Deep Transfer Feature Based Convolutional Neural Forests for Head Pose Estimation

Yuanyuan Liu, Zhong Xie, Xi Gong(✉), and Fang Fang

Faculty of Information Engineering, China University of Geosciences, Wuhan 430074, China
gongxi_cug@126.com

Abstract. In real-world applications, factors such as illumination, occlusion, and poor image quality, etc. make robust head pose estimation much more challenging. In this paper, a novel deep transfer feature based on convolutional neural forest method (D-CNF) is proposed for head pose estimation. Deep transfer features are extracted from facial patches by a transfer network model, firstly. Then, a D-CNF is devised to integrate random trees with the representation learning from deep convolutional neural networks for robust head pose estimation. In the learning process, we introduce a neurally connected split function (NCSF) as the node splitting strategy in a convolutional neural tree. Experiments were conducted using public Pointing'04, BU3D-HP and CCNU-HP facial datasets. Compared to the state-of-the-art methods, the proposed method achieved much improved performance and great robustness with an average accuracy of 98.99% on BU3D-HP dataset, 95.7% on Pointing'04 and 82.46% on CCNU-HP dataset. In addition, in contrast to deep neural networks which require large-scale training data, our method performs well even when there are only a small amount of training data.

Keywords: Convolutional neural network · Random forest
Transfer network · Head pose estimation

1 Introduction

Head pose estimation is the key step in many computer vision applications, such as human computer interaction, intelligent robotics, face recognition, and recognition of visual focus of attention [14,28]. The existing techniques achieve satisfactory results in well-designed environments. In real-world applications, however, factors, such as illumination variation, occlusion, poor image quality, etc., make head pose estimation much more challenging [19,22]. Hence, we propose a deep transfer feature based convolutional neural forest (D-CNF) method to estimate head pose estimation in unconstrained environment.

S. Satoh (Ed.): PSIVT 2017, LNCS 10799, pp. 5–16, 2018.
https://doi.org/10.1007/978-3-319-92753-4_1

A general head pose estimation framework appeared in most of previous works can be divided into two major steps, one is the feature extraction and the other is classifier construction [13]. Extracting robust facial features and designing effective classifier are the two key factors in unconstrained head pose estimation. For feature extraction, based on different features, several methods for the problem can be briefly divided into two categories, facial local feature and facial global feature based methods. The former methods usually require high image resolution for facial local feature identification, such as eyes, eyebrows, nose or lips [13,27], etc. These methods can provide accurate recognition results relying on accurate detection of facial feature points and high quality images. The latter methods based on facial global feature usually use texture features from an entire face to estimate head poses [1,4,17], etc. It may be good for dealing with low resolution image but not robust to occlusion and illumination. In the real-life scene, the various illumination occlusion, low image resolution and wide scene make facial local feature extraction difficult. In order to extract robust high-level features for head pose estimation, we address the problem based on globe deep transfer feature representation.

For the head pose classifier construction, most of the traditional classifiers, such as Support vector machine (SVM), Random forest (RF), Bays classifier and convolutional neural network (CNN), together with some unsupervised learning techniques are employed in the head pose estimation [17,21,25]. Recent years, CNN and RF become popular learning algorithms for head pose estimation in some real-life applications. CNN has an ability to automatically learn high-level feature representations from raw image data [11,16,20,24,30]. CNN achieves huge success in face recognition [23] and object multi-classification [26]. However, a limit for CNN is that the learning procedure needs a large amount of datasets and GPUs [6,9,15]. RF is a popular method given their capability to handle large training datasets, high generalization power and speed, and easy implementation [2,3,5,7]. In this paper, we are interested in constructing an effective head pose classifier using a limited amount of image data with a hybrid deep convolution networks enhanced decision forest. Our method aims at improving both accuracy and efficiency. The pipeline of our proposed D-CNF is depicted in Fig. 1. The deep transfer feature is extracted by transfer CNN model to suppress the influence of illumination, occlusion, and low image resolution, firstly. Then, head poses are estimated by the trained D-CNF model.

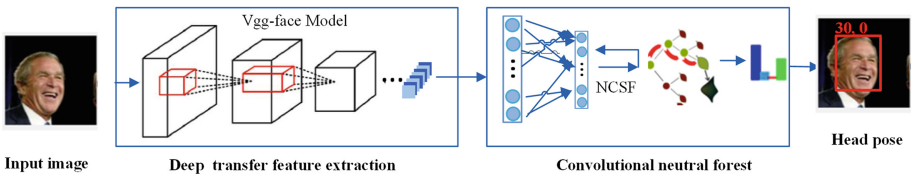


Fig. 1. The pipeline of D-CNF for head pose estimation

Our contributions include the following:

1. We propose a deep transfer feature based convolutional neural forest method (D-CNF) for head pose estimation in unconstrained environment, which unifies classification trees with the representation learning from deep convolution networks, by training them in an end-to-end way.
2. We introduce a neurally connected split function (NCSF) as new split node learning in a D-CNF tree. The D-CNF method can achieve fast and accurate recognized results in the limited amount of image data, rather than a large amount of data by CNN.
3. We propose a robust deep transfer feature representation based on a pre-trained CNN model.

The rest of this paper is organized as follows: Sect. 2 presents our D-CNF method in details. Section 3 discusses the experimental results using publicly available datasets. Section 4 concludes this paper with a summary of our method.

2 Deep Transfer Feature Based Convolution Neural Forests for Head Pose Estimation

In this section, we address the D-CNF approach for head pose estimation in unconstrained environment. First, we present robust deep feature representation based on facial patches, which can reduce the influence of various noises, such as over-fitting, illumination, low image resolution, etc. Then, we describe the framework of D-CNF training procedure for head pose estimation in details. Finally, we give the D-CNF prediction for head pose estimation in unconstrained environment.

2.1 Deep Transfer Feature Representation

We extract deep transfer feature from facial patches with a pre-trained CNN model, i.e., Vgg-face [23]. We employ the Vgg-face architecture that is pre-trained with the LFW and YTF face datasets [23] to derive deep high-level feature representation, as shown in Fig. 2. The model includes 13 convolution layers, 5 max-pooling layers, and 3 fully connected layers. The deep transfer feature is described as:

$$y^j = \max(0, \sum_i x^i w^{i,j} + b^j), \tag{1}$$

where y^j is the j^{th} output feature value of the first fully connected layer, x^i is the i^{th} feature map of the last convolution layer, $w^{i,j}$ indicates the weight between the i^{th} feature map and the j^{th} output feature value, and b^j donates the bias of the j^{th} output feature value. The deep transfer feature is used to train a two-layer network through back propagation, which can transfer the original Vgg-face feature to the pose feature.

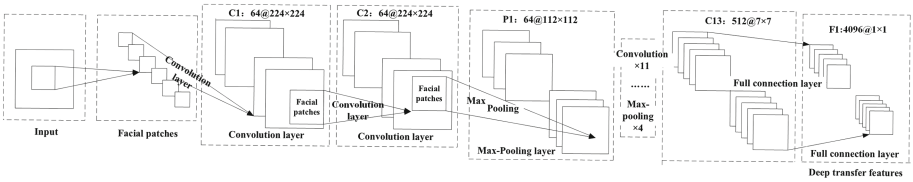


Fig. 2. The structure of pre-trained CNN network for deep feature representation. The trained network model includes 13 convolution layers, 5 max-pooling layers, and 3 full connection layers. In our work, we extract deep features from facial patches on the first connection layer.

2.2 D-CNF Training

In this paper, we propose a fast and efficient D-CNF method for robust head pose estimation on limit training sets, which is unifies classification trees with the representation learning from deep convolution networks, by training them in end-to-end way. The training of a traditional decision tree of a random forest (RF) consists in a recursive procedure, which starts from the root and iteratively builds the tree by splitting nodes [2]. The proposed D-CNF is also an ensemble of convolution neural trees, where split nodes are computed by the proposed neural connected split function (NCSF). The proposed NCSF can improve the learning capability of splitting node by deep neural learning representation, thus to improve the discrimination and efficiency of a tree. The detail training procedure is given as below.

Learning Splitting Nodes by NCSF. For facial patches, we extract a set of deep transfer features P, $P = \{P_i\}$ and $P_i = \{y^j\}$. We propose a NCSF-f_n to reinforce the learning capability of a splitting node by deep neural learning representation. Each output of f_n is brought in correspondence with a splitting node $d_n(P_i; Y)$,

$$d_n(P; Y) = \sigma(f_n(P; Y)), \tag{2}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function and Y is the decision node parametrization.

We employ a Stochastic Gradient Descent (SGD) approach to minimize the risk with respect to Y:

$$Y^{(t+1)} = Y^{(t)} - \frac{\eta}{|B|} \sum_{(P,\pi)\in B} \frac{\partial L}{\partial Y}(Y^{(t)}, \pi; P), \tag{3}$$

where $\eta > 0$ is the learning rate, π is facial expression label and B is a random subset (a.k.a. mini-batch) of samples. The gradient with respect to Y is obtained by chain rule as follows:

$$\frac{\partial L(Y, \pi; P)}{\partial Y} = \sum_{n\in N} \frac{\partial L(Y, \pi; P)}{\partial f_n(P; Y)} \cdot \frac{\partial f_n(P; Y)}{\partial Y}. \tag{4}$$

Hence, the gradient term that depends on the neutral decision tree is

$$\frac{\partial L(Y,\pi;P)}{\partial f_n(P;Y)} = -(d_n^{N_r}(P;Y) + d_n^{N_l}(P;Y)), \tag{5}$$

where given a node N in a tree and N_r and N_l denote its right and left child, respectively.

To split a node, Information Gain (IG) is maximized:

$$\tilde{\varphi} = \arg\max_{\varphi}(H(d_n) - \sum\nolimits_{S\in\{N_r,N_l\}} \frac{|d_n^S|}{|d_n|}(H(d_n))), \tag{6}$$

where $\frac{|d_n^S|}{|d_n|}, s \in \{N_r, N_l\}$ is the ratio between the number of samples in $d_n^{N_l}$ (arriving at the left child node), set $d_n^{N_r}$ (arriving at the right child node), and $H(d_n)$ is the entropy of d_n.

Learning Leaf Nodes. Create a leaf l when IG is below a predefined threshold or when a maximum depth is reached. Otherwise continue recursively for the two child nodes $d_n^{N_l}$ and $d_n^{N_r}$ at the splitting node step. For a leaf node in a conditional D-CNF tree, it stores the conditional multi-probability $p(\pi|\theta, y)$. Therefore, we simplify the distribution over head poses by a multivariate Gaussian Mixture Model (GMM) [17] as in:

$$p(\theta, l) = N(\theta; \overline{\theta}, \Sigma_l^{\theta}), \tag{7}$$

where $\overline{\theta}$ and Σ_l^{θ} are the mean and covariance of leaves' head pose probabilities, respectively.

2.3 D-CNF for Head Pose Estimation

This section provides the prediction procedure of the D-CNF for head pose estimation. Deep transfer feature patches pass through the trees in a trained D-CNF. All feature patches end in a set of leaves of the forest. In the leaves of a D-CNF forest, there are muti-probabilistic models of head poses. We simplify the distributions over multi-probabilities by adopting multivariate GMM as:

$$p(\theta|l) = N(\theta; \overline{\theta}, \Sigma_l^{\theta}), \tag{8}$$

where $\overline{\theta}$ and Σ_l^{θ} are the mean and covariance of leaves' head pose probabilities, respectively.

While Eq. 8 models the probability for a feature patch p_i ending in the leaf l of a single tree, the probability of the forest is obtained by averaging over all trees:

$$p(\theta|P) = \frac{1}{T}\sum_t p(\theta|l_t(P)) \tag{9}$$

where l_t is the corresponding leaf for the tree T_t, T is the number of trees in D-CNF.

3 Experimental Results

3.1 Datasets and Settings

To evaluate our approach, three challenging face datasets were used: Pointing'04 dataset [10], BU3D-HP dataset [31], and CCNU-HP dataset in the wide classroom [17]. These datasets were chosen since they contained unconstrained face images with poses ranging from −90° to +90°. The Pointing'04 head pose dataset is a benchmark of 2790 monocular face images of 15 people with variations of yaw and pitch angles from −90° to +90°. For every person, 2 series of 93 images (93 different poses) are available. The CCNU dataset was collected included an annotated set of 38 people with 75 different head poses from an overhead camera in the wide scene. It contains head poses spanning from −90° to 90° in horizontal direction, and −45° to 90° in vertical direction. The multi-view BU3D-HP database contains 100 people of different ethnicities, including 56 females and 44 males with variations of yaw angles from −90° to +90°.

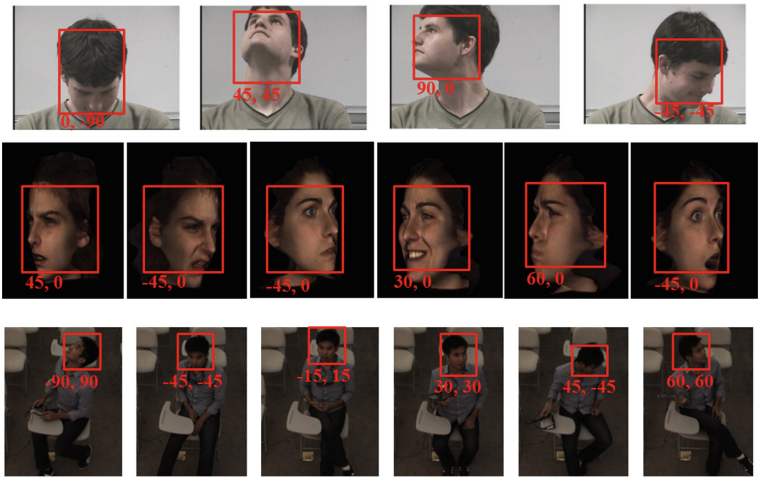


Fig. 3. The examples of head pose estimation on Pointing'04, BU3D-HP and CCNU-HP datasets. Top row: results of Pointing'04. Middle row: results of the BU3D-HP dataset. Bottom row: results of CCNU-HP dataset.

The examples of head pose estimation on Pointing'04, BU3D-HP and CCNU-HP datasets are shown in Fig. 3. The D-CNF method can achieve fast and accurate recognized results in limited amount of image data, rather than a large amount of data by CNN. Our method was trained with 2000 images from Pointing'04 dataset, 15498 images from BU3D-HP dataset and 2121 images from CCNU-HP dataset. In evaluation, we used 870 images from Pointing'04 dataset, 5166 images from BU3D-HP dataset and 707 images from CCNU-HP dataset. The experiments were conducted in a PC with Intel(R) Core(TM) i7-6700 CPU@

4.00 GHz, RAM 32 GB, NVIDIA GeForce GTX 1080 (2). We use the Caffe framework [12] for the transfer CNN and deep feature representation.

3.2 Experiments on Pointing'04 Datasets

Figure 4 shows the head poses estimation results on Pointing'04 datasets in yaw and pitch rotations, respectively. The average accuracy on 9 yaw head poses and 9 pitch head poses is 95.6%. As it is shown, the highest accuracy is 98.4% of 90° in the yaw rotation. The lowest accuracy is 92.6% of −45° in the pitch rotation, due to more occlusion in a face area.

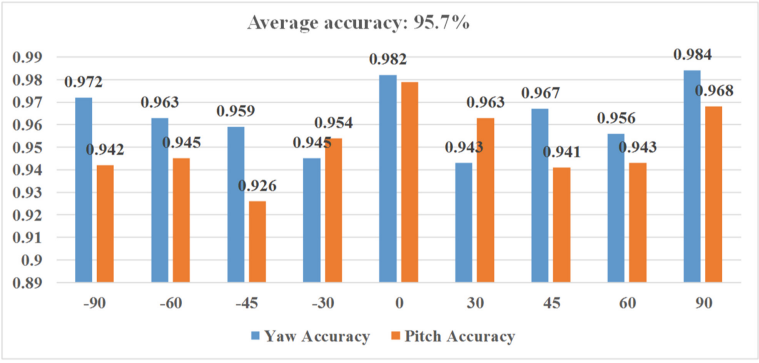


Fig. 4. Head pose estimation on Pointing'04 datasets in the yaw and pitch rotations

In comparison with the state-of-the-art head pose estimation methods, we conducted experiments using the MSHF [18], Multivariates label distribution (MLD-wj) [29], CNN(6convs+2fc) [15], multi-class SVM (M-SVM) [22] and HF [8] on Pointing'04 head pose dataset. The same training and testing datasets were used, and we employed a 4-fold cross-validation. Table 1 lists the average accuracy and error across using these methods. MLD-wj [29], CNN [15] and HF [8] yielded comparable results with an accuracy of approximately 70% in yaw and pitch rotations. MLD-wj [29] proposed to associate a multivariate label distribution to each image for head pose estimation in yaw and pitch rotations. MSHF [18] proposed a hybrid structure hough forest to 25 class head pose estimation and achieved the second highest accuracy of 84%. HF [8] improved random forests with Hough voting for real-time head pose estimation. M-SVM [22] produced similar accuracy in the range of 60%. Our proposed D-CNF exhibited the best performance with the accuracy of 95.7% in yaw and pitch rotations. In addition, the standard deviation of D-CNF indicates that D-CNF achieved the greatest consistency with a smallest STD. It is evidential that our D-CNF improved the head pose estimation with great robustness.

Table 1. Accuracy (%) and average error (in degrees) using different methods on Pointing'04 dataset.

Methods	Yaw	Pitch	Yaw + Pitch	STD
MSHF [18]	92.3	90.7	84.0	3.5
MLD-wj [29]	84.30	86.24	72.3	4.9
CNN [15]	83.52	86.94	71.83	5.5
HF [8]	82.3	84.86	70.54	5.2
SVM [22]	80.6	82.5	60.46	5.7
D-CNF	99.05	94.36	95.7	0.8

3.3 Experiments on Multi-view BU3D-HP Dataset

Each image in the BU3D-HP dataset is automatically annotated with one out of the nine head pose labels ({−90°, −60°, −45°, −30°, 0°, +30°, +60°, +75°, 90°}). We train a D-CNF of 50 neural trees using 15498 head pose images. Figure 5 shows the confusion matrix of head pose estimation on BU3D-HP dataset. The D-CNF estimated 9 head pose classes in the horizontal direction and achieved the average accuracy of 98.99%. Examples of the estimated head pose are shown in Fig. 3.

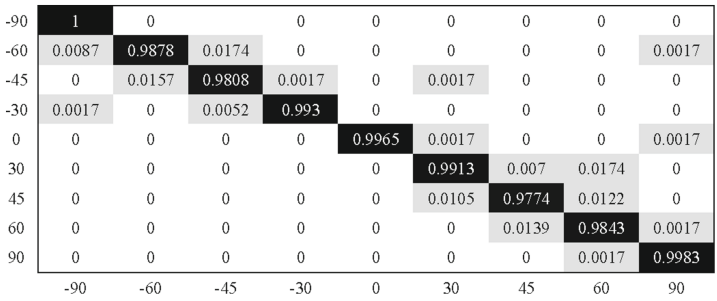


Fig. 5. Confusion matrix of head pose estimation on BU3D-HP dataset.

The average accuracy of our D-CNF method is compared with that of CNN, Zheng GSRRR [33], and SIFT + CNN [32] in Table 2. The CNN in this experiment contains three convolution layers followed by three max-pooling layers and two fully connected layers. Each filter is of size 5×5 and there are 32, 64, and 128 such filters in the first three layers, respectively. The input images are rescaled to 224 by 224.

The accuracy of the CNN on BU3D-HP dataset is 69.61% as presented in Table 2. The accuracies achieved with SIFT using algorithms proposed in [32, 33] are 87.36% and 92.26%, respectively. Our method achieves 98.99% which is

Table 2. Accuracy (%) and STD using different methods on multi-view BU3D-HP dataset.

Methods	Features	Poses	Accuracy	STD.
CNN	Image	9	69.61	0.9
Zheng GSRRR [33]	Sparse SIFT	9	87.36	0.8
SIFT + CNN [32]	SIFT	9	92.26	0.7
D-CNF	Deep transfer feature	9	98.99	0.5

competitive to the methods above. The lowest STD. of 0.5% using our method also proved the robustness of the proposed D-CNF.

3.4 Experiments on CCNU-HP Dataset in the Wide Scene

In this case, we evaluated the proposed D-CNF on CCNU-HP dataset in the wide scene. For evaluation, a 4-fold cross-validation was conducted. In our experiments, we annotate the dataset into 5 classes in the yaw rotation as Fig. 6(a) and 4 classes in the pitch rotation as Fig. 6(b). The final classified classes are 20 categories in the wide scene dataset.

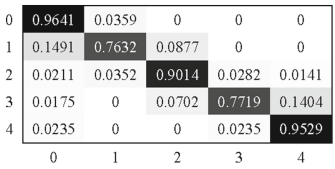
(a) The annotation categories of the yaw angels. The first row are the yaw angles in the dataset and the second row are the annotation class,

(b) The annotation categories of the pitch angels. The first row are the pitch angles in the dataset and the second row are the annotation class.

Fig. 6. The annotation categories of the yaw and pitch angels in the experiments. (a) The annotation classes in the yaw rotation, (b) The annotation classes in the yaw rotation.

Figure 7 shows the confusion matrixs of head pose estimation on CCNU-HP dataset in the yaw and pitch rotations, respectively. The D-CNF achieved the average accuracy of 88.54% in the yaw rotation and 76.38% in the more challenging pitch rotation. Examples of the estimated head pose are shown in Fig. 3.

Table 3 lists the average accuracy and error across on more challenging CCNU-HP datasets using four state-of-the-art methods. The average accuracy of the CNN on CCNU-HP dataset is 59.52% as presented in Table 3. The second highest accuracy is achieved 77.9% with combined features using D-RF method. Our method achieves 82.46% which is competitive to the methods above.



(a) The confusion matrix of yaw angles (b) The confusion matrix of pitch angles

Fig. 7. Confusion matrixs of head pose estimation on CCNU-HP dataset. (a) The matrix of yaw angles, (b) The matrix of pitch angles.

Table 3. Accuracy (%) using different methods on CCNU-HP dataset.

Methods	Features	Yaw	Pitch	Yaw + Pitch
CNN	Image	65.25	53.79	59.52
Gabor + RF	Gabor	75.42	67.57	71.5
D-RF [17]	Combined features	85.6	70.19	77.90
D-CNF	Deep transfer feature	88.54	76.38	82.46

4 Conclusion

This paper described a novel deep transfer feature based convolutional neural enhanced forests (D-CNF) method for head pose estimation in unconstrained environment. In this method, robust deep transfer features are extracted from facial patches using transfer CNN model, firstly. Then, the D-CNF integrates random trees with the representation learning from deep convolutional neural networks for head pose estimation. Besides, a neural connected split function (NCSF) is introduced to D-CNF to split node learning. Finally, a prediction procedure of the trained D-CNF can classify head pose in unconstrained environment. Our method can perform well in limit number of datasets owing to transferring pre-trained CNN to fast decision node splitting in a Random Forest. The experiments demonstrate that our method has remarkable robustness and efficiency.

Experiments were conducted using public Pointing'04, BU3D-HP and CCNU-HP datasets. Our results demonstrated that the proposed deep feature outperformed the other popular image features. Compared to the state-of-the-art methods, the proposed D-CNF achieved improved performance and great robustness with an average accuracy of 98.99% on BU3D-HP dataset, 95.7% on Pointing'04 dataset, and 82.46% on CCNU-HP dataset. The average time for performing a head pose estimation is about 113 ms.

Compared to CNN method from popular deep learning, our method achieved the greatest performance on limited number of datasets. In future, we plan to investigate on-line learning methods to achieve real-time estimation by integrating head movement tracking.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61602429), China Postdoctoral Science Foundation (No. 2016M592406), and Research Funds of CUG from the Colleges Basic Research and Operation of MOE (No. 26420160055).

References

1. Ahn, B., Park, J., Kweon, I.S.: Real-time head orientation from a monocular camera using deep neural network. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 82–96. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16811-1_6
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
3. Bulo, S.R., Kontschieder, P.: Neural decision forests for semantic image labeling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 81–88 (2014)
4. Chu, X., Ouyang, W., Li, H., Wang, X.: Structured feature learning for pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4715–4723 (2016)
5. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2578–2585. IEEE (2012)
6. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: a deep convolutional activation feature for generic visual recognition. In: ICML, vol. 32, 647–655 (2014)
7. Fanelli, G., Yao, A., Noel, P.-L., Gall, J., Van Gool, L.: Hough forest-based facial expression recognition from video sequences. In: Kutulakos, K.N. (ed.) ECCV 2010. LNCS, vol. 6553, pp. 195–206. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35749-7_15
8. García-Montero, M., Redondo-Cabrera, C., López-Sastre, R., Tuytelaars, T.: Fast head pose estimation for human-computer interaction. In: Paredes, R., Cardoso, J.S., Pardo, X.M. (eds.) IbPRIA 2015. LNCS, vol. 9117, pp. 101–110. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19390-8_12
9. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
10. Gourier, N., Hall, D., Crowley, J.: Estimating face orientation from robust detection of salient facial features in pointing. In: International Conference on Pattern Recognition Workshop on Visual Observation of Deictic Gestures, pp. 1379–1382 (2004)
11. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 34–50. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_3
12. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Proceedings of the 22nd ACM International Conference on Multimedia
13. Wu, J., Trivedi, M.M.: A two-stage head pose estimation framework and evaluation. Pattern Recogn. **41**, 1138–1158 (2008)
14. Kim, H., Sohn, M., Kim, D., Lee, S.: Kernel locality-constrained sparse coding for head pose estimation. IET Comput. Vis. **10**(8), 828–835 (2016)

15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
16. Liu, X., Liang, W., Wang, Y., Li, S., Pei, M.: 3D head pose estimation with convolutional neural network trained on synthetic images. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 1289–1293. IEEE (2016)
17. Liu, Y., Chen, J., Shu, Z., Luo, Z., Liu, L., Zhang, K.: Robust head pose estimation using dirichlet-tree distribution enhanced random forests. Neurocomputing **173**, 42–53 (2016)
18. Liu, Y., Xie, Z., Yuan, X., Chen, J., Song, W.: Multi-level structured hybrid forest for joint head detection and pose estimation. Neurocomputing **266**, 206–215 (2017)
19. Ma, B., Li, A., Chai, X., Shan, S.: CovGa: a novel descriptor based on symmetry of regions for head pose estimation. Neurocomputing **143**, 97–108 (2014)
20. Mukherjee, S.S., Robertson, N.M.: Deep head pose: gaze-direction estimation in multimodal video. IEEE Trans. Multimedia **17**(11), 2094–2107 (2015)
21. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 607–626 (2009)
22. Orozco, J., Gong, S., Xiang, T.: Head pose classification in crowded scenes. In: British Machine Vision Conference, London, UK, pp. 1–3, 7–10 September 2009
23. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC, vol. 1, p. 6 (2015)
24. Patacchiola, M., Cangelosi, A.: Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. Pattern Recogn. **71**, 132–143 (2017)
25. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. arXiv preprint arXiv:1603.01249 (2016)
26. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: XNOR-Net: imagenet classification using binary convolutional neural networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 525–542. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_32
27. Schwarz, A., Lin, Z., Stiefelhagen, R.: HeHOP: highly efficient head orientation and position estimation. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8. IEEE (2016)
28. Wu, S., Kan, M., He, Z., Shan, S., Chen, X.: Funnel-structured cascade for multi-view face detection with alignment-awareness. Neurocomputing **221**, 138–145 (2017)
29. Xin, G., Xia, Y.: Head pose estimation based on multivariate label distribution. In: IEEE Conference on Computer Vision and Pattern Recognition, Ohio, USA, pp. 1837–1842, 24–27 June 2014
30. Xu, X., Kakadiaris, I.A.: Joint head pose estimation and face alignment framework using global and local CNN features. In: Proceedings of the 12th IEEE Conference on Automatic Face and Gesture Recognition, Washington, DC, vol. 2 (2017)
31. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: 2006 7th International Conference on Automatic Face and Gesture Recognition, FGR 2006, pp. 211–216. IEEE (2006)
32. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., Yan, K.: A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans. Multimedia **18**(12), 2528–2536 (2016)
33. Zheng, W.: Multi-view facial expression recognition based on group sparse reduced-rank regression. IEEE Trans. Affect. Comput. **5**(1), 71–85 (2014)