Neurocomputing 173 (2016) 42–53

Contents lists available at ScienceDirect

Neurocomputing

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

Robust head pose estimation using Dirichlet-tree distribution enhanced random forests ☆

Yuanyuan Liu [a,b,c], Jingying Chen [a,b,*], Zhiming Su [a,b], Zhenzhen Luo [a,b], Nan Luo [a], Leyuan Liu [a,b], Kun Zhang [a,b]

[a] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China
[b] Collaborative & Innovative Center for Educational Technology (CICET), China
[c] Wenhua College, Wuhan, China

ARTICLE INFO

Article history:
Received 16 June 2014
Received in revised form
10 March 2015
Accepted 23 March 2015
Available online 4 August 2015

Keywords:
D-RF
HPE
Combined texture
Geometric features
Patch classification
Composite weighted voting

ABSTRACT

Head pose estimation (HPE) is important in human–machine interfaces. However, various illumination, occlusion, low image resolution and wide scene make the estimation task difficult. Hence, a Dirichlet-tree distribution enhanced Random Forests approach (D-RF) is proposed in this paper to estimate head pose efficiently and robustly in unconstrained environment. First, positive/negative facial patch is classified to eliminate influence of noise and occlusion. Then, the D-RF is proposed to estimate the head pose in a coarse-to-fine way using more powerful combined texture and geometric features of the classified positive patches. Furthermore, multiple probabilistic models have been learned in the leaves of the D-RF and a composite weighted voting method is introduced to improve the discrimination capability of the approach. Experiments have been done on three standard databases including two public databases and our lab database with head pose spanning from -90° to 90° in vertical and horizontal directions under various conditions, the average accuracy rate reaches 76.2% with 25 classes. The proposed approach has also been evaluated with the low resolution database collected from an overhead camera in a classroom, the average accuracy rate reaches 80.5% with 15 classes. The encouraging results suggest a strong potential for head pose and attention estimation in unconstrained environment.

1. Introduction

Real-time, robust HPE algorithms are very important and an active research topic in computer vision as [1,2]. Knowing human's head poses can provide important cues concerning visual focus of attention and analyzing human's behavior. Also, head pose is crucial to applications like video surveillance, intelligent environments, human machine interfaces and affection recognition [3–6]. Due to its practical signification and challenges, there is a fair amount of work developed fast and reliable algorithms for head pose estimation. However, most of the work has reported good results in constrained environment, the performance could be decreased due to the high variations in unconstrained environment, such as, facial appearance, poses, illumination, occlusion, expression and make-up. Hence, a Dirichlet-tree distribution enhanced Random Forests approach (D-RF) is proposed in this paper to estimate head pose efficiently and robustly in unconstrained environment.

Based on different features, several methods for the problem can be briefly divided into two categories, facial geometric feature and facial texture feature based methods. The methods based on facial geometric features usually require high image resolution for facial feature identification, such as eyes, eyebrows nose or lips [7–9]. These methods can provide accurate estimation results relying on accurate detection of facial feature points and high quality images. Other based on facial texture approaches usually use texture feature from an entire face to estimate head pose [10–13]. It may be good for dealing with low resolution image but not robust to occlusion. In the real life scene, the various illumination, occlusion, low image resolution and wide scene make the head pose estimation difficult. In order to estimate head pose in unconstrained environment, we address the problem based on combined geometric and texture features.

More recently, classification and regression are very popular methods for head pose estimation on low resolution images such as neural networks (NN) [14], support vector machines (SVM) [15,16], nearest prototype matching [7] or random forests [17,10,8,18]. Gourier et al. [14] used an auto-associative network to learn the mapping for

☆Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author at: National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China.
E-mail addresses: jane19840701@hotmail.com (Y. Liu), chenjy@mail.ccnu.edu.cn (J. Chen), happyszm@foxmail.com (Z. Su), 13720269596@139.com (Z. Luo), luonanccnu@hotmail.com (N. Luo), lyliu@email.ccnu.edu.cn (L. Liu), zhk@mail.ccnu.edu.cn (K. Zhang).

http://dx.doi.org/10.1016/j.neucom.2015.03.096
0925-2312/

head pose estimation on low resolution images. A simple winner-takes-all process was applied to select the head pose which prototype gives the best match in NN. They achieved a precision of 10.3 degrees in the yaw angle and 15.9° in the pitch angle only on the Pointing'04 database [9]. Orozco et al. [16] trained a multi-class Support Vector Machine for pose classification in crowed scenes. The distance features of each pixel of a head to the mean appearance templates of head images at different poses have been proposed to train a multi-class SVM for head pose classification. The performance on crowd public space and low resolution videos reached 80% accurate rate in 4 head poses classification. In [7], Wu and Trivedi proposed a two stage framework for continuous head pose estimation based on a finer geometrical structure. In the first stage, coarse head poses were classified by nearest prototype matching method, and refined head poses were estimated with a complex geometrical structure in second stage. The total accuracy was 75.4% in yaw and pitch angles. Recently, multi-class random forests become a very popular method in the field owing to their capability to handle large training datasets, their high generalization power and speed, and the relative ease of implementation.

Random forests are a family of ensemble classifiers introduced by Breiman in 2001 [19], which can be used either for multi-class classification [17,10,13], regression [8,11], or even both at the same time [12,18]. Fanelli et al. [18] proposed regression random forests for real-time head pose estimation from depth cameras. They reached 89% accurate rate with head and nose tip successful localization in high quality depth images. Some works [8,20] showed the power of RF in mapping image features to votes in a generalized Hough space [21] or to real-valued functions. Random forests have been combined with the concept of Hough transform for object detection and action recognition. These methods use two objective functions for optimizing the classification and the Hough voting properties of the random forests. Huang et al. [13] proposed Gabor feature based multi-class random forest method for head pose estimation. In order to enhance the discriminative power, they employed LDA technique for node tests. The successful accuracy reached 89% in public high resolution databases. Dantone et al. [12] proposed conditional random forests to estimate head pose under various conditions only in the horizontal direction. They used prior knowledge of some global variable to constrain output. In this case the global variable was the orientation of the head, divided into 5 classes. The accuracy rate reached 72.3% with five head pose classes in the wild database. Hence, head pose estimation in the wild and unconstrained environment is still a challenge and a significant problem.

To improve the accuracy and efficiency in the wild and unconstrained environment, a Dirichlet-tree distribution algorithm is introduced into random forest framework to estimate head pose in this paper. The idea of the paper is to use prior knowledge of some global variable to constrain output based Dirichlet-tree distribution. The Dirichlet-tree distribution was proposed by Minka [22]. It is the distribution over leaf probabilities that result from the prior on branch probabilities. Minka proved the high accuracy and efficiency of the distribution. Some researchers use a Dirichlet-tree distribution in multi-objects tracking [23] facial feature detection [24] and affective computing [25]. In this work, D-RF is proposed to estimate head poses in a coarse to fine way in various and unconstrained environment.

This paper is an extension of a paper presented at conference [10]. The main and different contributions from the conference paper are as follows. First, in order to improve classification, more powerful combined texture and geometric features (i.e., Gabor feature-based PCA, Sobel, LBPH and two geometric features) from positive facial patches are extracted to estimate head pose with D-RF, instead of only the texture features (Gabor feature-based PCA and gray values) used in the ICPRAM paper. Second, in the previous ICRPAM paper, single probabilistic model has been learned in leaves of the D-RF and a GMM method was used to vote the leaves. In this paper, multiple probabilistic models (i.e., head pose angles and two geometric offset vectors) have been learned in leaves of the new D-RF, and a composite weighted voting method that is composited of the classification and regression voting measures is introduced into probabilities voting to improve the discrimination capability of the approach. Third, an additive confidence parameter *pf* in the composite weighted voting method has been used to control the number of positive patches through geometric offset vectors stored at leaves, which can be used to eliminate the influence due to face deformation and wide range head poses. Finally, more detailed experiments have been done on three standard databases and our low resolution database collected from an overhead camera in a classroom, the average accuracy rate reaches 76.2% with 25 classes in standard databases and 80.5% with 15 classes in our collected low resolution database, respectively.

2. D-RF for head pose estimation

The flowchart of the proposed approach is given in Fig. 1. In the first stage, facial patches are extracted and classified to positive/negative patches from detected facial areas, and combined texture and geometric features from positive facial patches have been extracted. In the second stage, a more accurate D-RF approach with combined texture and geometric features is proposed based our previous work to estimate head pose in the horizontal and vertical directions. The proposed D-RF consists of four layers. D-L1 and D-L2 are two layers in the horizontal direction, D-L1 represents coarse classification while D-L2 is refined classification. In D-L2, the yaw angle has been estimated based on the classified result of D-L1. D-L3 and D-L4 are two layers in the vertical direction, D-L3 represents horizontal refined classification and vertical coarse classification, while D-L4 represents final refined classification in two freedom head poses. In each leaf of the D-RF, there are multiple probabilistic models including patch class probability, head poses and two geometric offset vectors. A composite weighted voting method is used to obtain final head pose parameter based multiple probabilistic models in leaves. Finally, the final head pose angles have been obtained in the D-L4 layer of D-RF. Details are given in the following.

2.1. Positive facial patch extraction

A facial area is first detected by Adaboost with Haar-like feature [26], which may include some noise for head pose estimation, such as hair, neck and occlusion. In order to eliminate noise, the facial area is segmented into foreground and background areas. The foreground areas include positive patches and negative patches, where the positive patches contribute to estimate head pose while the negative patches including occlusion or noise may introduce errors for the task.

To segment the background, the detected facial area which is normalized as 125∗125 pixels is divided into 6∗6 non-overlapping squares, and histogram distributions of the squares are computed as shown in Fig. 2. We analyze the uniformity of histogram distributions of the patches and segment most of the background patches.

200 patches are randomly extracted from the rest of facial area with background removed, which include positive and negative facial patches. The positive and negative patches are classified using RF [17,19]. In order to model the random tree, positive facial patches are labeled as 1 and the negative facial patches are labeled as 0. A tree T grows up based on Gabor features and histogram of the labeled patches. The training and testing are similar to RF