Neurocomputing 133 (2014) 404–415

Contents lists available at ScienceDirect

Neurocomputing

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

Saliency-guided improvement for hand posture detection and recognition

Yuelong Chuang, Ling Chen*, Gencai Chen

College of Computer Science, Zhejiang University, P.R. China

A R T I C L E I N F O

Article history:
Received 13 May 2013
Received in revised form
4 November 2013
Accepted 2 December 2013
Communicated by Qi Li
Available online 10 January 2014

Keywords:
Hand posture detection and recognition
Image saliency
Skin model

A B S T R A C T

To detect and recognise hand postures against complex backgrounds, we propose a novel model that is constructed by the integration of image saliency and skin information. Although a skin model is a simple and efficient strategy by which to locate skin regions within images, it is easily affected by complex backgrounds, e.g. skin-like background regions and various lighting conditions. To solve this problem, we propose a general image saliency detection method that is then integrated with skin information to improve the performance of hand posture detection. Lastly, a linear Support Vector Machine (SVM) is adopted to recognise hand postures according to the results of hand posture detection. In the experiment, we tested the performance of the proposed image saliency detection method over seven state-of-the-art methods. The saliency-based hand posture detection and recognition model was also evaluated. These experiments show that the proposed model has stable performance for a wide range of images.

1. Introduction

As one of the most important parts of the human body, hands provide the most important means for non-verbal interaction among people. Because hand posture is a natural and intuitive interaction modality, hand posture recognition can enable a wide range of applications in sign language recognition, human–computer interaction (HCI), human–robot interaction (HRI), and virtual reality (VR).

Various methods have been proposed to recognise hand postures in recent years. The first category is based on extra instruments [1] (e.g., gloves, magnetic sensors, acoustic or inertial trackers). Generally, data collected by extra instruments is more accurate than that by other methods. However, many physical restrictions would make users feel uncomfortable. In addition, the price of extra instruments is relatively high, which influences the popularity seriously. In contrast, the second category, namely, a vision-based approach, uses only camera(s) to capture a human hand's information without any physical restrictions. Because of its potential for providing more natural, unencumbered, and non-contact interaction, a vision-based approach is identified as a promising alternative to methods based on extra instruments. However, the presence of complex backgrounds makes it difficult for the recognition of hand postures.

The human vision system (HVS) can rapidly and effortlessly recognise objects within cluttered, natural scenes, which inspires the development of computational models of biological vision systems [2–4]. One of the most important abilities of the HVS is that it can serve as a filter for selecting only the interesting information related to current tasks for further processing, while ignoring irrelevant information. To alleviate the effect of backgrounds on vision-based methods, we propose a cognitive visual attention method to detect saliency information within images, which is then integrated with skin information to improve the performance of hand posture detection and recognition. The contributions of this study are as follows:

1. An isophote-based operator is adopted to detect potential structure and global saliency information related to each pixel;
2. An integral image-based operator is employed to compute local saliency information based on detected potential structures for three types of feature channels (colour, intensity, and orientation), which can avoid the problem of information loss existed in fusing processes across multiple channels;
3. The final saliency map is built up based on global and local saliency information without any prior knowledge or higher-level information;
4. The detected bottom-up salient information is integrated with top-down skin information to locate hand regions against complex backgrounds.

The remainder of this manuscript is organised as follows. Section 2 presents related work on image saliency detection and hand posture detection and recognition; Section 3 describes the framework of the proposed model, which includes two

* Corresponding author. Tel.: +86 13606527774.
E-mail address: Lingchen@zju.edu.cn (L. Chen).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.023

components: image saliency detection and saliency-based hand posture detection and recognition. The performance evaluation is given in Section 4, and brief conclusions are presented in Section 5.

2. Related studies

2.1. Image saliency detection

Generally, image saliency detection methods could be separated into two types: psychological theory based methods and mathematical methods.

According to findings from neurobiology and psychophysics, image saliency can be defined as regions that can be easily differentiated from their surroundings, these differentiators being colour, orientation, and intensity [4]. The key issues in detecting image saliency include the following: (1) how to determine the position and size for both the centre regions and neighbouring regions, and (2) how to compute differences between centre regions and adjacent regions for the three feature channels (colour, intensity, and orientation). Itti et al. [4] solve the above problems by building image pyramids and subtracting different scales of the image pyramid to determine centre–surround contrast. Frintrop et al. [5] build image pyramids by integral filters [6] and compute the centre–surround contrast by exhaustively searching at every level of each pyramid. To speed up the calculation, the integral image mechanism is employed. However, in [4,5], different methods are employed to compute saliency information for each feature channel, resulting in the problem that the fusion of the different feature channels with non-comparable properties is somewhat arbitrary. Instead of building image pyramids, Achanta et al. [7] build up the saliency map by filtering the original image in a raster scan fashion. However, determining reasonable values for both the centre and surrounding regions is rather difficult to solve, and the performance would be easily affected by the type of background in the image and the noise level.

In recent years, several mathematical methods have been proposed to compute image saliency. Achanta et al. [8] propose a frequency-tuned method to compute pixel saliency directly. Since the method considers only the first-order average colour, it could be insufficient to analyse the complicated situations that frequently occur in images with natural surroundings. Valenti et al. [9] adopt an isocentric feature approach to representing image saliency in a global manner. Since the isocentric feature is seriously influenced by the type of background in the image and the noise level, the results of [9] might contain a substantial amount of irrelevant information. Hou and Zhang [10] propose a novel method for detecting image saliency by exploring spectral components in an image. This method is very fast, but because it is based on global considerations, detailed information regarding salient objects could be overlooked. Schauerte et al. [11,12] propose a spectral-based image saliency detection for eye fixation prediction. However the method should define an appropriate colour/feature space.

2.2. Computer vision-based hand posture detection and recognition

It is natural to employ a skin model to detect and recognise human hand postures. Because of their rapid implementation, skin models have been popular in recent years [13–16]. However, there are several drawbacks to skin models. It is clear that different ethnicities have different skin features, which make skin-based hand detection difficult. In addition, skin models are sensitive to lighting conditions, for example, a skin region might represent different features under different lighting conditions. Lastly, skin-like regions that exist in backgrounds also affect the performance of skin models. Recently, Pisharady et al. [16] proposed a skin-based method to detect hand areas within images. The skin map computed by the proposed model could enhance the edges and shapes within the skin-coloured regions in images, which is used to extract texture and shape-based features by biologically inspired feature-extracting methods. To solve the drawbacks of skin models, these investigators introduce prototype patches extracted from training images in the process of detecting texture-shape features. Lastly, hand posture detection is implemented by the integration of texture-shape and skin features.

Except for skin-based methods, there are various methods proposed to detect and recognise hand postures. Kolsch and Turk [17] recognise hand postures based on the object recognition method proposed by Viola and Jones [6]. The method is applied for the recognition of six hand postures. Bretzner et al. [18] recognise hand postures by using a hierarchical hand model, which consists of the palm and the fingers, and a multi-scale colour feature is introduced to represent hand shape. Ong and Bowden [19] propose a boosted classifier tree-based method for hand detection and recognition. The method includes two layers: the top layer for hand detection and the branches in the bottom layer for hand shape classification. In [20], view-independent recognition of hand postures is considered. The suitability of a number of classifiers is investigated to make the method view independent. The study combine supervised and unsupervised learning strategies to propose a learning approach called Discriminant-EM (D-EM). Triesch and Malsburg [21] address the complex background problem in hand posture recognition using elastic graph matching. However, the accuracy of the method is only 85.8% within their hand posture database. Flores et al. [22] use a self-organizing neural network: its topology determines hand postures, and its adaptation dynamic through time determines hand gestures. To reduce the complexity of recognising hand postures, several restrictions are enforced on backgrounds: Kim and Fellner [23] detect fingertip locations under a black-light environment; Chang [24] and Chang et al. [25] recognise hand postures under a simple and black environment; Bhuyan et al. [26] propose a recognition method under a lighting background environment.

3. The framework of hand posture detection and recognition

As described in Section 2, skin models face problems when detecting hand postures. Pisharady et al. solved these problems by introducing prototype patches extracted from training images in the process of detecting texture-shape features, which could result in the fact that the model must be re-trained over a different dataset. In this section, we first propose a general image saliency detection method, which is then integrated with skin information to locate hand areas in a non-training manner. We then adopt a visual cortex-based feature extraction method [3] and a linear SVM to recognise hand postures according to the results of hand area detection. Fig. 1 shows the framework of the proposed hand posture detection and recognition model.

3.1. Image saliency detection

We propose an isophote-based method to detect the salient pixels in images. In our method, an isophote-based operator is employed to capture the potential structure and global saliency information related to each pixel. The potential structure is used to determine the centre–surround contrast that is then combined with global saliency to determine the final saliency information. Moreover, the integral image is employed to compute the centre–surround contrast, which is conducive to the fusion of all of the feature channel maps.