

두자 접촉 숫자열의 분할-자유 인식

최순만* · 오일석

전북대학교 컴퓨터학과

Segmentation-free Approach to the Recognition of Two Touching Numerals

Soon-Man Choi and Il-Seok Oh

Department of Computer Science, Chonbuk National University

요 약

숫자열 인식은 문서 자동화 분야에서 매우 중요하다. 기존의 방법들은 숫자열을 낱자 단위로 분할하는 단계와 분할된 숫자들을 인식하는 두 단계로 이루어져 있다. 이들 방법으로는 접촉 유형의 많은 변형 때문에 만족할 만한 결과를 얻을 수 없다. 본 논문은 두자 접촉 숫자열을 분할-자유 방법으로 인식하는 방법을 제안한다. 이 방법은 두자 접촉 숫자열을 하나의 패턴으로 간주하여 인식한다. 즉, 인식 대상이 되는 부류가 100개(00, 01, 02, ..., 98, 99)이다. 훈련 데이터는 NIST 데이터베이스에의 낱자 단위 숫자들을 조합하여 합성하였고, 테스트 데이터는 NIST의 숫자열 필드에서 두자 접촉한 숫자열 민을 추출하여 사용하였다.

1. 서론

숫자는 일반적으로 금융 기관의 분표나 관공서의 각종 민원 서류, 일상 생활에서 사용되는 양식 등에서 금액이나 키 또는 분류키로서 많이 사용되고 있다. 때문에 문서 처리 자동화에서 일반적인 다른 문자에 비해 그 중요도가 상대적으로 높다고 할 수 있다. 그래서 필기 숫자에 대한 인식 연구가 활발히 진행되어 상당한 인식률과 신뢰도를 갖는 낱자 숫자 인식이 발표되었다 [1, 2, 3].

최근 들어서는 낱자 단위 인식과 더불어 연속 필드에 대한 인식 필요성이 크게 대두되고 있는데, 문서 처리 자동화를 위해서는 연속 필드를 인식하는 기술이 필수적이기 때문이다. 일상 생활에서 흔히 사용하고 있는 많은 양식에는 숫자를 낱자 단위로 기입하는 형태가 아닌 하나의 긴 박스나 밑줄에 연속적으로 필기하는 형태가 대부분인데 이를 인식하기 위해서는 연속 필드를 낱자 단위나 단어 단위로 분할해주는 기술이 이 분야의 핵심이라 할 수 있다. 즉, 분할기를 이용하여 연속 필드를 낱자 단위로 분할한 후 분할된 것을 낱자 단위 인식기로 인식하는 것이다. 현재 이에 대한 많은 연구들이 행되어 왔다.

Shudhar은 아래 의곽선에서 시작점을 잡아 위로 진행하며 검은검을 만나면 주위 상황에 따라 왼쪽 또는 오른쪽으로 적어 나가며 분할 경로를 찾는 방법을 사용하였다 [6]. 이 알고리즘은 분할과 인식이 완전히 독립되어 있고 패턴의 아래 부분의 모양만을 사용하므로 근본적으로 한계를 안고 있다. Matan 등은 수직 방향 투영에서 분할 후보 지점을 선택한 후 인식기의 인식 결과에 따라 최종 분할 지점을 결정하는 방법을 사용하였다 [7]. 인식과 분할을 동시에 수행하는 순환 신경망을 제안되어 있다 [8]. 이 방법은 숫자열의 공간 정보뿐만 아니라 화소들 간의 공간적인 의존성도 학습에 고려하여 분할 및 인식 성능 향상을 모색하였다. Nishida와 Mon은 세션화된 컴퓨터에서 특정점들을 중심으로 선분들로 인식한 후, 이들 선분의 개개 조합을 인식기에 입력하여 인식 결과에 따라 분할하는 방법을 제안하였다 [9]. Fenrich는 연속 필드를 분할 방법을 순서적으로 적용해가며 인식 결과에 따라 분할을 결정하는 방법을 사용하였다 [10]. 패턴의 윗

과 아랫 부분의 특징점 쌍을 분할 후보 점으로 하여 이들을 적당한 점수식으로 정렬하여 최고점을 얻은 경로를 최종 분할 경로로 사용하는 방법도 개발되어 있다 [11, 12, 13, 18]. 이 방법에서는 점수식을 어떻게 만드느냐가 가장 중요한 문제이다. 이러한 방법들은 주로 경험적인 방법들을 사용하며 나름대로의 장점과 한계를 가지고 있다. 단순한 형태의 숫자열 분할에는 어느 정도 만족할 만한 성능을 보이고 있으나, 두 개의 인접 숫자가 두곳 이상에서 접촉하거나, 입력 영상에 잡음이 섞인 경우 등 좀더 복잡한 형태의 숫자열 분할에는 많은 문제점을 안고 있다.

또다른 문제점으로는, 숫자열 분할 성공률이 Westall 에서 82%[11], Shi 에서 78%[14]등으로 낮다는 것이다. 또한 분할이 성공적으로 이루어졌다 하더라도 분할된 낱자에 리가취(ligature) 같은 여분의 획이 포함되는 경우가 발생하여 이를 낱자 인식이 인식이 어렵다는 것이다. 결국 이들 문제로 인해 연속 필드의 최종 인식률(분할 성공률 × 숫자 1 인식 성공률 × 숫자 2 인식 성공률)이 낮아 전체적인 성능을 저하시키고 있다.

최근의 또다른 연구로 Wang에서는 두자 접촉한 숫자열을 분할하지 않고 인식하는 문제가 논의되고 있다 [5]. 이 방법은 패턴의 일부만으로도 그 패턴을 인식할 수 있다는 원리를 이용한 것으로 두자 접촉 숫자열 중앙의 일정 부분을 잘라내고 나머지 일부만을 인식하는 방법을 다루었다.

본 논문에서는 두자 접촉한 숫자열 인식 문제를 다루며, 접근 방법은 위에서 언급한 분할-기반 방법이 아닌 분할-자유 방법을 사용한다. 즉, 두자 접촉한 패턴을 낱자 단위로 분할하지 않고 하나의 단일 패턴으로 간주하여 통째로 인식한다. 즉, 훈련과 인식 대상이 되는 부류 수가 100개이며, 각각은 00, 01, 02, ..., 50, 51, ..., 98, 99 등이다. 두 자로 제한한 숫자열 필드에서 나타나는 접촉 길이 빈도 수에서 두자 접촉이 가장 두드러지게 나타나고, 부류수가 우리가 다룰 수 있는 범위내에 있기 때문이다. 높은 성능을 보장하고 두자 접촉 모델은 임의의 길이 숫자열 인식을 구현하는데 중요한 고도로 사용할 수 있다.

두자 접촉 숫자열 인식기의 훈련을 위해 훈련 데이터 집합

이 필수적인데, 100개 부류에 대한 충분한 양의 데이터 수집이 사실상 어려운 문제로, 본 논문에서는 낱자 패턴을 조합하여 훈련 패턴을 합성하였으며, 분류기는 부류별로 훈련이 가능한 모듈러 신경망 구조를 사용하였다. NIST(National Institute of Standards and Technology) 데이터베이스에서 낱자 숫자를 사용하여 훈련 데이터 집합을 합성하였고, 테스트 데이터는 NIST의 숫자열 필드에서 두자 접촉한 숫자열 1,374개를 추출하여 사용하였다.

2. 두자 접촉 숫자열 합성

두개의 낱자 숫자 패턴을 A와 B라 할 때 A와 B를 하나로 합한 AB와 BA 패턴을 합성할 수 있다. 그림 21은 낱자 숫자 A와 B, 그리고 합성된 두자 접촉 숫자열 AB와 BA를 보여주고 있다.



그림 21 낱자 숫자와 합성된 숫자열

숫자열 패턴을 합성하기 위해서는 틀을 이용한 방법과 연결 요소를 이용한 방법들을 생각할 수 있는데 본 논문에서는 후자의 방법을 이용하여 훈련 데이터를 합성하였다. 좌측의 숫자 A의 연결 요소의 개수가 n이고 우측의 숫자 B의 연결 요소 개수가 m일 때, AB의 연결 요소 개수의 합이 n+m보다 작을 때까지 한쪽의 숫자를 평행 이동시키는 방법이다. 즉 양쪽의 숫자들이 서로 연결 요소로 접촉이 이루어지도록 하는 방법이다. 여기서 평행 이동 시킬 때, 이동 축을 어디에 위치시키느냐에 따라서 여러 종류의 숫자열을 합성할 수 있다. 그림 2.2는 이 방법을 이용하여 만들어진 예이며, a)는 낱자 패턴의 위선, b)는 아래선, 그리고 c)는 중앙선에서 이동한 것이다.



그림 2.2 연결 요소를 이용한 접촉 숫자 합성

본 논문에서는 위의 방법을 이용하여 합성한 데이터 집합을 분류기의 훈련에 사용하였다. 사용된 낱자 숫자는 NIST 데이터베이스의 낱자 숫자들이고, 숫자열 합성시 양쪽 낱자 단위의 높이의 차가 크면 이상한 모양의 패턴이 되므로 높이의 차가 ±30% 이내가 되도록 하였다. 그림 2.3은 합성된 예제 패턴들이다.

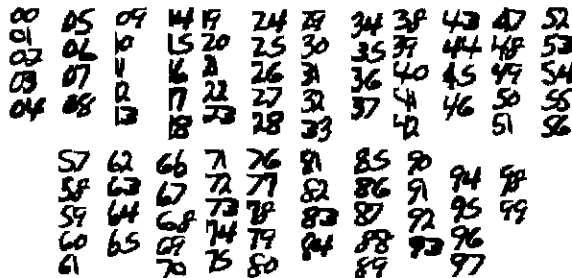


그림 2.3 합성된 두자 접촉 숫자열

이렇게 합성한 접촉 숫자열 중에는 그림 24에서 보는바와 같이 실제 발생할 수 없는 것들도 포함되어 있다. 현재 실험에서는 이러한 패턴들도 훈련집합에 포함되어 있다. 이러한 패턴의 제거, 그리고 같이 리가워가 많이 발생하는 부류들을 위한 리가워 삽입등의 작업을 통해 훈련 패턴 집합의 품질을 향상시키는 것이 향후 연구에서 아주 중요한 문제이다.

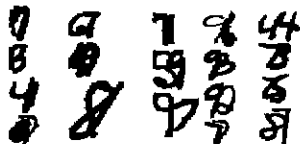


그림 2.4 저품질 합성 패턴

3. 두자 접촉 숫자열 인식기

3.1. 모듈러 신경망 인식기

인식기로는 모듈러 신경망을 사용하였다 [5, 17]. 그림 3.1은 100개의 부류중 하나의 특정 부류를 위한 부 망을 보여준다. 모듈러 구조를 이용하면 각각의 부류별로 훈련할 수 있는 장점이 있다. 이 네트워크 층간은 완전 연결되어 있으며, 1개의 입력층과 1개의 은닉층 그리고 1개의 출력층으로 구성되어 있다. 입력층은 20×16 명암 영상의 입력 데이터를 구성하기 위하여 320개의 노드, 은닉층은 4개의 노드, 출력층은 2개의 노드로 구성되어 있다. 훈련 알고리즘으로는 오류 역전파 알고리즘을 사용하였다.

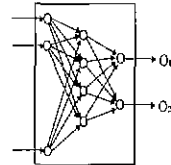


그림 31 모듈러 신경망을 위한 부 망

3.2. 특징 추출

특징을 위해서는 DDD(Directional Distance Distribution)를 사용하였다 [15]. 이 특징은 먼저 입력 패턴 P를 명암 배치를 R로 정규화 후 이진화한다. 이진화된 매쉬에서 각 화소별 좌우 원 집합과 점은 집합이리 부르는 두 개의 8비트 집합에, 관련 화소와 점은 화소 각각에서 8 방향에 대해 가장 가까운 점은 화소 계산후 배열을 1×4 블록 배치를 나눈 다음 평균을 구하여 최종적으로 320-치원의 DDD 특징 벡터를 얻는다.



a) 입력 패턴 b) 정규화된 명암 c) 이진화된 매쉬
P_{35×27} 매쉬 R_{20×16} 매쉬 R_{20×16}

4. 실험

제안된 방법의 실험을 위하여 인식기의 학습은 SUN Ultra Enterprise 2에서 이루어졌으며, C 언어를 사용하여 구현하였다. 훈련 데이터는 NIST 데이터베이스의 낱자 숫자를 이용하여 클래스별로 300자씩 총 30,000(300자×100클래스)자를 합성하였다. 합성된 데이터 예제는 그림 23과 같다. 제안된 방법의 테스트를 위해 NIST의 숫자열 필드에서 두자 접촉한 숫자열 1,374개를 추출하여 테스트 데이터로 사용하였다. 그림 4.1은 NIST에서 추출된 두자 접촉 패턴들이다.

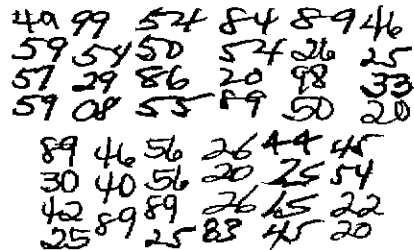


그림 4.1 NIST에서 추출된 테스트 데이터

표 41은 훈련 데이터에 대한 인식 성능 표이며 각각의 네모칸은 각 부류별로 올바르게 인식한 개수이다. 표 4.2는 테스트 데이터의 각 부류별 개수와 분류기에 의해 올바르게 인식된 개수를 나타내고 있다. 훈련 데이터와 테스트 데이터에 대한 인식률은 각각 97.3%와 64.4%를 보였다.

표 41 훈련 데이터에 대한 인식표

I \ R	0	1	2	3	4	5	6	7	8	9	Total
0	297	297	290	295	298	295	300	298	285	296	2,951
1	300	299	257	297	298	297	300	296	291	297	2,932
2	296	293	281	282	295	293	295	286	288	285	2,894
3	296	295	280	294	291	288	295	289	282	288	2,898
4	297	293	288	292	295	291	298	287	285	286	2,912
5	298	295	293	289	292	296	300	284	286	288	2,921
6	297	298	294	296	298	294	300	293	284	287	2,941
7	300	297	295	293	297	296	297	292	287	296	2,950
8	291	286	283	282	287	292	284	283	283	289	2,860
9	297	295	292	292	292	292	292	290	283	284	2,909
Total	2,969	2,948	2,892	2,912	2,943	2,934	2,961	2,898	2,854	2,896	29,168

표 42 테스트 데이터에 대한 인식표

I \ R	0	1	2	3	4	5	6	7	8	9	Total
0	29	1	9	12	5	15	8	2	7	17	105
1	0	0	7	2	4	6	6	2	4	12	64
2	0	0	0	3	1	0	5	2	0	5	18
3	50	23	31	39	2	10	32	1	21	17	261
4	30	20	27	26	8	27	24	0	11	7	177
5	17	0	4	5	6	17	16	6	13	14	119
6	25	4	9	10	11	14	18	3	9	8	112
7	29	4	9	10	11	14	18	3	9	8	112
8	25	4	9	10	11	14	18	3	9	8	112
9	14	2	2	3	3	6	22	4	0	3	79
0	14	2	2	3	3	6	22	4	0	3	79
1	3	0	2	2	2	2	2	2	1	1	16
2	3	0	2	2	2	2	2	2	1	1	16
3	14	0	4	17	11	15	10	5	20	154	282
4	9	0	4	11	7	11	7	5	6	1	83
5	0	0	1	4	3	7	10	0	4	1	40
6	0	0	1	4	3	7	10	0	4	1	40
7	0	0	1	4	3	7	10	0	4	1	40
8	0	0	1	4	3	7	10	0	4	1	40
9	0	0	1	4	3	7	10	0	4	1	40
Total	213	34	92	119	112	187	145	59	124	310	1,374
	140	26	68	69	77	120	107	28	72	180	885

표 43은 테스트 데이터의 좌우 숫자의 날자별 개수이며, 표 44는 날자 단위 인식률이다. 여기서 CC는 좌우 모두 인식한 경우, CE는 왼쪽 인식 오른쪽 오류, EC는 왼쪽 오류 오른쪽 인식, 그리고 EE는 모두 오류인 경우이다. 표 45는 테스트 데이터에서의 혼돈 행렬이며, C, E, R은 각각 정인식, 오류, 기각이다.

표 43 날자별 숫자 부류 개수

부류	0	1	2	3	4	5	6	7	8	9	Total
개수	318	52	353	238	331	412	224	87	375	358	2,748

표 44 날자 단위 인식률

	CC	CE	EC	EE	정인식 날자 개수	날자 정인식률
Train	29,168	360	305	130	59,001	98.5%
Test	885	198	191	100	2,159	78.6%

표 45 날자별 인식 혼돈 행렬

	0	1	2	3	4	5	6	7	8	9	C	E	R
0	227	29	12	2	4	6	27	3	5	3	227	91	0
1	0	35	5	0	4	1	1	4	2	0	35	17	0
2	6	7	289	18	2	13	9	5	1	3	289	64	0
3	1	3	13	184	2	19	3	4	6	3	184	54	0
4	1	14	9	1	277	1	14	2	5	7	277	54	0
5	4	5	1	37	3	351	3	0	5	3	351	61	0
6	1	5	11	0	5	3	196	1	2	0	196	28	0
7	0	6	7	1	2	0	0	64	1	6	64	23	0
8	6	22	14	9	12	17	3	6	259	27	259	116	0
9	6	18	4	1	27	5	0	15	2	277	277	81	0
	227	35	289	184	277	351	196	64	259	277	2,159	589	0
	125	109	176	172	161	165	160	140	129	152			
											78.6%	21.4%	0%

현재까지는 예비 실험으로서 낮은 성능을 얻었다. 하지만 훈련 데이터의 품질 향상, 보다 정교한 특징의 사용, 발생 불가능한 부류('1'과 같은)의 제거 등을 통해 성능을 높일 여지가 많이 남아있다. 이러한 점이 향후 연구로 설정되었다.

5. 결론 및 향후 연구 과제

본 논문은 두자 접촉 숫자열 인식 문제를 날자 단위로 분할하여 인식하는 분할-기반이 아닌 인식 관점에서 보는 완전 분할-자유 관점으로 신경망 인식기로 인식하는 문제를 다룬 것이다. 훈련을 위한 데이터로는 클래스별로 다량의 데이터 수집이 어려운 관계로 날자 단위의 숫자들을 이용한 합성 데이터를 사용하였다. 성능 평가를 위해 NIST 데이터베이스에서 두자 접촉 숫자열을 추출하여 테스트 데이터로 사용하였다. 앞으로의 연구 과제는 다음과 같다. 우선 인식을 측면을 보면 합성한 훈련 데이터에서는 97.3%를 보였으나, 실제 테스트 데이터에서는 64.4%의 낮은 인식률을 보였는데, 이에 대한 보완책이 필요하다. 즉, 훈련 데이터의 품질 향상, 최적의 특징 사용이나 특징 설계, 두자 접촉 숫자열 인식을 위한 최적의 네트워크 구조 설계가 필요하다. 또한, 인식 대상이 되는 부류수가 많아진 관계로 훈련 시간이 많이 소요되는데 훈련 시간 단축 문제도 향후 연구과제로 다루어져야 할 사항이다. 또한 임의 길이의 숫자열 인식을 위하여 날자 숫자 인식기와 두자 접촉 숫자열 인식을 이용하는 방안을 강구 하고 있다

참고 문헌

- [1] H. Bunke, P.S.P Wang, and H.S Baird (Edited), Document Image Analysis, World Scientific, 1994
- [2] H.S Baird, H. Bunke, and K. Yamamoto (Edted), Structured Document Image Analysis, Springer-verlag, 1992.
- [3] H. Bunke, Advances in Structural and Syntactic Pattern Recognition, World Scientific, 1992.
- [4] C.Y. Suen et al., "Computer recognition of unconstrained handwritten numerals," Proceedings of IEEE, Vol. 80, No. 7, pp. 1162-1180, July 1992.
- [5] X. Wang, V. Govindaraju, and S. Srihari, "Holistic recognition of touching digits," Proceedings of IWFHR VI, pp. 295-303, 1998.
- [6] M. Shridhar and A. Badreldin, "Recognition of isolated and simply connected handwritten numerals," Pattern Recognition, Vol. 19, No.1, pp.1-12, 1986.
- [7] O. Matan et al., "Reading handwritten digits: a ZIP code recognition system," IEEE Computer, Vol. 25, No. 7, pp. 59-63, July 1992.
- [8] 이정환, 이용재, "순환 신경망을 이용한 연속 필기 숫자열의 통합적 분할 및 인식," 한국정보과학회 1995 추계학술 발표회지, pp. 447-450, 1995.
- [9] H. Nishida and S. Mori, "A model-based split-and-merge method for character string recognition," in Document Image Analysis (Ed. by Bunke et al, World Scientific), pp. 209-226, 1994.
- [10] R. Fenrich, "Segmentation of automatically located handwritten words," Proceedings of IWFHR II, pp 33-44, 1991.
- [11] J.M Westall and M.S. Narasimha, "Vertex directed segmentation of handwritten numerals," Pattern Recognition, Vol. 26, No. 10, pp. 1473-1486, 1993
- [12] N.W. Strathy, C.Y. Suen, A. Krzyzak, "Segmentation of handwritten digits using contour features," Proceedings of ICDAR, pp. 577-580, 1992.
- [13] H. Fujisawa, Y. Nakano, and K. Kurino, "Segmentation methods for character recognition: from segmentation to document structure analysis," Proceedings of IEEE, Vol. 80, No. 7, pp. 1079-1092, 1992.
- [14] Z. Shi and V. Govindaraju, "Segmentation and recognition of connected handwritten numeral strings," Pattern Recognition, Vol. 30, No 9, pp. 1501-1505, 1997
- [15] 오일석, Chun Y. Suen, "광학 문자 인식을 위한 거리 측정," 정보과학회논문지 Vol 25, No. 7, pp 1028-1043, 1998.
- [16] M. Cheriet, Y. Huang, and C. Suen, "Background region-based algorithm for the segmentation of connected digits," Proceedings of 11th ICPR, pp. 619-622, 1992.
- [17] I.S. Oh, J.S. Lee, S.M. Choi, and K.C Hong, "Class-expert approach to unconstrained handwritten numeral recognition," IWFHR V pp 95-102, 1996.
- [18] 최순만, 홍기천, 오일석, "무제약으로 필기한 숫자 필기의 인식에 대한 연구," 제8회 영상 처리 및 이해에 관한 워크샵 발표논문집, pp. 1-6, 1996