Journal of Korea Multimedia Society Vol. 18, No. 11, November 2015(pp. 1268-1280)
http://dx.doi.org/10.9717/kmms.2015.18.11.1268

대규모 비디오 감시 환경에서 프라이버시 보호를 위한 다중 레벨 특징 기반 얼굴검출 방법에 관한 연구

이승호[†], 문정익[††], 김형일[†††], 노용만[††††]

Face Detection Using Multi-level Features for Privacy Protection in Large-scale Surveillance Video

Seung Ho Lee[†], Jung Ik Moon[††], Hyung-Il Kim[†††], Yong Man Ro[††††]

ABSTRACT

In video surveillance system, the exposure of a person's face is a serious threat to personal privacy. To protect the personal privacy in large amount of videos, an automatic face detection method is required to locate and mask the person's face. However, in real-world surveillance videos, the effectiveness of existing face detection methods could deteriorate due to large variations in facial appearance (e.g., facial pose, illumination etc.) or degraded face (e.g., occluded face, low-resolution face etc.). This paper proposes a new face detection method based on multi-level facial features. In a video frame, different kinds of spatial features are independently extracted, and analyzed, which could complement each other in the aforementioned challenges. Temporal domain analysis is also exploited to consolidate the proposed method. Experimental results show that, compared to competing methods, the proposed method is able to achieve very high recall rates while maintaining acceptable precision rates.

Key words: Face Detection, Multi-level Facial Feature, Privacy Protection, Large-scale Video Surveillance System

1. 서 론

최근 치안안전, 범죄수사에 등을 목적으로 한 비디오 감시 시스템이 주택가, 학교, 공항, 현금지급기(automatic teller machines, ATMs) 등 수 많은 장소에 설치되고 있다. 그런데 비디오 감시 시스템에서 촬영된 비디오 데이터에는 개인의 모습이 의도치 않게 촬영되고 무방비로 공유되는 것이 심각한 문제로 대두되고 있다[1]. 수 많은 영상 속에서 무방비로 드러난 개인의 얼굴은 신원(identity)을 노출하고 프라이버시를 침해 할 가능성이 매우 높다. 또한 최근 비디오 감시 시스템에 탑재된 자동 얼굴인식(automatic face recognition)에 의해 개인의 일상이 추적되고 오남용 될 수 있다.

※ Corresponding Author : Yong Man Ro, Address: (34141) N1-ITC Building, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, TEL : +82-42-350-3494, FAX : +82-42-350- 7619, E-mail : ymro@kaist.ac.kr
Receipt date : Mar. 26, 2015, Revision date : Sep. 1, 2015
Approval date : Sep. 11, 2015
[†] School of Electrical Engineering, Korea Advanced Institute Science and Technology (KAIST)
(E-mail : leesh09@kaist.ac.kr)
[††] School of Electrical Engineering, Korea Advanced Institute Science and Technology (KAIST)
(E-mail : iksciting@kaist.ac.kr)
[†††] School of Electrical Engineering, Korea Advanced Institute Science and Technology (KAIST)
(E-mail : hyungil.kim@kaist.ac.kr)
[††††] School of Electrical Engineering, Korea Advanced Institute Science and Technology (KAIST)
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2011-0011383)

비디오 감시 시스템 환경에서 촬영되는 개인의 프라이버시 보호하기 위한 가장 대표적인 방법 중 하나는 비디오 프레임들에서 얼굴을 마스킹(masking) 하는 것이다[2-4]. 얼굴 마스킹은 촬영된 사람이 누구인지 식별(identifying)할 수 없도록 얼굴영역에 블러링(blurring) 또는 암호화(encription) 등을 적용하는 영상처리 기법이다. 방대한 수의 얼굴들에 대해 일일이 프라이버시 보호를 적용해주기 위해서는 완전 자동화(fully automatic)된 기술이 필요하다[1]. 특히 얼굴 마스킹을 위해 얼굴영역을 자동으로 지정할 수 있는 얼굴검출(face detection) 기술이 핵심이고 필수적이다. 그런데 얼굴 마스킹을 위한 얼굴검출은 다음과 같은 이유로 매우 도전적(challenging)이다: (1) 얼굴의 해상도가 매우 낮거나 가림(occlusion)이 있는 경우 얼굴검출이 매우 어렵다. 하지만 이러한 경우에도 대부분 사람의 눈으로는 얼굴을 보고 누구인지 여전히 식별 가능(identifiable)하기 때문에 반드시 검출되고 마스킹이 되어야 한다[1]. (2) 감시 시스템에서 취득되는 비디오의 양이 방대하고 즉각적인 얼굴 마스킹을 위해 고속의 알고리즘 개발 또는 고속화를 위한 대량의 컴퓨팅 리소스(computing resource) 사용을 필요로 한다[1].

얼굴검출은 얼굴추적(face tracking)이나 얼굴인식을 위한 필수적인 전처리 단계로서 많은 방법들이 제안되었다. Viola와 Jones가 제안한 비올라 존스(Viola-Jones) 방법[5]은 현재까지도 매우 널리 사용되고 있는 가장 대표적인 얼굴검출 방법이다. 이 방법은 Haar-like 특징으로 정의되는 약 분류기(weak classifier)들의 선형조합(linear combination)으로써 강 분류기(strong classifier)를 생성한다[5]. 강 분류기들이 직렬로 연결된 캐스케이드(cascade) 분류기에서 모든 강 분류기에서 얼굴로 판별된 스캔 윈도우(scan window)만이 최종 얼굴영역으로 분류된다[5]. 비올라 존스 방법[5]은 캐스케이드 구조와 적분영상(integral image)의 사용에 의해 계산속도를 향상시키면서 비교적 안정적인 검출성능을 보인다. 저해상도 얼굴을 효과적으로 검출하기 위해 Zhang은 Haar-like 특징 대신 변형된 census transform 특징[7]을 캐스케이드 구조와 적분영상에 기반한 프레임워크에 적용하였다[6]. 이 방법은 기존의 9비트 이외에 3비트를 할당하여 컬러 정보를 추가적으로 인코딩하여 얼굴검출에 활용한다[6]. 앞서 언급한 방법들[5, 6]은 분류기 학습에 사용된 얼굴의 특성(예: 얼굴의 포즈(pose))과 테스트 영상에 존재하는 얼굴의 특성이 유사한 경우에 높은 얼굴검출 성능을 보인다. 반면에 특성이 다른 경우에는 단일한 얼굴 특징 사용에 의해 얼굴검출을 놓치거나 오검출(false positive)이 발생하기 쉽다는 한계점을 갖는다. 다양한 변화를 갖는 얼굴을 모두 포함하여 하나의 검출기를 학습시킬 수 있으나, 다양한 입력 영상에 존재하는 얼굴 특징을 찾아 분류하기에는 매우 많은 약 분류기들이 필요하게 되며, 훈련과정이 매우 복잡하게 된다[8]. 단일한 얼굴 특징 사용의 제한된 성능을 향상시키기 위해 El-Barkousky는 돌출 맵(saliency map)과 얼굴 피부 컬러에 기반한 특징을 사용한 얼굴검출 방법을 제안하였다[9]. 돌출 맵 사용은 얼굴과 같은 오브젝트(object)가 영상 내에서 도드라져 보인다는 가정에 근간한다. 이 방법은 비올라 존스 방법을 통한 분석, 돌출 맵을 통한 분석, 얼굴 피부 컬러를 이용한 분석 결과들을 융합하여 최종 얼굴검출 결과를 얻는다[9]. 이 방법은 제한된 DB에서 비올라 존스 방법보다 월등히 높은 검출성능[정확도(precision) 대비 리콜(recall)]을 보였다[9]. 하지만 돌출 맵은 단조롭지 않은 배경에서 오브젝트와 배경을 분간하기 어려워 효용성이 급격히 떨어지는 문제점이 있다.

얼굴의 포즈변화에서의 강인성을 위해 Li[10]는 플로트 부스트(float boost)를 이용해 얼굴 포즈별로 학습된 얼굴검출기들로 구성한 다중 뷰(multi-view) 얼굴검출 방법을 제안하였다. 이 얼굴검출 방법은 세 개 레벨의 피라미드 구조를 갖는다. 첫 번째 레벨에서는 포즈와 관계없이 개략적으로 얼굴인지 아닌지를 판별한다. 두 번째 레벨에서는 3개의 포즈(정면, 왼쪽 측면, 오른쪽 측면) 중 한 개의 포즈로 판별하고 세 번째 레벨에서는 7개 중 하나의 포즈로 더 세밀하게 판별한다. 이 방법은 다양한 포즈의 얼굴을 검출할 수 있다는 장점을 갖지만 조명변화나 가림에 의해 저하된(degraded) 얼굴영역에 대해서는 검출이 어렵다는 단점을 갖는다. 또한 다중 레벨의 피라미드 구조를 구성하기 때문에 테스트 할 때 계산속도가 매우 느리다.

본 논문에서는 앞에서 언급한 기존 얼굴검출 방법들의 한계점들을 극복하고 대규모 감시 비디오에서 얼굴 마스킹에 적합한 얼굴검출 방법을 제안한다. 제안 방법은 대규모 감시 비디오에서 두 가지 요구사항

을 충족시키고자 한다. 첫째, 제약되지 않은(unconstrained) 환경에서 취득된 비디오에서 얼굴의 마스킹 및 프라이버시 보호를 위해 중요한 높은 리콜(recall)을 달성한다. 둘째, 대규모 비디오에서 효율적인 얼굴 마스킹을 위한 고속 얼굴검출을 달성한다. 이 요구사항들과 관련하여 제안하는 얼굴검출 방법의 기술적 특장점은 아래와 같다.

- 얼굴검출에서 단일한 특징에 기반한 분류의 한계점을 극복하기 위해 비디오 프레임 내 다중 레벨의 공간 특징에 기반한 얼굴검출을 제안한다. 제안 방법은 컬러와 같은 로우 레벨(low-level) 특징, 얼굴의 눈과 입 등 국부적 정보에 해당하는 미드 레벨(mid-level) 특징, 얼굴의 전체(holistic) 구조적 정보(정면얼굴과 측면얼굴)와 같은 하이 레벨(high-level) 특징들의 분석을 융합함으로써 달성된다. 각 레벨의 특징은 영상 환경에 따라 장단점을 갖고 상보적인 정보를 제공하기 때문에 제약되지 않은 비디오에서도 높은 리콜을 달성할 수 있다.

- 비디오에 등장하는 개인의 얼굴은 시간에 따라 그 모습(appearance)이 연속적으로 변하며 심한 포즈 변화나 가림을 포함한다. 이 때문에 어떤 비디오 프레임에서는 얼굴영역을 판별하기에 충분한 단서(cue)를 포함하지만 그렇지 않은 비디오 프레임도 존재한다. 이 특성을 고려하여 제안 방법에서는 정방향 얼굴추적(forward face tracking)과 역방향(backward face tracking)을 사용한다. 양방향 얼굴추적에 의해, 충분한 단서를 갖는 비디오 프레임에서 검출된 얼굴영역 정보가 불충분한 단서를 갖는 전방 및 후방 비디오 프레임들에 전파된다. 이는 공간 특징에 기반한 얼굴검출에서 검출하지 못한 얼굴영역을 검출하여 리콜이 낮은 비디오 프레임들의 리콜을 높이는데 효과적이다.

- 다중 레벨 특징에 기반한 얼굴검출의 고속화를 위한 관심영역(Region-of-Interest, ROI) 검출을 제안한다. 이 방법은 이전 프레임에서 검출된 얼굴영역들의 정보를 이용하여 관심영역의 분류 모델을 적응적으로 생성한다. 따라서 주어진 프레임에서 얼굴에 해당되지 않은 불필요한 관심영역의 크기를 최소화 시킬 수 있다. 관심영역의 감소는 얼굴검출의 계산속도 향상뿐만 아니라 오검출(false positive) 또한 감소시킬 수 있다.

제안 방법의 효용성을 검증하기 위해 두 가지 데이터베이스(DB), McGill real-world face video DB[11]와 실환경에서 촬영된 CCTV DB에서 실험을 수행하였다. 실험결과, 제안 방법은 비교하는 최신 논문의 얼굴검출 방법, 상용화 얼굴검출 방법들보다 높은 리콜률(recall rate)과 고속 얼굴검출을 달성하였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 얼굴검출 방법의 전체 구성 및 개요를 설명한다. 3장에서는 제안하는 얼굴검출 방법을 자세히 설명한다. 4장에서 제안하는 얼굴검출 방법의 성능을 평가하고 5장에서 결론을 맺는다.

2. 제안 방법 개요

Fig. 1은 제안하는 얼굴 검출방법의 전체 구성도를 나타낸다. 제안 방법은 전처리 필터링(pre-filtering), 공간 도메인에서의 얼굴분석, 시간 도메인에서의 얼굴분석으로 구성된다.

전처리 필터링 단계의 목적은 입력된 비디오 프레임에서 얼굴이 존재할 영역을 짧은 시간 내에 선별함으로써 얼굴검출에 소요되는 계산시간을 대폭 줄이고 오검출을 감소시키는 것이다. 관심영역 검출에서는 피부색 픽셀 검출을 통해 피부색 픽셀들을 포함하는 직사각형 영역을 입력 비디오 프레임 내에서 선별한다. 만약 관심영역이 얼굴분석에서 검출 가능한 얼굴의 최소 픽셀사이즈(예: 20×20 픽셀)보다 작지만 육안으로 식별이 가능한 경우(예: 12×12 픽셀[1])에는, 얼굴을 검출하기 위해 관심영역의 사이즈를 늘린다(리사이징에 대한 자세한 기준은 3.1절 수식 (3) 참조).

공간 도메인에서의 얼굴분석 단계에서는 비디오 프레임 내 다중 레벨의 공간 특징에 기반한 얼굴검출을 수행한다. 이는 컬러와 같은 로우 레벨 특징, 얼굴의 눈과 입 등 국부적 정보에 해당하는 미드 레벨 특징 그리고 얼굴의 전체(holistic) 구조적 정보[정면(frontal)얼굴과 측면(profile)얼굴]와 같은 하이 레벨 특징들의 분석을 융합함으로써 달성된다.

시간 도메인에서의 얼굴분석 단계는 얼굴영역을

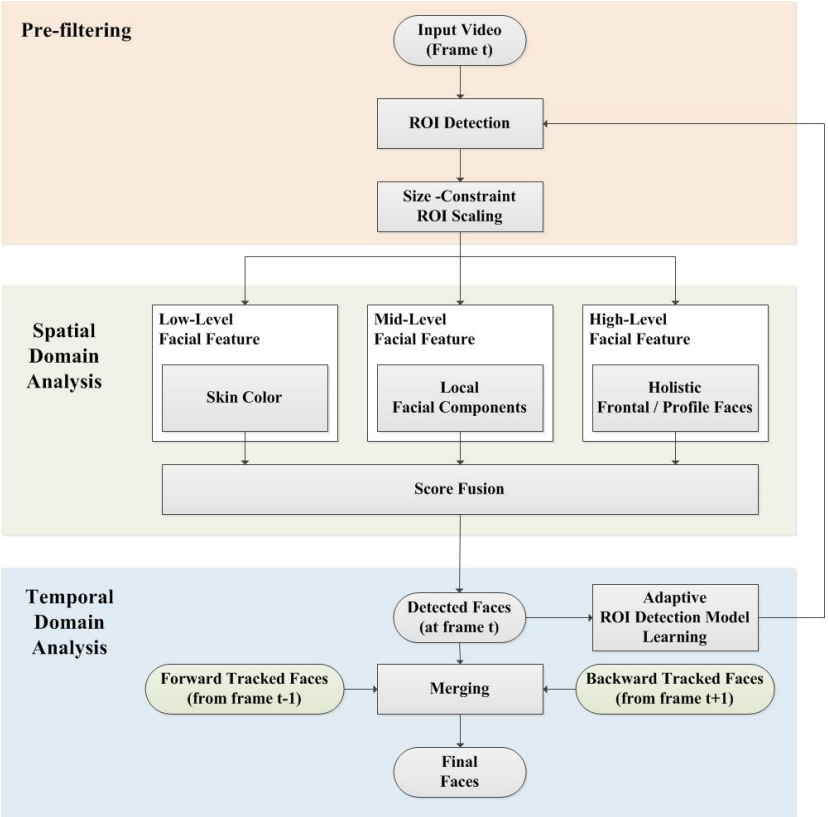

Fig. 1. Overall framework of the proposed method.

판별하기에 충분하지 않은 정보를 포함하는 경우 공간 도메인에서의 얼굴분석을 보완할 수 있다. 구체적으로, 정방향 얼굴추적과 역방향 얼굴추적을 통해 충분한 단서를 갖는 비디오 프레임에서 검출된 얼굴영역 정보가 불충분한 단서를 갖는 전방 및 후방 비디오 프레임들에 전파된다. 이 방법은 공간 특징에 기반한 얼굴검출에서 검출하지 못한 얼굴영역들을 추가로 검출할 수 있다.

3. 제안하는 얼굴검출 방법

3.1 전처리 필터링

전처리 필터링 과정에서는 관심영역을 검출하고 비 관심영역 내에서의 얼굴분석 과정을 생략할 수 있다. 이를 통해 비디오 프레임의 전체 영역을 분석하는 것에 비해 계산속도를 대폭 향상시킬 수 있고, 오검출을 감소시킴으로써 정확도(precision)을 향상시킬 수 있다. 본 논문에서는 관심영역 검출을 위해 피부색(skin color) 정보를 사용한다. 피부색 정보를 이용한 관심영역 검출은 피부색 픽셀 검출(skin pixel detection), 모폴로지 연산(morphological operation), 세그멘테이션(segmentation)의 순서로 이루어진다. 피부색 픽셀 검출 과정에서는 비디오 프레임을 구성하는 모든 픽셀들을 관찰하여 해당 픽셀의 피부색 여부를 판별한다. 여기서 피부색 픽셀을 판별하는 방법으로서 RGB 색공간(color space)에 기반한 방법을 사용하였다[12]. 피부색 픽셀 판별 규칙은 다음과 같다.

$$\begin{aligned}&(R>95)\text{and}(G>40)\text{and}(B>20)\\&\text{and}(\max\{R,G,B\}-\min\{R,G,B\}>15)\\&\text{and}(|R-G|>15)\text{and}(R>G)\text{and}(R>B).\end{aligned} \quad (1)$$

수식 (1)을 이용하여 얻은 피부색 맵(skin pixel map) 예를 Fig. 2(b)에 나타냈다. 다음으로 효과적인 세그멘테이션을 위해 피부색 맵에 세 가지 모폴로지 연산[닫힘(closing), 팽창(erosion), 침식(dilation)]들을 순차적으로 적용한다. 닫힘 연산은 근접한 위치에

(a) (b) (c) (d)

Fig. 2. Result of each step in ROI detection. (a) Input video frame, (b) skin pixel map, (c) morphological operation, (d) final segmentation result(bounding boxes indicate ROIs).

있는 컴포넌트들을 연결시켜준다(5×5 커널 사용). 그리고 침식, 팽창처리는 각각 12×12 커널을 사용하여 잡음이나 불완전한 엣지 등을 제거한다. 모폴로지 연산을 피부색 맵에 적용하여 얻은 결과를 Fig. 2(c)에 나타냈다. 마지막으로 피부색 픽셀들을 포함하는 직사각형의 관심영역들을 결정하는 세그멘테이션을 수행한다(Fig. 2(d))[13].

수식 (1)의 피부색 픽셀 판별 규칙에 기반한 관심영역 검출은 적용이 매우 간단하고 다양한 조명의 얼굴들(Fig. 3(a))의 피부색을 검출한다. 하지만 얼굴이 아닌 배경영역도 관심영역으로 판별할 수 있기 때문에 계산속도 측면에서 효용성이 낮다. 본 논문에서는 불필요하게 검출된 관심영역 영역을 최소화시키기 위해, 이전 비디오 프레임의 얼굴검출 결과를 피부색 모델의 학습에 사용하는 적응적 관심영역 검출 모델 학습(adaptive ROI detection model learning) 방법을 제안한다. 이 방법은 동일한 장소에서 촬영된 얼굴들은 서로 유사한 조명 특성을 가지며(Fig. 3(b))와 같이 연속 프레임 사이에 얼굴의 조명 특성이 급격히 변하지 않는다는 가정에 근간한다. 적응적 관심영역 검출 모델 학습에 기반한 분류는 베이시안 분류기(Bayesian classifier)를 이용하여 비디오 프레임 내 주어진 픽셀 x가 $p(x|skin) \geq p(x|nonskin)$인 경우 피부색 픽셀로 분류한다. 이때 베이시안 분류기를 학습시키기 위한 양(positive) 샘플 픽셀의 집합 P와 음(negative) 샘플 픽셀의 집합 N은 수식 (2)와 같이 정의된다.

$$P = \{(i,j) | ((i,j) \in Face) \cap ((i,j) \in Skin))\}, \quad (2)$$
$$N = \{(i,j) | ((i,j) \in Face^C) \cap ((i,j) \in Skin^C))\},$$

여기서 (i,j)는 픽셀의 좌표이다. $Face$는 각각 현재 프레임(시간 t)의 바로 이전 프레임(시간 t-1)에서 최종 얼굴로 판별된 얼굴영역이다. 그리고 $Skin$은 수식 (1)의 피부색 픽셀 판별 규칙에 의해 피부색으로 판별된 픽셀들의 집합영역이다. 수식 (1)과 (2)를 이용하여 다음과 같은 조건에 따라 적응적 관심영역 검출 모델 학습을 적용한다.

1) 만약 이전 프레임(시간 t-1)에서 얼굴이 검출되지 않은 경우 수식 (1)의 피부색 픽셀 판별 규칙에 의해 현재 프레임에서 모든 픽셀을 검사하여 피부색 픽셀과 피부색이 아닌 픽셀로 분류한다.

2) 만약 이전 프레임(시간 t-1)에서 적어도 한 개

(a)

(b)

Fig. 3. (a) face examples acquired under different kinds of illumination, (b) face examples acquired under similar illumination in the same place.

이상의 얼굴이 검출되면 현재 프레임(시간 t)에 적응적 관심영역 검출 모델 학습을 적용한다. 구체적으로 P와 N에 의해 학습된 베이시안 분류기를 통해 현재 프레임에서 모든 픽셀을 검사하여 피부색 픽셀과 피부색이 아닌 픽셀로 분류한다.

3) 만약 2)의 조건이 충족되면 피부색 픽셀 분류 과정 이후 모폴로지 연산과 세그멘테이션을 앞에 언급한 방법과 동일하게 적용한다.

Fig. 4는 고정된 규칙에 기반한 분류(fixed rule based classification)와 적응적 관심영역 검출 모델에 기반한 분류를 통해 얻어진 피부색 픽셀 맵의 비교를 보여준다. 적응적 관심영역 검출 모델에 기반한 분류의 경우, 이전 프레임(시간 t-1)에서 검출된 조명이 유사한 얼굴들을 이용하여 분류기를 학습시킴으로써 고정된 규칙에 기반한 분류에 비해 더 정확하고 세밀한 관심영역 검출을 수행할 수 있다. 따라서 얼굴검출 계산속도를 높일 뿐만 아니라 오검출을 줄일 수 있다.

대부분의 얼굴영역 검출방법들은 20×20 픽셀 등 지정한 최소 크기 이상의 픽셀 크기를 갖는 영상들로 분류기를 학습시키기 때문에 테스트 과정에서 최소 크기보다 작은 얼굴영역의 경우 검출을 적용하기 어렵다. 그런데 일반적으로 사람의 눈으로 식별 가능한 얼굴은 얼굴영역 검출방법들이 대상으로 하는 최소 얼굴 크기보다 작다(예: 12×12 픽셀[1]). 따라서 얼굴영역 검출방법이 처리할 수 없는 얼굴 크기보다는 작지만 사람의 눈으로 식별 가능한 얼굴영역을 검출하기 위해서 다음의 조건을 만족하는 관심영역의 크기를 늘린다.

$$S^{EYE} \le \min(S^{ROI_W}, S^{ROI_H}) < S^{DETECTOR} \quad (3)$$

수식 (3)에서 S^{EYE}는 사람의 눈이 식별 가능한 최소 얼굴영역의 가로와 세로 픽셀 크기, $S^{DETECTOR}$는 정면 및 측면 얼굴영역 검출방법들이 검출할 수 있는 최소 얼굴영역의 가로와 세로 픽셀 크기이다. 그리고 S^{ROI_W}과 S^{ROI_H}는 각각 관심영역의 가로 픽셀 크기와 세로 픽셀 크기이고, min(,)는 두 개의 인자 중 값이 작은 인자의 값을 반환하는 함수이다. 수식 (3)의 규칙에 부합하는 관심영역은 가로와 세로 중 값이 작은 것을 $S^{DETECTOR}$의 픽셀 크기만큼 리사이징을 통해 늘리고 리사이징을 할 때는 가로 세로 픽셀의 비(ratio)를 유지한다.

3.2 공간 도메인에서의 얼굴분석

제약되지 않은 비디오 감시 환경에서는 단일한 얼굴의 특징만을 이용하여 정확히 얼굴을 검출하기 어렵다. 특정 환경에서 검출이 잘 되는 얼굴 특징은 다른 환경에서 그 효용성이 급격히 떨어질 수 있기 때문이다. 본 절에서는 제약되지 않은 다양한 환경에서 취득된 비디오 프레임에서 안정적인 얼굴검출이 수행될 수 있는 방법을 제시한다. 구체적으로, 얼굴 마스킹에서 필수적인 높은 리콜을 충족시키기 위한 상호 보완적인(complementary) 다중 레벨의 특징 분석과 그들의 융합방법을 설명한다.

먼저 입력받은 관심영역의 각 픽셀위치에 정의되는 12×12 픽셀 이상의 다양한 크기(예: 30×30 픽셀)의 정사각형 스캔 윈도우에서 로우 레벨, 미드 레벨, 하이 레벨에 해당하는 세 가지 다른 종류의 특징 분석을 각각 수행하고 분석 결과에 해당하는 스코어 값들을 생성한다. 각 특징 분석에 대한 구체적인 설명은 다음과 같다.

1) 로우 레벨 특징 분석: 로우 레벨 특징으로서

(a) (b) (c)

Fig. 4. (a) previous frame (at time t−1) with positive (skin) sample pixel regions (inside the white bounding boxes) and negative (non−skin) sample pixels (outside the white bounding boxes), (b) skin pixel map at time t, obtained by fixed rule based classification (using Eq. (1)), (c) skin pixel map at time t, obtained by adaptive ROI detection model based classification.

피부색 정보를 이용한다. 관심영역 내 주어진 스캔 윈도우에서 피부색 픽셀이 차지하는 비율(ratio)을 계산하여 로우 레벨 특징의 스코어 값으로 정한다. 이 때 피부색 픽셀 판별은 3.1절에 설명한 방법을 사용한다. 피부색 정보를 이용한 로우 레벨 특징은 회전(rotation)이 있는 얼굴, 저해상도 얼굴, 블러(blur)가 있는 얼굴 등을 검출하는데 매우 유용하다. 하지만 얼굴과 동일한 색상을 가지는 신체 일부(예: 손, 팔, 다리 등)에 의해 정확도(precision)가 매우 낮다. 이 단점을 보완하기 위해 더 높은 레벨의 특징 분석이 필요하다.

2) 미드 레벨 특징 분석: 미드 레벨 특징으로서 국부적 얼굴 컴포넌트(즉, 양쪽 눈과 입) 검출 결과를 이용한다. Haar-like 특징과 부스팅 분류기(boosting classifier)의 직렬연결로 구성된 캐스케이드(cascade)[5]를 이용하여 양쪽 눈과 입의 검출을 각각 수행하고 총 3개의 컴포넌트 중 몇 개의 컴포넌트가 검출되었는지 비율을 계산하여 스코어 값으로 정한다. 국부적 얼굴 컴포넌트 검출을 이용한 미드 레벨 특징 분석은 로우 레벨 특징 분석에 비해 정확도가 높고 신체 일부(예: 손이나 머리카락)에 의해 얼굴의 일부(예: 입 영역)가 가려지더라도 눈과 같은 다른 컴포넌트들은 검출 가능하기 때문에 가림이 있는 얼굴검출에 유용하다. 하지만 블러된 얼굴이나 저해상도 얼굴에 대해서는 리콜이 낮다는 한계점을 가진다.

3) 하이 레벨 특징 분석: 하이 레벨 특징으로서 전체 얼굴영역 검출결과를 이용한다. Haar-like 특징과 부스팅 분류기의 연결로 구성된 캐스케이드에 기반한 비올라 존스 얼굴검출 방법[5]을 이용하여 정면 얼굴검출(frontal face detection)을 수행한다. 그러면 관심영역 내에서 특정 픽셀 위치에 얼굴로 판별되는 스캔 윈도우들의 위치가 중복되는 경우 얼굴로 판별될 가능성이 높다. 따라서 주어진 스캔 윈도우와 중복되는 얼굴로 판별된 스캔 윈도우들의 개수를 스코어 값으로 정의한다. 좌우로 회전된 얼굴의 리콜을 높이기 위해 정면 얼굴검출 결과로 중복되는 스캔 윈도우가 존재하지 않을 경우, 좌측 프로파일(profile) 얼굴영상들로 학습된 비올라 존스 얼굴검출과 우측 프로파일 얼굴영상들로 학습된 비올라 존스 얼굴 검출을 각각 수행하고 정면 얼굴검출과 동일한 방법으로 스코어 값을 정의한다. 전체 얼굴영역 검출 결과를 이용한 하이 레벨 특징 분석은 로우 레벨이나 미드 레벨의 특징 분석에 비해 정확도가 높다. 하지만 분류기 학습에 사용된 얼굴영상과 조명이나 각도 특성이 다른 얼굴이 테스트에 존재하는 경우 검출의 효용성이 급격히 저하되는 문제점이 있다.

세 가지 종류의 특징 분석을 통해 얻은 스코어 값들은 연결(concatenation)되어 3차원 특징벡터(feature vector)가 생성된다. 이는 패턴인식(pattern recognition)에서 가장 널리 사용되는 분류기 중 하나인 support vector machine(SVM)[15]의 입력으로 사용되어 주어진 스캔 윈도우가 얼굴인지(클래스 1) 얼굴이 아닌지(클래스 2)를 최종 판별한다. 이 때 세 가지 특징 분석이 스코어 융합에 동일하게 기여(contribution)하도록 최소-최대 정규화(min-max normalization)를 수행한다[16]. 패턴이 선형적으로 구분되지 않는 경우를 고려하여 선형(linear) 커널[15] 대신 비선형의 radial basis function(RBF) 커널[15]을 사용한다. SVM 분류기를 학습하기 위해 labeled faces in the wild(LFW) DB[17]에서 수집된 13,233개의 얼굴영상들(클래스 1 학습용)과 MIT-CSAIL DB[18]에서 수집된 2,000개의 비 얼굴 영상들(실내, 실외의 오브젝트나 배경들에 해당함, 클래스 2 학습용)을 사용한다. LFW DB는 5,749명의 매우 방대한 수의 인물들을 포함한다[17]. 또한 제약되지 않은 웹이나 뉴스 기사로부터 취득한 얼굴 영상이므로 얼굴의 각도나 조명이 매우 다양하고 악세사리나 손에 의해 가림도 존재한다[17]. 이러한 얼굴의 다양성은 제안한 얼굴검출 방법의 학습을 실제적(practical)이고 일반화(generalized) 시켜준다. 이는 적용 범위를 넓히고 환경이 바뀔 때마다 SVM 분류기를 일일이 재학습 시키는 번거로움을 덜어준다. Fig. 5는 SVM 분류기 학습용 얼굴 및 비 얼굴 영상들의 예제를 나타낸다. 본 논문에서 사용된 SVM 분류기의 모든 파라미터는 크로스 밸리데이션(cross validation)을 통해 최적화 된다.

3.3 시간 도메인에서 얼굴분석

비디오 프레임 내에서 얼굴 특징들이 불충분한 경우(예: 얼굴 회전각이 너무 커서 눈과 입이 보이지 않거나 가림이 매우 심한 경우 등), 공간 도메인에서의 얼굴분석을 통해 검출하지 못하는 얼굴들이 존재한다. 본 절에서는 공간 도메인에서의 얼굴분석을 보완하기 위해 순방향과 역방향 얼굴 추적에 기반한

Fig. 5. Examples of the training samples for (a) face (class 1) and (b) non-face (class 2).

시간 도메인에서의 얼굴분석을 설명한다. 순방향 얼굴추적은 현재 프레임(시간 t)과 다음 프레임(시간 t+1) 사이에서 수행되고 역방향 얼굴추적은 현재 프레임과 이전 프레임(시간 t-1) 사이에서 수행된다. 본 논문에서는 순방향과 역방향 얼굴추적 방법으로서 mean-shift 알고리즘[19]을 사용한다.

Fig. 6은 시간 도메인에서 얼굴분석을 이용한 얼굴검출 방법을 나타낸다. 시간 t에서의 현재 프레임(current frame)에서는 공간 도메인에서 얼굴영역 검출(FD)의 결과로 얻은 얼굴영역 좌표, 시간 t-1 프레임에서 현재(t) 프레임으로 순방향으로 얼굴 추적(FFT)을 수행한 결과로 얻은 얼굴영역 좌표, 그리고 시간 t+1 프레임에서 현재(t) 프레임으로 역방향으로 얼굴추적(BFT)을 수행한 결과로 얻은 얼굴영역 좌표들을 모두 통합한다. 현재 프레임(시간 t)에서 공간 도메인과 시간 도메인의 얼굴분석을 통해 특정 위치에서 얼굴이 중복 검출되는 경우 통합된 얼굴의 좌표는 공간 도메인의 얼굴분석을 통해 얻은 좌표로 지정한다. 이 방법을 통해 현재(시간 t)와 다음 프레임(시간 t+1) 사이에서 수행되는 순방향 얼굴추적과 현재와 이전 프레임(시간 t-1) 사이에서 수행되는 역방향 얼굴추적의 타겟(target)을 보다 정확하게 업데이트 함으로써 드리프트 현상[20]을 방지할 수 있다. 현재 프레임 기준(시간 t)에서 역방향 얼굴추적을 수행하는 프레임은 총 T개로 시간 (t-T+1)의 프레임까지이다.

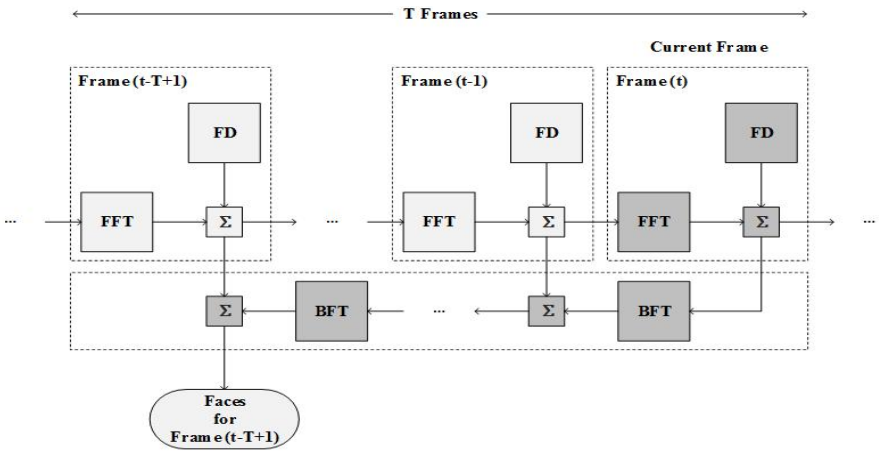

Fig. 6. Block diagram for forward and backward face tracking in temporal domain. (FD: face detection resulted from spatial domain analysis, FFT: forward face tracking, BFT: backward face tracking).

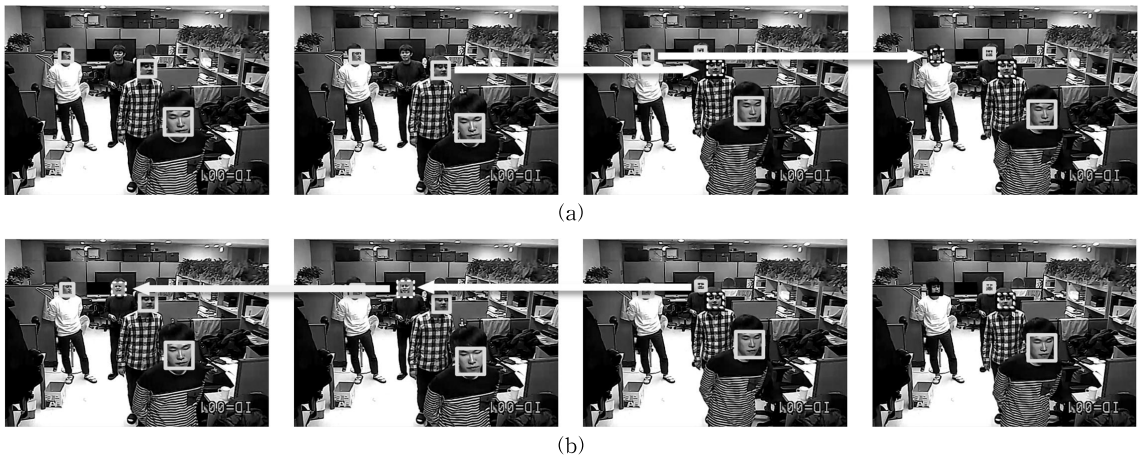

Fig. 7. Illustration of the tracking procedure based on the imaginary face detection result with solid line boxes indicating the result only from face detection in spatial domain, and dot line boxes indicating the additional detection using forward and backward tracking, respectively. (a) face detection result with forward face tracking, (b) face detection result with both forward and backward tracking.

Fig. 7은 시간 도메인에서 얼굴분석의 효용성을 보여준다. Fig. 7(a)는 공간 도메인에서 수행된 얼굴 검출의 결과인 얼굴들(실선 박스 표시)과 순방향 추적에 의해 검출된 얼굴들(점선 박스 표시)을 추가로 나타낸다. 순방향 추적은 초반부의 비디오 프레임에서는 검출이 되었지만 후반부의 비디오 프레임에서 검출이 되지 않은 얼굴영역들을 검출하는데 유용하다는 것을 보여준다. Fig. 7(b)는 역방향 추적에 의해 검출된 얼굴들(점선 박스 표시)을 추가로 나타낸다. 역방향 추적 기반 얼굴영역 검출은 후반부의 비디오 프레임에서는 검출이 되었지만 초반부의 비디오 프레임에서 검출이 되지 않은 얼굴영역들을 검출하는데 유용하다는 것을 보여준다.

4. 실 험

제안하는 얼굴검출 방법의 효용성을 검증하기 위해 두 가지 종류의 데이터 셋을 사용하였다: (1) McGill real-world 얼굴 DB[11]와 (2) 실제 감시환경에서 CCTV 카메라로 취득한 비디오[본 논문에서는 CCTV 감시 비디오(CCTV surveillance video)라고 통칭함]. McGill DB는 32명의 다른 인물에 해당하는 32개의 비디오 시퀀스로 구성된다. 비디오 프레임 사이즈는 640x480 픽셀이고 각 비디오 시퀀스는 300개의 프레임을 포함한다. Fig. 8에서 볼 수 있듯이 McGill DB에서는 다양한 종류의 배경으로 이루어져 있으며 등장인물은 얼굴의 빠르고 큰 폭의 포즈변화(예: 좌우 90도 이내), 스케일 변화, 오브젝트(object)나 손과 같은 신체일부에 의한 가림, 표정변화 등을 포함한다. CCTV 감시 비디오에서는 연구실 내에서 총 7명의 인물이 등장한다. 등장인물들은 자연스럽게 합류하여 서로 바라보며 대화를 나누고 자리를 뜨기도 한다. 따라서 얼굴 포즈변화와 스케일 변화를 포함한다(Fig. 9). 또한 앞사람의 머리에 가려져 가림

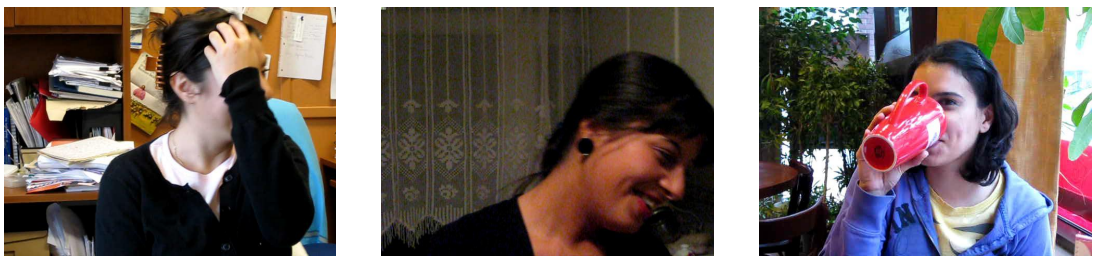

Fig. 8. Example video frames from Public McGill real-world video database.

Fig. 9. Example video frames from CCTV surveillance video.

도 존재한다. 이 비디오는 총 2,719개의 프레임을 포함하며 프레임 사이즈는 1,280×720 픽셀이다.

제안하는 얼굴검출 방법에서 미드 레벨 특징분석(눈과 입 검출)과 하이 레벨 특징분석(정면, 측면 얼굴검출)은 둘 다 공개 라이브러리인 OpenCV[22]의 내장함수를 이용하였다. Fig. 1에서 병렬로 구성된 공간 특징들의 추출 분석 속도를 향상시키기 위해 멀티스레딩(multithreading)[23]을 적용하였다. McGill DB의 경우 매우 심한 포즈변화에 의한 드리프트를 방지하기 위해 얼굴 추적 시 매 프레임마다 얼굴검출 결과를 타겟 업데이트에 이용하였다. CCTV 감시 비디오의 경우 얼굴검출 정확도와 계산속도의 트레이드 오프(trade-off)를 고려하여 네 프레임마다 얼굴검출 결과를 타겟 업데이트에 이용하도록 설정하였다. 역방향 얼굴추적에 사용한 시간축 윈도우 사이즈 T는 3으로 설정하였다. CCTV 감시 비디오는 프레임을 640×480 사이즈로 리사이징하여 제안 방법의 입력으로 사용하였다.

본 실험에서는 비올라 존스 얼굴검출[5], 돌출 맵에 기반한 얼굴검출[9], VeriLook 상용화 얼굴검출[21] 방법들이 제안 방법과의 비교를 위해 사용되었다. 비교를 위해 ROC 커브(receiver operating characteristic curve)를 이용하였다. 여기서 ROC 커브의 가로축은 오검출 개수, 세로축은 리콜률을 나타낸다. 유사한 오검출 개수 범위 내에서 리콜률을 계산하여 동작지점(operating point)을 표시하였다.

Fig. 10은 McGill DB에 대한 ROC 커브를 보여준다. 비올라 존스 얼굴검출은 네 가지 비교 방법 중 가장 낮은 리콜률을 보였다. 이 낮은 리콜률은 단일한 Haar-like 특징과 단일한 모델 사용에 기인한다. 특히 McGill DB는 매우 큰 폭의 포즈변화를 포함하기 때문에 특정 범위의 포즈에 대해서 학습된 모델은 이에 매우 취약할 수밖에 없다. 돌출 맵 기반 얼굴검출은 돌출 맵과 피부색 정보를 사용하여 조금 더 향상된 리콜률을 보였다. 하지만 여전히 80% 이하의 낮은 리콜률을 보였다. 돌출 맵은 배경이 복잡한 텍스처로 구성되어 있는 경우 오브젝트와 배경의 분간이 어려워 그 효용성이 급격히 저하될 수 있다. 제안하는 방법은 매우 심한 포즈변화와 가림이 존재하는 McGill DB에서 83-90%의 높은 리콜률을 달성하였다. Fig. 11과 같이 CCTV 감시 비디오의 경우에도 McGill DB와 비슷한 성능의 경향성을 보였다.

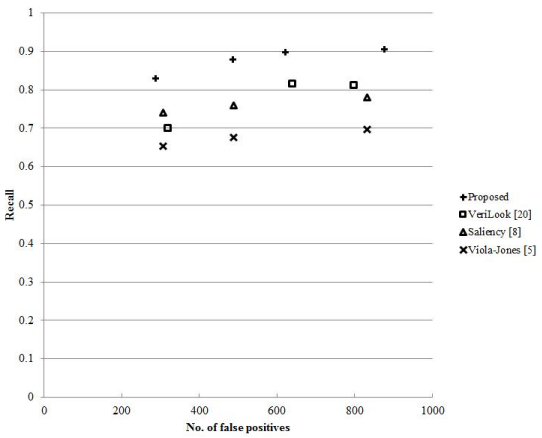

Fig. 10. ROC curve for McGill real-world video database.

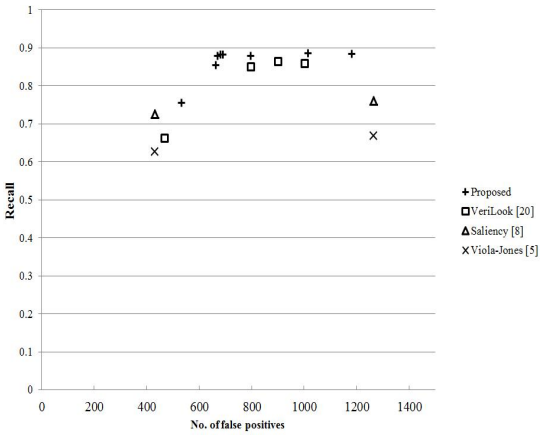

Fig. 11. ROC curve for CCTV surveillance video.

3.50GHz 펜티엄 CPU와 32GB RAM의 PC와 Visual Studio 2010을 이용하여 제안 방법의 동작속도를 측정하였다. 우선 전처리 필터링, 공간 도메인에서의 얼굴분석, 시간 도메인에서의 얼굴분석에 소요되는 시간은 각각 15ms, 71ms, 10ms이다. 앞에서 언급한 바와 같이 얼굴 추적의 타겟 업데이트를 위해 네 프레임마다 공간 도메인에서의 얼굴분석이 사용된다. 타겟 업데이트를 하는 프레임을 처리하는데 소요되는 계산시간은 96ms(=15ms+71ms+10ms)이다. 그리고 타겟 업데이트를 하지 않는 프레임을 처리하는데 소요되는 계산시간은 25ms(=15ms+10ms)이다. 네 프레임마다 타겟 업데이트를 하는 경우 프레임 당 평균 계산시간은 42.75ms[=(96+25×3)/4]이다. 이는 초당 23-24프레임을 처리할 수 있으며 24fps (frames per second) 비디오 기준으로 실시간 동작이 가능하다.

5. 결 론

본 논문에서는 제약되지 않은 환경에서 촬영된 감시 비디오에서 얼굴 프라이버시 보호용 얼굴 마스킹을 위한 얼굴검출 방법을 제안하였다. 제안 방법은 제약되지 않은 다양한 촬영환경에서 강인한 얼굴검출 성능(특히 높은 리콜)을 만족시키기 위해 비디오 프레임 내 다중 레벨의 얼굴 특징 분석을 이용한다. 다중 레벨의 얼굴 특징은 적용 환경에 따라 각기 다른 장단점을 갖기 때문에 감시 비디오가 갖는 특성(조명이나 인물의 얼굴 각도 등)이 다양하더라도 서로 상호보완적으로 작용하여 안정적인 얼굴검출 성능을 유지할 수 있다. 또한 비디오의 연속 프레임들 간에 순방향와 역방향 얼굴추적을 수행함으로써 비디오 프레임 내에 얼굴정보가 불충분할 때에도 얼굴을 검출하여 리콜을 높일 수 있다. 심한 포즈 변화와 가림이 존재하는 공인 DB와 실제 환경에서 촬영된 CCTV 비디오에서 비교실험을 수행하였다. 그 결과, 상용화 제품을 포함한 최신 얼굴검출 방법들에 비해 오검출 대비 높은 리콜을 보였다. 또한 초당 약 23-24 프레임의 실시간 동작이 가능하여 대규모(large-scale) CCTV 감시 시스템에 효율적으로 적용될 수 있음을 보였다.

본 논문에서는 계산속도 향상을 위해 피부색 픽셀 판별에 기반한 관심영역 선택 방법을 사용하였다. 하지만 이 방법은 야간 영상과 같은 매우 어두운 조명에 적용하기 어렵다. 향후 주제로서 야간 영상에서 피부색 정보를 대체할 관심영역 선택방법에 대해 연구할 것이다.

REFERENCE

[1] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, and A. Bissacco, "Large-scale Privacy Protection in Google Street View," *Proceeding of IEEE International Conference Computer Vision,* 2009, pp. 2373-2380.

[2] J. Crowley, J. Coutaz, and F. Berard, "Things that See," *Communications of the ACM*, Vol. 43, No. 3, pp. 54-64, 2000.

[3] C. Neustaedter and S. Greenberg, *Balancing Privacy and Awareness in Home Media Spaces*, University of Calgary, Department of Computer Science, Canada, 2003.

[4] M. Boyle, C. Edwards, and S. Greenberg, "The Effects of Filtered Video on Awareness and Privacy," *Proceedings of the 2000 ACM Conference Computer Supported Cooperative Work*, 2000, pp. 1-10.

[5] P. Viola and M.J. Jones "Robust Real-time Face Detection," *International Journal of Computer Vision,* Vol. 57, No. 2, pp. 137-154, 2004.

[6] J. Zhang, G.A. Ramirez, and O. Fuentes, "Face Detection in Low-resolution Color Images," In: *Image Analysis and Recognition*, Springer Berlin Heidelberg, 2010, pp. 454-463.

[7] B. Froba and A. Ernst, "Face Detection with the Modified Census Transform," *Proceeding of IEEE International Conference on Automatic Face and Gesture Recognition,* 2004, pp. 91-96.

[8] Y.-H. Jung, Y.-M. Song, and Y.-H. Ko, "Inclined Face Detection using JointBoost Algorithm," *Journal of Korea Multimedia Society*, Vol. 15, No. 5, pp. 606-614, 2012.

[9] A. EL-Barkousky, H. Rara, and A. Farag, "Face Detection at a Distance using Saliency Maps," *Proceeding of IEEE Computer Society Conference Computer Vision and Pattern Recognition Workshops,* 2012, pp. 31-36.

[10] S.Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical Learning of Multi-view Face Detection," *Proceeding of Computer Vision-ECCV 2002*, Springer Berlin Heidelberg, 2002, pp. 67-81.

[11] M. Demirkus, J.J. Clark, and T. Arbel, "Robust Semi-automatic Head Pose Labeling for Real-world Face Video Sequences," *Multimedia Tools and Applications,* Vol. 70, No. 1, pp. 495-523, 2014.

[12] P. Peer, B. Batagelj, J. Kovac, and F. Solina, "Color-based Face Detection in the "15 Seconds of Fame" Art Installation," *Proceeding of A Special Issue of Integrated Computer-Aided Engineering,* 2003, pp. 38-47.

[13] S. Suzuki and K. Abe, "Topological Structural Analysis of Digitized Binary Images by Border Following," *Computer Vision, Graphics, and Image Processing,* Vol. 30, No. 1, pp. 32-46, 1985.

[14] P. Langley, W. Iba, and K. Thompson, "An Analysis of Bayesian Classifiers," *Proceedings of the Tenth National Conference on Artificial Intelligence,* 1992, pp. 223-228.

[15] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Knowledge Discovery and Data Mining,* Vol. 2, No. 2, pp. 121-167, 1998.

[16] A. Jain, K. Nandakumar, and A. Ross, "Score Normalization in Multimodal Bioimetric Systems," *Pattern Recognition,* Vol. 38, No. 12, pp. 2270-2285, 2005.

[17] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*, *Technical Report 07-49*, UMass, 2007.

[18] A. Torralba, K.P. Murphy, and W.T. Freeman, "Sharing Features: Efficient Boosting Procedures for Multiclass Object Detection," *Proceeding of IEEE Computer Society Conference on. Computer Vision and Pattern Recognition,* 2004, pp. II-762-II-769.

[19] K. Fukunaga and L.D. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *IEEE Transactions on Information Theory,* Vol. 21, No. 1, pp. 32-40, 1975.

[20] M. Kim and S. Kumar, "Face Tracking and Recognition with Visual Constraints in Real-World Videos," *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2008, pp. 1-8.

[21] Neurotechnologija: Verilook SDK, *neurotechnologija Biometrical and Artificial Intelligence Technologies*, http://www.neurotechnologija. com.

[22] G. Bradski, The OpenCV Library, *Dr. Dobbs Journal of Software Tools,* 2000.

[22] Multithreading in C/C++, http://www.boost.org/doc/libs/1_38_0/doc/html/thread.html (accessed Mar., 25, 2015).

이 승 호

2011년 KAIST 전기 및 전자공학과 공학석사
2012년 University of Toronto, Dept. of EECS 방문연구원
2015년 KAIST 전기 및 전자공학부 공학박사
관심분야 : Face analysis, Object recognition, Multimedia retrieval and indexing

문 정 익

2015년 KAIST 전기 및 전자공학과 공학석사
현재 SK Hynix 연구원
관심분야 : Image and video processing, Object detection, Object recognition

김 형 일

2013년 KAIST 전기 및 전자공학과 공학석사
현재 KAIST 전기 및 전자공학부 박사과정
관심분야: Image and video processing, Pattern recognition and machine learning, Face recognition

노 용 만

1992년 KAIST 전기 및 전자공학과 공학박사
1987년 Columbia University 연구원
1996년 University of California, Berkeley 연구원
2006년 University of Toronto 초빙교수
현재 KAIST 전기 및 전자공학과 정교수
관심분야 : Image and video processing, 3D video processing, Image and video feature extraction/recognition, Object detection/recognition, and medical image processing