한국측량학회지, 제 29권 제 2호, 2011. 4 : pp. 113~121

CRITIC 방법을 이용한 형상유사도 기반의 면 객체 자동매칭 방법

A new method for automatic areal feature matching based on shape similarity using CRITIC method

김지영[1] · 허 용[2] · 김대성[3] · 유기윤[4]
Kim, Jiyoung · Huh, Yong · Kim, Dae Sung · Yu, Kiyun

Abstract

In this paper, we proposed the method automatically to match areal feature based on similarity using spatial information. For this, we extracted candidate matching pairs intersected between two different spatial datasets, and then measured a shape similarity, which is calculated by an weight sum method of each matching criterion automatically derived from CRITIC method. In this time, matching pairs were selected when similarity is more than a threshold determined by outliers detection of adjusted boxplot from training data. After applying this method to two distinct spatial datasets: a digital topographic map and street-name address base map, we conformed that buildings were matched, that shape is similar and a large area is overlaid in visual evaluation, and F-Measure is highly 0.932 in statistical evaluation.

Keywords : Areal feature matching, Similarity, CRITIC method, Building

초 록

본 연구에서는 기하학적 정보를 바탕으로 생성된 유사도 기반의 면 객체 자동 매칭 방법을 제안하였다. 이를 위하여 서로 다른 공간자료에서 교차되는 후보 매칭 쌍을 추출하고, CRITIC 방법을 이용하여 연동 기준별 가중치를 자동으로 생성하여 선형조합으로 추출된 후보 매칭 쌍 간의 형상유사도를 측정하였다. 이때, 훈련 자료에서 조정된 상자도표의 특이점 탐색을 적용하여 도출된 임계값 이상인 경우가 매칭 쌍으로 탐색된다. 제안된 방법을 이종의 공간 자료(수지치도 2.0과 도로명주소 기본도)의 일부 지역에 적용한 결과, 시각적으로 형상이 유사하고 교차되는 면적이 넓은 건물 객체가 매칭 되었으며, 통계적으로 F-Measure가 0.932로 높게 나타났다.

핵심어 : 면 객체 매칭, 유사도, CRITIC 방법, 건물

1. 서 론

1.1 연구 목적

최근에는 구글, 다음, 네이버 등의 포털 사이트의 지도 서비스, 내비게이션 등을 통하여 다양한 공간정보가 실생활에서 사용되고 있으며, 보다 양질의 다양한 공간정보에 대한 사용자 요구가 증대되고 있다. 이에 포털 사이트뿐 아니라 공공기관에서도 API로 제약적이지만 공간자료를 사용자 기호에 맞게 재생산할 수 있는 환경을 제공하고 있다. 그러나 각 시스템 별로 구축목적에 따라 공간자료와 속성자료의 정의 및 묘사 방법이나, 구축 시기 및 갱신 주기 등이 상이하여 서로 다른 공간자료를 통합적으로 활용하는데 한계가 있다(Huang 등, 2010; Wenjing 등, 2008). 즉, 공간자료의 사용목적에 따라 스키마, 데이터 모델 및 정확도, 포맷, 축척 등이 서로 상이하고, 공간자료별로 구축 시기나 갱신 주기가 상이하여 같은 지역의 공간자료라

1) 서울대학교 대학원 공과대학 건설환경공학부 박사과정(E-mail: soodaq@snu.ac.kr)
2) 정회원 · 서울대학교 대학원 공과대학 건설환경공학부 공학박사(E-mail : hy7808@snu.ac.kr)
3) 정회원 · 건국대학교 신기술융합학과 전임연구원 공학박사(E-mail : mutul94@snu.ac.kr)
4) 교신저자 · 정회원 · 서울대학교 공과대학 건설환경공학부 부교수(E-mail: kiyun@snu.ac.kr)

하더라도 도형 정보 및 속성 정보가 다르다. 이런 이유로 서로 다른 공간자료의 공유는 어려운 실정이다. 그러나 서로 다른 공간 데이터 공유를 통해 사업의 중복추진을 예방할 수 있고, 단순 정보만 제공하는 기능을 넘어 새로운 부가가치를 창출할 수 있다. 따라서 공간자료들의 공유를 위해서는 기 구축된 공간자료들 간의 연동이 선행되어야 한다.

서로 다른 공간자료의 연동 방법은 사용된 기준에 따라 기하학적 방법론, 위상학적 방법론, 의미론적 방법론으로 구분할 수 있다(Tong 등, 2009). 기하학적 방법론은 거리, 방위, 위치, 형상 등의 기하학적 특성의 유사도를 측정하여 연동하는 기술로, 가장 많이 사용되는 방법론이다. 위상학적 방법론은 선형객체 사이의 인접성, 폴리곤 간의 위치 관계 등의 위상정보를 이용하고, 의미론적 방법론은 두 객체의 명칭과 같은 속성정보에 대한 유사도를 이용한 방법이다(Cohen, 2000; Safra 등, 2006; Tong 등, 2009). 이들 방법론 중에서 현재 우리나라에 구축되어 있는 공간자료 특히 건물을 관리하는 시스템의 경우는 속성정보에 대한 메타 데이터나 데이터베이스에 대한 명세서 등이 명확치 않아 의미론적 방법론을 적용하는데 한계가 있다. 일례로 건물명의 경우, 경기도 수원시 팔달구 일부 지역에 해당하는 수치지도2.0에는 37.2%, 도로명주소 기본도에는 26.6%의 건물에만 건물명이 부여되어 있으며, 부여된 건물명 또한 상이한 경우가 있어 의미론적 방법론에 활용할 수 있는 동일한 건물명의 비율은 더 낮을 것이다. 다음으로, 위상학적 방법론은 주변 객체와의 위상관계의 파악이 어려운 폴리곤이나 포인트 객체에서는 많이 이용되지 않고 있다.

따라서 본 연구에서는 기하학적 방법론을 활용하여 서로 다른 공간자료를 연동하는 방법을 제안하고자 한다.

표 1. 면 객체를 직접 이용한 연동 방법을 제안한 선행연구

선행연구	연동 기준	임계값 설정
Ali (2001)	• 하우스도르프 거리: 면 객체의 내부를 포함한 면 객체간의 하우스도르프 거리 • 면적의 거리(Areal distance): 주어진 면 객체의 수학적 거리 • 형상 측정치: 수학적 모멘트를 사용하여, 각 벡터 면 객체를 이진영상으로 변환하여 계산	
Samal 등 (2004)	기하정보와 위상정보를 이용한 연동 방법 제안하였으며, 기하정보 연동 기준만을 정리하면, • 위치유사도: 회전, 이동, 스케일에 변하지 않는 최소사각형(MBR, Minimum bounding rectangle)의 무게중심 이용 • 형상유사도: 면 객체의 중심을 맞추고, 하나의 면 객체를 이루는 선형에 일정 버퍼영역을 설정하여 해당 영역 내에 비교하고자하는 면 객체의 형상이 얼마나 포함되는지를 계산 • 통합유사도: 각 유사도의 평균가중치를 사용	• 통합유사도를 바탕으로 유사도행렬 생성 • 통합유사도의 경우 가중치는 훈련이나 사용자가 임의 산정
Wenjing 등 (2008)	• 위치유사도: Samal 등(2004)과 동일 • 형상유사도: 무게중심에서 각 면 객체 경계에 있는 점 객체사이의 거리를 이용 • 면적유사도: 사람이 눈으로 매칭을 판단할 때 많이 사용되는 특성으로 면 객체의 면적 • 통합유사도: Samal 등(2004)과 동일	• 임계값을 설정하지 않음 • 통합유사도의 경우 가중치는 훈련이나 사용자가 임의 산정
Fu and Wu (2008)	• 통합유사도: 중복면적비가 0.5보다 큰 매칭 후보쌍 중에서, 두 면 객체의 면적, 중복면적비, 중심간 거리 등을 통합한 연동 기준	• 임계값 선정에 대한 명확한 근거가 제시되지 않음
Huang 등 (2010)	• 중복면적: MBR을 생성하여 중복되는 면적비 계산 • 무게중심간 거리: 두 객체 사이의 유클리디안 거리 • 형상비의 차: 형상비(shape ration)는 면 객체의 MBR과 면 객체의 면적비 간의 차 • 대각방향의 차: 면 객체 MBR의 대각방향을 계산하고, 이들 간의 차 • 통합 기준(synthesized criterion): 두 면 객체 사이의 거리와 중복면적을 통합하여 정의된 기준	• 임계값을 설정과 관련된 명확한 근거가 제시되지 않음 • 각각의 연동 기준에 대하여 임계값 설정

이때, 공간자료 중 그 활용도가 높으며, 2012년부터 시행될 도로명주소 사용에 대응하기 위하여 국가기본도인 수치지도2.0과 도로명주소 기본도의 건물 객체를 그 대상으로 한다.

1.2 이론적 고찰

기하학적 방법을 이용한 연동에서는 점이나 선형 객체를 이용한 매칭이 대부분이며, 면 객체를 직접 연동하는 연구는 미비한 실정이다(Guo 등, 2008; Huang 등, 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 이는 면 객체의 일부분이 매칭이 된 경우로 매칭된 면 객체의 일부가 면 객체와 연결되지 않는다(Liu, 2006). 따라서 면 객체를 직접 이용한 연동 기술 개발이 필요할 것이다.

면 객체의 연동은 동일한 지역의 공간자료간의 유사(similarity)와 차이(difference)를 명확히 하는 과정으로 볼 수 있다. 즉, 효율적인 공간정보의 연동을 위해서 다양한 연동 기준(criterion)을 바탕으로 두 객체간의 유사와 차이를 계량화하고, 계량화된 연동 기준을 훈련(training)이나 사용자의 결정으로 선정된 임계값을 바탕으로 유사한 정도를 판단하게 된다. 표 1과 같이 면 객체를 이용한 선행 연구를 연동 기준과 임계값 선정 측면에서 살펴보았다.

2. 제안된 방법

2.1 CRITIC를 이용한 유사도 기반 매칭

좌표변환만을 수행한 서로 다른 공간자료의 면 객체를 직접 이용하여, 사용자의 개입이 최소화된 유사도 기반의 면 객체 매칭 방법을 제안한다. 제안된 방법의 연구 흐름은 그림 1과 같다. 포맷과 좌표변환 후 중첩 분석(intersect)을 통하여 후보 매칭 쌍을 추출한다. 다음으로 CRITIC (CRiteria Importance Through Intercriteria Correlation) 방법으로 자동 생성된 가중치를 이용하여 후보 매칭 쌍 간의 형상유사도를 측정하고, 설정된 임계값을 기준으로 매칭 쌍을 탐색한다. 이때, 임계값은 훈련 자료에서 조정된 상자도표(adjusted boxplot)방법을 적용하여 설정된다.

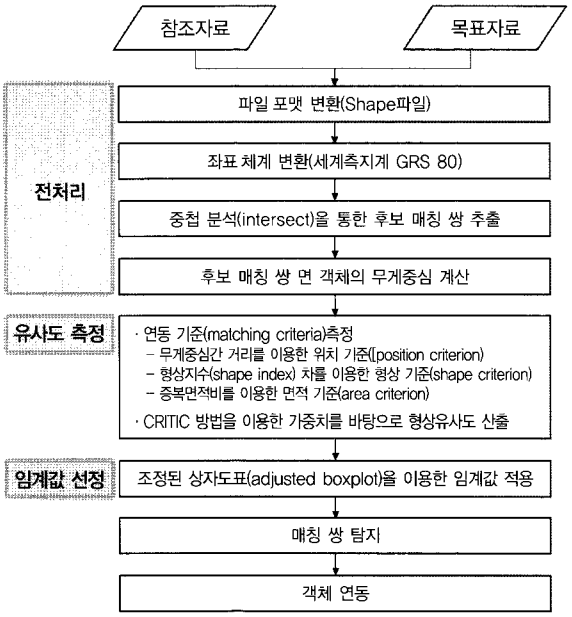

그림 1. CRITIC를 이용한 유사도 기반 면 객체 매칭 흐름도

2.2 전처리

면 객체를 직접 연동하는 선행연구에서는 전처리로 연동을 하고자 하는 공간자료의 파일 포맷, 좌표 등을 통일하고, affine변환을 통해 대상 공간자료 간의 위치 오차가 최소가 되도록 지도 정렬을 수행하였다. 그러나 affine변환을 위해서는 기준점을 추출해야 한다.

최근 우리나라의 경우 GRS80 타원체를 기준으로 하는 세계측지계를 공간자료 획득, 관리 및 표현의 기준으로 사용하고 있는데, 『국토지리정보원 고시 제2003-497호』에서 제시한 Molodensky-Badekas모델에 의하여 결정된 국가좌표변환계수와 변환식으로 좌표 변환을 수행 시 20~40cm 가량의 지역적 편의를 유발된다(조재관 등 , 2008). 즉, 좌표 변환 시 발생하는 계통적 오차는 건물 객체의 크기에 비하면 작은 수치로 좌표 변환만을 수행한 후 중복된 건물 형상정보를 활용하여 유사도를 측정하는데 문제가 없는 것으로 사료된다. 따라서 본 연구에서는 파일 포맷을 통일하고 좌표 변환만을 수행하여, 사용자의 개입을 최소화하였다.

다음으로, 동일한 좌표로 변환된 참조자료와 목표자료에 중첩 분석을 수행하여, 교차되는 건물 객체를 매칭 가능성이 있는 객체로 판단하여 후보 매칭 쌍을 추출하고, 추출된 후보 면 객체에서 무게중심을 구한다.

2.3 CRITIC 방법을 이용한 형상유사도 생성

2.3.1 연동 기준

본 연구에서는 좌표 변환만을 수행하였으므로, 연동 기준을 선정함에 있어, 거리, 위치 등과 관련된 기준만을 고려할 필요가 있다. 이에 선행연구의 연동 기준 중에서 면 객체의 대표점인 무게중심간의 거리를 이용한 위치기준(position criterion), 면 객체의 형상적 특징을 설명하는 형상지수(shape index)를 이용한 형상 기준(shape criterion), 중복면적비를 이용한 면적 기준(area criterion)을 연동 기준으로 선정하였다.

(1) 위치 기준

전처리 단계에서 구해진 참조자료 $P_A(X_A, Y_A)$와 목표자료 $P_B(X_B, Y_B)$의 무게중심 즉, 면 객체 대표점간의 차는 면 객체의 무게중심사이의 거리를 의미하며, 유클리드 거리로 구한다. 이렇게 구해진 무게중심간 거리의 최대값을 이용하여 정규화시킨 값이 바로 위치 기준으로 후보 매칭 쌍의 위치 기준이 클수록 대응 객체는 유사하다고 볼 수 있다(식 (1)).

$$\text{위치기준}(A,B) = 1 - \frac{\text{무게중심간거리}(P_A, P_B)}{\max(\text{무게중심간 거리}(P_{A_{all}}, P_{B_{all}}))} \quad (1)$$

$$\text{여기서, 무게중심간 거리}(P_A, P_B) = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}$$

(2) 형상 기준

Burghardt and Steiniger(2005)에 의하면 면 객체의 면적과 둘레는 양의 상관관계를 보이며, 이때, 면적과 둘레를 크기(size)라 하면, 크기와 형상지수 사이에도 양의 상관관계가 나타난다고 한다. 즉, 면적과 둘레가 크면 형상지수도 크다는 의미로, 이는 형상지수가 면 객체의 형상특성을 나타낸다고 볼 수 있다. 이렇게 구해진 형상지수 차의 최대값을 이용하여 정규화시킨 값이 바로 형상 기준으로 후보 매칭 쌍의 형상 기준이 클수록 대응 객체는 유사하다고 볼 수 있다(식 (2)).

$$\text{형상기준}(A,B) = 1 - \frac{\text{형상지수차}(A,B)}{\max(\text{형상지수차}(A_{all}, B_{all}))} \quad (2)$$

$$\text{여기서, 형상지수차}(A,B) = \left| \frac{\text{둘레}(A)}{2\sqrt{\pi \times \text{면적}(A)}} - \frac{\text{둘레}(B)}{2\sqrt{\pi \times \text{면적}(B)}} \right|$$

(3) 면적 기준

대칭차(symmetric difference)는 두 면 객체의 전체 면적에서 중복되는 면적을 차하는 경우로, 대상 면 객체간의 중복면적이 넓을수록 대칭차는 작은 값을 갖게 된다. 따라서 중복면적비는 대상 면 객체의 각각의 면적의 합에 대한 대칭차의 비로 정의하였다. 이렇게 구해진 중복면적비의 최대값을 이용하여 정규화시킨 값이 바로 면적 기준으로 후보 매칭 쌍의 면적 기준이 클수록 대응 객체는 유사하다고 볼 수 있다(식 (3)).

$$\text{면적기준}(A,B) = 1 - \frac{\text{중복면적비}(A,B)}{\max(\text{중복면적비}(A_{all}, B_{all}))} \quad (3)$$

$$\text{여기서, 중복면적비}(A,B) = \left| \frac{\text{면적}(A \cup B) - \text{면적}(A \cap B)}{\text{면적}(A) + \text{면적}(B)} \right|$$

2.3.2 형상유사도

여러 기준에서 최적의 대안을 선택하는 다기준 의사결정에서 주관적 방법으로 가중치를 결정하는 것은 항상 유용한 방법은 아니며, 객관적 방법에서 많이 사용되고 있는 표준편차 방법과 평균 방법은 속성들의 상호관계를 고려하지 않고, 동일한 가중치를 부여한다는 한계가 있다(Wang and Luo, 2010). 반면에 Diakoulaki 등(1995)이 제안한 CRITIC 방법은 각 기준의 표준편차뿐만 아니라 기준들 간의 상관관계를 고려함으로써 가중치를 결정하는 방법이다. 따라서 본 연구에서는 CRITIC 방법을 이용하여 사용자가 가중치를 선정하는 것이 아니고, 자동으로 가중치를 산출하여 세 가지의 연동 기준의 선형조합을 유도한다. 이때, 연동 기준들의 선형조합을 형상유사도라 한다(식 (4)).

$$\begin{aligned} &\text{형상유사도}(A,B) \\ &= \omega_1 \times \text{위치기준}(A,B) + \omega_2 \times \text{형상기준}(A,B) + \omega_3 \times \text{면적기준}(A,B) \end{aligned}$$

$$\text{여기서, } \omega_j = \frac{C_j}{\sum_{k=1}^{m} C_k}, \; C_j = \sigma_j \times \sum_{k=1}^{m} (1 - r_{jk}) \quad (4)$$

C_j: 각 매칭 기준의 정보량
σ_j: 각 매칭 기준의 표준편차
r_{jk}: 매칭 기준간의 상관관계

CRITIC 방법은 크게 4 단계로 구성된다.

1 단계: 각 연동 기준별 표준화(normalization)
2 단계: 다기준에서 발생할 수 있는 강도 차를 평가하기 위하여, 연동 기준들 사이의 상관계수 산출
3 단계: 상관계수와 각 매칭 기준별 표준편차를 이용하여 각 연동 기준 내 정보량(amount of information) 산출
4 단계: 가중치 산출

후보 매칭 쌍의 (A, B)가 일정한 값보다 큰 경우 대응 객체는 유사하다고 볼 수 있다. 이때, "대응 객체가 유사하다, 유사하지 않다"의 기준을 얼마로 해야 하느냐는 문제가 있다.

3. 임계값 선정

3.1 조정된 상자도표

2.3.2에서 언급한 바와 같이 형상유사도를 바탕으로 대상 객체간의 매칭을 판단하기 위해서는 유사하다는 기준, 즉 임계값이 필요가 있다. 각 대상 객체별로 구해진 형상유사도의 분포를 살펴보았을 때, 극단적인 큰 값이나 작은 값이 나타나는 구간이 임계값이 될 수 있다. 즉, 형상유사도 분포에서 특이점(outlier) 범위에 속하는 경우로 판단할 수 있다. 따라서 본 연구에서는 특이점 판별에 널리 사용되는 상자도표(boxplot)를 활용하여 임계값을 선정한다. 그러나 형상유사도는 그 값이 클수록 대응 객체가 유사하므로 정규분포보다는 치우진 분포(skewed distribution)가 나타날 것이다. 치우침이 있는 분포에 기존의 상자도표에서 정의하는 펜스(fence)로 특이점을 판별할 경우 정규관측값(regular observation)이 특이점으로 과잉 분류되는 문제가 있다(Hubert and Van der Veeken, 2008).

따라서 본 연구에서는 Hubert and Vandervieren (2008)이 제안한 치우침이 있는 분포가 반영된 조정된 상자도표(adjusted boxplot)를 이용하였다. 식 (5)의 하한펜스(low fence)와 상한펜스(high fence) 사이 즉, 펜스 사이를 벗어난 비정상적으로 극단적인 관측값이 특이점으로 판별된다.

$$\begin{matrix}\left[Q_1 - 1.5e^{-4MC}IQR;\ Q_3 + 1.5e^{3MC}IQR\right], \text{if } MC \geq 0 \\ \left[Q_1 - 1.5e^{-3MC}IQR;\ Q_3 + 1.5e^{4MC}IQR\right], \text{if } MC < 0\end{matrix} \quad (5)$$

여기서, $IQR = Q_3 - Q_1$: 사분위범위

(Q_1 : 제25분위수, Q_3 : 제75분위수)

조정된 상자도표의 구간에서 MC(medcouple)는 왜도(skewness)를 측정하는 강건한 측정치로, $MC \geq 0$ 인 경우는 대칭 분포이거나 오른쪽으로 치우진 분포를, $MC < 0$ 인 경우는 왼쪽으로 치우친 분포를 의미한다. Brys 등(2004)이 제안한 MC는 식 (6)과 같이 정의된다.

$$MC = \underset{x_i \leq Q_2 \leq x_j}{med} \left\{ \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i} \right\} \quad (6)$$

여기서, Q_2: 관측값의 중앙값(제50분위수)

3.2 훈련 자료에서 임계값 설정

수동으로 정 매칭 쌍과 오 매칭 쌍을 추출하여 훈련 자료를 생성한다. 각 훈련 자료별로 형상유사도 분포를 살펴보고, 각 분포에 조정된 상자도표를 적용하여 펜스를 정의한다. 다음으로, 정 매칭 쌍 분포와 오 매칭 쌍 분포의

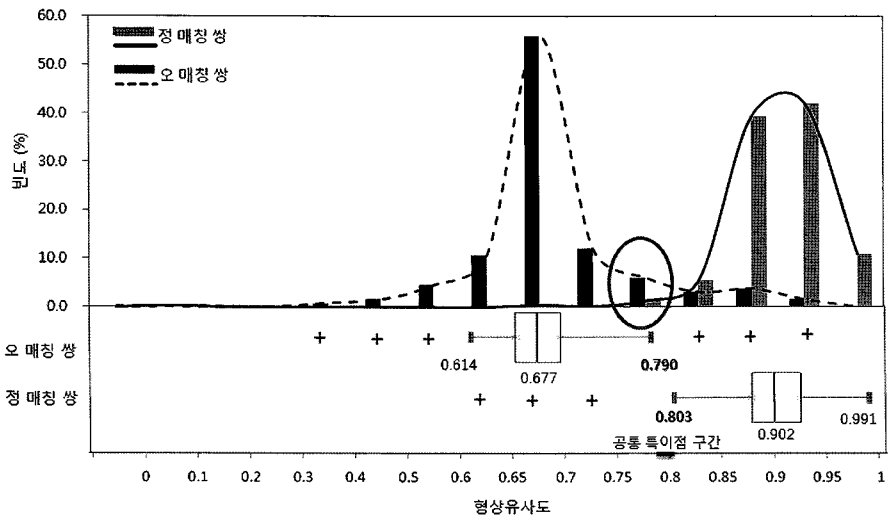

그림 2. 훈련 자료의 히스토그램과 조정된 상자도표

표 2. 훈련 자료로 생성된 조정된 상자도표의 주요 특성값

구 분	상한펜스	Q1	Q2	Q3	하한펜스	IQR	MC
정 매칭 쌍	0.991	0.878	0.902	0.926	0.803	0.047	-0.020
오 매칭 쌍	0.790	0.658	0.677	0.701	0.614	0.044	0.102

각 펜스 밖의 영역 중에서 공통인 영역을 특이점이 관찰되는 영역으로 판별한다. 이렇게 찾아진 공통 특이점 구간에서 정 매칭 쌍의 형상유사도 빈도가 오 매칭 쌍의 형상유사도 빈도보다 낮아지는 구간을 찾고, 해당 구간의 평균을 임계값으로 설정한다.

수치지도2.0 37709094도엽의 건물 레이어와 동일 영역의 도로명주소 기본도 건물 레이어에서 형상이 유사하고, 많이 중복되어 있는 1:1로 매칭된 객체를 정 매칭 쌍으로 수동 추출하였으며, 오 매칭 쌍은 중복되는 면적이 작고 형상이나 크기도 상이한 객체를 추출하였다. 그 결과 정 매칭 쌍 451개, 오 매칭 쌍 132개를 훈련 자료로 활용하였다.

수동으로 추출된 정 매칭 쌍과 오 매칭 쌍에 대하여, 본 논문에서 제안하는 방법으로 형상유사도를 구하고, 히스토그램과 조정된 상자도표를 그려보았다(그림 2). 형상유사도는 그 값이 클수록 대상 객체가 유사하다고 판단할 수 있다. 그러므로 조정된 상자도표에서 정 매칭 쌍의 경우는 형상유사도가 작을수록 유사하지 않은 경우에 해당되므로, 하한펜스 미만인 구간이 잠재적인 특이점일 가능성이 있다. 오 매칭 쌍의 경우는 형상유사도가 클수록, 즉 상한펜스 초과인 관측값이 특이점일 수 있다. 따라서 정 매칭 쌍과 오 매칭 쌍에서 공통 특이점 구간은 오 매칭 쌍의 상한펜스와 정 매칭 쌍의 하한펜스의 사이 즉, 0.790 < 형상유사도 < 0.803이다(표 2). 다음으로 공통 특이점 구간 중 0.75~0.8구간에서 정 매칭 쌍의 빈도가 오 매칭 쌍의 빈도보다 낮아지므로, 0.790~0.803의 평균인 0.797을 임계값으로 설정하였다. 따라서, 형상유사도 ≥ 0.797이면 대상 객체는 유사하고, 형상유사도 < 0.797이면 대상 객체는 유사하지 않다고 판별하게 된다.

4. 적용 및 평가

4.1 실험 자료

실험에 사용된 참조자료는 2006년 촬영되고, 2007년에 갱신된 수치지도2.0 37709081도엽의 건물 레이어와, 이 도엽에 해당되는 2010년 8월 갱신된 도로명주소 기본도 건물 레이어를 목표자료로 사용하였다(표 3). 이때, 수치지도2.0 37709081도엽의 일부 영역이 실험 영역으로, 실험 영역 내에 존재하는 건물 객체는 수치지도2.0과 도로명주소 기본도에 각각 1,051개, 1,156개가 있다. 이들 건물 객체를 이용하여 제안된 방법을 적용하였다.

4.2 실험 결과

파일 포맷은 Shape파일, 좌표는 GRS80 타원체 TM좌표로 통일한 후, 수치지도2.0 건물 객체와 교차되는 도로명주소 기본도 건물 객체를 후보 매칭 쌍으로 선정하였다. 그 결과 수치지도2.0 건물 객체 1,051개 중에서 교차되는 도로명주소 기본도의 건물 객체가 없는 경우는 45개, 교차되는 건물 객체가 있는 경우는 1,006개로, 1:1로 매칭이 되는 경우가 745개, 1:N의 매칭관계가 261개로 나타났다. 결과적으로 1개의 수치지도2.0 건물 객체가 여러 개의 도로명주소 기본도의 건물객체와 중복 매칭되는 경우가 있어 후보 매칭 쌍으로 1,355쌍이 생성되었다.

CRITIC 방법을 적용하여 산출된 세 가지 연동 기준간의 상관계수(σ)와 가중치(ω)는 표 4와 같으며, 본 실험 자료의 경우는 면적 기준이 형상유사도를 결정하는데 중요하게 나타나고, 다음으로 형상 기준, 위치 기준의 순서였다. CRITIC 방법으로 유도된 후보 매칭 쌍의 형상유사도에 훈련 자료에서 구해진 임계값 0.797을 적용한 결과, 전

표 3. 수치지도2.0과 도로명주소 기본도

구 분	수치지도2.0	도로명주소 기본도
파일 포맷	NDI파일(공간), NDA파일(속성)	SHAPE 파일
좌표 체계	GRS80 타원체 TM좌표	Bessel 타원체 UTM-K좌표
갱신주기	권역별 주기갱신 (일반권역 4년, 광역도시권 2년)	수시갱신

표 4. CRITIC 방법에 의해 산출된 상관계수(σ)와 가중치(ω)

	위치 기준	형상 기준	면적 기준	σ	ω
위치 기준	1	0.54	0.83	0.120	0.203
형상 기준	0.54	1	0.46	0.106	0.284
면적 기준	0.83	0.46	1	0.271	0.513

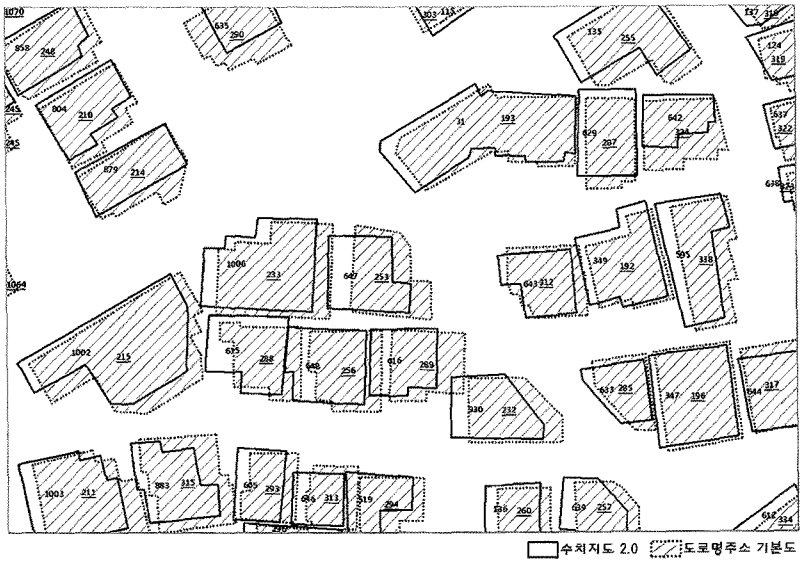

그림 3. 탐색된 매칭 쌍 중 정 매칭 쌍 일부 (숫자 아래 밑줄이 있는 것이 도로명주소 기본도의 인덱스)

체 1,335쌍의 후보 매칭 쌍 중 842쌍이 매칭 쌍으로 탐색되었다.

제안된 방법으로 탐색된 매칭 쌍에 대한 평가를 위하여, 본 연구에서는 실험 자료에서 수동으로 실제로 맞는 결과 즉, 정 매칭 쌍(846쌍)을 생성하였다. 탐색된 매칭 쌍 중에서 정 매칭 쌍에 787쌍이 포함되었으며, 수치지도2.0 건물 객체 중에서 도로명주소 기본도에서 교차되는 건물 객체가 있는 전체 1,006개 중에서 787개 즉, 78.2%의 객체가 도로명주소 기본도와 연동되었다. 이들 매칭 쌍을 시각적으로 살펴본 결과 형상이 유사하면서 교차되는 면적이 넓은 객체가 매칭이 된 것을 알 수 있었다. 특히, 건물이 밀집된 주택가에서도 비교적 매칭이 잘 수행된 것을 볼 수 있었다(그림 3).

다음으로 제안된 매칭 방법의 성능측정은 온톨로지 매칭의 결과를 평가하는 척도인 정확도(precision)와 재현율(recall)을 이용한 F-Measure로 수행하였다(Yatskevich 등, 2006, Giunchiglia 등, 2005). F-Measure는 정확도와 재현율을 같은 가중치를 두고 통합한 값으로 클수록 더 좋은 결과를 나타낸다(식 (7)).

$$F\text{-}Measure = \frac{\text{정확도} \times \text{재현율}}{0.5 \times \text{정확도} + 0.5 \times \text{재현율}} \quad (7)$$

$$\text{여기서, 정확도} = \frac{\text{탐색된 매칭쌍 중 수동추출된 정매칭쌍 개수}}{\text{제안된 방법에 의해 탐색된 매칭쌍 개수}}$$

$$\text{재현율} = \frac{\text{탐색된 매칭쌍 중 수동추출된 정매칭쌍 개수}}{\text{수동추출된 정매칭쌍 개수}}$$

제안된 매칭 방법의 경우 표 5와 같이 정확도, 재현율, F-Measure가 각각 0.935, 0.930, 0.932로 높게 나타났다. 다시 말하면, 제안된 매칭 방법이 정 매칭 쌍을 93%정도 탐색하고, 탐색된 매칭 쌍 중 93.5%정도가 정확한 매칭 쌍이라는 의미이다.

마지막으로 본 제안된 방법으로 매칭되지 않은 사례를 살펴보았다. 본 연구에서 가정을 하고 있는 즉, 기하학적 형상이 유사하지 않았으며(그림 4의 동그라미), 면 객체의 묘화나 갱신주기 등의 차이로 하나의 면 객체가 2개 이상의 객체와 교차되는 경우(그림 4의 나머지) 매칭이 되지 않는 것을 볼 수 있었다. 따라서 향후 하나의 면 객체가 2개 이상의 객체와 교차되는 경우에 대한 추가 연구가 진행될 필요가 있을 것이다.

표 5. 탐지된 매칭 쌍에 대한 통계적 평가

	제안된 방법으로 탐색된 매칭 쌍	수동 추출된 정 매칭 쌍	탐색된 매칭 쌍 중 정 매칭 쌍	정확도	재현율	F-Measure
매칭결과	842쌍	846쌍	787쌍	0.935	0.930	0.932

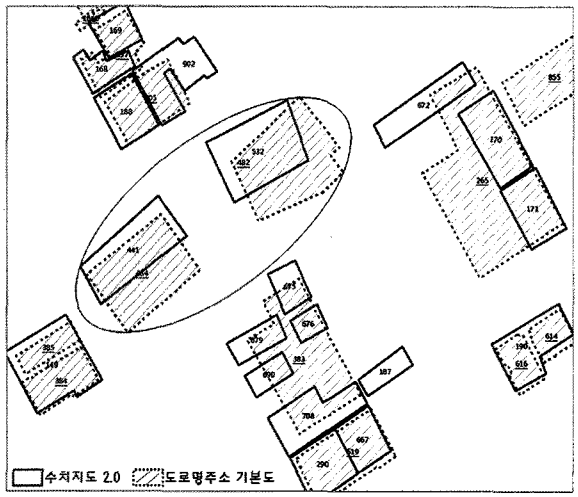

그림 4. 제안된 방법으로 매칭되지 않은 면 객체 일부

5. 결론 및 향후 연구

본 연구에서는 사용자 개입이 최소화된 면 객체의 기하학적 정보를 바탕으로 생성된 유사도 기반의 매칭 방법을 제안하였다. 이를 위하여 포맷과 좌표가 통일된 서로 다른 공간자료에서 교차되는 후보 매칭 쌍을 추출하고, 추출된 후보 매칭 쌍에 대하여 CRITIC 방법을 이용하여 자동으로 유도된 선형조합 즉, 형상유사도를 측정한다. 이때, 훈련 자료에서 조정된 상자도표의 특이점 탐색을 적용하여 도출된 임계값 0.797 이상인 경우가 매칭 쌍으로 탐색된다. 제안된 방법을 수지치도2.0과 도로명주소 기본도의 일부 지역에 적용하고, 시각적 평가와 통계적 평가를 수행하였다. 시각적으로 평가한 결과 형상이 유사하고 교차되는 면적이 넓은 건물 객체가 매칭된 것을 알 수 있었으며, 건물 밀도가 높은 주택가에서도 매칭이 잘 되는 것을 볼 수 있었다. 제안된 매칭 방법의 정확도는 0.935, 재현율은 0.930, F-Measure가 0.932로 높게 나타났다. 그러나 제안된 방법으로 매칭되지 않는 면 객체는 시스템의 구축 목적으로 인해 객체의 묘화방법이 상이하고, 다른 갱신주기로 하나의 객체가 다른 공간자료 여러 개와 대응이 되는 경우였다.

따라서 향후 형상유사도를 바탕으로 서로 다른 공간자료의 유사성을 판단하는 기준이 되는 임계값을 선정하는데 있어 다양한 매칭관계를 고려하여 제안된 방법의 향상을 도모해야 할 것이다.

감사의 글

본 연구는 국토해양부 첨단도시기술개발사업-지능형 국토정보기술혁신사업과제의 연구비지원(07국토정보C04)에 의해 수행되었습니다.

참고문헌

조재관, 최윤수, 권재현, 이보미 (2008), GIS 기본도 및 DB의 세계측지계 좌표변환 정확도 분석에 관한 연구, 한국지형공간정보학회지, 제 16권, 제 3호, 한국지형공간정보학회, pp. 79-85.

Ali, A. B. H. (2001), Positional and shape quality of areal entities in geographic databases: quality information aggregation versus measures classification", *Proceeding of ECSQARU '2001 Workshop on Spatio-Temporal Reasoning and Geographic Information Systems*, Toulouse, pp. 1-16.

Brys, G., Hubert, M. and Struyf, A. (2004), A robust measure of skewness, *Journal of Computational and Graphical Statistics*, Vol. 13, No. 4, pp. 996-1017.

Burghardt, D. and Steiniger, S. (2005), Usage of principal component analysis in the process of automated generalisation, *Proceedings of ICA Conference, A Coruña (Spain)*.

Cohen, W. W., Ravikumar, P. and Fienberg, S. E. (2003), A comparison of string distance metrics for name-matching tasks, *Proceedings of the IJCAI - 2003 Workshop on Information Integration on the Web (IIWeb-03)*, Acapulco, Mexico, pp. 73-78.

Diakoulaki, D., Mavrotas, G. and Papayannakis, L. (1995), Determing ovjective weights in multiple criteria problems: the CRITIC method, *Computers & Operational Research*, Vol. 22, No. 7, pp. 763-770.

Fu, Z. and Wu, J. (2008), Entity matching in vector spatial data, *Proceedings of the XXIth ISPRS Congress*, 3-11 Jul 2008 Beijing, CHINA, pp. 1467-1472.

Giunchiglia, F., Shvaiko, P. and Yatskevich, M. (2005), Semantic schema matching, *Proceedings of the 13th International Conference on Cooperative Information Systems*.

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*, 2008, Vol. 25, No.5, pp. 380-382.

Huang, L., Wang, S., Ye, Y., Wang, B. and Wu, L. (2010), Feature matching in cadastral map integration with a case study of Beijing, *Proceedings of 2010 18th International Conference on Geoinformatics*, Peking University, Beijing, China, pp. 1-4.

Hubert, M. and Vandervieren, E. (2008), An adjusted boxplot for skewed distributions, *Computational Statistics and Data Analysis*, Vol. 52, pp. 5186-5201.

Hubert, M. and Van der Veeken, S. (2008), Outlier detection for skewed data, *Journal of Chemometrics*, Vol. 22, pp. 235-246.

Liu, Z. (2006), *The research on areal feature matching among the conflation of urban geographic databases*, Master thesis, University of Wuhan, Wuhan.

Safra, E., Kanza, Y., Sagiv, Y. and Doytsher, Y. (2006), Integrating Data from Maps on the World-Wide Web, *Proceedings of the 6th International Symposium on Web and Wireless Geographical Information Systems (W2GIS)*, Hong Kong, China, (Springer), Lecture Notes in Computer Science, 4295, pp. 180-191

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.

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.

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.

Wenjing, T., Yanling, H., Yuxin, Z. and Ning, L. (2008), Research on areal feature matching algorithm based on spatial similarity, *Proceedings of Control and Decision Conference (CCDC 2008)*, China, pp. 3326-3330.

Yatskevich, M., Giunchiglia, F. and Avesani, P. (2007), A large scale dataset for the evaluation of matching systems, *Proceedings of ESWC 2007*.

Yuan, S. and Tao, C. (1999), Development of conflation components, *Proceedings of the International Conference on Geoinformatics and Socioinformatics*, Ann Arbor, Michigan, USA, pp. 1-13.

Zhang, Q. (2002), *Research on feature matching and conflation of geographic databases, PhD dissertation*, University of Wuhan, Wuhan.

(접수일 2011. 01. 31, 심사일 2011. 02. 28, 심사완료일 2011. 03. 02)