한국지형공간정보학회
(Journal of the Korean Society for Geospatial Information Science)
Vol.24 No.1 March 2016 pp.89-98

ISSN: 1598-2955 (Print)
ISSN: 2287-6693(Online)
http://dx.doi.org/10.7319/kogsis.2016.24.1.089

연구논문

이종의 공간 데이터 셋의 면 객체 자동 매칭 방법

Automated Areal Feature Matching in Different Spatial Data-sets

김지영* · 이재빈**
Kim, Ji Young · Lee, Jae Bin

要 旨

본 연구에서는 축척과 갱신 주기가 상이한 이종의 공간 데이터 셋을 융합하기 위하여 사용자의 개입을 최소화하면서 다대다 관계에도 적용이 가능한 기하학적 방법론 기반의 면 객체 자동 매칭 방법을 제안하였다. 이를 위하여 첫째, 포함함수가 0.4 이상인 객체(노드)는 인접행렬에서 에지로 연결되었고, 이들 인접행렬의 곱을 반복적으로 수행하여 다대다 관계를 포함하는 후보 매칭 쌍을 선정하였다. 다대다 관계인 면 객체들은 알고리즘으로 생성된 convex hull로 단일 면 객체로 변환하였다. 기하학적 매칭을 위하여, 매칭 기준을 설정하고, 이들을 유사도 함수를 이용하여 유사도를 계산하였다. 다음으로 변환된 유사도와 CRITIC 방법으로 도출된 가중치를 선형 조합하여 형상 유사도를 계산하였다. 마지막으로 훈련자료에서 모든 가중치에 대한 정확도와 재현율을 나타낸 PR 곡선의 교차점인 EER로 임계값을 선정하고, 이 임계값을 기준으로 매칭 유무를 판별하였다. 제안된 방법을 수치지도와 도로명주소기본도에 적용한 결과, 일부 다대다 관계에서 잘못 매칭되는 경우를 시각적으로 확인할 수 있었으나, 통계적 평가에서 정확도, 재현율, F-measure가 각각 0.951, 0.906, 0.928로 높게 나타났다. 이는 제안된 방법으로 이종의 공간 데이터 셋을 자동으로 매칭하는데 그 정확도가 높음을 의미한다. 그러나 일부 오류가 발생한 다대다 관계인 후보 매칭 쌍을 정확하게 정량화하기 위해서 포함함수나 매칭 기준에 대한 연구가 진행되어야 할 것이다.

핵심용어 : 기하학적 매칭, 형상 유사도, CRITIC, EER

Abstract

In this paper, we proposed an automated areal feature matching method based on geometric similarity without user intervention and is applied into areal features of many-to-many relation, for confusion of spatial data-sets of different scale and updating cycle. Firstly, areal feature(node) that a value of inclusion function is more than 0.4 was connected as an edge in adjacency matrix and candidate corresponding areal features included many-to-many relation was identified by multiplication of adjacency matrix. For geometrical matching, these multiple candidates corresponding areal features were transformed into an aggregated polygon as a convex hull generated by a curve-fitting algorithm. Secondly, we defined matching criteria to measure geometrical quality, and these criteria were changed into normalized values, similarity, by similarity function. Next, shape similarity is defined as a weighted linear combination of these similarities and weights which are calculated by Criteria Importance Through Intercriteria Correlation(CRITIC) method. Finally, in training data, we identified Equal Error Rate(EER) which is trade-off value in a plot of precision versus recall for all threshold values(PR curve) as a threshold and decided if these candidate pairs are corresponding pairs or not. To the result of applying the proposed method in a digital topographic map and a base map of address system(KAIS), we confirmed that some many-to-many areal features were mis-detected in visual evaluation and precision, recall and F-Measure was highly 0.951, 0.906, 0.928, respectively in statistical evaluation. These means that accuracy of the automated matching between different spatial data-sets by the proposed method is highly. However, we should do a research on an inclusion function and a detail matching criterion to exactly quantify many-to-many areal features in future.

Keywords : Geometric Matching, Shape Similarity, CRITIC, EER

Received: 2016.02.24, accepted: 2016.03.15

* 정회원 · 서울대학교 건설환경종합연구소 연구교수(Member, Research Professor, Institute of Construction and Environmental Engineering(ICEE), Seoul National University, soodaq@snu.ac.kr)

** 교신저자 · 정회원 · 국립목포대학교 공과대학 토목공학과 부교수(Corresponding Author, Member, Associate Professor, Dept. of Civil Engineering, Mokpo National University, lee2009@mokpo.ac.kr)

1. 서 론

많은 공공 및 민간의 공간정보가 GIS에 저장되고 관리되면서 유지 및 갱신 비용 절감을 위하여 이종의 데이터 셋을 연동(conflation)하는 것은 중요한 일이 되었다. 이때 GIS별로 대상으로 하는 객체나 구축 시기 등이 상이하여 이종의 데이터 셋에서 일치하는 객체를 탐지하는 것이 주요한 부분이다. 일치하는 객체를 탐지하기 위한 일반적인 방법은 수동으로 분석하는 방법이다(Kokla, 2006). 그러나 이들 방법을 대용량의 데이터에 적용하기란 쉽지 않으며, 객체 기반의 분석을 통하여 자동으로 일치하는 객체를 탐지하는 방법이 요구된다(Duckham and Worboys, 2005). 자동으로 객체 기반 분석을 통하여 일치하는 객체를 탐지하는 방법은 사용된 기준에 따라 기하학적 방법론, 위상학적 방법론, 의미론적 방법론으로 구분할 수 있다(Tong et al., 2009). 기하학적 방법론은 거리, 방위, 위치, 형상 등의 기하학적 특성의 유사도를 측정하여 일치하는 객체를 탐지하는 기술로, 가장 많이 사용되는 방법론이다. 위상학적 방법론은 선 객체 사이의 인접성, 면 객체 간의 위치 관계 등의 위상정보를 이용하고, 의미론적 방법론은 두 객체의 명칭과 같은 속성정보에 대한 유사도를 이용한 방법이다(Kim et al., 2011; Safra et al., 2006; Tong et al., 2009). 이들 중에서 대부분의 공간정보는 속성정보가 명확하지 않은 경우가 많아 의미론적 방법론을 적용하는데 한계가 있어 기하학적 방법론에 대한 연구가 활발히 진행되고 있다.

기하학적 방법을 이용한 연동에서는 점이나 선형 객체를 이용한 매칭이 대부분이며, 면 객체를 직접 연동하는 연구는 미비한 실정이다(Guo et al., 2008; Huang et al., 2010; Zhang, 2002). 그러나 실세계의 객체들은 건물, 토지, 하천 등 면으로 되어 있으며, 이들 면형 객체를 점이나 선형 객체로 분할하여, 즉 면의 무게중심이나 선분을 이용하여 연동을 수행할 경우 해당 점이나 선형이 매칭이 되어도 이것은 면 객체의 일부로 면 객체와 연결되지 않는다(Liu, 2006). 따라서 면 객체를 직접 이용한 연동 기술 개발이 요구되며, 이때 하나의 연동 기준으로 정확하게 면 객체 간의 기하학적 특성을 평가할 수 없다. 하나의 척도는 하나의 특성만을 설명하기 때문에 면 객체의 여러 척도를 결합할 필요가 있다(Bel Hadj Ali, 2001). 따라서 기하학적 방법론에 기반을 둔 효율적인 면 객체의 연동을 위해서 여러 매칭 기준을 이용하여 두 면 객체의 형상이나 위치 특성을 계량화하고, 계량화된 매칭 기준을 결합하여 면 객체간의 유사한 정도를 판단하게 된다.

Huang et al.(2010)은 중첩면적, 무게중심간 거리, 형상비의 차, 방향의 차, 융합된 기준(synthesized criterion)을 이종의 면 객체에 반복적으로 적용하여 후보 매칭 쌍을 판별하고, 주변에 여러 개의 면 객체가 있는 경우는 해당 면 객체를 병합을 하면서 최종 매칭 면 객체를 탐지하였다. 이때, 일치하는 객체를 판별함에 있어 다중의 매칭 기준(criterion) 각각에 대한 임계값이 요구되는 문제가 발생한다. Wenjing et al.(2008)은 다중의 매칭 기준에 대한 여러 임계값을 설정하는 문제를 해결하고자 다중의 매칭 기준을 융합(fusion)하여 매칭을 수행하였다. 즉, 위치 유사도, 형상 유사도, 크기 유사도의 가중 평균에 의하여 종합 유사도(overall similarity)를 계산하였다. Samal et al.(2004)도 상황-독립(context-independent) 유사도를 산출하는 과정에서 동일한 방법으로 종합 유사도를 계산하였다. 그러나 종합 유사도를 산출하기 위하여 사용된 매칭 기준별 가중치는 훈련(training)을 통해서나 연구자가 임의로 할당하였다. 또한 Fu and Wu(2008)과 Shao and Tong (2010)는 면 객체의 매칭 기준을 가중 선형 조합하여 종합 유사도를 산출하였으며, 이들 연구에서는 가중치의 합이 1이 되겠음 사용자가 매칭 기준별로 가중치를 임의로 부여하였다. 그러나 이들 종합 유사도를 이용한 경우에는 임계값 이외에 종합 유사도를 산출하는데 필요한 가중 선형 조합의 가중치를 선정함에 있어 사용자의 개입이 요구되며, 주관적인 방법으로 가중치를 결정하는 것이 항상 유용한 것은 아니다(Wang and Luo, 2010). 이에 사용자의 주관적인 개입을 최소화하기 위한 매칭 기법이 연구되었다. Bel Hadj Ali(2001, 2002)은 다중의 매칭 기준에 자동 계층 분류 방법(automatic hierarchical classification method)을 적용하여 동일한 형상과 위치 불일치 특성이 나타나는 클래스를 추출하였다. 그러나 여전히 분류 트리(classification tree)에서 클래스를 형성하기 위한 레벨을 사용자의 경험에 의해 결정해야 한다는 한계가 있다.

판별모델을 적용하여 여러 기준을 결합하였으나 매칭 유무를 판별하는 과정에서 여전히 사용자의 개입이 요구되는 문제를 해결하기 위하여 확률 기반의 자동화된 임계값을 산정하는 방법이 연구되었다. Beeri et al.(2004), Beeri et al.(2005), Safra et al.(2010)은 정확도(precision)와 재현율(recall) 그래프의 교차 지점과 오차 개수(error count)의 최소값을 고려하여 임계값을 산정하였다. 나아가 Tong et al.(2009)은 기하학적 기준, 공간관계 기준, 속성 기준에 Beeri et al.(2004)이 제안한 확률 기반으로 종합 유사도를 산정하고, 해당 확률이 최대인 면 객체 쌍을 매칭으로 판별하였다. 즉

임계값을 고려하지 않는 매칭 기법을 제안하였다. 그러나 이들 확률 기반 매칭은 이종의 면 객체가 일대일(one-to-one)이나 일대이(one-to-two) 관계인 면 객체간의 매칭만을 판별한다는 한계가 있다.

선행연구를 분석한 결과, 기하학적 방법론에 의한 일치하는 객체를 탐지하는 것, 즉 면 객체간 매칭은 위치, 형상, 크기 등의 기준들을 이용하여 유사도를 산출하고, 이들 유사도를 결합하여 매칭인 면 객체를 판별하는 과정으로 이루어진다. 이때, 가중치나 임계값을 선정함에 있어 사용자의 개입이 최소화된 정확한 매칭 방법이 요구되며, 일대일 관계뿐만 아니라 다대다(many-to-many) 관계를 고려해야 한다. 따라서 본 연구에서는 사용자의 개입을 최소화하면서 다대다 관계에도 적용이 가능한 기하학적 면 객체 매칭 방법론을 제안하고자 한다.

2. 기하학적 방법론 기반의 면 객체 자동 매칭 기법

이종의 데이터 셋에서 일치하는 객체를 탐지하는 것은 데이터 마이닝 분야의 중요한 연구 주제인 이진 클래스 판별 문제와 같다. 즉, 이종의 데이터 셋에서 객체가 '일치한다' 또는 '일치하지 않는다'의 2개의 클래스로 판별하는 것이다. 일반적으로 이진 클래스의 판별은 분석자에 의하여 정의된 훈련자료를 분석하여 판별모델을 정립하고 실험자료에 적용하여 그 성능을 평가한다. 그러나 이종의 공간데이터 셋에서는 묘화 방법이나 축척, 구축시기의 차이로 어느 데이터 셋의 한 개의 면 객체가 다른 데이터 셋의 여러 개의 면 객체들과 중첩되거나 여러 개의 면 객체들이 여러 개의 면 객체들과 중첩되는 경우가 나타난다. 따라서 훈련자료를 분석하여 정립된 판별모델을 실험자료에 적용하여 그 성능을 평가하기 위해서는 여러 개의 면 객체들은 단일 객체로 변형되어야 한다.

따라서 사용자 개입이 최소화된 기하학적 방법 기반의 면 객체 매칭을 위하여 본 연구에서는 Fig. 1과 같이 훈련자료(training data)에서 매칭 기준을 세우고, 이들을 유사도 함수를 이용하여 유사도로 변환한다. 다음으로 유사도들을 융합하여 종합 유사도 즉, 형상 유사도(Shape Similarity)를 산출하고, 임계값을 이용하여 판별하는 단계가 요구된다. 이렇게 도출된 판별모델을 실험자료에 적용하여 판별모델의 정확도를 평가한다.

2.1 후보 매칭 객체 선정

후보 매칭 객체를 선정하기 위하여 von Goesseln and Sester(2003)가 제안한 인접행렬의 곱(multiplication)을 적용한다. 먼저, 인접행렬 $C(i,j)$을 생성하기 위하여 이종의 데이터 셋에서 참조자료의 면 객체 A와 목표자료의 면 객체 B를 노드(node)로 하고, 포함함수 $I(A,B)$가 0.4이상인 이웃하는 두 면 객체를 에지(edge)로 연결한다. 이를 수식으로 나타내면 Eq. (1)과 같다.

$$C(i,j)=\begin{cases}1 & if\ I(A,B)\geq 0.4\\ 0 & otherwise\end{cases} \qquad (1)$$

$$where, I(A,B)=\frac{Area(A\cap B)}{Min(Area(A),Area(B))}$$

인접행렬을 이용하여 후보 매칭 객체를 찾기 위하여 Eq. (2)와 같이 자신 노드와의 인접(self-adjacency)을 포함하는 행렬 C'을 생성하고, 행렬 C'에서 0인 성분(entry)이 변하지 않을 때까지 거듭제곱을 수행한다. 이때, 계산을 용이하게 하기 위하여 0이 아닌 성분은 그 값을 1로 변환한다. 여기서, m은 참조자료의 면 객체 A의 수이고, n은 목표자료의 면 객체 B의 면 객체 수이다.

$$C'=\begin{bmatrix}I_{m\times m} & C\\ C^T & I_{n\times n}\end{bmatrix} \qquad (2)$$

결과 행렬의 행 계수(row rank)는 면 객체나 객체들의 수와 같다. 즉, 일차 종속(Linear-dependence)을 제

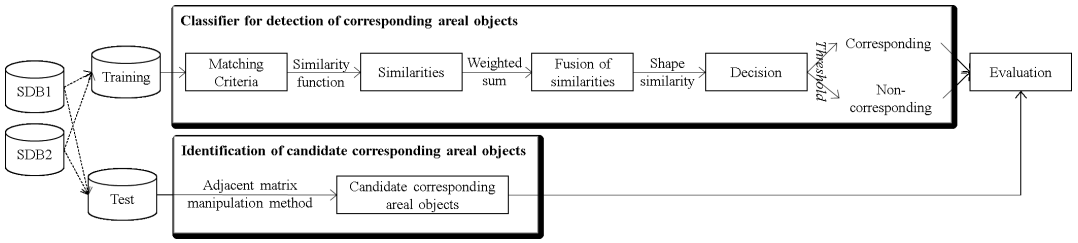


Figure 1. Process for automated areal feature matching in different spatial data-sets

거한 후 주대각선 성분만이 1인 행에 해당하는 면 객체가 일대일 관계인 후보 매칭 객체이고, 주대각선 성분 외의 다른 성분이 1인 연결되는 행들에 해당하는 면 객체들이 다대다 관계인 후보 매칭 객체가 된다. 이때, 다대다 관계인 면 객체들은 Huff and Batsell(1977)이 제안한 curve-fitting 알고리즘을 적용하여 이들 면 객체들을 포함하는 하나의 convex hull을 추출한다. 이를 통하여, 여러 개의 면 객체들이 하나의 convex hull, 즉 단일 면 객체로 변형된다. 결과적으로 다대다 관계인 후보 매칭 객체가 일대일 관계인 후보 매칭 객체로 변환됨으로써, 일대일 관계인 후보 매칭 객체가 선정된다.

2.2 판별 모델

2.2.1 단계 1: 매칭 기준

기하학적 매칭을 위하여 이종의 데이터 셋에서 정의된 후보 매칭 객체 쌍 $(A,B)_i$의 형상 치밀도, 형상 방향성과 교차 면적비가 매칭 기준으로 사용된다.

• 형상 치밀도(Shape Compactness, SC)

면 객체의 형상의 척도로 Eq. (3)으로 산출되며, 면 객체의 형상이 원과 비슷할수록 그 값이 1이 된다. 각 후보 매칭 객체 쌍의 형상 치밀도가 작을수록 해당 후보 매칭 객체 쌍은 유사하다.

$$SC_{(A,B)_i} = \left| \frac{Perimeter(A_i)}{2\times\sqrt{(\pi\times Area(A_i))}} - \frac{Perimeter(B_i)}{2\times\sqrt{(\pi\times Area(B_i))}} \right| \quad (3)$$

• 형상 방향성(Shape Orientation, SO)

각 후보 매칭 객체의 장축(longer side)과 진북의 사이각(θ_{A_i}, θ_{B_i})을 해당 면 객체의 방향으로 정의한다. Eq. (4)로 산출되며, 각 후보 매칭 객체 쌍의 형상 방향성이 작을수록 해당 후보 매칭 객체 쌍은 유사하다.

$$SO_{(A,B)_i} = \left| \cos(\theta_{A_i}\times\frac{\pi}{180}) - \cos(\theta_{B_i}\times\frac{\pi}{180}) \right| \quad (4)$$

• 교차 면적비(Intersection Ratio, IR)

Eq. (5)로 산출되며, 후보 매칭 객체 쌍이 서로 교차하는 면적이 클수록 교차 면적비가 작아진다. 즉, 후보 매칭 객체 쌍의 교차 면적비가 작을수록 해당 후보 매칭 객체 쌍은 유사하다고 볼 수 있다.

$$IR_{(A,B)_i} = \frac{Area(A_i \cap B_i)}{Area(A_i \cup B_i)} \quad (5)$$

2.2.2 단계 2: 유사도 함수

매칭 기준으로부터 얻은 거리 값에서 유사도를 취득하기 위해서는 정규화 과정이 수행되어야 한다. Eq. (6)과 같이 후보 매칭 객체 쌍$(A,B)_i$의 유사도 $\Psi(A,B)$는 거리 값 $\Delta(A,B)$ 을 [0,1] 사이로 1에서 1차 정규값을 뺀 값으로 정의되며, 이때 정규화 지수(U)는 후보 매칭 객체 쌍의 거리 값 중 최대값을 사용한다(Samal et al., 2004).

$$\Psi(A,B)_i = 1 - \frac{\Delta(A,B)_i}{U} \quad (6)$$

2.2.3 단계 3: 유사도들의 융합

2.2.1에서 정의한 기준들에 정규화 방법을 적용하여 거리 값을 유사도로 변환한 후, 유사도의 결합은 다중 속성 의사 결정(Multiple Attribute Decision Making, MADM)과 관련이 있다. MADM에서 가중 선형 조합에 따라 각 유사도가 융합된 종합 유사도, 즉 형상 유사도를 산출할 수 있다. Eq. (7)과 같이 가중치(w_j, $j=1,\cdots,c$)와 매칭 기준(c)으로부터 산출된 정규화한 값인 유사도들(SC, SO, IR)의 가중 선형 함수로 형상 유사도(SS)가 산출된다.

$$SS = w_1\times SC + w_2\times SO + w_3\times IR \quad (7)$$

가중치는 일정한 상황 하에서 다른 매칭 기준들과의 상대적 중요도를 나타내는 값으로 할당될 수 있다. 즉 가중치가 클수록 종합 평가에서 그 매칭 기준이 더 중요함을 의미한다. 사용자 개입 없이 각 기준에 유의미한 가중치를 할당하기 위하여, Diakoulaki et al.(1995)이 제안한 CRITIC(Criteria Importance Through Intercriteria Correlation) 방법을 적용하였다. 이 방법은 표준편차뿐만 아니라 상관관계를 고려하여 가중치를 결정하는 방법으로, 각 유사도 간의 표준편차와 상관관계를 사용하여 정보의 량(amount of information) C_j 를 산출하고, 이는 Eq. (8)을 따른다.

$$C_j = \sigma_j \times \sum_{k=1}^{c}(1-r_{jk}),\ j=1,\cdots,c \quad (8)$$

여기서, σ_j은 j번째 유사도의 표준편차이고, r_{jk}은 j

와 k번째 유사도 사이의 상관관계를 의미하며, 가중치 w_j는 Eq. (9)와 같이 벡터 $C=[C_j], j=1,\cdots,c$의 성분의 비로 산출된다.

$$\omega_j = \frac{C_j}{\sum_{k=1}^{c} C_k} \tag{9}$$

2.2.4 단계 4: 판별

매칭의 판별은 매칭인지 매칭이 아닌지 레이블을 붙이는 것이다. 이들 판별을 위해서는 임계값이 요구되며, 데이터 마이닝에서 널리 사용되는 Equal Error Rate(EER)로 산출될 수 있다. EER은 모든 임계값에 대한 정확도와 재현률을 그래프로 나타낸 PR 곡선(Fig. 2(a))에서 정확도와 재현율이 교차(trade-off)되는 지점으로, 정확도와 재현율이 일치하는 EER이 임계값이 된다(Bengio et al., 2005). 이때, 실제와 예측 간의 관계를 나타내는 혼동행렬(confusion matrix)로 이진 클래스의 판별모델의 결과를 표현할 수 있으며(Table 1), 클래스의 분포가 치우친 분포에서는 모델의 성능을 평가하는 지수로 널리 사용되는 정확도와 재현율은 Eq. (10)과 Eq. (11)로 구해진다(Davis and Goadrich, 2006).

Table 1. Confusion matrix

Actual \ Predicted	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

$$V_p = \frac{TP}{TP+FP} \tag{10}$$

$$V_r = \frac{TP}{TP+FN} \tag{11}$$

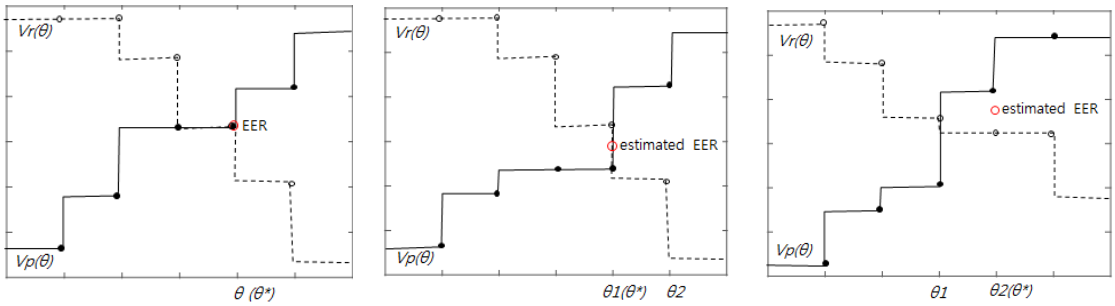
그러나 실제 데이터에서는 Fig. 2(b), (c)와 같이 정확도와 재현율의 분포가 연속적이거나 교차되는 지점이 존재하지 않는 경우가 있을 수 있다. 따라서 EER(θ^*)은 Eq. (12)로 계상된다.

2.3 평가

제안된 판별모델 평가는 정보검색(information retrieval)에서 널리 사용되는 평가지수인 F-measure를 사용한다(Yatskevich et al., 2007). Eq. (13)과 같이 F-measure($F_{0.5}$)은 2.2.절에서 설명된 정확도(V_p)와 재현율(V_r)을 같은 가중치로 두고 통합한 값으로 유도되고, 그 값이 클수록 더 좋은 매칭 결과를 나타낸다.

$$\theta^* = \begin{cases} \arg min_\theta |V_p(\theta) - V_r(\theta)| & \text{if } V_p(\theta) = V_r(\theta) \\ \dfrac{V_p(\theta_1) + V_r(\theta_1)}{2} & \text{if } V_r(\theta_1) - V_p(\theta_1) \le V_p(\theta_2) - V_r(\theta_2) \\ \dfrac{V_p(\theta_2) + V_r(\theta_2)}{2} & otherwise \end{cases} \tag{12}$$

여기서, $\theta_1 = \max_\theta(V_p(\theta) \le V_r(\theta))$, $\theta_2 = \min_\theta(V_p(\theta) \ge V_r(\theta))$ 이다.



(a) EER exits (b) EER is estimated at θ_1 (c) EER is estimated at θ_2

Figure 2. Estimation of EER in PR curve

$$F_{0.5} = \frac{V_p \times V_r}{0.5 \times V_p + 0.5 \times V_r} \quad (13)$$

3. 적용 및 결과

3.1 적용 데이터 셋

본 연구에서 사용된 데이터 셋은 2007년에 갱신된 1:5,000 축척의 수치지도(Digital Topographic Map, DTM)와 2012년 9월에 제작된 1:1,000 축척의 도로명 주소기본도(Korea Address Information System, KAIS)이다. 이들 데이터의 커버리지는 대한민국의 서울시 주변 대략 370m × 400m이고, 도심지를 포함한다. 이종의 공간 데이터 셋을 세계측지계로 일치시킨 후 아핀(affine) 변환을 수행하였으며, 이때 평균제곱근(Root Mean Square Error, RMSE) 오차는 1:5,000 축척의 수치지도 위치오차 허용범위 0.5m 보다 작은 0.48m이다. 이들 데이터 셋에서 수동으로 매칭 쌍과 비매칭 쌍을 추출하고, 이들 중 Fig. 3과 같이 각 전체 데이터 셋의 2/3는 임계값을 추정하기 위하여 훈련자료로 사용되었으며, 나머지 실험자료에 제안된 방법이 적용되었다.

3.2 임계값 설정

제안된 판별모델에서는 면 객체 쌍이 매칭여부를 판별하기 위하여 형상유사도에 대한 임계값이 필요하다. 훈련자료에 제안된 방법으로 형상 유사도를 산출하고 앞의 2.2.4에서 제안한 PR 곡선에서 EER을 구함으로

Figure 3. Data-sets used in this research after affine transformation

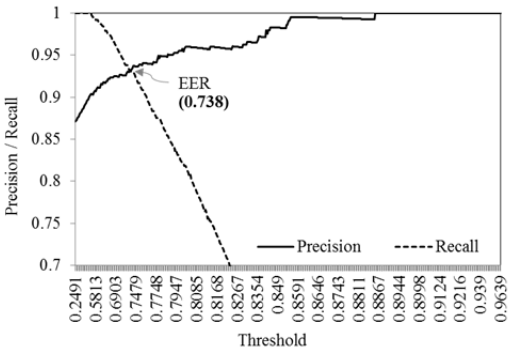


Figure 4. PR curve and EER(0.738) in training data

써 임계값을 추정하였다.

PR 곡선을 그리기 위해서 이종의 데이터의 각 쌍의 형상유사도와 판별결과가 필요하다. 이를 위하여 훈련자료에서 수동으로 매칭 쌍과 비 매칭 쌍을 추출하였다. 이때, 여러 개의 객체가 여러 개의 객체와 쌍이 된 경우에는 2.1에서 언급된 curve-fitting 알고리즘을 적용하여 여러 개의 면 객체들을 하나의 면 객체로 변환하였다. 다음으로 일대일로 변환된 훈련자료에서 구해진 형상 유사도와 수동으로 부여된 판별결과를 이용하여 Fig. 4와 같은 PR 곡선을 만들었다. Fig. 4와 같이 임계값이 변하면서 계산되는 정확도와 재현율을 이용하여, 이들이 교차되는 EER인 0.738이 임계값으로 선택되었다. 따라서 실험자료에서 판별모델을 적용하여 후보 매칭 쌍의 형상 유사도가 0.738 이상이면 매칭, 그렇지 않으면 비매칭으로 판별하게 된다.

3.3 결과

3.3.1 시각적 평가

실험자료 중첩하고 포함함수가 0.4 이상인 객체를 이용하여 앞의 2.1절에서 제안된 방법으로 여러 개의 면 객체들이 하나의 면 객체나 여러 개의 면 객체들과 중첩되는 경우는 여러 개의 객체를 둘러싸는 convex hull을 추출하여 단일 객체로 변환하였다. Table 2와 같이 포함함수가 0.4 이상인 중첩된 객체가 후보 매칭 쌍 중에서 일대일 관계가 174쌍이고, 그 외 여러 개의 면 객체들이 중첩되는 M:1, 1:N, M:N 관계가 49쌍이다. 여러 개의 면 객체들이 중첩되는 경우는 Fig. 5와 같이 convex hull을 생성하여 단일 면 객체로 변환하고 새로운 ID를 부여하였다. 정확도 평가를 위하여 단일 객체의 기존 ID와 새로운 ID를 별도의 테이블로 저장하였다.

다음으로 단일 객체로 변환된 후보 매칭 쌍에 3가지 매칭 기준을 Table 3과 같은 가중치로 선형 조합하여 형상 유사도를 산출하였다. 3가지 매칭 기준에 대하여

Table 2. Identified candidate corresponding object pairs in test data

Cardinality of candidate corresponding object	1:1	M:1	1:N	M:N	Total
No. of pairs	174	6	40	3	223

Table 3. Weighting and correlation coefficients for test data based on the CRITIC method

	SO	SC	IR	Std.	Weighting
SO	1	0.021	0.145	0.274	0.466
SC	0.021	1	0.239	0.161	0.260
IR	0.145	0.239	1	0.183	0.274

Table 4. Detected corresponding object pairs (No. of pairs)

Cardinality of candidate corresponding object	1:1	M:1	1:N	M:N	Total
Corresponding object	141	6	34	2	183
Non-corresponding object	33	0	6	1	40

순서대로 0.466, 0.26, 0.274의 가중치가 부여되었다.

마지막으로 훈련자료에서 추정된 임계값 0.738을 적용한 결과, 전체 각 후보 매칭 쌍의 형상 유사도가 0.738 이상인 객체 쌍 183개는 매칭으로, 0.738 미만인 객체 쌍 40개는 비 매칭으로 판별되었다. Table 4와 같이 일대일 관계에서 매칭 쌍 중에서 비 매칭 쌍으로 판별된 33쌍은 Fig. 6(a)와 같이 객체 쌍의 방향이 상이하거나 중첩되는 면적이 작은 경우가 일부 있었다. 그러나 형상 치밀도로 미세한 면 객체의 형상이 다른 것까지 구분되지 못하는 한계도 나타났다. 이에 향후 보다 면 객체의 작은 형상 차이도 정량화할 수 있는 기준을 적용할 필요가 있을 것이다. 다대다 관계에서 매칭 쌍 중에서 비 매칭 쌍으로 판별된 7쌍은 Fig. 6(b)의 오른쪽 아래와 같이 후보 매칭 쌍을 잘못 선정하여 비 매칭으로 판별되었으며, 나머지 경우는 convex hull을 생성하였지만 그 형상이 상이한 경우가 대부분이었으며, 일부는 일대일 관계와 동일한 이유로 잘못 판별되었다.

3.3.2 통계적 평가

실험자료에 대하여 수동으로 판별된 매치 쌍과 비교하여 그 정확도는 Eqs. (10), (11) and (13)을 이용하여 정확도, 재현율, 그리고 F-measure로 평가된다.

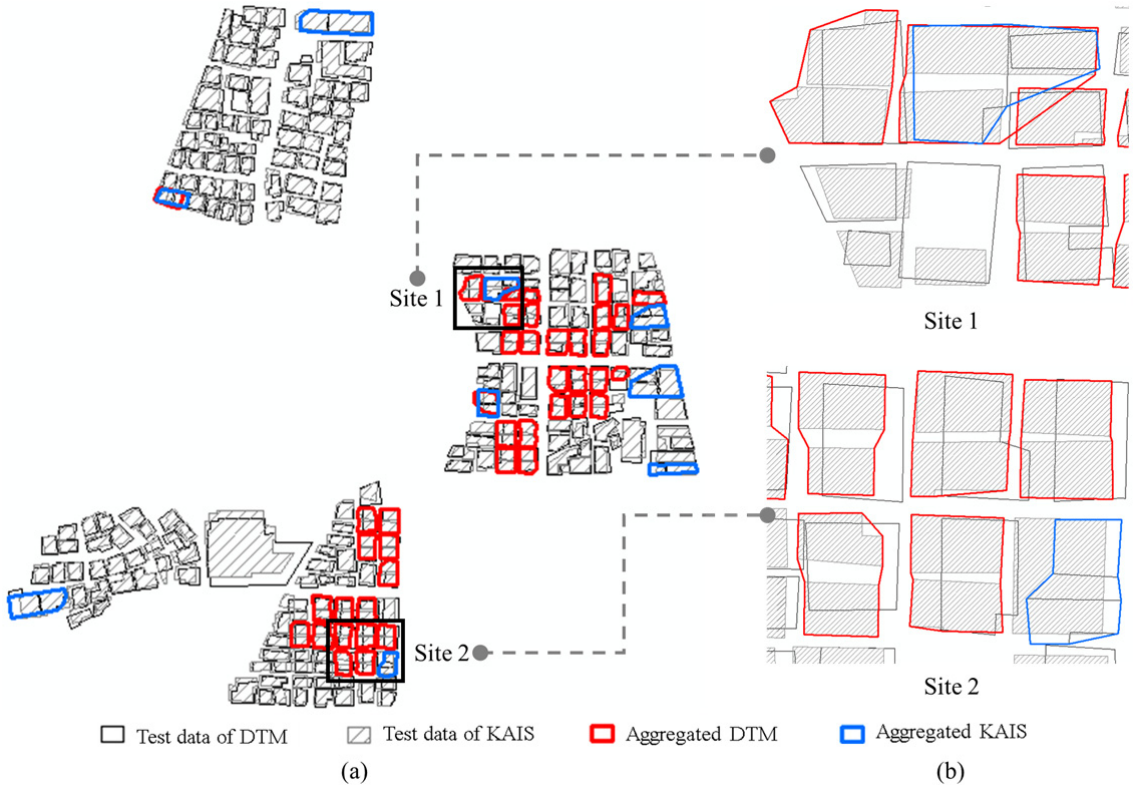


Figure 5. Identification of candidate corresponding object pairs at test data: (a) candidate corresponding object pairs of test data (left); (b) enlarged sample sites (right)

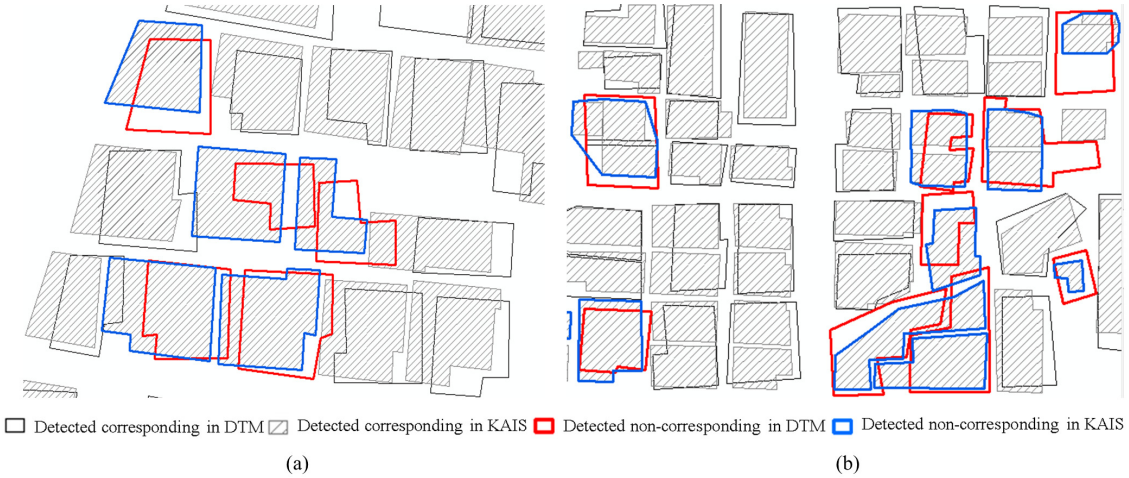


Figure 6. Mis-detected corresponding object pairs among non-corresponding object pairs in the reference data: (a) 1:1 relations; (b) M:1, 1:N or M:N relations

Table 5. Confusion matrix of test data (No. of pairs)

Target \ Reference	Corresponding	Non-Corresponding
Corresponding	174	9
Non-Corresponding	18	22

이를 위하여 Table 5와 같은 혼동 행렬을 구성하였으며, 참인 매칭 중에서 174쌍이, 참인 비 매칭 쌍 중에서는 22쌍이 정확하게 판별되었다. 결과적으로 제안된 방법으로 판별된 매칭 쌍의 정확도, 재현율, F-measure는 차례대로 0.951, 0.906, 0.928이다.

4. 결 론

본 연구에서는 사용자의 개입을 최소화하면서 다대다 관계에도 적용이 가능한 기하학적 면 객체 매칭 방법을 제안하였다. 이를 위하여 이종의 데이터를 동일한 좌표계로 변환하고, 아핀 변환을 수행하였다. 이때, RMSE은 0.48m이다. 전처리된 이종의 공간정보를 중첩하여 포함함수가 0.4이상인 객체 쌍에 인접행렬의 곱을 적용하여 후보 객체 쌍을 탐지하였다. 이렇게 탐지된 후보 객체 쌍에서 구해진 형상 치밀도, 형상 방향성, 교차 면적비를 정규화하여 유사도로 변환하고, CRITIC 방법으로 산출된 가중치를 곱하여 형상 유사도를 산출하였다. 훈련자료의 PR 곡선의 EER로 도출된 임계값 0.738을 적용하여, 실험자료에서 매칭 객체 쌍을 탐지하였다. 마지막으로 매칭 결과를 시각적인 측면과 통계적인 측면에서 평가하였다. 시각적으로 평가한 결과 다대다 관계에서는 후보 매칭 쌍이 잘못 선정되어 매칭 쌍이 비 매칭 쌍으로 판별된 사례가 있었으며, 그 외에는 가중치가 큰 형상 방향성이 작은 면 객체 쌍이나 중복되는 면적이나 모양의 차이가 큰 면 객체 쌍에서 잘못 판별되는 경우가 관찰되었다. 통계적으로 평가한 결과, 혼동 행렬에서 계산된 정확도, 재현율, F-measure가 0.951, 0.906, 0.928로 높게 나타났다. 결과적으로 사용자의 개입이 없는 제안된 매칭 방법이 이종의 공간정보에서 일대일 관계인 객체 쌍뿐만 아니라 다대다 관계인 객체 쌍에서도 일치하는 객체를 판별하는데 효율적이라는 것을 알 수 있었다.

그러나 시스템의 구축 목적이나 축척에 따른 면 객체의 묘화 방법이 상이하거나 다대다 매칭이 많이 발생할 수 있어 다대다 관계인 후보 매칭 쌍을 선정하는 포함함수나 형상의 작은 차이를 보다 정확하게 정량화할 수 있는 매칭 기준에 대한 연구가 필요하다.

감사의 글

본 논문은 중소기업청에서 지원하는 2015년도 산학연협력 기술개발사업(No. C0276754)의 연구수행으로 인한 결과물임을 밝힙니다.

References

1. Beeri, C., Doytsher, Y., Kanza, Y., Safra, E. and Sagiv, Y., 2005, Finding corresponding objects when integrating several geo-spatial datasets, Proc. of the 13th annual ACM international workshop on Geographic information systems, ACM, New York, USA, pp. 87-96.
2. Beeri, C., Kanza, Y., Safra, E. and Sagiv, Y., 2004, Object fusion in geographic information systems. Proc. of the 13th International Conference on Very Large Data Bases(VLDB 2004), Morgan Kaufmann, Toronto, Canada, pp. 816-827.
3. Bel Hadj Ali, A., 2001, Positional and shape quality of areal entities in geographic databases: quality information aggregation versus measures classification, Proc. of ECSQARU 2001 workshop on Spatio-temporal reasoning and geographic information systems, Toulouse, France, pp. 1-16.
4. Bel Hadj Ali, A., 2002, Moment representation of polygons for the assessment of their shape quality, Journal of Geographical Systems, Vol. 4, No. 2, pp 209-232.
5. Bengio, S., Maréthoz, J. and Keller, M., 2005, The expected performance curve, Proc. of the ICML'05 workshop on ROC analysis in machine learning, Bonn, Germany, pp. 43–50.
6. Davis, J. and Goadrich, M., 2006, The relationship between precision-recall and ROC curves, Proc. of the 23rd International Conference on Machine Learning(ICML 2006), Pittsbrugh, USA, pp. 233-240.
7. Diakoulaki, D., Mavrotas, G. and Papayannakis, L., 1995, Determing objective weights in multiple criteria problems: the CRITIC method, Computers & Operational Research, Vol. 22, No. 7, pp. 763-770.
8. Duckham, M. and Worboys, M., 2005, An algebraic approach to automated geospatial information fusion, International Journal of Geographical Information Science, Vol. 19, No. 5, pp. 537–557.
9. Fu, Z. and Wu, J., 2008, Entity matching in vector spatial data, Proc. of the XXIth ISPRS Congress,

ISPRS, Beijing, China, pp. 1467-1472.
10. Guo, L., Cui, T., Zheng, H. and Wang, H., 2008, Arithmetic for area vector spatial data matching on spatial direction similarity, Journal of Geomatics Science and Technology, Vol. 25, No. 5, pp. 380-382.
11. Huang, L., Wang, S., Ye, Y., Wang, B. and Wu, L., 2010, Feature matching in cadastral map integration with a case study of Beijing, Proc. of 2010 18th International Conference on Geoinformatics, IEEE, Beijing, China, pp. 1-4.
12. Huff, D. A. and Batsell, R.R., 1977, Delimiting the areal extent of a market area, Journal of Marketing Research, Vol.15, pp. 581-585.
13. Kim, J., Huh, Y., Kim, D. S. and Yu, K., 2011, A new method of automatic areal feature matching based on shape similarity using CRITIC method, Journal of the Korean Society of Surveying, Vol. 29, No. 2, pp. 113-121.
14. Kokla, M., 2006, Guidelines on geographic ontology integration, Proc. of the ISPRS technical commission II symposium, ISPRS, Vienna, Austria, pp. 67-72.
15. Liu, Z., 2006, The research on areal feature matching among the conflation of urban geographic databases, Master's thesis, University of Wuhan.
16. Safra, E., Kanza, Y., Sagiv, Y. and Doytsher, Y., 2006, Integrating Data from Maps on the World-Wide Web, Proc. of the 6th International Symposium on Web and Wireless Geographical Information Systems(W2GIS 2016), Springer LNCS 4295, Hong Kong, China, pp. 180-191.
17. Safra, E.., Kanzab, Y., Sagiv, Y., Beeri, C. and Doytsher, Y., 2010, Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets, International Journal of Geographical Information Science, Vol. 24, No. 1, pp. 69-106.
18. Samal, A., Seth, S. and Cueto, K., 2004, A feature-based approach to conflation of geospatial sources, International Journal of Geographical Information science, Vol. 18, No. 5, pp. 459-489.
19. Shao, S. and Tong, C., 2010, A Matching method for multi-characteristic vector elements of complex polygon, 2010 International Conference on Multimedia Technology (ICMT), IEEE, Ningbo, China, pp. 1-4.
20. Tong, X., Shi, W. and Deng, S., 2009, A probability-based multi-measure feature matching method in map conflation, International Journal of Remote Sensing, Vol. 30, No. 20, pp. 5453-5472.
21. Wang, Y. and Luo, Y., 2010, Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making, Mathematical and Computer Modelling, Vol. 51, pp. 1-12.
22. von Goesseln, G. and Sester, M., 2003, Change detection and integration of topographic updates from ATKIS to geoscientific data sets. Proc. of International Conference on Next Generation Geospatial Information, Boston, USA, pp. 69-80.
23. Wenjing, T., Yanling, H., Yuxin, Z. and Ning, L., 2008, Research on areal feature matching algorithm based on spatial similarity, Proc. of Chinese Control and Decision Conference(2008 CCDC), IEEE, Yantai, China, pp. 3326-3330.
24. Yatskevich, M., Giunchiglia, F. and Avesani, P., 2007, A large scale dataset for the evaluation of matching systems, Proc. of 4th European Semantic Web Conference, http://citeseer.ist.psu. edu/viewdoc/summary?doi=10.1.1.103.8240
25. Zhang, Q., 2002, Research on feature matching and conflation of geographic databases, Doctoral thesis, University of Wuhan.