제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

Virtual ID 사용을 위한 유사한 우편 영상 추출 방법

박상철, 정창부, 손화정, 김수형
전남대학교 전산학과
e-mail:sanchun@iip.chonnam.ac.kr

Mail Image Filtering Method for Use of Virtual ID

Sang-Cheol Park, Chang-Boo Jeong, Hwa-Jeong Son,
Soo-Hyung Kim
Dept. of Computer Science, Chonnam National University

요 약

우편물을 배달하기 위해서 집배원은 오전 시간의 대부분을 이용하여 배달 경로에 따라 우편물을 정렬한다. 우리나라의 자동화 시스템은 순로구분의 전단계까지만 수행하는데 그치고 있으나 외국의 순로구분 자동화 시스템은 바코드나 사용자 태그를 이용하여 순로구분을 수행한다. 본 논문에서는 영상 기반특징과 인식 기반 특징인 Virtual ID 사용을 위한 우편 영상 검증의 과정으로 처리 속도를 향상 시킬 수 있도록 유사한 영상 혹은 동일 DM 발송 우편 영상을 추출해 내는 2가지 방법을 제안한다. 첫째는 영상 기반 특징을 추출하여 신경망을 사용하고, 두 번째는 우편 영상의 문자열의 Bound Box를 추출하여 이들의 겹침정도를 이용하여 유사성을 판별한다. 실험을 통해 제안한 두가지 방법이 유용함을 입증하였다.

1. 서론

우체국에서 우편물을 집배원 별로 구분하여 각 집배원에게 할당한 후, 배달이 이루어지기 위해서는 집배원의 배달경로에 따른 우편물의 정렬이 먼저 이루어진다. 일반적으로 집배원들은 동이 구분된 우편물을 수령 후 여러 명이 모여서 자신의 배달구역의 우편물을 분류하여 나누어 가진 후, 이 우편물을 자신의 배달 경로에 맞게 다시 분류하여야 한다. 만약 1명의 집배원 분의 최적 배달 경로 데이터베이스를 만드는데 5일이 소요되고 10,000명분을 하려면 50,000일이 걸린다. 이는 1년을 300일로 계산하면 166.7년이 되고 한 명의 1년 인건비가 1,500만원이라면 추정 인건비는 약 25억이 산출된다.

우편 기술 선진국에서는 우편배달 순로구분의 자동화를 위해 집중국에서 우편물의 인식 결과를 바코드로 변환하여 우편물에 인쇄하고 배달국에서 모든 처리 과정과 순로구분 과정에서 이를 활용한다. 이 선진국들은 우편배달 순로구분 시스템을 구축하기 위해 오래 전부터 꾸준히 국가차원에서 투자하여 경제적·기술적 효과를 누리고 있다. 이에 비해 국내 우편배달 순로구분 시스템은 수작업에 의존하고 있다.

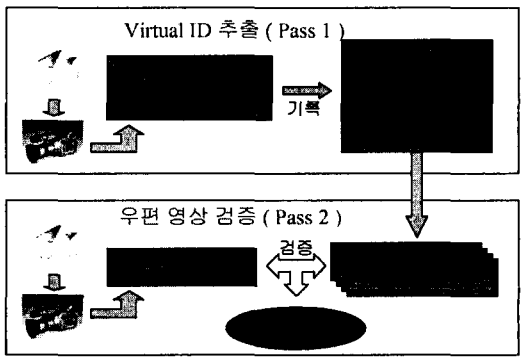

그림 1. Virtual ID 사용을 위한 우편물 검증 과정

우편 영상을 자동으로 구별하는 장치에서는 일반적으로 바코드나 사용자 Tag를 사용하는데 반하여 본 논문에서는 영상 기반특징과 인식 기반 특징인

Virtual ID 사용을 위한 우편 영상 검증의 과정으로 처리 속도를 향상 시킬 수 있도록 유사한 영상 혹은 동일 DM 발송 우편 영상을 추출해 내는 2가지 방법을 제안한다. 첫째는 영상 기반 특징을 추출하여 신경망을 사용하고, 두 번째는 우편 영상의 문자열의 Bound Box를 추출하여 이들의 겹침정도를 이용하여 유사성을 판별한다.

국내의 우편배달 순로구분 자동화 시스템이 구축되어 있지 않은 상황에서 Virtual ID 개념을 도입한 우편배달 순로구분 시스템은 순로구분의 저비용 효과와 우편물 분류에 있어 큰 신뢰성의 이득을 가져다 줄 것으로 사료된다.

2. 관련 연구

Virtual ID를 이용한 우편물 검증 기술은 외국 기업인 Solystic [1]만이 보유하고 있는 기술이다. 이들이 사용한 특징은 전체 우편 영상의 가로·세로 크기, 그레이 레벨의 평균·분산, 엔트로피 등이고 부분 영상의 특징으로 전체 영상에서 사용한 특징을 그대로 사용하고 있다. 그리고 마지막으로 우편 영상의 주소 열을 구성하는 단어의 개수로 우편물을 검증하고 있다. 그 업체의 자료에 의하면 95.5%의 성능을 주장하고 있으나 실제는 80%의 검증 성능을 보인다.

국내에서는 아직까지 우편배달 순로구분 자동화의 주제로 수행된 과제는 없으며 Virtual ID를 사용한 우편물 검증 관련 논문은 본 논문이 유일하다.

3. 신경망을 이용한 방법

3.1 수행 방법

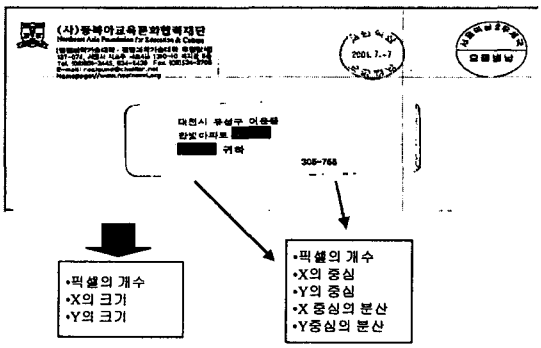

그림 2. 영상 기반 특징 추출

유사한 우편물 영상과 동일 DM 발송 우편물 영상은 크기와 영상을 이루는 화소의 개수가 근소한 차이를 갖는다. 따라서 영상을 수직 4등분, 수평 3등분의 Grid로 구분 [2]하여 각 Grid를 이루는 화소의 개수, 중심 그리고 중심의 흩어진 정도를 계산하여 이 값들의 거리(Distance)가 적은 영상이 서로 유사한 영상이라고 판별할 수 있다. 따라서 두 영상을 이루는 특징의 거리 값을 신경망의 입력으로 하여 두 영상의 유사성을 판별하였다.

그림 2에서 알 수 있듯이 전체 영상에서 3가지 특징과 12개 지역 영상에서 5가지 특징을 추출하여 총 63가지의 특징을 추출하였다. 비교되는 두 영상의 특징을 추출하고 이들의 유클리디안 거리를 계산하였다. 이 거리 값들의 분포가 정규분포를 이룬다고 가정하고 0과 1사이의 값으로 정규화한 후 신경망에 입력하였다.

3.2 실험 및 결과

실험 장비는 Pentium-Ⅳ PC, 2.4GHz CPU, 1GB Memory를 사용하였고, 실험 데이터는 한국전자통신연구원에서 제공한 415장의 영상을 이용하였다. Pass 2의 임의의 영상과 Pass 1에서 Pass 2에 있는 동일한 영상을 포함한 7장의 영상을 비교하기 위해 3587개의 쌍을 만들었다. 두 영상이 같은 영상이거나 같은 DM 발송의 우편 영상인 경우 즉, 수취인 주소만 다른 경우를 유사한 영상으로 판단한다. 실험 데이터의 60%인 2154개 쌍(유사한 쌍 1231개, 유사하지 않은 쌍 923개)은 신경망의 훈련에 사용하였고, 40%(206개 영상에서 얻어진 쌍)인 1433개의 쌍은 신경망 [3]의 훈련 정도를 판단하는데 사용하였다.

표 1. 신경망을 이용한 유사 영상 추출 결과

테스트 영상 개수	206 장
동일한 영상인데 유사하지 않다라고 검증된 영상 개수	2 장
유사한 영상으로 검증된 평균 영상 개수	3.83 장
한 장의 영상(7쌍)을 검증하는데 소요되는 시간	124 ms

다음은 표 1에서 동일한 영상인데 유사하지 않다라고 에러를 나타낸 우편 영상들이다. 그림 3과 4는

Pass1과 Pass2의 과정에서 창봉투의 스캔이 불규칙해서 서로 유사하지 않다라고 판단되었다.

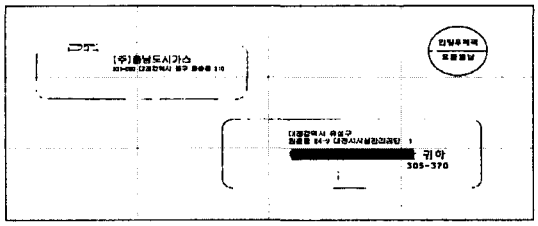

그림 3. Pass 1의 82번 영상

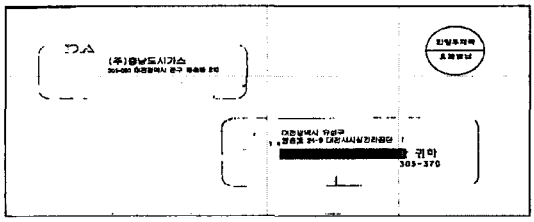

그림 4. Pass 2의 82번 영상

그림 5와 6은 우편 영상이 스캔될 때 한쪽으로 많이 밀려서 나타난 잡음으로 인해 서로 유사하지 않다라고 판별되었다.

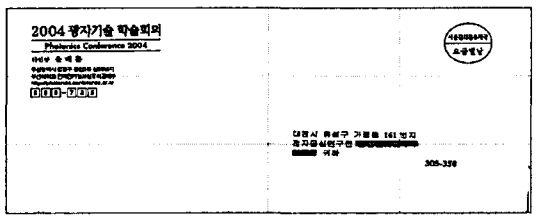

그림 5. Pass 1의 117번 영상

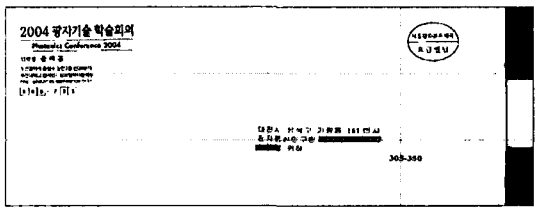

그림 6. Pass 2의 117번 영상

4. 문자열의 Bound Box를 이용한 방법

4.1 수행 방법

Pass 1과 Pass 2에서 얻어진 우편물 영상이 동일 영상이라면 이들의 겹침 정도가 다른 영상에 비해 큰 값을 가질 수 있음에 근거하여 발신인 주소, 수신인 주소, 우표 영역 그리고 광고 영역을 구성하는 문자열 바운드 박스를 추출하여 이들 문자열 바운드 박스의 겹침 정도에 따라 유사성을 판단하였다. 문자열 바운드 박스는 ETRI에서 제공한 문자열 추출 DLL [4]을 이용하여 획득하였다.

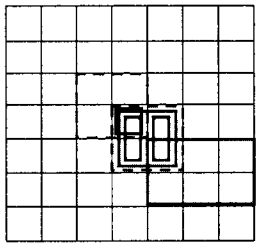

그림 7. 문자열 Bound Box의 중첩

추출된 문자열중에서 에러를 유발하는 Bound Box(BB)는 제거한다. 첫 번째로 BB의 넓이가 영상 넓이의 1/2보다 크거나 같은 경우, 두 번째로 BB의 높이가 영상 높이의 1/2보다 크거나 같은 경우, 마지막으로 BB와 영상 테두리 사이의 거리가 임계치 이하인 경우이다. 바운드 박스의 중첩 정도는 식 2.1에 의해 계산된다.

$$Sim = \frac{(\sum Box_1 \cap Box_2) \times 2 \times 100}{\sum Box_1 + \sum Box_2} \qquad \text{(식 2.1)}$$

4.2 실험 및 결과

실험 장비는 3.1과 동일하며 실험 데이터는 3.1의 415장의 영상에서 219장을 사용하였다.

표 2. BB의 중첩 정도에 따른 유사 영상 추출 결과

A \ B	70	80	90
20	0	1	13
	3.99	3.88	3.47
10	0	1	7
	4.0	3.90	3.58
5	0	1	7
	4.0	3.90	3.58
0	1	3	7
	3.97	3.87	3.52
소요시간	484 ms		
BB 제거 안한 경우	7	17	32
	3.94	3.7	3.13

A : BB제거를 위한 테두리와의 거리에 대한 임계치

B : Sim_{Max} 대비 유사율 임계치

Sim_{Max} : 7개 영상 쌍 중에서 가장 높은 유사도

셀 :

X.xx	False Negative
X.xx	다음 단계로 넘길 영상의 평균

실험 결과에 의하면 BB제거를 위해 BB와 테두리와의 거리가 5~10이 적당하다. Sim_{Max} 대비 유사율 임계치가 80이상이 되면 유사한 영상으로 추출되는 평균 영상 개수가 작아지는 이점이 있으나 에러가 커진다. 반면 70에 가까워지면 에러가 작아지는 관계에 있다.

그림 8과 9는 우편 영상을 스캔할 때 영상이 기울어져 Pass1과 Pass2의 영상이 상당한 차이를 보이고 있다. 따라서 Pass2의 영상 테두리쪽에 나타는 잡음과 Pass2의 왼쪽 아래의 모서리쪽 BB와 아래면의 BB는 제거 대상이 된다.

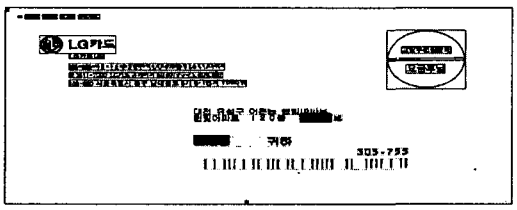

그림 8. Pass 1의 161번 영상

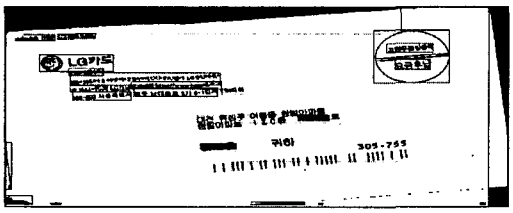

그림 9. Pass 2의 161번 영상

5. 결론 및 향후 연구 과제

본 논문에서는 영상 기반특징과 인식 기반 특징인 Virtual ID 사용을 위한 우편 영상 검증의 과정으로 처리 속도를 향상 시킬 수 있도록 유사한 영상 혹은 동일 DM 발송 우편 영상을 추출해 내는 2가지 방법을 제안하였다. 첫째는 영상 기반 특징을 추출하여 신경망의 입력 입력 데이터로 사용하여 유사한 영상을 추출하였고, 두 번째는 우편 영상의 문자열의 Bound Box를 추출하여 이들의 겹침정도를 이용하여 유사성을 판별하였다. 이들이 결과는 영상 기반 특징을 이용한 경우 주어진 영상 데이터에서 7장의 영상을 비교하여 유사한 영상을 추출한 경우 평균 3.8에서 4.0개의 영상이 서로 유사하다고 판별하였으며, 그 수준에서 에러는 0%에 가까운 성능을 보였다.

Virtual ID사용을 위한 우편물 검증의 결과는 Pass2의 임의의 영상과 Pass1의 7장의 영상과 비교하여 그중 동일한 영상을 고르는 문제이기 때문에 앞으로 수취인의 주소 영역을 이용하여 단어의 개수를 비교하거나 문자의 특징을 비교하여 동일 우편 영상을 검증하는 방법을 연구할 계획이다.

Acknowledgement

본 연구는 한국전자통신연구원의 "Virtual ID 사용을 위한 우편 영상 검증 기술 연구" 과제의 지원으로 수행되었음.

참고문헌

[1] http://www.solystic.com/

[2] Nievrgelt, J., H. Hinterberger and C. Sevcik. The grid file: an adptable, symmetric, multikey file structure. Proc. of the ACM TODS, pp. 38-71, March. 1984.

[3] E. Gose, R. Johnsonbaugh and S. Jost, *Pattern Recognition and Image Analysis*, Prentice Hall, 1996.

[4] http://www.etri.re.kr/

[5] 김수형, 박상철 외 4명, Virtual ID 사용을 위한 우편 영상 검증 기술 연구 연구 과제 중간보고서, 연구보고서, 한국전자통신연구원, 2004.