Pattern Recognition Letters 45 (2014) 145–153

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# Head pose estimation using image abstraction and local directional quaternary patterns for multiclass classification ☆

ByungOk Han [1], Suwon Lee [1], Hyun S. Yang *

*Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 29 August 2013
Available online 6 April 2014

*Keywords:*
Head pose estimation
Image abstraction
Multiclass classification

ABSTRACT

This study treats the problem of coarse head pose estimation from a facial image as a multiclass classification problem. Head pose estimation continues to be a challenge for computer vision systems because extraneous characteristics and factors that lack pose information can change the pixel values in facial images. Thus, to ensure robustness against variations in identity, illumination conditions, and facial expressions, we propose an image abstraction method and a new representation method (local directional quaternary patterns, LDQP), which can remove unnecessary information and highlight important information during facial pose classification. We verified the efficacy of the proposed methods in experiments, which demonstrated its effectiveness and robustness against different types of variation in the input images.



## 1. Introduction

The natural interaction between people and computers is an important research topic, which has recently attracted considerable attention. This research area addresses natural user interfaces (NUI) between humans and computers. A NUI is a human–machine interface that does not employ input devices. NUI is a natural interaction method that resembles communication between people. Various related research areas aim to achieve non-intrusive and natural human–computer interaction (HCI), such as face recognition, facial expression recognition, activity recognition, and gesture recognition. As a starting point of these techniques, head pose estimation is a crucial technology that aims to predict human intentions to facilitate the use of non-verbal cues for communication via NUIs. Thus, people can estimate the orientation of another persons head to understand whether they want to interact with them.

Head pose estimation is a technique that aims to determine three-dimensional (3D) orientation properties from an image of a human head. In 3D space, objects have geometric properties with six degrees of freedom for rigid body motion, i.e., three rotations and three translation vectors. Head pose estimation methods are usually designed to extract head angular information in terms of the pitch and yaw rotations of a facial image. The pitch and yaw are more difficult to estimate than other properties, such as the roll angle, two-dimensional (2D) translation, and scale, which can be calculated easily using 2D face detection techniques, because of occlusions by features such as glasses, beards, hair, and head angle changes. Different identity, illumination, and facial expression conditions are also serious hindrances when extracting the angular properties of head images.

## 2. Related works

In recent years, many methods have been developed for estimating 3D human head poses using facial RGB images. These studies can be categorized into methods based on classification and those based on regression with machine learning techniques. These methods simply aim to determine whether the pose space is discrete or continuous. The advantages of these classification approaches are that they are comparatively simple and control pose training datasets can be used for training sessions. These methods can then be expanded to a larger pose set by users at any time [1]. Furthermore, the training dataset only requires human head images with corresponding labels for the head angle information. However, this approach can only estimate designated and discrete head poses. By contrast, the regression methods used for human head pose estimation can obtain continuous information related to head poses. However, it is difficult to develop an exact function for robust head pose estimation because of the complexity of the non-linear and linear mappings that connect the facial images and pose labels. From an image representation perspective, head

☆ This paper has been recommended for acceptance by S. Wang.

* Corresponding author. Tel.: +82 42 350 7727; fax: +82 42 867 3567.
*E-mail address:* yang@paradise.kaist.ac.kr (H.S. Yang).
[1] Tel.: +82 42 350 7727; fax: +82 42 867 3567.

http://dx.doi.org/10.1016/j.patrec.2014.03.017
0167-8655/

pose estimation can be divided into two categories: appearance-based methods and geometric feature-based methods. Hence, methods can be classified based on the characteristics of the description vectors used for training. Appearance-based approaches obtain texture information from a facial image, whereas geometric feature-based approaches manipulate positional information related to the facial features, such as the eyes, eyebrows, nose, and mouth. The first method exploits pixel values in the actual facial image, so it is greatly affected by various factors that change images and it is necessary to employ an effective noise removal algorithm, such as illumination normalization or face alignment. The second method finds facial features using model-based algorithms such as the active shape model [2], the active appearance model [3], or the constrained local model [4]. Feature vectors can be generated from location information using several facial feature detectors, which are trained using another training set to facilitate 2D location detection. The feature vectors produced from positional information can be used as a supervised learning framework. They can also be employed directly to estimate facial poses. For example, a triangle obtained from three points, i.e., two eyes and a nose, can be used for pose estimation simply by calculating a triangle projected onto the image plane. However, the geometric approach has the disadvantage that the facial feature locations in all images must be labeled manually to generate the training dataset. However, this is an intuitive method for estimating head poses because it uses location information.

The approach we develop in the present study is based on concepts derived from multiclass classification and an appearance-based method. To compress useful visual content and to remove unnecessary information, we propose a new approach based on image abstraction. Image abstraction was originally developed for artistic purposes based on automatic stylization. It was also used to communicate information in a previous study [5]. Thus, image abstraction can provide important perceptual information during the recognition process, by simplifying and compressing the visual content. A related study [6] applied image abstraction methods to coarse head pose estimation algorithms, which were simple and accurate. To the best of our knowledge, this is the first study to apply image abstraction to head pose estimation. However, they only considered an estimation process based on various head poses in terms of variations in identity and they did not consider variations in illumination or facial expressions, whereas we aimed to develop a technique that was robust to such variations. Moreover, they only explain about a binary image produced by their image abstraction algorithm. To ensure that the method is more accurate and more robust to variations in the images, we propose a novel binary representation method, referred to as local directional quaternary patterns (LDQP), which describes the binary images produced by our image abstraction method. In this study, we extend our previous research using an image abstraction method, which generates a facial sketch image from a contour image with a cartoon-like effect [7]. We explain our new facial sketch image representation method, and we evaluate the effectiveness of the image abstraction method and representation methods in various experiments. The remainder of this paper is organized as follows. Section 3 provides an overview of the framework of our system. Section 4 presents the details of our image abstraction method and Section 5 explains our representation method. Our research results are given in Section 6 and we present our conclusions in Section 7.

## 3. System overview

Humans can recognize head poses by detecting simple sets of edges, similar to cartoon faces. People are capable of identifying simple head poses because they can abstract the features of faces intuitively. In particular, people innately recognize the shapes, configurations, or contours of trained features such as eyes, noses, mouths, eyebrows, foreheads, and chins. Thus, people can remember abstracted images of heads by inference from trained data. The basic concept of our system was designed and implemented from this perspective (Fig. 1).

To classify facial poses, it is necessary to train a classifier with facial images and their corresponding pose labels using a facial pose database created during the training session. First, Viola–Jones face detection algorithm detects coarse frontal and profile faces from images in the facial pose database. If the image does not contain a face, it is removed from the training data. The facial images are normalized after the exception handling process and the image abstraction and the representation processes are applied. A classifier is trained for multiclass classification using the binary images produced by the image abstraction algorithm. After the training session, a facial test image is vectorized using similar methods to those employed in the previous training session and the output pose is estimated.

## 4. Image abstraction

Image abstraction removes unnecessary information and emphasizes the main contents by reinterpreting scene information. This process can help viewers to capture specific visual information. We use an image abstraction method to interpret facial images. Our image abstraction method is shown in Fig. 2. The proposed algorithm performs GrabCut segmentation [8] using the rectangular area of a face. Next, to generate a cartoon-like effect, bilateral filtering [9] is applied to remove some of the noise, after which a difference of Gaussian (DoG) method [10] extracts the contours from the face image.

### 4.1. GrabCut algorithm

To obtain a cartoon-like representation of a face, it is necessary to extract the face region. Thus, face region and background region segmentation are required for image abstraction. A rough rectangular region containing the face can be obtained using a face detection algorithm. However, the rectangular region generated by the face detection algorithm still includes background pixels, which need to be separated from the facial image. Thus, another algorithm is required to obtain more precise results. In this case, we use the GrabCut algorithm, which segments the foreground from the background using some of the pixels in an image. The GrabCut algorithm only requires that some of the pixels are labeled as foreground or background pixels. This algorithm is widely used for extracting the foreground by partial labeling. The only input required by the algorithm is a rough segmentation of the foreground and the background. This can be achieved using the rectangular region produced by the face detection algorithm. The inputs of the GrabCut algorithm are the face region that corresponds to a target object, $R_f$, the background region, $R_{bg}$, and the unknown region, $R_u$. The algorithm aims to separate the face region ($R_f$) from the unknown region ($R_u$) using information from the background region ($R_{bg}$), as shown in Fig. 3. This is achieved using the Graph Cut algorithm [11] and a Gaussian mixture model (GMM).

The basic procedure of the GrabCut algorithm is as follows:

(i) A user input is obtained that contains a face region ($R_f$), a background region ($R_{bg}$), and an unknown region ($R_u$). The unknown region can contain face or background information.
(ii) The pixels in the background region and the face region are modeled using the GMM.