Journal of Korea Robotics Society (2017) 12(3):313-321
https://doi.org/10.7746/jkros.2017.12.3.313 ISSN: 1975-6291 / eISSN: 2287-3961

다중크기와 다중객체의 실시간 얼굴 검출과 머리 자세 추정을 위한 심층 신경망

Multi-Scale, Multi-Object and Real-Time Face Detection and Head Pose Estimation Using Deep Neural Networks

안 병 태[1], 최 동 걸[†], 권 인 소[2]

Byungtae Ahn[1], Dong-Geol Choi[†], In So Kweon[2]

Abstract One of the most frequently performed tasks in human-robot interaction (HRI), intelligent vehicles, and security systems is face related applications such as face recognition, facial expression recognition, driver state monitoring, and gaze estimation. In these applications, accurate head pose estimation is an important issue. However, conventional methods have been lacking in accuracy, robustness or processing speed in practical use. In this paper, we propose a novel method for estimating head pose with a monocular camera. The proposed algorithm is based on a deep neural network for multi-task learning using a small grayscale image. This network jointly detects multi-view faces and estimates head pose in hard environmental conditions such as illumination change and large pose change. The proposed framework quantitatively and qualitatively outperforms the state-of-the-art method with an average head pose mean error of less than 4.5° in real-time.

Keywords Head Pose, Deep Learning, Convolutional Neural Network

1. 서 론

Human-robot interaction (HRI), 지능형 자동차, 보안 시스템 등에서 가장 빈번하게 수행하는 작업 중 하나가 얼굴 인식, 표정 인식, 졸음 감지 등과 같은 얼굴관련 응용작업이다. 이런 작업들을 할 때, 사용자의 얼굴 자세를 정확하게 추정하는 것이 중요하다.

컴퓨터 비전의 관점에서 머리 자세 추정이란 입력된 얼굴영상으로부터 위치와 방향(roll, pitch, yaw)을 유추하는 과정인데, 현존하는 방법론들은 크게 두 가지로 분류할 수 있다: 첫째는 얼굴의 수학적 모델을 사용하여 입력 영상에 가장 잘 정합되는 가상의 얼굴모델을 생성하여 그 모델의 자세를 추정하는 방법들이고, 둘째는 외형 정보를 바탕으로 입력 영상을 특정 자세로 분류해내는 방법들이다.

생성형(generative) 방법들은 기하학적 단서들이나 가변적인 얼굴 모델을 사용한다. 이 방법들은 개별적인 범주가 아니라 연속적인 값들을 출력하고, 머리 자세뿐만 아니라 다양한 응용을 위한 얼굴 특징점의 위치까지 얻을 수 있다는 장점이 있다. 하지만 이런 방법들은 얼굴의 특징점 검출에 매우 의존하고 있기 때문에, 얼굴 특징점 검출이 어려운 환경-예를 들어 머리의 자세나 표정의 변화, 가려짐, 그리고 저해상도의 영상-에서는 머리 자세 추정 역시 매우 취약하다.

Received : Mar. 10. 2017; Revised : Apr. 27. 2017; Accepted : Jul. 28. 2017
※This research was supported by the Ministry of Trade, Industry & Energy and the Korea Evaluation Institute of Industrial Technology (KEIT) with the program number of "10060110".
†Corresponding author: Post doctor, Department of Electrical Engineering, KAIST, Daejeon, Korea (dgchoi@rcv.kaist.ac.kr)
[1]Ph.D candidate, Robotics Program, KAIST (btahn@rcv.kaist.ac.kr)
[2]Professor, Department of Electrical Engineering, KAIST (iskweon@kaist.ac.kr)

Copyrightⓒ KROS

식별형(discriminative) 방법들은 얼굴 전체의 시각 특징들과 함께 기계학습법을 사용한다. 이 방법들은 비교적으로 큰 머리 자세변화와 저해상도 영상에 강인하다. 하지만 대부분의 방법들이 특정 간격으로 나누어진 얼굴 영상들을 클래스화하여 사용하고, 새로운 입력 영상을 해당 범주로 분류하는 방식이다. 따라서 그 추정 값들은 연속적인 값이 아닌 큰 간격으로 범주화 되어있다 (보통 10° 이상).

본 연구는 식별형 방법론 중 하나인 깊은 신경망과 다중 작업 학습(multi-task learning) 방식을 사용하여 이러한 문제를 해결하였고, 실시간 계산에 유리한 그레이스케일 저해상도 영상을 사용한 강인한 머리 자세 추정법을 제안한다. 또한 제안된 방식은 정성적, 정량적 평가를 통하여 기존의 방식들에 비하여 우수성을 입증하였다.

2. 관련 연구들

본 섹션에서는 접근법에 따라 관련 연구들을 분류하고, 각 범주에서 대표적인 알고리즘들의 장단점을 논의한다.

2.1 생성형(generative) 방법들

일반적으로 얼굴은 형태(shape)와 외형(appearance)의 두 가지 가변적인 항으로 수학적으로 모델링 된다. 모델링 기반의 방법은 모델을 사용하여 여러 자세와 표정의 가상의 얼굴 모형을 생성해 낼 수 있기 때문에, 가상모델이 얼굴영상에 정합이 잘 된다면 자세의 정확도도 높고 연속적인 머리 자세 값의 출력이 가능하다. Active shape models (ASMs)[1]와 active appearance models (AAMs)[2]는 아주 유명한 통계적 얼굴 모델(statistical models of face)이다. 이 연구들은 얼굴 특징점 검출을 위해 제안되었지만, 점차 머리 자세 추정 연구로 확장되었다[3]. Vicente 등[4]은 모델 기반 접근법의 상용프로그램 IntraFace[3]를 이용하여 얼굴 특징점과 자세를 검출하고 운전자의 시선을 판별하는 하는 시스템(eyes off the road, EOR)을 개발하였다. 그들이 이용한 IntraFace 시스템은 supervised descent method (SDM)를 제안하여 parameterized appearance models (PAMs)의 비선형 최소 제곱(non-linear least squares, NLS) 문제를 해결함으로써, 25 fps의 속도로 평균 오차 4.6° 정도의 성능을 보였다.

하지만 이러한 모델 기반의 방법들은 특징점 검출에 머리 자세의 정확도가 많이 의존되기 때문에 흔들림, 노이즈, 저해상도, 그리고 조명변화 및 큰 자세의 변화에 매우 민감하다.

2.2 식별형(discriminative) 방법들

식별형 방법들은 2D 영상에서의 외형과 3D 머리 자세 사이의 관계를 직접적으로 추론한다. 이는 머리 자세 추정에 사용되는 중요한 요소들을 암묵적으로 학습하는 회귀자(regressor)에 의해 이뤄진다. Balsaubramanian 등[5]과 Foytik 등[6]은 얼굴영상의 고차원 공간을 저차원 매니폴드로 대응하는 매니폴드 임베딩 프레임웍을 제안했다. Foytik은 매니폴드 임베딩 프레임웍과 머리 자세를 coarse-to-fine 방법으로 추정하는 두 단계로 이루어진 방법을 제안했다. 하지만 결과의 오차는 13° 보다 크다. Zhu 등[7]은 얼굴 검출, 머리 자세, 얼굴 특징점 검출을 위한 통합모델을 제안했다. 그들은 yaw각의 회전에 따른 외형변화의 관계를 찾기 위해 mixture of tree-structured part models를 사용하였다. 그러나 그것이 통합 프레임웍이긴 하지만, 단지 입력영상을 몇 개의 범주화된 yaw 각도로만 분류해줄 따름으로 한계가 분명하다. 또한 VGA해상도 영상을 처리하는데 수초에서 수십초의 많은 처리 시간이 필요하다.

이런 개별적 분류 인덱스를 사용하는 접근법들과는 반대로, 몇 개의 연구들은 연속적인 머리 자세 추정을 위해서 기계학습 기법에 깊이(depth) 정보를 활용하였다. Breitenstein 등[8]은 거리영상(range image)을 기준 자세들로 정렬(align)함으로 머리 자세를 추정하였으며, GPU기반의 구현으로 10 fps의 처리 속도를 보였다. Fanelli 등[9]은 실시간으로 머리 자세 추정을 위해 랜덤 포레스트 기반의 보팅 프레임웍을 도입하였으며, 영상, 깊이, 그리고 머리자세 값으로 구성된 database (DB)도 제공하였다.

깊이 정보 때문에 이 방식은 밤에도 상대적으로 강인한 결과를 보이고 3차원 얼굴 모델을 얻을 수 있는 장점이 있지만, 이를 위해서는 Kinect 같은 특별한 장치가

필요하고 적외선을 사용하는 센싱 방식 때문에 실내에서만 사용할 수 있다는 단점이 있다.

2.3 깊은 신경망(deep neural network)

그래픽 프로세싱 유닛(GPU)이 발전하고 빅데이터에 대한 접근이 편리해 지면서, 딥러닝 기술이 활발하게 각광을 받아왔다. 이러한 방법들 중에, Convolutional Neural Network (CNN)[10]이 컴퓨터비전 업무에 성공적으로 적용되어왔다. 최근에, Deep Neural Network (DNN)들이 얼굴 관련 어플리케이션에도 널리 사용되고 있다.

Sun 등[11]은 DNN을 coarse-to-fine 방식의 얼굴 특징 세트 검출에 도입하였다. 하나의 DNN과 두 개의 얕은 신경망으로 이루어진 3단계 케스케이드 구조를 제안하였다. 그들은 또한 절대값 정류(absolute value rectification)와 국소 가중치(local weight) 공유 같은 기술의 영향도 분석하였다.

Li 등[15]은 CNN을 얼굴 검출 작업에 적용하였다. 그들은 보다 정교한 얼굴 검출을 위하여 전체 알고리즘을 검출단계와 보정단계로 나누었고, 각 단계에서 얕은 신경망 두 개와 깊은 신경망 하나를 연결(cascading)하는 방식으로 전체 알고리즘을 구성 하였다.

Ahn 등[12]은 DNN을 사용하여 시각 정보와 3차원 머리 방향의 사이의 관계를 학습하는 방법을 제시하였다. 그들은 다양한 DNN 매개변수들을 실험하고 선정하여 머리 방향 추정을 위한 DNN 구조를 개발했다. 추정된 값의 안정화를 위해 후처리 방식으로 파티클 필터를 이용한 추적을 적용하였다. 제안된 방법은 얼굴이 바르게 검출되었다는 가정 하에, DNN을 이용하여 실시간 이상의 속도로 3° 정도의 오차를 보이는 비교적 정확한 머리 방향 값을 추정하였다. 그러나 그들의 방법에는 머리 자세 추정에 필수적인 얼굴검출 단계가 빠져있으며, 그들이 사용한 평가방법은 훈련세트와 평가세트데이터 간에 독립성을 보장하지 못하기 때문에 모든 환경에서 그들이 제안한 성능이 나오기는 어렵다는 단점이 있다.

3. 다중 작업 심층 신경망

본 섹션에서는, 먼저 우리가 제안한 실시간 머리 자세 추정 알고리즘 전체를 소개하고, 후에 멀티태스크 학습 네트워크와 학습 데이터셋에 관한 자세한 내용을 설명한다.

3.1 제안된 알고리즘의 개요

Fig. 1은 우리가 제안한 실시간 머리 자세 추정 알고리즘의 개요이다. 우리는 얼굴 검출, 정확한 바운딩 박스 추출, 그리고 머리 자세 추정 세 가지 문제를 다중 작업 학습(multi-task learning)을 통해 학습된 공통의 특징을 사용함으로써 해결한다. 전체 알고리즘의 과정을 간단히 설명하면 다음과 같다. GPU를 사용한 빠른 연산을 위해 입력영상 전체를 바운딩 박스의 크기에 비례하는 건너뜀(stride) 값과 함께 많은 바운딩 박스들로 쪼개어 검출 윈도우 묶음(batch of detection windows)을 만든다. 이때 바운딩 박스는 멀티스케일 검출을 위해 다양한 크기로 구성한다. 이를 네트웍의 입력크기인 $48 \times 48 \times N$ 크기로 변환하고 학습된 다중 작업 네트워크에 입력한다. 제안된 다중 작업 네트워크에서 첫 번째 작업인 얼굴 검출을 위한 이진 분류(binary classification)가 수행되

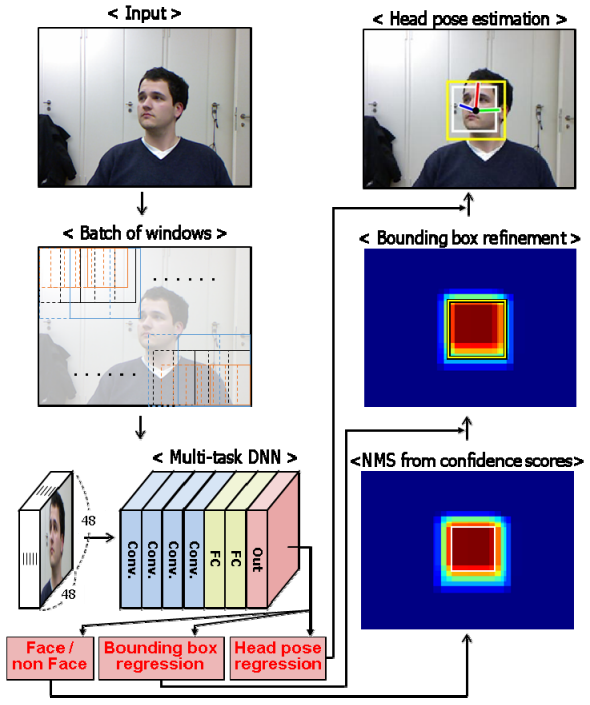


Fig. 1. Overview of the proposed real-time head pose estimation framework using DNN

고, 그 결과 검출 윈도우 각각의 점수 값에 따른 히트맵을 얻게 된다. 히트맵에 중복처리를 방지하기 위하여 non-maximum suppression (NMS)이 적용되고 초기 위치의 바운딩 박스를 얻는다. 입력영상에서 초기 바운딩 박스 위치의 영상을 제안된 다중 작업 네트워크의 두 번째 작업인 바운딩 박스 회귀를 수행하며, 그 결과 머리 자세 추정에 가장 적절한 위치와 크기의 바운딩 박스를 얻는다. 최종적으로 제안된 다중 작업 네트워크의 3번째 작업인 머리 자세를 추정한다. 여기서, 상기의 세 가지 작업들에 모두 제안된 다중 작업 네트워크에서 학습된 일련의 특징 세트(feature set)가 공통으로 사용된다.

우리의 연구와 가장 관련이 있는 연구는 DNN을 사용하여 2D 영상과 3D 머리 방향간의 대응 관계를 학습한 Ahn 등[10]의 연구이다. 하지만 그들이 제안한 네트워크는 얼굴이 정확히 검출됐다고 가정하고, 단지 3D 머리 방향만을 추정하는 것이므로 다양한 어플리케이션들에 실제적으로 사용하는데 한계가 있다. 본 연구에서는 관련 연구를 더욱 확장시켜 얼굴 판단, 최적의 바운딩 박스 검출, 그리고 5D (x, y, roll, pitch, yaw) 머리 자세 추정까지의 세가지 작업들에 사용되는 공통의 특징 세트를 학습하는 다중 작업 네트워크를 제안하며 Fig. 2에 구조와 매개변수들이 나타나 있다.

네트워크의 입력은 *48* × *48* 크기의 그레이 스케일 영상이다. 특징세트 추출 부분은 4개의 컨볼루션층, 3개의 풀링층, 그리고 하나의 공유되는 완전 연결 층(shared fully connected layer)과 각 작업별로 하나씩 완전 연결층이 추가로 구성이 되어있다. 활성 함수로 정류된 선형 유닛(rectified linear unit, ReLU)이 사용되었고, 풀링층에서는 최대값 선택(max-pooling)이 사용되었다. 4개의 컨볼루션층에 이어 나오는 완전 연결 층은 다중 작업 추정(multi-task estimation) 부분에서 공유하는 64차원의 특징벡터를 생성한다.

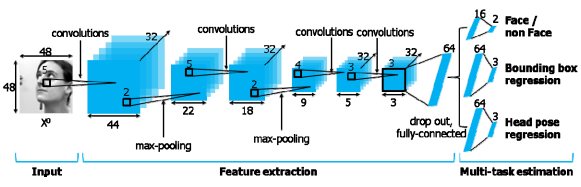


Fig. 2. Multi-task learning DNN and its parameters

3.2 얼굴 검출

얼굴 검출 방식은 슬라이딩 윈도우와 이진 분류에 의한 검출 방식이며, 출력 층의 비용 함수는 소프트맥스(soft-max) 함수를 사용한다. 입력 영상을 꼼꼼히 스캔하기 위해서 여러 스케일로(본 연구에서는 세 가지) $s_w \times s_w$ 크기의 검출 윈도우를 s_t 크기의 건너뜀(stride) 값으로 $w \times h$ 크기의 입력 영상을 검출 윈도우 묶음으로 변환한다. 이를 *48* × *48* × *n* 크기로 변환하고 이진 분류를 수행한다. 여기서 $n = [\{(w-s_w)/s_t\}+1] \times [\{(h-s_w)/s_t\}+1]$. 생성된 히트맵의 확률 값을 바탕으로 정사각형의 초기 바운딩 박스를 생성한다. 많이 중복된 검출 윈도우를 제거하고 보다 정확한 바운딩 박스를 얻기 위해 NMS를 수행한다. 본 연구에서 NMS의 목적은 인접한 여러 영역들이 바운딩 박스로 검출된 경우, 그 중 이진분류의 점수가 극대인 바운딩 박스만을 남기고 나머지를 제거하는 것이다. 그렇지만 최종 자세 추정의 성능향상을 위해서 본 연구에서는 NMS의 결과를 최종 바운딩 박스로 사용하지 않고, 보다 정확한 바운딩 박스의 위치 추정을 위해 다중 작업 네트워크에 바운딩 박스 회귀(bounding box regression) 작업을 추가하였다.

3.3 바운딩 박스 회귀(bounding box regression)

[3]과 [7] 등에서 제안한 머리 자세 추정의 전통적인 방법론은 먼저 입력영상으로부터 얼굴을 검출하고, 검출된 윈도우로 자세 추정을 수행한다. 그러나 단순히 검출된 윈도우로 자세 추정을 수행하게 되면 성능이 저하될 수 있다. 왜냐하면 얼굴 검출기는 일반적으로 실제값(ground-truth) 윈도우에서 50% 이상만 걸치면 참 값(positive)으로 검출되었다고 판단하기 때문이다. 자세 추정단계에서는 얼굴의 주요 특징들의 검출 유무가 매우 중요한데, 비록 얼굴이 검출되었다고 한들 Fig. 3과 같이 얼굴의 중요한 특징들(눈, 코, 입, 턱 등)이 검출 윈도우에 포함되지 않으면 자세 추정 결과의 정확도가 많이 떨어질 수밖에 없다. 또한 실시간 어플리케이션에서 안정적이지 않은 검출 윈도우 위치에 의해서 자세 추정 결과도 그에 따라 시간 도메인에서 매우 불안정한

Fig. 3. Samples overlapped with ground-truth box over 50%

결과가 얻어진다. 즉 다시 말해 비록 얼굴 검출 성공률이 높을 지라도 자세 추정의 성능은 매우 낮기 쉽다. 따라서 자세 추정에 있어서는 무엇보다 정확한 위치의 바운딩 박스의 추출이 중요하다. Fanelli 등[9]은 깊이 센서(Kinect)를 사용하여 깊이 정보를 기반으로 얼굴 영역을 검출함으로써 원천적으로 이 문제를 해결하였다. 본 연구는 카메라 영상만을 사용하는 방법론으로써 이 문제를 해결하기 위해 바운딩 박스 회귀 기법을 도입하였다. 얼굴 검출 박스 회귀 이슈는 얼굴 검출, 자세 추정과 같이 얼굴관련 작업이기 때문에 동일한 특징세트를 공유한다고 가정하고 다중 작업 네트워크를 구성하였다. 다중 작업 네트워크는 세 가지의 작업 모듈로 이루어져 있다. 첫 번째 모듈은 초기 바운딩 박스의 위치를 출력한다. 그 위치의 영상이 두 번째 모듈인 바운딩 박스 회귀 네트워의 입력이 되어 새로운 바운딩 박스의 위치가 추정되는 방식이다. R-CNN[13]에서도 별도의 CNN을 사용해 클래스 별 바운딩 박스 회귀를 사용하였다.

훈련 단계에서 입력은 N개의 훈련쌍이다$(\boldsymbol{i}, \boldsymbol{g})$. 여기서 $\boldsymbol{i} = (i_x, i_y, i_z)$이다. i_x, i_y는 초기 바운딩 박스의 좌상단의 픽셀 좌표이고 i_z는 정사각형 바운딩 박스의 한 변의 크기이다. 각 실제 값(ground-truth) 바운딩 박스는 같은 방법으로 $\boldsymbol{g} = (g_x, g_y, g_z)$ 이다. 목표는 초기 바운딩 박스 $\boldsymbol{i}$를 실제 값(ground-truth) 박스 $\boldsymbol{g}$로 대응하는 변환을 학습하는 것이다.

그 변환은 $d_x(\boldsymbol{i})$, $d_y(\boldsymbol{i})$, 그리고 $d_w(\boldsymbol{i})$ 세 개의 항으로 수식화 한다. 처음 두 개는 바운딩 박스 $\boldsymbol{i}$의 좌상단 지점의 스케일 독립적인 이동(scale-invariant translation)을 나타내고, 세 번째 항은 바운딩 박스 $\boldsymbol{i}$의 넓이의 log공간에서의 이동이다. 이 함수들에 의해 입력 바운딩 박스 $\boldsymbol{i}$는 아래의 변환 함수에 의해 예측되는 실제 값 (ground-truth) 바운딩 박스 $\hat{g}$로 변환된다.

$$\begin{aligned}\hat{g}_x &= i_w d_x(i) + i_x \\ \hat{g}_y &= i_w d_y(i) + i_y \\ \hat{g}_w &= i_w e^{d_w(i)}\end{aligned} \tag{1}$$

훈련쌍$(\boldsymbol{i}, \boldsymbol{g})$에 대한 회귀의 목적 함수 $\boldsymbol{t}$는 아래와 같이 정의된다.

$$\begin{aligned}t_x &= (g_x - i_x)/i_w \\ t_y &= (g_y - i_y)/i_w \\ t_w &= \log(g_w / i_w)\end{aligned} \tag{2}$$

3.4 머리 자세 회귀

Fig. 2에서 3번째 작업인 머리 자세 회귀의 손실 함수는 아래와 같은 유클리디안 거리를 사용한다.

$$E(X^k; W) = \sum_k \left\| g^k - f(X^k; W) \right\|_2^2 \tag{3}$$

여기서 k는 훈련 샘플의 인덱스이고, f는 추정되는 머리 자세(roll, pitch, yaw)이다. $\boldsymbol{W}$는 컨볼루션 필터의 가중치들 즉, 본 네트워크에서 학습되는 특징 세트이고 $\boldsymbol{g}^k$는 실제 값(ground-truth) 머리 자세이다.

Fig. 4는 첫 번째 층의 학습된 필터 중 몇 개를 나타낸 것이다. 특징필터가 구조를 띄고 있고, 그 형태가 서로 연관(correlated)되어있지 않으며, 출력을 보면 얼굴의

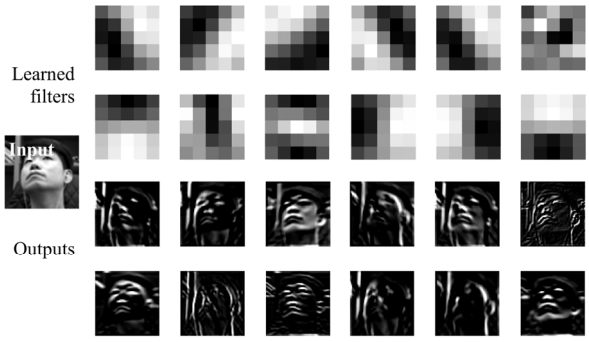


Fig. 4. Some learned features and their corresponding outputs

방향을 추정하는데 중요한 부분들(눈, 코, 머리, 턱 등)이 강조되어 있는 것을 볼 수 있다. 여기서, 학습된 특징들로 필터링 된 영상을 보면 강조 되어 있는 부분들이 얼굴 검출, 바운딩 박스 위치 및 크기 조절, 그리고 머리 자세 추정에 공통적으로 사용할 수 있는 특징들임을 확인할 수 있다.

3.5 훈련 데이터셋

본 연구에서는 총 3가지의 데이터셋을 사용하였다: Biwi Kinect Head Pose Database (BKHP)[9], 본 연구에서 추가로 획득한 컬러 영상 데이터(RCVface), 그리고 AFLW 데이터셋[14]. BKHP는 20명(4명은 두 번 등장하나 옷과 헤어스타일이 다름)의 상반신을 녹화한 15,678장의 영상으로 이루어져있다. 깊이 맵으로 만든 개별 사용자별 3D 템플릿을 구축하고, 실시간 추적기로 3D 템플릿을 사용자에게 정합하여 3D 모델의 머리 자세를 기록하는 방식으로 실제 값(ground-truth) 머리 자세 정보를 획득하였다. 개별 사용자의 3D 템플릿은 오프라인으로 구축한다. BKHP 데이터의 머리 방향은 roll ±50°, pitch ±60°, 그리고 yaw ±75° 범위를 커버한다. 한편 본 연구는 BKHP 데이터에서 제공하는 깊이 정보는 사용하지 않고, 컬러 영상만을 그레이스케일로 변환하여 사용하는데 그 컬러 영상들의 배경은 대부분 흰색으로 이루어진 고정된 두 가지의 배경으로만 이루어져있다. 이를 토대로 학습을 한다면 특징들이 동질(homogeneous)의 배경에 과정합(over-fitting)되어 다양한 배경에 대해서는 제대로 된 성능을 낼 수가 없다. 그리고 24명의 사람 역시 일반적인 얼굴 검출기로 사용하기엔 데이터의 다양성과 양이 너무 적다. 이를 해결하기 위해 RCVface 데이터셋과 AFLW 얼굴 검출 데이터셋 두 가지를 추가하였다. RCVface 데이터는 앞서 설명한 BKHP와 똑같은 형식으로 다양한 배경에서 추가로 23명, 약 30,000장을 획득하였다. 한편 대부분의 얼굴 검출용 DB들이 정면 얼굴이 압도적으로 구성되어있는 반면 AFLW 데이터셋은 다양한 배경에 다양한 자세로 구성되어있어 본 연구에 추가하기에 적합한 데이터셋이다.

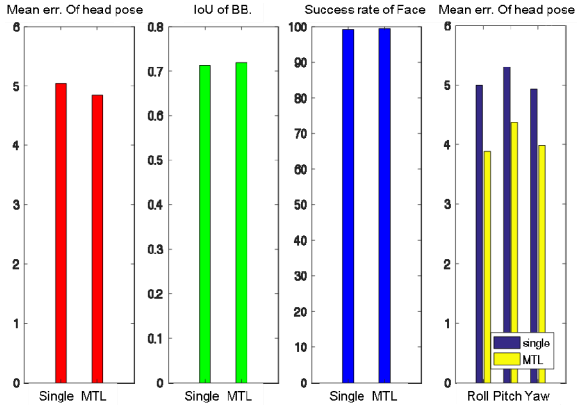
4. 실험 결과

본 섹션에서는 제안된 알고리즘의 성능을 정량적 그리고 정성적으로 검증하기 위하여 여러 가지 측면에서 실험을 수행한다. 먼저 제안된 다중 작업 깊은 신경망 방법론의 타당성을 검증하고, 머리 자세 추정에서 제안된 알고리즘의 성능을 테스트 하며, 이를 최신 기술의 성능과 비교한다. 그리고 여러 가지 상황에서 제안된 알고리즘의 수행 결과를 영상으로 보여준다.

평가에 사용된 DB는 BKHP, RCVface, AFLW이다. 공개 되어있는 머리 자세 DB들의 경우 대부분이 공통적인 기준이 없이 자기들만의 방식으로 실제 값(ground-truth)을 할당(annotate) 하였기에 실제 평가에 활용하기에 적합하지 않다. 따라서 머리 자세 추정 성능은 본 연구에 활용한 DB의 평가세트(validation set)로 평가하고 동일 DB를 사용한 최신기술의 성능과 비교한다.

제안 된 다중 작업 깊은 신경망 방법론의 타당성을 검증하기 위해, 세 가지 작업들에 대해 공통의 특징 세트를 학습한 제안된 방법과, 개별의 작업마다 단독의 DNN을 이용하여 각각의 특징 세트를 학습한 방법을 비교하였다. 학습에 사용된 DB는 BKHP, RCVface, 그리고 AFLW로 동일하며, 머리 자세 평가에 사용된 DB는 BKHP와 RCVface DB에서 훈련에 사용되지 않은 개체(사람)들이며, 이진 얼굴 분류(binary face classification) 평가에 사용된 DB 역시 BKHP, RCVface, 그리고 AFLW에서 학습에 사용되지 않은 평가 세트(validation set)이다.

Fig. 5(a)는 단독의 DNN을 각각 훈련 시 평가세트에 대하여 수렴한 훈련 오차 값을 나타낸다. 왼쪽은 머리 자세의 평균오차(mean error), 중간은 바운딩 박스의 intersection over union (IoU) 값(학습시에는 식 (2)를 비용 함수로 사용하였지만 평가에는 직관적 이해를 위해 IoU값으로 나타내었다.), 그리고 오른쪽은 얼굴 분류의 성공률이다. Fig. 5(a)를 보면 동일한 DB에 대해 단독 DNN들과 다중 작업 DNN의 훈련 시 수렴한 훈련 오차 값이 거의 동일한 것을 볼 수 있다. 이는 동일한 DB로 동일한 구조의 네트워크(다중작업 DNN도 내부의 개별 작업 모듈들은 각 작업에 대한 단독 DNN과 동일한 구조)을 훈련시킨 당연한 결과이다. 한편 Fig. 5(b)는 Fig.



(a) (b)

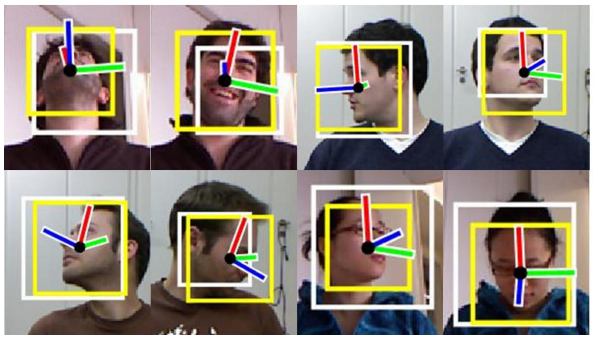
Fig. 5. Comparison results of cascading three individual DNNs and the proposed multi-task DNN. (a) Results on validation set consisting of *48 × 48 px* grayscale patches for each task (b) Results of head pose estimation.

Table 1. Comparison results of Fanelli et al.'s method[9] and ours on head pose estimation

	Mean err ± std dev. (°)		
	Roll	Pitch	Yaw
Fanelli stride 5	5.4 ± 6.0	3.5 ± 5.8	3.8 ± 6.5
Fanelli stride 10	5.5 ± 6.2	3.6 ± 6.0	4.0 ± 7.1
Fanelli stride 15	5.5 ± 6.2	3.8 ± 6.4	4.2 ± 7.8
Ours w/o BB reg.	4.8 ± 4.9	5.4 ± 4.8	5.7 ± 5.5
Ours with BB reg.	**3.9 ± 4.1**	**4.4 ± 4.0**	**4.0 ± 4.1**

1에 표현되어있는 제안된 알고리즘의 순서를 따라서 실험한 결과이다. 개별적으로 학습한 단독 DNN들을 연결한 방식보다 다중 작업 방식의 머리 자세 평균 오차 값이 약 22% 정도로 성능이 향상되었다. Fig. 5(a)와 비교하여 보면, 다중 작업 DNN 방식이 특징이 개별 작업의 훈련 데이터에 편향(biased)되지 않아 전체 시스템의 성능이 향상되었음을 알 수 있다.

우리의 방법을 식별형 접근법의 최신기술인 Fanelli 등[9]이 제안한 방법과 비교하였다. 그들은 깊이 센서 Kinect와 함께 random forest regression을 사용하여 머리 자세를 추정하였다. Table 1은 머리 자세의 평균 오차와 표준편차에 대한 비교 결과를 보여준다. 평균 오차와 표준편차 모두를 고려해 볼 때, 본 연구에서 제안한 방법의 성능이 최신기술보다 우수하다. [9]의 방법은 랜덤으로 추출된 패치들의 내부 깊이 값(internal depth value)을



(a) Head pose estimation results on BKHP dataset

(b) Head pose estimation results on web photos

Fig. 6. Multi-view face detection and head pose estimation results from the proposed multi-task DNN

특징으로 삼고 비교하여 머리 자세를 보팅(voting)하는 반면에, 우리의 DNN 기반의 접근법은 수많은 훈련 영상들로부터 자동으로 학습된 필터들을 사용한다. 그 결과 중요한 상위레벨 정보들(눈, 코, 턱 등의 상대적인 위치)을 암시적으로 추출하게 된다. 뿐만 아니라 적외선을 사

용하는 깊이 센서들은 낮에 실외환경에서는 사용할 수가 없는 반면, 우리의 방법은 낮과 밤 구분 없이 gray-scale 영상만으로 동작이 가능하다.

몇몇의 추정 결과들을 Fig. 6에 나타내었다. 우리는 roll, pitch, yaw 각의 표현에 전통적인 오른손 Cartesian 좌표계를 사용한다. 정의에 따라 roll과 pitch는 각각 x축(파란색)과 y축(녹색)에 대하여 시계방향으로 회전하는 각을 의미하며, yaw는 z축(빨강색)에 대하여 시계반대방향으로 회전하는 각을 나타낸다. 흰 박스는 초기 바운딩 박스이며 노란색 박스는 바운딩 박스 회귀 결과이다. 3.2섹션과 Fig. 1에서 설명하고 표현한 히트맵과 NMS를 방법을 통해 영상 전체에 있는 얼굴들이 동시에 검출된다. 따라서 다중객체의 처리도 단일객체의 검출 및 자세추정과 거의 같은 처리속도를 보인다. 다양한 표정과 자세에서도 잘 동작하는 것을 볼 수 있다.

우리의 방법으로 얼굴 검출 및 머리 자세 추정을 하는 실험 구현의 구체적 사양은 다음과 같다: 슬라이딩 윈도우의 크기는 3가지 스케일로 하였고, 건너뜀 값은 슬라이딩 윈도우 크기의 1/4크기로 하여 320 × 240 해상도 영상의 경우에 매 프레임당 총 206개의 윈도우들이 추출되어 약 40 fps의 처리속도로 실시간 적용이 가능하다. 테스트 시스템 환경은 3.6 GHz Intel Core i7 CPU와 Nvidia™ GeForce GTX 1080 GPU 이다. 슬라이딩 윈도우 크기의 1/4크기의 건너뜀 값은 다른 연구들에 비해 비교적 큰 값인데, 큰 건너뜀 값을 사용 할 지라도 바운딩 박스 회귀 단계 덕분에 적절한 바운딩 박스 위치와 크기가 추정된다. 따라서 비교적 적은 연산량만으로도 준수한 성능을 얻게 된다.

5. 결 론

본 논문에서 우리는 실시간으로 사용자의 머리 자세를 추정할 수 있는 다중 작업 깊은 신경망(multi-task deep neural network) 프레임웍을 제안하였다. 제안된 시스템은 단독 DNN 기반의 방법론과 비교하여 다중 작업 학습 기반의 방법론이 다양한 데이터들에 대해 과정합되지 않고 더 나은 정확도를 보였다. 또한 제안된 방법을 최신기술[9]과 성능을 비교하여 우월성을 입증했다.

향후에는 본 연구를 기반으로 몇 가지 추가 연구를 계획하고 있다. 본 논문에서는 깊은 신경망의 학습을 위하여 실제 촬영한 이미지 데이터만 사용하였는데, 성능 개선과 다양한 응용에 손쉬운 적용을 위하여 합성(synthetic) 데이터를 활용하는 딥러닝 구조를 연구할 것이다. 또한 사용자의 얼굴 검출 및 머리 자세뿐만 아니라 동일한 특징을 사용할 것이라고 예상되는 감정인식 같은 얼굴 관련 응용들을 추가하여 보다 다양한 작업들을 위한 보다 일반적인 특징 학습을 위한 깊은 신경망 구조를 연구 할 것이다.

References

[1] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, “Active shape models-their training and application,” *Computer Vision and Image Understanding*, vol. 61, pp. 38-59, 1995.

[2] T.F. Cootes, G. Edwards, and C. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681-685, 2001.

[3] F.D. la Torre, W.S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, “Intraface,” *in IEEE International Conference on Automatic Face and Gesture Recognition*, vol. 1, pp. 1-8, 2015.

[4] F. Vicente, Z. Huang, X. Xiong, F.D. la Torre, W. Zhang, and D. Levi, “Driver gaze tracking and eyes off the road detection system,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2014-2027, 2015.

[5] V. N. Balasubramanian, J. Ye, and S. Panchanathan, “Biased manifold embedding: a framework for person-independent head pose estimation,” *in IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[6] J. Foytik and V.K. Asari, “A two-layer framework for piecewise linear manifold-based head pose estimation,” *International Journal of Computer Vision*, vol. 101, pp. 270-287, 2013.

[7] X. Zhu and D. Ramanan, “Face detection, pose estimation and landmark localization in the wild,” *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2879-2886, 2012.

[8] M.D. Breitenstein, D. Kuettel, T. Weise, and L. van Gool, “Real-time face pose estimation from single

range images," *in IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[9] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L.V. Gool, "Random forests for real time 3d face analysis," *International Journal of Computer Vision*, vol. 101, pp. 437-458, 2013.

[10] Y. Lecun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541-551, 1989.

[11] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3476-3483, 2013.

[12] B. Ahn, J. Park, and I.S. Kweon, "Real-time head orientation from a monocular camera using deep neural network," *in Asian Conference on Computer Vision*, pp. 82-96, 2014.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014.

[14] M. Koestinger, P. Wohlhart, P.M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," *in IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[15] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5325-5334, 2015.

안 병 태

2007 금오공과대학교 전자공학부(학사)

2011 성균관대학교 바이오메카트로닉스 전공(석사)

2012~현재 KAIST 로봇공학학제전공 박사과정

관심분야: HRI, Deep Learning

최 동 걸

2005 한양대학교 전자컴퓨터공학부(학사)

2007 한양대학교 전자전기제어계측공학과(석사)

2016 KAIST 로봇공학학제전공(박사)

2016~현재 KAIST 정보전자연구소 박사 후 연구원

관심분야: Sensor fusion, Autonomous robotics, AI

권 인 소

1981 서울대학교 기계설계학과(학사)

1983 서울대학교 기계설계학과(석사)

1990 Carnegie Mellon Univ. Robotic Institute(박사)

1991~1992 일본 도시바 중앙연구소 연구원

1992~현재 KAIST 전기 및 전자공학과 교수

관심분야: Computer Vision