Neurocomputing 148 (2015) 455–466

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# VoD: A novel image representation for head yaw estimation

Bingpeng Ma [a], Rui Huang [b,*], Lei Qin [c]

[a] School of Computer and Control Engineering, University of China Academy Science, Beijing, China
[b] Huazhong University of Science and Technology, Wuhan, China
[c] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

*Article history:*
Received 8 November 2013
Received in revised form
30 March 2014
Accepted 8 July 2014
Communicated by Jinhui Tang
Available online 19 July 2014

*Keywords:*
Head yaw estimation
Image representation
Fisher vectors
Metric learning

ABSTRACT

Building on the recent advances in the Fisher kernel framework for image classification, this paper proposes a novel image representation for head yaw estimation. Specifically, for each pixel of the image, a concise 9-dimensional local descriptor is computed consisting of the pixel coordinates, intensity, the first and second order derivatives, as well as the magnitude and orientation of the gradient. These local descriptors are encoded by Fisher vectors before being pooled to produce a global representation of the image. The proposed image representation is effective to head yaw estimation, and can be further improved by metric learning. A series of head yaw estimation experiments have been conducted on five datasets, and the results show that the new image representation improves the current state-of-the-art for head yaw estimation.



## 1. Introduction

During the last decades, there has been a significant progress in the face recognition research. However, one of the most challenging factors influencing the robustness and accuracy of face recognition is pose variation. To achieve robustness to pose variation, one might have to process face images differently according to their poses. Therefore, head pose estimation has been an active research topic for many years.

More precisely, head pose estimation essentially means the computation of three types of rotations of the head: yaw (looking left or right), pitch (looking up or down) and roll (tilting left or right). Among them, the roll rotation can be computed easily by the relative positions of the feature points, but the other two rotations are rather difficult to estimate. As the estimation of the yaw rotation has many important applications, it attracts more attention than pitch estimation [1], with more research data available. Therefore, in this paper, as most previous works have done, we focus on the challenging problem of estimating the head yaw pose from the input face images.

In head pose estimation, one of the crucial steps is to extract image representation characterizing the pose. Generally speaking, the proposed visual features can be roughly categorized into *global* and *local* features. While global features encode the holistic configuration of the image, local features encode the detailed traits within a local region. In the literatures, many methods combine both global and local features as they play different roles in the visual perception process. Among them, perhaps the most commonly used one is the Bag-of-Words (BoW) model [2] in which local descriptors extracted from an image are first mapped to a set of visual words and the image is then represented as a histogram of visual word occurrences. Recently, the Fisher vectors [3], which encode higher order statistics of local descriptors, improved the BoW model greatly for image classification. Instead of encoding only the frequency of visual word occurrences, Fisher vectors encode how the parameters of the model should be changed to represent the image. It can be seen as an extension of BoW, and has been shown to achieve the state-of-the-art performance for several challenging object recognition and image retrieval tasks [4,5].

Inspired by these exciting advances, we present a novel image representation for head yaw estimation in this paper. More specifically, the proposed image representation encodes a new type of concise 9-dimensional local descriptors with Fisher vectors to describe the head images, called *fisher Vector of local Descriptors*, VoD for short. The VoD representation has been experimentally validated on five head pose datasets (FacePix, Pointing'04, MultiPIE, CAS-PEAL and our own dataset). The results on these datasets show that the proposed representation outperforms the state-of-the-art.

The contribution of this paper is three-fold. Firstly, we proposed a 9-dimensional local attribute vector which can be applied in the Fisher vector method. The 9-dimensional vector is extracted at each pixel, which contains the coordinates, intensity, the first- and second-order derivatives and the magnitude and orientation of the gradient of the pixel. Compared with the SIFT feature used in the traditional Fisher vectors, the computational efficiency of the 9-dimensional local descriptor is significantly improved. More

* Corresponding author.
*E-mail addresses:* bpma@ucas.ac.cn (B. Ma), ruihuang@hust.edu.cn (R. Huang), lqin@jdl.ac.cn (L. Qin).

http://dx.doi.org/10.1016/j.neucom.2014.07.019
0925-2312/

importantly, despite its conciseness, the descriptor preserves enough information essential to head pose estimation.

Secondly, the proposed local descriptors are encoded and aggregated into Fisher vectors to form the new VoD representation. To keep the spatial structure of the head in the global representation, we divide a head image into many rectangular bins and compute one VoD per bin.

Finally, we further improved the discriminative ability of VoD by supervised metric learning. Considering the great success of Keep It Simple and Straightforward Metric Learning (KISSME) [6], we train kVoD from VoD using KISSME under a supervised setting. The final product improved the accuracy of head pose estimation greatly over the state of the art.

The remainder of this paper is organized as follows: in Section 2, we introduce the related methods for head pose estimation; In Section 3, the proposed representation is introduced in detail. Experiments on five challenging datasets are shown in Section 4 to demonstrate the effectiveness of the proposed representations. Conclusions are drawn in Section 5 with some discussions on the future work.

## 2. Related work

Head pose estimation from images is a challenging problem due to large variations of illumination, facial expressions, subject variability, occlusions, noise and perspective distortion. A generic (i.e., person-independent) algorithm for head pose estimation has to be robust to such factors. There exists a large amount of literatures on this topic, see [7] for a review. Broadly speaking, most previous work can mainly be categorized into three groups: algorithms based on facial features [8–10], model-based algorithms [11,12], and appearance-based algorithms [1,13].

For the algorithms based on facial features, the 3D face structure is exploited along with a priori anthropometric information in order to define the head pose. The elliptic shape of the face, the mouth–nose region geometry, the line connecting the eye centers, the line connecting the mouth corners and the face symmetry are some of the geometric features used to estimate the head pose. This category of algorithms has a major disadvantage: they are sensitive to the misalignment of the facial feature points, while the accurate and robust localization of facial landmarks remains an open problem, especially for the non-frontal faces.

Using the 3D structure of human head, the model-based algorithms build a priori known 3D models for human faces and attempt to match the facial features such as the face contour and the facial components of the 3D face model with their 2D projections. Once the correspondences from 3D to 2D are found between the input data and the face model, conventional pose estimation techniques are exploited to provide the head pose. The main problem for these algorithms is that it is difficult to precisely build the head model for different persons and to define the best mapping of the 3D model to the 2D face image.

The appearance-based algorithms typically assume that there exists a certain relationship between the 3D face pose and some properties of the 2D face image and infer the relationship by using a large number of training images and statistical learning techniques. Intuitively, these appearance-based algorithms can naturally avoid the drawbacks of the algorithms based on facial features and the model-based algorithms. Therefore, they have attracted more and more attention. In these algorithms, instead of using facial landmarks or face models, the whole image of the face is used for pose estimation.

Generally speaking, there are two steps in appearance-based algorithms: feature extraction and classification. For feature extraction, the subspace-based algorithms have been widely used since they can reduce the data dimensionality. Specifically, Gong et al. studied the trajectories of multi-view faces in linear Principal Component Analysis (PCA) feature space [14,15]. They used two Sobel operators (horizontal and vertical) to filter the training images. PCA was then performed to reduce the dimensionality of the training examples. Finally, Support Vector Machine (SVM) regression was utilized to construct two pose estimators for the pitch and yaw angles. Darrell et al. computed a separate eigenspace for each face under each possible pose [16]. The head pose was determined by projecting the input image onto each eigenspace and selecting the one with the lowest residual error. In some sense, this method can be formulated as a Maximum A Posteriori (MAP) estimation problem. Li et al. exploited Independent Component Analysis (ICA) and its variants, subspace analysis and topographic ICA for pose estimation [17]. ICA takes into account higher order statistics required to characterize the view of objects and suitable for the learning of view subspaces. Wei et al. proposed that the optimal orientation of the Gabor filters can be selected for each pose to enhance pose information and eliminate other distractive information like variable facial appearance or changing environmental illumination [13]. In their method, a distribution-based pose model was used to model each pose cluster in Gabor eigen-space. Haj et al. created a system based on a kernelized variant of Partial Least Squares (PLS) that was insensitive to data misalignment, while achieving excellent accuracy on several datasets [18]. Their work shows that regression tools can be very effective for the case of estimating the orientation of a face.

Besides the traditional subspace-based algorithms, since the set of the face images with various poses intrinsically form a manifold in the image space, manifold learning [19–21] for head pose estimation is thus getting popular recently [22–26]. In [22], by thinking globally and fitting locally, Fu and Huang proposed to use the graph embedded analysis method for head pose estimation. They first constructed the neighborhood weighted graph in the sense of supervised locally linear embedding [19]. The unified projection was calculated in a closed-form solution based on the graph embedding linearization, and then they projected new data into the embedded low-dimensional subspace with the identical projection. To overcome the disadvantage that most embedding based methods are unsupervised in nature and do not extract features that incorporate class information, in [27], Huang et al. presented the method Supervised Local Subspace Learning ($SL^2$), which learns a local linear model from a sparse and non-uniformly sampled training set. The authors argued that $SL^2$ was robust to under-sampled regions, over-fitting and image noise. In [28], the authors presented a two layer system (coarse/fine). They assumed that for local patches of the latent manifold, neighborhood-dependent linear functions can be used to effectively describe the modes of variation that correspond to pose changes. Then, they modeled the global nonlinear pose manifold in terms of local linear transforms.

After extracting the representation of the face images, classifiers are trained and then used to determine the actual pose of an input image. Besides the above-mentioned SVMs, some widely used classifiers in pattern recognition, such as neural networks, Bayesian approaches, and Boosting, have all been applied in head pose estimation. In [29], a neural network-based approach was presented in which a multi-layer perception was trained for each pose angle (pan and tilt) by feeding it with preprocessed face images captured by a panoramic camera. In [30,31], based on a Bayesian formulation, Ba et al. proposed an algorithm that couples head tracking and pose estimation in a mixed state particle filter framework. In [32], the authors used Boosting regression and simple Haar-type features to estimate the head pose.

More recently, 3D sensing technologies are becoming ever more affordable and reliable. More and more researchers used the additional depth information to overcome some problems inherent of methods based on 2D data [33,34].