FACIAL IMAGE FEATURES FOR HEAD POSE ESTIMATION

S. Anishchenko[1,2], **A. Labantcev**[1], **I. Shepelev**[1]
[1]A.B.Kogan Research Institute for Neurocybernetics, Southern Federal University,
Rostov-on-Don, Russia, sergey.anishenko@gmail.com
[2]School of Engineering and Information Sciences, Middlesex University,
London, NW4 4BT, UK.

In computer vision context the head pose estimation is the process of extraction of the information about the head pose from an image of the face. Most common approach is to extract features from a facial image and predict pose using machine learning tools. In this research five different features was evaluated and compared to be used for the head pose prediction with the multilayer perceptron.

Introduction

Head pose estimation is the common task for many applications, such as human-machine interaction, biometry, medical image processing, operator fatigue estimation etc.

Many approaches have been proposed to estimate head pose based on facial images [1]. In general it can fall into two major groups: feature-based and model-based.

The first group algorithms normally consist of two steps: facial feature extraction and head pose prediction using machine learning tools. It is important to use features which are invariant to various parameters (such us illuminance level etc.) and allow to gather higher accuracy. This research is devoted to the evaluation and comparison of the five different facial images feature for the task of head pose estimation. The multilayer perceptron was used for construction of mapping between the image features and the head pose. The head pose was characterized by three angles: roll, yaw and pitch.

Video database

The public available video database with ground truth indicating head pose on each frame [4] was used in this research. The only one person' video sequences (n=6) was considered (the name of video in database is jal[sequence number], see Fig. 1). The facial feature points were marked up manually (Fig. 3.).

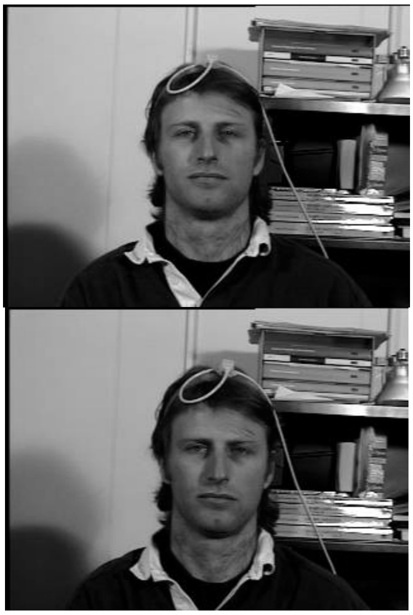

Fig. 1. An example of the frames with the same labels in ground truth. On the top the first frame of Jal3 is shown; on the bottom – the first frame of Jal8. It is obvious that head pose is different on the frames while labels in ground truth indicating same poses.

To ensure that ground truth is correct the video and labels was preprocessed in the following manner. The frames where same head pose is indicated by labels were compared. It was revealed that ground truth should be corrected because it indicates same pose on the frames which are actually different (Fig. 1).

To compute correction coefficient the most similar facial frames was detected on sequences. For example on the sequence named in database jal3 and jal8 the frames number 2 and 103 respectively was detected as the most similar (Fig. 2). Ground truth for those frames was (-0,593; -1,384; -1,011) and (12,041 ; 3,098; 1,384) respectively. Thus, labels of sequence jal8 should be corrected by adding (-12,634; -4,482; -2,395) to the roll, yaw and pitch respectively. The similarity between frames was computed by comparing angles between lines connected facial landmarks.

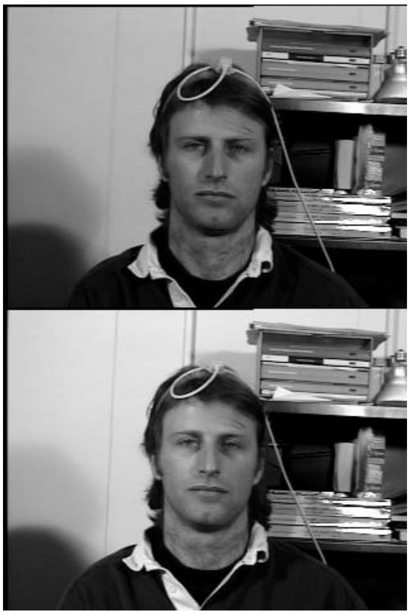

Fig. 2. The top and bottom photo is the most similar frames from video sequences jal3 (frame number 2) and jal8 (frame number 103) respectively. It is obvious that head is in the same pose on the both frames while ground truth indicates different pose.

To produce training and test set for this research the most similar frames was detected on each clip. Then ground truth was corrected to indicate same head pose on that frames. After that procedure the head pose in the overall video sequences varied in the ranges shown in Table 1.

Table 1. Ranges of angles in the ground truth.

	Interval
Roll	[-18,9; 23,2]
Yaw	[-16,3; 31,1]
Pitch	[-27,5; 17,1]

Facial features

Five sets of features for the face pose prediction were evaluated. The first one is the set of angles (n=8) between lines connected facial landmarks (eyes corners, nose tip, nose basement, see Fig. 2). All others features are based on Histogram of Oriented Gradients (HOG), but computed in the different ways.

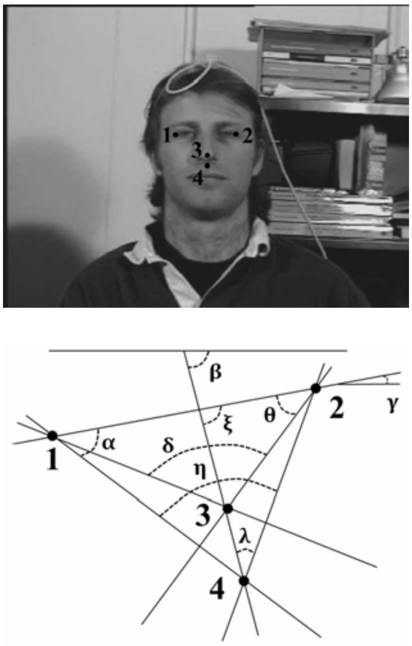

Fig. 3. Facial landmarks (top); and its scheme with shown landmarks (1, 2, 3) and angles (α, β, γ, δ, η, θ, ξ, λ) used as feature description of the face pose (bottom).

The second type of the feature is the HOG, computed inside the circumscribed rectangle of the facial landmarks. Gradient direction was quantized with step 22.5^{o}, thus, the result feature vector dimension was 16. Further, the rectangle was divided by 4x4 grid and HOG was computed in each cell. Histograms then were combined in one MultiHOG (MHOG), respectively the dimension of this feature vector was 16*4=64.

Next two feature vectors were achieved in the same way as HOG and MHOG but region of interest was detected by another method. Particular the colour segmentation algorithm described earlier in [3] was used. The features called CHOG and CMHOG.

Described feature vectors were extracted from each frame (n=732) and, further, along with ground truth were used for training and testing artificial neural network. The precision of pose prediction was analyzed to evaluate features.

Neural network model

The multilayer perceptron (MLP) with one hidden layer [2] was used for prediction of the head pose. The number of the inputs of the neural network was varied according to the dimension of the feature space being tested in the computational experiments. Since each of the head angle was predicted by the separate neural network, the output was the only one neuron. The number of neurons in the hidden layer was fixed for all computational experiments and equal to 24. Backpropagation algorithm was used for training. The stopping criterion was the follows. The network was trained until an error percentage *e* of head angle prediction for all training exemplars is less or equal to 10:

$$e \le e_{stop},$$

where

$$e = \frac{\left| y - y^{d} \right|}{y_{\max} - y_{\min}} \cdot 100\%, \quad (1)$$

$$e_{stop} = 10,$$

y is the actual output of the network corresponding to one of head angle to be predicted, y^d is the desired output, i.e. correct value, y_{max} is the maximum value of the angle and y_{min} is the minimum one.

Computational experiments results

In the computational experiments the five groups of features to predict three angles of head rotation was tested. 10-fold cross-validation technique was used and the results are presented in Table 2.

In the columns of the table the averaged values of the training and test accuracy are shown. It was computed using Eq. 1.

The feature space dimension is specified in the Table 2 and it defines the number of neural network inputs directly. An exception was the first features. Since the MLP could not reach the desired accuracy of prediction on training set, i.e. the stopping criteria could not been satisfied, the dimension of vector was changed in the following ways. Taking into account that training cases are statistically dependent, for this features the training set was represented as time series. I.e. to process current feature vector the previous ones can be also taken into account to achieve more reliable prediction. The minimum number of previous vectors (i.e. delay line), which allow to reach the desired accuracy of training was founded. Since the delay line was equal to nine, thus, the number of neural network inputs was 8*9=72.

Table 2. Cross-validation results.

Spherical coordinates	Training accuracy, %	Test accuracy, %	Test accuracy, degree
Feature: angles between lines connected facial landmarks (dimension – 8, dimension for time series representation (delay line=9) - 72).			
roll	96.3	94.6	2.3
yaw	94.5	92.9	3.4
pitch	96.1	94.5	2.5
Feature: HOG (dimension - 16)			
roll	96.4	94.9	2.1
yaw	95.6	92.1	3.7
pitch	96.5	95.1	2.2
Feature: MHOG (dimension - 64)			
roll	96.5	94.7	2.2
yaw	95.8	**94.3**	2.7
pitch	96.6	95.3	2.1
Feature: CHOG (dimension - 16)			
roll	96.7	95.9	1.7
yaw	96	93.7	2.9
pitch	96.4	95.6	1.9
Feature: CMHOG (dimension - 64)			
roll	96.8	**96.2**	1.6
yaw	96.7	**94.3**	2.7
pitch	96.8	**95.9**	1.8

Conclusion

In this research a set of five feature vectors for head pose prediction was evaluated with MLP using cross-validation technique. The first type of feature was the set of angles between lines connected facial landmarks. To reach desired prediction accuracy it was represented as time series. All other features were based on HOG.

The NN was trained until maximum error in the training set reached 10%. Then averaged error was analyzed (Table 2.).

The results show that yaw angle is the most difficult for prediction. Based on the test accuracy of this angle we conclude that MHOG and CMHOG features outperform the others. Comparing prediction accuracy for all of three angles it can be concluded that

CMHOG is the most suitable feature for head pose estimation.

The performance of all HOG-based features was better than facial landmark-based one because the image resolution is small and distance between some facial landmarks is less than 5 px, therefore, small changes in head pose are not reflected distinctively in the landmarks position on the images.

Acknowledgments: The work is supported by the Russian Foundation for Basic Research, grant 11-01-00750a.

References

1. Erik Murphy-Chutorian and Mohan Manubhai Trivedi, “Head pose estimation in computer vision: A survey”. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, issue 4, April 2009: 607–626.
2. Bishop C.M. Neural Networks for Pattern Recognition. Oxford University Press. 1995.
3. S. Anishenko, D. Shaposhnikov, R. Comley, X. Gao. A colour based approach for face segmentation from video images under low luminance levels. // In Proc. of the 11th IASTED International Conference on Computer Graphics and Imaging (CGIM 2010) Feb. 17-19, 2010, Innsbruck, Austria. - pp. 184-189.
4. Cascia E. L., Sclaroff S., Athitsos V. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. // Pattern Analysis and Machine Intelligence, 22(4), 2000.