깊은 신경망을 통한 사용자 머리 자세 추정

안병태, 권인소
한국과학기술원

Head Orientation Estimation using Deep Neural Networks

Ahn Byungtae, Kweon In So
Korea Advanced Institute of Science and Technology
e-mail: btahn@rcv.kaist.ac.kr, iskweon77@kaist.ac.kr

요 약

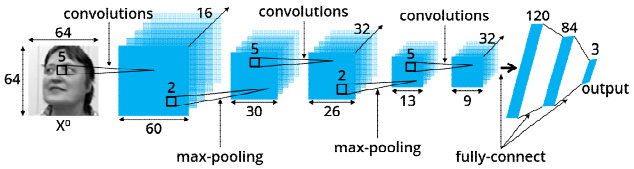
사람의 머리 자세 추정에 문제는 HRI 분야에서 중요한 문제이다. 기존의 대표적인 머리 자세 추정방법은 3차원 얼굴 모델을 정합하여 추정하는 방법과, 적외선 깊이 센서를 이용한 방법 등이 있다. 하지만 3차원 얼굴모델을 활용하는 방법은 큰 자세의 변화, 조명변화, 표정변화, 가려짐, 입력 영상의 해상도등에 많은 영향을 받는 제한적인 단점이 있고, 적외선 깊이 센서를 쓰는 방법은 특정 센서가 필요하며, 실외환경에서는 활용이 불가능하다. 본 논문에서는 깊은 신경망을 활용하여 영상정보만으로 머리 자세를 추정하는 방법론을 제안하며, Biwi Kinect Head Pose Database로 실험한 결과 영상정보만을 활용하여 평균오차 4.8°정도의 최신기술에 필적하는 정확도를 보여주었다.

1. 서론 및 관련연구

로봇과 사람의 소통을 위해서 사람의 행동이나 생태의 인지에 관한 연구는 오랫동안 이어져왔다. 그 중, 사람의 시선검출, 얼굴 인식, 동작인식 게임, 운전자의 졸음 검출 등의 응용에서 사람의 머리 자세 추정은 아주 중요한 요소기술이다. 컴퓨터비전 분야에서, 머리 자세 추정은 입력영상의 머리모양에서 방향벡터를 추론하는 일련의 과정인데, 대표적으로 3차원 얼굴 모델을 정합하여 추정하는 방법론과[1], 깊이 센서를 사용한 방법론[2] 등이 있다. 더 자세한 내용은 [3]에 나와 있다.

3차원 얼굴 모델 정합을 통한 머리 자세 추정은 [1], 얼굴과 머리의 특징점의 실제 3차원 위치정보를 가지고 학습을 통해 3차원 통계적 모델을 생성하고, 입력영상에서 비용함수가 최소로 되는 점들을 찾아서 모델을 정합하는 방식이다. 3차원 모델을 가지고 있기 때문에 머리 자세 추정 뿐만 아니라 얼굴 특징점의 위치 정보도 얻을 수 있는 장점이 있지만, 큰 자세의 변화나, 조명변화, 표정변화, 가려짐, 입력 영상의 해상도 등에 많은 영향을 받는 단점이 있다. 최근에 많이 사용되고 있는 깊이 센서를 이용한 머리 자세 추정 방법은[2], 실시간으로 비교적 다양한 자세변화에 강인한 성능을 보이지만, 특별한 깊이 센서가 있어야하며, 적외선을 사용하기 때문에 실외환경에서는 활용이 불가능 하다는 단점이 있다.

본 연구에서는, 최근 많은 발전을 이룩한 깊은 학습 방법을 이용하여 머리 자세를 추정한다. 특별히



[그림 1] DNN structure for head orientation estimation

Deep Neural Networks (DNN)를 사용하여 특정 센서의 필요없이 입력 영상으로부터 실시간으로 자세, 조명, 표정, 가려짐, 저해상도에 비교적 강인한 머리 자세 추정 방법을 제안한다. Convolutional Neural Networks (CNN)는[4] 다층 퍼셉트론과 같은 간단한 신경망의 모든 문제들을 풀기위해 제안되었다. CNN은 입력 영상으로부터 유의미한 속성들을 뽑아내는 전방향(feed-forward) 네트웍이다. CNN은 국부적 필터 적용, 가중치 공유, 그리고 sub-sampling 이 세 가지 방법을 채택하므로 이동, 회전, 스케일변화와 같은 간단한 기하학적 변환에 독립적이다[4].

2. 머리 자세 추정을 위한 DNN 구조

본 연구에서 제안된 DNN의 구조를 그림 1에 나타내었다. 고안된 깊은 신경망 구조를 가지고 식(1)의 손실함수를 최소화하는 학습을 수행하였다.

$$E(X_i^0;W)=\sum_i \| Y_i - f(X_i^0;W)\|_F^2 \qquad (1)$$

여기서 i 는 훈련샘플의 인덱스이고, X^0와 Y는 각각 훈련 샘플 영상에서 추정된 머리의 각도(roll, pitch, yaw)와 실제 각도이다. W는 파라미터들 집합이다.

제안된 깊은 신경망은 3개의 컨볼루션 층을 가지고 있다. 처음 두 개의 컨볼루션 층의 뒤에는 최댓값 통합(max-pooling)층이 있고, 세 번째 컨볼루션 층의 뒤에는 전체 연결된(fully-connect) 층들이 있다. 입력영상의 크기는 64x64이고, 모든 훈련샘플들은 그레이 스케일로 히스토그램 정규화(histogram normalization)를 통해 밝기 값이 보정된 영상이다. 5x5크기의 학습된 필터에 의해서 첫 번째 컨볼루션 층은 16개의 채널을 출력하고 두 번째 세 번째 컨볼루션 층은 32개의 채널을 출력한다. *l* 번째 컨볼루션 층은 식(2)와 같이 정의된다.

$$X_u^{l+1} = \sigma\left(\sum_{u=1}^{I} W_{uv}^l \otimes X_u^l + b_v^l\right) \qquad (2)$$

여기서 W_{uv}^l와 X_u^l는 각각 필터와 영상 패치를 의미하고, u, v는 각각 입력과 출력 채널의 인덱스이다. 예를 들어, 첫 번째 층의 경우 $u=1$, $v=1\sim16$ 이다. 따라서 X_v^{l+1}는 v번째 채널의 출력인데 이것은 $l+1$층의 입력이 된다. $\sigma(x) = \max(0,x)$는 활성함수로써 교정된 선형 유닛함수(ReLU)이고, $\otimes$는 컨볼루션 연산자이다. 바이어스 벡터들은 b로 나타내었다. 식(1)은 활성함수 때문에 비선형 함수가 되는데 [5]에서처럼 통계적 기울기 하강(stochastic gradient descent; SGD)을 사용한 역전파(back-propagation) 방법으로 학습하였다.

3. 실험 및 결론

실험에 사용된 데이터셋은 Biwi Kinect Head Pose Database[2]로, 20명의 15,678장의 영상에서 머리자세에 대한 실제 값을 제공한다. DB는 ±75° pitch각, ±60°의 yaw각의 범위의 자유로운 머리 움직임 데이터가 있고, 그 중 13,500장은 학습에 사용하였고 2,178장은 테스트에 사용하였다. 그림 2에 DB중 몇 장을 나타내었다.

실험 결과를 깊이 정보와 random forests를 사용한 최신 기술인 [2]와 비교하여 표 1에 나타내었다.

[그림 2] Biwi Kinect Head Pose DB

[표 1] Mean of Euler angles estimation errors

Method	Yaw (°)	Pitch (°)	Roll (°)
[2]	4.0	3.6	5.5
proposed	3.3	8.4	2.8

최신 기술[2]와 비교하여 볼 때, yaw각과 roll각에서는 훨씬 정확한 성능을 보였고 pitch각에서는 비교적 낮은 정확도를 보여주었다. 이는 특정 깊이센서가 요구되고 또 실외에서는 사용할 수 없는 [2]에 비하여 값싼 webcam의 영상만으로도 좋은 성능을 나타내는 장점을 보여주며, pitch각의 정확도가 비교적 낮은 것은 아직 깊은 학습의 구조와 파라미터가 최적화 되지 않았기 때문이다. 향후 연구에서 컨볼루션 층의 개수, 학습될 필터의 개수, 학습률, 모멘텀 등 다양한 변수를 조정하여 최적화를 거치면 더욱 정확한 성능을 보일 것이라 기대되어진다.

본 논문은 깊은 학습을 이용한 머리 자세 추정 방법론을 제안하였다. 이를 위해 깊은 컨볼루션 신경망을 사용하여 구조를 설계하였고, 이를 사용하여 최신기술과 비교하여 필적할만한 성능을 보였다. 향후 구조를 최적화 및 개선하면 더 나은 성능을 보일 것으로 기대되며, 앞으로 머리자세 추정 분야에서 주목받는 방법론이 될 것으로 기대되어진다.

후기

이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2010-0028680).

참고문헌

[1] J. Xiao, S. Baker, I. Matthews, T. Kanade, “Real-Time Combined 2D + 3D Active Appearance Models,” CVPR, pp.535-542, 2004.

[2] G. Fanelli, M. Dantone, J. Gall, A. Jssati, L. Van Gool, “Random Forests for Real Time 3D Face Analysis,” IJCV, pp.437-458, 2013.

[3] E. Murphy-Chutorian, M. Trivedi. “Head pose estimation in computer vision: A survey,” TPAMI, vol. 31, no. 4, pp.607-626, 2009.

[4] Y. LeCun, R. Bottou, Y. Bengio, P. Haffner, “Gradient-based Learning Applied to Document Recognition,” Proc. IEEE, vol. 86, no. 11, pp.2278-2324, 1998.

[5] A. Krizhevsky, I. Sutskever, G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” NIPS, pp.1106-1114, 2012.