ISSN 2395-1621

Survey on Head Pose Estimation and Head Pose Classification

[#1]Priyanka Jadhav, [#2]Mrs. V. L. Kolhe

[1]jadhav.priyanka96@gmail.com
[2]vlkolhe@gmail.com

[#12]Department of Computer Engineering

D. Y. Patil College of Engineering, Akurdi Pune.

ABSTRACT

In Computer vision systems head pose estimation in low resolution and gaze estimation are not easy tasks. Human gazing direction makes accurate classification difficult. Head pose estimation faces some problems for non-frontal head poses. In this paper, we discuss different techniques for head pose estimation. Head pose estimation and gaze estimation are useful in visual surveillance and Human Computer Interaction. Classification of frontal and non-frontal images can be done using head pose estimation. This information is also useful in human-to-human/scene interaction. In this paper, we discuss the problems in head pose estimation; head pose classification and present an organized survey describing the evolution of the field.

Index Terms: Head pose estimation, Head pose classification, RGB-D data, Gaze direction.

ARTICLE INFO

Article History

Received: 8th January 2017

Received in revised form :

8th January 2017

Accepted: 10th January 2017

Published online :

11th January 2017

I. INTRODUCTION

Modelling human head pose is a challenging problem in computer vision and signal processing. The headpose signal gives meta-information about communicative gestures, salient regions, crowd behavioral dynamics and tracking, and anomaly detection. It is helpful in close range of domains where eye tracking is not possible. Head pose estimation and gaze direction are important in many applications like visual surveillance, human computer interaction for the analysis of human behaviour. The gaze is nothing but the combination of head pose and eye location.

Head pose estimation is the process of determining the orientation of human head in digital imagery. In the context of computer vision, the ability to infer the orientation of a human head relative to the view of a camera is the head pose estimation[1]. Head pose estimator can demonstrate invariance in images or image sequences. Head pose estimation can be applied on different algorithms for identification of frontal and non-frontal faces[2].

Head pose estimation and gaze direction provides impor-tant mete information about communication gesture, crowd behavioural dynamics and tracking, anomaly detection. Head pose estimation is important in domains where close-level iris/ eye tracking is not possible. Human head pose estimation can be done by using 3 degrees of freedom (DOF) on three dimensional co-ordinate system in which 3 dimensions are roll, pitch and yaw as shown in Fig. 1

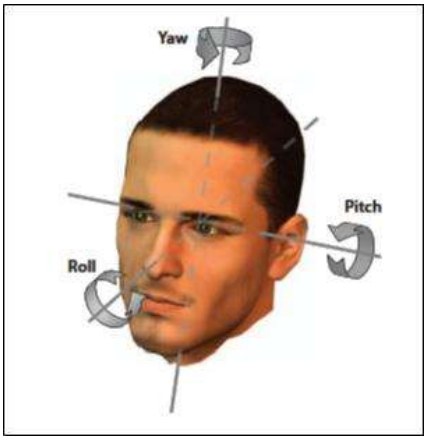

Fig. 1. Three degrees of Freedom of human head by using roll, pitch and yaw[1].

Head pose estimation is linked with gaze estimation. It is the ability to characterize the direction and focus of a

persons eyes. Head pose provides a coarse indication of gaze that can be estimated when close level iris/eye tracking is not possible (like low-resolution imagery, or in the presence of eye-occluding objects like sunglasses)[1].

In many real life applications, automatic and robust algo-rithms for head pose estimation can be useful. The goal of systems like human-computer interfaces or a necessary pre-processing analysis such as identification or facial expression recognition is accurately localizing the head and its orientation. But it is a challenging task in computer vision community to develop fast and reliable algorithms for head pose estimation. To overcome such problem the recent development and avail-ability of 3D sensing technologies are useful.In the regression based approaches, head pose estimation problem can be solved using some parameters like depth data[3].

In recent years, head pose estimation can be done using RGB or RGB-D i. e. RGB depth data. This depth data can be used to enhance RGB-based head pose estimation and human tracking. In low resolution, using RGB-D data classification of human gazing direction problem can be solved. In this, a convolutional neural network based model is used for head pose estimation and classification[2].

Section II presents a survey of different methods of head pose estimation and head pose classification. Section III presents material i. e. different datasets for head pose esti-mation and head pose classification which is useful in gaze direction where close level iris/eye tracking is not possible.

II. LITERATURE SURVEY

The various methods have been developed for head pose estimation are as follows :

A. Head Pose Estimation

Evolutionary taxonomy consists of the following categories which describe the conceptual approaches that are used to estimate head pose:

- Appearance Template Methods : In these methods, compare a new image of a head to a set of examples (each labeled with a discrete pose) to find the most similar view.

- Detector Array Methods : These methods train a series of head detectors and assign a discrete pose to the detector with the greatest support.

- Nonlinear Regression Methods : Nonlinear regression tools are used to develop a functional mapping from the image or feature data to a head pose measurement.

- Manifold Embedding Methods : seek low-dimensional manifolds. It models the continuous variation in head pose. New images can be embedded into manifolds and used for embedded template matching or regression.

- Tracking Methods : It recovers the global head pose changes from the observed movement between video frames.

- Hybrid Methods : It combines one or more of above mentioned methods to overcome the limitations inherent in any single approach[1].

For head pose estimation, the manifold must be modeled. A new sample is embedded into the manifold. This is a low-dimensional embedding and it is used for head pose estimation with techniques like regression in the embedded space or embedded template matching. Any dimensionality reduction algorithm can be a challenge lies in creating an algorithm which recovers head pose ignoring other sources of image variation. Principal component analysis (PCA) and its nonlinear kernelized version KPCA are the two dimensionality reduction techniques. Head pose estimation can be done using PCA. By projecting an image into a PCA subspace and comparing the results to a set of embedded templates. In KPCA, primary components are relate to pose variation[1].

In previous work, Robertson and Reid proposed a feature based vectors model on skin detection to classify head poses in 8 different orientations. This technique was extended by Benfold et al.[4]. They proposed algorithm for classification of head poses in low resolution and mapping between colours and labels. But all template based methods faces two problems : localize head poses in low resolution is difficult, another is non-frontal head poses may appear more similar. To avoid such problems some researchers proposed different feature space for representing head images.

A Neural Network based approach was proposed by Stiefelhagen[5] to estimate horizontal and vertical head ori-entation of a person from facial expressions. Non-linear re-gression like high-dimensional manifold based approach was proposed to determine head pose and face images in various pose angles[6]. Chen and Odobez[7] proposed multi-level Histogram of Oriented Gradients (HOG) based method for head-pose and body-pose estimation in surveillance videos.

On the other hand, on the HCI side of the problem solutions are limited to 2 meter distance from the sensor

along with near-frontal head-poses. An iterative closest point (ICP) based mesh fitting method has been proposed for head pose detection[8], [9]. Work on head pose regression has been introduced for scene and human interaction understanding[10]. This work focuses on head-pose regression and interaction detection in 2D movie/ tv-series scenes.

Recently, manifold based metric learning methods have been applied to head pose estimation[11]. In other approaches, the spherical nature of the view manifold of objects is used as a strong prior for manifold learning[12].

Head pose estimation can be useful in many applications like anomaly detection, crowd behavioral dynamics. Head pose estimation provides a interface for computing. Some existing examples includes control to computer mouse using head pose movements, respond to pop-up dialog boxes with head nods and shakes or use head gestures to interact with embodied agents.

B. Deep Learning and Convolutional Neural Network

Deep learning, especially convolutional neural networks(CNNs) are used to learn non-linear representations from input data and have been especially successful on images[13] and audio[14]. But, this is in contrast to traditional computer vision pipelines like HOG[7]. These features would be used as input to machine learning framework like support vector machines (SVM) for achieving classification or regression. In [13], trained a large, deep convolutional neural network is used to classify the 1.2 million high-resolution images in the ImageNet in 1000 different classes[15].

On the other hand, CNNs are supervised, discriminative and have mostly surpassed the Deep Belief Networks(DBNs) in terms of accuracy on large labelled datasets like the Imagenet[13]. CNNs are deep models which belongs to fully connected networks. CNNs are also applied in multi-modal RGB-D domains. In [15], author introduced a fusion of RGB-D channels and transfer learning for initialisation of the weights of the green, blue and depth channels with filters learned from the depth channel. But this form of early fusion is not very helpful. RGB-D networks are generally trained with late fusion[16], [17].

CNNs are used to train large scale labelled training data. The number of parameters in the convolution layers are orders of magnitude lower than the fully connected layer. Separate CNN will be trained on RGB and depth modularities based on the architecture. Networks will be modified by changing

Rectified Linear Unit non-linearities (RELU) with Parametric Rectified Linear Unit (PRELU) and their corresponding weight initialization.

III.MATERIALS AND CLASSIFICATION METHODS

A. Materials

Material refers to the datasets used in different papers. There are many data sets for evaluating head pose estimation which are mentioned below.

1) Oxford town centre dataset

Oxford town centre dataset provides higher resolution images that were up to 20 pixels in diameter. For these images, classifiers are based on gradient histograms and colour differences were robust to small errors in the head location. When combined with tracking, both the locations and gaze directions of pedestrians could be estimated in real-time[18].

2) BIWI Kinect head-pose dataset

The database contains 24 sequences acquired with a Kinect sensor. 20 people (some were recorded twice - 6 women and 14 men) were recorded with turning their heads, sitting infront of the sensor, at one meter of distance. For each sequence, the corresponding .obj file represents a head template of the neutral face of that specific person. For each frame, a rgb.png and A depth.bin files are provided, containing color and depth data[19].

3) Caviar shopping centre dataset

The CAVIAR head pose dataset is resized to 50 50 pixels, come from a set of sequences which have 1500 frames on average, acquired from a real surveillance camera located in a shopping centre in Lisbon. The dataset is composed by non-occluded head images for a total number of 21326 examples and 366 occluded examples[20].

4) HIIT Head Orientation dataset

The HIIT dataset contains 6 classes and 2000 examples. The size of the samples is 50x50 pixels, without margin around the heads. Dataset has a stable background and no occlusions, so that it represents the ideal scenario[21].

5) Multi-PIE Face Database

In CMU Multi-PIE face database 337 people recorded in four sessions over the span of five months whivh contains more than 750,000 images. Subjects were imaged under 15 view points and 19 illumination conditions and it displays a range of facial expressions. High resolution frontal images were acquired in this datadase. In total, the database

contains more than 305 GB of face data. The Content page describes the database in more detail[22].

The datasets also vary in resolution from very high (BIWI) to very low (Caviar).

B. Head Pose Classification

Classification of Head poses can be done under the different class labels. One of the above dataset is used for head pose classification. Different classification methods are as follows :

- K-clusters Regression Forests :

This method is based on the standard random forests for regression. It introduces more flexible node split algorithm. The splitting rule of K-cluster Regression Forests at each node consists of clusters of training samples as multiple groups and learns the decision function to distinguish the samples in the same cluster from others as a classification problem. It splits data using the predicted cluster label by the trained classifier[23].

- Multivariate Label Distribution (MLD) :

This is a recently proposed classification method. It captures the correlation between neighboring poses in the label space. Based on standard Label Distribution Learning (LDL), Multivariate Label Distribution is extended to model the two-dimensional output of head pose estimation (i.e., yaw and pitch angles of head viewing direction). MLD can be treated as multi-label learning with correlated labels[24].

- Support Vector Machines (SVM) :

SVM is widely used classifier and it separates the classes with a largest possible distance between them. It is used together with the kernel trick that implicitly maps the data into a high dimensional kernel space and also the linear kernel is widely used with large data sets[25].

- Artificial neural networks (ANN) :

ANNs are powerful, nonlinear models and can learn complex relationships between variables. This network is used in various machine learning problems including image recognition and optical character recognition. ANN treats the multi-label encoding of the classes in a straight-forward manner and does not require any multi-category heuristics. The neural network topology consists of 2 hidden layers having 200 and 70 neural units respectively with sigmoid activation functions. The output is a softmax layer of size 22[26].

Deep Belief Networks (DBN) are used when unsupervised data is used for classification. Convolutional neural networks (CNNs) are used for classification of supervised and discriminative data[13]. CNNs have also been applied in the multimodal RGB-D domain. RGB-D networks are trained with late fusion and modalities are learned separately and combined in the classifier phase[15].

TABLE I
EFFECT OF STACKING THE CLASSIFIERS[26].

Model	Public MAE	Private MAE
500-tree random forest	6.156	6.546
5-nearest neighbor	6.826	7.460
Logistic regression	6.694	6.949
Extremely randomized trees (with stacking)	4.772	4.718

- Ensemble Methods and Stacked Generalization :

Stacked generalization is a tool. This tool have a classifier and it is used to train a pool of classifiers and feed their outputs to a predictor. Predictor is used as randomized trees classifier is trained on the training data with augmented features. A separate models are trained using yaw and pitch angles. Table 1 shows the effect of stacking in terms of Public MAE and Private MAE for the individual models and the stacked ensemble[27].

C. Gaze Estimation Methods

The different gaze estimation methods are available as follows :

- Geometric based methods

These methods depend on the detection of local features. In most of the methods gaze annotated samples are collected and used to determine user specific parameters describing the eyeball geometry or a direct mapping to the point of regard. IR illumination and sensing are methods in this category. These methods are useful in face motion captures and eye tracking. Recent methods apply a similar techniques to RGB-D data[28].

- Appearance based methods

In these methods, direct mapping method is modelled from full eye image to gaze parameters. It is useful in low-resolution gaze sensing. These approaches avoid the local features tracking. In [29] author proposed a method and used to train a Support Vector Regression (SVR) model using the stacked eye image pixels and the appearance feature vector are used.

- Head pose invariant gaze estimation

This approach is used to rectify the eye images into a canonical (frontal) head viewpoint and scale regardless of the actual head pose by exploiting the calibrated RGB-D input data, and the gaze estimation is done in canonical view. In addition to this, 3D Morphable Model (3DMM) is introduced to create a user-specific 3D facial template and generate a large variety of possible face shapes using a relatively small set of coefficients[30].

- Person invariant gaze estimation

In this method, Person invariant classifier is used when training data is not available for the test subject to learn an appearance to gaze regression model. The cross-user eye image alignment problems are solved by using Person invariant gaze estimation[30].

IV. CONCLUSION

Head pose estimation is process to recover information gap between people and computers. Human activity gives information about intent, motivation, people's attentio i the world. Most of the head pose estimation techniques faces the problem due to lighting, background and camera, etc. Viewing the progress in head pose estimation in real life. In recent years, people have become much more aware of the need for comparison metrics which emphasize pose variation than image variation. The different methods are used for head pose estimation like appearance based methods, hybrid methods, model-based tracking and so on. These methods will require standard datasets to understand the potential of dataset.

Head pose estimation is also a challenging task in low reso-lution and/or using frontal and non-frontal head poses. Human attention modelling and head-pose estimation can be done using low-resolution and high-resolution in various domains using different methods. Head pose estimation systems will play a key role in the creation of intelligent environments. Head pose classification methods are used to classify head pose images. It is useful in human gazing direction. Using gaze estimation human-human/scene interaction detection can be achieved. Gaze directions are useful in anomaly detection, crowd behavioural dynamics, group detection and a scene based on focus of attention.

ACKNOWLEDGMENT

The authors would like to thank the publishers and re-searchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

REFERENCES

[1] Murphy-Chutorian, Erik, and Mohan Manubhai Trivedi. "Head pose estimation in computer vision: A survey."IEEE transactions on pattern analysis and machine intelligence 31.4 (2009): 607-626.

[2] Mukherjee, Sankha S., and Neil Martin Robertson. "Deep Head Pose: Gaze-Direction Estimation in Multimodal Video."IEEE Transactions on Multimedia 17.11 (2015): 2094-2107.

[3] Fanelli, Gabriele, Juergen Gall, and Luc Van Gool. "Real time head pose estimation with random regression forests."Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[4] Benfold, Ben, and Ian Reid. "Colour Invariant Head Pose Classification in Low Resolution Video."BMVC. 2008.

[5] Stiefelhagen, Rainer. "Estimating head pose with neural networks-results on the pointing04 ICPR workshop evaluation data."Pointing04 ICPR Workshop of the Int. Conf. on Pattern Recognition. 2004.

[6] Balasubramanian, Vineeth Nallure, Jieping Ye, and Sethuraman Pan-chanathan. "Biased manifold embedding: A framework for person-independent head pose estimation."2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007.

[7] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection."2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. IEEE, 2005.

[8] Mora, Kenneth Alberto Funes, and Jean-Marc Odobez. "Gaze estimation from multimodal kinect data."2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012.

[9] Cazzato, Dario, Marco Leo, and Cosimo Distante. "An investigation on the feasibility of uncalibrated and unconstrained gaze tracking for human assistive applications by using head pose estimation."Sensors 14.5 (2014): 8363-8379.

[10] Fanelli, Gabriele, Juergen Gall, and Luc Van Gool. "Real time head pose estimation with random regression forests."Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[11] Peng, Xi, et al. "From circle to 3-sphere: Head pose estimation by instance parameterization."Computer Vision and Image Understanding 136 (2015): 92-102.

[12] Ma, Bingpeng, et al. "CovGa: a novel descriptor based on symmetry of regions for head pose estimation."Neurocomputing 143 (2014): 97-108.

[13] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks."Advances in neu-ral information processing systems. 2012.

[14] Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups."IEEE Signal Processing Magazine 29.6 (2012): 82-97.

[15] Alexandre, Lus A. "3D object recognition using convolutional neural networks with transfer learning between input channels."Intelligent Au-tonomous Systems 13. Springer International Publishing, 2016. 889-898.

[16] Gupta, Saurabh, et al. "Learning rich features from RGB-D images for object detection and segmentation."European Conference on Computer Vision. Springer International Publishing, 2014.

[17] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convo-lutional networks for semantic segmentation."Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[18] Benfold, Ben, and Ian Reid. "Guiding Visual Surveillance by Tracking Human Attention."BMVC. Vol. 2. No. 6. 2009.

[19] Fanelli, Gabriele, et al. "Random forests for real time 3D face analy-sis."International Journal of Computer Vision 101.3 (2013): 437-458.

[20] Available on 'https://sites.google.com/site/diegotosato/ARCO/caviar'

[21] Availableon 'https://sites.google.com/site/diegotosato/ARCO/iit'

[22] Gross, Ralph, et al. "Multi-pie."Image and Vision Computing 28.5 (2010): 807-813.

[23] Hara, Kota, and Rama Chellappa. "Growing regression forests by clas-sification: Applications to object pose estimation."European Conference on Computer Vision. Springer International Publishing, 2014.

[24] Geng, Xin, and Yu Xia. "Head pose estimation based on multivariate label distribution."Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[25] Scholkopf, Bernhard, and Alexander J. Smola. "Learning with kernels: support vector machines, regularization, optimization, and beyond."MIT press, 2002.

[26] LeCun, Yann, et al. "Gradient-based learning applied to document recognition."Proceedings of the IEEE 86.11 (1998): 2278-2324.

[27] Wolpert, David H. "Stacked generalization."Neural networks 5.2 (1992): 241-259.

[28] Guestrin, Elias Daniel, and Moshe Eizenman. "General theory of remote gaze estimation using the pupil center and corneal reflections."IEEE Transactions on biomedical engineering 53.6 (2006): 1124-1133.

[29] Noris, Basilio, Jean-Baptiste Keller, and Aude Billard. "A wearable gaze tracking system for children in unconstrained environments."Computer Vision and Image Understanding 115.4 (2011): 476-486.

[30] Funes-Mora, Kenneth A., and Jean-Marc Odobez. "Gaze Estimation in the 3D Space Using RGB-D Sensors."International Journal of Computer Vision: 1-23.

