Neural Networks 81 (2016) 29–38

Contents lists available at ScienceDirect

Neural Networks

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

Real-time object tracking based on scale-invariant features employing bio-inspired hardware

Shinsuke Yasukawa [a], Hirotsugu Okuno [b], Kazuo Ishii [a], Tetsuya Yagi [b],*

[a] Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4, Hibikino, Wakamatsu, Fukuoka, 808-0196, Japan

[b] Graduate School of Engineering, Osaka University, 2-1, Yamadaoka, Suita, Osaka, 565-0871, Japan

HIGHLIGHTS

- We developed a real-time vision system with analog/digital mixed architecture.
- The system consists of an analog MOS transistor resistive network (RN) and an FPGA.
- The RN conducts multi-scale filtering in real time with a low power consumption.
- The FPGA finds scale-invariant key points by frequency-band parallel processing.
- The system was combined with a PC to track a moving target of a varying scale.

ARTICLE INFO

Article history:
Received 24 November 2015
Received in revised form 27 April 2016
Accepted 9 May 2016
Available online 17 May 2016

Keywords:
Intelligent vision sensor
Bio-inspired architecture
Scale-invariant feature
Real-time object tracking
Resistive network
Hardware acceleration

ABSTRACT

We developed a vision sensor system that performs a scale-invariant feature transform (SIFT) in real time. To apply the SIFT algorithm efficiently, we focus on a two-fold process performed by the visual system: whole-image parallel filtering and frequency-band parallel processing. The vision sensor system comprises an active pixel sensor, a metal-oxide semiconductor (MOS)-based resistive network, a field-programmable gate array (FPGA), and a digital computer. We employed the MOS-based resistive network for instantaneous spatial filtering and a configurable filter size. The FPGA is used to pipeline process the frequency-band signals. The proposed system was evaluated by tracking the feature points detected on an object in a video.

1. Introduction

Real-time object tracking is one of the most important requirements for autonomous mobile robots. Vision-based object tracking requires selected points to be tracked and corresponding points to be searched in each frame. These tasks can be achieved effectively by using a local feature-point detector and descriptor. Methods that use such a detector and descriptor have been employed in a wide range of applications, and they are known to be effective (Schmid & Mohr, 1997; Su et al., 2012; see also Tuytelaars & Mikolajczyk, 2007 for an overview).

These methods extract feature points from an image according to particular criteria, and they describe a feature vector at each feature point. To apply feature points for object tracking in an algorithm implemented in a robot, the feature points need to be invariant to scale. This is because, in an image obtained by the mobile robot's vision sensor system, the apparent size of objects changes over time depending on the distance between the object and the vision sensor system. In addition, the vision sensor system used by a mobile robot must ensure low power dissipation with compact hardware.

One feasible algorithm for extracting and describing scale-invariant features is the scale-invariant feature transform (SIFT) (Lowe, 2004). The SIFT algorithm extracts and describes the scale-invariant local features of an object using multiple band-pass-filtered images. A couple of previous studies have demonstrated the effectiveness of the SIFT algorithm for object recognition (Ihara, Fujiyoshi, Takagi, Kumon, & Tamatsu, 2009; Lowe, 2004, for example). However, it is difficult for conventional digital image processing systems to extract SIFT features from an image in real time with low power consumption and compact hardware. There are two reasons for this. First, the SIFT algorithm uses multiple

* Corresponding author.
E-mail addresses: s-yasukawa@brain.kyutech.ac.jp (S. Yasukawa), h-okuno@eei.eng.osaka-u.ac.jp (H. Okuno), ishii@brain.kyutech.ac.jp (K. Ishii), yagi@eei.eng.osaka-u.ac.jp (T. Yagi).

http://dx.doi.org/10.1016/j.neunet.2016.05.002

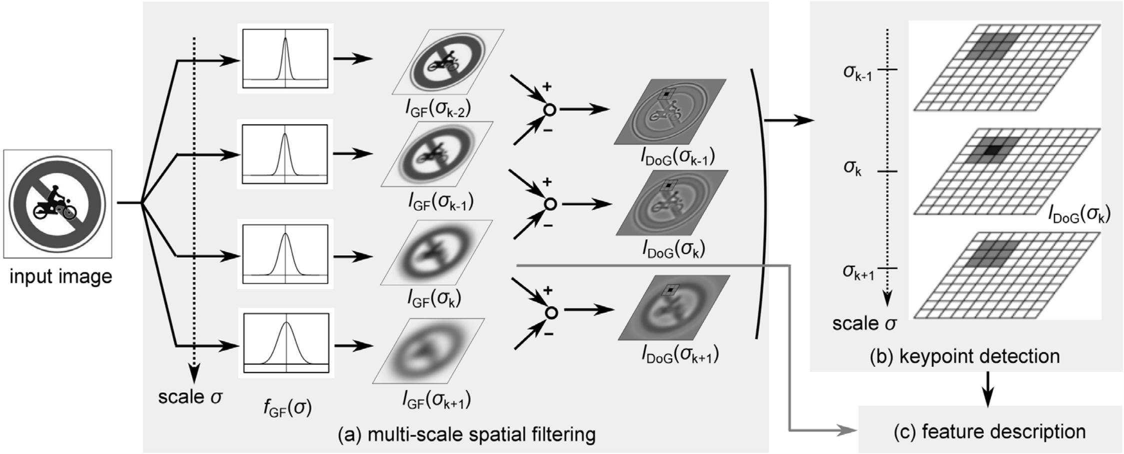

Fig. 1. Processing flow diagram for the SIFT algorithm. The SIFT algorithm consists of (a) multi-scale spatial filtering, (b) SIFT keypoint detection, and (c) feature description. In this image, $I_{GF}(\sigma_k)$ represents a Gaussian-filtered image, which is obtained by applying the Gaussian filter $f_{GF}(\sigma_k)$ to the input image. Further, $I_{DoG}(\sigma_k)$ represents a DoG-filtered image, and σ_k represents the scale parameter.

spatial low-pass-filtered images, and their sequential generation incurs high computational costs. Second, feature detection and description are executed repeatedly for multiple filtered images, which have high computational time requirements.

Much efforts have been devoted to improving the efficiency of the SIFT algorithm, that are focusing on improving the software algorithm and developing hardware accelerators. Software-based approaches include Fast approximated SIFT (Grabner, Grabner, & Bischof, 2006), SURF (Bay, Ess, Tuytelaars, & Van Gool, 2008), BRISK (Leutenegger, Chli, & Siegwart, 2011) and ORB (Rublee, Rabaud, Konolige, & Bradski, 2011). These methods aim to reduce the computational time by sacrificing the quality of the extracted features. Previous studies have also developed fast Gaussian filtering techniques, performing large Gaussian filtering with much less computation (Deriche, 1990; Farneback & Westin, 2006; Robinson, 2012; Sugimoto & Kamata, 2015; Unser, Aldroubi, & Eden, 1993; Wells, 1986). Although these efforts are successful at reducing the computational cost of the SIFT algorithm, these costs remain high for high-resolution inputs in particular.

Hardware accelerators dedicated to the SIFT algorithm have been also designed. These systems exploit the fast processing time of graphics processing units (GPUs), field-programmable gate arrays (FPGAs), or application-specific integrated circuits (ASICs). A GPU system has been applied to three-dimensional object recognition using the SIFT algorithm (Sinha, Frahm, Pollefeys, & Genc, 2006). Although GPUs offer high-speed parallel processing, GPU-based systems are unsuitable for portable applications, owing to their excessive power dissipation. An ASIC with multiple processing elements, each of which executes the SIFT algorithm selectively to predetermined regions of interest, has also been fabricated and applied to object recognition. In other previous studies, hardware accelerators that execute fast spatial filtering have been implemented on an FPGA (e.g., Bonato, Marques, & Constantinides, 2008; Huang, Huang, Ker, & Chen, 2012) or an ASIC (Lee et al., 2011; Su et al., 2012). Although application-specific hardware components improve the computational speed, they still suffer from high computational costs intrinsic to spatial filtering with full digital image processing.

Biological visual systems can be a good model for designing architectures to compute the SIFT algorithm efficiently. Spatial band-pass filtering is known to be a process executed at the early stages of the retinal neuronal circuit revealed in the response of bipolar cells (Kaneko, 1973). A neuronal architecture that applies such band-pass filtering was modeled as the difference between the outputs of two resistive networks (RNs) with a different spatial extent (Yagi, Ariki, & Funahashi, 1989; Yagi, Ohshima, & Funahashi, 1997). The difference between the two layers yields a spatial-band-passed image whose pass band is determined by the spatial-frequency characteristics of the two layers. Such band-pass architecture with two layers of RNs was implemented in a complementary metal-oxide semiconductor (CMOS) circuit (Boahen & Andreou, 1992; Kameda & Yagi, 2003; Matsumoto et al., 1992), inspired by analog very-large-scale integrated (aVLSI) image sensors known as silicon retinas (Mead & Mahowald, 1988). The RN is considered to be the most suitable solution to minimizing power consumption during spatial filtering (Mead, 1990; Poggio & Koch, 1985).

In a previous study, we proposed a SIFT computation algorithm that employs RN filters rather than Gaussian filters to extract keypoints. These filters play a key role in finding the corresponding points in images of different scales and rotation, based on computer simulations (Yasukawa, Okuno, & Yagi, 2012). However, it is unfeasible to implement multiple RNs with analog integrated circuit technology to realize scalable multi-band pass filters. Moreover, it is simply unrealistic to implement a SIFT algorithm with analog hardware. A novel system combining the advantages of analog metal-oxide semiconductor (MOS)-based RNs and compact digital hardware is needed for applying the SIFT algorithm to real-time image processing.

In this paper, we propose a novel architecture of a vision sensor system consisting of an analog MOS-based RN and an FPGA to execute SIFT computations in real time and constructed a prototype system based on the architecture. We demonstrate the efficiency of the prototype system by applying it to real-time object tracking based on scale-invariant features. The key factors of the efficiency are pixel-parallel filtering performed by the analog MOS-based RNs, and frequency-band parallel processing performed by the FPGA. Experiment results show that the scale-invariant features extracted by the proposed system are applicable to object tracking in which the apparent size of the target varies.

2. SIFT algorithm

We used the SIFT algorithm to extract local feature points from an input image. In this section, we only briefly review the SIFT algorithm, because its details are described in Lowe (1999, 2004).

Fig. 1 shows a flow diagram for the SIFT algorithm. The SIFT algorithm consists principally of multi-scale spatial filtering (Fig. 1(a)), SIFT keypoint detection (Fig. 1(b)), and a description of the feature vector around SIFT keypoints (Fig. 1(c)).