논문 2005-42CI-5-8

고차원 공간에서 유클리드 거리의 효과적인 근사 방안

(An Effective Method for Approximating the Euclidean Distance in High-Dimensional Space)

정 승 도*, 김 상 욱**, 김 기 동***, 최 병 욱**

(Seungdo Jeong, Sang-Wook Kim, Kidong Kim, and Byung-Uk Choi)

요 약

고차원 공간상의 벡터들 간의 유클리드 거리를 빠르게 계산하는 것은 멀티미디어 정보 검색을 위하여 매우 중요하다. 본 논문에서는 고차원 공간상의 두 벡터들 간의 유클리드 거리를 효과적으로 근사하는 방법을 제안한다. 이러한 근사를 위하여 두 벡터들의 놈(norm)을 사용하는 방법이 기존에 제안된 바 있다. 그러나 기존의 방법은 두 벡터간의 각도 성분을 무시하므로 근사 오차가 매우 커지는 문제점을 가진다. 본 연구에서 제안하는 방법은 기준 벡터라 부르는 별도의 벡터를 이용하여 추정된 두 벡터간의 각도 성분을 그들을 위한 유클리드 거리 근사에 사용한다. 이 결과, 각도 성분을 무시하는 기존의 방법과 비교하여 근사 오차를 크게 줄일 수 있다. 또한, 제안된 방법에 의한 근사 값은 유클리드 거리 보다 항상 작다는 것을 이론적으로 증명하였다. 이는 제안된 방법을 이용하여 멀티미디어 정보 검색을 수행할 때 착오 기각이 발생하지 않음을 의미하는 것이다. 다양한 실험에 의한 성능 평가를 통하여 제안하는 방법의 우수성을 규명한다.

Abstract

It is crucial to compute the *Euclidean distance* between two vectors efficiently in high dimensional space for multimedia information retrieval. In this paper, we propose an effective method for approximating the Euclidean distance between two high-dimensional vectors. For this approximation, a previous method, which simply employs norms of two vectors, has been proposed. This method, however, ignores the *angle* between two vectors in approximation, and thus suffers from large approximation errors. Our method introduces an additional vector called a *reference vector* for estimating the angle between the two vectors, and approximates the Euclidean distance accurately by using the estimated angle. This makes the approximation errors reduced significantly compared with the previous method. Also, we formally prove that the value approximated by our method is always smaller than the actual Euclidean distance. This implies that our method does not incur any *false dismissal* in multimedia information retrieval. Finally, we verify the superiority of the proposed method via performance evaluation with extensive experiments.

Keywords: multimedia information retrieval, feature vector, distance approximation, lower bound function, high-dimensional space

I. 서 론

초고속 정보통신망의 대중화로 인해 멀티미디어 데이터가 급격히 증가하고 있다. 멀티미디어 분야의 중요한 이슈 중 하나는 사용자가 원하는 데이터를 정확하고 빠르게 찾을 수 있는 효과적인 검색 기능을 제공하는 것이다.

멀티미디어 응용에 대한 많은 기존의 연구에서 멀티미디어 데이터는 특징 벡터로 표현되며, 멀티미디어 데이터베이스는 다차원 공간상의 특징 벡터들의 집합으로 간주된다. 따라서 멀티미디어 정보 검색은 다차원 공간

* 학생회원, 한양대학교 전자통신컴퓨터공학과
(Dept. of Electronics and Computer Engineering, Hanyang University)
** 정회원, 한양대학교 정보통신대학
(College of Information and Communications, Hanyang University)
*** 정회원, 강원대학교 산업공학과
(Dept. of Industrial Engineering, Kangwon National University)
※ 이 논문은 한양대학교 교내 연구비(HY-2003)와 제주대학교 IT 연구센터(텔리매틱스 요소 기술 연구) 연구비의 부분적인 지원을 받았습니다.
접수일자: 2005년7월5일, 수정완료일: 2005년9월6일

상에서 사용자의 질의를 만족하는 특징 벡터를 찾아내는 것으로 정의 된다. 효과적인 멀티미디어 정보 검색을 위해서는 많은 특징들이 사용되어야 하므로 특징 벡터는 일반적으로 수십에서 수백 차원의 고차원 벡터의 형태로 나타난다[3][7][8][9].

사용자의 질의는 질의 벡터와의 차이가 유사 허용치(ε) 이하인 특징 벡터들을 찾는 범위 질의(range query)와 질의 벡터와 가장 유사한 k개의 특징 벡터들을 찾는 k-최근접 질의(k-nearest neighbor query)로 분류된다[2][3][6]. 또한, 많은 연구들에서는 벡터간의 유사도 기준으로 유클리드 거리(Euclidean distance)를 사용하고 있다[1][3][7].

그러나 특징 벡터는 고차원 벡터이므로 두 특징 벡터간의 유클리드 거리를 계산하는 시간이 전체 검색 시간에 중요한 비중을 차지하게 된다. 질의 벡터와 전체 데이터 벡터간의 실제 유클리드 거리를 모두 계산하지 않고, 최종 질의 결과에 포함될 가능성이 높은 후보들만을 미리 구할 수 있다면 전체 검색에 필요한 시간을 크게 줄일 수 있을 것이다. 따라서 대부분의 기존 연구들은 정답 가능성을 가진 후보 벡터들의 집합을 파악하는 전처리 과정(pre-processing step)과 후보 집합내의 벡터들을 대상으로 실제 정답 여부를 확인하는 후처리 과정(post-processing step)으로 나누어 문제를 다루고 있다[1]. 이러한 검색 방식을 이단계 검색(two step searching)이라 한다. 반면, 질의 벡터와 전체 벡터 간의 거리를 모두 계산하여 정답을 결정하는 방식을 완전 검색(exhaustive searching) 방식이라 한다.

이단계 검색 방식이 항상 모든 정답들을 찾는 것을 보장하기 위해서는 전처리 과정에서 반드시 정답들을 모두 포함하는 후보 집합을 형성해야 한다. 착오 기각(false dismissal)이란 정답이 전처리 과정에서 후보 집합에서 제외됨으로써 최종 정답 집합에서 제외되는 경우를 말한다. 착오 채택(false alarm)은 정답이 아니면서 전처리 과정에서 후보 집합 내에 포함됨으로써 불필요하게 후처리 과정을 거치게 되는 경우를 말한다. 착오 채택은 후처리 과정에서 제거되지만 착오 채택의 개수가 많으면 실제 거리를 계산해야 하는 벡터가 많아지기 때문에 후처리 과정의 수행 시간이 증가하게 된다. 따라서 멀티미디어 정보 검색에서는 정답을 보장하는 동시에 수행 시간을 줄이기 위해 착오 기각은 없애고 착오 채택은 최소화하는 것을 목표로 한다.

유클리드 거리 계산의 효율성을 높이고자 시도하였던 기존의 연구들은 거리를 근사하기 위하여 Cauchy-Schwartz 부등식을 사용한다[4][5]. Cauchy -Schwartz 부등식은 두 벡터의 내적(inner product)의 상한(upper bound)을 정의하기 위하여 벡터의 놈(norm)을 사용한다. 두 벡터에 대한 유클리드 거리 함수는 두 벡터의 내적을 포함하고 있기 때문에 Cauchy-Schwartz 부등식을 이용하여 유클리드 거리 계산 함수에 대한 하한(lower bound)을 정의할 수 있다. 따라서 Cauchy-Schwartz 부등식을 이용한 하한 함수를 이용하여 전처리를 수행함으로써 정답을 모두 포함하는 후보 집합을 형성할 수 있다.

이 경우, 두 벡터 간 내적의 계산은 각 벡터의 놈에 해당하는 두 수의 곱으로 대체되기 때문에 연산량을 크게 줄일 수 있다는 것이 장점이다. 그러나 이 하한 함수의 결과 값은 실제 유클리드 거리와 차이가 매우 크다는 단점을 가지고 있다. 본 논문에서는 실제 유클리드 거리와 이를 근사하는 함수에 의한 거리의 차를 근사 오차라고 정의한다. 근사 오차가 커지는 경우 후보 집합에 포함되는 착오 채택의 개수가 증가하므로 후처리 과정을 지연시키는 결과를 초래한다.

각 벡터의 일차 놈을 사용하는 경우의 단점을 해결하기 위하여 k차 놈을 사용하여 유클리드 거리를 근사하는 크기 근사(magnitude approximation) 기법이 제안되었다[4][5]. 그러나 이 기법은 제안된 거리 근사 함수에서 사용하는 적절한 계수를 선정하기가 어렵다는 단점이 있다. 이는 각 계수가 데이터 분포와 차수에 의존하여 선정되기 때문이다. 또한, 제안된 근사 함수는 하한 함수가 아니므로 착오 기각이 발생한다는 심각한 단점을 가지고 있다.

벡터의 놈은 대칭성(symmetric property)을 갖는다. 이러한 특징은 서로 다른 벡터가 동일한 놈을 가지는 것을 허용하며, 결과적으로 근사 오차를 증가시키는 요인이 된다. k차 놈의 대칭성에 의한 근사 오차를 줄이기 위하여 각 벡터가 가지는 형태 정보를 반영하여 유클리드 거리를 근사하는 거리 근사 기법인 형태 근사(shape approximation) 기법이 제안되었다[6]. 이 방법은 크기 근사 기법과 동시에 사용함으로써 후보 집합을 상당히 줄일 수 있는 장점이 있다. 그러나 이 기법 역시 유클리드 거리에 대한 하한 조건을 만족하지 않기 때문에 착오 기각을 발생시키는 단점을 갖는다.

본 논문에서는 유클리드 거리를 효과적으로 근사할 수 있는 기법에 관하여 다룬다. 두 벡터간의 유클리드 거리는 각 벡터의 놈과 두 벡터가 이루는 각도 성분을 이용하여 구할 수 있다. 그러나 기존의 일부 연구에서

는 각도 성분을 무시함으로써 거리 근사 오차가 커지는 단점을 가진다. 또다른 일부의 연구에서는 거리에 대한 하한 조건을 보장하지 못하는 단점을 가지고 있다[4][5][6]. 본 논문에서 제안하는 기법은 유클리드 거리 계산에 있어서 각도 성분을 고려하여 거리 근사 오차를 줄임으로써 착오 채택의 개수를 효과적으로 줄일 수 있다. 따라서 후처리 과정의 수행 시간을 크게 줄일 수 있다. 뿐만 아니라 유클리드 거리에 대한 하한 조건을 만족하도록 함으로써 착오 기각을 발생시키지 않는 것을 보장한다.

본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 관련 연구로서 유클리드 거리에 대한 근사 함수들을 소개하고, 장단점을 논의한다. 제 Ⅲ장에서는 제안하는 거리 근사 기법에 관하여 자세히 다룬다. 제 Ⅳ장에서는 본 논문에서 제안하는 기법을 적용할 수 있는 응용에 관하여 다룬다. 제 Ⅴ장에서는 제안된 기법의 우수성을 다양한 성능 평가를 통하여 검증한다. 끝으로, 제 Ⅵ장에서는 본 논문을 요약하고, 결론을 내린다.

Ⅱ. 관련 연구

본 장에서는 두 벡터간 유클리드 거리 함수의 변형을 이용하여 검색 속도를 향상시키고자 시도한 기존 연구를 살펴보고, 그 문제점을 지적한다.

1. Cauchy-Schwartz 부등식을 이용한 근사

$$< X, Y >= \sum_{i=1}^{N} x_i \cdot y_i \le \|X\| \|Y\| \quad (1)$$

$$D(X, Y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} = \sqrt{\|X\|^2 + \|Y\|^2 - 2 < X, Y >} \quad (2)$$

$$D(X, Y) \ge D_{cs}(X, Y) = \sqrt{\|X\|^2 + \|Y\|^2 - 2\|X\| \|Y\|} \quad (3)$$

식 (1)은 두 벡터의 내적 $< X, Y >$과 놈(norm)의 관계를 표현하는 Cauchy-Schwartz 부등식을 나타낸다. 즉, Cauchy-Schwartz 부등식은 주어진 두 벡터 내적의 상한(upper bound)을 정의한다. 식 (2)는 두 벡터 간의 유클리드 거리 함수 $D(X, Y)$를 정의하고 있으며, 두 벡터의 놈과 내적을 이용하는 식으로 표현된다. 유클리드 거리 함수의 내적 항을 식 (1)에 따라 벡터 놈의 곱으로 대체하면 식 (3)에서와 같이 Cauchy-Schwartz 부등식을 사용한 유클리드 거리 함수의 하한(lower bound) $D_{cs}(X, Y)$를 정의할 수 있다.

이러한 결과를 멀티미디어 정보 검색에 적용하기 위해서 모든 데이터 벡터에 대하여 각 벡터의 놈을 요약 정보로써 저장해 둔다. 질의 벡터가 들어왔을 때 질의 벡터의 놈과 저장되어 있는 각 데이터 벡터의 놈을 이용하여 두 벡터간의 유클리드 거리를 식 (3)의 오른쪽 항을 이용하여 근사한다. 근사된 거리가 유사 허용치(ε) 이하인 경우 해당 데이터 벡터를 후보 집합에 포함시킨다. 식 (3)의 오른쪽 항이 유클리드 거리의 하한 함수이기 때문에 근사된 거리는 실제 거리보다 항상 작으며, 이 결과 착오 기각이 발생하지 않는다. 뿐만 아니라 놈만을 이용하여 계산하므로 계산량이 차원 수에 비례하여 증가하는 유클리드 거리 계산에 비하여 계산 시간을 크게 줄일 수 있다.

그러나 Cauchy-Schwartz 부등식을 이용하는 방식은 벡터간의 거리를 단 하나의 놈 값만 사용하여 근사하기 때문에 근사 오차가 크다는 단점을 갖는다. 특히, 벡터의 차원이 커질수록 근사 오차가 커지고, 결과적으로 실제 거리는 크지만 근사된 거리가 유사 허용치 보다 작아져 착오 채택되는 후보들의 개수가 크게 증가하게 된다. 이는 후처리 과정의 수행 시간을 증가시키는 요인이 된다.

2. 크기 근사

크기 근사 기법은 Cauchy-Schwartz 부등식을 사용했을 때 근사 오차가 크다는 단점을 해결하기 위하여 제안된 방법이다[4][5]. 크기 근사 기법에서 사용하는 근사 내적 함수는 식 (4)와 같다.

$$< X, Y >^k$$

$$\approx b_1 \Psi_1(X)\Psi_1(Y) + b_2 \Psi_2(X)\Psi_2(Y) + \ldots + b_k \Psi_k(X)\Psi_k(Y) \quad (4)$$

$$\Psi_k(X) = x_1^k + x_2^k + \ldots + x_n^k \quad (5)$$

식 (4)에서 Ψ_k는 식 (5)와 같이 벡터의 k차 놈의 k제곱을 나타내고, b_i는 근사 함수의 각 항을 위한 최적 계수를 나타낸다. 최적 계수는 최소 자승 기법을 이용하여 구할 수 있으며, 제안된 내적 근사 함수를 식 (2)에 대입함으로써 유클리드 거리를 근사할 수 있다. 그러나 최적 계수는 데이터의 분포 특성과 k값에 따라 각각 달라진다. 또한, 제안된 거리 근사 함수는 유클리드 거리에 대한 하한을 보장할 수 없다[4][5]. 즉, 제안된 거리 근사 함수에 의해 후보 집합을 구성하는 경우, 정답들이

제외되는 착오 기각이 발생한다는 중대한 단점을 가진다. 따라서 이 기법은 착오 기각을 허용하는 일부의 응용에 한하여 적용이 가능하다.

3. 형태 근사

식 (5)는 대칭성(symmetric property)을 갖고 있다. 즉, 각 속성(attribute)의 순서만 다른 모든 벡터들은 동일한 k차 놈을 갖는다. 이는 두 벡터간의 유클리드 거리를 근사할 때 발생하는 오차의 주요 요인이 된다. 형태 근사는 놈의 대칭성에 의한 근사 오차를 줄이기 위한 방법으로 제안되었다[6]. 하나의 벡터를 속성이 서로 겹치지 않는 부분 그룹으로 분리하고, 부분 그룹별 k차 놈을 개별적으로 계산함으로써 벡터의 각 항의 배치 형태 정보를 일부 반영한다. 이 결과, 전혀 다른 두 벡터가 유사한 k차 놈을 가지게 됨으로써 후보로 선택되는 단점을 극복할 수 있다. 그러나 형태 근사 기법 역시 유클리드 거리 함수에 대한 하한 함수가 아니기 때문에 착오 기각이 발생한다는 단점을 가지고 있다.

III. 제안하는 기법

본 장에서 두 벡터간의 각도 성분을 근사하여 유클리드 거리 함수에 적용함으로써 좀 더 실제 유클리드 거리와 가까운 값을 반환하는 근사 함수를 제안한다.

1. 기본 개념

유클리드 거리 함수를 다시 정리해 보면 식 (6)과 같다. 식 (3)에서 Cauchy-Schwartz 부등식을 이용한 거리 추정 함수는 두 벡터간의 각도 성분을 전혀 고려하지 않고 있다. 각도 성분을 고려하지 않는 것이 Cauchy-Schwartz 부등식을 사용한 거리 함수에 의한 근사 오차의 직접적인 원인이다.

$$D(X, Y) = \sqrt{\|X\|^2+\|Y\|^2-2<X, Y>} = \sqrt{\|X\|^2+\|Y\|^2-2\|X\|\|Y\|cos\theta} \quad (6)$$

본 논문에서는 두 벡터간의 각도를 근사함으로써 두 벡터 간 유클리드 거리를 보다 정확하게 근사하는 방법을 제안한다. 식 (6)에서와 같이 주어진 두 벡터간의 유클리드 거리는 각 벡터의 놈과 두 벡터 사이의 각도를 이용하여 구할 수 있다. 그러나 질의 벡터가 주어지기 전까지는 질의 벡터와 데이터 벡터가 이루는 각도를 미리 파악할 수 없다. 질의 벡터가 주어진 후에 모든 데이터 벡터들과의 각도를 각각 계산해야 한다면, 검색 시간이 매우 길어지게 될 것이다. 본 연구에서 제안하는 기법의 핵심 아이디어는 데이터베이스 구축 단계에서 이 각도 성분을 근사할 수 있는 사전 정보를 기록함으로써 검색 시간을 단축시키자는 것이다.

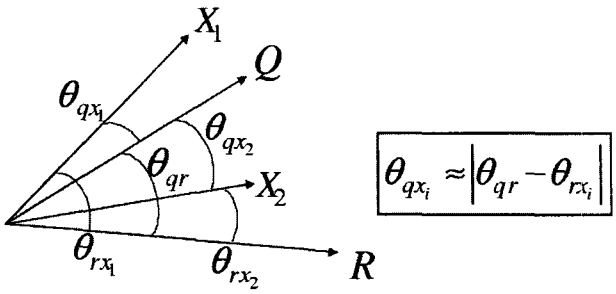


그림 1. 기준 벡터를 이용한 각도 근사 방법
Fig. 1. Angle approximation using a reference vector.

$$\overline{\theta_{qx_i}} = |\theta_{qr} - \theta_{rx_i}| \quad (7)$$

$$\theta_{rx_i} = cos^{-1}\left(\frac{< X_i, R >}{\|X_i\|\|R\|}\right) \quad (8)$$

$$\overline{D}(Q, X_i) = \sqrt{\|X_i\|^2+\|Q\|^2-2\|X_i\|\|Q\|\cos\overline{\theta_{qx_i}}} \quad (9)$$

본 논문에서 제안하는 각도 근사 방법을 그림 1에 나타내었다. 제안하는 방법은 기준 벡터라는 개념을 사용한다. 그림에 나타난 바와 같이, 임의의 기준 벡터 R이 하나 주어졌을 때, 기준 벡터와 i번째 데이터 벡터 X_i 간의 각도 θ_{rx_i}를 알 경우, 질의 벡터와 기준 벡터간의 각도 θ_{qr}만을 계산하면 질의 벡터와 i번째 데이터 벡터와의 각도 θ_{qx_i}는 식 (7)과 같은 단순한 연산만으로 근사할 수 있다. 따라서 본 논문에서는 이와 같은 방식을 이용한 각도 근사를 위해 기준 벡터를 미리 정하고, 이 기준 벡터와 모든 데이터 벡터들과의 각도를 사전에 구하여 요약 정보로서 저장해 둔다. 이 각도는 식 (8)과 같이 각 벡터의 놈과 내적을 이용하여 계산할 수 있다. 식 (8)은 질의 벡터와 전혀 상관없기 때문에 질의 벡터가 주어지지 않아도 사전에 모든 데이터 벡터에 관하여 계산할 수 있다. 질의 처리 과정에서 유사 허용치 ε과 비교할 거리를 구하는 거리 근사 함수는 식 (9)와 같다.

2. 요약 정보의 구성

요약 정보 구성 단계에서는 각 멀티미디어 데이터의 n차원의 특징 벡터로부터 질의 처리 과정에서 근사에

사용될 정보를 미리 구하여 이들을 데이터베이스에 저장한다. 멀티미디어 데이터로부터 특징 벡터는 이미 추출되어 벡터 형태로 구성되어 있다고 가정한다.

먼저, 전체 데이터에 적용할 기준 벡터를 선정한다. 다음으로, n차원의 각 데이터 벡터에 대하여 놈과 기준 벡터와의 각도를 구한다. 즉, 데이터베이스에 저장되는 요약 정보는 각 데이터 벡터 당 $< X_i , \theta_{rx_i} >$로 구성되는 엔트리들의 집합과 기준 벡터이다.

3. 질의 처리

질의 처리 단계는 사용자가 제시하는 질의 벡터와 유사 허용치 ε내에 포함되는 데이터 벡터들을 찾아내는 과정이다. 이 과정에서 미리 작성된 요약 정보를 활용함으로써 효율성을 높일 수 있다. 먼저, 질의 벡터와 기준 벡터의 놈을 구하고 식 (8)을 사용하여 질의 벡터와 기준 벡터의 각도 θ_{qr}을 계산한다. 다음으로, 요약 정보를 이용하여 각 데이터 벡터와의 각도를 근사한다. 각 데이터 벡터와의 각도는 연산량이 많은 내적을 통하여 계산하는 것이 아니라, 단순히 식 (7)을 사용하여 근사한다. 다음으로, 식 (9)에서 제시한 거리 근사 함수를 이용하여 거리를 근사하고, 이것이 유사 허용치 ε보다 작을 경우 이 데이터 벡터를 후보로 선택한다. 후처리 과정에서는 후보 집합에 포함된 각 데이터 벡터를 대상으로 질의 벡터와의 실제 유클리드 거리를 구한 후, 정답 여부를 결정한다.

4. 논의 사항

본 절에서는 먼저 제안한 기법이 착오 기각을 발생시키지 않음을 보이고, 아울러 기존 기법에 비하여 착오 채택되는 벡터의 개수도 줄일 수 있음을 보인다.

정리 1:

그림 2에서 or, ox, oq가 각각 형성하는 사이 각인 α, β, θ 중 어느 두 각을 더한 값은 나머지 한 각 보다 항상 크거나 같다.

증명:

α, β, θ 중 가장 큰 각을 α라고 가정 하자. 그림 2에서 벡터 or과 벡터 ox가 이루는 2차원 평면은 유일하게 결정된다. 이 평면을 P라고 정의 할 때, 평면 P에

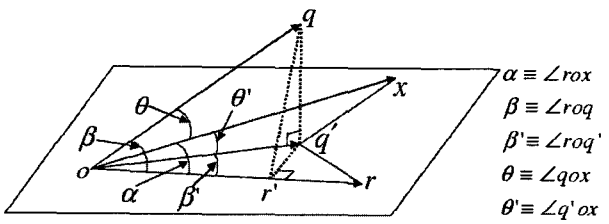


그림 2. 기준 벡터 R, 질의 벡터 Q, 그리고 데이터 벡터 X간의 관계

Fig. 2. The relationship among a reference vector R, a query vector Q, and a data vector X.

대한 q의 사상(projection)을 q'으로 놓자. 또한 q'에서 or에 내린 수선의 발을 r'이라 하자. 여기서, 삼수선의 정리에 의해 $r'q$와 or은 직교한다. 따라서 삼각형 $or'q$와 삼각형 $or'q'$은 직각 삼각형이다. 그러므로 $\cos\beta = |or'|/|oq|$, $\cos\beta' = |or'|/|oq'|$ 이다. 여기서 β'은 선분 oq'과 선분 or이 이루는 각도 한편, oq와 평면 P가 이루는 각을 δ라고 하면 $|oq'| = |oq|\cos\delta$ 이며, $|oq'| \leq |oq|$이 항상 성립한다. 여기서, 등호가 성립하는 경우는 δ가 0, 즉 q가 평면 P상에 있는 경우이다. 따라서 $\cos\beta \leq \cos\beta'$이다. β와 β'은 선분의 사이 각들이므로 0보다 크거나 같고, π보다 작거나 같다. 또한 cos 함수는 해당 범위에서 감소함수이므로 $\beta \geq \beta'$이 성립한다. 유사한 과정을 거치면 $\theta \geq \theta'$도 쉽게 보일 수 있다. 여기서 θ'은 oq'과 ox가 이루는 각도로서, 본 논문에서 추정하고자 하는 θ에 대한 추정각 $\bar{\theta}$이 된다. 한편 α는 θ'과 β'의 합이므로 $\beta+\theta \geq \beta+\theta' = \alpha$ 이 성립하고 등호는 q가 평면 P상에 있는 경우에 성립한다.

4차원 이상의 벡터 공간에서도 서로 독립인 벡터 두 개는 2차원 평면을 결정할 수 있으므로, 이 정리는 유용하다. □

따름 정리 1:

제안하는 기법에 의하여 근사한 질의 벡터와 데이터 벡터 간의 각은 실제 각보다 항상 작거나 같다. 즉, $\theta \geq |\alpha - \beta|$.

증명:

정리 1에 의해 $\theta+\beta \geq \alpha$, $\theta+\alpha \geq \beta$은 모두 성립한다. 따라서 $\theta \geq \alpha-\beta$와 $\theta \geq \beta-\alpha$도 성립하며, $\theta \geq |\alpha-\beta|$도 성립한다. □

따름 정리 2:

근사 각을 적용한 거리 근사 함수에 의해 계산된 거리는 실제 거리보다 항상 작거나 같다.

$$< X, Y > = \|X\| \|Y\| \cos\theta \leq \|X\| \|Y\| \cos\theta' \quad (10)$$

증명:

식 (10)은 두 벡터의 내적과 각도 성분의 관계를 나타낸다. 식에서 보여주는 바와 같이 코사인 함수는 1보다 작고 단조 감소 함수이기 때문에 근사 각 θ'이 실제 각 θ보다 작거나 같을 경우 실제 내적보다 항상 크거나 같다. 따라서 식 (9)와 같은 근사 각을 적용한 거리 근사 함수에 의해 계산된 거리는 실제 거리보다 항상 작거나 같다. 즉, 식 (9)는 유클리드 거리에 대한 하한 함수이다. □

따름 정리 2는 본 논문에서 제안한 근사 함수가 실제 유클리드 거리에 대한 하한 함수임을 증명한 것이다. 이는 제안하는 근사 함수를 이용하여 후보 집합을 형성할 경우 착오 기각이 발생하지 않음을 의미한다.

실제 유클리드 거리 함수 D와 제안하는 거리 근사 함수 $\overline{D}$와 Cauchy-Schwartz 부등식을 이용한 거리 근사 함수 D_{cs}와의 관계를 정리하면 식 (11)과 같다.

$$\begin{aligned} &D(X, Y) \geq \overline{D}(X, Y) \geq D_{cs}(X, Y) \\ &where, D(X, Y) = \sqrt{\|X\|^2 + \|Y\|^2 - 2\|X\| \|Y\| cos\theta} \\ &\overline{D}(X, Y) = \sqrt{\|X\|^2 + \|Y\|^2 - 2\|X\| \|Y\| \cos\overline{\theta}} \\ &D_{cs}(X, Y) = \sqrt{\|X\|^2 + \|Y\|^2 - 2\|X\| \|Y\|} \end{aligned} \quad (11)$$

따라서 제안하는 거리 근사 함수는 실제 유클리드 거리 함수에 대한 하한 함수임과 동시에 Cauchy-Schwartz 부등식을 이용한 근사 함수에 비해 실제 유클리드 거리 값에 가까운 값을 나타낸다. 따라서 착오 기각이 발생하지 않을 뿐만 아니라 Cauchy-Schwartz 부등식을 이용한 경우에 비교하여 후보의 개수를 줄일 수 있다. 최악이 경우라 할지라도 Cauchy-Schwartz 부등식을 이용한 경우와 동일한 후보 집합이 형성되고, 근사 각이 실제 각과 정확히 같아지는 최선의 경우는 정답만을 포함하는 후보 집합을 형성하게 된다.

Ⅳ. 응 용

본 논문에서 제안하는 거리 근사 기법은 멀티미디어 객체(object)가 특징 벡터 형태로 표현되고, 유사도 기준으로서 유클리드 거리를 사용하는 경우에 모두 적용이 가능하다. 예를 들어, 히스토그램 기반의 영상 검색은 원본 영상으로부터 히스토그램을 특징 벡터로 추출하고 저장한다. 또한, 요약 데이터로 기준 벡터와 더불어 각 영상에 대하여 <특징 벡터의 놈, 기준 벡터와 이루는 각도>의 쌍을 저장한다. 영상 검색에 적용되는 데이터베이스의 계층 구조를 그림 3에 나타내었다.

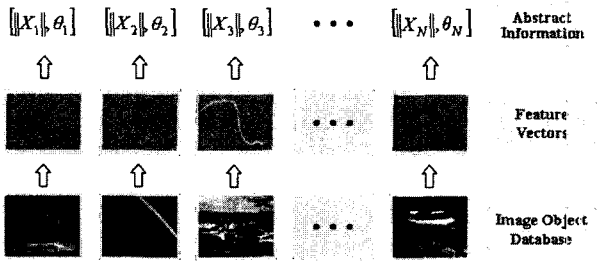


그림 3. 영상 데이터베이스에 대한 계층 구조
Fig. 3. Hierarchical structure for image database.

질의 영상이 주어지면 먼저 질의 영상으로부터 히스토그램을 특징 벡터로 추출하고, 추출된 특징 벡터와 저장되어 있는 기준 벡터간의 각을 계산한다. 다음으로 요약 정보를 이용하여 각 데이터 벡터와 질의 벡터 간의 각을 근사하고, 제안된 거리 근사 함수에 적용하여 질의 벡터와 모든 데이터 벡터 간의 근사 거리를 계산한다. 근사된 거리가 유사 허용치 이하인 벡터들을 대상으로 후보 집합을 결정한다. 마지막으로, 후처리 과정을 통하여 후보 집합으로부터 최종 정답을 결정하고, 대응되는 실제 영상을 질의 결과로 제시한다.

Ⅴ. 성능 평가

본 장에서는 제안하는 거리 근사 기법의 성능을 평가하기 위한 실험 환경과 실험 결과를 제시한다.

1. 실험 환경

본 논문에서는 성능을 평가하기 위해 합성 데이터와 실제 데이터를 사용하였다. 합성 데이터는 10,000개의 데이터로 구성되며, 데이터 벡터내의 각 차원 값으로 [0, 1]의 구간 내의 실수 값을 갖는다. Corel 영상 데이터는 총 68,040장의 영상으로 구성되어 있다[10]. 특징 벡

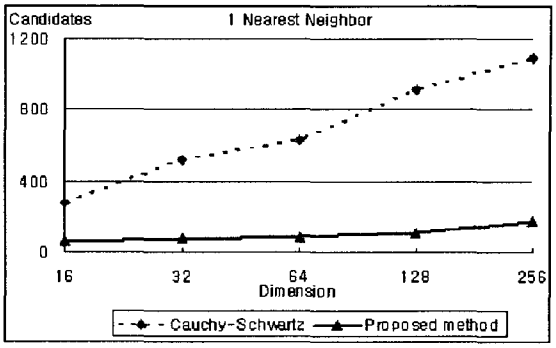
터는 수평, 수직, 그리고 두 대각선에 해당하는 4 방향의 Co-occurrence Texture 정보로써, 한 방향의 Co-occurrence Texture는 이차 각 모멘트(second angular moment), 대비(contrast), 역 차분 모멘트(inverse difference moment), 엔트로피(entropy)로 구성된다. 따라서, Corel 영상 데이터의 특징 벡터는 16차원의 벡터로 표현된다. 제안하는 기법의 성능을 Cauchy-Schwartz 부등식을 사용하는 근사 기법의 성능과 비교 평가한다. 한편, 기존 연구에서 다루었던 크기 근사 기법이나 형태 근사 기법은 착오 기각을 발생시키는 기법이다. 따라서 본 논문에서 추구하는 목표를 달성할 수 없는 기법이므로 성능 평가를 위한 비교 대상에서 제외하였다.

본 논문에서는 두 가지 평가 척도에 대하여 실험하였다. 첫째로, 전처리 과정의 효율성을 평가하기 위하여 후보 개수를 비교하였다. 이는 착오 채택이 얼마나 많이 발생하는 가에 대한 척도로서, 본 실험에서는 총 100개의 임의의 질의 벡터에 대하여 후보 개수를 각각 구하고 평균을 취하였다. 또한, 차원 수의 변화와 최종 정답 수의 변화에 따른 후보 개수의 변화를 비교하였다. 다음으로, 전처리 시간을 포함한 전체 질의 처리 시간을 비교하였다. 이는 실제 검색 성능에 대한 평가 척도가 된다. 질의 처리 시간에 대한 비교 역시 차원 수의 변화 및 최종 정답 수의 변화에 따른 변화를 비교하였다. 여기서, 전체 처리시간은 총 100개의 임의의 질의 벡터에 대한 모든 처리 시간을 취하였다. 본 실험에서는 주기억장치에 데이터 벡터를 저장하였다. 성능 평가를 위한 하드웨어 플랫폼은 2.8G Pentium Ⅳ와 512MB의 주기억장치가 장착된 PC이며, 소프트웨어 플랫폼은 MS Windows 2000 및 Visual C++6.0이다.

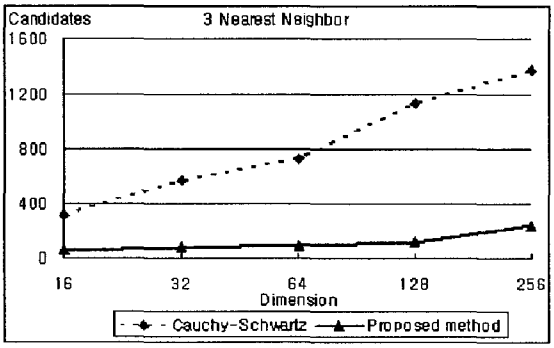
2. 실험 결과

그림 4는 k-NN 검색에 대하여 Cauchy-Schwartz 부등식을 이용한 경우와 제안하는 거리 근사 기법을 사용한 경우의 후보 개수를 비교한 결과이다.

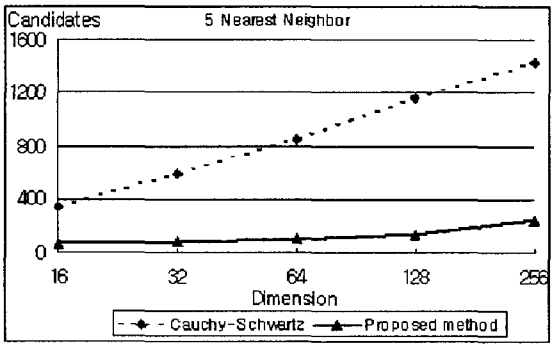
Cauchy-Schwartz 부등식을 사용한 근사 함수는 실제 거리에 대한 하한 함수이기는 하지만 질의 벡터와 데이터 벡터간의 각도 성분을 완전히 무시하므로 상대적으로 후보 개수가 많아진다. 반면에 제안하는 근사 기법은 각도 성분을 추가적으로 사용함으로써 전처리 과정 후에 선택되는 후보의 개수를 상당히 줄일 수 있었다. 16차원의 1-NN의 경우 실험 중 후보 개수가 가장 적었고, 평균적으로 총 데이터의 0.62%만 실제 거리



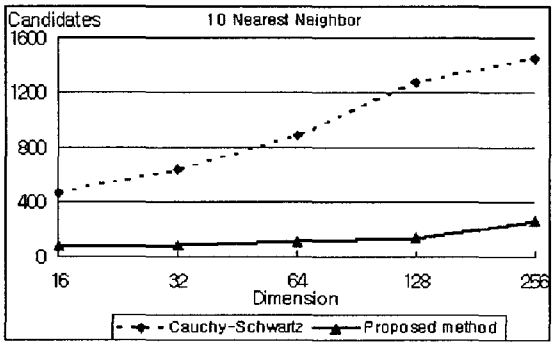
(a) 1-NN



(b) 3-NN



(c) 5-NN

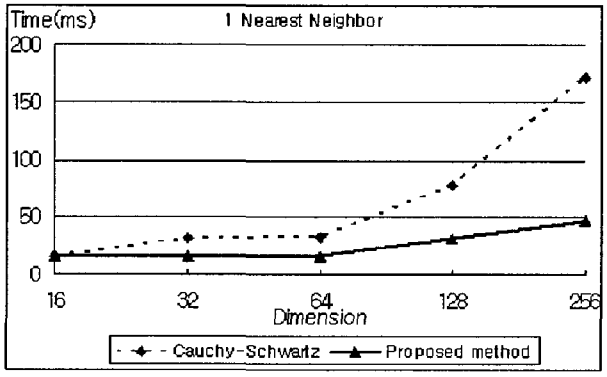


(d) 10-NN

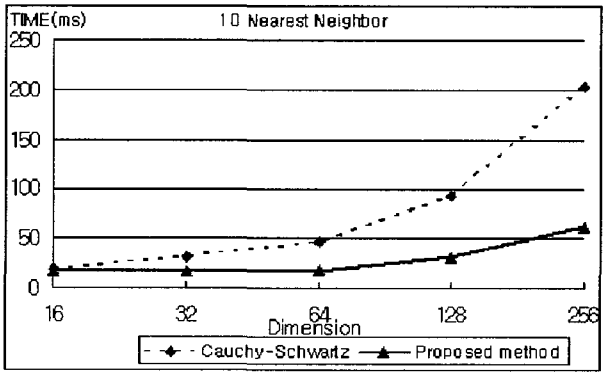
그림 4. 차원 수 변화에 따르는 후보 개수의 변화
Fig. 4. The number of candidates according to varying numbers of dimensions.

를 구하였다. 256차원의 데이터에 대한 10-NN의 경우에서도 실제 거리를 구한 데이터는 단지 평균 2.6%에 불과하였다. Cauchy-Schwartz 부등식을 사용한 근사 기법과 비교하여 제안하는 기법은 차원에 따라 상대적으로 80%에서 90%까지 후보 개수를 더 줄일 수 있으며, 약 4.5배에서 9.5배의 성능 향상 효과가 있음을 확인하였다. 뿐만 아니라, Cauchy-Schwartz 부등식을 사용하는 경우에는 차원이 증가함에 따라 후보 개수가 급격히 증가하지만 제안하는 기법은 차원이 증가하더라도 후보 개수가 비슷한 수준을 유지함을 볼 수 있다.

Cauchy-Schwartz 부등식을 사용한 근사 함수는 실제 거리에 대한 하한 함수이기는 하지만 질의 벡터와 데이터 벡터간의 각도 성분을 완전히 무시하므로 상대적으로 후보 개수가 많아진다. 반면에 제안하는 근사 기법은 각도 성분을 추가적으로 사용함으로써 전처리 과정 후에 선택되는 후보의 개수를 상당히 줄일 수 있었다. 16차원의 1-NN의 경우 실험 중 후보 개수가 가장 적었고, 평균적으로 총 데이터의 0.62%만 실제 거리를 구하였다. 256차원의 데이터에 대한 10-NN의 경우



(a) 1-NN



(b) 10-NN

그림 5. 차원 수 변화에 따른 질의 처리 시간의 변화
Fig. 5. The query processing time according to varying numbers of dimensions.

에서도 실제 거리를 구한 데이터는 단지 평균 2.6%에 불과하였다. Cauchy-Schwartz 부등식을 사용한 근사 기법과 비교하여 제안하는 기법은 차원에 따라 상대적으로 80%에서 90%까지 후보 개수를 더 줄일 수 있으며, 약 4.5배에서 9.5배의 성능 향상 효과가 있음을 확인하였다. 뿐만 아니라, Cauchy-Schwartz 부등식을 사용하는 경우에는 차원이 증가함에 따라 후보 개수가 급격히 증가하지만 제안하는 기법은 차원이 증가하더라도 후보 개수가 비슷한 수준을 유지함을 볼 수 있다.

그림 5는 특징 벡터의 차원에 따라 총 100개의 질의 벡터에 대한 질의 처리를 수행하는 시간을 보여주고 있다. 그림 5에서는 1-NN과 10-NN에 대한 실험 결과만을 보였으나, 다른 경우에도 거의 비슷한 경향을 보였다. Cauchy-Schwartz 부등식을 이용하는 경우, 차원이 증가할수록 후보들의 개수가 급격히 증가하기 때문에 후처리에 대한 부담으로 인해 전체 수행시간이 길어짐을 알 수 있다. 반면, 제안하는 기법의 경우, 각도 근사에 필요한 연산량이 상대적으로 많기 때문에 전처리 과

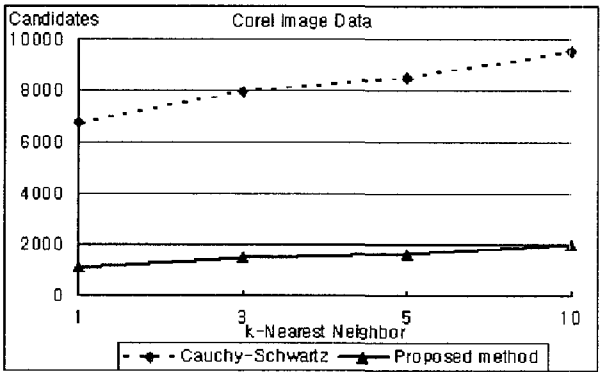


그림 6. 최종 정답 수 변화에 따른 후보 개수의 변화
Fig. 6. The number of candidates according to varying numbers of final results.

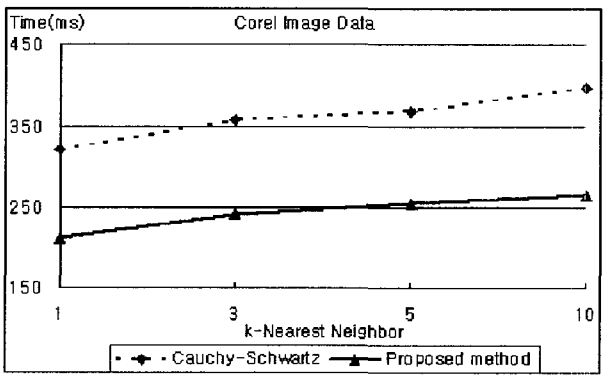


그림 7. 최종 정답 수 변화에 따른 질의 처리 시간의 변화
Fig. 7. The query processing time according to varying numbers of final results.

정만을 비교할 경우 오히려 더 많은 시간이 걸리지만, 선택되는 후보 개수를 크게 줄일 수 있으므로 후처리 과정의 처리 시간이 줄어들어 전체 처리 시간은 작아진다. 결과적으로, 차원이 증가 할수록 전체 처리 시간의 개선 효과는 커짐을 보여주고 있으며, 최대 약 3.6배까지 성능 향상 효과가 있음을 보였다.

그림 6과 7은 실제 데이터인 Corel 영상 데이터에 대하여 최종 정답 수 변화에 따른 질의 처리 시간의 변화를 나타낸 실험 결과이다. 각각 유사 허용치에 따른 후보 개수와 질의 처리 시간을 나타낸다. 실제 데이터에 대해서도 합성 데이터에서와 유사한 경향을 보인다. 후보 개수에 대한 실험에서, 1-NN의 경우 실제 거리를 구한 데이터의 개수는 평균 1.56%에 불과하고 기존 기법과 비교하여 약 4.8배에서 6.4배까지 성능 향상 효과가 있음을 볼 수 있다. 질의 처리 시간에 대한 실험 결과도 합성 데이터에서와 마찬가지로 전처리 과정은 각도 계산이 필요하기 때문에 시간이 소모되지만 실제 거리를 구해야 하는 후보 개수가 적기 때문에 전체 질의 처리 시간이 짧아졌고, 기존 기법과 비교하여 약 1.5배의 성능 향상 효과가 있음을 확인하였다. 그림 6과 7을 비교해 보면 두 실험 결과가 매우 유사한 형태를 보인다. 결과적으로, 후보 개수가 질의 처리 시간에 결정적인 영향을 줌을 알 수 있다.

VI. 결 론

멀티미디어 정보 검색에서는 유사한 정도의 기준으로 유클리드 거리를 널리 사용한다. 고차원 공간 내 질의 벡터와 데이터 벡터 간에 유클리드 거리를 계산하는 시간은 전체 검색 시간 중 가장 큰 부분을 차지한다. 따라서 이러한 유클리드 거리의 계산 시간을 줄임으로써 전체 질의 처리 시간을 크게 단축시킬 수 있다.

기존 연구 결과로서 Cauchy-Schwartz 부등식을 이용하여 고차원 공간상의 두 벡터들 간의 유클리드 거리를 근사하는 방법이 제안된 바 있다. 이 방법은 두 벡터들의 놈(norm)만을 사용하여 유클리드 거리를 근사하는 방법이다. 그러나 이 방법은 두 벡터간의 각도 성분을 무시하므로 근사 오차가 매우 커지는 문제점을 가진다. 이러한 근사 오차를 줄이기 위하여 크기 근사와 형태 근사 기법이 제안된 바 있으나, 이 두 방법은 근사 거리가 유클리드 거리의 하한을 보장하지 못하므로 질의 처리 시 착오 기각을 발생시킨다는 심각한 문제점을 가진다.

본 연구에서는 유클리드 거리를 효과적으로 근사하는 새로운 방법을 제안하였다. 제안하는 방법은 기준 벡터라 부르는 별도의 벡터를 이용하여 추정된 두 벡터 간의 각도 성분을 그들을 위한 유클리드 거리 근사에 사용한다. 이 결과, 각도 성분을 무시하는 기존의 방법과 비교하여 근사 오차를 크게 줄일 수 있다. 또한, 제안된 방법은 유클리드 거리의 하한 함수를 사용함으로써 질의 처리 시 착오 기각을 발생시키지 않음을 보장한다. 이러한 제안된 방법의 특성을 이론적으로 증명하였다. 또한, 다양한 실험에 의한 성능 평가를 통하여 제안하는 방법의 우수성을 규명하였다.

실험 결과에 의하면, 제안된 기법은 Cauchy-Schwartz 부등식을 이용한 기존의 기법과 비교하여 4.5배에서 6.5배까지의 성능 개선 효과를 가지는 것으로 나타났다. 특히, 차원이 높아질수록 또한 유사 허용치가 작을수록 제안된 방법의 성능 개선 효과가 더욱 커지는 경향을 보였다. 실제 정보 검색 환경의 특성을 고려할 때, 이러한 경향은 매우 바람직한 것이다.

참 고 문 헌

[1] R. Agrawal, C. Faloutsos, A. Swami, "Efficient Similarity Search in Sequence Database," in *Proc. of the 4th Int'l Conference on Foundations of Data Organization and Algorithms*, pp. 69-84, Oct. 1993.

[2] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is ''Nearest Neighbor'' Meaningful?," in *Proc. the 7th International Conference on Database Theory (ICDT '99)*, pp. 217-235, Jan. 1999.

[3] C. Bohm, S. Berchtold and D. A. Keim, "Searching in High-Dimensional Spaces-Index Structures for Improving the Performance of Multimedia Databases," *ACM Computing Surveys (CSUR)* Vol. 33, Issue 3, pp. 322-373, Sep. 2001.

[4] O. Egecioglu and H. Ferhatosmanoglu, "Dimensionality Reduction and Similarity Computation by Inner Product Approximations," in *Proc. the 9th ACM Interational Conference on Information and Knowledge Management*, pp. 219-226, Nov. 2000.

[5] O. Egecioglu, "Parametric Approximation Algorithms for High-dimensional Euclidean Similarity," in *Proc. of the 5-th European Conference on Principles of Data Mining and Knowledge Discovery, (PKDD 2001)*, pp. 79-90,

Sep. 2001.
[6] U. Y. Ogras and H. Ferhatosmanoglu, "Dimensionality Reduction Using Magnitude and Shape Approximations," in *Proc. the Twelfth International Conference on Information and Knowledge Management*, pp. 99-107, 2003.
[7] C. Faloutsos, R. Barber, M. Flickner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and Effective Querying By Image Content," in *Journal of Intelligent Information Systems*, Vol. 3 No.3/4 pp. 231-262, Jul. 1994.
[8] T. Seidl, and H.-P. Kriegel, "Efficient User-adaptable Similarity Search in Large Multimedia Databases," in *Proc. 23rd Int. Conf. on Very Large Databases*, pp. 506-515, Aug. 1997.
[9] R. Weber, H. J. Schek, and S. Blott, "A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces," in *Proc. 24rd International Conference on Very Large Data Bases (VLDB '98)*, pp. 194-205, Aug. 1998.
[10] http://kdd.ics.uci.edu/databases/CorelFeatures/CorelFeatures.html

저 자 소 개

정 승 도(학생회원)
1999년 한양대학교 전자·전자통신·전파공학과 학사 졸업
2001년 한양대학교 전자통신전파공학과 석사 졸업
2001년~현재 한양대학교 전자통신 컴퓨터공학과 박사과정 재학 중
<주관심분야 : 컴퓨터비젼, 생체인식, Image-based Rendering, Augmented Reality>

김 상 욱(정회원)
1989년 서울대학교 컴퓨터공학과 학사 졸업
1991년 한국과학기술원 전산학과 석사 졸업
1994년 한국과학기술원 전산학과 박사 졸업
현 한양대학교 정보통신대학 정보통신학부 부교수
<주관심분야 : 데이타베이스 시스템, 저장 시스템, 데이타 마이닝, 멀티미디어 정보 검색, 공간 데이터베이스/GIS, 주기억장치 데이타베이스, 트랜잭션 관리>

김 기 동(정회원)
1989년 서울대학교 산업공학과 학사 졸업
1991년 서울대학교 산업공학과 석사 졸업
1997년 서울대학교 산업공학과 박사 졸업
현 강원대학교 산업공학과 교수
<주관심분야 : 생산정보시스템, 스케쥴링, 인공지능>

최 병 욱(정회원)
1973년 한양대학교 전자공학과 학사 졸업
1978년 일본 경응의숙(KEIO)대학 전기공학과 석사 졸업
1981년 일본 경응의숙(KEIO)대학 전기공학과 박사 졸업
현 한양대학교 정보통신대학 정보통신학부 교수
<주관심분야 : 영상처리, 멀티미디어 공학>