Microprocessors and Microsystems 46 (2016) 21–33

Contents lists available at ScienceDirect

Microprocessors and Microsystems

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

An efficient hardware-oriented stereo matching algorithm

Giuseppe Cocorullo, Pasquale Corsonello*, Fabio Frustaci, Stefania Perri

Department of Informatics, Modeling, Electronics and System Engineering, University of Calabria, Rende, Italy

ARTICLE INFO

Article history:
Received 5 August 2015
Revised 30 August 2016
Accepted 21 September 2016
Available online 21 September 2016

Keywords:
Stereo matching
VLSI architectures
Computer vision
FPGA

ABSTRACT

This paper presents a novel stereo matching algorithm that synergistically exploits in an original way the Adaptive Census Transform and the Support Local Binary Pattern approaches. For the first time, the Support Local Binary Pattern technique is applied to efficiently process large windows of pixels without excessive computational requirements, and thus allowing easy design of specific integrated circuits. Results obtained for conventional benchmark image sets demonstrate that, despite the simplifications adopted to make the novel algorithm hardware-friendly, the method proposed here can reach qualities higher than its competitors. Several hardware implementations have been carried out and characterized on the Xilinx Virtex FPGA chips. For 640×480 stereo images and a disparity range equal to 60, the proposed architecture guarantees an average error in computing the map as low as 9% with a throughput rate up to 68 frames per second. The cheapest version of a system designed as reported in this paper occupies less than 49,000 slices, 112 DSPs and 32 BRAMs.

1. Introduction

Stereovision is a technique widely used in several applications, such as obstacle detection, robot navigation, surveillance and many others [1–4], to create a three-dimensional reconstruction of an observed scene. In order to do this, depth information is calculated by applying the triangulation theory [5] to two different perspective images (also called stereo images) captured by two cameras that observe the same scene from two slightly different points of view.

A stereovision system retrieves depth from disparity that is the displacement between corresponding (matching) points in the acquired stereo images. Corresponding points can be recognized through several efficient matching algorithms [6–30], each characterized by its own accuracy and computational complexity. Global and semi-global approaches, such as those proposed in [7–12], are desirable when achieving very high accuracy is the main objective, even if it is obtained at the expense of the high computational complexity that makes standalone hardware implementations difficult to carry out without significantly sacrificing their accuracy. However, recently few attempts have been done in designing efficient stereo systems that exploit the global approach based on dynamic programming [13]. On the contrary, local approaches, such as those demonstrated in [14–26,29–33], are typically more appropriate when the realization of standalone stereovision systems is required together with reasonable accuracies. These algorithms locate corresponding pixels in the stereo images by measuring the dissimilarity between windows of pixels, usually named aggregation windows.

This paper proposes a novel hardware-oriented local matching algorithm that jointly exploits the Adaptive Census Transform (ACT) [26] and the Support Local Binary Pattern (SLBP) [25] approaches to compute accurate disparity maps. The SLBP approach was purposely modified to process windows larger than those originally adopted in [25], but limiting its computational complexity.

Quality tests performed on benchmark image sets [34] demonstrate that the method proposed here reaches accuracy levels significantly higher than competitors [19–22,25,26,29].

Several hardware implementations of the proposed algorithm have been carried out and characterized on Xilinx Virtex FPGA chips [35,36]. Results demonstrate that a throughput rate up to 68 640×480 frames per second can be reached with an average error rate of ~11.5%, by requiring ~70,800 Virtex-6 Slices. Whereas, average error rates lower than 9% can be obtained by using a Virtex-7 XC7V980T or XC7V2000T chip, and achieving a speed rate of 45 fps or 68 fps, respectively.

The remainder of the paper is organized as follows: Section 2 provides a brief background about the local matching algorithms selected as the counterparts; in Section 3, the novel algorithm presented here is described; accuracy measurements and comparison results are provided in Section 4; circuit designs of the basic computational modules required to hardware implement the novel algorithm are then presented in Section 5; finally, conclusions are drawn in Section 6.

* Corresponding author. Fax: +390984494834.
E-mail address: p.corsonello@unical.it (P. Corsonello).

http://dx.doi.org/10.1016/j.micpro.2016.09.010
0141-9331/

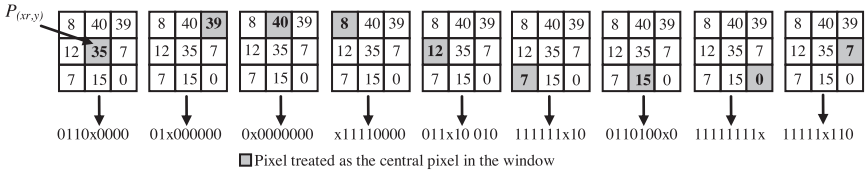

Fig. 1. Example of a transformation based on the SLBP approach.

2. Related works

In order to compute disparity maps, stereo image pairs acquired by two distinct cameras placed at a distance b from each other, called baseline, are preventively rectified following the computational steps detailed in Faugeras' book [5]. Then, for each pixel in the right image its corresponding (matching) point in the left image is found, or vice versa.

As it is well known, area-based local methods require three computational steps to find matching pixels. In the first step, for each pixel $P_{(xr,y)}$ in the right image and for each possible disparity d, with d varying between the minimum value d_min and the maximum value d_max, a pixel-based matching cost $MC(P_{(xr,y)},Q_{(xl,y)})$ between $P_{(xr,y)}$ and the pixel $Q_{(xl,y)}$ in the left image, with $xl = xr + d$, is computed. In the second step, for each pixel $P_{(xr,y)}$ in the right image and for each possible disparity d, the dissimilarity between $P_{(xr,y)}$ and $Q_{(xl,y)}$ is measured by aggregating the pixel-based matching costs computed in the previous step around the pixels of interest. The size of the aggregation windows obtained in this way is indicated as $W \times W$. In the last step the disparity $\delta_{(xr,y)}$ leading to the minimum dissimilarity value is recognized.

The absolute difference and the Hamming distance are two of the most used metrics to compute $MC(P_{(xr,y)},Q_{(xl,y)})$ [19–22,25,26,29–33,37–39]. When the Hamming distance is exploited, the Census transform is preliminary performed considering a proper number of pixels around the pixel of interest. The size of the support windows centred at $P_{(xr,y)}$ and $Q_{(xl,y)}$ is indicated as $W_T \times W_T$. The elements of the $(W_T^2 - 1)$-bit Census Vectors (CVs) are determined by comparing each pixel within the support window with its centre pixel [38]. The CVs related to $P_{(xr,y)}$ and $Q_{(xl,y)}$ are then compared in terms of Hamming distance and $MC(P_{(xr,y)},Q_{(xl,y)})$ is computed as the number of different elements in the CVs.

The Mini-Census transform has been recently proposed in [21]. For each $W_T \times W_T$ support window a $(W_T + 2)$-bit binary Mini-Census Vector (MCV) is calculated. Then, MCVs are compared in terms of Hamming distance to compute $MC(P_{(xr,y)},Q_{(xl,y)})$.

Alternative area-based algorithms, such as those proposed by Yoon and Kweon in [37], compute dissimilarities aggregating the pixel-based matching costs and exploiting the so called adaptive support-weight approach. It is based on the conjecture that neighbouring pixels within an aggregation window provide different supports according to their similarity and geometric proximity with respect to the central pixel. Thus, the weight of a pixel within an aggregation window is computed considering its Euclidian colour distance and its spatial distance from the centre pixel.

3. Inspiring algorithms

3.1. Background on support local binary pattern

The Support Local Binary Pattern (SLBP) method has been recently proposed by Nguyen et al. in [25] to reduce the sensitivity of the matching cost estimation to the intensity differences between the left and right images in real scenes. This algorithm measures the similarity between the generic pixel $P_{(xr,y)}$ within the right image and the pixel $Q_{(xl,y)}$ within the left image by processing their 3×3 windows. For each window, nine 8-element CVs are computed considering each pixel in the window in turn as the centre pixel. The similarity between $P_{(xr,y)}$ and $Q_{(xl,y)}$ is then measured summing the Hamming distances computed for the corresponding Census Vectors. As the final step, a winner-takes-all approach is applied to locate the pixel $Q_{(xl,y)}$ at the disparity $\delta_{(xr,y)}$ that is the least dissimilar from $P_{(xr,y)}$.

Let $P_{(xr,y)}$ be the centre pixel of a 3×3 window in the right image. $R_{(x',y')}$ being the neighbour pixels in the window, with $x' = xr - 1,\ldots,xr + 1$ and $y' = y - 1,\ldots,y + 1$, the SLBP approach consists in considering $R_{(x',y')}$ in turn as the centre pixel. In this way, 9 CVs would be computed each consisting of 8 elements. Fig. 1 illustrates an example with the nine 8-bit CVs generated. As detailed in [38], the generic bit of a CV is obtained by comparing the neighbour pixel $R_{(x',y')}$ with the centre pixel $P_{(xr,y)}$. The output bit is set to 1 if $R_{(x',y')} \geq P_{(xr,y)}$, otherwise it is set to 0. Each CV contains the value x to indicate the superfluous comparison between the current centre pixel with itself.

3.2. Background on Adaptive Census Transform

The recently proposed Adaptive Census Transform (ACT) approach [26] is particularly suitable for hardware realization of disparity maps computation engines. It exploits adaptive support weights in the image transformation step.

Such transformation produces vectors of integer numbers, thus the pixel-based matching cost is defined as their Sum of Absolute Differences (SAD). Finally, the adaptively weighted sum of SADs is used as the dissimilarity metric.

For each $W_T \times W_T$ support window centred at $P_{(xr,y)}$, with $W_T = 2s_T + 1$ and s_T being the radius of the window, the Adaptive Census transform generates a W_T^2-element weighted census vectors $WCV_{(xr,y)}$. The kth element of each WCV is defined by the piecewise function reported in (1a), with: $x' = xr - s_T,\ldots, xr + s_T$; $y' = y - s_T, \ldots, y + s_T$; k varying between 0 and $W_T^2 - 1$, and k depending on xr, x', y and y' as shown in (1b).

$$WCV_{(xr,y)}(k) = \begin{cases} -w(R,P) & | \quad P_{(xr,y)} > R_{(x',y')} \\ w(R,P) & | \quad P_{(xr,y)} \leq R_{(x',y')} \end{cases} \tag{1a}$$

$$k = \left(s_T + xr - x'\right) + \left(s_T + y - y'\right) + \left[W_T \cdot \left(s_T + xr - x'\right) - \left(xr - x'\right)\right] \tag{1b}$$

The term $w(R,P)$ is the support weight related to the neighbour pixel $R_{(x',y')}$ in the processed window and calculated referring to (2), where Δc and Δg are the colour and spatial distance between $P_{(xr,y)}$ and $R_{(x',y')}$, respectively, and γ_c and γ_p are tuning constants.[1]

$$w(R,P) = e^{-\left(\frac{\Delta c}{\gamma_c} + \frac{\Delta g}{\gamma_p}\right)} \tag{2}$$

[1] As discussed by Yoon et al. in [30], γ_c is an experimental parameter determined by the perceptual difference between colours, whereas γ_p depends on the radius of the support window.