Pattern Recognition Letters 112 (2018) 241–248

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Extracting discriminative features using task-oriented gaze maps measured from observers for personal attribute classification

Masashi Nishiyama[a,b,*], Riku Matsumoto[a], Hiroki Yoshimura[a], Yoshio Iwai[a,b]

[a] Graduate School of Engineering, Tottori University, 101 Minami 4-chome, Koyama-cho, Tottori 680-8550, Japan
[b] Cross-informatics Research Center, Tottori University, 101 Minami 4-chome, Koyama-cho, Koyama-cho, Tottori 680-8550, Japan

ARTICLE INFO

Article history:
Received 23 February 2018
Available online 2 August 2018

MSC:
68T45
68U10
68T99
68U99

Keywords:
Gaze map
Feature extraction
Personal attributes

ABSTRACT

We discuss how to reveal and use the gaze locations of observers who view pedestrian images for personal attribute classification. Observers look at informative regions when attempting to classify the attributes of pedestrians in images. Thus, we hypothesize that the regions in which observers' gaze locations are clustered will contain discriminative features for the classifiers of personal attributes. Our method acquires the distribution of gaze locations from several observers while they perform the task of manually classifying each personal attribute. We term this distribution a task-oriented gaze map. To extract discriminative features, we assign large weights to the region with a cluster of gaze locations in the task-oriented gaze map. In our experiments, observers mainly looked at different regions of body parts when classifying each personal attribute. Furthermore, our experiments show that the gaze-based feature extraction method significantly improved the performance of personal attribute classification when combined with a convolutional neural network or metric learning technique.

1. Introduction

Personal attributes such as gender, clothing, and carried items, which are of interest in the field of soft-biometrics [6,7,27,32], help the collection of statistical data about people in public spaces. Furthermore, personal attributes have many potential applications, such as video surveillance and consumer behavior analysis. In general, pedestrians captured on video or in still images are used for personal attribute classification. Researchers have proposed several methods for automatically classifying personal attributes in pedestrian images; for example, techniques involving convolutional neural networks (CNNs) [22,25,29,30] and metric learning [21,41] have been proposed. The existing methods can extract discriminative features for personal attribute classification and obtain high accuracy when many training samples containing diverse pedestrian images are acquired in advance. However, the collection of a sufficient number of training samples is very time consuming. Unfortunately, the performance of the existing methods has been found to decrease when the number of training samples is small.

People correctly and quickly classify personal attributes. We believe that people have the visual ability to extract features from an individual. For instance, people correctly classify gender from facial images [3,4]. In the research field of cognitive science, Yarbus [38] reported that human observers can recognize personal attributes in a scene image with high accuracy when they are given different tasks such as remembering the clothes worn by the individuals or estimating their ages. In this interesting research, he noticed that the observers paid attention to different regions in the scene when they tackled a different task even though they viewed the same image. Recently, researchers have made some efforts to analyze the role of task in various applications [13,14,19]. Based on these observations, we hypothesize that people pay attention to different informative regions in pedestrian images while tackling various tasks of personal attribute classification.

It may be possible to reproduce human visual abilities via an algorithm on a computer with a small number of training samples such that the classification performance is equivalent to that of humans. With respect to object recognition, several existing methods for mimicking human visual abilities have been proposed [12,33,40]. To mimic human visual ability, the existing methods exploited a saliency map computed from low-level features in a given image using techniques such as those described in [17,39,42]. However, the use of the saliency map does not sufficiently represent human visual abilities because of the deep mechanisms of human vision.

An increasing number of pattern recognition studies, specifically those attempting to mimic human visual ability, have mea-

* Corresponding author at: Graduate School of Engineering, Tottori University, 101 Minami 4-chome, Koyama-cho, Tottori, 680-8550, Japan
E-mail addresses: nishiyama@tottori-u.ac.jp, masashi1@m.ieice.org (M. Nishiyama).

https://doi.org/10.1016/j.patrec.2018.08.001
0167-8655/

sured the gaze locations of observers [11,18,31,36,37]. These gaze locations have great potential for the collection of informative features during various recognition tasks. Very recently, state-of-the-art techniques [26,28] have demonstrated that gaze locations can help to extract informative features for the attribute classification of fashion clothing and face images. However, these existing methods do not consider how to treat the case in which observers tackle different tasks for body attributes in the same pedestrian image. We believe that the informative region of the body for each classifier is significantly different for each task of personal attribute classification.

In this paper, we consider the challenging case in which participants in an experiment are given different tasks of personal attribute classification while viewing the same pedestrian images. We confirm whether or not test participants look at different regions when tackling each task. We determine whether or not the gaze locations measured from the participants play an important role in the personal attribute classification. To this end, we generated a task-oriented gaze map from the distribution of gaze locations recorded while participants viewed images to complete each task of manually classifying personal attributes. The high values in a task-oriented gaze map correspond to regions that are frequently viewed by participants. We assume that these regions contain discriminative features for each classifier of a personal attribute because they appear to be useful when the participants are tackling each task of personal attribute classification. When extracting features to learn the classifier, larger weights are given to the regions of the pedestrian images that correspond to the attention regions of the task-oriented gaze maps. The experimental results indicate that our method improves the accuracy of feature extraction when using a CNN or metric learning technique with a small number of training samples.

This paper is organized as follows. Section 2 describes related work, Section 3 describes the generation of task-oriented gaze maps, and Section 4 describes feature extraction using the maps. Our concluding remarks are given in Section 5.

2. Related work

To mimic human visual ability, existing methods [12,33,40] involve the saliency maps of object images with representations of the regions that draw visual attention. Walther et al. [33] combined a recognition algorithm with a saliency map generated from low-level features of gradients of color and intensity using [17]. Researchers have developed techniques [12,40] that use the object labels of images in addition to the low-level features of objects to generate saliency maps. Furthermore, existing methods [39,42] add image boundary information in low-level features to generate saliency maps with high accuracy. However, the use of low-level features to generate a saliency map does not sufficiently represent human visual abilities. Our method exploits the use of gaze locations instead of a saliency map to increase the performance of personal attribute classification.

Existing methods [11,18,31,36,37] aim to design an algorithm that is close to the human visual ability by measuring gaze locations from observers. Xu et al. [37] generated saliency maps of facial images using prior gaze locations from participants who viewed the images. They reported that the generated saliency maps represented high-level features corresponding to the facial feature points of the eyes, nose, and mouth. Furthermore, gaze locations are used in applications involving action recognition or image preference estimation. Fathi et al. [11] classified actions by simultaneously inferring regions where gaze locations were gathered via an egocentric camera. Xu et al. [36] demonstrated that the use of gaze tracking information (such as fixation and saccade) significantly helps the task of egocentric video summarization. Sugano et al. [31] estimated more highly preferable images using gaze locations and low-level features. Karessli et al. [18] classified objects using only gaze features without object labels for zero-shot learning. Additionally, Sattar et al. [28] predicted the category and attribute of fashion clothing images by embedding gaze distributions in the pooling layers of a CNN. Murrugrra-Llerena et al. [26] classified the attributes of shoe and face images using a binary masking of gaze distributions. However, the existing methods do not consider the variation of gaze locations with respect to body regions when participants tackle several different tasks using pedestrian images. We attempt to observe the variation of gaze locations for different tasks of personal attribute classification. Based on the gaze locations measured with respect to body regions, we develop a method for extracting features to improve the performance of personal attribute classification.

3. Generating task-oriented gaze maps

3.1. Gaze locations in personal attribute classification

Here, we consider the regions of pedestrian images that are frequently looked at by observers when manually classifying personal attributes. For instance, Hsiao et al. [15] found that observers looked at a region around the nose when they identified individuals from a facial image. In the case of gender classification, we believe that the human face plays an important role. However, a pedestrian image contains not only a face but also a body. Yarbus [38] found that observers look at a different region in a scene image when tackling each task of personal attribute classification. However, he did not analyze the distributions of gaze locations in pedestrian images for personal attribute classification. Thus, we attempt to discern the regions of pedestrian images that tend to collect gaze locations from observers when given several different manual personal attributes classification tasks. Note that we assume that the alignment of the pedestrian images has already been completed using a pedestrian detection technique such as [9,16]. The details of our method are also described below.

3.2. Generation algorithm

To generate a task-oriented gaze map, we use a gaze tracker to acquire gaze locations while a test participant views a pedestrian image on a screen. We prepare T tasks, P participants, and N pedestrian images. Given a gaze location (x_f, y_f) in a certain frame f, gaze map $g_{t,p,n,f}(x, y)$ is labeled 1 when $x = x_f, y = y_f$; otherwise, it is labeled 0, where p is a participant, t is a task, and n is a pedestrian image. Note that the participant not only looks at point (x_f, y_f) on each pedestrian image, but also the region surrounding this point. Thus, we apply a Gaussian kernel to the measured gaze map $g_{t,p,n,f}(x, y)$. To determine the size k of the Gaussian kernel, we use the following equation:

$$k = \frac{2dh}{l} \tan \frac{\theta}{2}, \tag{1}$$

where d is the distance between the screen and the participant, θ is the angle of the region surrounding a measured gaze point, l is the vertical length of the screen, and h is the vertical resolution of the screen. Fig. 1 illustrates the parameters used to determine the kernel size. We assume that each pixel on the screen is square. We aggregate each $g_{t,p,n,f}(x, y)$ to $g_{t,p,n}(x, y)$ to represent the distribution of gaze locations in a certain pedestrian image as

$$g_{t,p,n}(x, y) = \sum_{f=1}^{F_{t,p,n}} k(u, v) * g_{t,p,n,f}(x, y), \tag{2}$$

where $F_{t,p,n}$ is the time taken to classify personal attribute by a participant, $*$ is the convolution operator, and $k(u, v)$ is a Gaus-