|  |
| --- |
| Knn알고리즘을 통한  서류 자동 인식에 관한 연구 |
|  |
| 김종은  동덕여자대학교 경영학과(빅데이터 연계전공)  e-mail : jongeun0122@gmail.com |
|  |
| A Study on KNN Algorithm  for Automatic recognition of documents |
|  |
| Jong-Eun Kim  Dept. of Business Administration (with BigData), Dongduk Womens University |
|  |
| **요 약**  KNN알고리즘, 인공지능, 이미지 자동인식, 자동인식, 서류 자동인식 |

**Ⅰ. 서 론**

최근 인공지능(AI)을 활용한 연구가 각 분야에서 활발히 진행되고 있으며, 컴퓨터 비전 분야

또한 인공지능이 빈번하게 활용되고 있는 것을 볼 수 있다.

특히, 여러 이미지의 집합 속에서 유의미한 특징(feature)을 추출해야 하는 자동인식 분야에서 인공지능은 처리 속도를 개선하고 모델을 생성해 더욱 효율적인 결과로 이끈다.

손 글씨 인식(MNIST), 차량 인식과 이미지 자동 인식 분야에서 기본적으로 활용되는 머신러닝 알고리즘으로는 KNN(K-Nearest Neighbor) 알고리즘이 있다.

따라서 스마트 항만 서류 자동인식에서 활용한 KNN 알고리즘에 대해 논하도록 하겠다.

**Ⅱ. 본론**

1. **이미지 처리와 KNN(K-Nearest Neighbor)**

KNN(K-Nearest Neighbor) 알고리즘은 데이터 내 벡터가 뿌려져있다고 가정할 때, 하나의 벡터를 기준으로 임의로 설정 거리 값 K에 해당하는 가까운 거리에 위치한 벡터끼리 묶는다. 이때 거리 계산 시 사용하는 수식으로는 유클리디언 거리 (Euclidean distance) 와 맨하탄 거리 (Manhattan Distance) 등 다양한 방법이 있다.

KNN 알고리즘은 위와 같은 거리 공식을 반영해 묶여진 집합들을 바탕으로 모델을 생성하고, 새로운 벡터가 들어왔을 때 기존 모델을 기반으로 새로운 벡터를 분류 및 예측하는 매커니즘이다.

이러한 원리를 갖고 있는 KNN 알고리즘은 패턴 분류에 활용되는 가장 기초적인 통계 알고리즘이라 할 수 있다.

1. **이미지 처리 과정**

KNN알고리즘은 모델 생성 및 학습 단계에서 활용되며 전체 단계를 정리해보면, 다음과 같다. 1) 항만 서류 텍스트 이미지를 온라인으로 업로드한 후 .jpg 혹은 .png 이미지 형태로 저장 및 배열화(change to matrix) 시킨다.

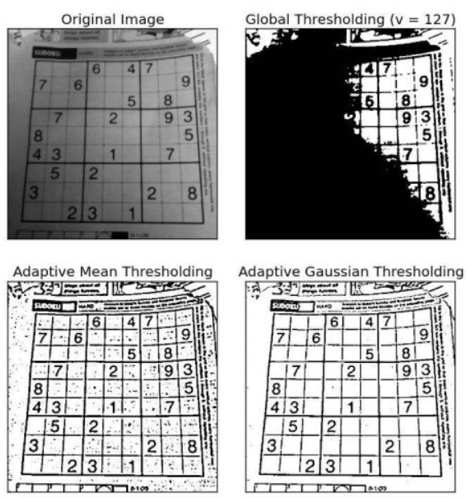
2) 저장한 이미지를 이진화(binary) 한다

3) 이진화한 이미지로 segmentation 및 분류 알고리즘을 만든다.

4) KNN 알고리즘을 활용해 분류

1) & 2) 이미지 업로드 및 이진화

항만 서류 이미지는 차량 자동 인식, 얼굴 자동 인식 등과 달리 픽셀 값의 대비가 뚜렷하며 노이즈가 적다. 따라서 사용하는 필터 또한 노이즈 제거보다는 윤곽선 검출(Edge Detection)에 효과적인 소벨필터(Sobel Filter), 케니필터(Canny Filter)와 임계치(Threshold)를 활용한 윤곽선 검출 필터인 적응임계처리(adaptive Threshold)를 활용해 윤곽선(Edge)을 검출했다.[그림1],[그림 2]



4) KNN 알고리즘을 통한 분류

이번 연구에서는 유클리디언 거리를 사용했다. 그 이유는 유클리디언 거리가 이미지의 너비(width) 및 길이(length)만 필요로 하며 거리 계산을 하기 때문이다.[1]

= , for all ,

[수식] 유클리디언 거리

KNN 알고리즘을 활용하여 데이터셋를 학습시킨 후 나온 학습률은 약 90.2%로 좋은 성능을 기록한다.

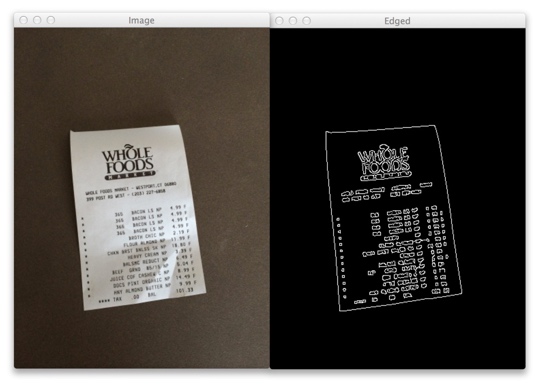
참고문헌

[1] Kolountzakis, Mihail N., and Kiriakos N. Kutulakos. "Fast computation of the Euclidian distance maps for binary images." *Information processing letters* 43.4 (1992): 181-184.

[2] <https://webnautes.tistory.com/1254> - OpenCV Pyhthon 강좌-영상이진화,adaptive threshold 이미지

[3] <https://www.pyimagesearch.com/2014/09/01/build-kick-ass-mobile-document-scanner-just-5-minutes/>,How to Build a Kick-Ass Mobile Document Scanner in Just 5 Minutes, canny filter 이미지

[그림1] adaptive threshold 예시 [2]



[그림2] canny filter예시 [3]

3) segmentation

특징 추출을 위해 주성분분석(Principal Component Analysis;PCA)과 독립성분분석(Independent Component Analysis;ICA) 을 사용했으며 20,000 개의 한글 서류 데이터 셋을 학습시켰다.