컬러코드를 이용한 스캔 문서 분류 자동화

안상길*, 최병욱*
*한양대학교 전자컴퓨터통신공학과
e-mail : ask1004@hanyang.ac.kr

Automating Scanned Document Classification Using ColorCode

Sang-Kil Ahn*, Byung-Uk Choi*
*Dept. of Electronics Computer Engineering, Hanyang University

요 약

디지털 형태의 문서가 널리 퍼지고 끊임없이 증가함에 따라 이를 자동으로 가공하고 처리하는 문서자동분류의 중요성이 널리 인식되고 있다. 본 논문에서는 복합기에서 컬러코드를 인식하는 모듈을 탑재하여 스캔된 문서를 자동으로 분류하는 시스템을 제안하고자 한다. 복합기에서 컬러코드가 부착된 종이문서를 스캔한 다음 그 컬러코드를 추출하여 인식하고 해당 컬러코드와 관련된 문서관리정보에 따라 스캔문서를 복합기 내부의 지정 폴더에 저장하거나 다른 곳으로 전달하는 시스템이다. 이렇게 함으로써 종이문서를 전자화하는 과정에서 수작업으로 분류하는 시간을 줄일 수 있고 또한 사람에 의한 오류를 줄일 수 있다는 장점이 있다.

1. 서론

기업 내의 비효율적인 업무 중 하나는 자료의 부재 또는 부실에 따른 것이다. 자료의 양이 증가함에 따라 자료를 찾는 시간이 늘어나고 자료를 찾지 못해 다시 만드는 경우도 허다하다. 통합적인 자료관리의 부재가 그 원인이다. 금융기관이나 법원의 경우 종이문서를 사용하는 경우가 많아 그에 대한 보관비용과 인건비에 대한 부담은 클 수 밖에 없다. 이러한 불편함을 없애기 위해 2005 년 10 월부터 시행된 전자거래기본법 개정으로 전자문서의 법률상 효력이 명확해지고 이를 보관하기 위한 공인전자문서보관소 제도가 도입되었다[1]. 이로써 종이문서 보관과 관련된 불필요한 비용과 번거로움이 점차 사라지고 전자화된 데이터를 활용하는 다양한 솔루션과 비즈니스 모델을 활성화시킬 수 있는 기회가 만들어졌다.

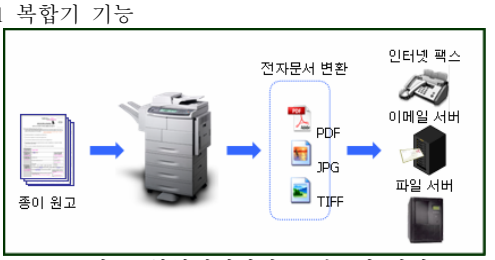
문서를 효율적으로 관리하기 위해서는 자료검색 기능이 잘 갖춰져 있어야 하며 이를 위해서는 체계적인 문서분류가 필수적이다. 문서의 자동분류는 1960 년대 정보검색의 한 분야로 연구되기 시작하였다. 1980 년대 말까지는 주로 이론적인 연구에 머물러 있었으며, 실제 응용시스템도 전문가가 수작업을 통해 생성해낸 규칙을 기반으로 한 방법을 통해 주로 구현되었다. 그러나 1990 년대에 접어들어 컴퓨터가 널리 보급되고 인터넷이 발전함에 따라 디지털 형태의 정보가 급격히 증가하기 시작하여 정보의 과잉현상이 나타나게 되었다. 따라서 많은 양의 정보를 자동으로 가공하여 분류하는 문서자동분류 분야의 중요성이 널리 인식되기 시작하였으며, 현재에 이르기까지 다양한 이론과 방법들이 깊이 있게 연구되고 있다[2].

본 논문에서는 복합기에서 스캔문서를 분류하는데 있어 종이문서의 고유양식 정보를 이용하며, 컬러코드를 이용하여 그 고유양식을 구분하고 스캔된 문서를 자동으로 분류하는 시스템을 설계하고자 한다. 일반적으로 복합기를 이용하여 종이문서를 스캔하여 전자문서 형태로 보관할 경우, 복합기에서 스캔한 문서에 아무런 의미 없는 일련번호로 파일이름을 붙여 저장한다. 따라서 이 파일들을 적절히 분류하고 다른 저장소로 배치하기 위해서는 사람의 수작업이 필요하며 일일이 내용을 다시 확인해야 하는 불편함이 있다. 이런 불편함을 해결하기 위해 종이문서에 그 문서양식에 해당하는 컬러코드를 미리 인쇄하거나 별도로 부착하여 복합기에서 그 컬러코드를 인식하고 해당 문서관리정보에 따라 자동 분류하는 시스템을 제안한다. 여기에서 다른 바코드 대신 컬러코드를 사용하는 이유는 컬러코드가 다른 바코드에 비해 인식하기 쉽고 인식률 또한 높다는 장점이 있기 때문이다.

2. 관련 연구

이 장에서는 일반적인 복합기에서의 스캔문서 처리 기능과 컬러코드에 대하여 소개하고자 한다.

2.1 복합기 기능

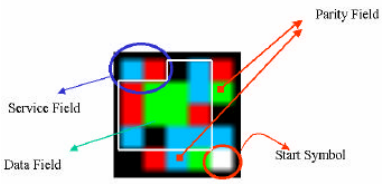


(그림 1) 복합기에서의 스캔문서 처리

복합기는 인쇄, 복사, 스캔, 팩스 등의 기본 기능을 가지고 있으며, 최근에는 다양한 네트워크 프로토콜을 지원하고 또한 하드 디스크를 내장함으로써 강력한 문서관리기능을 제공한다. 종이문서를 전자문서로 디지털화하는 스캔 기능은, 복합기에 탑재된 스캐너를 통해 종이문서를 이미지로 스캔하여 파일 형식에 따라 적절한 압축을 수행한 다음 파일로 변환하여 메모리에 저장하고 그 파일을 다시 사용자가 원하는 곳으로 전송할 수 있다. 이때 파일 이름은 정해진 규칙에 따라 만들어지는데 사용자가 복합기의 GUI 패널을 통해 직접 입력할 수도 있다.

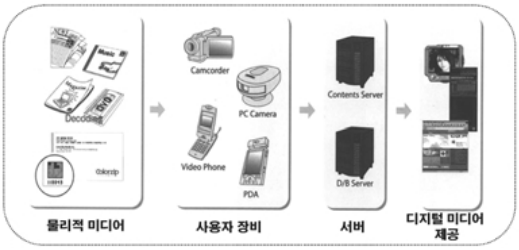
2.2 컬러코드

일반적인 2D 이미지 코드와 차별성을 갖는 컬러코드는 1999 년 연세대학교에서 개발하였으며, 저가 장비에서도 인식이 가능하며 다양한 디자인을 가지고 있다. 코드를 인식하는 방법, 매트릭스 형태, 패리티 형태 등에 따라 다양한 종류의 코드가 있으며, 적용되는 분야의 특성에 맞게 코드가 사용된다[3].



(그림 2) 컬러코드 구조

컬러코드는 임의의 크기로 사용할 수 있는데, 일부 셀을 패리티 영역이나 참조 영역으로 이용된다. 위의 그림처럼 5x5 크기인 경우 실제 데이터는 14 셀을 이용할 수 있다[4].



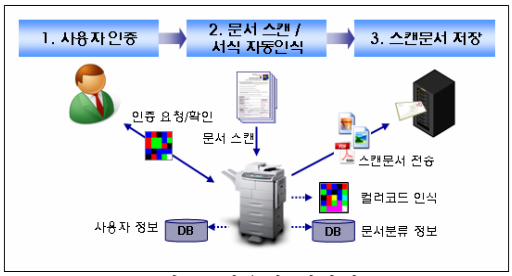
(그림 3) 컬러코드 서비스

컬러코드는 다양한 분야에 적용되어 사용되고 있으며, 특히 저널 분야, 교육 분야, 엔터테인먼트 분야, 개인정보 서비스, 광고 분야, 모바일 분야 등에 사용된다[5].

3. 스캔 문서 분류 자동화 시스템 설계

스캔 문서 분류 자동화 시스템은 컬러코드에 따라 스캔된 문서를 자동으로 분류해 주는 기능을 제공한다.

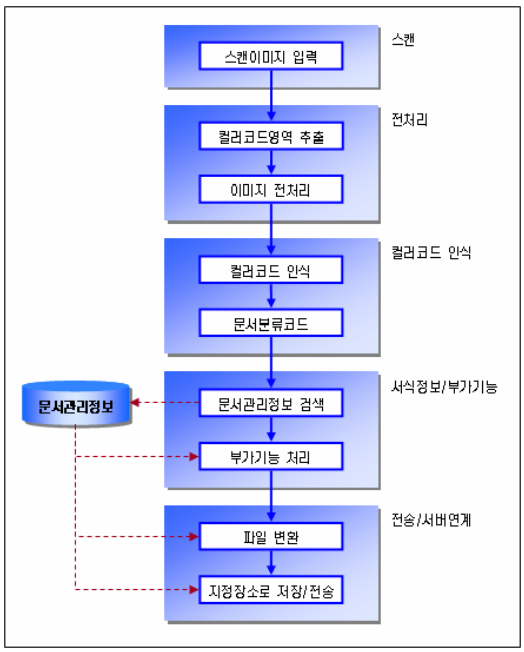
3.1 시스템 설계



(그림 4) 사용자 시나리오

먼저 사용자는 복합기에서 사용자 인증을 해야 하는데, 명함에 인쇄된 컬러코드를 스캔하여 사용자 ID를 인식한 다음 GUI 패널에서 패스워드를 입력한다. 복합기는 입력된 사용자 정보를 내부 사용자 DB 에서 확인하고 스캔 기능의 사용을 승인한다. 사용자가 종이문서를 올려놓으면 한 장씩 스캔하여 각 페이지에 부착된 컬러코드를 인식하고, 문서관리정보 DB 에서 문서분류정보에 따라 자동으로 분류한다. 모든 스캔이 끝나면 GUI 패널에 스캔한 이미지와 함께 자동 분류된 결과를 보여주고 사용자로부터 확인을 받는다. 사용자는 분류 결과가 잘못되었을 경우 수정할 수 있으며 문제가 없으면 복합기 내부에 저장하거나 다른 파일 서버로 전송한다.

스캔 문서에 대한 좀더 상세한 처리는 아래와 같다.



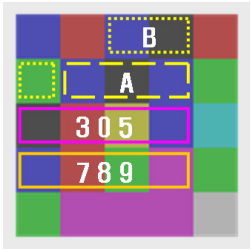
(그림 5) 스캔문서 처리 흐름도

종이문서를 스캔한 이미지는 복합기 내 메모리에

저장된다. 그런 다음 컬러코드 영역을 추출하여 노이즈 제거나 기울임 보정 등의 전처리를 하고, 에지 검출을 통해 컬러코드의 크기를 한 다음 각 셀의 컬러 값을 인식하여 정해진 코드체계에 따라 디코딩한다. 디코딩된 코드 값은 문서양식의 고유 ID 로 이 값을 이용하여 문서를 자동으로 분류할 수 있다. 이 문서 ID 를 문서관리정보 DB 의 인덱스로 이용하여 해당 정보를 검색하고 그 정보에 따라 추가적인 부가기능을 수행한다. 예를 들면, 문서 내 특정 영역을 추출하여 OCR 을 처리한 다음 메타데이터를 생성하는 것이다. 마지막으로 스캔 이미지 및 관련 추출정보를 한 파일 또는 여러 파일로 변환한 후 지정된 내부 하드디스크 폴더에 저장하거나 이메일 등을 이용하여 외부 파일 서버로 전송한다.

3.2 문서분류코드 및 컬러코드 설계

컬러코드의 크기는 가변적이므로 문서분류코드에 맞게 정하면 된다. 그림 2 에서의 데이터 영역인 14 셀을 하나의 문서양식번호로 직접 매핑할 수도 있지만, 일반적인 문서분류는 트리 형태의 구조를 갖는다. 여기에서는 문서분류를 대분류, 소분류, 일련번호와 같이 3 단계로 나누고, 대분류에는 아스키 2 문자를, 소분류와 일련번호는 3 자리 숫자를 사용한다. 예를 들어 은행에서 새로운 계좌를 개설하기 위한 신청서에 대하여 대분류는 은행, 소분류는 계좌개설신청서, 일련번호는 신청순서로 분류하고 각각에 해당하는 코드를 “BA”, “305”, “789”와 같이 지정할 수 있다.



(그림 6) 문서분류코드와 컬러코드

셀 컬러로 8 색을 사용할 경우 셀마다 3 비트를 표현할 수 있다. 아스키 문자 하나를 표현하기 위해서는 8 비트가 필요하고 컬러코드에서는 3 셀이 사용된다. 그리고 3 자리 십진수를 표현할 때에는 최대값이 999 이므로 10 비트가 필요하고 컬러코드에서는 4 셀이 사용된다. 따라서 “BA-305-789”의 경우 총 14 셀로 표현이 가능하다.

4. 실험 및 평가

스캔 문서 분류 자동화 시스템을 구현하는데 가장 중요한 부분은, 복합기 모델에 상관없이 컬러코드를 정확히 인식해야 한다는 것과 인식된 컬러코드 값에 따라 많은 데이터에서 해당 정보를 빠르게 찾는 것이다.

4.1 컬러 인식

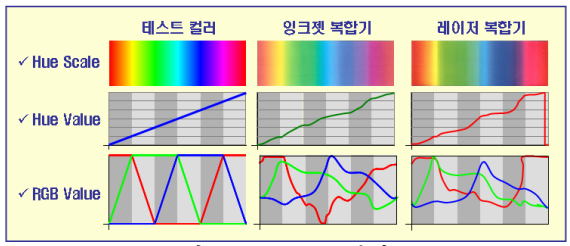
컬러코드를 인쇄한 다음 다시 스캔할 경우 컬러 값은 많이 달라진다. 특히 프린터의 경우 RGB 컬러 대신 CMYK 컬러를 사용하기 때문에 색역 자체가 줄어들고 잉크나 토너의 색상에 영향을 많이 받는다. 또한 프린터마다의 서로 다른 스크리닝 방식에도 영향을 받는다. 따라서 일반 스캐너처럼 RGB 컬러에 대한 정규화만으로는 정확한 컬러를 구할 수 없다.

기존의 연구에서는 컬러를 인식할 때 주로 RGB 값을 사용했는데 얼마나 정확하게 인식할 수 있는지 실험하였다. 실험은 컬러 테스트 이미지를 잉크젯 복합기인 HP PhotoSmart 2610 과 레이저 복합기인 삼성 CLX-2161K 에 600 dpi 로 인쇄한 후 200 dpi 로 스캔한 이미지를 사용하였다. 아래는 컬러코드에 사용되는 기본 8 가지 색을 인쇄한 결과이다. 육안으로 봤을 때 레이저 복합기가 잉크젯 복합기보다 좀더 밝고 선명하게 보인다.

<표 1> 기본 8 색 인쇄 결과

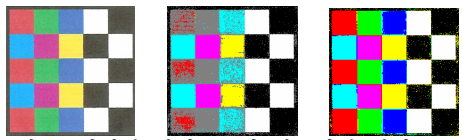
	잉크젯 복합기	레이저 복합기
인쇄 결과		
60배 확대		

아래 그래프는 Hue 컬러의 인쇄 결과에 대한 Hue 값과 RGB 값을 그래프로 분석한 것이다. 임의의 컬러에 대한 Hue 값은 어느 정도 선형적으로 비례하지만 RGB 값은 차이가 많이 난다는 것을 알 수 있다.



(그림 7) Hue / RGB 분석

아래 그림에서 두번째 컬러코드는 기존 연구에서의 RGB 을 수정없이 적용한 결과이며, 세번째 컬러코드는 Hue 값을 단순히 6 등분하여 처리한 결과이다.

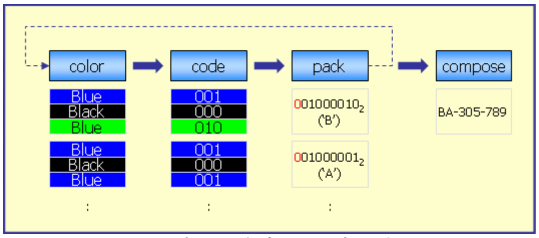


(그림 8) 입력된 컬러코드와 정규화된 컬러코드

복합기에 따라 인쇄할 때 또는 스캔할 때 컬러가 달라지므로 그에 영향을 덜 받는 컬러모델을 선택하는 것이 중요한데 실험을 통해 RGB 값보다는 Hue 값을 이용하는 게 훨씬 더 좋다는 것을 알 수 있다.

4.2 컬러코드 디코딩

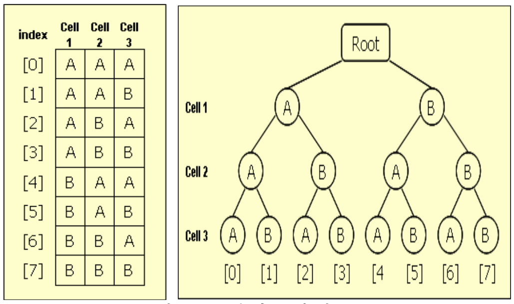
컬러코드를 이용하기 위해서는 항상 인코딩과 디코딩이 필요하다. 디코딩의 경우는 아래 그림처럼, 먼저 셀 컬러를 코드값으로 바꾸고 여러 셀의 코드값을 하나로 합쳐 단위 코드값을 만들고 다시 여러 개의 단위 코드값을 합쳐 하나의 컬러코드 값을 만들어야 한다. 특히 사용자 인증이나 컨텐츠 서비스처럼 컬러코드가 빈번히 사용되는 경우 디코딩하는데 많은 시간이 소요된다.



(그림 9) 컬러코드 디코딩

어떤 값을 컬러코드로 인코딩할 때 인코딩된 셀 컬러들을 컬러 문자열로 표현하여 저장하고, 나중에 컬러코드를 해석할 때 셀 컬러에 따라 그 문자열을 검색한다면 디코딩하는 시간을 단축할 수 있다. 그림 6에서 나온 예처럼 "BA-305-789"를 인코딩하면 컬러 문자열 "BRBKR-GBKBG-KRYBC-BRGMG-GMMMW"로 표현할 수 있다. (R=Red, G=Green, B=Blue, C=Cyan, M=Magenta, Y=Yellow, W=White, K=Black)

아래 그림은 컬러코드의 크기 S 가 3 셀이고 컬러 C 는 2 가지만 사용할 경우 표현할 수 있는 컬러코드의 조합($N = C^S = 8$)을 나타낸다.



(그림 10) 컬러문자열 구조

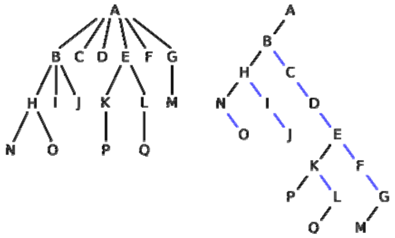
왼쪽 그림처럼 하나의 컬러코드마다 하나의 컬러 문자열을 할당되어 문자열 순으로 소팅된 경우, 이진 탐색을 이용하더라도 다음과 같은 시간이 소요된다.

$$O(t) = \log N * S = \log C^S * S = S^2 * \log C \ldots (1)$$

오른쪽 그림처럼 트리 구조로 만들면 좀더 시간을 줄일 수 있다.

$$O(t) = S * C \ldots\ldots\ldots\ldots(2)$$

만약 컬러 C 가 3 이상일 경우는 다음 그림 11 과 같이 이진 트리로 변환해 주어야 구현하기도 쉽다.



(그림 11) Encoding *n*-ary trees as binary trees

그리고 오른쪽 형제 노드를 다시 완전 이진 트리로 만들면 시간은 좀더 줄어든다.

5. 결론 및 향후 연구과제

본 논문에서 제안한 컬러코드를 이용한 스캔 문서 분류 자동화 시스템은, 사람의 수작업 시간과 오류를 줄이기 위해서, 컬러코드가 부착된 종이문서를 복합기에서 스캔할 때 그 컬러코드를 인식하여 문서를 자동 분류하는 기능을 제공하는 것을 목적으로 한다.

실험을 통하여 컬러코드를 인식할 때 RGB 값보다 Hue 값을 이용하는 게 훨씬 효과적이라는 것을 보였고, 디코딩 속도를 빠르게 하기 위해 컬러값 배열을 그대로 이용하는 방법을 제안하였다.

차후의 연구방향으로는, RGB 나 Hue 이외의 다른 컬러 모델을 이용한 컬러 인식에 대한 추가적인 검증과 작은 크기의 컬러코드도 정확히 인식할 수 있는 방법에 대한 연구가 필요하다.

참고문헌

[1] 공인전자문서보관소, http://www.ceda.or.kr
[2] 이지행, 조성배, "전자우편 문서의 자동분류를 위한 다중 분류기 결합", 정보과학회논문지:소프트웨어 및 응용 제 29 권 제 3 호, pp.192-201, 2002
[3] 컬러코드, http://www.colorzip.com
[4] 신은동, "컬러 태그를 이용한 새로운 인터넷 인터페이스 설계 및 응용", 연세대학교 대학원 컴퓨터과학 석사학위, 2000
[5] 한탁돈, "컬러 코드", TTA 저널 제 84 호, pp.104-110, 2003
[6] 조맹섭, 디지털 컬러 프로세싱, 도서출판 국제, 2006