Stereo-based Spatial and Temporal Feature Matching Method for Object Tracking and Distance Estimation

Young-Chul Lim, Chung-Hee Lee and Jonghwan Kim
Daegu Gyeongbuk Institute of Science & Technology, 50-1 Sang-Ri, Hyeonpung-Myeon, Dalseong-Gun, Daegu, Korea

Keywords: Stereo Vision, Object Tracking, Distance Estimation, Feature Tracking, Feature Matching.

Abstract: In this paper, we propose a stereo-based object tracking and distance estimation method using spatial and temporal feature matching scheme. Our work aims to track an object robustly and to estimate its distance accurately without stereo matching processing, which requires a considerable amount of processing time and numerous processing resources. Our method combines temporal feature matching and spatial feature matching schemes. Our experimental results demonstrate that the proposed method can provide good object tracking and distance estimation performance in real-world environments.

1 INTRODUCTION

Stereo-based detection and tracking methods have been widely researched and applied in various fields over the past few decades. Stereo vision system can detect and track an object reliably while also estimating the distance of the object accurately. Generally, stereo matching processing is required to obtain a 3D depth image from two rectified 2D images. Stereo matching performs a process known as brute-force corresponding searching, a complex and time-consuming task. Moreover, this task is vulnerable to the illumination difference between left and right images, which often causes matching errors under external environments. The erroneous matching result results in poor detection and tracking performance.

Feature-based object tracking attempts to find corresponding features using a distinctive feature extraction method such as Harris corner detection (Harris and Stephens, 1988), scale-invariant feature transform (SIFT) (Lowe, 2004), and speeded up robust features (SURF) (Bay et al., 2008), and features from accelerated segment test (FAST) (Rosten et al., 2010). Feature-based object tracking establishes these correspondences in consecutive frames and estimates the transform matrix of the feature pairs. The Kanade-Lucas-Tomasi (KLT) feature tracker using Harris corner detection (Jianbo and Tomasi, 1994) has been widely used in many feature-based object tracking applications. The method basically depends on the sum of squared differences in the window. It often fails to track the features during illumination changes. SIFT is robust against rotation, translation, scaling, and illumination changes. SURF can perform similarly to SIFT while processing much faster. When these feature detectors are used for feature tracking, the interest points of the two images are matched by descriptor comparisons. A descriptor is a vector with a fixed size of its floating point values, which represent the direction and magnitude of the gradient around the key point. Therefore, these features need much more time to extract the key points and to match their descriptors compared to the KLT method. Recently, the binary robust independent elementary features (BRIEF) (Calonder et al., 2012) method using binary strings as a feature descriptor was proposed to reduce substantially the computation amount while yielding higher matching rates under certain restricted conditions.

Feature-based object tracking involves the following procedure. The region of interest (ROI) of the target object is located in the first frame, either manually or automatically, using an object-specific detector. The corresponding feature candidates of the previous frame are estimated using the above-mentioned feature extraction method. Transform matrixes such as a homography matrix (eight degrees of freedom (DOF)) or an affine matrix (6 DOF) are estimated using RANdom SAmple Consensus (RANSAC) (Torr and Murray, 1997) over the set of matching candidates to minimize model estimation error due to the outlier feature

Lim Y., Lee C. and Kim J. (2013).
Stereo-based Spatial and Temporal Feature Matching Method for Object Tracking and Distance Estimation.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 492-496
DOI: 10.5220/0004333904920496

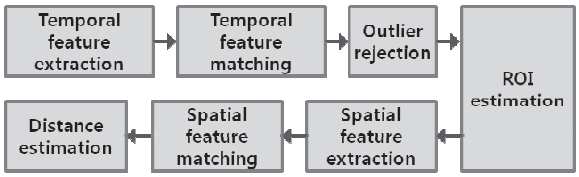

Figure 1: Block diagram of the proposed method.

pairs. However, if the number of outlier features increases due to the misaligned ROI, the RANSAC algorithm fails to estimate the model parameters correctly.

Most feature-based object tracking methods were used in the 2D image plane. In these methods, projection errors may occur when the motion parameters of the features are expressed by a transform matrix. Feature tracking and feature clustering in 3D are regarded as the similar problem (Agrawal et al., 2005). In this paper, we propose an object tracking method which uses an integrated spatial and temporal feature matching scheme. The proposed method offers enhanced tracking performance by means of temporal feature matching while accurately estimating the object distance by means of spatial feature matching.

The rest of our paper is organized as follows. We describe proposed method in Section 2. First, we introduce an overview of our method and the spatial and temporal feature matching method. Experimental results and an analysis of real-world image sequences are presented in Section 4. Finally, Section 5 concludes this paper and discusses future works.

2 THE PROPOSED METHOD

2.1 Overview of the Proposed Method

Our framework consists of temporal feature extraction, temporal feature matching, outlier rejection, ROI estimation, spatial feature extraction, spatial feature matching, and distance estimation, as shown in Figure 1. Before the spatial feature matching process, calibration and rectification processes are required to align the epipolar line. Temporal features are extracted in the ROI of the previous reference image and candidate region of the current reference image using the FAST detector (Rosten et al., 2010). The BRIEF method (Calonder et al., 2012) is used as a descriptor of the FAST features due to its speed and robustness. Temporal features are matched by the Hungarian algorithm (Kuhn, 1955). The outliers are removed, and the transform matrix and ROI are estimated using the prior disparity information of the features and the RANSAC algorithm. Spatial features are extracted in the search region of the current corresponding image and then matched between left and right features. The epipolar constraint and prior disparity information reduce the spatial matching errors. Finally, the 3D position of the tracking object is calculated using an inverse perspective map (IPM) (Lim et al., 2010).

2.2 Temporal and Spatial Feature Matching

Temporal features are extracted in the search regions of the current reference images using the FAST detector. The search region ($\Re_t^T$) of temporal features is determined by the motion model of the target object (Figure 2).

$$X_t^T = X_{t-1}^T + V_{t-1}^T \Delta t,\ \ S_t^T = S_{t-1}^T + S_{v_{t-1}}^T \Delta t,$$
$$\Re_t^T = \left(X_t^T, S_t^T\right), X_t^T = \left(u_t^T, v_t^T\right), V_t^T = \left(v_{u,t}^T, v_{v,t}^T\right), \quad (1)$$
$$S_t^T = \left(w_t^T, h_t^T\right), S_{v_t}^T = \left(s_{v_{x,t}}^T, s_{v_{y,t}}^T\right),$$

X_t^T and V_t^T respectively denote the center of the search region and the velocity of the object's motion, and S_t^T and $S_{v_t}^T$ likewise denote the scale (width and height) of the search region and the variance of the ROI size. Additionally, Δt is the frame rate. The BRIEF descriptor is used to match the features of reference and corresponding features. The Hungarian algorithm is used for the globally optimal one-to-one feature matching.

$$\hat{A} = \arg\min_A \sum_{j=1}^{n}\sum_{i=1}^{m} c_{ij} a_{ij},$$
$$C = \begin{bmatrix} c_{11} & \cdots & c_{1m} \\ \vdots & \vdots & \vdots \\ c_{n1} & \cdots & c_{nm} \end{bmatrix}, A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}, \quad (2)$$
$$\sum_{i=1}^{n} a_{ij} = 1\ , \sum_{j=1}^{m} a_{ij} = 1\ ,$$

In this equation, C is the cost matrix and A is the assignment matrix, which should be mutually exclusive. The cost matrix is calculated by the Hamming distance between the descriptors of the two features. A transform matrix is estimated using matched feature pairs.

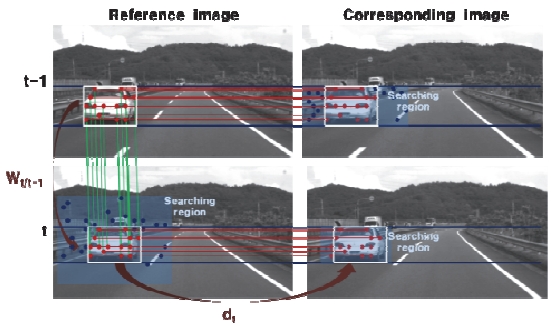

Figure 2: Spatial and temporal matching method.

$$\overline{W}_{t/t-1} = \arg\min_{W_{t/t-1}} \sum_{k=1}^{p} \left(W_{t/t-1} f_{t-1,R}^{k} - f_{t,R}^{k} \right)^2, \quad (3)$$

where $f_{t-1,R}^{k}$ and $f_{t,R}^{k}$ are the k^{th} matched features in the previous and current reference images, and $W_{t/t-1}$ denotes the affine transform matrix. The RANSAC algorithm and prior disparity information are used to remove any outlier features and to estimate the ROI more robustly.

Spatial features are extracted in the ROI region of the current reference image and in the search region of the current corresponding image. The search region of the spatial features is determined in a valid disparity range ($\lfloor d_{t,\min}, d_{t,\max} \rfloor$) using the predicted object distance ($\bar{z}_t$) and velocity ($\hat{v}_{z_{t-1}}$) (Figure 2),

$$d_{t,\min} = round\left(\frac{b\alpha}{\bar{z}_t - \sigma_z} \right), \quad d_{t,\max} = round\left(\frac{b\alpha}{\bar{z}_t + \sigma_z} \right),$$
$$\bar{z}_t = \hat{z}_{t-1} + \hat{v}_{z_{t-1}} \Delta t, \quad \hat{z}_{t-1} = \frac{b\alpha}{\hat{d}_{t-1}} \quad (4)$$

where b and α are the baseline of the stereo camera and the focal length expressed in pixel units, and $\hat{d}_{t-1}$ and σ_z correspondingly represent the disparity estimated in the previous images and the reliability of the predicted distance. In spatial feature matching, a 2D corresponding search problem can be reduced to a 1D searching problem due to the epipolar constraint.

$$\hat{d}_t = \arg\min_{d_t} \sum_{i=1}^{l} \sum_{j=1}^{m} \left(f_{t,R,x}^{i,j} - d_t - f_{t,C,x}^{i,j} \right)^2, \quad (5)$$

where $f_{t,L,x}^{i,j}$ and $f_{t,R,x}^{i,j}$ denote the horizontal position of the matched feature of the i^{th} row and the j^{th} column in the reference and corresponding images.

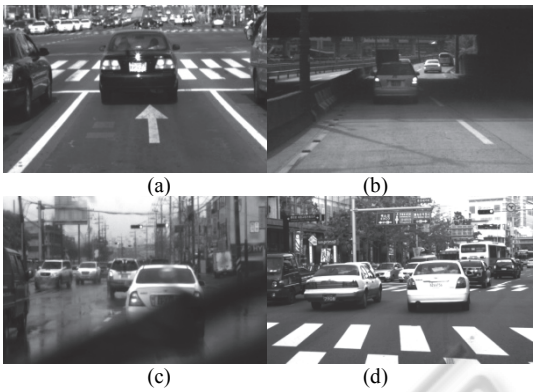

Figure 3: Test datasets: (a) scene 1 (total 101 frames): size change, (b) scene 2 (total 177 frames): illumination change, (c) scene 3 (total 100 frames): partial occlusion due to windshield, (d) scene 4 (total 200 frames): cluttered environment.

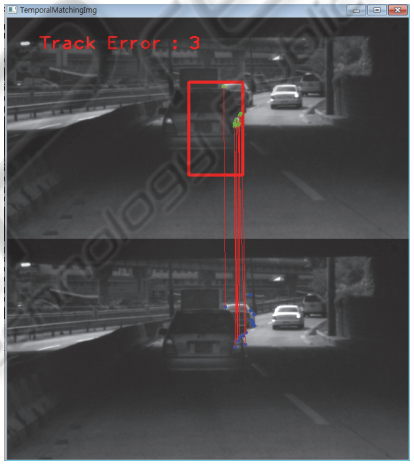

Figure 4: Tracking error in scene 2 due to the severe lighting condition.

The global position of the object in the current frame is calculated by the following equation:

$$\hat{z}_t = \frac{b(\alpha \cos\theta - y_d \sin\theta)}{d}$$
$$y_d = y_{pl} - y_0 = y_{pr} - y_0. \quad (6)$$

where $\hat{z}_t$ is the longitudinal distance of the target object, y_{pl} is the vertical positions of the left image coordinates, y_0 is the optical center, θ is the angle between the Z direction and the optical axe of the cameras, α is the focal distance expressed in the units of pixels.

4 EXPERIMENTAL RESULTS

We implement our method with visual C++ 9.0 and the OPENCV 2.2 library. Our method is tested and verified by test datasets which are captured from

Table 1: Experimental results.

	Scene 1	Scene 2	Scene 3	Scene 4
# of Failure	0	5	0	0
Precision	0.726	0.683	0.803	0.855
Process time	483 ms	49.1 ms	153 ms	275 ms

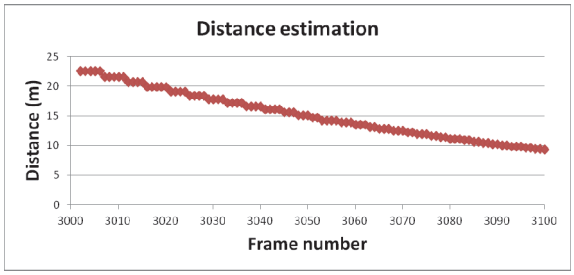

Figure 5: Distance estimation (scene 1).

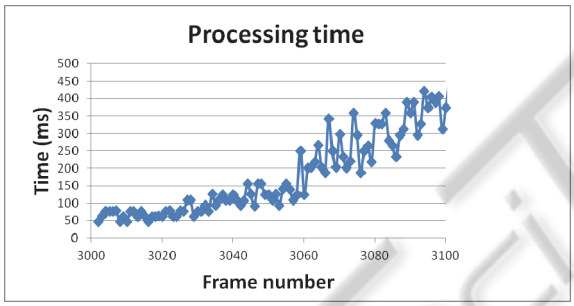

Figure 6: Processing time (scene 1).

a few real and challenging road environments, as shown in Figure 3. The moving vehicles are manually initialized in the first frame, after which the trackers estimate the ROI of the target object. If a tracker fails to estimate the position of the target, the errors are counted and the ROI is reinitialized by the ground truth.

As shown in Table 1, the experimental results demonstrate that our method demonstrates robust tracking performance except in scene 2. In scene 2, our method often fails to track objects in the tunnel (Figure 4). Our method has a shortcoming under severe lighting conditions. The processing times are highly dependent of the number of features. The distance estimation results and the processing time for scene 1 are illustrated in Figure 5 and Figure 6, respectively.

5 CONCLUSIONS

In this paper, we proposed a stereo-based spatial and temporal matching method that can track an object robustly and estimate its global position accurately without dense stereo matching processing. Our experimental results verified that the proposed method is capable of accurately estimating distances and robustly tracking objects. However, severe illumination often causes tracking failures. In addition, the processing time increases drastically if the number of features increases. Our future work will center on a more robust feature matching algorithm and methods that reduce the processing time.

ACKNOWLEDGEMENTS

This work was supported by the DGIST R&D Program of the Ministry of Education, Science and Technology of Korea.

REFERENCES

Agrawal, M., Konolige, K., and Iocchi, L., 2005. Real-Time Detection of Independent Motion using Stereo. in Proc. of IEEE Workshop on Motion and Video Computing, Vol. 2, pp. 207-214.

Bay, H., Ess, A., Tuytelaars, T., and Gool, L. V., 2008. SURF: speeded up robust features. Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346-359.

Calonder, M., Lepetit, V., Özuysal, M., Trzcinski, T., Strecha, C., and Fua, P., 2012. BRIEF: Computing a Local Binary Descriptor Very Fast. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34 no.7, pp. 1281-1298.

Harris, C. and Stephens, M., 1988. A combined corner and edge detector. In Alvey Vision Conference, pp. 147-151.

Lim, Y. C., Lee, Mh, Lee, C. - H., Kwon, S., and Lee, J.-H., 2010. Improvement of stereo vision-based position and velocity estimation and tracking using a stripe-based disparity estimation and inverse perspective map-based extended Kalman filter. Optics and Lasers Engineering. vol. 48, no.9, pp. 859-868.

Jianbo, S. and Tomasi, C., 1994. Good features to track. In Proc. of IEEE Computer Vision and Pattern Recognition, pp. 593-600.

Kuhn, H. 1955. The Hungarian method for the assignment problem. Naval Research Logistic, vol. 2, pp. 83-97.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110.

Rosten, E., Porter, R., and Drummond. T., 2010. Faster

and Better: A Machine Learning Approach to Corner Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no.1, pp. 105-119.

Torr, P. H. S. and Murray, D. W., 1997. The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix. International Journal of Computer Vision, vol. 24, no. 3, pp. 271-300.