Neurocomputing 157 (2015) 296–305

Contents lists available at ScienceDirect

Neurocomputing

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

Robust hand tracking via novel multi-cue integration

Xiaoqin Zhang [a,*], Wei Li [b], Xiuzi Ye [a], Stephen Maybank [c]

[a] Lab of Innovative Design and Intelligent Computing, Wenzhou University, Wenzhou, China
[b] Taobao (China) Software Company Limited, Hangzhou, China
[c] Department of Computer Science and Information Systems, Birkbeck College, London, UK

ARTICLE INFO

Article history:
Received 1 October 2014
Received in revised form
15 December 2014
Accepted 9 January 2015
Communicated by Ran He
Available online 19 January 2015

Keywords:
Hand tracking
Multi-cue integration
Feature selection
RCD

ABSTRACT

In this paper, we present a robust real-time hand tracking system via multi-cue integration. In practice, the motion information of the hand, such as optical flow, is hard to exploit, because images of hands lack texture. As a result, the integration of the color and motion cues using conventional integration algorithms is difficult. Here, we integrate the motion and color cues from a novel feature point selection view. The hand is tracked using feature points, and the integration is realized during the feature points generation and selection process. In the generation process, a bounding box estimated by the color cue is used to provide a region for the feature points generation. Then, the *RCD* (*Representative*, *Compact* and *Diverse*) criteria are proposed to control the feature point selection process. After the selection process, the feature points are tracked using estimates of the motion of each feature point. The centroid of the feature points in each frame is adopted as the position of the hand. The experimental results show that our integration algorithm outperforms tracking algorithms that only use a single cue. Also the proposed tracking algorithm is more robust in complex environments than other state-of-art algorithms.

1. Introduction

Hand tracking is one of the most important tasks in human computer interaction [1,2]. It is the foundation of high level tasks such as hand gesture recognition. The successful understanding of hand gestures relies on the results of hand tracking.

The goal of hand tracking is to locate the position of a hand in an image sequence. Unlike other tracked objects such as human form, and face, the hand can undergo large deformations due to the joint movement, so the shape information of the hand is not reliable. Only color and motion cues are available and reliable for the hand tracking. Dorin et al. [3] adopt the color histogram as a feature for tracking an object, and use mean shift iterations to locate the position of the object. MOG (Mixture of Gaussians) [4–6] are applied to model the color distribution of the object. However, the color information is reliable only when the object color is discriminative against the background. If the color of the background is similar to the object or there are other objects with the similar color around the target object, then color based tracking algorithms will fail.

Motion is another useful cue for object tracking. Optical flow [7] which aims at estimating the translation of the pixels in the object is the most widely adopted motion cue. When calculating the optical flow, it is usually assumed that the brightness is constant for the corresponding pixels between consecutive frames and an image registration technique is used to estimate the motion of the pixels through the Newton–Raphson iteration [8]. However, not all the pixels of the object contain robust motion information. In [9,10], a corner detector is proposed to detect the points at which there are large changes in image intensity. An auto-correlation matrix is defined to find the corners. Shi and Tomasi [11] extend this idea to select the pixels which are suitable to tracking. Good feature points for tracking should meet two criteria: (1) the eigenvalues of the auto-correlation matrix should both be large; (2) the changes of the appearance between the first and the current frame are not significant. However, tracking with feature points is only available when a number of good feature points can be detected in the target region. Because the hand image is lack of texture, few feature points that satisfy the criterion (1) can be detected. As shown in Fig. 1(a), the points detected using the Shi and Tomasi's algorithm are too few in number to provide reliable tracking results. Also if the appearance of the hand changes, the brightness constancy assumption fails. As a result, a reliable process to generate and select new feature points is needed for hand tracking.

A single type of cue is not enough to provide reliable tracking results in noisy and challenging environments. So, multi-cue integration to make full use of all the cues is widely researched [12–17]. An integration scheme which assumes the different cues are conditionally independent is the most common method. Jochenet and Malsburg

* Corresponding author.
E-mail address: zhangxiaoiqnnan@gmail.com (X. Zhang).

http://dx.doi.org/10.1016/j.neucom.2015.01.002
0925-2312/

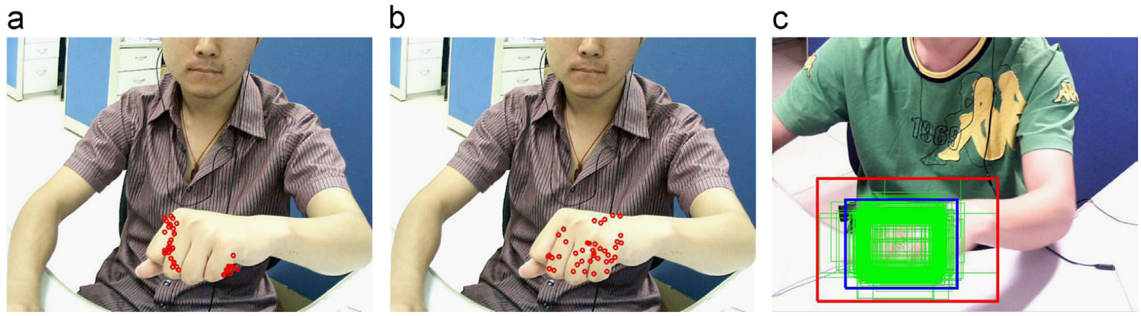

Fig. 1. (a) The feature points derived using Shi and Tomasi's algorithm. (b) The feature points derived using our method. (c) The small green bounding boxes are the particles generated in the current frame. The red bounding box is the smallest bounding box that contains all the particles. The blue rectangle is the tracking result. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

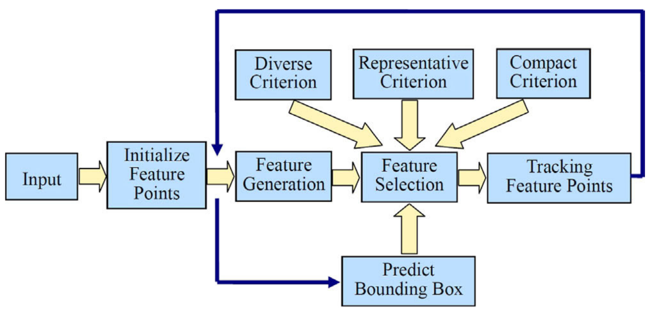

Fig. 2. The flowchart of our algorithm.

[12] propose an adaptive integration for different cues. The weights for each cue are adjusted according to the residual errors of different cues. To make the object cues supplementary to each other, cue dependence is introduced. In [13], Wu and Huang adopt a co-inference learning to make the cues interact with each other. The multi-cue fusion is realized during the sampling process within the particle filter framework. The particles generated for one cue are based on the particles with large weights obtained using another cue. The partition sampling technique introduced in [14,15] shares a similar idea with Wu and Huang's work. The integrations in these two papers are conducted in the sample generation process. Moreno-Noguer et al. [16] explore the dependence of the object cues during the measurement process. The integration is constructed by propagating through a cascade of Bayesian filters.

Although there is a vast number of papers for multi-cue integration, design of a robust hand tracking system through multi-cue integration is still an open and challenging problem. The main challenge in multi-cue integration is that the overall tracking accuracy is greatly reduced if the tracking performance with one cue is poor. Meanwhile, because the hand lacks texture and may change appearance during the tracking, the motion cue for a hand is hard to exploit. So the integration of motion and color cues using the algorithms discussed above is difficult. To overcome the above difficulty, we propose an integration mechanism based on feature point selection. We track the hand using feature points, and the optical flow which carries the motion cue is used to track the feature points. The feature generation and re-selection processes are important parts of the tracker. A color cue is used to guide these two processes. In our approach, color and motion cues are fused in a natural way without any necessity for tuning the relative weight of the different cues.

At the beginning of the tracking, suitable feature points are generated. However, the following problems are encountered: (1) the feature points generated in the first frame are too few for reliable tracking; (2) these points are concentrated in small regions of the image with high texture; (3) as the tracking proceeds, some of the feature points are lost because of changes in the appearance of the hand. More feature points need to be generated as the tracking proceeds. Feature selection criteria are required to select useful feature points and control their distribution.

Our integration is embodied in the generation and selection process. The color cue is integrated with the motion cue through the following steps. First, during the generation process a bounding box which provides a rough estimation of the hand region is applied to confine the locations of the new features points. The particle filter [18] which uses color cue is adopted to obtain the bounding box. Second, during the selection process, *RCD* (Representative, Compact and Diverse) criteria are proposed to select feature points from the set of initial feature points and the set of newly generated feature points. The *RCD* criteria make full use of the color cue to select the feature points which are suitable for the tracking. As shown in Fig. 1(b), the feature points selected using our criteria are diversely and compactly distributed in the hand area. After the selection process, we track the feature points using the constancy of pixel brightness. The motions of the points are estimated by minimizing the change in brightness between consecutive frames.

The rest of the paper is organized as follows. In Section 2, a flowchart is presented to give a clear view of our integration framework. Then different parts of our algorithm are detailed. The experimental results to validate our method are presented in Section 3, followed by a conclusion in Section 4.

2. Proposed algorithm

To avoid confusion, all the symbols used in this paper are chosen as follows: (1) lowercase letters for scalars $(x, y, z, \ldots)$; (2) bold lowercase letters for vectors $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$; (3) capital letters for matrices $(X, Y, Z, \ldots)$.

2.1. Overview

The flowchart of our algorithm is presented in Fig. 2. In the first frame, the tracking is initialized using a feature point detection algorithm. Since a hand image has little texture, few feature points are detected. A feature points generation process is conducted to obtain enough feature points for tracking. Then we select a subset of these feature points using our *RCD* criteria. The subset is highly representative and contain rich motion information. During the tracking process, some feature points are lost because of changes in the appearance of the hand. The feature generation and selection processes are conducted iteratively in each frame.