SELF-RECTIFICATION AND DEPTH ESTIMATION OF STEREO VIDEO IN A REAL-TIME SMART CAMERA SYSTEM

Xinting Gao, Richard Kleihorst, Peter Meijer, and Ben Schueler

NXP Semiconductors, Corporate I&T / Research
High Tech Campus 32, 5656AE Eindhoven, The Netherlands
{xinting.gao, richard.kleihorst, peter.b.l.meijer, ben.schueler}@nxp.com

ABSTRACT

This paper presents a self-rectification stereo vision system based on a real-time, low power, and wireless smart camera platform. The proposed self-rectification method is suitable for an embedded parallel stereo system, where the epipolar lines are parallel to the image scan lines. The stereo images are first aligned by applying 1D signature matching. Then the alignment is refined based on the quality of the disparity measurement. The rectification method can be applied both offline and online. The major advantage of this rectification method is that no clean background is needed during the rectification process. After the rectification, the conjugate epipolar line is collinear. The dense matching method is implemented to achieve the depth map. This depth map provides a tool for segmentation. The application runs in an SIMD video-analysis processor, IC3D, at 30fps and handles disparity up to 37 pixels in CIF (320x240 pixels) mode.

Index Terms — self-rectification, depth estimation, stereo vision, real-time, embedded system, smart camera

1. INTRODUCTION

Through passive sensing with relatively low cost sensors, stereo vision views the same scene with two or more cameras and extracts the 3D information of the scene. This 3D information is useful in many computer vision systems such as video surveillance and robot vision. In the WiCa system, it can also be used as a basic segmentation tool to implement foreground extraction.

For a binocular stereo vision system, the centers of the two camera sensors and the 3D point determine the epipolar plane that intersects the two image planes at the epipolar lines. The stereo rectification aligns the conjugate epipolar lines collinear with each other and parallel to the horizontal scan lines, which implies that the corresponding points lie on the conjugate image lines. Thus, the stereo rectification simplifies the searching process from two dimensions to one dimension. To take the assumption that the stereo images have been rectified, the stereo analysis algorithm is simplified and easy to be implemented in a real-time system [1]. However, the vertical disparity may cause serious error in matching process if the stereo images are not rectified very well [2]. Consequently, it is important to apply a suitable rectification method for such kind of real time stereo system. For most of the existing embedded real-time stereo vision systems, the rectification is achieved with the help of the PC [1, 3]. A coordinate lookup table is saved and used. In the WiCa 1.1 system, the proposed self-rectification stereo vision system is implemented. The rectification consists two steps. Firstly, a 1D signature based matching is applied to get an initial vertical translation parameter. Secondly, the quality of the disparity measurement is used to refine the rectification parameter. Using the rectification result, the stereo image can be rectified and a dense disparity estimation method is also implemented in the WiCa stereo vision system. The correspondence searching is based on a local similarity measurement. Left-to-right checking and reliability checking is applied to reduce the errors. The system handles up to 37 pixels disparity at 30fps in CIF (320x240 pixels) mode. The schematic of the stereo vision system in WiCa is shown in Fig. 1.

The implemented depth estimation method is intended to be used as a segmentation tool in high level video processing, for example, in the application of gesture analysis [4]. For such applications, the accuracy of the depth map is not demanding. The stereo vision system presented using the low cost sensors and the high performance SIMD processor pro-
vides a robust segmentation tool that is relatively invariant to light condition and scene contents. Comparing to other technology for depth estimation, e.g., time-of-flight [5], the WiCa stereo system has the advantages of low power consumption, as active infrared modulated light is not needed, low cost sensors and lenses and portability.

The paper is organized as follows. In Section 2, the configuration of the WiCa smart camera system is introduced. Section 3 describes the stereo vision system in WiCa, that includes the theoretical background of the stereo rectification, the proposed self-rectification method and the disparity estimation method. Section 4 demonstrates the results of the WiCa stereo system and Section 5 gives the conclusion.

2. SYSTEM INTRODUCTION OF THE WICA SYSTEM

WiCa is a high performance wireless smart camera system that is suitable for real-time video processing. Due to the advances of integration, WiCa provides a low cost, low power and embedded programmable platform. Being an embedded wireless system, WiCa 1.1 is small-size (7.5 x 9.5 x 2.5cm). WiCa can be portably used with batteries as power supply. Due to the advantages of the WiCa, it can be applied in ambient computing, mobile smart vision system and consumer electronics. Stereo vision is one of the major tools of the WiCa system. A detailed introduction about the WiCa system can be found in [6]. In the following, we give a description of the components as far as it is relevant to the stereo vision implementation.

2.1. WiCa 1.1 hardware

The WiCa 1.1 smart camera system contains the following components: two VGA color image sensors, an SIMD processor - IC3D, an 8051 microcontroller, a communication module and RAM module (256KBx8). The RAM in WiCa 1.1 is SRAM which is accessed as a Dual Port RAM (DPRAM) using a CPLD. The RAM functions as the communication buffer between the two processors (IC3D and 8051) and enables them to work in their own clock domains. It also serves as a loop-back frame buffer for the IC3D. Fig. 2 shows the WiCa 1.1 board (the 8051 is on the other side of the board). The two VGA color image sensors provide either CIF (320x240) or VGA (640x480) images to the stereo system. The baseline of the two sensors is 4cm. The functional diagram of WiCa 1.1 is shown in Fig. 3. The LCD is only used for debugging the embedded system.

The IC3D is one of the SIMD processors of Philips’ and NXP’s Xetal series. Xetal is designed specifically for low level video processing. Fig. 4 shows the basic internal blocks of the IC3D. The main component of the chip is the Linear Processor Array (LPA) with 320 RISC processors. Each of the processors has simultaneous read and write access within

Fig. 2. WiCa 1.1: a smart camera system.

Fig. 3. Functional diagram of WiCa 1.1.

Fig. 4. Architecture of the IC3D.
<table>
<thead>
<tr>
<th>Features</th>
<th>IC3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array Width</td>
<td>320 processors</td>
</tr>
<tr>
<td>Number of Line Memories</td>
<td>64</td>
</tr>
<tr>
<td>Data Path Width</td>
<td>10-bit</td>
</tr>
<tr>
<td>Number of Video Input</td>
<td>3 or 4</td>
</tr>
<tr>
<td>Number of Video Output</td>
<td>3</td>
</tr>
<tr>
<td>Performance</td>
<td>50GOPS (10bits MAC)</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>15mW/GOP</td>
</tr>
<tr>
<td>Power Supply</td>
<td>1.2 ... 1.8 Volt</td>
</tr>
</tbody>
</table>

Table 1. Specifications of IC3D.

one clock-cycle to its own memory position or that of its left and right neighbors in the parallel line memory. Both the memory address and the instruction of the processors are shared in SIMD sense, i.e. one memory address or one instruction is applied to the whole 320 pixels simultaneously. The line-memory block can store 64 lines of 320 pixels. Each pixel is up to 10 bits. The Global Control Processor (GCP) controls the IC3D and does global DSP operations. The peak pixel performance of the IC3D is around 50GOPS at 80MHz. Despite its high performance, IC3D is an inherently low power processor. It consumes around 15mW/GOP. As a massive parallel processor, IC3D is advantageous for a dense matching algorithm [7]. Therefore, a dense matching based depth estimation method is implemented in WiCa 1.1. Table 1 shows some specifications of the IC3D.

The DPRAM is another important component in the stereo vision application. One of the stereo image is written to DPRAM. It is transformed and read back during the rectification and depth estimation. DPRAM works as the asynchronous connection between IC3D and 8051. The size of the memory is 64KB for each bank and WiCa 1.1 has 8 banks in total. One image of 256x256 pixels can be directly stored into one DPRAM bank.

The host controller 8051 is suitable for high-level processing and communicates to the outside world.

2.2. WiCa 1.1 software

Programs for the WiCa 1.1 are written in a language called XTC. It is similar to C, but with implicit parallel data-types and without pointers. For next generation, WiCa 1.2, the programming language C will be used and all XTC programs can be converted directly.

2.3. Setup of the hardware

Fig. 5 shows the setup of the system. The LCD is for debugging display. The WiCa is connected to the PC through a USB port. The program is compiled by a specific compiler. A PC program called “WiCaEnv” is used to configure the system and upload the program from the PC to the WiCa board.

3. WiCa stereo vision system

In this section, the WiCa self-rectification stereo vision system is introduced. For the dense depth estimation methods based on epipolar geometry, most of the existing papers assume that the two images/videos are already rectified. In practice, the system is rectified by either manually adjusting the hardware or using some existing tools with the assistance of the PC [1, 3]. As an embedded system, this is not convenient, especially when it is applied in a mobile situation. On the other hand, rectification is necessary since the vertical disparity may cause severe error in the dense depth estimation process based on the epipolar line constraint. In the WiCa stereo system, a self-rectification method and a dense depth estimation method are implemented.

3.1. Theoretical background of the stereo rectification

Generally speaking, both correction and rectification are needed to align the two epipolar lines collinear as demonstrated in Fig. 6. The correction process removes distortion, e.g. lens distortion, while the rectification process transforms the two corrected images and make the conjugate epipolar lines collinear with each other and parallel to the image scan lines.

To illustrate the process more clearly, we take a self-rectification method proposed by Papadimitriou and Dennis in [2] as an example. The self-rectification process is illustrated in Fig. 7. The baseline is taken as the \( X \)-axis of the world coordinate system. It is assumed that the camera systems have the coplanar \( X \)-axis and \( Z \)-axis and parallel \( Y \)-axis. Then the rotation of the two cameras is around the \( Y \)-axis. For a convergent stereo rig, the angles between the optical axes and the \( Z \)-axis, \( \phi_l \) and \( \phi_r \), are generally nonzero before rectification. The rectification process is to rotate and shift the two cameras. The rotation makes the \( \phi_l \) and \( \phi_r \) equal to zero, i.e., the two optical axes of the rectified cameras are orthogonal to the baseline and thus parallel to each other. This process makes the epipolar lines parallel to the images’ horizontal scan lines. The translation shifts one image vertically to make the conjugates epipolar lines collinear with each other, i.e., removes the vertical disparity. The horizontal disparity is caused by
Align image to satisfy epipolar constraint
Eliminate distortion
Left image corrected
Left image rectified
Original left image
Original right image
Align image to satisfy epipolar constraint
Eliminate distortion
Right image corrected
Right image rectified
Original right image
Original image
Image corrected
Image rectified

**Fig. 6.** Image correction and rectification to satisfy the collinear epipolar constraint.

depth and should not be be removed in the rectification. In Fig. 7, the 3D point $M$ is projected to the image planes $I_l$ and $I_r$ as a image pixel $(x_l, y_l)$ and $(x_r, y_r)$, respectively. The rectification process projects the point $M$ into the rectified image planes $I'_l$ and $I'_r$ with coordinates $(x'_l, y'_l)$ and $(x'_r, y'_r)$, respectively. They are related as follows.

$$x'_l = f_l \cdot \frac{x_l - x_{lo}}{f_l \cdot \cos \phi_l + (x_l - x_{lo}) \cdot \sin \phi_l}$$  \hspace{1cm} (1)$$

$$y'_l = \frac{y_l - y_{lo}}{f_l \cdot \cos \phi_l + (x_l - x_{lo}) \cdot \sin \phi_l}$$  \hspace{1cm} (2)$$

where, $f_l$ is the focal length of the left camera, while $(x_{lo}, y_{lo})$ is the center of the image point of the left image. The equations for the right image can be derived in a similar way. To remove the vertical disparity, we get $y'_l = y'_r$, i.e.,

$$y' - y_{lo} = y_r - y_{ro} = d_y$$  \hspace{1cm} (3)$$

where, $d_y$ is a constant that represents the vertical shift between the left and right images.

In [2], several pairs of corresponding points in the stereo images are identified and used to solve Eq. (3). By applying the parameters to the stereo system, the stereo images are rectified.

**3.2. Offline self-rectification of WiCa stereo vision system**

In the WiCa 1.1 system, the two cameras are rigidly connected to the WiCa board as shown in Fig. 8. We have measured that WiCa 1.1 system has a proper parallel stereo rig, i.e., the pan angles, $\phi_l$ and $\phi_r$ in Eq. (3) equals to zero. In such case,

$$y_l - y_{lo} = y_r - y_{ro}$$

i.e.,

$$(y_l - y_{lo}) = (y_r - y_{ro}) = d_y$$  \hspace{1cm} (3)$$

The proposed self-rectification system includes two parts. Firstly, a 1D signature based matching method is implemented to achieve the rough vertical match between the left and right images. Secondly, the vertical shift parameter is refined iteratively based on the quality of the disparity measurement. In

**Fig. 7.** The illustration of stereo rectification.
the following, the two parts are explained in detail.

For the 1D signature based matching, the two images are projected horizontally by summing the central parts of the video lines as follows.

\[ s_l(i) = \sum_{j=m}^{n} I_l(i, j), \quad (4) \]

\[ s_r(i) = \sum_{j=m}^{n} I_r(i, j), \quad (5) \]

where \( s_l \) and \( s_r \) are the 1D signatures, and \( I_l \) and \( I_r \) are the intensity values of the left and right images, respectively. Then the two signatures are shifted and matched to each other.

\[ C(t) = \sum_{i=p}^{q} |s_l(i) - s_r(i - t)|. \quad (6) \]

Here, \( C(t) \) is the similarity measurement of the two 1D signatures. For an integer \( t \in [-r, r] \), the rectification parameter \( t_y \) is the shift value \( t \) that corresponds to the minimum value of \( C(t) \) as follows.

\[ t_y = \min_{t} \{ C(t) | -r \leq t \leq r \}. \quad (7) \]

Apply \( t_y \) to the stereo images, the two images are roughly rectified. In the second step, the rectification is improved by comparing the quality of disparity estimation. As we know, a better rectification leads to a better quality of disparity measurement. The quality of disparity estimation is defined as the sum of the each correlation measurement corresponding to its disparity as follows.

\[ Q(t) = \sum_{(i,j)} corr(i, j, t(i, j)). \quad (8) \]

Here, \( corr(i, j, t(i, j)) \) is the correlation value corresponding to the searched disparity. By applying \( t_y, t_y + 1, \) and \( t_y - 1 \) to rectify the stereo images, we obtain three quality measurements. The accurate rectification corresponds to the minimum of the quality measurement. This process is iterated until the result converges. The flowchart of the rectification is shown in Fig. 9. Comparing to the rectification method presented in [8], the major advantage of the proposed self-rectification method is that there is no restriction of the scene during the rectification process.

3.3. Depth estimation of WiCa stereo vision system

For the dense depth estimation methods, Scharstein and Szeliski gave a detailed overview in their seminal paper [9]. Due to the simplicity and relatively good performance of the SAD (sum of absolute difference) [10], we choose SAD as the similarity measurement. For each shift within the disparity searching range, the SAD, \( C(i, j, t) \), is computed. Then, the Winner Takes All (WTA) method is applied in the disparity computation, i.e., the disparity corresponding to the minimum SAD during the searching is taken as the correct one.

There are many factors that can result in wrong matching, i.e., the wrong disparity value is achieved. Two refinement processing steps are adopted in the system. First, the reverse searching is conducted, which is particularly effective for wrong matching caused by occlusion around the border area of an object. This is called left-right checking. Second, the reliability is checked for each matching pixel. During the matching process, the correlation corresponding to the second best matching, \( corr_{2nd} \), is recorded as well as the best matching result, \( corr_{best} \), for each pixel. If the two values are too close to each other, it is taken that the pixel is within a homogeneous region, making the matching result unreliable.

The details of the implementation of the depth estimation method are described in [8].

3.4. Online self-rectification of WiCa stereo vision system

The above offline rectification method can be applied directly online by carefully distributing the computational load. WiCa system supports VGA(640x480) mode of video processing. At present, the stereo vision system runs in CIF(320x240) mode, which means that we only compute the disparity map for half of the total video lines. Therefore, the self-rectification can be applied in the other half of the total video lines. The flowchart of the disparity estimation with an online self-rectification method is demonstrated in Fig. 10. The online rectification is running iteratively within every second.

4. IMPLEMENTATION AND RESULTS

For the offline self-rectification, the two images are captured and written into two DPRAM banks. Both the 1D signa-
Compute the 1D signatures of the two images
Rectify the image according to $t_y / (t_y \pm 1)$
Compute the two disparity maps based on the above rectification
Compute the quality of the two disparity maps
Decide the new translation and refinement parameters

ON

Read the snapshot images from the two dpram banks
Compute the 1D signatures of the two images
Shift and match the 1D signatures of the two images
Get the vertical translation $t_y$ between the two images

REFINE

Read the snapshot images from the two dpram banks
Rectify the image according to $t_y / (t_y \pm 1)$
Compute the two disparity maps based on the above rectification
Compute the quality of the two disparity maps
Decide the new translation and refinement parameters

REFINE/OFF

Fig. 9. The flowchart of the offline self-rectification.

Odd video lines
Online rectification?
Yes

Frame = = 1?
No

Rectification based on 1D signature matching
Refine rectification based on quality of disparity map

Disparity estimation

Even video lines
Rectify the image according to $t_y$
Compute disparity map
Apply right-to-left checking
Apply the reliability check
Adjust the disparity range and display

Fig. 10. The flowchart of the disparity estimation.
Table 2. The resource usage of the WiCa 1.1 system for the depth estimation application running at 30fps with 37 disparity levels in CIF mode.

<table>
<thead>
<tr>
<th>Resources</th>
<th>used/available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program Memory</td>
<td>983/2048</td>
</tr>
<tr>
<td>Number of Line-memories</td>
<td>47/64</td>
</tr>
<tr>
<td>Instructions/video line</td>
<td>4436</td>
</tr>
<tr>
<td>IC3D Power Consumption</td>
<td>350mWatts</td>
</tr>
</tbody>
</table>

ture based method and the subsequent refinement read the snapshots from the two DPRAM banks. For the online self-rectification, 1D signature based method is applied with down sampling of the input video lines, i.e., the maximum length of $s_l$ and $s_r$ in Eq. (4) and Eq. (5) is 120. The result can be improved in the refinement process.

By applying the proposed self-rectification method and disparity estimation method, the stereo vision system achieves quite satisfactory results as shown in Fig. 11 and 12, respectively. The rectification method vertically aligns the left image with the right one.

Table 2 shows the resource usage for the real-time depth estimation application. In CIF mode, video data is read at every odd line and can be processed in either odd lines or even lines. In Table 2, the instructions/video line is measured at the odd line. For the even lines, the only operation is to write the upper image into the DPRAM. The full computational power at the odd video lines is used to achieve the 37 disparity levels. However, IC3D has a computational power of 480 video lines per frame. The computational power for the even lines is still available and can be used by other programs. The IC3D power consumption in table 2 is calculated according to the above analysis.

5. CONCLUSIONS

In this paper, we presented a self-rectification stereo vision method in a smart camera system - WiCa 1.1. The proposed self-rectification method includes two steps: 1D signature based matching and quality of disparity measurement based refinement. The proposed method is suitable for a parallel stereo system like WiCa and can be applied both offline and online in the application. The dense depth estimation method was also implemented in the system. By taking advantages of the SIMD processor, the system runs at 30fps and achieves satisfactory results for disparities up to 37 pixels in CIF mode.

6. ACKNOWLEDGEMENT

The authors would like to thank all the team members for Xetal project, especially Joost Hart, Herman Budde, Alexander Danilin and Zoran Zivkovic for their assistance and discussions during the work.

7. REFERENCES

Fig. 11. Results of the rectification: (a) original left image, (b) original right image, and (c) rectified left image.

Fig. 12. Results of depth estimation: (a) depth estimation without checking, (b) depth estimation with the left-to-right checking, and (c) depth estimation with both the left-to-right checking and the reliability checking.