신문광고영상 데이터베이스구축을 위한 유사영상 분류 시스템

Development System of Mimicking Image Classification for Newspaper Advertisements Database Construction

김기현, Kihyun Kim*, 김광태, Kwangtae Kim**, 박현우, Hyunwoo Park**,
이동훈, Donghoon Lee**, 윤태수, Taesoo Yun***

요약 본 논문에서는 광고영상에서 동일한 광고의 다수 매체(신문, 잡지)에 실리는 판형을 효율적으로 관리하는 데이터베이스 시스템을 구축하기 위한 유사광고를 분류하는 시스템을 제안한다. 현재, 신문광고를 데이터베이스화하는 작업은 사람이 직접 매체를 하나씩 스캐너를 이용하여 영상데이터를 획득한 후 포토샵이나 그림판과 같은 이미지 편집 툴을 이용하여 광고영역을 잘라내고 저장하고, 각 광고에 따른 날짜정보, 매체정보, 페이지정보, 광고가 실린 면의 종류, 크기정보 등을 일일이 기록, 저장하기 때문에 비능률적이고 비효율적인 업무형태로 많은 시간과 인력의 낭비를 초래하고 있다. 이러한 문제를 해결하기 위하여 디지털 카메라를 이용하여 신문영상을 획득하고, 영상 전처리 과정을 통하여 광고후보영역을 추출하며, 신문매체광고가 가지는 특성에 따라 광고후보영역을 분류한다. 따라서 본 시스템은 모든 광고영상의 유사성을 비교하여 신규광고인지, 기존의 광고인지를 분류하여 데이터베이스화 한다.

Abstract In this paper, we propose a system of mimicking image classification for building database system. It better manages the format which recording the same advertisements multimedia in advertisements image. Recently, the work of converting database is made directly by the people. This work doses the media scanning, image editing and saving, and saving of advertising information (date, the media, the page and size, so far). Therefore, it is wasted a lot of time and manpower as inefficient business. To solve these problems, first of all we gain an image by digital camera, extract and classify candidate area of advertisements. Accordingly, our system saves database to comparison of the mimicking of all advertising and classify whether area of image is the new or existing advertising.

핵심어: Newspaper Advertisements, Mimicking Image Classification, Database, Image processing

1. 서론

현대 사회에서 정보 전달의 중요한 매체이며, 막대한 양의 정보기록 문서인 신문이나 잡지에는 막대한 양의 다양한 광고가 게재되고 있다. 그리고 수많은 광고들은 소비자 분석과 제품분석, 경쟁사 분석을 위해 파일로 변환하여 저장을 하고 데이터베이스화하는 작업이 이루어지고 있다[1]. 그러나 매체에 게재된 수많은 광고들을 파일로 변환하고 데이터베이스화하는 작업은 쉬운 일이 아니다. 현재, 광고를 데이터베이스화 하는 작업은 사람이 직접 매체를 하나씩 스캐너를 이용하여 영상데이터를 획득한 후 포토샵이나 그림판과 같은 이미지 편집 툴을 이용하여 광고영역을 잘라내고 저장하고, 각 광고에 따른 날짜정보, 매체정보, 페이지정보, 광고가 실린 면의 종류, 크기정보 등을 일일이 기록, 저장하는 실정이다. 이러한 과정은 비능률적이고 비효율적인 업무형태

*주저자 : 동서대학교 디자인&IT전문대학원 영상콘텐츠학과 석사과정 email: khkim@dit.dongseo.ac.kr
**공동저자 : 동서대학교 디자인&IT전문대학원 영상콘텐츠학과 석사과정 email: ktkim@dit.dongseo.ac.kr
**공동저자 : 동서대학교 RIC 센터 연구원 email: phw1010@gdsu.dongseo.ac.kr
**공동저자 : 동서대학교 디자인&IT전문대학원 영상콘텐츠학과 교수 email: dhl@dongseo.ac.kr
***교신저자 : 동서대학교 디자인&IT전문대학원 영상콘텐츠학과 교수 email: tsyun@dongseo.ac.kr

로 많은 시간과 인력의 낭비를 초래하고 있다(그림1 참조). 또한 신문매체의 경우 일반 스캐너로는 스캔 받을 수 없는 크기이므로 대형 스캐너 구입으로 인한 과다한 데이터베이스 구축비용이 발생하고 있다.

그림 1. 광고분류업체의 작업방식

기존의 연구에서는 대부분 입력된 문서영상으로부터 문자영역과 비문자 영역만을 추출하여 문서를 분류하는 형태이므로 문서로부터 특정 이미지 검출이나 공문서의 구조적 유형 분석과 같은 다양한 응용에는 한계가 있다[3]. 특히 신문이나 잡지와 같은 정보매체를 유형별로 자동으로 인식하고 분류하고 저장되는 연구는 미흡하다.

따라서 본 논문에서는 다양한 광고매체를 카메라 영상으로 입력 받아, 광고 후보영역을 추출하고, 추출한 후보영역을 광고의 패턴 특성을 이용하여 광고 영역으로 분류한 다음, 분류된 광고 이미지를 기존 데이터베이스와 비교·검색한 후 새롭게 분류된 기존의 광고와 유사도를 검사하여 기존에 존재하는 광고와 같은 광고인지 새로운 광고인지를 데이터베이스에 등록시킴으로써, 담당인력 축소 및 처리시간을 단축시킬 수 있는 시스템을 제안한다.

2절에서는 시스템이 어떻게 구성되어 있는지에 관하여 설명을 하고, 3절에서는 유사영상분류시스템을 구현한 방법에 관해서 서술한다. 그리고 4절, 5절은 실험결과와 결론 및 향후과제를 서술한다.

2. 시스템 구성

본 논문에서 소개하는 시스템은 크게 영상추출모듈, 영상분류모듈, 유사광고영상비교모듈의 세 가지로 구성된다(그림 2 참조). 영상추출모듈은 영상처리기술을 이용하여 입력받은 광고매체영상에서 광고후보영역을 추출하는 모듈이며, 영상분류 모듈은 추출된 광고후보영역을 기사와 광고영상을 구분하여 광고영상을 추출하는 영역이며, 유사광고영상비교모듈은 저장 된 데이터베이스를 검색하여 비교할 광고영상의 후보를 선별한 후, 선별 된 영상들끼리 유사영상인지를 비교하는 모듈이다. 그리고 위 과정들에 대한 광고영상비교 결과를 데이터베이스에 저장한다.

그림 2. 시스템 구성도

3. 유사영상분류시스템 구현

3.1 영상 획득을 위한 환경설정

신문영상을 얻기 위한 방법에는 스캐너를 이용한 스캔방식과 디지털 카메라를 이용하여 얻는 방법이 있다. 스캐너를 이용한 방식은 신문영상의 정확한 데이터를 얻을 수 있지만, 신문지 크기에 맞는 스캐너를 사용하려면 스캐너의 구입비용이 많이 든다. 그리고 신문지 크기를 스캔하려면 많은 시간이 걸리기 때문에 하루에도 수십 부씩 쏟아져 나오는 신문매체를 모두 스캔하기에는 힘들다. 카메라를 이용하는 방식은 신문지를 전체를 볼 수 있는 높이에서 촬영을 한 후 영상데이터를 바로 뽑아 낼 수 있기 때문에 신문영상을 획득하는데 많은 시간이 소요 되지 않고, 질 좋은 영상을 획득할 수 있다. 본 시스템에서는 영상을 획득하기 위한 카메라의 역할을 수행하는 장비로 캐논 EOS 40D를 사용하였으며, 렌즈의 화각은 1:2.8의 넓은 화각을 촬용할 수 있는 렌즈를 사용하였다(그림3 참조).

그림 3. 카메라를 이용한 신문 영상 촬영

3.2 영상추출 모듈

3.2.1 영상전처리 단계

인식하고자하는 광고이미지 외에는 모두 잡음이므로 식별자와 배경을 분리하는 작업이 필요할 뿐만 아니라, 영상의 종류가 다양하고 임계값이 일정하지 않기 때문에 광고영역 추출을 위한 전처리 단계로서 canny알고리즘[5]을 이용하여 윤곽선을 추출하여 물체와 배경을 분리하고, Thesholding으로 이진화하여 배경과 Object의 경계를 확실하게 분리한다(그림4 참조).

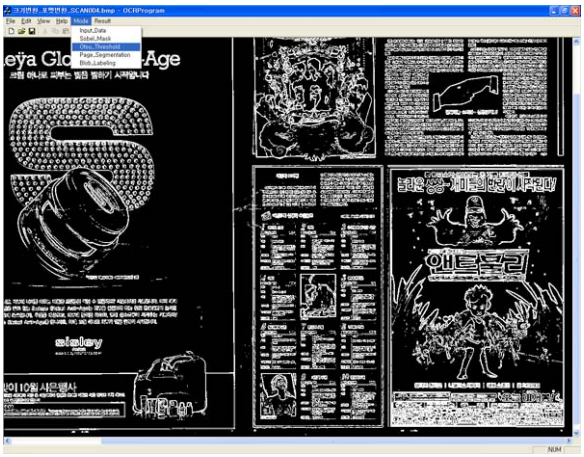


그림 4. 전처리단계완료영상

3.2.2 영상처리 단계

윤곽선을 추출한 영상에서 광고영역 추출을 위해 이웃하는 픽셀의 간격이 적정값일때 빈 픽셀을 채우는 Page Segmentation 기법과 채워진 블록들의 번호를 매기고 일정 크기가 아닌 즉 광고영역으로 분류되지 않는 블록을 제거하는 Blob Labeling 기법을 구현한다.

Page Segmentation[6] 기법은 영상처리 모듈에서 추출된 윤곽선정보로 근접해 있는 물체들을 연결시켜 하나의 블록을 만드는 것으로 예를 들어 한 라인의 이진화정보가 다음과 같이 있을 때 1001000010001100 1과 1사이의 간격 파라미터를 3으로 준다면 정보는 1111000011111111로 변환된다(그림5 참조).

Blob Labeling 기법은 블록화 된 물체들 중 광고영역과 아닌 영역을 추출하기 위해 사용하는 기법으로 일정한 크기의 블록(광고후보영역)일 경우 번호를 매겨 정보를 저장하고 나머지(잡음)는 제거하는 기능을 가진다(그림 6참조).

3.3 영상분류 모듈

추출된 광고영역들 중에는 광고이미지 외에 사설이나 기사사진 등의 잡음이 있을 수 있으므로 추출된 영역을 분류하는 단계가 필요하다. 다음에서는 매체상의 광고이미지의 특성을 분석하고, 광고영역을 분류할 수 있는 패턴을 정의한다. 매체광고의 기본적인 특성은 다음과 같다.

매체광고 전면 규격은 가로 12컬럼(37cm)과 높이 15단(51cm)이며, 매체 광고의 기본단위는 가로 1컬럼(약 3cm)×높이 1단(약 3.4cm)이다. 일반적으로 매체광고는 하단에 위치하며 면별로 크기가 정해져 있으며, 1면의 경우는 반드시 4단 광고이어야 하고, 2~5면, 사회면의 경우는 5단 광고를 게재해야 한다.

또한 기사나 사설과 같은 후보영역 같은 경우에는 글과 배경이 주를 이루기 때문에 광고와는 그 패턴 또한 틀리다. 그러므로 사설이나 기사 같은 경우에는 영상 전처리과정을 거친 후 Page Segmentation기법을 가로만 적용 시킨다. 다음으로 후보영역의 높이에 대하여 가로크기를 세 등분 하여 projection을 하면 검정과 흰색이 자주 반복 되는 패턴을 찾을 수 있다(그림7 참조). 이와 같은 과정을 거쳐 선별된 광고이미지는 인덱싱하여 광고가 실린 신문의 날짜, 광고영상의 크기, 광고가 실린 신문의 페이지번호, 면의 종류, 컬러 여부 등의 정보를 데이터베이스에 저장 한다.

그림 5. Page Segmentation

그림 6. Blob Labeling

(a)

(b)

그림 7. (a)는 기사를, (b)는 광고를 가로 방향으로 Page Segmentation한 영상

3.4 영상데이터베이스 Indexing 모듈

같은 광고 영상을 찾아야 할 때 모든 광고를 비교하면 오랜 시간이 걸리기 때문에 데이터베이스에 저장된 내용을 기반을 두어 검색을 함으로써 영상 검색 속도를 줄이기 위해서이다. 신문 광고에서의 데이터베이스 검색 방법은 다음과 같다.

① 같은 날짜이고 같은 매체명은 검색하지 않음; 한 매체에서 같은 광고를 두 번 이상 내지 않기 때문에 같은 매체명에 대해서는 검색을 하지 않아도 된다. 그러나 서로 다른 날짜일 경우에는 같은 매체에 같은 광고를 낼 수 있기 때문에 검색을 해야 한다.

② 광고 영상의 컬러 유무; 같은 광고 영상이라 하더라도 컬러와 흑백은 다른 광고 이므로 같은 색상의 영상만 검색한다.

③ 가로와 세로의 크기가 ±2cm 일 경우에만 검색; 광고영상이 보기에는 같은 광고라 하더라도 광고영상의 크기가 같지 않을 경우에는 다른 광고이므로 크기를 비교한다. 여기서 오차 범위를 2cm까지 두는 것은 광고영상의 한 블록의 크기가 최소 3.3cm이기 때문이다.

④ 이전에 검색 한 내용은 검색하지 않음; 이전에 이미 검색한 내용은 검색하지 않고 이전 검색 내용의 결과에 유사 광고의 횟수를 증가 시키면 됨으로 새로 입력 된 광고 영상만 검색한다.

3.5 광고영상 비교 모듈

광고영상의 각 화소들은 RGB 컬러 모델의 R,G,B 값으로 구성된다. 이 RGB 컬러모델은 명도, 채도와 같은 색상을 표현하는 고유의 값 외의 값도 포함하기 때문에 영상의 각 화소를 HSI 컬러 모델의 색상(HSI)값으로 변환한다. 각 화소들의 색상 값을 범위에 따라 그룹화하고, 영상의 각 화소를 HSI 값으로 변환하는 공식은 다음과 같다.

$$H = \cos^{-1}\left\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)+(R-B)(G-B)]^{1/2}}\right\} \quad (1)$$

$$S = 1 - \frac{3[\min(R, G, B)]}{(R+G+B)} \quad (2)$$

$$I = \frac{1}{3}(R+G+B) \quad (3)$$

위 식 (1)은 HSI 컬러 모델의 색상(HSI)을 계산하는 식으로 R, G, B 각각은 입력 영상의 한 화소에 대한 Red, Green, Blue 채널에 해당되는 값을 나타내며 범위는 0부터 255까지이다. 식 (2)는 채도(Saturation)를 계산하는 식으로 여기서 min(R, G, B)는 입력 영상의 한화소의 Red, Green, Blue 채널 중에서 가장 작은 값을 말한다. 식 (3)은 명도(Intensity)를 계산하는 식으로 입력 영상의 각 화소에 대한 Red, Green, Blue 채널의 평균값이다. 인간은 시각 정보를 인지하는 세 가지 지각 변수 중에서 채도나 명도보다는 색상 정보인 Hue에 더 민감하다. 따라서 본 시스템에서의 광고영상을 비교할 때 Hue의 값의 유사도를 이용하였다. 히스토그램의 성분 중 가장 높은 히스토그램을 비교하여 ±10의 오차 범위에서 유사도를 측정하고 유사도 측정 결과 같은 광고영상은 데이터베이스에 다시 저장된다.

4. 실험 결과

실험에 사용한 매체는 경향신문, 국민일보, 동아일보, 매일경제 4가지의 신문 매체를 4일치를 사용하였다. 먼저 신문매체를 카메라를 이용하여 영상을 획득 한 후 3절의 2번의 과정을 거쳐 광고후보영역을 선별하고, 3번의 과정을 거쳐 광고영역을 검출한 후 그 내용을 데이터베이스에 저장하였다(그림8 참조). 이와 과정을 진행하여 데이터베이스에 약 300개정도의 광고매체를 저장하였다.

위의 본문의 내용을 이용하여 조건을 만족하는 광고 영상을 데이터베이스를 이용하여 검색을 함으로써 짧은 시간에 총 300장의 이미지 영상 중 10장을 미리 선별 할 수 있었다(그림9 참조). 그리고 광고영상비교모듈을 이용하여 10장의 광고영상을 모두 비교하여 유사광고영상을 찾은 후 데이터베이스에 저장하였다. 300개의 모든 영상과 영상을 비교하여 데이터베이스에 저장하는 것보다 저장된 데이터베이스를 검색한 후 이미지를 검색하여 데이터베이스에 저장하는 것이 빠른 결과를 볼 수 있었다(그림10 참조).

5. 결론 및 향후과제

하루에 나오는 신문의 종류는 무수히 많고, 각 신문매체에는 막대한 양의 다양한 광고가 게재되어 있다. 수없이 막대한 광고를 스캐너를 이용하여 영상파일을 획득하고 포토샵이나 그림판과 같은 이미지 편집 툴을 이용하여 광고영상을 획득 분류하는 일은 많은 시간과 인력의 낭비를 초래한다. 하지만 본 시스템에서는 스캐너 장비가 아닌 디지털 카메라를 이용함으로써 신문영상데이터를 획득하는 시간과 비용을 줄였다. 그리고 신문영상에서 광고영역 후보군을 추출하고 분류함으로써 기존의 포토샵과 같은 이미지 편집 툴에서 사용하지 않고 짧은 시간에 분류, 저장을 함으로써 이미지 편집 툴의 전문가가 아니더라도 쉽게 작업할 수 있었으며, 이미지를 데이터베이스화 하여 이미지 정보에 대한 검색이 더욱 빨라졌다. 그리고 신규광고인지, 기존의 광고인지에 대한 유사이미지 검색알고리즘을 통하여 사람의 눈으로 일일이 비교하지 않고, 본 시스템으로 비교, 검색함으로써 사람이 직접 할 때보다 더욱 정확한 데이터베이스를 구축할 수 있었다.

향후 기사패턴과 광고패턴을 연구하고, 패턴인식에 적용하여 기사와 광고, 이미지를 분류할 수 있는 연구가 필요하다.

(a)

(b)

그림 8. (a)는 광고후보영역을 추출한 결과이고, (b)는 광고영역을 추출한 결과이다.

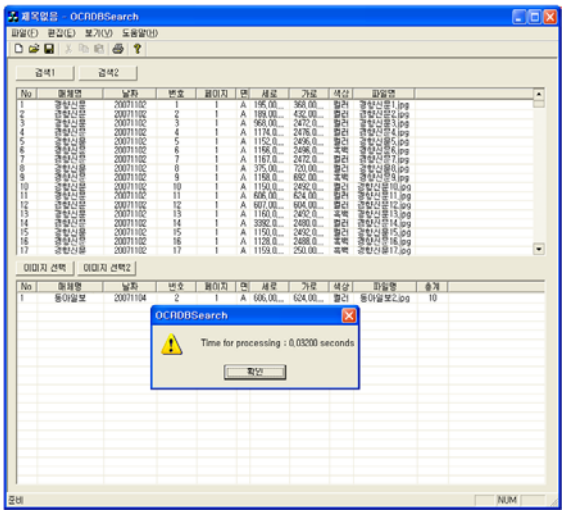
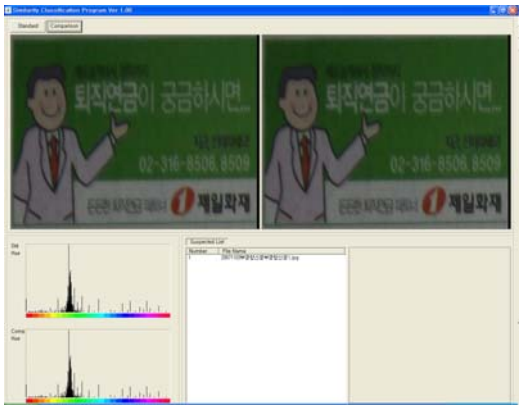
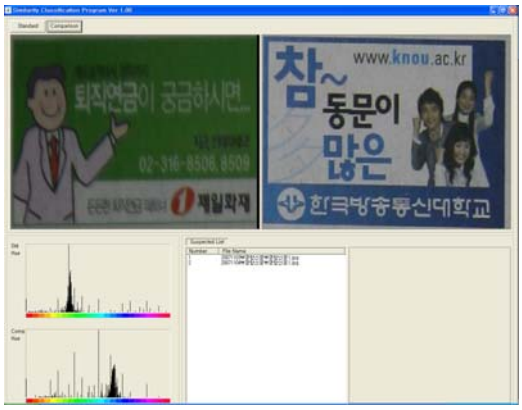


그림 9. 데이터베이스 검색결과



(a)



(b)

그림 10. (a)는 유사영상을 비교한 결과이고, (b)는 데이터베이스의 조건은 일치하지만 유사하지 않은 영상의 비교한 결과이다.

6. 참고문헌

[1] 이미영, "디지털 광고매체 가치에 대한 광고주 인식 조사 연구", 한국지역언론학회, 언론과학연구 제6권 2호, 2006. 6, pp. 221 ~ 255 (35pages)

[2] 이은주, 정성환, "Color N×M-grams를 이용한 영상 분류", 한국 정보처리학회 추계학술발표 논문집, 제5권, 제2호, pp.158-162, 1998.

[3] 모문정, 김욱현, "문서 영상의 영역 분류와 회전각 검출", 정보처리학회논문지B 제9-B권 제4호 2002.8.

[3] F. Y. Shih and S. S. Chen, "Adaptive document block segmentation and classification", IEEE Trans, Cybernetics, Vol.26, N0.5, 1996.

[4] J. L. Chen and H. J. Lee, "An efficient algorithm for form structure extraction using strip projection", Patten Recognition 31(9), pp.1353-1368, 1998.

[5] Z. lu, "Detection of text regions form digital engineering drawings", IEEE Trans, Pattern Analysis and Machine Intelligence, Vol.20, No.4, 1998.

[6] L. Y. Tseng and R. C. Chen, "Recognition and data extraction of form documents based on three types of line segments", Pattern Recognition 31(10), pp.1525-1540, 1998.

[7] J Canny, "A Computational Approach to Edge Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 6, 1986, pp. 679-698.

[8] Faisal Shafait, Daniel Keysers, Thomas M. Breue, "Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images ", ICPR 2006, International Conference on Pattern Recognition, pages 872-875