

분할 특징 추출에 의한 양식 문서의 분류

정현철*, 이종현*, 최영우**, 김재희*

* 연세대학교 전기 및 컴퓨터 공학과

** 숙명여자대학교 전산학과

Classification of Form-based Documents

by Partitioned Feature Extraction

Hyunchul Jung, Jonghyon Yi, Yeongwoo Choi, Jaihie Kim

Intelligent Vision Lab. Department of Electrical & Computer Engineering, Yonsei University

Department of Computer Science, Sookmyung Women's University

hcjung01@hotmail.com

Abstract

Specially, form-based documents are easily understood, quickly processed and thus used more than the general documents. In this paper, a method to classify the documents with minimum features is proposed, not like former methods which use all possible features. To apply this characteristics, a document was first partitioned to areas of certain shape and size, then features were extracted from the partitioned area. It is also possible to sort the partitioned area by using the fact that each partitioned area has the different significance in the point of feature. In conclusion, by using proposed method of extracting features from partitioned document, the processing time decreases due to search area reduction.

1. 서론

문서 종류의 다양함과 양의 증가로 인해, 자동으로 문서를 처리하려는 연구의 중요성이 부각되고 있다. 특히 일정한 형태를 가지고 있는 양식 문서는 정보의 전달과 이해에 효과적이기 때문에 문서 중에 차지하는 비율이 크다. 본 논문의 연구 대상인 양식 문서의 분류란 문서마다 다른 위치에 있는 데이터를 인식 등의 목적으로 처리할 경우 문서의 양식이나 형태 정보 등을 알아내는 과정이라 할 수 있다. 본 논문에서 제시한 분할 특징 추출에 의한 문서 분류 방법은, 문서 내의 모든 특징(feature)이 추출된 다음에 문서를 분류하는 기준의 방법과는 다르게, 문서를 여러 영역으로 분할하여 영역 단위로 특징 추출이 가능하여 일정 개수의 영역의 특징 정보로 문서의 분류가 충분히 되었다고 판단되면 더 이상 특징을 추출할 필요가 없어지므로 특징 추출에 소요되는

시간을 줄일 수 있다.

또한 문서를 분할하여 특징을 추출함으로 영역 중에는 선분이 있는 영역도 있고, 선분이 없는 빈 영역도 있을 수 있다. 선분이 있는 영역은 문서의 선분 특징을 나타내며, 선분이 없는 빈 영역은 그 위치의 영역에는 선분이 없다고 볼 수 있고, 따라서 빈 영역 자체로의 특징으로 이용 가능하다. 기존의 문서 분류 방법에서 사용된 특징은 선분 특징(Segment-oriented)이나 영역 특징(Region-oriented) 중에 하나만을 이용하였으나 본 논문에서의 문서 분류는 영역을 분할하여 선분과 영역 특징을 병행하여 사용한다. 선분 특징만을 이용한 기존의 방법으로는 가로선과 세로선을 추출한 다음 문서를 분류하는 방법[1]과 연상 그래프(Association graph)를 이용하는 방법[2] 등이 있으며, 영역 특징을 주로 이용하는 방법은 선분들로 둘러싸인 박스를 이용하는 방법[3]이 있다.

문서의 영역마다 가지고 있는 특징의 두드러짐이 다르다는 점을 이용하여 본 논문에서는 특징의 두드러짐을 기준으로 영역을 정렬하였고, 정렬된 분할 영역의 위치에서 우선적으로 특징 비교를 하여 문서의 모든 특징을 추출하지 않아도 문서 분류가 가능하다는 것을 기술하였다. 문서를 분류하는 과정에서 처리 시간의 대부분이 특징을 추출하는 과정에서 소요된다는 점을 고려할 때, 본 논문에서 제안한 분할 영역 단위에서의 특징 추출 방법이 효과적임을 알 수 있다.

2. 영상 전처리

2.1 이진화

본 논문에서는 스캐너를 사용하여 입력받은 신용카드 매출전표의 그레이 영상을 대상으로 하며 처리의 단순화를 위해 이진화를 한다. 영상들의 대부분이 부분적인 영역에서 영상의 밝기 변화가 작기 때문에 전체 영상

분할 특징 추출에 의한 양식 문서의 분류

에 하나의 임계값만을 결정하여 이진화를 수행하는 전역적 이진화 방법을 적용하였다.

본 논문에서 사용된 전역적 이진화 방법 중에 하나인 p-tile 이진화 방법은 영상의 전체 화소에서 일정한 비율까지를 개체 화소로 간주하여 임계값을 결정한다. 입력 영상의 개체 화소의 비율이 p 라 할 때,

$$p = \sum_{i=0}^t P(i) \text{를 만족하는 } t \text{가 임계값이며, } P(i) \text{는 } i \text{에서의 영상의 밝기 값의 빈도이다.}$$

2.2 기울기 보정

본 논문의 연구 대상인 전표 영상은 해상도가 낮고 다양한 형태의 잡영이 포함되어 있다. 또한 해상도가 낮기 때문에 선분의 끊어짐이 많이 발생한다. 이러한 특성을 고려하여 RLS 결과 영상에 Hough 변환을 적용하여 기울기를 보정하였다. RLS은 끊어진 선분을 연결하는 역할을 하고 Hough 변환은 문서 내에 가장 뚜렷한 선분의 기울어짐을 이용해 문서의 기울기를 구할 수 있다라는 장점이 있다.

3. 분할 특징 추출에 의한 문서 분류

문서를 분류하기 위한 기준의 방법은 선분이나 영역 특징을 추출할 때 문서의 모든 특징을 추출한 다음 등록 문서와 특징 비교를 한다. 그렇지만 문서 분류는 비교 대상의 등록 문서의 특징 공간으로 제한되어 있다는 특성이 있다. 따라서 본 논문에서는 문서 분류의 특성을 고려하여 처리 시간을 줄일 수 있는 분할 특징 추출에 의한 문서 분류 방법을 제안한다.

3.1 문서 영역 분할

분할 영역에서의 특징 추출 및 비교를 위해, 문서를 여러 영역으로 분할하였다. 하지만 분할하는 영역의 모양이 사각형일 경우 가로선과 세로선의 특징이 사각형의 각 변의 방향과 일치하기 때문에 문서의 이동(rigid movement)에 의한 특징 추출의 어려움 발생하면 분할 영역에서의 특징이 크게 달라질 수 있다. 그러므로 그림 1과 같이 분할 영역의 모양을 변의 기울기가 45°인 마름모로 설정하였다.

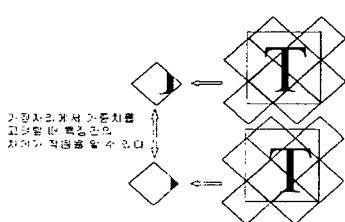


그림 1. 마름모 모양의 분할 영역

마름보는 가장자리에서 좁아진다는 특성이 있어 폭

에 따라 가중치를 설정한다면 사각형의 분할 모양에서 생기는 문제점을 극복할 수 있다. 그리고 분할 영역의 크기는 문서 내의 문자나 선분의 최소 거리의 특성 등을 고려해 설정하였다.

3.2 분할 영역에서의 특징 추출

전체 문서의 영역에서 선분을 추출하는 방법과는 달리 본 논문에서 사용한 분할 영역 안에서의 선분은 분할 영역의 크기가 작기 때문에 탐색 공간도 줄어들어 간단한 선분 추출 알고리듬을 이용해 추출할 수 있고 결과적으로 전체 처리 시간을 줄일 수 있다.

가로선의 추출 방법과 세로선 추출 방법이 비슷하고 단지 처리 방향만 다르기 때문에 여기서는 가로선 탐색 및 추출 방법에 대해서만 기술하겠다. 그럼 2처럼 마름모의 좌우 경계선과 3pixel의 두께로 설정된 탐색 영역을 위에서 아래 방향으로 이동해 가면서, 탐색 영역 안에서 선분의 유무를 결정한다. 탐색 영역 안에서 선분의 유무는 끊어진 구간의 개수와 탐색 영역의 좌우 절이와 비례한 임계값보다 작은 경우 탐색 영역 안에는 선분이 있는 것으로 판단하고 임계치보다 큼 경우는 없는 것으로 판단하였다.

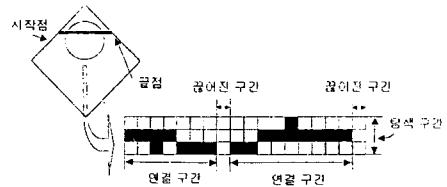


그림 2. 선분 추출 방법

3.3 분할 영역의 거리값 결정 및 문서 분류

비교 대상의 분할 영역에서 거리값은 영역 내에 선분의 유무와 개수에 관계없이 결정할 수 있어야 한다. 비교 대상의 분할 영역에서 오직 하나의 선분만이 존재할 때 특징의 거리값(M)을 구하는 방법은 그림 3과 같다. 선분이 하나일 경우 특징 거리값(M)은 두 선분 사이

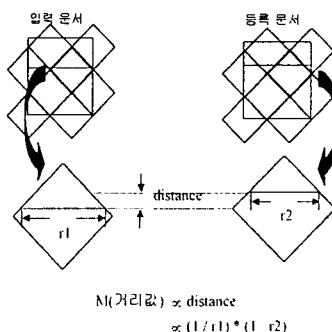


그림 3. 선분이 있는 경우
분할 영역에서의 거리값(M)

이의 거리와 비례하고 두 선분의 길이의 곱과는 반비례하도록 설정하였다. 하지만 r_1, r_2 값이 0으로 접근하면 특정 거리값이 발산할 수 있기 때문에 식(1)을 이용하였다.

$$M = \frac{d}{(r_1+1)*(r_2+1)} \quad (1)$$

d : 비교하고자 하는 선분의 distance,

r_1, r_2 : 선분의 비율

그리고 비교 대상의 분할된 영역 안에 여러 개의 선분이 있을 경우의 특정 거리값은 선분별로 비교하고자 하는 선분과 가장 가까운 선분을 선택하고, 식 (1)와 같은 방법으로 산출한 특정 거리값을 더한 후에 분할 영역 내의 선분 개수로 나누어 정규화하였다.

또한 선분이 있는 영역과 없는 영역간의 특정 거리값은, 선분이 없는 영역에서 마름모의 가장자리에 선분의 길이가 0인 가상의 선분을 설정한 다음 식(1)을 이용하여 산출하였다.

3.4 분할 영역에서의 변별력 설정

문서의 분할 영역마다 다른 문서와 구별짓는 특징의 두드러짐이 다를 수 있다. 이러한 특징의 두드러짐을 분할 영역에서의 변별력이라 정의하였으며, 분할 영역별 변별력을 식(2)을 이용하였다.

n번째 등록 문서의 분할 영역(i, j)에서의 변별력 $E_n(i, j)$ 는

$$E_n(i, j) = \sum_{k=0}^{k=D-1} M_k^2(i, j) \quad (2)$$

D : 총 등록 문서의 개수

$M_k(i, j)$: n번째 등록 문서와 k번째 등록 문서의 분할 영역(i, j) 거리값(M)

($0 \leq n \leq D - 1, 0 \leq k \leq D - 1$)

식 (2)에서 알 수 있듯이 다른 등록 문서와의 특정 거리값이 큰 영역일수록 분할 영역(i, j)에서의 변별력이 커지며 따라서 특징의 두드러짐도 커짐을 알 수 있다. 그러므로 변별력이 큰 영역에서 우선적으로 분할 영역의 거리값을 산출하면, n번째 문서와 동일한 양식을 띠고 있는 문서일 경우, n번째 등록 문서와의 거리값은 작게 증가하고 다른 등록 문서의 거리값들은 크게 증가함으로써 입력 문서가 n번째 등록 문서와 동일한 양식임을 좀 더 빠르게 알 수 있다.

3.5 분할 특징 추출의 순서 정렬

모든 분할 영역들을 사용하여 입력된 문서를 분류할 때는 등록 문서들마다 산출된 모든 분할 영역의 거리값을 더해서 최소가 되는 등록 문서를 선택한다. 최소의 분할 영역만을 이용하여 문서를 분류하기 위해선 가능한

빨리 입력 문서와 동일한 등록문서의 영역 거리값의 합이 다른 등록 문서마다의 영역 거리값의 합보다 충분히 작아져야 한다. 그러므로 입력과 다른 등록 문서의 거리값의 증가는 클수록 그리고 동일한 등록 문서의 거리값의 증가는 작을수록, 적은 분할 영역의 개수를 이용해 서도 문서를 분류할 수 있다. 분할 영역에서의 변별력이 클수록 입력과 동일한 등록 문서의 거리값과 다른 등록 문서에서의 거리값이 평균적으로 크게 차이가 난다.

분할 영역의 위치를 선택하기 위해서는, 먼저 등록 문서별로 변별력에 따라 정렬된 분할 영역의 위치 정보를 저장할 큐(queue)가 필요하다. 그러한 큐는 그림 4와 같이 현재 영역의 위치에서 산출한 거리값을 이용해 다음 특징을 추출할 분할 영역의 위치를 얻는데 사용되어 진다. 임의의 분할 영역의 위치 $p_k = (i_k, j_k)$ 라고 할 때, 다음 특징을 추출할 분할 영역의 위치 p_{k+1} 라 하자.

i) $p_0 = (0, 0)$ 이다.

ii) $p_k = (i_k, j_k)$ 라고 하면,

$$\begin{aligned} p_{k+1} &= (i_{k+1}, j_{k+1}) \\ &= Get(Q_d) \end{aligned}$$

$$(d = index(\min\{\sum M_0, \dots, \sum M_{l-1}\}))$$

$M_i : (i_k, j_k)$ 에서 입력과 i번째 등록 문서와의 거리값

$Get(Q_d) : d$ 번째 등록문서의 큐에서 영역의 위치 획득
($0 \leq d \leq D - 1$)

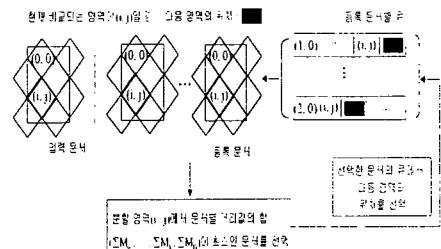


그림 4. 분할 특징 추출의 순서 설정

즉, 현재 위치까지 구한 분할 영역의 거리값의 합이 최소가 되는 등록 문서를 찾은 다음, 찾은 등록 문서의 큐에서 다음 특징의 추출 위치를 얻는다.

그림 4에서 등록 문서별로 저장된 큐에는 변별력의 크기에 따라 순서대로 분할 영역의 위치가 저장되어 있기 때문에 특징의 두드러짐이 큰 영역을 우선적으로 추출하고 비교함으로써 초기에 최소 거리값을 나타내는 문서는 결국에도 최소 거리값을 유지하도록 한 것이다. 그러므로 분할 영역의 특징 추출의 순서를 설정하고 그 순서에 따라 특징을 추출하여 문서를 분류하면 모든 특장을 추출할 필요가 없다. 일부의 분할 영역의 특징을 추출하여도 비교하여도 모든 영역을 이용하여 특징을 분류할

때의 문서분류의 에러율과 거의 비슷하다는 것을 본 논문에서는 실험적으로 검증하였다.

4. 실험 및 결과

IBM PC 호환 기종 Pentium Pro(233 MHz)에서 Visual C++언어로 구현하여 실험하였다. 실험 데이터는 그림 5와 같이 신용카드 매출 전표를 대상으로 200dpi로 영상화한 Low-Quality 그레이 문서 영상을 이용하였다.

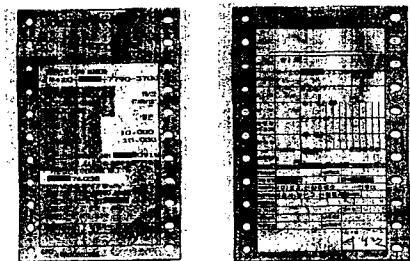


그림 5. 신용 카드 매출 전표의 예

실험에 사용된 양식 문서는 6종류이며 실험 장수는 총 348장이다. 모든 분할 영역의 특징을 추출한 다음, 문서를 분류하는데 평균 0.35초의 처리 시간이 소요되었고 96.6%의 문서 분류의 성공률을 나타내었다.

특징을 추출하는 분할 영역의 개수에 따른 문서 분류의 에러율 변화를 실험해 보았다. 특징을 추출하는 분할 영역의 개수가 증가함에 따라 에러율이 감소하리라 예측할 수 있다. 그림 6에서는 특징 추출에 사용되는 분할 영역의 위치를 설정하는 방법 중에서, 임의로 위치를 설정하는 무작위(random) 방법과 본 논문에서 제안한 큐(queue) 방법을 비교하였다. 그림 6에서 보면 큐 방법에 의한 방법이 무작위 방법보다 문서 분류의 에러율이 초기에 급속하게 감소하는 것을 알 수 있다. 또 모든 분할 영역의 개수의 43.7%인 분할 영역을 이용하여도 모든 분할 영역을 이용할 때의 문서 분류 에러율과 비슷한 에러율을 나타내었다.

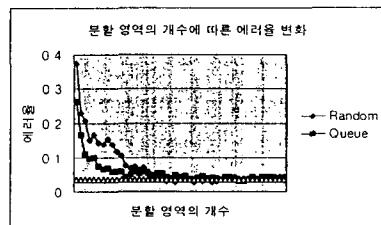


그림 6. 특징을 추출하는 영역의 개수에 따른 에러율 변화

5. 결론

분할 특징을 이용하여 문서 분류를 했을 경우 전체 문서의 43.7%만을 이용하면 전체의 특징을 모두 사용했을

때 나타나는 에러율과 비슷하게 나타난 것을 알 수 있다. 다시 말해 본 논문에서 실험 대상으로 이용한 신용 카드 매출 전표의 경우, 특징을 추출하는 처리 시간을 절반 이상 줄이면서 문서 분류의 에러율은 비슷하게 유지할 수 있다는 것을 실험적으로 검증하였다. 특징을 추출하는 알고리듬을 정교하게 개선하고 분할 영역을 정렬하는 알고리듬을 개선하면 좀 더 적은 개수의 분할 영역만을 이용해서도 문서 분류가 가능하리라 본다.

6. 참고문헌

- [1] Toyohide Watanabe, Hiroyuki Naruse, "Structure Analysis of Table-form Documents on the Basis of the Recognition of Vertical and Horizontal Line Segments", Proc. of 1st Int. Conf. on Document Analysis and Recognition, IEEE Computer Society, pp.638-646, 1991.
- [2] Yasuto Ishitani, "Model Matching Based on Association Graph for Form Image Understanding", Proc. of 3rd Int. Conf. on Document Analysis and Recognition, IEEE Computer Society, pp.287-292, 1995.
- [3] Osamu Itoh and David S. Doermann, "Robust Table-form Structure Analysis Based on Box-Driven Reasoning", Proc. of 3rd Int. Conf. on Document Analysis and Recognition, IEEE Computer Society, pp.218-221, 1995.