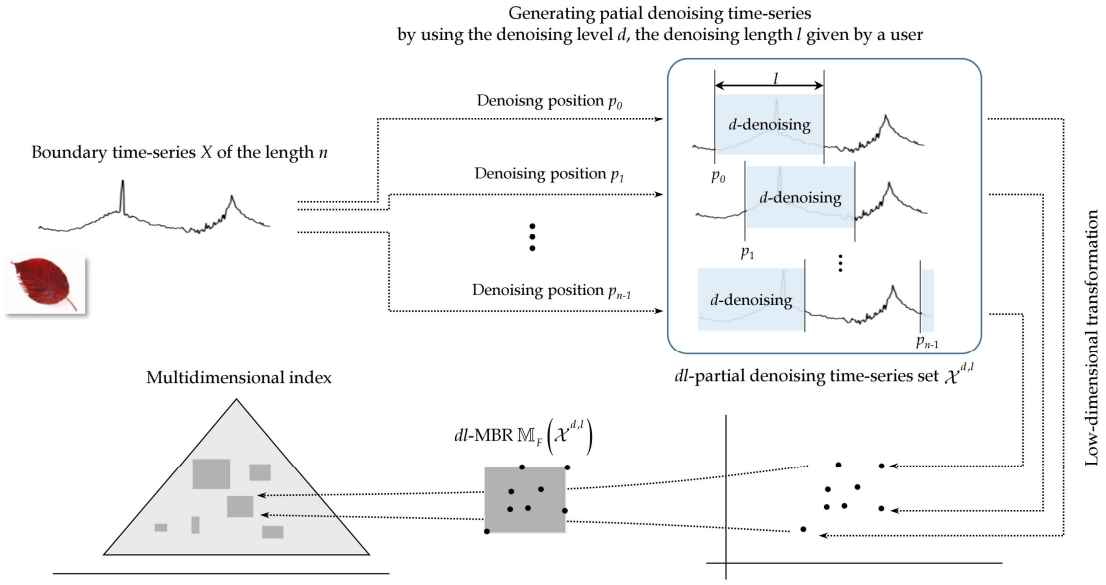


Fig. 3. An example of dl-MBR construction in a multidimensional index.

$$PDD(X,Y,d,l) = \min_{p=0}^{n-l} D\left(X,\tilde{Y}_p^{d,l}\right) = \min_{p=0}^{n-l}\sqrt{\sum_{i=0}^{n-l}\left|x_i-\tilde{y}_{p,i}^{d,l}\right|^2}. \quad (3)$$

여기서, $D(\cdot)$는 유클리디안 거리이다. 즉, $D(X,\tilde{Y}_p^{d,l})$는 X와 $\tilde{Y}_p^{d,l}$ 사이의 유클리디안 거리이다. 따라서, 부분 노이즈제거 거리의 개념을 사용하여 부분 노이즈제거 윤곽선 매칭의 문제를 다시 정형적으로 재정의한다. 즉, 부분 노이즈제거 윤곽선 매칭에서 주어진 질의 (윤곽선) 시계열 Q, 허용치 ε, 노이즈제거 레벨 d, 노이즈제거 길이 l이 있을 때, 만약 부분 노이즈제거 거리 $PDD(Q, T, d, l)$에 있는 데이터 이미지의 데이터 (윤곽선) 시계열 T가 ε보다 작거나 같으면 즉, $PDD(Q, T, d, l) \le \varepsilon$이면, T는 Q와 '유사하다'라고 한다. 역시, 이미지 데이터베이스로부터 '모든 유사한 이미지들을 찾았다'라고 부른다[6].

색인 기반의 부분 노이즈 제거 윤곽선 매칭을 위해 본 논문에서는 데이터 시계열을 효과적으로 색인에 저장하고 검색할 수 있는 방법을 제안한다. 식 (2)를 보면, 부분 노이즈제거 시계열에서 노이즈가 제거된 서브시퀀스를 제외하고 모든 엔트리들은 노이즈제거 위치가 다를지라도 다른 부분 노이즈제거 시계열에서 자주 사용된다. 즉, 원본이 같은 하나의 윤곽선 시계열로부터 생성된 부분 노이즈제거 시계열은 서로 유사하다. 이런 관찰은 색인에서 각각의 부분 노이즈제거 시계열 대신에 모든 부분 노이즈제거 시계열을 감싸는 MBR을 사용하는 색인 과정을 가능하게 한다. R*-tree와 같은 다차원 색인에서 효과적인 색인을 위해, 본 논문에서는 먼저 다음과 같이 부분 노이즈제거 시계열 집합을 정의한다.

정의 1. 길이 n의 주어진 윤곽선 시계열 X일 때, *dl-부분 노이즈제거 시계열 집합*인 $\tilde{\chi}^{d,l}$은 n개의 부분 노이즈제거 시계열 집합 $\{\tilde{X}_p^{d,l} | 0 \le p \le n-1\}$으로 정의한다.

다음으로, 본 논문에서는 l-부분 노이즈제거 시계열 집합 안에 있는 부분 노이즈제거 시계열로부터 변환된 모든 저차원 점들을 감싸는 MBR을 구성하고 다음과 같이 정의한다.

정의 2. 주어진 윤곽선 시계열 X이 있을 때, dl-부분 노이즈제거 시계열 집합 $\tilde{\chi}^{d,l}$에서 저차원 변환 $F(\cdot)$에 의해 부분 노이즈제거 시계열 $\tilde{X}_p^{d,l}$으로부터 변환된 $F(\tilde{X}^{d,l})$을 감싸는 MBR을 dl-MBR, $\{F(\tilde{X}_p^{d,l}) | 0 \le p \le n-1\}$이라 하고, $\mathbb{M}_F(\tilde{\chi}^{d,l})$이라 표기한다.

그림 3은 다차원 색인에서 저차원 변환된 부분 노이즈제거 시계열로부터 구성된 dl-MBR의 예제이다. dl-MBR 구성의 이해를 돕기 위해, 그림은 2차원 공간에서 점과 MBR들을 나타낸다. 부분 노이즈제거 시계열이 유사하지 않을 때, dl-MBR의 범위는 매우 넓다. 이는 dl-MBR이 얼마나 꽉 조이는지에 따라 매칭 성능이 결정됨을 의미한다. 제4장에서는 dl-MBR을 사용하여 제안한 색인 기반 접근법이 얼마나 효과적으로 부분 노이즈제거 윤곽선 매칭을 수행하는지를 실험을 통해 알아본다.

본 논문에서는 dl-MBR을 사용하는 것이 색인 기반의 부분 노이즈제거 윤곽선 매칭에서 착오 기각을 발생하지

Algorithm 2 *BuildBasicIndex* ($\mathcal{T}$, d, l)

Input: Boundary time-series database $\mathcal{T}$, Denoising level d, Denoising length l

1: **for each** data time-series $T \in \mathcal{T}$ **do**
2: Make an f-dimensional MBR $\mathbb{M}_F(\cdot)$ that is initially empty;
3: **for each** denoising position $p \in [0, n-1]$ of T **do**
4: Make a partial denoising time-series $\widetilde{T}_p^{d,l}$ from T;
5: **end-for**
6: Construct a set $\widetilde{\mathcal{T}}^{d,l}$ of all partial denoising time-series;
7: Construct a set of f-dimensional points from $\widetilde{\mathcal{T}}^{d,l}$ by using the low-dimensional transformation $F(\cdot)$;
8: Construct dl-MBR $\mathbb{M}_F\left(\widetilde{\mathcal{T}}^{d,l}\right)$ by bounding all f-dimensional points;
9: Make a record $< T\text{-}ID, \mathbb{M}_F\left(\widetilde{\mathcal{T}}^{d,l}\right) >$, and store it into the index;
10: **end-for**

Fig. 4. The basic index-building algorithm for partial denoising boundary matching.

않는다는 것을 정형적으로 증명한다. 먼저, 질의 시계열과 dl-MBR간의 거리는 질의 시계열과 부분 노이즈제거 시계열들 간 거리의 하한임을 다음과 같이 증명한다.

보조정리 1. 주어진 두 개의 윤곽선 시계열 X and Y이 있을 때, X와 Y의 부분 노이즈제거 시계열 $\widetilde{Y}_p^{d,l}$와의 거리의 하한 조건은 다음과 같이 식 (4)를 만족한다.

$$\forall p, D\left(F(X), \mathbf{M}\left(\Upsilon^{d,l}\right)\right) \le D\left(X, \tilde{Y}_p^{d,l}\right) \quad (4)$$

증명. $\widetilde{Y}_p^{d,l}$는 dl-부분 노이즈제거 시계열 집합의 정의에 의해 $\tilde{\Upsilon}^{d,l}$의 원소이다. 따라서, $\widetilde{Y}_p^{d,l}$는 고차원 MBR $\mathbf{M}\left(\tilde{\Upsilon}^{d,l}\right)$에 포함된다. 그로므로, X에서 $\widetilde{Y}_p^{d,l}$까지의 거리는 X와 $\mathbf{M}\left(\tilde{\Upsilon}^{d,l}\right)$의 거리보다 같거나 같다.

마침내, 본 논문에서는 다차원 색인에서 저차원 점과 dl-MBR과의 거리를 사용하는 것이 착오 기각을 발생하지 않음을 다음과 같이 보인다.

정리 1. 주어진 두 개의 윤곽선 시계열 X와 Y가 있을 때, 저차원 점 $F(X)$와 dl-MBR $\mathbf{M}\left(\tilde{\Upsilon}^{d,l}\right)$와의 거리 $D\left(F(X), \mathbf{M}\left(\Upsilon^{d,l}\right)\right)$가 X와 $\tilde{\Upsilon}^{d,l}$의 부분 노이즈제거 시계열 $\widetilde{Y}_p^{d,l}(0 \le p \le n-1)$의 하한 조건을 만족하면, 다음과 같이 식 (5)가 성립한다.

$$\forall p, D\left(F(X), \mathbf{M}_F\left(\Upsilon^{d,l}\right)\right) \le D\left(X, \tilde{Y}_p^{d,l}\right) \quad (5)$$

증명. 먼저, 식 (6)은 Y가 부분 노이즈제거 시계열 $\widetilde{Y}_p^{d,l}$로 대체한 것을 제외하고 Faloutsos et al. [16]에서 보조정리 1과 동일하다.

$$\forall p, D\left(F(X), F\left(\tilde{Y}_p^{d,l}\right)\right) \le D\left(X, \tilde{Y}_p^{d,l}\right) \quad (6)$$

한편, 식 (7)은 식 (4)로부터 유도될 수 있다.

$$\forall p, D\left(F(X), \mathbf{M}_F\left(\tilde{\Upsilon}^{d,l}\right)\right) \le D\left(F(X), F\left(\tilde{Y}_p^{d,l}\right)\right) \quad (7)$$

따라서, 정리의 식 (5)은 식 (6)과 (7)에 의해 명백히 성립한다.

정리 1은 저차원 점 $F(X)$와 Y의 dl-MBR $\mathbf{M}_F\left(\tilde{\Upsilon}^{d,l}\right)$ 사이의 거리가 사용자에게 주어진 허용치 ϵ에 작거나 같다면 그런 Y로 구성된 후보들은 착오 기각이 발생하지 않음을 의미한다. 정리 1에 기반하여 본 논문에서는 색인 구축과 색인 기반 매칭 알고리즘을 각각 제시한다.

그림 4는 부분 노이즈제거 윤곽선 매칭을 위한 기본 색인 구축 알고리즘을 나타낸다. 알고리즘 입력 값은 윤곽선 시계열 database T, 노이즈제거 레벨 d, 노이즈제거 길이 l이며, 출력 값은 다차원 색인이다. 라인 2에서 9에서는 데이터 시계열 T로부터 부분 노이즈제거 시계열들을 생성하는 과정, 저차원 변환, MBR 구성을 진행한다. 라인 2에서는 먼저 f-차원 MBR을 초기화한다. 라인 4에서는 노이즈제거 위치 p가 변함으로써 T로부터 부분 노이즈제거 시계열을 생성한다. 라인 6에서는 모든 부분 노이즈제거 시계열들을 구성하는 집합을 구축한다. 라인 7에서는 저차원 변환 $F(\cdot)$를 사용함으로써 부분 노이즈제거 시계열을 f-차원 점들로 변환한다. 라인 8에서는 이 점들을 감쌈으로써 f-차원 MBR을 구성한다. 라인 9에서는 마지막으로 Y의 Y-ID 식별자와 함께 색인에 이 MBR을 저장한다. 각각의 데이터 시계열은 라인 2에서 9까지 반복하여 색인을 구축한다.

그림 5는 색인 기반 매칭 알고리즘을 나타낸다. 입력 값은 질의 시계열 Q, 허용치 ϵ, 노이즈제거 레벨 d, 노이즈제거 길이 l이다. 출력 값은 질의 시계열과 유사한 데이터 시계열들이다. 알고리즘을 보면, 라인 2에서는 먼저 저차원 변환 $F(\cdot)$을 사용함으로써 질의 시계열을 f-차원 점으로 변

Algorithm 3 $IndexBasedMatching\,(Q,\,\epsilon,\,d,\,l)$

Input: Query time-series Q, tolerance ϵ, Denoising level d, Denoising length l
Output: The result set $\mathcal{R}$

```
1: R := ∅
2: Transform Q to an f-dimensional point F(Q) by using the transformation F(·);
3: Make a range query using F(Q) and ε;
4: Construct a candidate set C by evaluating the range query on the index;
5: for each candidate set C ∈ C do      // Start the post-processing step
6:     if PDD(Q,C,d,l) ≤ ε then
7:         R := R ∪ {C};
8:     end-if
9: end-for
10: return R;
```

Fig. 5. The index-based boundary matching algorithm supporting partial denoising.

환한다. 라인 3에서는 $F(Q)$와 허용치 ε을 사용하여 f-차원 범위 질의를 구성한다. 라인 4에서는 다차원 색인에서 범위 질의를 검색하고, 질의 시계열과 잠재적으로 유사한 후보 시계열 집합을 구성한다. 이 후보집합은 진짜 유사한 시계열에 더하여 착오 알람을 포함한다. 마침내, 라인 6에서는 데이터베이스로부터 질의 시계열로부터 그들의 부분 노이즈제거 거리를 계산하고 진짜 데이터 시계열들을 검색함으로써 착오 알람을 제거하는 단계인 후처리 과정을 수행한다.

IV. Experimental Evaluation

본 논문에서는 실험을 위해 Kim et al. [6]에서 사용한 합성 윤곽선 데이터셋을 사용하였다. 이 데이터셋은 모두 웹에서 수집한 원본 이미지로부터 각각 길이와 위치를 변화하여 생성된 아홉 가지 다른 부분 노이즈들이 포함된 길이 360의 102,590 개 윤곽선 시계열으로 구성되어 있다. 부분 노이즈는 가우시안 노이즈 모델 [26]을 사용하여 생성하였다. 논문에서 원본 이미지는 일만 개만을 사용하였으나 [3, 6, 10, 11], 하나의 이미지에 여러 개의 윤곽선 오브젝트를 포함될 수 있기 때문에 *CCD* 방법을 사용하여 추출된 윤곽선 시계열의 개수는 더 많게 된다.

수행된 실험 환경은 다음과 같다. 하드웨어는 2.0GHz Intel Core 2 Duo CPU, 2.0GB RAM, 500GB 하드디스크를 장착한 IBM 호환 PC이다. 소프트웨어는 CentOS 6.3 운영체제를 기반으로 C/C++ 언어를 사용하여 본 논문에서 제안한 색인 구축 알고리즘과 매칭 알고리즘을 구현하였다. 다차원 색인은 R*-tree를 사용하였으며, 색인과 데이터 페이지 크기는 4,096 바이트으로 설정하였다. 저차원 변환은 PAA를 사용하였으며, 저차원 변환을 이용하여 각 시계열을 72개의 특성으로 추출하였다.

본 논문에서는 제안한 색인 방법을 사용한 부분 노이즈제거 윤곽선 매칭의 성능 향상을 확인한다. 성능 실험에서는 부분 노이즈제거를 지원하는 윤곽선 매칭 알고리즘들의 걸린 시간을 비교한다. 즉, 제안한 색인 기반 매칭 알고리즘인 BBI와 색인 사용하지 않은 두 개의 시퀀스 매칭 알고리즘인 NIV-OG와 NIV-OP와의 걸린 시간을 비교한다. NIV-OG은 Kim et al. [6]에 의해 제안된 단순한 매칭 알고리즘이며, NIV-OP은 NIV-OG의 최적화된 알고리즘이다. 실험에서 질의 시계열은 일만 개의 원본 이미지들 중 백 개의 윤곽선 시계열을 임의로 선정하여 사용한다. 그 다음으로, 백 개의 질의 시계열으로부터 매칭에 걸린 시간을 측정하여 실험 결과로서 그 평균 값을 사용한다. 한편, 제안한 색인 기반 매칭 방법은 기존의 모양 매칭(shape matching)의 심각한 오버헤드가 발생하기 때문에 모양 매칭과 직접 비교하기가 쉽지 않다. 예를 들어, 모양 매칭에서 가장 널리 사용되고 있는 모양 문맥 매칭(shape context matching)[25]은 하나의 질의 시계열을 처리하는데 평균 약 676 초가 걸리는 반면에, 제안한 색인 기반 매칭 방법은 평균 약 0.9 초 걸린다. 따라서, 대용량 윤곽선 이미지 데이터베이스를 대상으로 모양 문맥 매칭을 사용하는 것은 매우 어렵다. 반면에, 제안한 색인 기반 매칭 방법은 대용량 윤곽선 이미지 데이터베이스를 대상으로 사용하기 적합하다.

그림 6은 제안한 매칭 알고리즘의 확장성을 보여준다. 그림 6을 보면, 제안한 색인 기반 매칭 알고리즘이 기존 매칭 알고리즘에 비해 성능을 크게 향상시켰다. (그래프의 *Y*축이 로그-스케일임에 유의한다.) 성능 실험은 데이터 개수에 따른 확장성을 보여준다. 좀 더 자세히 보면, NIV-OG와 NIV-OP인 경우에는 순차검색이므로 데이터 개수가 증가할수록 걸린 시간 또한 비례하여 증가하는 반면에, 색인 기반 매칭 알고리즘인 BBI는 로그 함수 형태로 증가함을 알 수 있다. 이는 제안한 알고리즘이 대용량 이미지 데이터베이스에 더 적합함을 의미한다. 따라서, 제안

한 색인 기반 매칭 알고리즘은 실제 대용량 이미지 데이터베이스일지라도 단지 몇 초안에 결과를 얻을 수 있다. 실험 결과, BBI는 NIV-OG와 NIV-OP에 비해 약 258.2 배에서 2580.8 배까지 성능을 향상시켰다. 결과적으로, 본 논문에서 제안한 색인 기반의 접근법은 대용량 윤곽선 이미지 데이터베이스를 대상으로 부분 노이즈제거 윤곽선 매칭을 수행하는데 효율적인 방법임을 확신한다.

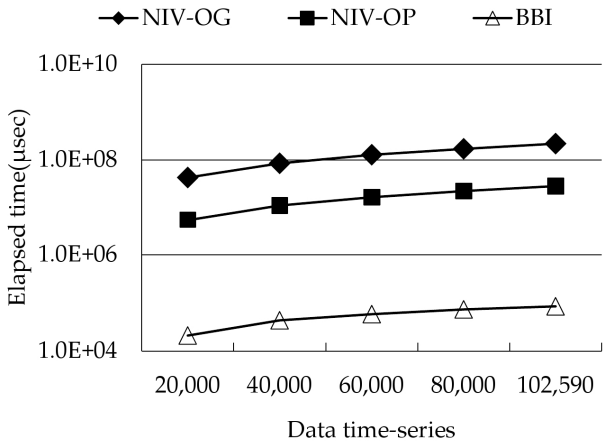


Fig. 6. The scalability of partial denoising boundary matching algorithms(d=24, ℓ=72).

V. Conclusions

본 논문에서는 다차원 색인을 사용하여 부분 노이즈제거 윤곽선 매칭에서 여러 가지 노이즈제거 매개변수들을 지원하는데 발생하는 오버헤드 문제를 해결하였다. 본 논문의 공헌으로는 다음과 같다. 먼저, R*-tree와 저차원 변환인 PAA를 사용하여 윤곽선 매칭에서 부분 노이즈제거를 지원하는 효율적인 색인 기반 접근법을 제안했다. 둘째, 임의의 노이즈제거 매개변수를 지원하고 단일 색인 구축 알고리즘을 제시하고, 이의 정확성을 정형적으로 증명했다. 셋째, 실험을 통해 색인 기반 매칭 알고리즘이 기존 매칭 알고리즘보다 우수함을 보였다. 제안한 방법의 한계점은 윤곽선이 폐곡선을 형성하는 물체 이미지에 적합하다. 윤곽선 연결이 짧게 끊어진 부분은 선형 보간을 사용하여 윤곽선을 추출하지만, 길게 끊긴 부분이나 뾰족한 부분은 논문에 사용된 *CCD* 방법으로 적절하게 시계열을 추출하기가 어렵다. 따라서, 향후 작업으로 다른 모양 특성에 대한 새로운 해결책을 연구할 예정이다. 또한, 실험 결과는 색인 기반의 매칭 알고리즘이 기존의 매칭 알고리즘보다 수백에서 수천 배나 더 우수한 것으로 나타났다. 향후 연구로는 구축 성능을 좀 더 향상하는 방법과 새로운 접근법에 대한 연구를 수행할 예정이다.

REFERENCES

[1] R. Agrawal, C. Faloutsos, and A. Swami, “Efficient Similarity Search in Sequence Databases,” In Proc. of the 4th Int'l Conf. on Foundations of Data Organization and Algorithms, Chicago, Illinois, pp. 69-84, Oct. 1993.

[2] Y. Cai, H. Tong, W. Fan, P. Ji, and Q. He, “Facets: Fast Comprehensive Mining of Coevolving High-Order Time Series," In Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Sydney, Australia, pp. 79-88, Aug. 2015.

[3] W.-K. Loh, S.-P. Kim, S.-K. Hong, and Y.-S. Moon, “Envelope-based Boundary Image Matching for Smart Devices under Arbitrary Rotations,” Multimedia Systems, Vol. 21, No. 1, pp. 29-47, Feb. 2015.

[4] Y.-S. Moon and B. S. Lee, “Safe MBR-Transformation in Similar Sequence Matching,” Information Sciences, Vol. 270, No. 2, pp. 28-40, June 2014.

[5] J. Paparrizos and L. Gravano, “k-Shape: Efficient and Accurate Clustering of Time Series,” In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, Melbourne, Australia, pp. 1855-1870, May/June 2015.

[6] B.-S. Kim, Y.-S. Moon, and J.-G. Lee, “Boundary Image Matching Supporting Partial Denoising Using Time-Series Matching Techniques,” Multimedia Tools and Applications, Vol. 76, No. 6, pp. 8471-8496, Mar. 2017.

[7] Y.-S. Moon and W.-K. Loh, “Triangular Inequality -based Rotation-Invariant Boundary Image Matching for Smart Devices,” Multimedia Systems, Vol. 21, No. 1, pp. 15-28, Feb. 2015.

[8] A. Dalton, S. Patel, A. R. Chowdhury, M. Welsh, T. Pang, S. Schachter, G. Olaighin, and P. Bonato, “Development of a Body Sensor Network to Detect Motor Patterns of Epileptic Seizures,” IEEE Trans. on Biomedical Engineering, Vol. 59, No. 11, pp. 3204-3211, Nov. 2012.

[9] M. Bashir and F. Kempf, " Advanced Biometric Pen System for Recording and Analyzing Handwriting," Journal of Signal Processing Systems, Vol. 68, No. 1, pp. 75-81, July 2012.

[10] B.-S. Kim, Y.-S. Moon, M.-J. Choi, and J. Kim, “Interactive Noise-Controlled Boundary Image Matching Using the Time-Series Moving Average Transform,” Multimedia Tools and Applications, Vol. 72, No. 3, pp. 2543-2571, Oct. 2014.

[11] Y.-S. Moon, B.-S. Kim, M. S. Kim, and K.-Y. Whang,

"Scaling-Invariant Boundary Image Matching Using Time-Series Matching Techniques," Data & Knowledge Engineering, Vol. 69, No. 10, pp. 1022-1042, Oct. 2010.

[12] F. Korn, H. V. Jagaciish, and C. Faloutsos, "Efficiently Supporting Ad Hoc Queries in Large Data Sets of Time Sequences," In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, Tucson, Arizona, pp. 289-300, May 1997.

[13] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys, Vol. 40, No. 2, pp. 34-94, Apr. 2008.

[14] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. C. Jain, "Content-based Image Retrieval at the End of the Early Years," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1349-1380, Dec. 2000.

[15] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: An Efficient and Robust Access Method for Points and Rectangles," In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, pp. 322-331, Atlantic City, New Jersey, May 1990.

[16] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time-Series Databases," In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, pp. 419-429, Minneapolis, Minnesota, May 1994.

[17] W.-S. Han, J. Lee, Y.-S. Moon, S.-W. Hwang, and H. Yu, "A New Approach for Processing Ranked Subsequence Matching based on Ranked Union," In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, Athens, Greece, pp. 457-468, June 2011.

[18] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering Similar Multidimensional Trajectories," In Proc. of the 18th IEEE Int'l Conf. on Data Engineering, San Jose, California, pp. 673-684, Feb./Mar. 2002.

[19] L. Chen, M. T. Ozsu, and V. Oria, "Robust and Fast Similarity Search for Moving Object Trajectories," In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, Baltimore, Maryland, pp. 491-502, June 2005.

[20] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*(Ed. 2), Springer Texts in Statistics, 2006.

[21] S. Berchtold, C. Bohm, and H.-P. Kriegel, "The Pyramid-Technique: Towards Breaking the Curse of Dimensionality," In Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, pp. 142-153, Seattle, Washington, June 1998.

[22] K.-P. Chan, A. W.-C. Fu, and C. T. Yu, "Haar Wavelets for Efficient Similarity Search of Time-Series: With and Without Time Warping," IEEE Trans. on Knowledge and Data Engineering, Vol. 15, No. 3, pp. 686-705, Jan./Feb. 2003.

[23] R. Lesch, Y. Caille, and D. Lowe, "Component Analysis in Financial Time Series," In Proc. of the IEEE Int'l Conf. on Computational Intelligence for Financial Engineering, New York, New York, pp. 183-190, Apr. 1999.

[24] M. Krawczak, G. Szkatula, "An Approach to Dimensionality Reduction in Time Series," Information Sciences, Vol. 260, No. 2, pp. 15-36, Mar. 2014.

[25] S. Belongie, J. Malik, J. Puzicha "Shape Matching and Object Recognition Using Shape Contexts," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 4, pp. 509-522, Apr. 2002.

[26] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*, 4th Ed., Cengage Learning, 2014.

[27] P. Suetens, P. Fua, and A. J. Hanson, "Computational Strategies for Object Recognition," ACM Computing Surveys, Vol. 24, No. 1, pp. 5-62, Mar. 1992.

[28] Y.-S. Moon and J. Kim, "Efficient Moving Average Transform-based Subsequence Matching Algorithms in Time-Series Databases," Information Sciences, Vol. 177, No. 23, pp. 5415-5431, Dec. 2007.

[29] R. Arandjelovi´c and A. Zisserman, "Three Things Everyone Should Know to Improve Object Retrieval," In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2911-2918, Providence, Rhode Island, June 2012.

Authors

Bum-Soo Kim received his Ph. D. (2013) degrees in computer science from Kangwon National University. From 2013 to 2017, he was a postdoctoral researcher in Korea Advanced Institute of Science and Technology (2013 and 2015), Kangwon National University (2014), and Korea University (2016-2017). He is currently a postdoctoral researcher in Department of Future Technology and Convergence Research from Korea Institute of Civil Engineering and Building Technology (KICT). His research interests include time-series data mining, construction bigdata analysis, and data mining applications.