Korean Society of Surveying, Geodesy, Photogrammetry, and Cartography 2004 / Suwon / Korea

상대적 밀도를 이용한 LiDAR 데이터의 Outlier 검출
Outlier Detection from LiDAR Data based on the Relative Density

문지영[1] · 이임평[2] · 김성준[3] · 김경옥[4]
Moon, Jiyoung · Lee, Impyeong · Kim, Seongjun · Kim, Kyoung-ok

[1] 서울시립대학교 대학원 지적정보학과 석사과정(E-mail:venus@uos.ac.kr)
[2] 서울시립대학교 도시과학대학 지적정보학과 교수(E-mail:lplee@uos.ac.kr)
[3] 서울시립대학교 도시과학대학 지적정보학과 학사과정(E-mail:sidus7953@naver.com)
[4] 한국전자통신연구원 텔레매틱스연구단 공간정보연구팀 팀장(E-mail:kokim@etri.re.kr)

Abstract

LiDAR data often include outliers, the points being signficantly separated from other points and so seeming not to be measured from physical surfaces. Outliers should be removed before processing further the data for applications. Many methods have been developed for other data rather than LiDAR data as a part of data mining processes but their straightforward application to LiDAR data did not provide satisfactory results. In this study, we have thus modified one of such methods by considering the properties of LiDAR data and developed a method based on the relative point density. The proposed method have been applied to simulated and real data. The results confirms its promising performance with respect to the processing time and the detection accuracy.

1. 서 론

다수의 개체로 이루어진 데이터 집합에서 일부의 개체들은 밀집된 분포를 보이기도 하고 다른 일부의 개체들은 희박한 분포를 보이기도 하며 어떤 개체들은 다른 개체들로부터 확연히 분리된 위치에 있기도 한다. Outlier는 데이터의 군집에서 멀리 떨어져 존재하는 관측값인데 이것이 데이터의 통계적 분석에 강한 영향을 미쳐 소수의 Outlier로 인해 전체의 분석 결과가 왜곡될 수도 있다. 항공 LiDAR 측량은 지표면이나 지상의 사물의 물리적 표면에서 측정된 다수의 점들의 집합을 생성한다. 측정된 점들 중에는 실제 물리적 표면에서 측정된 점이라고 보기 어려운 이상점들이 간헐적으로 존재한다. 예를 들어, 지표면보다 확연히 아래에 있는 점이나 주위의 다른 점들에 멀리 떨어져 독립적으로 존재하는 몇 개의 점의 집합 등이 포함될 수도 있다. 이러한 이상점들을 LiDAR 데이터에 포함된 outlier라고 하며, LiDAR 데이터를 DEM의 생성, 건물의 추출, 영상과의 융합 등과 같은 여러 가지 용도의 활용을 위해 처리하기 전에 반드시 제거해야 한다.

LiDAR 데이터에서 outlier를 검출하는 알고리듬은 1) 데이터에 존재하는 모든 이상점을 outlier로 분류하고, 2) 이상점이 아닌 점을 outlier로 분류하지 않고, 3) 대용량의 점을 빠른 속도로 처리할 수 있어야 한다. 이에 본 연구는 이러한 세 가지 조건을 만족하는 outlier 검출 방법을 LiDAR 데이터의 고유한 특성을 면밀히 고려하여 개발하는 것을 목표로 한다. 이를 위한 본 연구의 개발 전략은 기존의 일반적인 데이터에서 outlier를 검출하는 방법을 조사하고, 조사된 방법 중에 공간 데이터에 가장 적합한 방법을 선택하고, 선택된 방법을 라이다 데이터 특성에 적합하도록 보완하는 단계로 구성된다. 본 논문은 이어서 2장에서 상대적 밀도를 고려한 outlier 검출 방법의 원리와 실용적인 근사 방법을 기술하고, 3장에서 검출방법을 구현하여 라이다 데이터에 적용한 실험 결과를 소개하고, 4장에서 결론 및 향 후 계획으로 마무리한다.

2. Outlier 검출의 원리와 실용적 근사방법

2.1 상대적 밀도를 고려한 outlier 검출의 원리

기존의 일반적인 데이터에 존재하는 outlier를 검출하는 연구는 통계적 분포에 기반을 둔 방법, 개체간의 거리를 이용한 방법, 개체의 분포 밀도를 이용한 방법 등이 있다. 이러한 여러 가지 방법 중에서 거리기반 접근법(Knorr 등, 1998)과 밀도기반 접근법(Breuning 등, 2000)은 LiDAR 데이터와 같은 공간 데이터에 적합한 방법으로 조사되었다. LiDAR 데이터는 일반적으로 지형의 형태에 따라 비균일한 점의 분포를 보이는데 이로 인해 동일한 데이터 셋이라도 한 점을 중심으로 주변점들의 분포로부터 계산한 국지적 점밀도가 중심점의 위치에 따라 크게 달라질 수 있다. 이로 인해 점의 국지적 밀도에 대한 절대적 기준으로 outlier 여부를 결정하는, 즉 국지적 밀도가 임계값 이하이면 outlier로 검출하는 방법은 만족할 만한 결과를 생성하지 못한다. 이러한 특성을 감안하여 중심점의 국지적 밀도를 주변점들의 국지적 점 밀도와 비교하는 소위 상대적 밀도를 기준으로 하는 방법을 선택하였다.

Papadimitriou 등 (2002)에 의해 개발된 MDEF (Multi-granularity DEviation Factor)는 Breunig 등 (2000)이 개발한 개체의 국지적 밀도인 LOF (Local Outlier Factor)의 복잡성을 해소하여 보다 빠르고 직관적인 계산을 가능하게 한다. MDEF는 중심점의 국지적 점밀도를 주변점의 국지적 점밀도와 정량적으로 비교하는 수치이다. 이를 정량적으로 정의하기 위해서 r-neighborhood와 αr-neighborhood의 개념을 도입한다. 그림 1에서 보는 바와 같이 3차원의 공간에서 한 점 p_i를 중심으로 반지름 r의 거리 안에 있는 점을 p_i의 r-neighborhood라고 하고, p_i로부터 $\alpha \times r$의 거리 이내에 포함되는 모든 점을 p_i의 αr-neighborhood라 한다. 여기서 α는 0에서 1사이의 값이며 2^{-b}(b는 자연수)의 값이 이용된다.

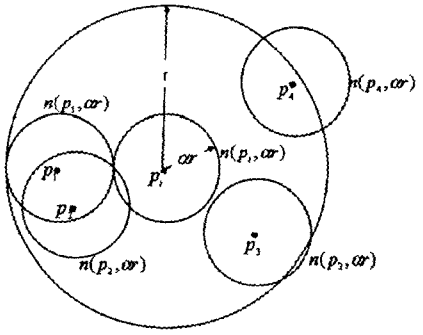

그림 1. r-neighborhood와 αr-neighborhood의 정의

p_i의 αr-neighborhood에 속하는 점의 개수를 $n(p_i, \alpha r)$로 표현한다. p_i의 r-neighborhood가 $\{p_1, p_2, \ldots, p_m\}$처럼 m개의 점을 포함한다고 하면, 각각의 점으로부터 다시 m개의 αr-neighborhood를 정의하여 포함되는 점의 개수의 평균값을 구하여 $\hat{n}(p_i, \alpha, r)$이라고 한다. 최종적으로 한 점 p_i의 MDEF는 다음과 같이 정의된다.

$$MDEF = 1 - \frac{n(p_i, \alpha r)}{\hat{n}(p_i, \alpha, r)} \tag{1}$$

MDEF는 $n(p_i, \alpha r)$이 $\hat{n}(p_i, \alpha, r)$에 비해 아주 작을 경우, 즉 p_i의 국지적 밀도가 p_i의 r-neighborhood로 정의되는 주변점의 국지적 밀도보다 현저히 낮은 경우에는 1에 가까운 값을 가지게 된다. 결국 라이다 데이터 셋의 모든 점에서 개별적으로 MDEF를 계산하여 1에 가까운 값을 가지면 이

러한 점들은 국지적 밀도가 상대적으로 주변점들의 국지적 밀도보다 낮기 때문에 outlier로 검출한다. 그렇지만, 이러한 방법은 모든 점에 대해서 $\hat{n}(p_i, \alpha, r)$를 계산해야 되기 때문에 시간이 오래 걸린다는 단점이 있다. 이를 보완하기 위한 실용적인 근사방법을 이어서 기술한다.

2.2 상대적 밀도를 고려한 Outlier 검출의 실용적 근사방법

2.2.1 Grid 데이터 구조

위의 검출 원리를 기본으로 하여 전체 데이터셋에 대해 빠른 계산을 하기 위해 점들을 계층을 가지는 여러 개의 grid로 저장한다. Cell의 크기가 d인 것부터 2배씩 증가시켜 결국 {d, $2d$, $4d$, $8d$, …}의 cell의 크기를 갖는 grid로 생성된다. 최하위 계층 grid의 cell 간격인 d를 합리적으로 결정하는 방법은 3.1절에서 소개하도록 한다. 최하위 계층의 grid가 한 방향으로 2^n개의 cell을 가진다면 한 단계의 상위 grid가 만들어질 때마다 cell의 크기는 두 배가 되고 한 축의 cell의 개수는 2배씩 감소하여 그림 2에서 보는 바와 같이 전체적으로 총 n개의 계층의 grid pyramid를 구성하게 된다. 최하위 계층의 grid는 각각의 cell의 범위에 포함된 점들에 대한 링크와 점의 개수를 저장하고, 그 외의 상위 계층의 grid는 각각의 cell의 범위에 포함된 점의 개수를 저장한다.

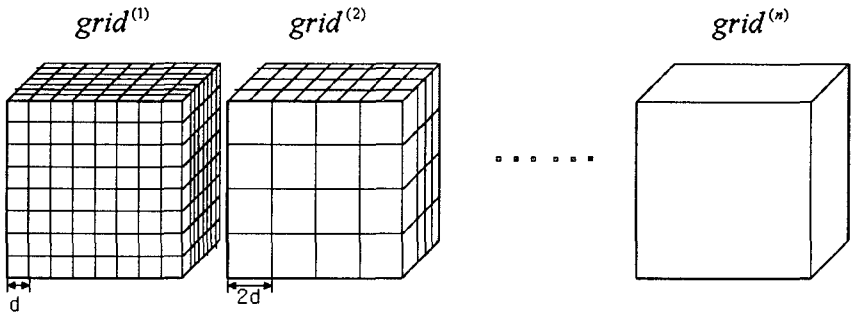

그림 2. Grid Pyramid

2.2.2 $\hat{n}(p_i, \alpha, r)$의 근사적 계산

만약 충분히 가까운 거리에 위치한 두 점(p_i, p_j)를 가정하면, 각각의 점에서 정의된 r-neighborhood는 상당히 유사하고 이로 인해 이러한 점들로부터 계산한 αr-neighborhood의 평균값인 $\hat{n}(p_i, \alpha, r)$, $\hat{n}(p_j, \alpha, r)$는 거의 동일한 값을 갖는다. 즉 가까운 거리에 위치한 점들은 서로 동일한 $\hat{n}(p_i, \alpha, r)$를 갖는다고 가정을 통해 개개의 점마다 독립적으로 계산하지 않음으로써 계산량을 크게 줄일 수 있다. 이러한 가정을 통해 $\hat{n}(p_i, \alpha, r)$에 대한 근사값을 구하는 과정은 아래와 같다.

충분히 가까운 거리에 위치한 점들이라 함은 2.2.1절에서 소개한 하나의 grid에서 동일한 cell안에 포함되는 모든 점들을 말하며, 이들의 αr-neighborhood는 점이 속한 cell안의 점들로 근사적으로 가정할 수 있으며 이로 인해 모두 같은 개수를 가진다. 원래의 정의에서 αr과 r은 반지름이 α배 차이가 나는 범위이다. α가 2^{-b}이므로 r-neighborhood에 포함되는 점의 개수는 αr-neighborhood보다 b단계 상위 grid의 cell안의 점의 수라고 한다. 따라서 $\hat{n}(p_i, \alpha, r)$의 근사값은 r-neighborhood범위 안의 α r-neighborhood의 점의 개수의 평균이다. 예를 들면, b단계 상위 grid의 한 cell에 속한 8^b개의 cell에서 각 cell의 점의 개수를 $\{C_1, C_2, \ldots\ldots C_{8^b}\}$라 하면 평균값은 식 (2)으로 구해진다. 즉, grid$^{(1)}$ 의 평균값은 grid$^{(b+1)}$의 cell을 기준으로 계산이 되고 grid$^{(n-b)}$의 평균값은 grid$^{(n)}$를 기준으로 계산되어 n-b개 grid의 모든 cell들이 $\hat{n}(p_i, \alpha, r)$의 근사값을 가지게 된다.

$$n(p_i, \alpha, r) \approx \tag{2}$$

하나의 점은 n개의 grid에 모두 속하고 이 점을 중심으로 각 계층의 grid마다 $\hat{n}(p_i, \alpha, r)$이 근사된다. 점에서 반지름 r의 거리에 포함되는 개체의 수인 $n(p_i, \alpha r)$과 근사된 $\hat{n}(p_i, \alpha, r)$을 이용하여 식 (1)에서 정의된 MDEF를 구할 수 있다. 이때 $n(p_i, \alpha r)$의 r은 $\hat{n}(p_i, \alpha, r)$의 grid의 cell간격을 2로 나눈 값이다. MDEF는 $\hat{n}(p_i, \alpha, r)$과 마찬가지로 하나의 점에서 n-b개가 구해진다. 이때 점이 하나도 없는 빈 cell들은 무시하고 점이 하나라도 들어있는 cell들만을 계산에 포함시키도록 한다. 이로써 전체 영역의 대부분을 빈 cell이 차지하고 있어도 평균이 작아지지 않아 outlier를 찾아낼 수 있다.

2.2.3 저밀도 점들의 분류

소수의 점이 데이터의 군집에서 떨어져있는 outlier는 높은 MDEF를 가진다. 하지만 여러 개의 점이 모인 작은 군집이라도 데이터셋 안에 큰 군집이 있을 경우에는 작은 군집에 포함된 점은 1에 가까운 MDEF를 가질 수 있다. LiDAR 데이터에서는 이러한 점들이 자연 지형지물이나 인공시설물의 데이터일 경우에 많이 나타난다. 따라서 이러한 점들의 검출을 배제시키기 위해 실제 검출을 원하는 outlier가 존재할 것으로 예상되는 저밀도 점들만을 분류한다.

Grid pyramid의 가장 하위 계층에서 한 cell안에 3점 이상인 경우와 3점 미만인 경우로 나누어 분류한다. 이 때 3점 이상인 cell에서는 MDEF가 높더라도 outlier가 아닌 것으로 간주한다. 따라서 하나의 셀에 3점 미만이 존재하는 저밀도 점들을 분류하여 이러한 점들에 대해서만 계산을 수행함으로써 outlier 검출에 소요되는 시간과 복잡성을 대폭 감소시킬 수 있다.

3. Outlier 검출 방법의 구현, 적용 및 토의

3.1 Outlier 검출 방법의 변수의 설정

앞서 제안된 Outlier 검출 방법을 적용하기 전에 주어진 LiDAR 데이터셋의 특성을 고려하여 세 가지 변수를 설정하여야 한다. 먼저 grid의 cell 간격 d를 결정해야 한다. cell의 간격은 건물이나 지상의 의미 있는 물체보다는 작아야 하지만 cell이 많아질수록 복잡성이 증가하므로 계산의 속도를 유지하기 위해서는 지나치게 작은 값은 피하는 것이 좋다. cell 간격은 2.5 m와 가까운 수이면서 데이터셋의 z 방향의 범위를 2의 거듭제곱으로 나눈 값으로 설정한다. z 방향으로 cell의 간격을 제한하는 것은 매 단계마다 2배씩 cell의 수가 줄어가는 grid 구조를 구성할 때 가장 상위 grid에서 한 개의 cell만 존재하게 하기 위함이다.

다음으로 α를 결정해야 하는데 α는 αr-neighborhood와 r-neighborhood의 거리의 비를 결정하는 상수이다. 이것은 데이터의 분포 특성과 사용자가 요구하는 outlier 검출의 감도를 고려하여 결정하여야 한다. 이 연구에서는 실험을 거쳐 α를 0.25로 사용하였다.

마지막으로 결과에 직접적으로 영향을 끼치는 임계값을 결정하여야 한다. 임계값은 outlier 여부를 결정하는 기준이 되는 MDEF 수치이다. 실험에서 $\hat{n}(p_i, \alpha, r)$이 $n(p_i, \alpha r)$의 1000배 이상인 점을 찾아내기 위해 임계값을 0.999로 하였다. 예를 들어, 33000/s의 간격으로 측정을 하는 LiDAR 장비가 고도 1000 m로 측량을 할 때 약 2000/㎡의 점밀도를 가진다. 0.999의 임계값은 2000/㎡의 밀도를 갖는 데이터에서 1㎡에 1개 또는 2개 정도의 점이 있을 때 outlier로 검출하는 정도이다. 2.2.2절에서 한 점에 대해 n-b개의 MDEF를 구하는데 하나의 MDEF라도 임계값보다 크다면 outlier로 결정하였다.

3.2 Outlier 검출 방법의 구현

Outlier 검출 방법은 grid의 생성, grid pyramid 구현, outlier 검출 등을 개별적인 클래스로 생성하여 C++을 이용하여 객체 지향 프로그램으로 구현하였다. 3차원 좌표형태의 결과에 대한 시각화는 Matlab을 활용하였다. 그림 3은 구현된 검출 방법을 간략하게 설명한 흐름도이다.

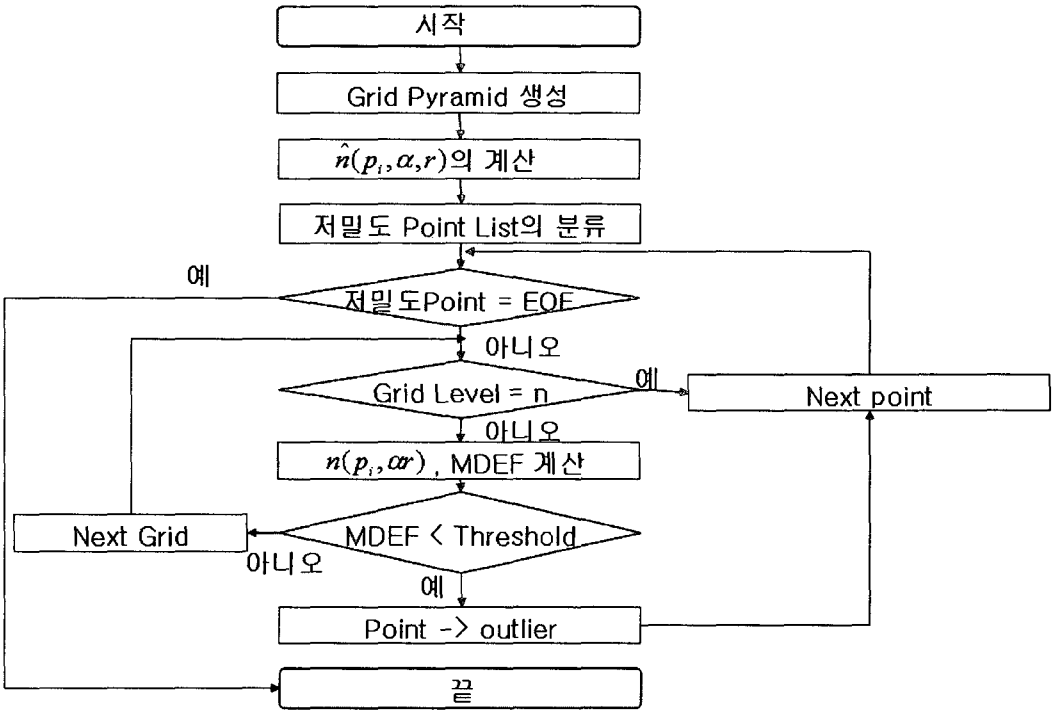

그림 3. 구현된 검출 방법의 흐름도

3.3 Outlier 검출 방법의 적용 결과 및 토의

3.3.1 가상 데이터에 적용

실험에 사용한 가상 데이터는 10000개의 점 데이터로서 각각 3차원 좌표를 가지고 있다. 실제 지형을 단순화하여 약간의 기울기를 가진 평면으로 가정하여 점들을 생성한 후 여기에 임의의 이상점을 20점정도 추가하였다. 실험 결과 23개의 점이 outlier로 검출되었는데, 여기에는 outlier로 의도적으로 추가된 20개와 평면 내에서 독립도가 높은 3개의 점이 포함되었다. 그림 3에서 실선으로 표시된 영역에 outlier로 검출된 점이 분포한 영역이다. 개체의 군집과 떨어진 모든 점들이 outlier로 검출된 것을 볼 수 있다.

3.3.2 실측 데이터에 적용

이러한 방법을 약 5만점 정도의 실측데이터에 적용시킨 결과 그림 4와 같이 89개의 점이 0.999이상의 MDEF를 갖는 outier로 검출되었다. 어떤 점에 대해 outlier로 볼 것인가 아닌가에 관해서는 데이터의 사용자가 판단해야 할 문제이므로 실험 결과의 정확도를 제시하기는 힘들지만 결과를 그림 4처럼 시각화하였을 때 지면에서 멀리 떨어진 점이 검출된 것을 볼 수 있었다.

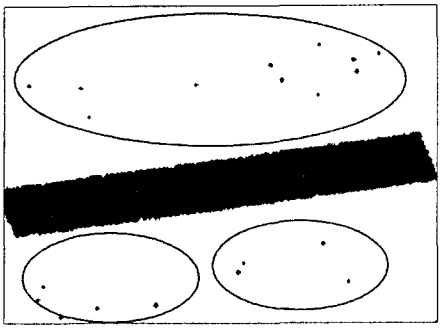

그림 3. 가상 데이터에 적용 결과

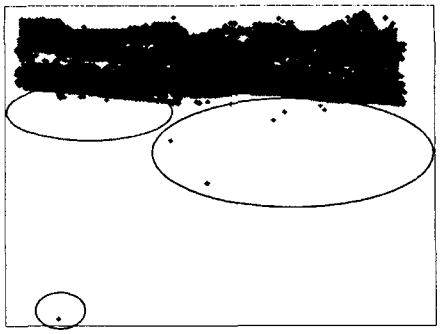

그림 4. 실측 데이터에 적용 결과

4. 결 론

위의 두 실험에서 계산에 소요되는 시간은 둘 다 1초 미만이었다. 더 큰 용량의 데이터를 처리할 경우 검출에 필요한 시간은 증가하겠지만 이 연구의 알고리듬은 선형적이므로 대용량 데이터에 적합하다고 할 수 있다. 또한 이 실험에서는 데이터의 손실을 최소화하고자 큰 임계값을 사용하였지만 데이터의 특성이나 사용자의 요구에 맞춰 임계값을 다르게 한다면 다른 종류의 데이터에도 적용 가능할 것이다.

향후 과제로써 해당 지역의 DEM이 확보되거나 LiDAR 데이터로부터 지면점을 추출한 경우 DEM이나 지면점으로부터 일정 범위를 outlier 대상에서 제외시킨다면 더욱 빠르고 효율적인 활용이 기대된다.

감사의글

이 논문은 2003년도 서울시립대학교 학술연구용 첨단장비 지원에 의하여 이루어진 것이며 이에 학교 당국에 감사드립니다. 본 연구에 많은 도움을 주신 인하대학교 토목공학과 조우석 교수님과 한국전자통신연구원 텔레매틱스연구단 공간연구정보팀 이영진 연구원에게 감사드립니다.

참고문헌

Becker, C. and Gather, U. (1997), The masking breakdown point of multivariate outlier identification rules, Department of statistics, University of Dortmund.

Knorr, E. M. and Raymond, T.Ng. (1998), Algorithms for mining distance-based outliers in large datasets, VLDB conference, New York.

Breunig, M. M., Kriegel, H. P., Raymond, T. Ng. and Sander, J. (2000), LOF: Identifying density-based local outliers, SIGMOD/PODS conference, Texas.

Papadimitriou, S. and Kitagawa, H. (2002), LOCI : Fast outlier Detection using the local correlation integral. IRP, TR.