위치 오차를 고려한 건물 데이터 셋의 매칭에 관한 연구

Study on Building Data Set Matching Considering Position Error

김 기 락* 허 용** 유 기 윤***
Ki Rak Kim Yong Huh Ki Yun Yu

요 약 최근 GIS 분야에서 공간 정보를 효과적으로 사용하기 위하여 다양한 원천 자료를 통합하는 것이 중요한 화두로 대두되고 있다. 일반적으로 공간 정보의 통합은 대응 공간 객체를 탐색하고 각 객체와 연동되어 있는 정보를 결합함으로써 수행된다. 하지만 어떤 공간 객체에 대응되는 다른 공간 객체를 탐색하는 것은 매우 어려운 문제로, 서로 다른 공간 객체를 탐색하기 위한 매칭 방법이 많이 연구되고 있다. 따라서 본 연구는 서로 다른 건물 데이터 셋의 통합 과정에서 좌표 변환 이후에도 잔존하는 국지적 위치 오차를 고려하여 대응 공간 객체를 탐색할 수 있는 방법을 개발하는 것을 목적으로 한다. 이러한 목적을 위해 두 지도를 좌표 변환하고 중첩 및 위치 오차가 유사한 단위 구역을 생성한 후, 위치 오차가 유사한 단위 구역 내의 건물들을 매칭하기 위하여 유사도와 ICP(iterative closest point) 알고리즘을 이용하였다. 그리고 이러한 제안된 방법의 활용 가능성을 실험을 통하여 알아보았다.

키워드 : 건물 매칭, 유사도, 위치 오차, ICP

Abstract Recently in the field of GIS(Geographic Information System), data integration from various sources has become an important topic in order to use spatial data effectively. In general, the integration of spatial data is accomplished by navigating corresponding space object and combining the information interacting with each object. But it is very difficult to navigate an object which has correspondence with one in another dataset. Many matching methods have been studied for navigating spatial object. The purpose of this paper is development of method for searching correspondent spatial object considering local position error which is remained even after coordinate transformation when two different building data sets integrated. To achieve this goal, we performed coordinate transformation and overlapped two data sets and generated blocks which have similar position error. We matched building objects within each block using similarity and ICP algorithm. Finally, we tested this method in the aspect of applicability.

Keywords : Building Matching, Similarity, Position Error, ICP

1. 서 론

정보통신분야의 급속한 발전은 전통적인 소수 전문 기관을 중심으로 구축되고 활용되었던 공간 정보를 단순한 일상생활 속에서도 활용할 수 있는 수준으로 일반화시켰다. 그 결과 다양한 분야에서 공간 정보를 구축하고 활용하기 시작하였다. 이러한 경향에 맞추어 동일한 지형지물에 대하여 구축된 공간 정보가 점차 증가함에 따라 이들 공간 정보를 통합하여 보다 풍부한 정보는 물론 기존에는 제공할 수 없었던 새로운 서비스를 제공하기 위한 연구가 수행되고 있다[3,8,14,15]. 공간 정보에서 정보의 최소 단위는 공간 객체로써 관심 지형지물을 하나의 객체로 관리한다. 따라서 일반적으로 공간 정보의 통합은 대응 공간 객체를 탐색하고 각 객체와 연동되어 있는 정보를 결합함으로써 수행된다[17].

† 본 논문은 국토해양부 첨단도시개발사업-지능형국토정보기술혁신 사업과제의 연구비지원(06국토정보B01)에 의해 수행되었음.

* 서울대학교 건설환경공학부 석사 kimki0@snu.ac.kr
** 서울대학교 건설환경공학부 박사 hy7808@snu.ac.kr
*** 서울대학교 건설환경공학부 부교수 kiyun@snu.ac.kr(교신저자)

하지만 어떤 공간 객체에 대응되는 다른 공간 정보의 객체를 탐색하는 것은 매우 어려운 문제이다. 동일한 지형지물일지라도 공간 정보의 활용목적이나 제작 과정의 기준에 따라 대응 공간 객체의 정보는 매우 상이할 수 있기 때문이다.

일반적인 매칭 기법들에서 동일한 지형지물을 표현하는 공간 객체는 두 공간 정보에서 유사한 위치에 유사한 형상으로 표현되어 있다는 가정을 이용한다. 예를 들어 건물 데이터 셋(data set)의 경우, 폴리곤 객체의 특징들을 추출하여 매칭에 활용하는데 두 공간 객체의 중첩 면적, 중심점 거리, 외곽선 비교 등을 통하여 유사도를 측정하고 가장 높은 유사도를 가지는 객체쌍을 매칭쌍으로 선택한다. 그러나 중첩 분석이나 중심점 거리와 같은 유사도 측정은 두 데이터 셋 사이의 위치 오차에 직접적인 영향을 받는다. 만약 통합할 두 공간 정보가 서로 다른 기관에서 다른 좌표 체계를 기준으로 작성되었다면 비록 좌표 변환과정을 통하여 정오차를 보정하였다 해도 국지적인 위치 오차는 여전히 잔존하게 된다. 여기서 국지적인 위치 오차란 좌표 변환 이후 중첩하였을 때, 임의의 작은 지역에서 여전히 존재하고 있는 오차를 의미한다. 이런 문제를 해결하기 위해서는 지상 기준점들을 이용한 국지적 좌표 변환을 전처리 과정에서 수행함으로써 대응 공간 객체들의 위치를 어느 정도 일치시켜야 한다[3,12]. 하지만 각 객체간의 국지적 위치 오차는 우연 오차의 성격을 가지므로 소수 지상 기준점만으로는 일치시키기 힘들며, 따라서 데이터 셋 집합에 대한 매칭 방법이 필요하다.

폴리곤 객체 매칭과 관련하여 건물 매칭에 대한 연구들을 살펴보면, 다음과 같다. Dunkars(2003)는 서로 축척이 다른 지도를 선택하여 실험하였으며, 건물 객체를 대축척에서는 폴리곤으로, 소축척에서는 하나의 상징(symbol)으로 가정하였다[5]. 이때, 사용할 수 있는 정보를 위치와 각도만으로 한정하여 그 유사성을 비교하였다.

Gösseln과 Sester(2004)는 폴리곤 객체 매칭에서 크게 네 가지 기준을 사용하였으며, Hausdorff 거리, symmetric 차이, compactness 차이, 각도 각각에 대하여 비교하였다[6]. Hausdorff 거리는 두 유한 집합 내의 모든 정점들 간의 비일치도를 측정하는 방법으로 일반적인 거리 함수를 이용한 최대최소(minmax) 기법을 이용한 거리이다. Symmetric 차이는 두 폴리곤의 면적이 얼마나 중첩되어 있는가에 따른 차이이고, compactness 차이는 폴리곤의 면적을 둘레로 나눈 비율의 차이이며, 각도는 폴리곤의 장축의 방향성을 의미한다. 각각의 매칭 기준에 대하여 0과 1사이로 결과가 나온 것들을 평균하여 그 값들을 비교, 차이가 가장 적은 쌍을 매칭쌍으로 선택하였다.

Wenjing 외 3인(2008)은 폴리곤 형상 정보에 대한 유사도를 조합하여 매칭쌍을 결정하였다[13]. 크게 세 가지를 비교하였는데, 거리, 형상, 면적을 기준으로 하였으며, 모든 유사도는 0과 1사이로 결과가 나오게 하였다. 우선 폴리곤 중심점을 이용하여 위치 유사도를 측정하였으며, 중심점과 외곽선을 이루는 점들 간의 거리차를 이용하여 형태 유사도를 측정하였다. 마지막으로 면적 유사도를 비교하였는데, 단순한 면적차를 이용하여 결정하였다. 매칭쌍은 각각의 유사도에 대해 가중 평균을 하여 가장 유사도가 높게 나오는 것을 선택하였다. 이 때, 가중치는 전문가가 선택하거나 training site에서 결정한다.

Huang 외 4인(2010)은 폴리곤 매칭을 위해 중첩 면적, 중심점 거리, 형상비(shape ratio), 방향성 등 네 가지 매칭 기준을 종합적으로 검토하여 매칭쌍을 선택하였다[7].

선행 연구들의 거리나 중첩을 기반으로 하는 유사도를 측정하는 경우, 초기 건물 객체의 위치에 크게 영향을 받는다. 따라서 유사도를 측정하기에 앞서 전처리로 좌표 변환을 하거나, 혹은 좌표 변환 후에도 존재하는 기하학적 편차(geometric deviation)가 있을 경우 이를 감소시키기 위하여 rubber sheeting이나 affine transformation을 수행한다[11,12]. 그러나 이러한 좌표 변환은 기준점을 선택하여 수동으로 수행해야 할 뿐 아니라, 미리 위치 오차에 대한 정보를 알고 있어야 한다. 또한 좌표 변환 이후에도 여전히 존재하는 국지적 위치 오차를 가지고 있는 상태에서 건물 데이터 셋의 유사도를 측정할 경우, 건물의 잘못된 위치 정보에 의해 유사도가 왜곡이 될 수 있다. 따라서 본 연구에서는 건물 데이터 셋의 건물 객체가 아닌 건물 집합 즉, 위치 오차가 유사한 단위 구역을 이용하여 이러한 구역 단위로 건물 매칭을 수행함으로써 국지적인 위치 오차에 강건한 유사도 기반의 건물 매칭 방법을 제안하고자 한다.

2. 매칭쌍 결정 기법

본 연구의 프로세스는 그림 1과 같이 우선, 좌표 변환된 두 지도를 중첩한 뒤 위치 오차가 유사한 단위 구역을 설정하고 유사도와 거리 오차에 대한 임계값을 설정하는 전처리 단계와 반복 과정을 통해 유사도를 이용한 매칭쌍 결정과 매칭쌍을 이용한 변환 함수를 결정하는 매칭쌍 결정 단계로 나누어진다.

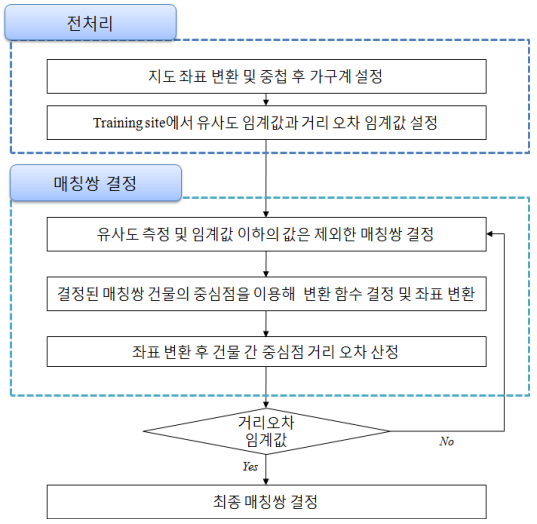

그림 1. 연구 흐름도

2.1 전처리 단계

실험 데이터는 내비게이션 지도와 도로명주소 지도로, 서로 좌표계가 다르므로 좌표 변환을 통하여 좌표를 일치시킨 뒤 중첩을 한다. 이 경우 국지적 위치 오차가 존재하는 지역이 발생하며, 이들 국지적 오차를 고려하기 위하여 도로명주소 지도의 법정동 경계 레이어와 도로 레이어를 이용하여 국지적 오차가 발생하는 최소 영역, 즉 위치 오차가 유사한 단위 구역을 설정한다[1,9].

위치 오차가 유사한 단위 구역은 크게 두 가지 목적을 위하여 설정되는데, 첫 번째로 위치 오차를 발견하고 고려하기 용이하게 한다. 사용되는 두 지도를 중첩하여 좌표계를 일치시킬 경우 위치 오차가 존재하는 경우와 존재하지 않는 경우가 함께 존재한다. 법정동 경계 내의 경우를 보면 이러한 위치 오차는 그 크기와 방향성이 국지적으로 나타나고 있어서 동시에 처리하기가 곤란하다. 이러한 점에서 위치 오차가 유사한 단위 구역을 설정하면 그 내부의 위치 오차만 고려하므로 처리하기가 편리하게 된다. 두 번째로 평균 거리 오차의 최소값 감소를 들 수 있다. 실험 과정은 평균 거리 오차의 최소값이 특정 임계값을 만족시킬 경우 종료하게 되는데, 이러한 오차의 최소값을 감소시킴으로서 좀 더 정확한 매칭이 가능해 진다.

위치 오차가 유사한 단위 구역을 결정하는 방법은 여러 가지가 있으나 본 연구에서는 레이어(layer)를 이용하여 폴리곤을 생성, 위치 오차가 유사한 단위 구역으로 정의하였다. 사용된 레이어는 도로명주소의 도로 레이어와 법정동 경계 레이어로서, 위치 오차가 유사한 단위 구역은 이 두 가지를 중첩시켜 만든 폴리곤을 의미한다. 내비게이션 지도의 도로 레이어가 훨씬 더 세밀하게 표현되어 있으나 실제 위치 오차가 유사한 단위 구역을 생성시킬 경우 기준선 초과 오류(overshooting), 기준점 미달 오류(undershooting), 중복부(sliver) 등의 오차가 발생하므로 사용하지 않았다.

위치 오차가 유사한 단위 구역 내의 건물 객체에 대한 매칭은 유사도를 기반으로 하며, 이때 유사도는 거리, 중첩 면적, 방향각, 형상비 등을 기준으로 측정한다. 다음으로 제안된 방법에서는 두 가지의 임계값이 필요하며, 초기 매칭쌍을 결정하기 위한 유사도 임계값과 실험 과정의 종료를 위한 거리 오차 임계값을 측정하여야 한다. 이 임계값들은 training site에서 미리 결정되는데, 건물들의 유사도 분포 및 거리 오차 분포를 통해 그 값이 결정된다.

2.2 매칭쌍 결정 단계

2.2.1 유사도 측정

건물 폴리곤 객체의 유사도를 측정하는 방법은 여러 가지가 있으며, 이를 조합하는 방법 역시 여러 가지가 있다.

선행연구에 제안된 유사도 중에서 Huang외 4인(2010)의 건물 객체 매칭 방법과 제안된 방법의 결과를 비교하기 위하여 본 연구에서는 이들이 제안한 중첩 면적, 중심점 거리, 형상비, 방향성 등 네 가지 유사도를 이용한다[7]. 측정된 유사도 각각의 조합을 위하여 모든 정보를 0과 1 사이로 정규화 시켰으며, 가중 평균을 구하여 종합 유사도를 측정하였다.

두 폴리곤 집합 A={a_i | 1≦ i ≦m}, B={b_j | 1≦ j ≦n}를 가정하고, 본 연구에서 사용된 유사도를 정의하면 다음과 같다.

첫째, 중심점 거리를 이용한다. 폴리곤 a_i의 중심이 (x_1, y_1)이고, 폴리곤 b_j의 중심이 (x_2, y_2) 이라면 중심점 간 거리 $dist(a_i, b_j)$는 식 (1)과 같이 정의된다.

$$dist(a_i, b_j) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

둘째, 중첩 면적을 이용한다. 본 연구에서 사용하는 중첩도 $overlaparea(a_i, b_j)$는 식 (2)와 같이 정의된다. 여기서, $area(a_i, b_j)$는 폴리곤 a_i와 폴리곤 bj의 중첩 면적이고 $area(a_i)$는 폴리곤 a_i의 면적이다.

$$overlaparea(a_i, b_j) = \frac{area(a_i, b_j)}{area(a_i)} \quad (2)$$

셋째, 형상비(S)를 이용한다. 형상비의 정의는 식 (3)과 같으며 형상비의 차이 $sr_{diff}(a_i, b_j)$는 식 (4)와 같다. 여기서, $area(a_{i_{MBR}})$는 폴리곤 a_i의 MBR (Minimum Boundary Rectangle)의 면적이다.

$$S(a_i) = \frac{area(a_i)}{area(a_{i_{MBR}})} \quad (3)$$

$$sr_{diff}(a_i, b_j) = |S(b_j) - S(a_i)| \quad (4)$$

넷째, 방향성의 차이를 이용한다. 방향성의 차이 $drct_{diff}(a_i, b_j)$는 식 (5)와 같다. 본 실험에서 방향성은 폴리곤의 가장 긴축을 주축으로 고려하여 이 축의 방향성을 폴리곤의 방향성으로 가정한다. 여기서, $drct(a_i)$, $drct(b_j)$는 각각 폴리곤 ai, bj의 방향성이다.

$$drct_{diff}(a_i, b_j) = |drct(b_j) - drct(a_i)| \quad (5)$$

정의된 유사도는 각 객체쌍 간의 차이값을 계산하고, 각각의 유사도를 계산된 차이값의 최대값으로 나누어 정규화를 수행한다. 정규화된 유사도는 식 (6)과 같은 방법으로 조합된다. 여기서, a, b, c, d는 유사도의 각각의 값을 의미하며 w_1, w_2, w_3, w_4는 가중치를 의미한다. 단, 중첩 면적의 경우 역수를 취하였다.

$$similarity = 1 - (w_1 * a + w_2 * b + w_3 * c + w_4 * d) \quad (6)$$

가중치의 경우 각각의 유사도 기준에 대한 중요도를 따져, training site에서 실험적으로 결정하였다. 가중치는 일반적으로 미리 결정하는데, training에 의하거나 전문가에 의해 획득된다[11,13].

2.2.2 반복 과정을 통한 매칭쌍 결정

위치 오차가 유사한 단위 구역 내에서 두 데이터 셋 집합의 전체적인 위치를 고려하기 위하여 변환 함수를 이용하며, 각각의 공간 객체에 대한 매칭쌍을 유사도로 측정한다. 유사도만을 이용하여 매칭쌍을 결정할 경우, 데이터 셋 간의 위치가 정확히 일치하지 않으면 잘못된 매칭쌍을 선택할 확률이 높다. 따라서 전체 데이터 셋의 위치를 고려하여야 하며, 이에 대한 방법으로 ICP(iterative closest point) 알고리즘[2]을 이용한다. ICP라는 용어는 Besl(1992)에 의해 발표된 논문에서 처음 사용되었으며 기본 과정은 다음과 같다. 입력되는 두 데이터 간의 가장 가까운 거리에 있는 점 쌍을 찾고, 찾아진 점 쌍 간의 거리를 최소화 시키는 변환 매개변수(transformation parameter)를 찾는다. 이러한 두 과정이 반복되며 매칭쌍을 찾아 좌표 변환을 한다. 실험은 반복적으로 수행되며 좌표 변환된 두 데이터 간의 평균 거리 오차가 설정한 임계값 이하가 될 때 종료된다.

본 연구의 실험 과정은 크게 다음 두 가지로 나누어진다. 첫째가 매칭쌍 결정 과정이고, 두 번째가 결정된 매칭쌍을 적용하여 변환 함수를 결정하는 과정이다. 매칭쌍은 유사도를 측정하여 결정하고, 변환 함수는 매칭쌍이 된 건물들의 중심점 좌표를 이용하여 평균 거리 오차가 최소가 되도록 결정한다.

위치 오차가 유사한 단위 구역 내의 건물 데이터 셋 집합 A에 대해 집합 B의 각각에 대한 매칭쌍 결정 방법은 그림 2와 같다. 여기서 $Th1$은 training site에서 결정된 유사도의 임계값을 의미하며,

```
데이터 셋 집합 A={ai | 1≦ i ≦m}와 B={bj | 1≦ j ≦ n}를 비교할 때,
for i = 1 : m
  for j = 1 : n
    - ai에 대해 모든 bj의 유사도를 각각 측정
    - ai에 대해 최대의 유사도를 가지는 bj를 선택
    - 각각의 매칭쌍을 결정
  end
end
if similarity(ai, bj) < Th1
    - 매칭쌍 집합에서 제외
else
    - 매칭쌍 집합에 포함
end
```

그림 2. 매칭쌍 결정 의사코드

similarity(a_i, b_j)는 매칭쌍으로 결정된 (a_i, b_j)의 유사도를 의미한다.

변환 함수의 결정을 위해 사용되는 매칭쌍들은 매칭쌍 집합에 포함된 값들이 이용된다. 따라서 결정된 매칭쌍 중에 일부만이 변환 함수 결정에 이용하게 된다. 이렇게 결정된 매칭쌍 건물들의 중심점 좌표를 이용하여 변환 함수를 결정하며, 결정된 변환 함수를 적용하여 국지적으로 좌표 변환을 반복적으로 수행한다. 각 매칭쌍 집합에 포함된 건물의 중심점 좌표는 점집합 {m_i}와 {n_i}가 되어 변환 함수 결정에 이용된다. 실험은 매칭되는 점 간의 평균 거리 오차가 임계값을 만족할 때 까지 반복하여 수행하며 그 방법은 그림 3과 같다. 여기서 *Th2*는 training site에서 결정된 거리 오차 임계값을 의미한다.

```
for k = 0 : 최대반복횟수
    - 초기값으로 매칭쌍으로 결정된 (ai, bj)를 이용
    - 매칭쌍의 중심점 좌표를 이용하여 오차를 최소
      화하는 회전행렬과 이동행렬 산정
    - 산정된 회전행렬과 이동행렬을 이용하여 좌표
      변환
    - 평균 거리 오차가 Th2보다 크면 매칭쌍의 재결
      정을 통해 변환 함수를 다시 결정
    - 평균 거리 오차가 Th2보다 작으면 반복을 종료
end
```

그림 3. 변환 함수 결정 의사코드

3. 실험 및 결과

3.1 실험 내용

본 연구의 실험 데이터는 내비게이션 지도와 도로명주소 지도의 건물 데이터 셋을 활용하였으며, 테스트는 ArcGIS와 MATLAB을 이용하여 수행하였다.

두 지도의 좌표계는 서로 다른 기준을 가지고 원점 및 단위가 다르기 때문에 투영 변환을 하여 도로명주소 좌표계로 일치시킨 후 중첩시켰다. 이 과정 후 결과를 확인해 보면 그림 4와 같이 좌표 변환 후 일부 지역에서 국지적 오차가 발생하는 것을 볼 수 있다. 이러한 형태는 전 지도에 걸쳐 일어나는 현상으로 특정한 형태를 가지고 있지 않아 처리에 어려움이 존재한다.

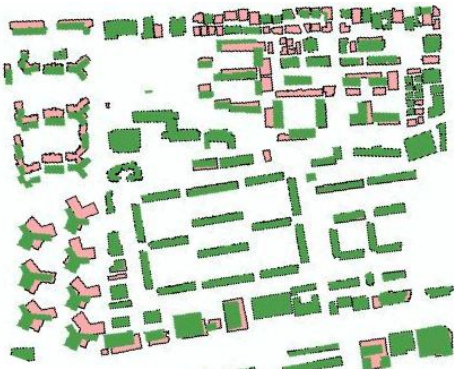

그림 4. 좌표 변환 후 중첩된 모습

본 연구에서 발생하는 위치 오차는 정오차와는 다르게 국지적으로 발생한다. 이러한 오차를 가능한 일정하게 만들고 크기를 감소시키기 위하여 실험 지역을 분할하였다. 따라서 매칭이 일어나는 전체 지역을 분할하였으며 이를 위해 매칭을 한정하는 임의의 지역, 즉 위치 오차가 유사한 단위 구역을 설정하였다.

매칭쌍을 결정하기 위해서, 객체간의 유사성을 판단할 기준이 필요하며, training site에서 기준으로 이용할 임계값을 설정하였다. 본 연구에서는 위치 오차에 관계없는 임의의 지역을 training site로 선정하여 실험하였다. 따라서 training site 내에서 위치 오차가 존재하는 경우와 그렇지 않은 경우가 동시에 존재한다. 각각의 건물들에 대해 유사도를 측정해 가장 큰 값을 가지는 건물쌍을 매칭쌍으로 선정하였다. Training site는 그림 5와 같은 지역을 선택하였다.

그림 5. 유사도의 임계값을 결정하기 위한 training site

조합된 유사도 값의 분포는 그림 6과 같으며 대표적인 통계값들은 표 1에 나온 값들과 같다. 그림

6에서 보면 유사도의 분포는 정규 분포를 이루지 않고, 두 부분에서 높은 값을 보이고 있다. 이러한 현상이 발생하는 이유는 언급하였듯이 training site 내에 위치 오차가 존재하는 지역 뿐 아니라 존재하지 않는 지역 역시 포함되어 있기 때문이다. 이러한 분포를 고려하여 임계값을 결정하며, 향후 본 실험에서 매칭쌍 결정에 이용하고자 한다[10]. 임계값이 커지면 유사도에 의해 결정되는 매칭쌍의 수가 적어져서 정확률은 높아지게 되나, 매칭쌍 중의 올바른 매칭쌍의 수 역시 적어져서 재현율이 낮아지게 된다. 반면, 임계치의 값이 작아지면, 반대로 정확률은 낮아지나 재현율은 높아진다. 이러한 정확률과 재현율의 trade-off 관계를 고려하여 본 연구에서는 가능한 많은 매칭쌍을 후보군에 넣기 위해서 하사분위수값(lower hinge)인 0.79를 임계값으로 설정하였다.

거리 오차에 대한 임계값 역시 같은 방식으로 설정하였다. 매칭쌍으로 결정된 두 데이터 셋의 건물들의 중심점 간 거리를 계산하여 거리 오차 분포를 산정하였다. 그리고 계산된 값들의 분포에서 하사분위수값을 역시 임계값으로 설정하였다.

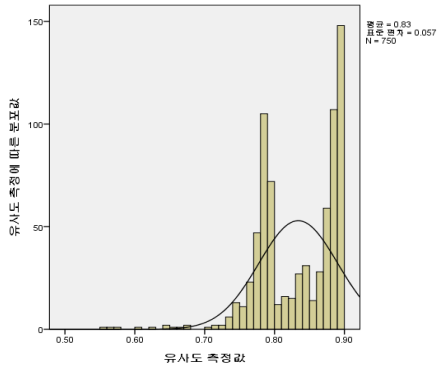

그림 6. 유사도 분포

표 1. 유사도 분포값

	평균	표준 편차	중간 값	lower hinge	upper hinge
유사도	0.83	0.06	0.84	0.79	0.89

3.2 실험 결과

우선 전국을 위치 오차가 유사한 단위 구역으로 나누었다. 이 구역 중 몇몇 부분들을 제안된 방법으로 매칭을 수행하였고, 그 결과는 온톨로지 매칭에서 성능을 평가하는 수단인 정확률과 재현율, F-측정값을 사용하여 확인하였다.

정확률과 재현율은 온톨로지(ontology) 매칭 방법론의 성능을 평가하는 수단 중의 하나로 F-측정값과 함께 사용되고 있다. 정확률은 매칭을 얼마나 정확하게 했는지를 알아보는 요소이고, 재현율은 매칭을 수행하였을 때 얼마나 많은 결과가 나타나는지를 알아보는 요소이다[16]. F-측정값은 일반적으로 정확률과 재현율을 같은 가중치를 두어 통합한 값으로 이 값의 크기가 클수록 더 좋은 결과를 나타낸다.

정확률과 재현율을 정의하기 위하여 Do와 Rahm (2002)은 그림 7과 같이 실제 발견된 매칭쌍과 실험에 의해 도출된 매칭쌍을 이용하여 식 (7)과 (8)로 도출하였다[4].

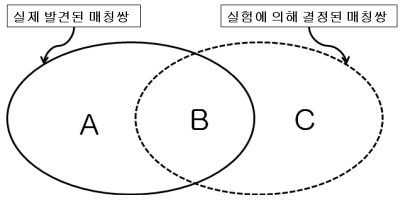

그림 7. 정확률과 재현율

$$\text{정확률} = \frac{B}{B+C} \tag{7}$$

$$\text{재현율} = \frac{B}{A+B} \tag{8}$$

따라서 본 연구에서 정확률과 재현율 및 F-측정값은 다음 식 (9), (10), (11)과 같이 정의한다.

$$\text{정확률} = \frac{\text{올바른 매칭쌍 개수}}{\text{실험에서 결정된 매칭쌍 개수}} \tag{9}$$

$$\text{재현율} = \frac{\text{올바른 매칭쌍 개수}}{\text{실제 발견된 매칭쌍 개수}} \tag{10}$$

$$F = \frac{2\times(\text{정확률})\times(\text{재현율})}{(\text{정확률})+(\text{재현율})} \tag{11}$$

정확률과 재현율을 산정하기 위하여 올바른 매칭쌍의 정의가 필요한데, 일반적으로는 단순 중첩이 되는 데이터 셋을 의미한다[11]. 그러나 위치 이동이 있을 경우 이러한 정의는 문제가 되므로 전체적인 건물의 위치들을 고려하여 정확한 이동이 되었

을 때의 알맞은 쌍을 올바른 매칭쌍으로 선택한다.

실험은 그림 8과 같이 위치 오차가 유사한 단위 구역 내에서 수행되었다. 각각의 실험 자료에 대하여, Huang외 4인(2010)이 제안한 매칭 방법(선행연구)과 본 연구에서 제안된 방법의 결과를 비교하여, 본 제안된 방법의 성능을 평가해 보았다(표 2)[7].

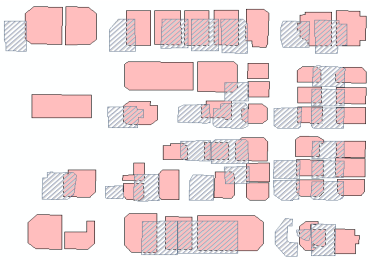

그림 8. 위치 오차가 존재하는 예(case 1)

표 2. 매칭쌍 결과에 대한 비교

	정확률		재현율		F-측정값	
	선행 연구	제안된 방법	선행 연구	제안된 방법	선행 연구	제안된 방법
case1	$\frac{11}{26}$	$\frac{18}{20}$	$\frac{11}{32}$	$\frac{18}{32}$	0.38	0.69
case2	$\frac{6}{16}$	$\frac{15}{16}$	$\frac{6}{26}$	$\frac{15}{26}$	0.29	0.71
case3	$\frac{6}{11}$	$\frac{4}{4}$	$\frac{6}{13}$	$\frac{4}{13}$	0.5	0.47
case4	$\frac{4}{8}$	$\frac{8}{8}$	$\frac{4}{10}$	$\frac{8}{10}$	0.44	0.89
case5	$\frac{6}{10}$	$\frac{7}{7}$	$\frac{6}{15}$	$\frac{7}{15}$	0.48	0.61
case6	$\frac{3}{8}$	$\frac{4}{4}$	$\frac{3}{12}$	$\frac{4}{12}$	0.3	0.5

표 2를 보면 제안된 방법에 의하여 매칭쌍을 결정하였을 경우 case 3을 제외한 나머지 실험 자료에서 보다 정확한 매칭쌍이 선택되었다. 좀 더 상세한 결과의 평가를 위하여 정확률과 재현율 및 F-측정값에 대하여 분석하였다.

우선 정확률만을 고려하면 본 연구에서 제안한 방법이 더 좋은 결과를 보인다는 것을 알 수 있다. 먼저 정확률의 정의에서 좋은 결과가 나왔다는 것은 올바른 매칭쌍이 더 많이 찾아졌거나, 혹은 결정된 매칭쌍이 더 적게 찾아 졌다는 것을 의미한다. 각각의 경우에 대해 매칭쌍을 정확히 알아보면 표 3과 같다. 정매칭은 선택된 매칭쌍 중 올바른 매칭쌍을 의미하고 오매칭은 선택된 매칭쌍 중 잘못된 매칭쌍을 의미한다.

표 3. 방법에 따라 결정된 매칭쌍의 분류

	선행 연구		제안된 방법	
	정매칭	오매칭	정매칭	오매칭
case1	11	15	18	2
case2	6	10	15	1
case3	6	5	4	0
case4	4	4	8	0
case5	6	4	7	0
case6	3	5	4	0

위의 표 3에서 알 수 있듯이 선행 연구에 의해 선택된 매칭쌍은 오매칭이 된 경우가 많은 것을 볼 수 있다. 이것은 중심점 간의 거리와 중첩 면적에 대한 유사도가 초기 건물들 간의 거리에 크게 영향을 받는데, 이를 고려하지 않고 바로 적용을 하였기 때문이다. 따라서 좌표를 일치시킨 후 중첩을 하였을 경우, 위치 오차가 존재한다면 선행 연구에 의한 방법은 오매칭이 많이 일어날 수밖에 없어서 정확률이 떨어지게 된다.

이에 반해 제안된 방법에 의한 매칭을 하였을 경우, 오매칭은 거의 일어나지 않는 것을 알 수 있다. 따라서 정확률이 좋아지는데, 특이한 점은 정매칭쌍의 개수가 반드시 늘어나지는 않았다는 것이다. Case3에서 보면 실제로는 올바른 매칭쌍의 개수가 줄어들었다. 그러나 정확률은 높은데 이는 결정된 매칭쌍이 4개로 선행 연구에서 찾은 매칭쌍보다 훨씬 더 적은데 기인한다. 이러한 결과가 나오는 이유는 임계값을 너무 높게 설정하였기 때문이라고 판단된다. 실험 과정에서 매칭 후보군을 선택하는데 임계값이 너무 높아서 가능한 후보군을 줄이기 때문이다. 또한 내부적으로 거리 오차의 값이 최소가 될 때 실험이 종료되는데, 그 과정에서 거리 오차를 줄이기 위하여 매칭쌍을 삭제하는 것도 이유가 될 것이다.

다음으로 재현율의 경우 이미 정해진 전체 매칭쌍의 개수에 대해 올바른 매칭쌍의 개수를 비교하므로 올바른 매칭쌍만이 영향을 미친다. 따라서 표 3에 나타나있듯이 제안된 방법에 의해 올바른 매칭쌍이 더 많이 결정되었으므로 대부분의 case에서 재현율이 높다는 것을 알 수 있다. 이는 그림 9를

통해 확인할 수 있다.

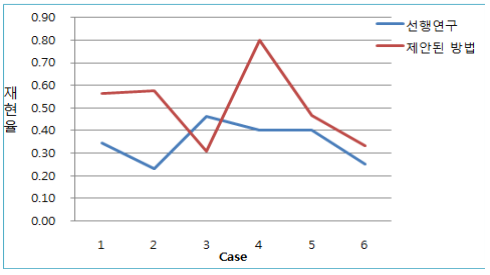

그림 9. 각 case에 따른 재현율 변화

Case3의 경우 선택된 올바른 매칭쌍이 선행 연구에 의한 것보다 적어 재현율이 낮게 나왔다. 이는 앞서 언급하였듯이 임계값에 의한 결과이므로 임계값을 더 낮게 조정한다면 더 좋은 결과를 얻을 수 있으리라 기대된다. 이 외의 경우에는 대부분 제안된 방법에 의한 재현율이 훨씬 높게 나온 것을 알 수 있다.

마지막으로 F-측정값을 살펴보면, 그 크기가 클수록 매칭이 잘 되었다는 것을 의미하며, 대부분의 case에서 제안된 방법의 F-측정값이 크게 나온 것을 알 수 있다. 따라서 제안된 방법이 보다 좋은 매칭 결과가 나왔다고 볼 수 있으며, case3의 경우에도 약간 작은 값이 나오나 크게 차이가 나지는 않는다. 전체적인 F-측정값의 경향을 확인해 보면 그림 10과 같다.

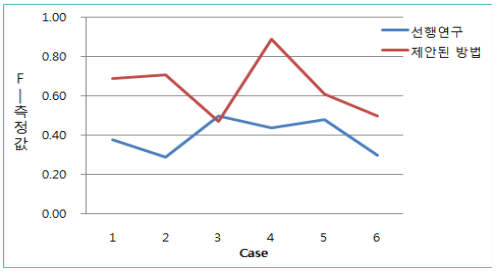

그림 10. 각 case에 따른 F-측정값 변화

위의 결과에서 보면, 제안된 방법에 의한 결과가 정확률 뿐 아니라, 재현율과 F-측정값이 선행연구와 비교 시 더 높게 나왔으며, 이는 국지적인 위치 오차가 존재할 경우 제안된 위치 오차가 유사한 단위 구역 단위의 건물 객체 매칭이 보다 정확한 매칭쌍이 탐색된다고 판단할 수 있다.

4. 결 론

기존 유사도 측정 알고리즘은 위치 오차가 존재하는 경우 유사도에 왜곡이 발생하여 잘못된 매칭쌍을 선택할 확률이 높아진다. 따라서 본 연구에서는 국지적인 위치 오차를 조정하며 유사도를 측정하는 방법을 제안하였다. 이를 위해 우선 위치 오차가 유사한 단위 구역을 설정한 뒤, 위치 오차가 유사한 단위 구역 내에 위치 오차가 있는 지역의 건물들을 실험 데이터 셋으로 추출하여 각 건물들에 대해 기하학적 유사도를 측정하였으며 오차에 대한 영향을 가능한 감소시키기 위하여 ICP 알고리즘을 이용하였다. 제안된 방법은 크게 매칭쌍 결정 과정과 변환 함수 결정 과정으로 나누어지는데, 초기 매칭쌍은 선행 연구에서 제안한 유사도를 측정하여 결정하였다. 결정 과정에서 특정 임계값 이하의 유사도를 가지는 매칭쌍은 제외하였으며, 이 임계값은 training site에서 결정하였다. 임계값 이상의 초기 매칭쌍을 적용하여 변환 함수를 결정하였으며, 매칭된 건물들의 중심점 간 평균 거리 오차가 training site에서 설정한 거리 오차 임계값에 도달할 때 실험이 종료되게 하였다. 알고리즘 내에서 반복적으로 건물들의 위치가 조정되면서 유사도 역시 변화되어 가며 매칭쌍을 재결정하는 과정을 거쳤다. 유사도만을 이용하여 매칭쌍을 결정하는 방법과 제안된 방법에 의해 매칭쌍을 결정하는 방법의 결과를 비교하였다. 정확률의 경우 제안된 방법에 의한 결과에서 대부분이 1(100%)이 나왔으며, 선행 연구에 의한 결과보다 높거나 같은 값을 보였다. 재현율의 경우도 case 3을 제외하고 모두 제안된 방법이 높게 나타났다. 이러한 선행 연구의 결과보다 낮게 나온 결과는 임계값에 의한 영향으로 볼 수 있는데 높은 임계값에 의한 결과로 판단된다. F-측정값은 재현율과 비슷한 양상을 보였으며, 대부분의 경우에서 제안된 방법이 선행 연구보다 높게 나왔다.

본 연구에서 사용된 유사도 임계값은 training site에서 결정된 값이다. 일반적으로 아주 작은 유사도를 가지는 경우가 아닌 이상 대부분의 경우 포함해야 하므로, 가능한 유사도 임계값을 작게 주어야 한다. 그러나 유사도 임계값이 작게 되면 정확률도 낮아질 수 있다. 따라서 향후 유사도 임계값에 따른 정확률과 재현율의 관계를 정량적으로 분석할 필요가 있을 것이다.

참 고 문 헌

[1] G. Achilleos, 2006, “Propagation of uncertainty within the affine transformation application on Contours.” International Journal of Geographical Information Science, Vol. 12, pp. 79-91.

[2] P. J. Besl, and N. D. McKay, 1992, “A method for registration of 3-D shapes,” Pattern analysis and machine intelligence, Vol. 14, No. 2, pp. 239-256.

[3] M. Butenuth, G. V. Gösseln, M. Tiedge, C. Heipke, U. Lipeck, and M. Sester, 2007, “Integration of heterogeneous gespatial data in a federated data-base,” ISPRS Journal of Photogrammetry & Remote Sensing, pp. 328 - 346.

[4] H. H. Do, S. Melnik and E. Rahm, 2002, “Comparison of schema matching evaluations,” In Proceedings of the workshop on Web and Databases.

[5] M. Dunkars, 2003, “Matching of Datasets,” Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science, pp. 67-78.

[6] G. V. Gösseln and M. Sester, 2004, “Integration of geoscientific data sets and the German digital map using a matching approach,” International Archives of Photogra- mmetry and Remote Sensing , Vol. 35, pp. 1249 - 1254.

[7] L. Huang, S. Wang, Y. Ye, B. Wang and L. Wu, 2010, “Feature matching in cadastral map integration with a case study in Beijing,” Geoinformatics, pp. 1-4.

[8] H. Mohammadi, 2008, “The Integration of multi-source spatial datasets in the context of SDI Initiatives,” the University of Melbourne in fulfillment of the degree of Doctor of Philosophy.

[9] İ. Özkal-Sanver, 2005, “Stability and Efficiency of partitions in matching problems,” Theory and Decision, Vol. 59, pp. 193-205.

[10] C. Reimann, P. Filzmoser and R. G. Garret, 2004, “Background and threshold: critical comparison of methods of determination,” The Science of the Total Environment, pp. 1-16.

[11] A. Samal, S. Seth and K.. Cueto, 2004, “A feature-based approach to conflation of geospatial sources,” International Journal of Geographical Information Science, Vol. 18, pp. 459-489.

[12] S. Volz, 2006, “An iterative approach for matching multiple representations of street data,” International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences, pp. 101 - 110.

[13] T. Wenjing, H. Yanling, Z. Yuxin and L. Ning, 2008, “Research on areal feature matching algorithm based on spatial similarity,” Control and Decision conference, pp. 3326-3330.

[14] 김영표, 한선희, 2001, “GIS 시장과 산업의 실태분석,” 한국GIS학회지, 제9권, 제3호, pp. 1-21.

[15] 김정옥, 허용, 이원희, 유기윤, 2009, “공간정보 플랫폼 구축을 위한 전자지도와 POI 정보의 매칭 방법,” 한국지형공간정보학회지, 제17권, 제4호, pp. 23-29.

[16] 안성준, 김우주, 박상언, 2007, “최적 온톨로지 매핑 방법론에 관한 연구,” 한국지능정보시스템학회 학술대회 논문집, pp. 457-462.

[17] 이동욱, 백성하, 김경배, 배해영, 2009, “공간 데이터 웨어하우스에서 효율적인 공간 데이터 적재를 위한 이기종 데이터 소스의 비중복 추출기법,” 한국공간정보시스템학회지, 제11권, 제2호, pp. 143 -150.

논문접수 : 2011.02.11
수 정 일 : 2011.03.21
심사완료 : 2011.03.25

김 기 락

2007년 서울대학교 지구환경시스템공학 공학사
2011년 서울대학교 건설환경공학 석사
관심분야는 GIS

허 용

2001년 서울대학교 지구환경시스템공학 공학사
2004년 서울대학교 지구환경시스템공학부 석박사통합과정 수료
2011년 서울대학교 건설환경공학 박사
관심분야는 공간정보통합 및 패턴인식

유 기 윤

1988년 연세대학교 토목공학 공학사
1990년 연세대학교 대학원 토목공학 공학석사
1998년 Ph.D. GIS, University of Wisconsin at Madison
1988년~2000년 건설교통부 사무관, 서기관
2000년~현재 서울대학교 건설환경공학부 부교수
관심분야는 GIS 및 위치기반서비스