A NEW GRADIENT-BASED FEATURE EXTRACTION METHOD FOR REAL-TIME DETECTION OF MOVING OBJECTS USING STEREO CAMERAS

Güray Sonugür#, Barış Gökçe*

Received on October 19, 2021
Presented by Ch. Roumenin, Member of BAS, on November 30, 2021

Abstract

In this study, a gradient-based feature extraction method has been developed that can be used to detect moving objects in real-time applications such as unmanned ground or air vehicles. Feature extraction methods should produce fast results in real-time applications, as results need to be obtained between successive frames of video sequences within a limited time. For this reason, various sized image blocks were used in the developed method. The arithmetic mean (AM), geometric mean (GM), median (MD), and local contrast (LC) methods were used to calculate block intensities. In the stereo video stream, depth maps were also divided into blocks along with successive frames’ R, G, and B channels. A novel feature extraction method was developed by calculating gradient-based relationships between adjacent blocks around the centre block. In experimental studies, the features extracted from stereo video frames using the proposed method were compared with Surf, Fast and Brisk methods according to their quantity, accuracy, and processing times, and more successful results were obtained. In addition, the moving object detection performance of the method was tested in real-time using an Unmanned Ground Vehicle.

Key words: digital cameras, feature extraction, gradient methods, stereo image processing, video surveillance, unmanned ground vehicle

Introduction. In order to detect moving objects between successive frames in video streams obtained from moving platforms, the noise caused by camera...
movement must first be compensated. One of the most common approaches used to address this problem is image registration [1]. This method can be divided into two classes, area-based and feature-based. The computational costs of area-based methods are quite high. Feature-based methods have lower computational costs than area-based methods and often produce more successful results [2]. Features are numerical data containing important information about the image [3]. In feature-based image registration methods, position changes of features between successive frames are calculated, and a transformation is performed according to the transposition value found. Assuming that the moving platform is a vehicle such as Unmanned Ground Vehicle (UGV) or Unmanned Aerial Vehicle (UAV), the volumes, orientations, and positions of moving objects will change during successive video frames [4]. Therefore, the extracted features for correct matching should be invariant to translation, rotation, and scale [5]. There are many feature extraction methods in the literature. The most common of these are Speeded-Up Robust Feature (Surf) [6], Fast and Harris Corner Detector [7,8], Binary Robust Invariant Scalable Keypoint (Brisk) [9], and ORB [10] methods.

When these methods are examined, it is observed that they have computational complexity that can lead to temporal problems when used in real-time movements. This problem restricts the use of these methods in real-time applications. These drawbacks are addressed by the developed GoSP (Gradient of Surrounding Pixels) technique. It can extract 3-D features because it can use the depth information obtained from stereo video sequences. Stereo camera is a type of camera that can extract per-pixel depth information of objects on the scene.

In this study, successive frames and depth maps were divided into various numbers of equal-sized blocks. In this way, the number of pixels used in calculations was reduced, and the computational speed was increased. The intensity value of each image block was calculated using the arithmetic mean (AM), geometric mean (GM), median (MD), and local contrast (LC) methods so that the blocks could accurately represent the original image. As a result, GoSP consists of three stages. In the first stage, the image frames and depth maps taken from the stereo camera were divided into blocks of various sizes. In the second stage, a gradient-based feature vector was created, and in the third stage, feature matching and image registration operations were performed. The main contributions of this study are as follows: (1) determining the most suitable method by trying different approaches in the reduction of image sizes with the blocking method; (2) using of depth maps in feature extraction processes; (3) success in feature extraction in flat areas where traditional methods have failed; (4) due to its low time consumption, it can be used for real-time applications.

**Methodology.** **Image blocking.** High-speed computers, in general, cannot be employed in real-time applications such as UGV and UAV. Therefore, to reduce the computational cost, it is necessary to reduce the size of high-quality image frames without losing the information they contain. In this study, image frames
were divided into blocks of various sizes, and their dimensions were reduced. The more the block sizes increase, the lower the image sizes, but the more distorted the image content.

**Extracting of GoSP.** In the developed feature extraction method, gradient-based relationships between the centre pixel and neighbour pixels were digitized. The gradient vector consists of two parts. These are amplitude $\nabla f$ and direction $(\theta)$. The amplitude and direction expressions are given in Eq. 1-2, respectively.

\[
\nabla f = \sqrt{(\frac{\partial f}{\partial x})^2 + (\frac{\partial f}{\partial y})^2}
\]

\[
\theta = \tan^{-1}\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right).
\]

The feature extraction is performed for all colour channels and depth maps in an RGB image. Computational time is short because of simple gradient-based calculations.

**Image registration.** Image registration is the geometric alignment of two images of the same scene captured from various angles, in different scales, and with different sensors \[11\]. It is mainly used in the detection and recognition of objects from video sequences taken with moving cameras. In video streams, the registration of images is achieved by aligning frames one by one in a cycle. For the alignment, the features extracted from image pairs are used. That process is called Feature Matching. The first image captured is selected as a reference in each cycle, and the next image is a sensed image.

The sensed image is then aligned to the reference image. Then the discrepancies between the reference and the sensed images are calculated by techniques such as Inter frame differences \[12\] and background estimation \[13\]. Such discrepancies are often caused by moving objects. The loop lasts until the end of the video stream. Feature matching is the most critical step in image registration. Because the more correctly matched features, the more accurate alignment will occur.

Nevertheless, there are translational, rotational, and scaling variations between the reference and the sensed images in the video stream.

**Experimental results.** Three locations were identified for use in experimental studies. The locations were called Location1 (Loc1), Location2 (Loc2), and Location3 (Loc3). Loc1 is indoor, Loc2 and Loc3 are outdoors. Loc3’s background is more complicated than others. A UGV was used in the test studies of the proposed method. The vehicle was moved with remote control, and visual data were retrieved from the Zed stereo camera. The performance of the entire system was measured in three steps. The first step is the image blocking action. This step calculated how well each blocking approach (AM, GM, MD, and LC) represented the original image. For calculation, the histograms of the original and blocked images were compared. In the second step, the proposed approach was
compared with other methods in terms of the number of features and time consumption. In the third step, feature matching and image recovering performances were compared with ground truth results.

**Feature extraction performance.** In this experiment, GoSP was compared to Surf, Fast and Brisk methods using 180x320 blocked RGB images and depth maps. As a result of the calculations, it was determined that an equal number of features (212 for 180×320 resolution) were extracted for each image from all locations using GoSP. On the other hand, using the Surf method, an average of 116 features could be extracted from all locations. Fewer features could be extracted by using the Fast and Brisk methods. When the results were analyzed based on the number of extracted features, it was discovered that conventional techniques had certain drawbacks. For example, traditional methods extracted fewer features from areas where the intensity variation is low, such as depth maps.

**Feature matching and image recovering.** In earlier tests, GoSP outperformed other techniques in terms of feature count. However, it is not sufficient to detect the features from only one frame in a video stream. Features should be detected along with successive frames and matched with those detected from the previous frame. For this reason, experimental studies were conducted to determine the correct matching performance of the GoSP by selecting frame pairs with 30 frame differences between one and another (e.g., FR₁ and FR₃₁, FR₂ and FR₃₂). The test videos were taken from Loc1, Loc2, and Loc3. The Zed stereo camera is recording at 60 fps, so the time between frames is 0.5 seconds. These two frames were divided into 2×2, 3×3, and 4×4 blocks with MD, AM, and GM methods. The results of the experimental test with the sample frame pair (Fr₁₄×₄ and Fr₃₁₄×₄) are seen in Fig. 1. For detailed figures: [https://github.com/akuMeka/A-New-Gradient-Based-Feature-Extraction-Method/tree/main/Figures](https://github.com/akuMeka/A-New-Gradient-Based-Feature-Extraction-Method/tree/main/Figures)

In GoSP, adequate feature matching has occurred to produce transformation for image recovery, even at low resolution. The K-Means clustering approach was used to determine the correct matches.

**Image registration and scale invariance performance.** In order to measure GoSP’s success in image registration, it is necessary to calculate the similarity between the recovered image and the reference image. The high similarity will show that the method was successful. There are varying amounts of translational, rotational, and scaling differences between consecutive frames in real-time video streaming. First, it is essential to recover the sensed image by developing a transforming model to compensate for these variations. In this study, the sensed image was recovered using the geometric transformation [14]. The correlation between the reference image and the sensed image can only be determined after this step. Cross-Correlation (CC), Sum of Squared Differences (SSD), and Mutual Information (MI) methods were used to measure similarity.

A dataset was created to measure the similarity between the sensed image recovered after the geometric transformation and the reference image. In order
to measure performance reliably, it should be compared with ground truth values. Therefore, an image obtained from Loc3 was translated, rotated, and scaled manually at the values given in Table 1.

The test results showed that the procedure worked well up to the 7th line but failed beyond that. Our scale invariancy limit is the values in line 7. A successful image recovery cannot be carried out using the method developed after these values. However, in real-time UGV movements, the maximum ranges of translation,
CC, MI, and SSD similarities between sensed images and reference images after geometric transformation

<table>
<thead>
<tr>
<th>GT</th>
<th>Translation (pixel)</th>
<th>Rotation (degree)</th>
<th>Scaling (coeff.)</th>
<th>CC</th>
<th>MI</th>
<th>SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{r,ref}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4.44</td>
<td>0</td>
</tr>
<tr>
<td>$F_{s,ref}$</td>
<td>5</td>
<td>5</td>
<td>0.7</td>
<td>0.939</td>
<td>2.248</td>
<td>2209.2</td>
</tr>
<tr>
<td>$F_{s,sensed}^1$</td>
<td>8</td>
<td>8</td>
<td>0.83</td>
<td>0.919</td>
<td>2.346</td>
<td>2346.2</td>
</tr>
<tr>
<td>$F_{s,sensed}^2$</td>
<td>11</td>
<td>11</td>
<td>0.96</td>
<td>0.901</td>
<td>2.531</td>
<td>2630.8</td>
</tr>
<tr>
<td>$F_{s,sensed}^3$</td>
<td>14</td>
<td>14</td>
<td>1.09</td>
<td>0.890</td>
<td>2.611</td>
<td>2901.8</td>
</tr>
<tr>
<td>$F_{s,sensed}^4$</td>
<td>17</td>
<td>17</td>
<td>1.22</td>
<td>0.879</td>
<td>2.712</td>
<td>3194.9</td>
</tr>
<tr>
<td>$F_{s,sensed}^5$</td>
<td>20</td>
<td>20</td>
<td>1.35</td>
<td>0.841</td>
<td>2.134</td>
<td>3485.6</td>
</tr>
<tr>
<td>$F_{s,sensed}^6$</td>
<td>23</td>
<td>23</td>
<td>1.48</td>
<td>0.839</td>
<td>2.584</td>
<td>3527.5</td>
</tr>
<tr>
<td>$F_{s,sensed}^7$</td>
<td>26</td>
<td>26</td>
<td>1.61</td>
<td>0.03</td>
<td>1.291</td>
<td>7202.1</td>
</tr>
</tbody>
</table>

rotation, and scaling changes are essential to determine the performance of the proposed method. For this reason, the translation, rotation, and scaling changes occurring in the successive frames of the videos obtained on asphalt, dirt, and muddy roads were measured with the UGV used in the study. The study results are given in Table 2. The method worked well even on the dirt road, where the highest differences were seen. The UGV was used at 4 m/s, and image data were collected only during forward motion. A total of 30 frames at intervals of 5 frames were measured. Translation values were taken on the horizontal axis only. It is shown that our technique can be utilized in real-time because the obtained values are inside the spectrum where the proposed method is scale-invariant.

Table 2
Average difference values

<table>
<thead>
<tr>
<th>Road Conditions</th>
<th>Average translational differences</th>
<th>Average rotational differences</th>
<th>Average scaling differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asphalt</td>
<td>5.1 px</td>
<td>3.1°</td>
<td>1.1</td>
</tr>
<tr>
<td>Dirt</td>
<td>12.7 px</td>
<td>8.9°</td>
<td>1.4</td>
</tr>
<tr>
<td>Muddy</td>
<td>8.4 px</td>
<td>4.6°</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Real-time performance. Finally, the success of the GoSP method in the real-time movement was tested. In the test run, the UGV was moved forward at two different speeds, 4 m/s and 5 m/s. Successful object detection rates in Frame Differencing [15] and Mixture of Gaussian [16] methods were used as performance metrics. GM was used as the blocking method and $4 \times 4$ as the block size.

4. Conclusions. A gradient-based feature extraction method for stereo video streams was developed and tested with a UGV in this study. Depth maps
Table 3  
Real-time performance of the GoSP method

<table>
<thead>
<tr>
<th>Speed</th>
<th>Frame differencing</th>
<th>Mixture of Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 m/s</td>
<td>90.4</td>
<td>86.3</td>
</tr>
<tr>
<td>5 m/s</td>
<td>81.9</td>
<td>79.8</td>
</tr>
</tbody>
</table>

obtained from stereo cameras were also used in this method. For reducing the computational and memory costs of the method, the RGB (left camera) and the depth map images were divided into blocks of certain sizes. Block intensity values were determined using the geometric mean that provides the results nearest to the original. The method developed was compared with the Surf, Fast and Brisk methods in terms of the number of features and time-consuming, and perfect results were obtained. Experiments on video sequences of varying background complexities extracted 1.8 times more features than the nearest method based on the number of features and 2.1 times faster than the nearest method based on the processing speed. In contrast to Surf, Fast, and Brisk methods, the method developed can easily extract features from flat areas of the image. In addition, the moving object detection performance of the method was tested in real-time using a UGV. The results showed that the GoSP method’s image registration is scale-invariant for the following ranges between two consecutive image frames. Horizontal and vertical translation $[-30, +30]$ pixels, Rotational $[-20^\circ, +20^\circ]$ degrees, Scale coefficient $[0.6, 1.48]$ backward and forward moving. Since these values are within the scale invariancy limits detected in the previous section for UAVs and UGVs, the method can be easily applied in real-time video streams. In tests performed with a UGV at various speeds, moving objects were detected with two different methods, and success rates ranging from 90.4% to 79.8% were reached.

REFERENCES


G. Sonugür, B. Gökçe


*Department of Mechatronics Engineering
Faculty of Engineering and Architecture
Necmettin Erbakan University
42140 Konya, Turkey
e-mail: b_gokce@erbakan.edu.tr

Department of Mechatronics Engineering
Faculty of Technology
Afyon Kocatepe University
03200 Afyonkarahisar, Turkey
e-mail: gsonugur@aku.edu.tr