IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.
www.ijiset.com
ISSN 2348 – 7968

Head Pose Estimation Using Convolutional Neural Networks

Neeta Malagavi[1], Vidya Hemadri[2],Dr. U. P. Kulkarni[3]

[1]Student, Departmentof Computer Science Engineering, SDMCET, Dharwad, India.
[2]Assistant Professor, Department of Computer Science Engineering, SDMCET, Dharwad, India.
[3]Professor, Departmentof Computer Science Engineering, SDMCET, Dharwad, India.

Abstract

Detection and estimation of head pose is fundamental problem in many applications such as automatic face recognition, intelligent surveillance, and perceptual human-computer interface and in an application like driving, the pose of the driver is used to estimate his gaze and alertness, where faces in the images are non-frontal with various poses. In this work head pose of the person is used to detect the alertness. Convolution neural networks are used to train and classify various head poses such as frontal, left, right, up and down. The system locates the distinctive features such as nose and eyes with a higher accuracy and provides robust estimation of the head pose.

Keywords—Face detection; Convolutional neural networks; Head pose estimation, Recognition

1. Introduction

Successful interaction with humans requires robust and accurate perception and tracking of their body parts to infer nonverbal signals of attention and intention. In order to establish a joint focus of attention, estimating the head pose is crucial since it usually coincides with the gaze direction. Furthermore, head pose estimation is also essential for analyzing complex meaningful gestures such as pointing gestures or head nodding and shaking.In computer vision content, head pose estimation is defined as the computation of the orientation of a human head relative to the camera. Head pose estimation playsthe important role in determining the accuracy of the face recognition system. Estimation of the pose angles is one of thetechniques in recognizing the non-frontal face image.

Most of the face recognition techniques are sensitive to pose variation. Face recognition across pose is very important in many applications especially dealing with uncooperative faces. Uncooperative face means the face image captured by the camera is without the notice of the subject. So therefore, most of the images which are captured are non-frontal face images.

This paper is organized as follows. The section IIreviews related work and convolutional neural networks. Section III presents the system design. Section IV presents system implementation. Finally section V defines experimental results.

2.Related Work

In [1] author presents a new Head pose estimation system that is capable of online operation in a car. This approach uses soft histograms of location specific gradient orientation as the input to a support vector regressor for each degree-of-freedom. It compares this method to a system that uses support vector regression applied to a feature vector created by principal component analysis of raw image gradients. The subcomponents of HPES are that the facial regions are found by three cascaded-Adaboost face detectors applied to the grayscale video images. The detected facial region is scale-normalized to fixed size and used to compute a Localized Gradient Orientation (LGO) histogram. The histogram is passed to two Support Vector Regressors (SVRs) trained for head pitch and yaw.The limitations are it evaluates and overcomes the difficulties that arise withlighting conditions in a moving car. The system outperforms the other approaches in absolute yaw error by a significant margin: 9.28 degreecompared to 14.90 degree and 12.19 degree during the daytime experiment, and 7.74 degree compared to 16.49 degree and 13.11 degree during the nighttime drives.

In [2] author proposes head pose estimation is based on Gabor eigenspace modeling. Gabor filter is used to enhance pose information and eliminate other distractive information like various face appearance or change in the environmental illumination. It discusses the selection of optimal Gabor filter's orientation to each of the pose, which will lead to more compact pose clustering. Then the distribution based pose model (DBPM) is used to model the pose cluster in Gabor eigenspace. Thus to each pose cluster, a 2Dimentional distance space is established where the distance from centroid could be used to estimate the head pose. The experimental results demonstrate the algorithm'srobustness and generalization. The study also tries algorithm on real scene sequences to detect human face and estimate its pose. In this way, the user can control an intelligent wheelchair just by his head poses.

In [3] author presents a geometrical approach based on the symmetrical properties of the face to achieve head pose estimation address the problem of head pose estimation from digital images which consists of locating a person's head and estimating the orientation of its three degrees of freedom (Yaw, Pitch and Roll). The approach selects a set of features from the symmetrical parts of the face and the size of the bilateral symmetrical area of the face is a good indicator of the Yaw head pose. The approach trains a Decision Tree model in order to recognize head pose with regard to the symmetrical areas. These approaches do not need the location of interest points on face and is robust to partial occlusion. The tests were performed on a different dataset from that used for training the model and the results demonstrate that the change in the size of the regions that contain a bilateral symmetry provides accurate pose estimation.

ISSN 2348 – 7968

In [4] author proposes estimation of the head pose is an important capability of a robot when interacting with humans since the head pose usually indicates the focus of attention. The approach proceeds in 3 stages. Firstly, a face detector roughly classifies the pose as frontal, left, or right profile. Then, the classifiers will train withAdaboost using Haar-like features; detect distinctive facial features such as nose tip and the eyes. Based on the positions of these facial features, the neural network finally estimates the 3 continuous angles of rotation to use the model the head pose. The approach is computationally highly efficient. The experiments are shown with the standard databases as well as with real-time images, the proposed system locates the distinctive features with a higher accuracy and provides robust estimates of the head pose.

In [5] author presents the head detection module consists of an array of Haar-wavelet Adaboost cascades. The pose of the head initially estimates the module which employs localized gradient orientation (LGO) histograms as input to support vector regressors (SVRs). The tracking module provides a fine estimate of the 3-D motion of the head using a new appearance-based particle filter for 3-D model tracking in an augmented reality environment. The implementation utilizes OpenGL-optimized graphics hardware to work efficiently and compute the samples in real time. To demonstrate the suitability of this system for real driving situations, it providesa comprehensive evaluation made with drivers of varying ages, race, and sex daytime and nighttime conditions (light illumination). To measure the accuracy of system, it compares the estimation results to a marker-based cinematic motion-capture system which is installed in the automotive test bed.

In [6] author presents Viola and Jones method to detect facial features inside the image, it works accurately and rapidly. This technique is used to detect facial features. Face detection procedure classifies images based on the value ofsimple features. The successful detection of facial features which works accurately and the next objective goal is to research the more extract details like eyes nose mouth etc. A face-detection algorithm focuses on the detection of frontal human faces. It analyzes image detection in which the image of a person is matched bit by bit. Image matches with the image and stores in database. If any of the facial feature changes in the database,that will invalidate the matching process.

In [7] author presents a hybrid system for convolutional neural network and logistic regression classifier (LRC) are combined. The CNN is used to train and detect and recognize face images. The logistic regression classifier is used to classify thefeatures learned by the convolutional network. Applying feature extraction using CNN to normalized data causes the system to cope with faces subject to pose and lighting variations. The Logistic Regression Classifier, which is a discriminative classifier, is used to classify the extracted features of face images. The architecture was tested by training the network to recognize face. The Images are normalized to lie in the range -1 to 1 by removing the mean value and dividing them by their standard deviation. Network can have only one input for image (e.g. no stereo images simultaneously).have to set connection matrix after the initialization.

The limitation of the previous works focuses of the weighted average of correct classification which is equal to 81.4%. The classification rate for class 3 which corresponds to the two poses: -15 degree and 15degree, is the lowest, because often these poses are closer to that of class 2 and class 4, the other poses are better classified, and their classification rate exceeds 80%.The proposed system works with all the adverse conditions like wearing spectacles and without wearing spectacles, during daytime and night time, with different hair styles etc. The classification rate exceeds 85%. The previous works the poses were not classified in the adverse conditions like wearing spectacles. The classification accuracy is more efficient.

2.1 Proposed system

The head pose estimation system consists of three major steps namely

- Image Acquisition
- Face Detection
- Head Pose Estimation

Image Acquisition

The major step of head pose estimation system is image acquisition; the images are collected from recorded videos clips. Acquisition is the pre-processing stage in which the video clips are converted into frames. Around 100 videos were collected of different facial attributes like with spectacles, without spectacles, different hair style conditions in different adverse conditions over timeline.

Face Detection

The Viola Jones algorithm [5] is used to detect facial features. It is one of the efficient face detector algorithms. The Viola-Jones algorithm uses Haar-like features, that is, a scalar product between the image and some Haar-like templates. The detected face is marked with the rectangle box as shown in the fig 1

Head Pose Estimation

The different head pose profiles namely right, left, straight, up and down is classified by using convolutional neural networks. The input images are taken from the database and they aretrained and tested by CNN. The samples of the head pose images are shown in the fig 2

ISSN 2348 – 7968

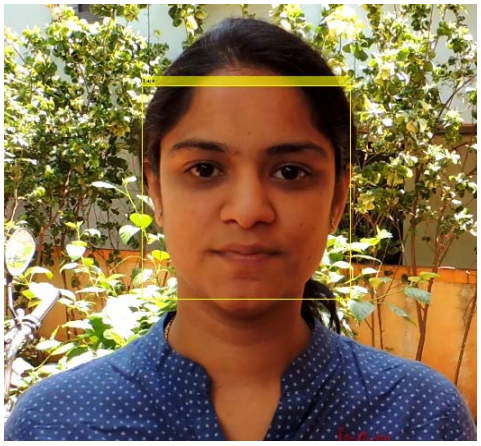

Fig 1: Detected Face

Fig 2 Samples of head pose images

2.2. Convolutional Neural Networks

Convolutional Neural Networks is a feed-forward network with the ability of extracting topological properties from the input image [6]. It extracts features from the raw data image and then a classifier classifies the extracted features. CNNs are invariance to distortions and simple geometric transformations like translation, image scaling, image rotation and compressing.

A convolutional layer is used to extract features from local receptive fields in the preceding layer. To extract different types of local features, a Convolutional Layer is organized in planes of neurons called feature maps which are responsible to detect a specific feature. In a network with a 5×5 convolution kernel each unit has 25 inputs connected to a 5×5 area in the previous layer, it is the local receptive field. A trainable weight is assigned to each connection, but all units of one feature map share the same weights. The feature which allows reducing the number of trainable parameters is called weight sharing technique and is applied in all CNN layers [7]. A reduction of the resolution of the feature maps is performed through the sub sampling layers. In a network with a 2×2 sub sampling filter such a layer comprises as manyfeature map numbers as the previous convolutional layer but with half the number of rows and columns.

- The major advantage of convolutional networks is the use of shared weight in convolutional layers, which means that the same filter (weights bank) is used for each pixel in the layer; this will both reduce required memory size and improves performance.
- Neural networks are very simple to implement.
- Neural networks cannot be retrained. If you add data later, it impossible to add to an existing network.

The general architecture of the system is as shown in fig 3

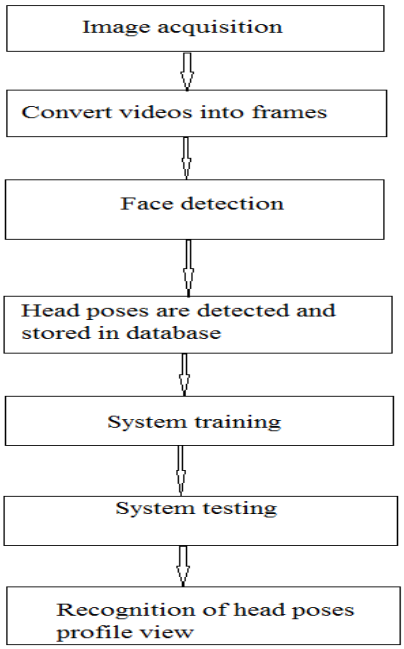

Fig 3: General Architecture

3.System Modeling and Design

3.1 CNN Structure

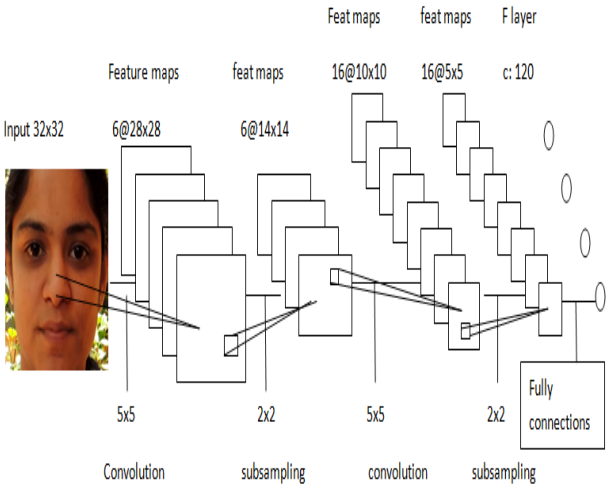

Fig 4: CNN Structure

ISSN 2348 – 7968

Table 1 Dimensions of the convolutional network

LAYER	TYPE	x	y	Receptive field of x	Receptive field of y
1	convolution	28	28	5	5
2	Subsampling	14	14	2	2
3	convolution	10	10	5	5
4	subsampling	5	5	2	2
5	convolution	1	1	5	5
6	Fully connected	1	1	5	5

3.2. Training

In the proposed Architecture, The data set contains 100 video samples of individuals. The CNN is composed of 8 layers: input layer, three convolutional layers and three sub sampling layer. Input layer is considered as 32×32. The dimension of the convolutional network is as shown in the table 1. The first convolutional layer has six feature maps of size 28×28 with a receptive field of 5×5. The second layer is sub sampling layer, contains six feature maps of size 14×14 with a receptive field of 2×2. The second convolutional layer has sixteen feature maps of size 10×10 with a receptive field of 5×5. The second sub sampling layer contains sixteen feature maps of size 5×5, with a receptive field of 2×2. The output layer is a fully connected layer with 120 feature maps with the size of 1×1, with a receptive field of 5x5. The CNN structure is as shown in fig 4.

```
Begin
    Read the video from the datbase
        Video=inputfile
        Read the frames of the video
            If(bbox=facedetector) then
                Facedetected(straight, left, right, up, down)
                Else
                    Show(face not detected)
                End
                For i <- 1 to size of(test features)
                If (classifier=CNN) then
                    K <- 1
                    CNN classify(trainfeatures, testfeatures)
                    Else
                        CNN predict(testfeatures,recognitior
                    End
                End
                If (class=testlabel(i))
                    Correct <- correct+1
                End
            End
        Accuracy <-correct/no of testsamples) * 100
End
```

. Fig 5 algorithm

4. Implementation

In this chapter the proposed algorithm for detection of head pose estimation based on different head pose views. It is implemented by MATLAB. The proposed algorithm is shown with the step by step implementation of the system.

The algorithm is as shown in the fig 5 shows how the implementation is been formed and how the test features and train features are classified. The prediction is done after classifying the images of frontal view, straight view, up profile view, down profile view, left profile view, and right profile view. Then the efficiency is calculated.

5. Experimental Analysis

Some of the experimental results of video clips are shown in the following table 2.

Table 2: Experimental Results

SL NO	VIDEO CLIP	DURATION	ATTRIBUTES	RECOGNITION
1	Video 1	0.07	Normal	100%
2	Video 2	0.07	Normal	100%
3	Video 3	0.07	Normal	98%
4	Video 4	0.07	Normal	98%
5	Video 5	0.07	Normal	98%
6	Video 6	0.07	Normal	100%
7	Video 7	0.31	Different hairstyle	94%
8	Video 8	0.07	Normal	100%
9	Video 9	0.07	Normal	100%
10	Video 10	0.07	Normal	100%
11	Video 11	0.29	Day time	100%
12	Video 12	0.07	Normal	94%
13	Video 13	0.07	Day time	98%
14	Video 14	0.07	Without spectacle	100%
15	Video 15	0.14	With spectacle	90%
16	Video 16	0.37	Night time	94%
17	Video 17	0.07	Normal	98%
18	Video 18	0.07	Normal	98%
19	Video 19	0.07	Normal	100%
20	Video 20	0.07	Normal	100%

ISSN 2348 – 7968

Based on the experimental study we say that 93.1% shows recognition. The experimental results are done on all the adverse conditions. The algorithm fails under very dark illumination (below 30lx). The accuracy is 93.1% according to experimental results

The epoch is the start time and iteration defines every single repetition of a process. During iterative training of neural networks an epoch is a single pass through the entire training set.

The root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. In the proposed method the RMSE value is 0.41, the accuracy is based on the training samples.

The MCR (misclassification rate) is 0.30. The images which are not recognized accurately are referred as misclassification rate
To measure the classification performance, we compute the detection rate which is given by

Detection rate =number of correct feature detections
Number of test images

The fig 6 shows the result of classified images accurately and calculates the frames of the video 11 from the table 2 is as shown in the fig 7.

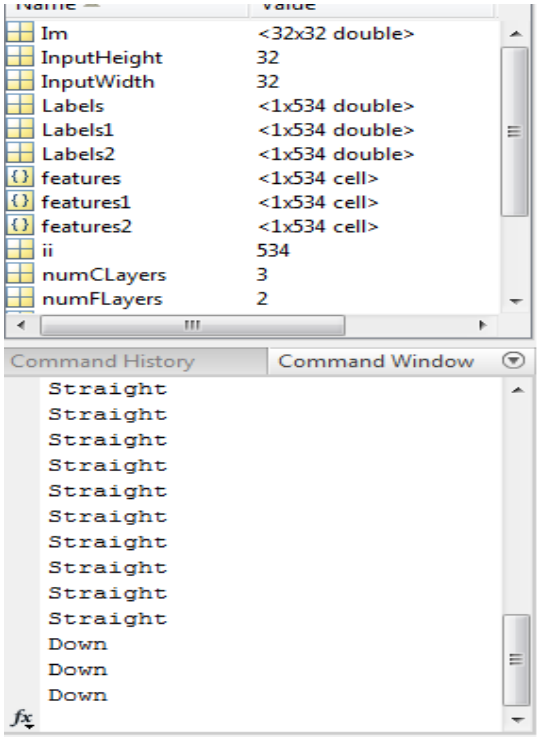

Fig 6: Classification

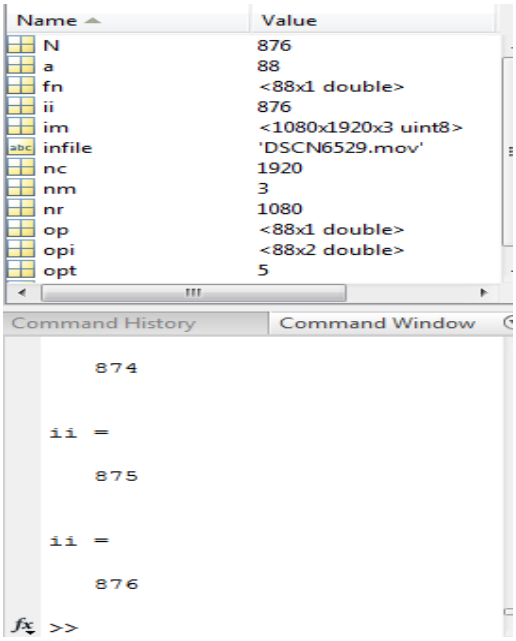

Fig 7: Results of frames

5.Conclusion

This project works well with low light illumination and any other environmental conditions. The processing is done under different adverse conditions like hair style, wearing spectacles, without wearing spectacles, different complexion. The head pose estimation is working with good efficiency and accuracy of above 85%. The dataset contains such fewer examples of all the environmental conditions.

References

[1] Erik Murphy-Chutorian, AnupDoshi, and Mohan ManubhaiTrivedi, “Head Pose Estimation for Driver Assistance Systems: A Robust Algorithm and Experimental Evaluation” Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference Seattle, WA, USA, Sept. 30 - Oct. 3, 2007.
[2] Yucheng Wei, LudovicFradet, Tieniu Tan, “Head Pose Estimation Using Gabor Eigenspace Modeling” National Laboratory of Pattern Recognition (NLPR), Institute of Automation
[3] AfifaDahmane, SlimaneLarabi, ChabaneDjeraba, Ioan Marius Bilasco, “Learning Symmetrical Model for Head Pose Estimation,”21st International Conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan.
[4] Erik Murphy-Chutorian, Member, IEEE, and Mohan ManubhaiTrivedi, “Head Pose Estimation and Augmented Reality Tracking: An Integrated System and Evaluation for

ISSN 2348 – 7968

Monitoring Driver Awareness," IEEE TRANSACTIONSON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 11, NO. 2, JUNE 2010

[5] StylianosAsteriadis, KostasKarpouzis, StefanosKollias, "Face Tracking and Head Pose Estimation using Convolutional Neural Networks" Image, Video, Multimedia Lab, National Technical University of Athens, Greece

[6] Ole Helvig Jensen, "Implementing the Viola-Jones Face Detection Algorithm," KongensLyngby 2008 IMM-M.Sc.-2008-93

[7] HuriehKhalajzadeh, Mohammad Mansouri and Mohammad Teshnehlab, "Face Recognition using Convolutional Neural Network and Simple Logistic Classifier" online conference, 2012

[8] Steve Lawrence, Member, IEEE, C. Lee Giles, Senior Member, IEEE, Ah Chung Tsoi, Senior Member, IEEE, and Andrew D. Back, Member, IEEE, "Face Recognition Convolutional Neural-Network Approach,"IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 8, NO.1, JANUARY 1997

[9] Jialue Fan, Student Member, IEEE, Wei Xu, Ying Wu, Senior Member, IEEE, and Yihong Gong, "Human Tracking Using Convolutional Neural Networks," IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 21, NO.10, OCTOBER 2010.

[10] Rainer Stiefelhagen, "Estimating Head Pose with Neural Networks - Results on the Pointing04 ICPR Workshop Evaluation Data" Interactive Systems Laboratories Universitat Karlsruhe (TH) Germany.