명함 영상에서의 E-mail 영역 검출 알고리즘
(An Algorithm of E-mail Region Extraction in a Calling Card Image)

신상철, 권미숙, 정재영
(Sang Chul Shin, Mi Sook Kwon, Jae Young Jung)
동양대학교 컴퓨터공학부

요 약 통신 수단의 발달로 인터넷을 이용한 E-mail이 활성화된 지금 명함에서 E-mail정보는 빠지지 않고 표기된다. 만약 수작업으로 관련 정보를 입력했던 것을 명함이미지에서 E-mail을 자동으로 추출한다면 유용할 것이다. 본 논문에서는 명함 영상에서 E-mail 영역을 검출하기 위한 텍스처 특성을 분석하여 텍스트 영역을 분할하고 연결화소를 이용한 개별문자 추출 방법을 통해 at symbol(@)을 인식하는 방법에 관하여 논한다.

1. 서 론

현대사회에 명함을 주고받는 것은 관례처럼 되어오고 있다. 어떤 직함을 가진 사람은 물론이고 그렇지 않은 사람도 자신을 광고하는 용도로 많이 사용하고 있는 것이 명함이다. 이러한 명함들은 받고 나면 명함첩에 잘 보관하는 경우도 있지만 대개의 경우 서랍이나 책 같은 곳에 보관함으로 분실되거나 훼손되는 경우가 많아서 실제로 필요할 때 제대로 사용할 수 없을 경우가 많다. 그리고 요즘엔 통신 수단이 전화 다음으로 E-mail이 많이 쓰이고 있고 대부분의 명함에 빠지지 않고 기입되어 있는 것이 E-mail이다. 명함을 한번 스캔해서 E-mail 부분을 추출해서 사용함으로써 수작업으로 E-mail을 입력해서 정보에 오류가 발생할 수 있는 것을 줄일 수 있다.

본 논문에서는 텍스처 특성을 분석하여 텍스트(문자)영역을 분할[1]하고, 연결화소를 이용한 개별문자 추출[2] 방법을 통해 at symbol(@)을 인식하여 명함이미지에서 E-mail 영역을 검출하는 방법에 관하여 논한다.

문자영역을 분할하는 기존의 방법으로는 색상 변화 빈도수를 이용하는 방법[3]과 R. Lienhart의 [4,5]에서의 비디오 프레임에 출현하는 문자 중에서 인위적 문자(artificial text)에 관심을 두고 우선 분리와 합병(Split-and-Merge)알고리즘으로 영역들을 분할한 다음 배경영역을 제거하기 위해 문자의 최대와 최소크기 값을 사용하는 방법이 있고 개별문자를 추출하는 기존의 방법으로는 투영에 의한 방법[6,7,8]과 외곽선 추적에 의한 방법[9]이 있다.

문자영역 분할에서의 색상 변화 빈도수를 이용하는 방법은 일정한 거리와 색상변화 빈도수에 대한 임계값을 설정하는 문제가 있고 R. Lienhart가 제안한 방법은 잡음이 많은 곳에서는 무수히 많은 객체들이 분할되므로 문자객체와 배경객체의 구분이 어렵다. 개별문자 추출에서의 투영에 의한 방법은 빠른 속도와 문서 영상의 분리 결과 문자 인식에 영향을 줄 수 있는 많은 정보를 얻을 수 있지만, 겹친 문자를 분리해 낼 수 없고, 자소 분리를 할 수도 없어 분리가 곤란한 경우는 강제 분리를 해야 하고, 외곽선 추적에 의한 방법은 겹친 문자를 분리 할 수 있지만 속도면에서 늦고, 인식에 이용할 획득된 정보가 적다는 문제점을 안고 있다.

본 논문에서는 각 방법들의 장점을 이용하여 문자열 부분의 기울기 특징과 텍스처 특성을 분석하여 문자영역의 명암값 과 크기에 상관없이 복잡한 배경에서 문자영역을 정확하게 분할하는 방법과 투영에 의한 방법과 외곽선 추적에 의한 방법의 장점을 이용하여 한번의 투영으로 연결화소를 이용한 개별 문자 추출 후 각각의 개별 문자의 중

심으로부터 4방향을 탐색하여 at symbol을 찾아냄으로서 E-mail 영역을 검출 할 수 있음을 실험을 통해서 보였다.

본 논문의 구성은 다음과 같다. 2장에서 본 논문에서 제안한 알고리즘을 이용해 E-mail 영역을 검출하는 방법을 기술하고 3장에서 실험 결과를 4장에서 결론 순으로 기술한다.

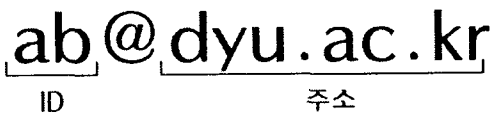
2. 제안한 알고리즘

입력받은 Gray명함 이미지의 배경으로부터 텍스트영역을 분할하기 위해 우선 일반 문자열의 특성과 E-mail 문자열의 특성을 살펴본 후 텍스트영역을 배경과 현격하게 구별할 수 있는 특징을 추출하여 텍스트영역을 배경으로부터 분할하는 과정은 기술하며, 분할된 텍스트영역에서 연결 화소를 이용하여 개별 문자를 추출한 후 각각의 개별 문자에 대해 at symbol의 특징 과 일치하는지 탐색하는 과정을 기술한다.

2.1 특징 추출

텍스트영역은 대부분 하나 이상의 문자열로 존재하며 문자열의 수평 방향으로 단일명암 값이 주기적(periodic)으로 존재하여 배경과 두드러지게 구분되는 특성이 있다. 즉 각 수평라인에 대해서 X축 방향인 왼쪽에서 오른쪽으로 텍스트영역과 배경 명암 값의 큰 차이가 주기적으로 반복되는 텍스처 특성이 존재한다. (그림 1)은 임의의 한글과 영문 문자열에 나타나는 수평방향의 텍스처 특성을 보여주고 있다.

(그림 1) 수평방향의 텍스처 특성



(그림 2) E-mail 주소 형식

(그림 2)는 E-mail 문자열의 전형적인 형태를 보여주고 있다. 즉 한글이 포함되어 있지 않은 영문과 숫자의 조합으로 된 최소한 7개의 문자 이상으로 되어있다는 현격한 차이가 존재한다는 점을 알 수 있다.

2.2 명암 간소화

명함의 Gray이미지는 매우 다양한 모양과 명암값으로 구성된 배경을 가지고 있으며 심지어 동일한 명암값으로 구성되어 있어야 할 텍스트영역의 명암값도 다양한 명암값으로 구성되어 있다. 따라서 기울기 특징을 추출하기 전에 텍스트영역이 변경되지 않는 범위 내에서 다양한 배경과 텍스트영역의 명암값을 단순화하여 텍스트영역 추출 오류를 최소화 할 수 있고 다음에 처리할 연결 화소처리에서 문자의 명암값과 배경의 명암값을 명확히 구분 할 수 있다.

명암 간소화 작업을 하는 방법으로 우선 이미지의 각 명암 단계에 대해서 식 (1)을 적용한다.

$$D(x,y) = \lfloor \frac{G \times O(x,y)}{K} \rfloor \times \frac{K}{G} \quad \text{식(1)}$$

D(x,y)는 각 픽셀의 명암간소화 결과, O(x,y)는 각 픽셀의 원래 명암값, K는 원래 명암단계의 수, G는 축소하고자 하는 명암단계의 수를 의미한다.

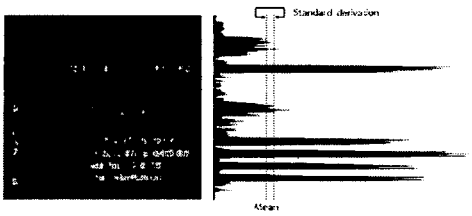
2.3 후보영역 검출

텍스트영역의 특징은 명암값이 간소화 된 이미지에 대해 각 수평 라인 화소들의 기울기 측면도(profiles)와 각 라인마다 기울기 크기의 합으로 표현할 수 있다. 기울기 크기는 x 방향 소벨마스크 Gx에 의해 구성되며 기울기 크기의 합을 수평방향에 투영(projection)하여 텍스트영역과 비 텍스트영역으로 구분할 수 있다.

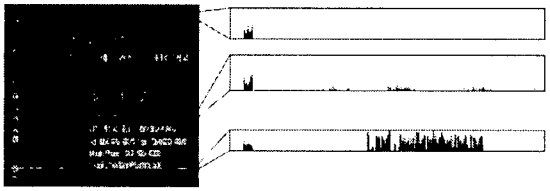
(그림 3)은 기울기 프레임에서 기울기 크기의 합에 대한 투영도를 보여주고 있으며 텍스트영역이 비 텍스트영역에 비해 현격히 높은 기울기 크기의 합을 가지고 있다는 점을 알 수 있다.

(그림 4)는 기울기 프레임에서 수평 라인 각 화소에 대한 기울기 측면도를 보여주고 있으며 텍스트영역의 기울기 값

빈도가 비 텍스트영역에 비해 높다는 것을 알 수 있다.



(그림 3) 기울기 투영도



(그림 4) 기울기 측면도

2.3.1 수평 영역 검출

기울기 투영도를 분석하면 후보영역의 Top 위치와 Bottom 위치를 분할할 수 있다. 기울기 투영도는 임계값 이상의 기울기 합이 텍스트영역 구간에서 수직 방향으로 연속성을 가지기 때문에 후보영역의 세로인 Y축 영역 구간을 결정하는데 사용된다. 텍스트영역을 2차원 (x,y) 좌표로 보았을 때 직사각형의 왼쪽 하단 (xi,yi)과 오른쪽 상단 (xj,yj)로 표현된다. 세로 영역 yi 와 yj를 결정하기 전에 우선 (그림 3)의 기울기 투영도에 나타낸 평균과 표준편차를 다음 식들에 의해 계산한다. width 와 height는 이미지의 가로와 세로 크기, G_X 는 소벨 수직방향 기울기 마스크, G_Y 는 이미지 세로 각 라인의 기울기 크기의 합, G_m 은 G_Y 의 대한 평균, G_σ 는 표준편차이다.

$$G_Y = \sum_{X=1}^{width-1} \sqrt{G_X{}^2}$$

$$G_m = \frac{1}{height-2} \times \sum_{Y=1}^{height-1} \sum_{X=1}^{width-1} \left(\sqrt{G_X{}^2}\right)$$

$$G_\sigma = \sqrt{\sum_{Y=1}^{height-1} (G_Y - G_m)^2 \times \frac{1}{height-2}}$$

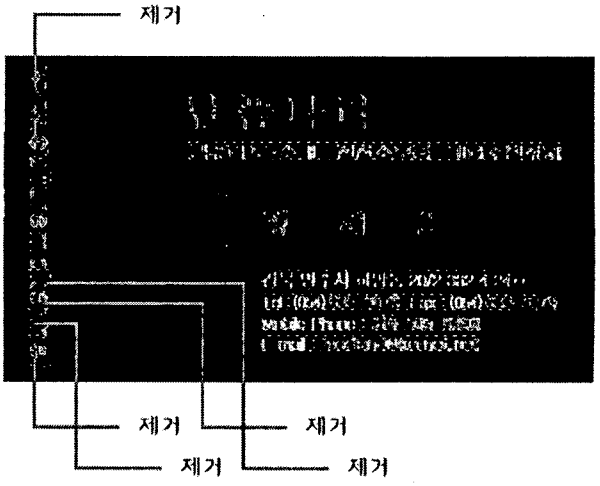
다음으로 $[\ G_m + G_\sigma < G_Y]$를 만족하는 G_Y 의 값들을 탐색한 후 연속적인 구간을 찾아내어 (yi,yj)의 쌍들을 후보영역의 세로인 Y축 영역구간으로 결정한다.

2.3.2 수직 영역 검출

기울기 측면도를 분석하면 후보영역의 Left 위치와 Right 위치를 분할할 수 있다. 기울기 측면도는 텍스트영역 구간에서 빈도수가 높기 때문에 후보영역의 가로인 X축 영역 구간을 결정하는데 사용된다. 결정된 yi 와 yj 범위 내에서 누적한 기울기 측면도를 보면 가로 영역 xi 와 xj 는 누적한 기울기 값이 평균값 이상이고 값들 사이의 거리가 임계값 이하이다, 즉 연속된 구간을 자막영역의 가로인 X축 영역구간으로 결정한다. 여기서 말한 임계값은 (yj - yi)*2 로 즉 글자들의 비율을 1:1로 봤을 때 높이 와 폭이 같기 때문에 높이의 2배 이상의 거리가 되면 분리된 문자라 인식할 수 있기 때문이다.

2.4 텍스트 영역 검출

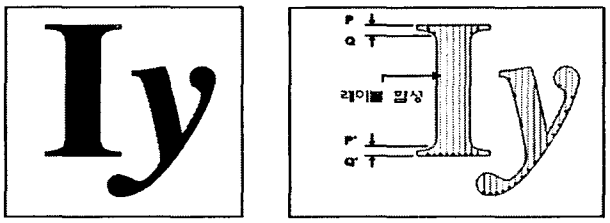
위의 같이 선택된 후보영역이 배경일 가능성을 배제할 수 없고 필요없는 후보영역을 제거하기 위해 수평방향으로 단일 명암값이 주기적으로 존재하는 텍스트영역의 특징을 이용 수평방향의 길이의 임계치를 적용시켜 임계치 이하이면 후보영역들을 (그림 5)와 같이 제거한다.
임계치로는 (yj - yj) * 7을 이용하는데 이는 위에서 설명했듯이 E-mail의 최소 길이를 반영한 것이다.



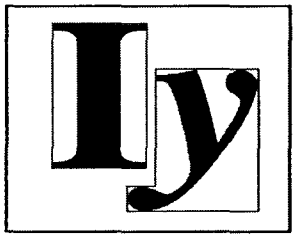
(그림 5) 후보영역 제거

2.5.1 연결 화소를 이용한 개별문자 검출

위의 절에서 추출된 텍스트영역을 가지고 연결 화소를 추출하기 위해서는 각 문자열을 세로로 투영하여 추적하는데, 우선 찾아진 점에서부터 이웃화소들과 연결되었는지 알아야 한다. 그 방법으로서 세로 투영시 배경은 0으로 레이블링하고 문자화소에 대해서는 연속 화소를 구하여 레이블링한다. 이러한 결과 각 레이블 값은 하나의 연결 화소 블록을 가지게 된다. 이렇게 추출된 연결 화소들을 분리 과정을 통하여 겹친 문자와 접합 문자 등을 분리 할 수 있다. (그림 6)에서는 연결 화소 추적을 이용하여 레이블링된 자소를 모두 개별적으로 추출한다.



(a) 입력영상 (b)레이블화



(c) 연결화소 추출

(그림 6) 레이블링 과정

아래 단계 3에서와 같이 세로 방향 연속 화소 구간 (P,Q)에서 왼쪽 이웃 화소들의 레이블 값을 참조한다. 레이블 값이 없으면 구간 (P,Q)에 새로운 레이블 값을 부여하고 레이블 값이 있으면 그 값을 구간 (P,Q)에 부여한다. 구간 (P,Q)에 레이블 값이 결정되면 그 값에 해당하는 연결 화소 블록을 구간 (P,Q)에 따라 수정한다. 만약 구간 (P,Q)의 왼쪽 이웃 화소들이 두 개 이상의 레이블 값을 갖는다면 가장 작은 값을 구간(P,Q)에 부여한다. 나머지 값들의 블록은 구간 (P,Q)의 블록에 합성한다.

단계1. 문서 영상에서 배경은 0 문자화소는 1로 나타내는 함수를 f(x,y)로 표현한다.

$$f(x,y)=\begin{cases}0 & \text{배경화소}\\ 1 & \text{문자화소}\end{cases}$$

단계2. f(x,y)=0 이면 , L(x,y)=0
f(x,y)=1 이면, 연속 화소 구간을 구한다. 세로 방향 연속 화소 구간은 P <= y <=Q에서 f(x,y)=1 이다.

단계3. 구해진 연결 화소 구간 (P,Q)의 왼쪽 이웃열의 레이블 값을 참조한다.

$$L(x,y)=\begin{cases}L(x-1,y) & \text{if } L(x-1,y)>0\\ 0 & \text{if } L(x-1,y)=0\end{cases}$$

$$M=\{L(x,y)/P\le y\le Q\}$$

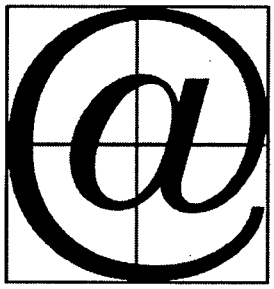
단계4. M의 원소가 0하나이면 구간(P,Q)에 NewLable을 부여하고 M이 0이외 두 개 이상의 원소를 가지면 하나의 원소를 선택하고 나머지는 합성한다.

(그림 6)에서는 위 식에서 획득한 정보를 이용하여 연결 화소를 이용하여 문자 분리 과정의 예를 보인다. (그림 6)(b)에서는 문자화소의 런값 P,Q와 P',Q' 값이 나중에 하나의 레이블로 합성되는 것을 볼 수 있고 (그림 6)(c)와 같은 결과를 얻는다. 이러한 각 레이블 값은 하나의 연결 화소 블록을 가지게 된다. 이렇게 추출된 연결 화소들을 분리 과정을 통하여 겹친 문자와 접촉 문자 등을 분리할 수 있다. 우선 접촉 문자를 분리하기 위해서는 분리할 블록을 선택해야 한다. 분리할 블록을 선택하는 것은 분리를 하는 방법보다 더욱 중요하고 본 논문에서 제안한 분리 블록을 선택하는 경우는 한 행에 대한 평균 문자 블록 가로폭의 최빈수(mode)를 기준 블록 폭이라고 하고 그 폭에 임계치를 주어서 그 이상이 되는 블록을 분리할 블록으로 판단하고 세로로 투영하여 얻은 문자 화소수들 중 최소 화소의 수를 가진 위치를 문자간 분리 위치로 선정하고 분리된 블록들에 대해서 다시 분리 될 블록의 조건에 맞는지를 확인한다.

2.5.2 E-mail 영역 검출

각각의 텍스트영역에 대한 레이블 테이블이 만들어 졌으면 하단부에 위치하고 있는 텍스트영역부터 검사를 한다. 아래쪽 영역부터 조사하는 이유는 명함에서 E-mail 영역은 하단부에 위치하는 경우가 많다. 그 만큼 아래쪽 영역부터 조사하는 것이 더 빠른 검색이 되기 때문이다. 그리고 아직까지는 E-mail 주소에 한글이 들어가는 경우는 없기 때문에 한글의 특징을 이용해서 작업 할 텍스트 영역을 제거할 수 있다. 즉 한글은 초성 중성 종성으로 이루어져 있기 때문에 레이블 테이블의 레이블 블록을 세로 방향으로 탐색 하여 3개 이상의 레이블 값이 겹치는 경우는 한글이 포

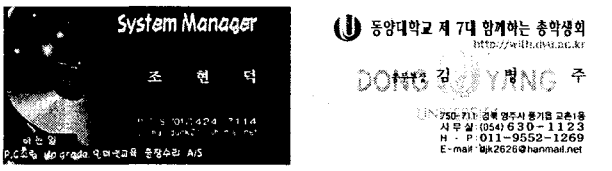
함된 행이므로 그 곳의 텍스트영역은 제외시킬 수 있다. 마지막으로 남아있는 텍스트영역의 레이블 테이블에 있는 레이블 중 최소값을 갖는 레이블 블록부터 순차적으로 Boundary Box을 씌워서 (그림 7)처럼 중심을 기준으로 수평, 수직 방향 즉 4방향으로 탐색을 하여 명암 변화 빈도수가 각 방향에서 2번씩 일어 날 때 at symbol이라 판단 할 수 있다.



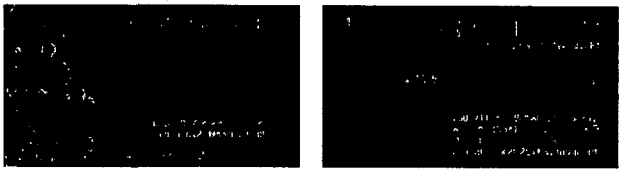
(그림 7) 수평, 수직방향 탐색

3. 실험 결과

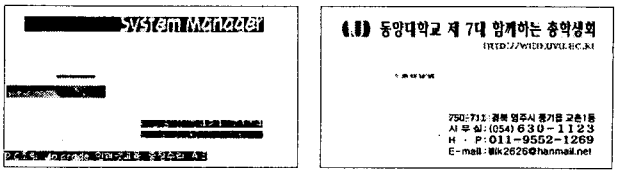
본 논문에서 제안한 방법을 테스트하기 위한 환경으로 Pentimu-III 700Mhz, RAM 256MB, Visual C++ 6.0을 이용하였으며, 실험 명함 이미지는 300dpi 의 해상도로 스캔한 다양한 형태의 명함 이미지 20장을 이용했다.



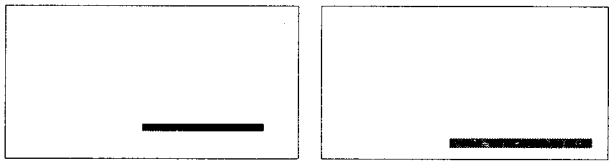
(그림 8) 원본 이미지



(그림 9) Edge 검출영상(소벨마스크)



(그림 10) 텍스트영역 추출



(그림 11) E-mail 영역 추출

위의 그림에서 보듯이 문자의 명암 값과 크기, 배경에 상관없이 E-mail 영역이 추출되는 것을 볼 수 있다. (그림 9)는 (그림 8)의 원본 이미지를 명암 간소화를 시킨 결과를 소벨마스크를 이용해 Edge 검출을 한 결과이다. 그림에서 보듯이 명암 간소화는 문자영역을 다치게 하지 않는 범위내에서 배경명암 값과 유사한 명암 값을 배경과 결합시키는 효과를 볼 수 있다. (그림 10)은 텍스처 특성을 분석하여 텍스트 영역을 분할한 결과이며 문자열 최소 길이를 임계치로 후보 영역을 제거한 결과로 적은 수의 문자열을 제거 한 것을 볼 수 있다.

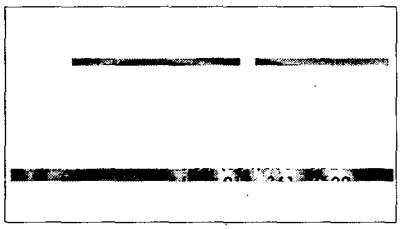
마지막으로 (그림 11)은 텍스트 영역 중에서 at symbol을 찾아 E-mail 영역을 검출한 결과를 보여 주고 있다. 이 실험에서 알 수 있듯이 그림과 문자들이 조합된 명함이미지에서도 정확히 E-mail 영역을 검출해 낼 수 있다는 것을 볼 수 있다.

하지만 (그림 10)에서도 볼 수 있듯이 텍스처 영역에 아직 남아있는 배경그림을 볼 수 있다.

이 배경은 나중에 처리할 개별문자 분리와 at symbol 검출에는 큰 영향을 미치지는 않지만 연산하는 속도에 약간에 영향을 미칠 것 이라 생각된다.

(그림 12) 실패한 명함 영상

(그림 13) Edge 검출영상(소벨마스크)



(그림 14) 텍스트영역 추출

(그림 12)는 E-mail 영역을 검출하지 못한 명함이미지로 배경이 전반적으로 물결무늬의 복잡한 형태로 되어있고 문자의 명암 값이 배경의 명암 값과 거의 차이가 없어 사람의 눈으로 보아도 식별해 내기가 까다로운 경우이어서 (그림 13)과 같이 많은 수의 Edge가 검출되어서 오 동작을 하게 되었다. (그림 14)는 (그림 13)영상에서 텍스트 영역을 검출 해낸 결과이다. 보는 바와 같이 텍스트 영역을 제대로 추출하지 못 하였다.

결과적으로. 실험한 20장의 명함 이미지 중 1장을 제외한 나머지 19장에서는 정확히 E-mail 영역을 검출해 내어 95%의 성공률을 보였다.

이번 실험 중 여러 가지 상황을 다 테스트 해보지는 못했지만 명함 이미지가 많이 손상되어 있거나 입력과정에서 문자 영역에 많은 잡음이 들어 갔을 때 는 영상을 향상 시키는 전처리 과정이 필요 할 것이다.

4. 결론

본 논문에서는 문자의 명암 값과 크기에 무관하게 복잡한 배경에서 문자영역을 분할하고 겹친문자에 상관없이 E-mail 영역을 검출하는 방법을 제안하였다. 본 논문에서 제안한 E-mail 검출 알고리즘은 영상에서 기울기 특징과 텍스처 특성을 분석하고 연결화소을 이용하여 개별문자를 추출 후 4방향 탐색을 통해 at symbol의 특징인 원안에 원이 존재하는 특징을 정확하게 검출해 내는 방법을 제안하였다. 본 논문은 기존 연구에서 문제시되고 있는 문자 영역 검출의 하나인 E-mail 영역 검출방법을 제안하였으며 보다 신뢰성 있는 결과를 얻기 위해 다양하고 많은 자료를 바탕으로 한 실험과 수정이 필요하고 여러 가지 상황에 따라 발생할 수 있는 잡음과 영상의 기울어진 문제점들이 고려해야 할 것 이다.

참 고 문 헌

[1] 임문철, 김우생, "비디오 분석을 위한 자막 프레임구간과 자막영역 추출", 한국정보처리학회 논문지, 제7권 11호, 2000.

[2] 김의정, 김태균, "오프라인 문서에서 개별문자 추출과 한자 인식에 관한 연구",한국정보처리학회 논문지, 제4권 5호, 1997.

[3] 임문철, 김우생, "비디오에서 색상변화 빈도수를 이용한 자막영역 추출기법", 한국멀티미디어학회, '99춘계학술발표 논문집, 제2권 1호,pp121-126, 1999.

[4] R. Lienhart and F. Stuber, "Automatic Text Recognition in Digital Videos", In Image and Video Processing IV 1996, Proc. SPIE 2666-20, January 1996.

[5] R. Lienhart, "Automatic Text Recognition for Video Indexing", In Proceedings of the ACM Multimedia 96, (Boston, MA, USA, November 11-18, 1996), S.11-20, November 1996.

[6] S. Liang, M. Ahmadi, M. Shridhar, "Segmentation of Characters in Printed Document Recognition", Proceeding 2nd International Conference on Document Analysis and Recognition, pp569-572, 1992.

[7] S. Tsujimoto, and H. Asada, "Resolving Ambiguity in Segmenting Touching Characters", Proceeding 1st International Conference on Document Analysis and Recognition, pp701-709, 1991.

[8] 이도엽, 김형재, 배익성, 이철희, 차의영, "변형된 Run Length Coding 기법을 이용한 이치화된 자동차 번호판 영상에서의 문자분리", 한국멀티미디어학회, '춘계학술발표 논문집, 1998

[9] 장명욱, 천대녕, 양현승, "연결화소를 이용한 문서 영상의 분할 및 인식", 한국정보과학회 논문지, Vol. 20, No. 12, pp.1741-1751, 1993