논문 2011-48CI-3-11

# 인터넷상에서 개인식별정보가 포함된 영상 검색을 위한 특징정보 분석에 관한 연구

## ( A Study on Features Analysis for Retrieving Image Containing Personal Information on the Web )

김 종 배*

( JongBae Kim )

요 약

정보통신 기술의 급격한 발전으로 인해 인터넷이 대중화됨에 따라 인터넷을 이용한 사이버 공간상에 정보의 상호교환, 전자상거래, 인터넷뱅킹 등의 사회 활동이 활발해지고 있다. 하지만, 인터넷 사용의 편리함을 추구하는 경향에 의해 개인식별용 증명서(주민등록증, 운전면허증, 여권, 학생증 등)들이 전자적인 매체로 표현되어 인터넷상에서 노출되는 경우가 빈번하게 발생하고 있다. 따라서 본 연구에서는 인터넷상에 노출된 개인정보가 포함된 이미지들을 효율적으로 검색하기 위한 방안을 제안한다. 제안한 방안은 이미지의 색상과 질감, 그리고 모양 특징정보들 중에서 개인식별정보가 포함된 이미지들에서 고유한 특징정보들을 분석하여 추출한 후 이를 이용하여 개인식별정보가 포함된 이미지들을 검색한다. 제안한 방안을 실험한 결과, 전체 개인식별정보가 포함된 이미지들 중에서 약 89%이상의 검색 성공률과 이미지 파일 당 수행시간은 약 0.17초가 소모되었다. 이러한 결과를 바탕으로 실제 인터넷상에서 개인식별정보가 포함된 이미지 파일들의 검색과 노출 여부 판단을 위한 시스템에 효과적으로 적용할 수 있다.

Abstract

Internet is becoming increasingly popular due to the rapid development of information and communication technology. There has been a convenient social activities such as the mutual exchange of information, e-commerce, internet banking, etc. through cyberspace on a computer. However, by using the convenience of the internet, the personal IDs(identity card, driving license, passport, student ID, etc.) represented by the electronic media are exposed on the internet frequently. Therefore, this study propose a feature extraction method to analyze the characteristics of image files containing personal information and a image retrieval method to find the images using the extracted features. The proposed method selects the feature information from color, texture, and shape of the images, and the images as searched by similarity analysis between feature information. The result which it experiments from the image which it acquires from the web-based image DB and correct image retrieval rate is 89%, the computing time per frame is 0.17 seconds. The proposed method can be efficiently apply a system to search the image files containing personal information and to determine the criteria of exposure of personal information.

**Keywords**: Content-based Image Retrieval, Features Histogram Matching Image Analysis

## Ⅰ. 서 론

최근 인터넷을 통해 홈쇼핑, 홈뱅킹, 교통편 예약 등 다양한 정보 서비스를 자신이 원하는 곳에서 제공받을 수 있게 되었다. 이러한 배경에는 1980년대 이후 급격한 정보통신 기술의 발전으로 인한 인터넷의 확산에 기

* 정회원, 서울디지털대학교 컴퓨터공학부
(Department of Computer Engineering, Seoul Digital University)

※ 이 논문은 2010년도 한국인터넷진흥원 연구사업과 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2010-0021071).
접수일자: 2011년2월9일, 수정완료일: 2011년5월12일

인할 수 있다. 정보통신 기술의 급격한 발전으로 말미암아 개인용 컴퓨터의 성능이 더욱 강력하게 발전하고 또한 인터넷의 보급에 따라 누구나 인터넷을 이용한 사이버 공간상에서 정보의 상호 교환, 전자상거래, 인터넷 뱅킹 등의 사회 활동이 컴퓨터를 통해 편리하게 활용하고 있다. 하지만, 사이버 공간 상에서 활발한 정보의 상호 교환으로 인해 자신의 비밀스러운 정보들을 알려야만 하는 경우가 있다. 예를 들어 신용카드번호, 병력, 자격증, 주민등록번호, 여권 번호, 학력 등과 같은 비밀 정보들을 남에게 알려주고 싶은 인터넷 사용자는 없을 것이 자명한 일이다. 그러나 아이러니하게도 수많은 비밀스러운 개인정보들이 사이버 공간상에 존재하고 있는 것이 사실이다. 그 이유는 PC 사용자의 개인정보들이 컴퓨터상에 저장하여 편리하게 사용하면서 인터넷을 통해 메일로 전송하거나, 웹 하드에 개인정보가 저장된 파일을 업로드하거나, 게시판에 무심코 개인정보를 게시하거나, 바이러스 혹은 해킹에 의해 인터넷 상으로 유출되어 개인정보가 노출되기 때문이다[1~4].

과거 정보통신부가 한국정보보호진흥원이 개발한 주민등록번호 검색 소프트웨어를 통해 구글 (Goggle) 데이터베이스에 접속하여 주민등록번호 노출 상황을 검점한 결과 약 90만 명의 주민등록번호가 인터넷 상에 존재하고 주민번호 전부가 노출된 경우는 약 9만 5천여 명이고, 주민번호 앞자리 6자리가 노출된 경우는 80만 8천여 명에 이른다고 발표하였다[5]. 그리고 지난해 말부터 방송통신위원회와 한국인터넷진흥원은 인터넷 상에서 노출되는 개인정보인 주민등록번호, 신용카드번호, 계좌번호, 운전면허번호 등을 신속하게 검색하여 검증하고 삭제 조치 등 노출된 개인정보에 대해 대응 할 수 있는 “개인정보 노출 대응시스템”을 구축 및 운영하고 있다[3, 6]. 운영 중인 개인정보 노출 대응 시스템은 웹사이트의 페이지들을 검색하여 웹 자원에서 정확하게 개인정보 노출 사실을 추출하기 위해 개인정보 유효성을 적용하고 위협지수와 사실지수를 산정하여 노출 위협수준을 측정하는데 사용하고 있다. 그러나 개인정보 노출 대응 시스템 구축 및 운영을 통하여 인터넷 상에 노출된 개인정보 검색 및 삭제로 개인정보 노출로 인한 오·남용 최소화에 기여 하였으나, 이미지파일상의 노출된 개인정보 노출에 대한 검색이 이루어지지 않는 문제점을 가지고 있다. 이러한 문제점을 해결하기 위해서는 인터넷 상에 존재하는 수많은 이미지파일들 중에서 개인정보가 포함된 이미지파일들을 효율적으로 검색할 수 있는 방안의 필요성이 제기되고 있다. 따라서 본 연구에서는 인터넷 상에서 노출된 개인정보가 포함된 개인식별용 이미지파일을 획득하여 이들이 가지고 있는 고유한 특징정보를 분석하는데 연구의 초점을 둔다. 이미지를 표현하는 수많은 특징정보들 중에서 일반 이미지들 대비 개인정보가 포함된 이미지들이 가진 특징정보만을 효율적으로 분석하여 수치화함으로써 이미지파일을 검색하는 시스템에 활용한다.

본 논문의 구성은 Ⅱ장에서 개인정보가 포함된 이미지파일들의 특징정보를 추출하고 일반이미지 파일들과 비교하여 파일 분류성능을 그래프로 제시한다. Ⅲ장에서는 이미지파일들로부터 추출된 특징정보의 효율성을 검증하기 위해 특징정보 기반의 이미지파일 검색 시스템에 적용하여 성능을 분석하고 다른 연구들과 비교 평가한 결과를 기술한다. 그리고 Ⅳ장에서 결론을 맺는다.

## 1. 개인정보가 포함된 이미지파일의 정의

인터넷 상에는 존재하는 다양한 형태의 이미지파일들 중에서 개인정보가 포함된 이미지파일들에는 주민등록증에서부터 학생증까지 다양한 형태의 개인을 식별할 수 있는 정보들이 이미지 형태로 존재하고 있다. 본 연구에서는 이러한 모든 유형의 파일을 검색하기에는 무리가 있어 개인 식별도가 높은 표. 1과 같은 유형의 파일들을 검색대상으로 정의한다. 그리고 표. 2를 만족하는 이미지파일들만을 검색 대상으로 사용한다. 일반적으로 이미지파일로부터 개인정보에 해당하는 문자정보를 식별하기

표 1. 개인식별이 가능한 신분의 유형
Table 1. Types of personal identification.

| 유형 | 개인식별이 가능한 정보 |
|---|---|
| 주민등록증 | 성명, 주민등록번호, 주소 |
| 운전면허증 | 성명, 주민등록번호, 주소, 운전면허번호 |
| 여권 | 성명, 여권번호, 주민등록번호 |
| 자격증 | 성명, 주민번호, 주소 |

표 2. 이미지파일 검색을 위한 파일 형식 조건
Table 2. Image files format for image-retrieval.

| 형식 | 값 |
|---|---|
| 크기 | 150 × 150 pixel 이상 |
| 문자 해상도 | 72 DPI 이상 |
| Bit plane | 24-bit |
| 포맷 | BMP, JPEG, TIFF, GIF, PCX, PGM, |

위해서는 최소한 일정크기 이상이어야 가능하다. 이는 가독성 측면에서도 일정 크기 이하에서는 문자를 판독하기 어렵기에 개인정보 이미지파일 검색 시스템에서도 이를 예외로 처리토록 한다. 또한 개인정보가 포함된 이미지파일 동향에서와 같이 개인정보 이미지파일의 대다수가 24bit 칼라 색상으로 표현되어져 있어 이러한 제약을 가지는 이미지파일들에서만 제안한 방안을 적용한다. 이러한 조건에 따라 개인정보가 포함된 이미지파일들의 획득은 이미지 검색엔진에서는 다양한 색인어를 입력하여 검색된 이미지파일 중에서 실제 노출된 개인식별용 이미지 파일 740여개를 획득하였다. 획득한 이미지들 중에서 표. 2에 해당하는 이미지 465개만을 분석에 사용하였으며 대부분 가독성을 높이기 위해 촬영하거나 캡처한 영상들이지만 그 개수가 적어 학습하는데 어려움이 존재하였다. 이는 각 웹 서비스 제공자들이 자체적으로 차단하거나 인터넷 사용자들이 개인정보에 대한 민감성을 충분히 인지하고 있어 웹상에 노출되는 개인정보 이미지파일의 업로드를 자제하는데서 기인한다.

## Ⅱ. 개인정보가 포함된 이미지파일의 특징분석

인터넷 상에 존재하는 이미지파일들 중에서 개인정보가 포함된 이미지파일을 검색하기 위해 사전에 학습한 개인정보 이미지파일들의 특징정보 사전과의 유사성 비교를 수행함으로써 가능하다. 하지만 개인정보 이미지파일들은 서로 유사한 상관관계를 가지고 있어 이들을 각각 분류하기 위한 특징정보 선택은 많은 어려움이 존재한다[7~10]. 따라서 본 연구에서는 개인정보 이미지파일과 일반 이미지파일을 분류하는데 초점을 두고 그 방안을 제시한다. 표. 1의 개인식별용 신분증들은 사진과 성명 및 주민등록번호가 뿐만 아니라 위변조 방지를 위해 홀로그램 및 다양한 로그 및 그래픽 무늬들이 배경과 바탕에 함께 표현되어 있다. 결국 개인식별용 신분증들은 증명서과 통지서와 달리 다양한 무늬 배경을 바탕으로 개인정보가 인쇄되어 있어 이를 문자 인식을 통해 개인정보 포함여부를 판별하기에는 상당한 어려움이 존재한다. 따라서 해당 개인정보가 포함된 이미지파일을 검색하기 위해서 이미지파일의 특징정보를 이용한 검색 방안을 본 연구에서는 제안하도록 한다. 이를 위해 본 연구에서는 인터넷 상에서 존재하는 개인정보 이미지파일들을 수집하여 이들의 공통적인 특징정보를 분석하여 특징정보 사전을 생선하고

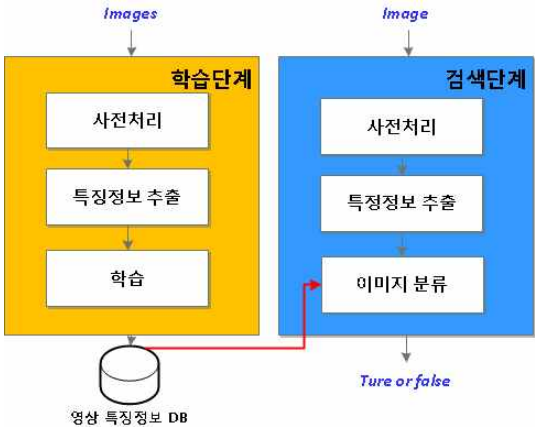



그림 1. 이미지파일 검색을 위한 처리 흐름도
Fig. 1. Processing flow for Image-retrieval.

이를 바탕으로 개인정보 이미지파일들을 검색한다. 그림. 1은 이미지파일 검색을 위한 처리 단계로써 먼저 이미지파일의 특징정보를 분석하여 특징정보 사전을 생성하고, 생성된 특징정보 사전을 이용하여 이미지파일을 검색한다. 특징정보 사전은 이미지파일에 대한 모양과 색상, 특징 등에 대해서 공통적으로 포함되는 정보를 담고 있는 사전을 의미한다. 이러한 특징정보 분석을 위해 이미지의 색상, 질감, 그리고 모양 특징정보들을 추출한다. 색상정보는 이미지의 칼라와 흑백 정보를 이용하고 질감 정보는 이미지 내의 픽셀값의 변화 정도를 통계학적 방법으로 측정하여 정보로 이용한다. 그리고 모양 정보는 이미지가 어려 영역의 집합으로 표현되어 있음을 가정하고 각 영역의 구조적 분석 방법에 의해 표현된 특징정보들을 사용한다[11~13]. 특징정

(가)

(나)

그림 2. 검색된 개인식별용 이미지들의 평균 이미지와 WANG의 이미지 데이터베이스 예제
Fig. 2. Average images of personal ID's and WANG's sample images.

표 3. 이미지 분석을 위한 특징정보
Table 3. Features for images analysis.

| 유형 | 특징정보 |
|---|---|
| 색상 | Gray / RGB / HSI / Color Moments / Gray Contrast |
| 질감 | Correlation / Autocorrelation / Cluster Shade Prominence / Dissimilarity / Entropy / Energy / Homogeneity / Maximum Probability Smoothness / Uniformity |
| 모양 | Third Moments |

보 분석을 위해서 사용한 실험 이미지는 그림. 2와 같다. 그림. 2(가)는 시계 방향으로 주민등록증, 운전면허증, 자격증, 사원증, 여권, 그리고 학생증의 평균 이미지를 표현한 이미지이고 (나)는 WANG 이미지 DB[14]에서 획득한 이미지들이다. 각각의 개인정보 이미지들은 다양한 조건하에 촬영된 그림 2(가)는 이미지의 형태를 띠고 있어 각각을 개별적으로 분석하는데 그 분산의 정도가 높아 실질적인 특징정보 비교 분석을 수행하는데 어려움이 발생하였다. 따라서 이미지간의 특징정보 비교분석을 위해 개인정보 이미지들을 각각 평균 이미지를 계산하였다. 평균 이미지는 획득된 샘플 이미지들의 크기를 정규화한 후 각각의 픽셀 위치에 해당하는 칼라 값을 합하여 그 평균을 계산한 이미지이다. 개인정보 이미지의 크기는 종류에 따라서 조금씩 차이를 가지고 있으나 대표적인 크기 비율에 따른 정규화를 수행하였다. 따라서 실험을 위해서는 전체적으로 1.6:1로 크기를 정규화하고, 이미지에 포함된 영상의 노이즈 감소를 위해 웨이블릿 기반의 노이즈 제거 방법을 적용한다[15]. 그리고 이미지파일들의 특징정보를 추출하기 위해 표. 3의 정보들을 사용하여 실험 이미지들과 Wang의 이미지들 간의 비교분석을 수행한다.

### 1. Gray / RGB / HSI 히스토그램

| | |
|---|---|
| 내용 | 이미지에서 해당 픽셀 값의 범위를 가진 분포를 측정 |
| 출력 값의 범위 | [0, G] G=밝기 값의 Level |
| 수식 | $p(r_k)=\frac{h(r_k)}{n}$, $r_k$=k번째 밝기 레벨, $h(r_k)$=$r_k$-레벨의 픽셀의 개수, $p(r_k)$ $r_k$-레벨이 나타날 확률 |

히스토그램은 이미지의 색상 분포를 표현하는 특징정보로써 데이터의 분포 상태를 알아보기 쉽게 분포의 모습을 한눈으로 확인하여 데이터가 어떤 값을 중심으로 어떤 분포를 가지는가를 확인이 가능한 정보이다. 그림. 3은 그림. 2의 개인정보 이미지들과 일반 이미지들의 2차원 RGB 칼라 히스토그램의 분포를 나타낸 것이다. 실선은 각각의 개인정보 이미지들의 히스토그램을 의미하고 막대그래프는 일반 이미지의 칼라 히스토그램으로 표현하였다. 그림과 같이 두 부류(개인정보 이미지 vs. 일반 이미지)의 히스토그램 비교를 통해 어느 정도 해당 부류를 유출할 수 있다. 예를 들어 개인정보 이미지들의 경우 검은색의 인쇄된 문자 정보가 다수 포함되어 있어 RGB 각각 모두 큰 밝기값을 가지고 있음을 확인할 수 있다. 그에 반에 일반 이미지들은 히스토그램의 가운데에 값에 대부분의 픽셀이 위치하고 있음을 알 수 있다. 이러한 단서를 바탕으로 향후 개인정보 이미지 파일 검색에 색상 히스토그램 정보가 유용하게 사용될 수 있음을 확인할 수 있다.

### 2. Gray Contrast

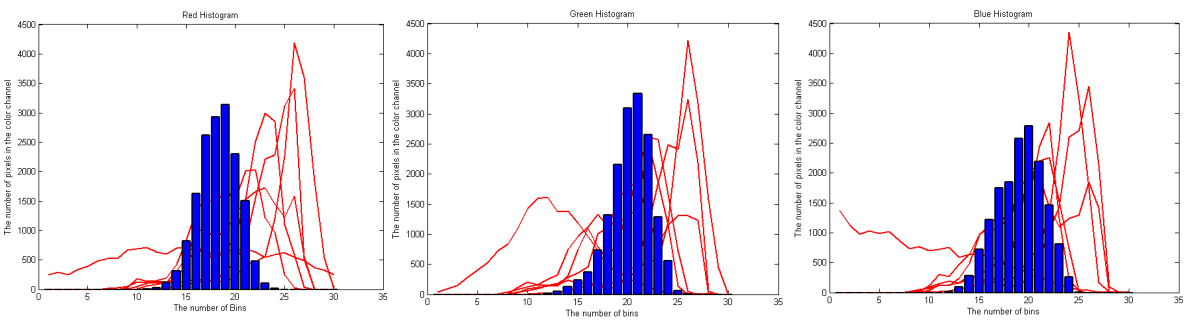
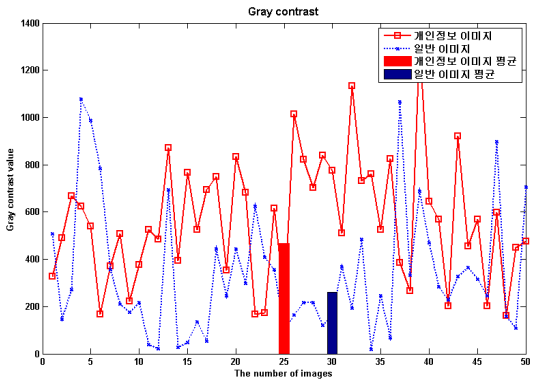


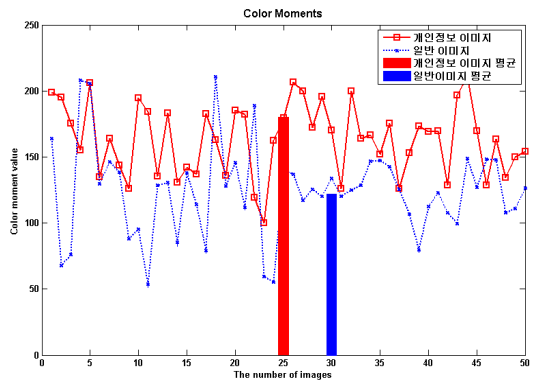

그림 3. 개인정보 vs 일반 이미지들의 RGB 칼라 히스토그램 분포(Bins=30)
Fig. 3. RGB color Histograms of images containing personal information vs generic images (Bins=30).

| 내용 | 한 픽셀과 주위에 인접한 픽셀들과의 밝기값 대비를 측정 |
|---|---|
| 출력 값의 범위 | [0, (size(*G*, 1)-1)$^2$] *G*=밝기 값의 Level Constant image는 Contrast는 0. |
| 수식 | $\sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 p_{ij}$, $p_{ij}$는 이미지 *G*의 *ij*번째 픽셀값의 확률 |

이미지내의 픽셀 밝기값의 변화 정도를 측정하는 특징정보로써 임의의 한 픽셀 밝기값이 전체 이미지에서 나타나는 확률을 계산하고 주위에 인접한 픽셀들과의 밝기 값 차이 정보를 표현하는 척도이다. 그림. 4(가)는 두 부류 이미지파일 사이의 Gray Contrast 비교 그림이다. 분석을 위해 개인정보 이미지 파일과 랜덤하게 선택된 일반 이미지파일들을 크기 정규화한 후 비교 분석 d,mf 수행한 결과, 대비는 밝기 픽셀의 변화 정보를 의미함으로써 개인정보 이미지파일보다는 일반 이미지 파일들의 대비가 평균적으로 약 12.48%의 높은 결과치




(가)




(나)

그림 4. 개인정보 vs 일반 이미지들의 Gray Contrast (가), 그리고 Color Moments(나) 분석 결과

Fig. 4. Result of gray co'trast and color moments of images containing personal information vs generic images.

를 제시하였다. 이로써 개인정보가 포함된 이미지 파일 검색을 위한 특징 정보로써 Gray Contrast 정보가 유용함을 확인할 수 있다.

## 3. Color Moments

| 내용 | 이미지의 색상 유사도를 측정 |
|---|---|
| 출력 값의 범위 | [0, 225] 5×5 Grid 마다 평균, 표준편차, 비대칭(skewness) 계산 |
| 수식 | $E_i=\sum_{N}^{j=1}\frac{1}{N}p_{ij}$, $\sigma_i=\sqrt{\left(\frac{1}{N}\sum_{N}^{j=1}(p_{ij}-E_i)^2\right)}$ $s_i=\sqrt[3]{\left(\frac{1}{N}\sum_{N}^{j=1}(p_{ij}-E_i)^2\right)}$ |

이미지를 5×5 grid로 표현하고 각 grid 내의 픽셀 값들로부터 칼라 모멘트를 계산한다. 즉, 영상내에서 칼라 정보의 치우침 정도를 분석한 것이다. 그림. 4(나)는 두 부류의 칼라 모멘트 값을 측정한 그래프이다. 그래프의 bin을 225로 출력하였다. 비교분석을 위해 각 대상 이미지들에서 칼라 모멘트값을 추출하고 이들의 평균값을 계산하여 그래프로 출력하였다. 비교분석 결과 개인정보 이미지파일들이 일반 이미지파일들보다 칼라 모멘트값이 평균적으로 10.04% 높게 나타났다. 비교적 근소한 차이의 성능 값을 나타내었지만 향후 두 부류 분류에 특징값으로 사용할 수 있음을 알 수 있다.

## 4. Autocorrelation

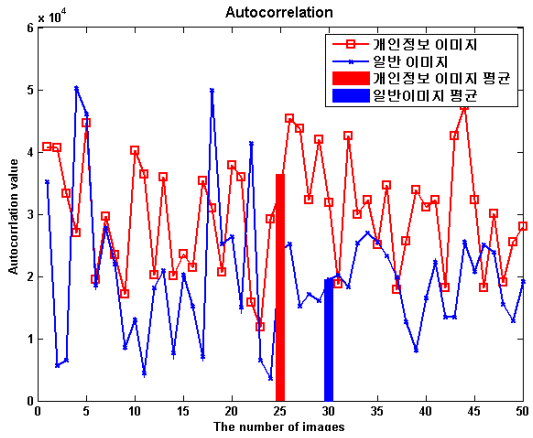
| 내용 | 평균 콘트라스트의 척도 |
|---|---|
| 출력 값의 범위 | [0, R] |
| 수식 | $S(x,y)=\sum_{i=1}^{M-x}\sum_{j=1}^{N-y}\frac{I(i,j)\times I(i+x,j+y)}{(M-x)\times(N-y)}$ |

일반적으로 이미지 내에 픽셀값들은 일정한 크기의 밝기 군을 형성한다. 만약 밝기군의 크기가 크다면 부드러운 질감을 가진 이미지가 될 가능성이 커고 그 반대의 경우는 거친 이미지일 가능성이 크다고 할 수 있다. 따라서 이미지에서 자기상관이란 임의의 픽셀 값이 이웃하는 픽셀 값들과 어느 정도의 관련이 있는지를 측정하는 척도로써 밝기 군에 대한 크기를 추정하는 척도이다. 결국 자기상관함수의 값이 커다는 것은 밝기 군의 크기가 상대적으로 크다는 것을 의미하는 것이 되고 만약 자기상관함수의 값이 작아지면 밝기 군의 크기가 작아짐을 의미한다. 이는 이미지가 얼마나 부러운가? 혹은

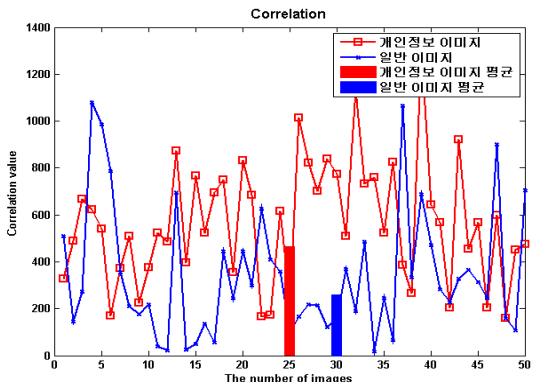
거친가의 질감 정도를 파악할 수 있다. 그림. 5(가)은 개인정보와 일반 이미지 파일의 자기상관함수의 척도값을 나타낸 그림이다. 비교 분석 결과 개인정보 이미지 파일들이 일반 이미지 파일들보다 자기상관함수의 값이 평균적으로 18.6% 높게 나타났다. 이것은 개인정보 이미지 파일들이 정형화된 형식에 따라 제작되어 일반 이미지파일들에 비해 비교적인 균일한 색상 값의 영역이 크다는 것을 의미한다.

### 5. Correlation

| | |
|---|---|
| 내용 | 전체 이미지 상에 존재하는 유사한 픽셀값을 가진 군들의 크기 측정 |
| 출력 값의 범위 | [0, R] |
| 수식 | $f = \sum_i \sum_j (i,j) p_{ij},$ |




(가)




(나)

그림 5. 개인정보 vs 일반 이미지들의 Autocorrelation (가)와 correlation(나) 분석 결과

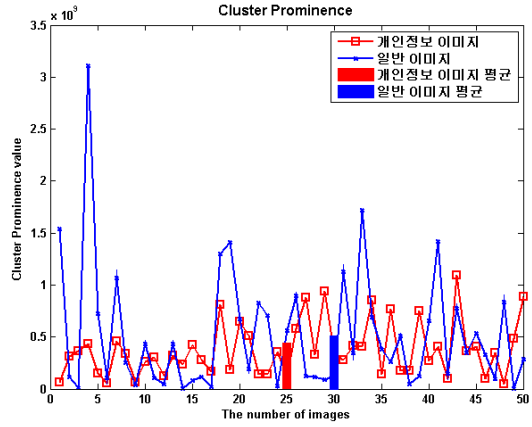
Fig. 5. Result of autocorrelation and correlation of images containing personal information vs generic images

이미지 파일의 픽셀 값이 위치한 곳에서 이웃하는 픽셀들과의 상호 관계를 측정하는 척도이다. 이웃하는 픽셀들 간의 어떤 선형적 관계를 갖고 있는지 분석하는 방법으로써 상관관계 값이 작을수록 무상관에 가깝고 선형적인 상관관계가 존재하지 않음을 의미한다. 상관관계 값이 커다는 것은 이미지의 색상 변화가 그 만큼 적다는 것을 의미함으로 부러운 질감을 가질 확률이 높아지게 된다. 그림. 5(나)는 개인정보와 일반 이미지 파일의 상관관계 척도값을 나타낸 그림이다. 비교 분석 결과 개인정보 이미지 파일들이 일반 이미지 파일들보다 자기상관함수의 값이 평균적으로 17.9% 높게 나타났다. 이것은 Autocorrelation 특징정보와 같이 개인정보 이미지 파일들이 정형화된 형식에 따라 제작되어 일반 이미지파일들에 비해 비교적인 균일한 색상 값으로 표현되어 있음을 의미한다. 그리고 자기상관함수의 특징정보만을 이용하여 두 부류의 이미지파일을 분류하게 되면 성능 오류 예측치가 약 26.3%±(%5) 정도로 예측되었다.

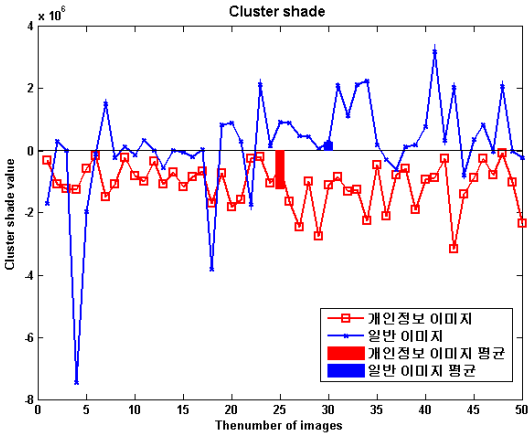
### 6. Cluster Shade / Prominence

| | |
|---|---|
| 내용 | 평균 콘트라스트의 척도 |
| 출력 값의 범위 | Cluster Shade: [-R, R], Cluster Prominence: [0, R] |
| 수식 | $f_c = \sum_i \sum_j (i+j-\mu_x-\mu_y)^3 p_{ij},$<br>$f_p = \sum_i \sum_j (i+j-\mu_x-\mu_y)^4 p_{ij},$ |

이미지 픽셀값의 분포가 비대칭인지를 측정하는 척도이다. 만약 Cluster Prominence 값이 크다는 것은 이미지 픽셀의 분포가 대칭적이지 않다는 것을 의미한다. 그 반대의 경우는 이미지 픽셀의 평균값을 중심으로 균형적인 픽셀 값 분포를 가지고 있음을 의미한다. 결국 이 척도를 사용하여 이미지 색상의 변화 여부를 파악할 수 있다. 그림. 6(가)와 (나)는 개인정보와 일반 이미지 파일의 Cluster Shade와 Prominence 척도값을 나타낸 그림이다. 비교 분석 결과 개인정보 이미지 파일들이 일반 이미지 파일들보다 Cluster shade와 prominence 값이 평균적으로 약 23.3%와 11.5%로 낮게 나타났다. 이것은 개인정보 이미지파일들의 픽셀 값의 변화가 일반 이미지파일들에 비해 변화가 작음을 의미한다. 하지만, 일반 이미지의 경우 하나의 샘플 이미지에서 Cluster prominence 값이 평균에 비해 3배 이상 높게 나타냈다. 이 값으로 인해 일반 이미지 파일들의 평균 Cluster prominence 값이 높게 나타났다. 결국 Cluster




(가)




(나)

그림 6 개인정보 vs 일반 이미지들의 Cluster Prominence(가) 와 Shade(나) 분석 결과

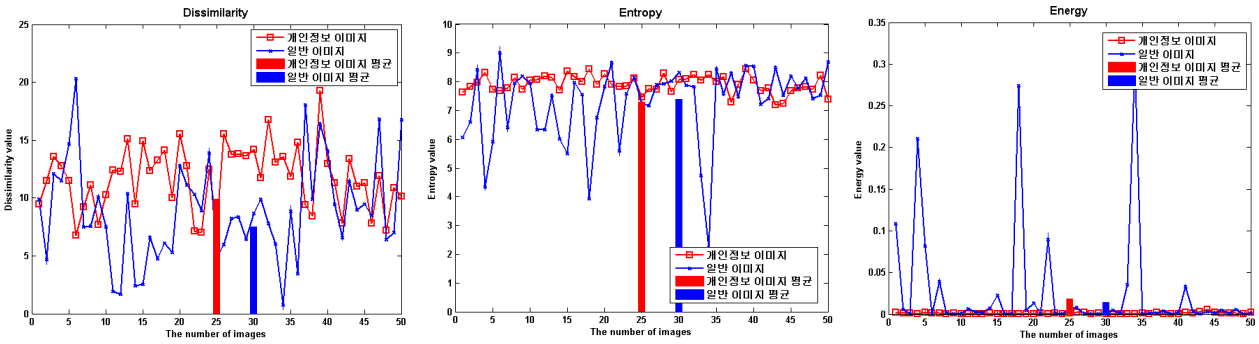
Fig. 6. Result of cluster prominence and shade of images containing personal information vs generic images.

prominence 특징 값으로만 두 부류 분류기의 특징정보로 사용하기에는 다소 무리가 있다고 판단된다. 그리고 Cluster Shade의 경우는 충분히 두 부류 분류기의 특징정보도 사용할 수 있는 정보이다. 개인정보와 일반 이미지 파일 사이에 충분한 상관관계를 표현하고 있으며 특히 두 분류 사이의 Cluster shade 값의 차이가 현저함을 그림. 6(나)를 통해 알 수 있다.

## 7. Dissimilarity / Entropy / Energy

| 내용 | 픽셀간의 비유사성 정도 측정 |
|---|---|
| 출력 값의 범위 | [0, R] |
| 수식 | $f=\sum_{i}^{K}\sum_{j}^{K}\lvert i-j\rvert \cdot p_{ij}$, $e=-\sum_{i=0}^{K-1}p(z_i)\log_2 p(z_i)$, $f=\sum_{i}^{K}\sum_{j}^{K}p_{ij}^2$ |

인접한 픽셀들 간의 색상 값의 비유사성 정도를 측정하는 척도들이다. Dissimilarity는 인접한 픽셀들 간의 색상 차이 정도를 측정, Entropy는 픽셀들 간의 무질서 정도를 측정, 그리고 Energy는 질감의 균일성을 측정하는 척도이다. 그림. 7(가)의 Dissimilarity 척도를 비교 분석한 결과 평균적으로 약 13.2%의 분류성능을 나타내었다. 하지만, Dissimilarity 특징정보만을 이용하여 두 부류의 이미지파일을 분류하게 되면 성능 오류 예측치가 약 42%±(%5) 정도로 저조하게 예측되었다. 이는 Dissimilarity는 두 부류 이미지들이 다양한 색상 값으로 구성되어 있어 낮은 분류 결과를 나타내었다. 비록 두 부류간 분류 성능이 좋은 결과를 제시하였더라고 성능 분류 오류율이 높아 실제 파일 검색에 적용하기에는 어려움이 존재한다. 따라서 분류 오류를 줄일 수 있는 다른 특징정보들과의 조합이 필요하다. 또한 그림. 7(나)의 Entropy 특징값 역시 분류 성능과 함께 분류 오




(가) (나) (다)

그림 7. 개인정보 vs 일반 이미지들의 Dissimilarity(가), Entropy(나), 그리고 Energy(다) 분석 결과

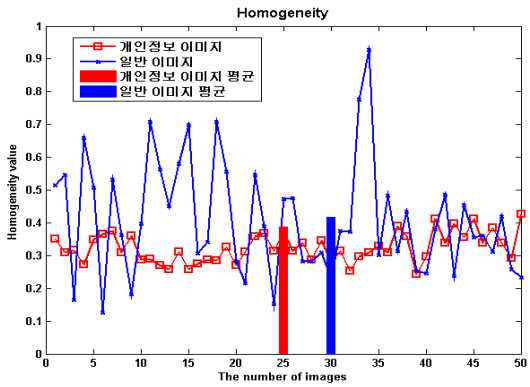
Fig. 7. Result of Dissimilarity, Entropy, and Energy of images containing personal information vs generic images.

류율이 역시 높아 실제 검색에 적용하는데 어려움이 존재한다. 그리고 그림. 7(다)의 Energy 특징정보의 경우는 이미지의 색상 변화가 적을 때 작은 값을 가지는 특성에 의해 개인정보 이미지가 대부분 균일한 색상 분포를 나타내고 있어 낮은 Energy 값을 출력한다. Energy 척도를 비교 분석한 결과 평균적으로 약 13.38%의 두 분류 성능을 나타내었고 오류율은 평균적으로 약 24%의 결과를 제시하였다.

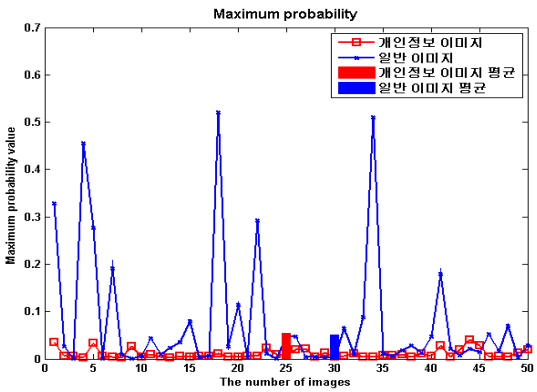
### 8. Homogeneity / Maximum probability

| | |
|---|---|
| 내용 | 이웃하는 픽셀과의 유사도 측정 / 픽셀값의 Max 빈도 측정 |
| 출력 값의 범위 | [0, R] |
| 수식 | $f=\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{p_{ij}}{1+\lvert i-j\rvert}$, $f=\underset{i,j}{MAX}\,p_{ij}$ |

Homogeneity는 이웃한 픽셀들 간의 색상값의 유사도를 측정하는 척도이다. 그리고 Maximum probability는 이미지내의 픽셀 값들 중에서 가장 높은 빈도를 가진 색상을 측정하는 척도이다. 즉, 이미지의 대표성을 측정하는 척도이다. 그림. 8(가)는 Homogeneity 척도를 비교 분석한 결과로써 평균적으로 약 10.7%의 두 분류 성능을 나타내었다. 개인정보 이미지파일의 경우는 일반 이미지들보다 유사한 영역으로 구성되어 있어 Homogeneity가 낮은 빈도를 제시하였고, 분류오류율은 평균적으로 약 29%의 성능을 제시하였다. 그림. 8(나)는 Maximum probability 척도를 비교 분석한 결과 평균적으로 약 10.5%의 두 분류 성능을 나타내었고, 분류 오류율은 평균적으로 약 38%의 성능을 제시하였다. 두 특징정보역시 다양한 특징정보들의 조합으로 높은 분류 성능을 나타낼 수 있을 것이다.




(가)




(나)

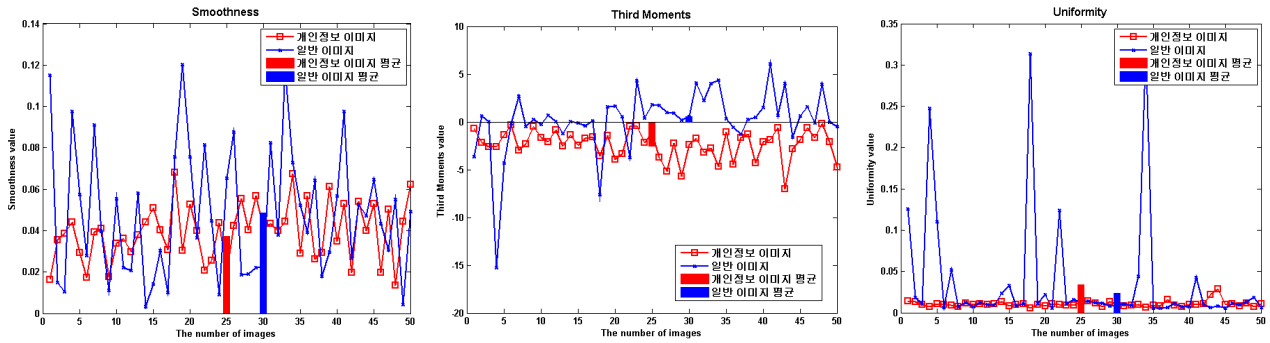
그림 8. 개인정보 vs 일반 이미지들의 Homogeneity(가)와 Maximum probability(나) 분석 결과

Fig. 8. Result of Homogeneity and Maximum probability of images containing personal information vs generic images.

### 9. Smoothness / Third Moments / Uniformity

| | |
|---|---|
| 내용 | 픽셀이 상대적인 부드러운 정도 측정, 히스토그램의 비대칭도를 측정, 균일도를 측정 |
| 출력 값의 범위 | [0, R], [-R R] , [0 R] |
| 수식 | $R=1-\frac{1}{1+\sigma^2}$ , $\mu_3=\sum_{i=0}^{K-1}(z_i-m)^3p(z_i)$, $U=\sum_{i=0}^{K-1}p^2(z_i)$ |

Smoothness는 이미지내의 영역에서 색상 값의 상대적인 부드러운 정보를 측정하는 척도이다. 임의의 영역이 일정한 색상 값을 가지고 있다면 0이고, 색상 값이 크게 벗어나는 영역에서는 큰 값에 접근한다. Third Moments는 이미지 파일의 색상값이 비대칭인 정도를 측정하는 척도이다. 대칭인 히스토그램은 0이고, 평균을 기준으로 오른쪽에 치우친 히스토그램에 대해서는 양의 값을 가지며, 왼쪽에 치우친 히스토그램에 대해서는 음의 값을 가진다. 그리고 Uniformity는 이웃한 픽셀들 간의 색상 값의 유사도를 측정하는 척도이다. 이 척도는 모든 그레이 레벨이 같을 때 최대이고, 그 이후부터는 감소한다. 그림. 9(가)은 Smoothness 척도를 분석한 결과 평균적으로 약 13.1%의 두 분류 성능을 나타내었고, 분류 오류율은 평균적으로 약 21.7%를 나타내었다. 그림. 9(나)은 Third Moments 척도를 비교 분석한 결과 평균적으로 약 20.7%의 분류성능을 나타내었고, 분류 오류율은 평균적으로 약 10%를 나타내었다. 이 척도가 분류성능이 높고 또한 오류율이 낮음을 제시하였




(가) (나) (다)

그림 9. 개인정보 vs 일반 이미지들의 Smoothness(가), Third Moments(나), 그리고 Uniformity(다) 분석 결과
Fig. 9. Result of Smoothness, third moments, and uniformity of images containing personal information vs other images.

다. 따라서 개인정보 노출 대응 시스템의 이미지 파일 검색을 위해 필히 포함되어야할 특징정보이다. 그리고 그림. 9(다)는 Uniformity 척도를 비교 분석한 결과 평균적으로 약 14.5%의 분류 성능을 제시하였다.

## Ⅲ. 실험 결과

제안한 방안을 실험하기 위해 영상 검색 실험 연구에 많이 이용하는 Wang 이미지 DB에서 획득한 이미지 915개와 개인정보가 포함된 이미지 465개를 이용하였다. 개인정보 이미지는 실제 웹상에서 다양한 상용으로 서비스 중인 인터넷 검색 엔진들을 사용하여 개인정보가 포함되어 있는 이미지파일을 수동으로 획득하였다. 획득된 개인정보 이미지들에서는 ISP 업체들에서 개인정보 영역을 마스킹한 이미지도 포함되어 있으며, 영상의 촬영 각도와 환경 등에 따라 다양한 변화를 가지고 있는 특징이 있다. 실험을 위해 사용한 PC 환경은 펜티엄 IBM호환 PC에서 윈도우 XP 환경(CPU: QuardCore 3.2GHz, RAM:4GM)의 Matlab언어로 구현하였다. 이미지 파일의 분류의 유사도 계산과 성능 평가는 식 (1)과 같다.

$$
\begin{aligned}
&S = \sqrt{|a-b|^2} \\
&Recall = \frac{True\ Positives}{Total} \\
&False\ Positives\ Rate(FPr) = \frac{False\ Positives}{Class\ Size} \\
&False\ Nagatives\ Rate(FNr) = \frac{False\ Nagatives}{Class\ Size}
\end{aligned}
\tag{1}
$$

식 (1)의 *Recall*은 전체 실험 이미지파일에서 바르게 분류한 이미지 파일의 분류한 개수를 의미한다. *FPr*은 해당 클래스의 이미지 파일이 아님에도 해당 클래스로

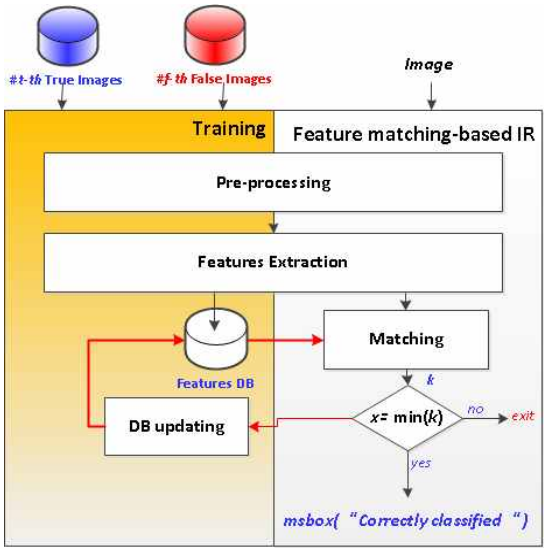



그림 10. 특징정보 매칭 기반의 이미지파일 분류 처리 흐름도
Fig. 10. Process flow of features matching based image retrieval.

분류한 개수이고(No인데 Yes라고 출력한 경우), *FNr*은 해당 클래스 임에도 해당 클래스가 아니라고 분류한 이미지 파일의 개수를 의미한다(Yes인데 No라고 출력한 경우). 그림 10은 개인정보가 포함된 이미지파일들이 가진 고유한 특징정보들의 조합을 사용하여 질의 이미지에 대한 분류 실험을 위한 처리 흐름도이다. 우선 개인정보가 포함된 이미지파일의 특징정보 학습을 위해 사전에 획득한 이미지들로부터 특징정보들을 추출하여 이를 특징 사전에 저장한 후, 입력된 이미지파일이 가진 특징정보가 특징사전과의 유사도 관계를 비교하여 분류한다. 따라서 각 특징정보들마다 이미지파일 분류 성능을 분석하고 이들 중에서 분류 성능이 임계치 이상인 특징정보들을 조합하여 특징정보 매칭 기반의 분류

에 사용한다. 그러므로 특징정보들을 기반으로 하여 실제 학습한 이미지들과의 유사성 분석을 통한 분류 성능 평가를 수행한다. 특징정보 유사도 기준 측정을 위해 제안한 방안은 최소 유사도를 가진 값의 인덱스를 확인하여 개인정보 이미지인지 아닌지를 판별한다. 그리고 입력된 이미지에 대한 특징정보를 특징정보 사전에 업데이트한다. 개인정보 이미지파일 분석을 위해 사용한 특징정보는 표. 3에서와 같이 총 17개의 특징정보들이다. 따라서 입력 이미지 당 17개 특징벡터 결과값이 출력으로 제시된다. 즉, 입력 이미지에서 Gray Contrast 값을 출력하여 이미지 사전에 비교 과정을 수행한 후 최소값이 특징정보 사전의 값과 유사하면 1을 출력하고 그 외의 경우는 0을 출력한다. 이러한 과정을 통해 17bit의 스트링을 출력하게 된다. 결국 17개의 bit값을 통해 해당 입력 이미지가 개인정보이미지 인지 아닌지를 판별하다. 만약 입력 영상이 개인정보 이미지라면 17bit 모두 1의 결과값을 제시하고, 만약 아니라면 0의 결과를 제시하는 경우가 최상의 경우가 될 것이다. 따라서 1의 개수에 따라 개인정보 이미지인지 아닌지를 결정하는 임계치 기준 설정이 필요하다. 제안한 방안에서 30%의 임계치를 기준으로 하여 특징정보 기반의 매칭 후 결과 bit string에 1의 개수가 12개 이상이면 개인정보가 포함된 이미지로 판별하고 만약 5개 미만이라면 일반 이미지로 판별한다. 제안한 방안의 성능을 실험하기 위해서 특징정보 매칭 기반의 분류 성능 분석, 노이즈(salt&pepper)에 대한 분류 성능 분석, 회전 변환에 의한 성능 분석, 그리고 크기 변환에 의한 성능 분석을 각각 실시하였다. 특징정보 매칭 기반의 분류 성능 분석에 있어 특징정보의 개수에 따라 30%의 개인정보 이미지 결정 기준을 적용하였다. 표. 4의 실험결과 특징정보 개수에 따른 성능 분석에서는 가능한 최대의 특징정보 조합을 사용하는 것이 좋은 분류 성능이 제시하였다. 노이즈에 대한 성능 분석 역시 최소 10%의 노이즈가 포함된 이미지에서도 최소한 90%의 분류 성능율을 제시하였다. 다만, 개인정보가 포함된 이미지파일의 경우 노이즈 영향에 의해 오분류가 다소 높게 나타냈다. 그리고 회전변환 크기 변환에 의한 분류 성능 측정에서는 여러 영상 변환에 불구하고 검색 성능에는 큰 영향을 주지 못하였다.

그리고 표. 5에서는 RGB[16]와 3차원 HSV 칼라 히스토그램 기반의 영상 검색 방법[17], Haar wavelet 도1AZ인에서 63-bit Haar 칼라 히스토그램 매칭을 통한 영상 검색 방법[18], 그리고 invariant features 히스토그램 매칭 기반의 영상 검색 방법[19]들과 제안한 방안을 영상 분류 평가하였다. 각 영상 방법들의 특징정보 매칭을 위해 히스토그램의 Bin의 수를 30으로 설정하였다. 그리고 Haar wavelet 변환을 2-level으로 설정하고, invariant features들에는 7개의 invariant moment 값들을 사용하였으면 최소한 4개 이상의 특징정보를 임계치 결정 기준으로 설정하였다. 또한 각 방안들의 유사도 분석을 위해서 유클리디안 최소거리계산하였다. 제안한 방안을 실험한 결과 개인정보가 포함된 이미지 465개 중에서 412개를 올바르게 분류하였으며 이중 53개를 일

표 4. 제안한 방안의 이미지 파일 분류 성능(%)
Table 4. Performances of the proposed method.

| 성능 | *Recall* | *FNr* | *FPr* |
|---|---|---|---|
| 방안 / feature 개수 | 특징정보 매칭 기반의 분류 성능 평가 | | |
| 17 | 90.80 | 11.39 | 8.08 |
| 13 | 84.06 | 20.21 | 13.77 |
| 9 | 83.33 | 24.94 | 12.45 |
| 5 | 74.78 | 66.88 | 4.04 |
| 방안 / 노이즈(%) | 노이즈에 대한 분류 성능 평가 | | |
| 5 | 92.02 | 15.91 | 3.93 |
| 10 | 90.14 | 24.08 | 2.62 |
| 20 | 79.78 | 56.29 | 1.85 |
| 30 | 73.26 | 73.33 | 3.06 |
| 방안 / 방향 | 회전변화에 대한 분류 성능 평가 | | |
| 45 | 90.21 | 12.25 | 8.52 |
| -45 | 89.85 | 11.61 | 9.39 |
| 90 | 91.30 | 16.34 | 4.80 |
| -90 | 91.59 | 15.91 | 4.59 |
| 방안 / 크기 | 크기 변환에 대한 분류 성능 평가 | | |
| 0.5배 확대 | 90.72 | 13.76 | 6.99 |
| 0.5배 축소 | 90.79 | 13.11 | 7.21 |
| 2배 확대 | 92.97 | 12.04 | 4.48 |
| 2배 축소 | 92.89 | 15.48 | 2.84 |

표 5. 다른 영상 검색 방안들과의 성능 평가
Table 5. Results of performances evaluation.

| 성능 / 특징정보 | *Recall* | *FNr* | *FPr* |
|---|---|---|---|
| 제안한 방안 | 90.80 | 11.39 | 8.08 |
| RGB [16] | 75.65 | 50.32 | 11.14 |
| 3차원 HSV [17] | 89.27 | 16.77 | 7.65 |
| Haar wavelet [18] | 87.89 | 26.66 | 4.69 |
| Invariant moments [19] | 81.88 | 37.41 | 8.30 |

반이미지로 오분류하였다. 그리고 일반 이미지들 중에서 개인정보 이미지로 오분류한 경우는 약 74개였다. RGB 칼라 특징정보 매칭을 통해 영상 검색을 제외한 나머지 방안들에서는 Recall이 80% 이상의 결과를 제시하였다. 하지만, invariant feature나 wavelet 도메인에서의 특징정보를 개인정보가 포함된 이미지를 검색에 사용하는 것으로 *FPr*과 *FNr*의 오류율을 증가시킴으로써 적합하지 않은 것을 알 수 있다.

## Ⅳ. 결 론

제안한 방안은 인터넷 상에서 개인정보가 포함된 이미지파일을 효율적으로 검색하기 위해 이미지에 포함된 특징정보들 중 개인정보 이미지에 특화된 특징정보를 분석하여 효율적으로 개인정보 이미지파일을 검색하는 방안이다. 제안한 방안을 실험한 결과 분류성공률은 약 89%의 결과를 제시하였으나 개인정보 이미지파일임에도 일반이미지 파일로 오분류된 경우가 약 10% 이상을 제시하였다. 결국 분류성공률은 높지만 오분류율을 더욱 낮추기 위한 방안이 요구되는 대목이다. 이를 위해 향후 연구로는 개인정보 이미지파일 분류를 위해 현재 분석된 17개의 특징 정보보다 더욱 다양한 특징정보의 추출 방법을 연구하고, 그리고 추출된 특징정보의 매칭 방법에 적응적으로 반응할 수 있는 분류 임계치 설정과 신경망과 SVM과 같은 학습기 기반의 이미지파일 분류기에 대한 연구를 수행하고자 한다.

## 참 고 문 헌


[1] 한승원, 이상진, 이강신, 차윤호, “개인정보 저장 형태에 따른 유출 탐지 방안”, 한국정보과학회지, vol. 27, no. 12, pp.42-49, 2009.

[2] 유진호, 지상호, 임종인, “개인정보 유.노출 사고로 인한 기업의 손실비용 추정”, 정보보호학회논문지, vol. 19, no. 4, pp.63-75, 2009.

[3] 최진영, 하태균, 이강신, 원유재, “개인정보 노출대응 체계”, 한국정보보학회지, vol. 19, no. 6, pp.9-14, 2009.

[4] 이상진 외 9명, “개인정보 유출 공격 탐지 방안”, KISA-WP-2009-005, 연구보고서, 한국인터넷진흥원, 2009.

[5] http://itnews.inews24.com/php/news_view.php?g_menu=020100&g_serial=218397

[6] http://www.gocj.net/news/articleView.html?idxno=23551

[7] 이강현, “영상 분할을 이용한 영역 기반의 내용 검색 알고리즘”, 대한전자공학회 논문지, 제44권, 5호, 1-11쪽, 2007년.

[8] 백영현, 문성룡, “컬러 질의 영상 검출을 위한 객체 기반 영상 검색”, 대한전자공학회 논문지, 제45권, 제3호, 97-102쪽, 2008년.

[9] 성중기, 천영덕, 김남철, “칼라 공간적 상관관계 및 국부 질감 특성을 이용한 영상 검색”, 전자공학회 논문지, 제42권, 5호, 103-114쪽, 2005년.

[10] 노형기, 황본우, 문종섭, 이성환, “내용 기반 영상 정보 검색 기술의 현황”, 대한전자공학회 논문지, 제25권, 8호, 798-806쪽, 1998년.

[11] W.T Chen, W.C. Liu, M. S. Chen, “Adaptive color feature extraction based on image color distributions”, IEEE Trans. on Image Processing, vol. 19,. no. 8, pp.2005-2016, 2010.

[12] W. Bian, D. Tao, “Biased discriminant euclidean embedding for content-based image retrieval”, IEEE Trans. on Image Processing, vol. 19, no. 2, pp.545-554, 2010.

[13] H. Yuan, X. P. Zhand, “Statistical modeling in the wavelet domain for compact feature extraction and similarity measure of images”, IEEE Trans. on Circuits and systems for Video Tech., vol. 20, no. 3, pp.439-445, 2010.

[14] http://wang.ist.psu.edu/docs/related/

[15] J.B, Kim and H.J. Kim, “Multiresolution-Based Watersheds for Efficient Image Segmentation”, Pattern Recognition Letter, vol. 24, vo. 1, pp. 473-488, 2003.

[16] A. K. Jain, and V. Aditya, “Image retrieval using color and shape”, Great Britain: Elsevier Science Ltd, 1995.

[17] T. Giannakopoulos, “Color-based image retrieval -Query by Example”, http://www.mathworks.com/matlabcentral/fileexchange/22030-image-retrieval-query-by-example-demo.

[18] A. Utenpattanant, et al., “Color descriptor for image retrieval in wavelet domin”, Pro. in ICACT, pp. 818-821, 2006.

[19] T. Deselaers, “Feature for image retrieval”, PhD Thesis, RWTH Aachen Univerisity, 2008.


### 저 자 소 개


김 종 배(정회원)
대한전자공학회 논문지
제 47CI편 제1호 참고