Engineering Applications of Artificial Intelligence 50 (2016) 168–176

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Brief paper

Discriminant sparse label-sensitive embedding: Application to image-based face pose estimation

F. Dornaika [a,b,*], C. Chahla [c,d], F. Khattar [a,e], F. Abdallah [c], H. Snoussi [d]

[a] University of the Basque Country UPV/EHU, San Sebastian, Spain
[b] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[c] Lebanese University, Lebanon
[d] University of Technology of Troyes, Troyes, France
[e] University of Pau and Pays de l'Adour, Bayonne, France

ARTICLE INFO

Article history:
Received 28 August 2015
Received in revised form
14 January 2016
Accepted 23 January 2016

Keywords:
Model-less face pose estimation
Manifold learning
Locality preserving projections
Sparse coding
Out-of-plan rotation

ABSTRACT

In this letter, the authors propose a new embedding scheme for image-based continuous face pose estimation. The main contributions are as follows. First, it is shown that the concept of label-sensitive Locality Preserving Projections, proposed for age estimation, can be used for model-less face pose estimation. Second, the authors propose a linear embedding by exploiting the connections between facial features and pose labels via a sparse coding scheme. The resulting technique is called Sparse Label sensitive Locality Preserving Projections (Sp-LsLPP). Third, for enhancing the discrimination between poses, the projections obtained by Sp-LsLPP are fed to a Discriminant Embedding that exploits the continuous labels. The resulting framework has less parameters compared to related works. It has been applied to the problem of model-less face yaw angle estimation (person independent 3D face pose estimation). It was tested on three databases: FacePix, Taiwan, and Columbia. It was conveniently compared with other linear and non-linear techniques. The experimental results confirm that the proposed framework can outperform, in general, the existing ones.

1. Introduction

Face image analysis has attracted increasing attention in the computer vision community. It is required for developing artificial systems able to perform intelligent behavior such as face recognition and annotation (Choi et al., 2011; Hu et al., 2011; Huang et al., 2014; Hwang et al., 2011; Lu et al., 2013), facial landmark annotation (Zhu and Ramanan, 2012), age estimation (Chao et al., 2013), or face pose estimation (Murphy-Chutorian and Trivedi, 2009). Human's facial pose is considered as an important cue of non-verbal communication. Indeed, humans can easily discover and understand other people's intentions easily by interpreting their head pose. However, in order to make a machine capable of interacting with the human's head movements and expressions, huge effort has to be done to estimate the pose from the pixel representation of a facial image in a robust and efficient way. The estimation process requires a series of processing steps to transform a pixel-based representation of a face into a high-level concept of direction. 3D face pose can play an important role in many applications (Wholer, 2013). For instance, it can be used in the domain of face recognition either by using hierarchical models or by generating a frontal face image. The head pose estimation refers to the specific task consisting of determining the position and/or the orientation of the head in an image (e.g. a facial one). This task is a challenging problem because there are many degrees of freedom that should be estimated.

Recently, researchers investigated the 3D sensing technologies for face pose estimation (Fanelli, 2011; Pashalis et al., 2012). Although this technology promises a lot in overcoming some of the problems of methods based on 2D data by using the additional depth information, it suffers of serious computational problems. It cannot handle large pose variations, it cannot run in real time and it needs manual initialization. Furthermore, they are not as scalable as the 2D sensors providing 2D images.

During the past years many techniques and algorithms have been proposed to estimate the pose of faces from images. A good survey about the proposed techniques can be found in Murphy-Chutorian and Trivedi (2009). The majority of work in 3D face pose estimation deals with tracking full rigid body motion. This requires the estimation of 6 degrees of freedom of the face/head in every video frame. This can be successful for a limited range of motion

* Corresponding author at: University of the Basque Country UPV/EHU, Manuel Lardizabal, 1 20018 San Sebastian, Spain.

http://dx.doi.org/10.1016/j.engappai.2016.01.035
0952-1976/

(typically $\pm 45^{\circ}$ out-of-plane) and only for relatively high resolution images (Whitehill and Movellan, 2008). Such systems typically rely on a 3D model that should be fitted to the person specific shape (Dornaika and Davoine, 2006; Unzueta et al., 2014). There is a tradeoff between the complexity of the initialization process, the speed of the algorithm and the robustness and accuracy of pose estimation. Although the model-based systems can run in real-time, they rely on frame-to-frame estimation and hence are sensitive to drift and require relatively slow and non-jerky motion. These systems require initialization and failure recovery. For situations in which the subject and camera are separated by more than a few feet, full rigid body motion tracking of fine head pose is no longer practical. In this case, model-less pose estimation can be used (Guo et al., 2008; Aghajanian and Prince, 2009). This approach can be performed on a single image at any time without any model given that some pose-classified ground truth data are previously learned (Fu and Huang, 2006; Ma et al., 2006a).

Geometric methods (Horprasert et al., 1996; Wang and Sung, 2007) rely heavily on the estimation of facial features, such as eyes, mouth corners, nose tip, etc. and use their relative position to estimate the pose using projective geometry. For example, if the eyes and the mouth form an isosceles triangle, then the image corresponds to a frontal view. The major disadvantage of these methods is to locate the features needed for estimation in a very precise and accurate way. They also need to handle missing facial features in some poses. Appearance template methods use similarity algorithms and compare a given image to a set of exemplars in order to discover the most similar image (Beymer, 1993; Niyogi and Freeman, 1996). Nevertheless, even if these methods have the advantage of not requiring a feature extraction step, they may suffer from noise caused by illumination and expression changes in addition to the need of high computational power since the matching process they use is based on pair-wise similarities. Classification-based methods (Huang et al., 1998) operate by training head pose classifiers through the distribution of the training images into a set of discrete head poses. However, both the appearance template and classification-based methods can only return discrete poses and are sensitive to non-uniform sampling in the training data. Regression-based methods (Ma et al., 2006b) allow us to obtain continuous pose estimates. Indeed, they use regression techniques (Drucker et al., 1996; Vinzi et al., 2008) in order to find the relationship between the face image and its corresponding pose and to earn continuous mapping functions between the face image and the pose space. The high dimensionality of the data represents an important challenge in this kind of methods because of the well-known “curse of dimensionality” problem (Bishop, 2006).

Many researchers use a dimensionality reduction step before the regression (Chun and Keles, 2007; Nilsson et al., 2007). The main disadvantage of these methods is that their performance deteriorates with bad head localization. The manifold embedding methods (Balasubramanian et al., 2007) consider face images as samples of a low-dimensional manifold embedded in the high-dimensional observation space (the space of all possible images). They try to find a low dimensional representation that is linked to the pose. After that, classification or regression techniques are applied to discover the pose. The main weakness of manifold embedding methods is that appearance variation is not only affected by pose changes but also by other factors such as lighting changes and identity.

As can be seen, manifold learning and machine learning approaches are the only tools that can solve the face estimation for the featureless and model-less cases. For example, the Synchronized Submanifold Embedding (SSE) was proposed in Yan et al. (2009) in order to project face images. This proposed algorithm is dually supervised by both identity and pose information. The submanifold of each subject is approximated as a set of simplexes constructed using neighboring samples, and the pose label is further propagated within all the simplexes by using the generalized barycentric coordinates. Then these submanifolds are synchronized by seeking the counterpart point of each sample within the simplexes of a different subject, and consequently the Synchronized Submanifold Embedding is formulated to minimize the distances between these aligned point pairs and at the same time maximize the intra-submanifold variance. Finally, for a new datum, a simplex is constructed using its nearest neighbors measured in the dimensionality reduced feature space, and then its pose is estimated as the propagated pose of the nearest point within the simplex.

1.1. Letter contribution

In this letter, the authors propose a new embedding scheme for image-based continuous 3D face pose estimation. The main contributions are as follows. First, the authors show that the concept of label sensitive Locality Preserving Projections, proposed for age estimation, can be used for modeless face pose estimation. Second, the authors provide a linear embedding by exploiting the connections between facial features and pose labels via a sparse coding scheme. The resulting technique is called Sparse Label Sensitive Locality Preserving Projections (Sp-LsLPP). Third, for enhancing the discrimination between poses, the projections obtained by the Sparse Label sensitive Locality Preserving Projections are fed to a Discriminant Embedding that exploit the continuous labels. The authors stress the fact that the proposed embedding method can be useful for many real-world problems for which data have continuous labels such as age estimation, face attractiveness scoring and facial emotion scoring. The letter is organized as follows. Section 2 reviews the linear manifold learning techniques as well as the label sensitive Locality Preserving Projections method. Section 3 presents the proposed framework. Section 4 presents a performance evaluation on two real face image datasets. Section 5 provides some concluding remarks.

2. Manifold learning: related work

The classic linear embedding methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) (Fukunaga, 1990), and Maximum Margin Criterion (MMC) (Li et al., 2006) are proved to be computationally efficient and suitable for practical applications, such as pattern classification and visual recognition. PCA projects the samples along the directions of maximal variances and aims to preserve the Euclidean distances between the samples. Unlike PCA which is unsupervised, LDA (Fukunaga, 1990) is a supervised technique. One limitation of PCA and LDA is that they only see the linear global Euclidean structure. In addition to the Linear Discriminant Analysis (LDA) technique and its variants (Dai and Yuen, 2007; Fukunaga, 1990), there is recently a lot of interest in graph-based linear dimensionality reduction. Many dimensionality reduction techniques can be derived from a graph whose nodes represent the data samples and whose edges quantify the similarity among pairs of samples (Sugiyama, 2007; Yan et al., 2007). Recent proposed methods attempt to linearize some non-linear embedding techniques. This linearization is obtained by forcing the mapping to be explicit, i.e., performing the mapping by a projection matrix. For example, Locality Preserving Projection (LPP) (Xu et al., 2010) and Neighborhood Preserving Embedding (NPE) (He et al., 2005) can be seen as linearized versions of Laplacian Eigenmaps (LE) and Local Linear Embedding (LLE), respectively. The main advantage of the