Expert Systems With Applications 000 (2015) 1–10

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Expert Systems with Applications
An International Journal
Editor-in-Chief
Binshan Lin

A multi-vision sensor-based fast localization system with image matching for challenging outdoor environments

Jongin Son, Seungryong Kim, Kwanghoon Sohn*

The School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Republic of Korea

ARTICLE INFO

Keywords:
Localization
Image matching
Affine invariance
Descriptors
Intensity-based

ABSTRACT

A sensor-based vision localization system is one of the most essential technologies in computer vision applications like an autonomous navigation, surveillance, and many others. Conventionally, sensor-based vision localization systems have three inherent limitations, These include, sensitivity to illumination variations, viewpoint variations, and high computational complexity. To overcome these problems, we propose a robust image matching method to provide invariance to the illumination and viewpoint variations by focusing on how to solve these limitations and incorporate this scheme into the vision-based localization system. Based on the proposed image matching method, we design a robust localization system that provides satisfactory localization performance with low computational complexity. Specifically, in order to solve the problem of illumination and viewpoint, we extract a key point using a virtual view from a query image and the descriptor based on the local average patch difference, similar to HC-LBP. Moreover, we propose a key frame selection method and a simple tree scheme for fast image search. Experimental results show that the proposed localization system is four times faster than existing systems, and exhibits better matching performance compared to existing algorithms in challenging environments with difficult illumination and viewpoint conditions.

1. Introduction

A localization system is a technology that estimates current location in dorter to run an autonomous navigation systems in cars and an unmanned monitoring robots both safely and consistently (Abdelhafez et al., 2008; Choi, Park, Kim, & Choe, 2012; Royer, Lhuillier, Dhome, & Lavest, 2008; Wang, Zha, & Copolla, April 2006). Conventionally, its performance depends mainly on estimating global positioning system(GPS) information. Various factors, however, such as the cover, radio wave reflection, and interference in the open field with obstacles influence GPS estimation performance. To solve this problem with GPS and increase localization accuracy, localization systems with various kinds of vision and GPS sensors have been popularly developed (Choi, Park, Song, & Kweon, 2011; Hays & Efros, 2008; Schindler, Brown, & Szeliski, 2007; Zamir & Shah, 2010). A localization system based on vision sensors is generally composed of two parts, the environmental map generation stage and the location estimation and correction stage (Choi et al., 2011).

First, the environmental map generation stage aims to describe and establish a target environment and proper data form, based on the both GPS and visual information. The most important part of this stage is the compaction of the large amount of images associated with the location information. In other words, a large amount of complex images and location information should be compactly packaged so as to reduce any potential issues involving high computational complexity.

Second, the location estimation and correction stage aims to find the target image that the most similar to the query image from the environmental maps. That is, the image matching technology used to estimate similarities between images generated from the environmental map and the query image is the most important part. For these reasons, illumination and viewpoint variations are the most important issues that must be overcome in order to provide satisfactory localization performance in the localization system.

In this paper, we propose a robust localization system that can be used to solve viewpoint and illumination variation problems while maintaining low computational complexity. The main contributions of this paper can be summarized as follows:

- Robustness in outdoor environments: we have developed a robust feature extraction method using virtual view images under changing viewpoint conditions and a feature description method that can be used to improve HC-LBP under changing scale and illumination conditions.
- Low computational complexity: we propose a fast image retrieval method for the localization system based on the key frame selection and the search range tree.

* Corresponding author. Tel: +82 2 2123 2879.
E-mail addresses: go3son@yonsei.com (J. Son), srkim89@yonsei.ac.kr (S. Kim), khsohn@yonsei.ac.kr (K. Sohn).

http://dx.doi.org/10.1016/j.eswa.2015.07.035
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

Please cite this article as: J. Son et al., A multi-vision sensor-based fast localization system with image matching for challenging outdoor environments, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.035

We evaluated our proposed localization system against various experimental datasets including simulation results for indoor and outdoor datasets totaling 42,933 frames.

The paper is organized as follows. We introduce related works for conventional image matching and localization systems in Section 2. In Section 3, we propose a robust image matching algorithm that is optimized to the proposed localization system. Finally, Section 4 shows the experimental results for challenging environments both indoors and outdoors, followed by conclusions in Section 5.

2. Related works

2.1. Visual image correspondence

2.1.1. Feature extraction

In general, image matching algorithms are composed of three steps: the feature extraction step, the feature description step, and the matching step. Firstly, the feature extraction step aims to determine reliable and repeatable key-points on an image. In the literature, these are, called interest points or feature points. Conventionally, Harris corner detection (Harris & Stephens, 1988), which is based on the response of a structural tensor, was the most popular key-point detector. Taking their cue from Harris detector, many methods have been proposed to detect affine invariant regions around points; these methods, including Harris–Affine (Mikolajczyk & Schmid, 2004) and Hessian–Affine detector (Mikolajczyk & Schmid, 2005). Maximally Stable Extremal Regions (MSERs) (Matas, Chum, Urban, & Stereo, 2002) was also proposed for the determinations of affine invariant key-points. The Anisotropic Binary Feature Transform (ABFT) framework was also proposed based on structure tensor space (Kim, Yoo, Ryu, Ham, & Sohn, 2013). However, these affine invariant detectors cannot find reliable regions due to difficulties in localization. The Scale Invariant Feature Transform (SIFT) has been one of the most popular approaches due to its high robustness under various environments (Lowe, 2004). It detects Different of Gaussian (DoG) points, which approximate the Laplacian of Gaussian (LoG). In order to reduce computational complexity, Bay et al. proposed the Speeded-Up Robust Features (SURF) algorithm (Bay, Ess, Tuytelaars, & Gool, 2008), which approximates to SIFT and outperforms other existing methods. Although these conventional algorithms show satisfactory performance, they still have high computational complexities. Recently, Rosten et al. proposed the Features from Accelerated Segment Test (FAST) feature detector (Rosten, Porter, & Drummond, 2010). Even though it provides satisfactory results under low geometric deformations, they have limitations for severe geometric deformation such as affine variations. To estimate affine invariant feature point, reliably Guoshen et al. have proposed a fully affine invariant framework, i.e., Affine-SIFT (ASIFT) (Guoshen & Morel, 2009) based on the matching in fully affine space. Although the SIFT has shown in reliable matching for various affine variations, it also provides dramatically many outliers and requires a high computational complexity. To overcome the problems of ASIFT, Yu, Huang, Chen, and Tan (2012) proposed the iterative solver to find homography matrix of two images, which the reference image is then matched with the simulated image. In Chen, Shao, Li, and Liu (2013), local stable regions are extracted from the reference image and the query image, and transformed to circular areas according to the second-order moment. However, these methods still require high computational complexities.

2.1.2. Feature description

The feature description step aims to describe each key-point on an image as a distinctive vector that represents the local support window of each key-point. The SIFT (Lowe, 2004) and the SURF (Bay et al., 2008) descriptor were based on the orientation of gradient histogram. Calonder (2011) proposed the Binary Robust Independent Elementary Features (BRIEF) feature descriptor from intensity variation tests. The combination of FAST detection and BRIEF description has been popular since they provide satisfactory performance and low computational complexity (Heinly, Dunn, & Frahm, 2012). In addition, Rublee, Rabaud, Konolige, and Bradski (2011) proposed the Oriented FAST and Rotated BRIEF (ORB), which addresses the rotation variant problem of BRIEF. Leutenegger, Chli, and Siegwart (2011) also proposed the Binary Robust Invariant Scalable Keypoints (BRISK), which is a scalespace FAST detector in combination with bit-string descriptors. Alahi, Ortiz, and Vandergheynst (2012) proposed the Fast Retina Keypoint (FREAK) inspired by the human visual system. However, they have limitations under various illumination conditions and changing viewpoint conditions. To provide the illumination robustness, local binary pattern (LBP) (Guo, Zhang, & Zhang, 2010) was proposed based on the intensity comparison. Furthermore, based on LBP, center-symmetric LBP (CS-LBP) and Haar-like Compact LBP (HC-LBP) (Kim, Choi, Joo, & Sohn, 2012) have been proposed. Multi Support Region Order Based Gradient Histogram (MROGH) was proposed based on overlapping regions using multiple support regions combined by intensity order pooling (Fan, Wu, & Hu, 2012). Furthermore, Local Intensity Order Pattern (LIOP) uses the intensity order pooling and the relative order of neighbor pixels to define the histogram (Wang, Fan, & Wu, 2011). Recently, Ballavia, Tegolo, and Valenti (2014) proposed shifting gradient local orientation histogram (sGLOH) to improve the discriminative power of histogram-based key point descriptors. To estimate the correspondence between multi-modal images, the Local Self-Similarity (LSS) (Shechtman & Irani, 2007) and LSS frequency (LSSF) (Kim, Ryu, Ham, Kim, & Sohn, 2014) descriptor were proposed based on the local internal layout of self-similarities. More recently, the Dense Adaptive Self-Correlation (DASC) descriptor was proposed to estimate dense correspondence under multi-modal and multi-spectral variations (Kim et al., 2015).

2.2. Vision-based localization systems

Recently, vision-based localization systems based on image matching, have been proposed (Choi et al., 2011; Hays & Efros, 2008; Knopp, Sivic, & Pajdla, 2010; Schindler et al., 2007; Zamir & Shah, 2010). These popular methods, measure the correspondence between a reference dataset of images containing GPS information and a query image in order to estimate the location of the query image. Choi et al. (2011) proposed a localization system based on the feature matching of SURF (Bay et al., 2008) and a KD-tree. Hays and Efros (2008) proposed a method for extracting coarse geometric information from a query image. Zamir and Shah (2010) utilized online public images as the reference database, using, a feature pruning method that incorporated geometric information, they were able to locate incorrectly matched features. Schindler et al. (2007) designed a scale localization system based on the bag of visual words model. It improves the conventional methods using a searching vocabulary tree. (Sattler, Leibe, & Kobbelt, 2010; 2012) designed a framework for identifying multi dimensional correspondences between the query and the reference database containing a large number of user shared images. These methods, however, are also limited to certain illumination and viewpoint conditions.

3. Illumination and viewpoint invariant localization system

3.1. Problem statement and overview

Fig. 1 shows a block diagram of the overall proposed system. The multiple vision sensors are mounted top of unmanned vehicles. The proposed system is composed of an environmental map generation stage and a location estimation and correction stage. In the environmental map generation stage, first, we use three images obtained from three cameras and GPS data to generate the environmental map.

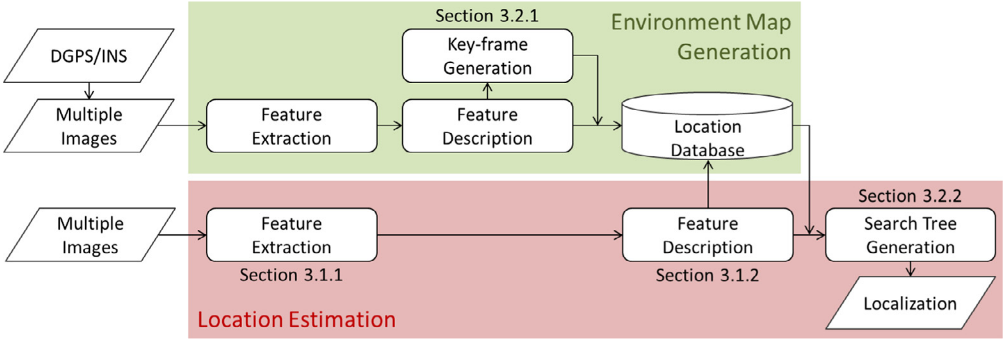

Fig. 1. Block diagram of the proposed localization system.

These are then merged into a single image by image stitching. Then, we extract the key feature point and feature point descriptor, These are, typically not robust to illumination and view point changes. Finally, we build the hierarchical environmental map using the relationships between feature points. The environmental map consists of key frames and non-key frames. As a representation of video sequence using proposed feature extraction and matching under time domain, which will be described in Section 3.3. For location estimation, the feature points of the query image are matched with the key frames from the environmental map. It matches the key frames with the non-key frames near them, and then finds the image that has the most similar feature points in the environmental map. We estimate the location using the correspondence between the matched database image and the query image.

3.2. Robust image matching

3.2.1. Virtual view feature extraction

We can establish the acquisition model for the camera images, have slight viewpoint shifts, as illustrated by the camera model (Guoshen & Morel, 2009; Juan & Gwun, 2009) and digital image acquisition. Let an image be $f : \mathcal{I} \rightarrow \mathbb{R}$ or $\mathbb{R}^3$, where $\mathcal{I} = \{i = (x_i, y_i)\} \subset \mathbb{N}^2$ is a discrete image domain. Assume that the image f is transformed by any affine transformation matrix A, which is representative of a viewpoint change, and can be characterized as follows:

$$A = H_\lambda R_1(\psi) T_t R_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}, \quad (1)$$

where ϕ and t are the latitude angles of the camera's optical axis and transition tilt, respectively. The ψ angle is the camera spin, and λ represents the zoom parameter. Since the tilt can be represented as $t = 1/\cos\theta$ for $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, the affine transformation model is controlled by two variations including θ and ϕ (Alahi et al., 2012). A more detailed definition was reviewed in (Alahi et al., 2012). This affine transformation matrix induces the coordinate of original image as

$$\mathcal{I}' = A \cdot \mathcal{I}. \quad (2)$$

Using this relationship, the affine transformed image f', can be derived from the original image,f ,using the affine matrix A, for which $f' : \mathcal{I}' \rightarrow \mathbb{R}$ or $\mathbb{R}^3$.

To build an affine image, we employ constant sampling along the longitude and latitude lines. Latitude θ is sampled using a geometric progression (e.g., 1, a, a^2,..., a^n, $a > 1$) and longitude ϕ is sampled in an arithmetic progression (e.g., 0, b/t,..., b/t, $b \simeq 72°$). The interval of sampling is established as $a = \sqrt{2}$, $kb/t < 180°$, $t_{\max} \approx \sqrt[4]{2}$ in light of the exactness and efficiency of creating a viewpoint in third dimensional coordinates, as shown in Fig. 2(a) (Guoshen & Morel, 2009). Note that we do not need to rotate all 180° since we use multiple-cameras with left and right viewing angles. Because of this, the complexity is significantly reduced due to the amount of matching by expanding from 30° to 60°, changing the sampling interval of the latitude and longitude from 7-levels to 3-levels, and only using the virtual viewpoint image of the one-sided image.

Fig. 2 (b) shows the virtual image optimized in the proposed localization system. To detect the feature point in the virtual viewpoint image, we employ the integral image scheme used in the SURF detector (Bay et al., 2008). Although our detection looks similar to ASIFT, it has an explicit advantage over ASIFT. Unlike ASIFT, we develop a feature extraction method that merges the integral image and Hessian matrix, and depending on the image resolution, we are able to eliminate the 3 × 3 sub-sampling methods. We will compared the proposed method with ASIFT the experimental section.

3.2.2. Scale invariant HC-LBP description

It self intensity order based local features rather than raw intensity have been proposed based on the observation that the intensity orders between pixels are invariant to monotonic changes in intensity (Kim et al., 2012). However, these methods are limited in terms of their scale change. To alleviate this problem, we propose a scale invariant Haar-like LBP descriptor, where the most appropriate scale is integrated according to the size of the region that constitutes the four descriptors. Specifically, for neighboring pixel q of center pixel p, we calculate the average intensities of 4 regions - up, down, left, and right, as shown in Fig. 3, and then calculate the difference between the mean of the up and down regions as $d_{ud}(q) = m_u(q) - m_d(q)$ and $d_{lr}(q) = m_l(q) - m_r(q)$, where $m_u(q)$, $m_d(q)$, $m_l(q)$ and $m_r(q)$ are the average intensities of the up, down, left, and right regions, respectively. $d_{ud}(q)$ is the difference between the means of the up and down regions, and $d_{lr}(q)$ is the difference between the means of the left and right regions. Comparing the two differences, we assign a pattern code Φ based on the code criteria as in Algorithm 1. In this way, we assign two codes $\Phi_1(q)$ and $\Phi_2(q)$ for a given pixel q. The number of possible codes is 8 as shown in Fig. 3. Although the proposed modified binary pattern can describe a point using simple code, statistical information around the feature point enables a powerful description with respect to illumination variation.

Finally, a robust grid pooling scheme is employed which subdivides the region around the feature point into 16 sub-regions as shown in Fig. 4; this is similar to a SIFT descriptor. A code histogram is established in each sub-region. Each region then makes an 8-level histogram., A 128-level histogram is made by connecting individual sub-region histograms. Finally, we can obtain feature descriptor $v(p)$ for pixel p.

The proposed method can reduce the memory usage of the HC-LBP notion of how to scale up lighting. This is, in addition to proposing a scale that considers Scale Invariant Haar-like feature - Local Binary

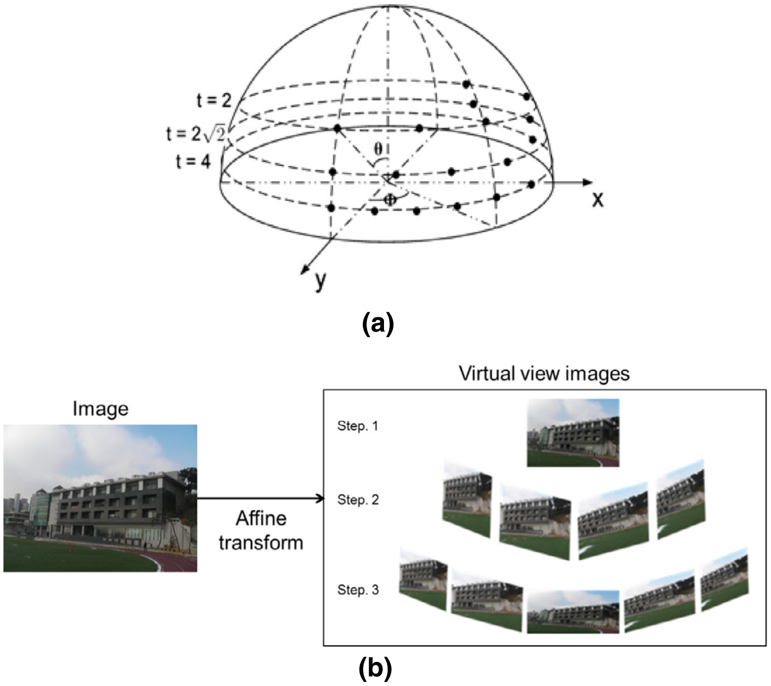

Fig. 2. Extraction from a feature point with insignificant changes after a viewpoint change. (a) The virtual viewpoint of the sampled affine movie and (b) the key map for the extraction of affine features.

Pattern (SIH-LBP). The proposed descriptor generation method can solve for both the various scales and changing illumination.

Algorithm 1. Scale-invariant HC-LBP descriptor

1: If both $d_{ud}(q)$ and $d_{lr}(q)$ are smaller than a certain threshold value, the pixel q is in a homogeneous region. We only assign a code if at least one difference is larger than the threshold.

$$d_{\max}(q) = \max\left(|d_{ud}(q)|, |d_{lr}(q)|\right) > T_{nd} \tag{3}$$

2: If d_{ud} is larger than d_{lr}, the parallel gradient is larger than the vertical gradient at point p. Otherwise, the vertical gradient is dominant at point p. According to the maximum between d_{ud} and d_{lr}, we classify the two cases.
3: After determining the dominant orientation between parallel and vertical, we assign the first code for point p according to the sign of a larger d.

$$\Phi_1(q) = \begin{cases} U(q), & d_{\max}(q) = |d_{ud}(q)| \& d_{up}(q) > T_{nd} \\ D(q), & d_{\max}(q) = |d_{ud}(q)| \& d_{up}(q) < -T_{nd} \\ L(q), & d_{\max}(q) = |d_{lr}(q)| \ \& d_{lr}(q) > T_{nd} \\ R(q), & d_{\max}(q) = |d_{lr}(q)| \ \& d_{lr}(q) < -T_{nd} \end{cases} \tag{4}$$

4: The second code is assigned using a smaller difference. If we compare the averages relating to a smaller difference, we assign a second code as follows:

$$\Phi_2(q) = \begin{cases} U(q), & \left|d_{ud}(q)\right| < |d_{lr}(q)| \& m_u(q) > m_d(q) \\ D(q), & |d_{ud}(q)| < |d_{lr}(q)| \& m_u(q) < m_d(q) \\ L(q), & |d_{ud}(q)| > |d_{lr}(q)| \& m_l(q) > m_r(q) \\ R(q), & |d_{ud}(q)| > |d_{lr}(q)| \& m_l(q) < m_r(q) \end{cases} \tag{5}$$

3.3. Location estimation

3.3.1. Key frame selection

In order to reduce memory and processing time, we apply key frame selection, which is used to process the third dimensional models using minimal structural calculations of the motion of the key frame (Ahmed, Dailey, Landabaso, & Herrero, 2010).

We match correspondence points between the nearest key frame and the current query image using the proposed feature extraction and matching algorithms. We then select this image as the key frame when the number of corresponding points is below a threshold value or when the number of correspondence points, which is the next nearest to the key frame, is few. We do this because the situation is recognized as the image that include many new feature points.

For example, we define the first image a key frame 1. Key frame 2 is at least T_{n1} of common interest points with key frame 1. Key frame n is at least T_{n1} of common interest points with key frame n-1 and at least T_{n2} of common interest points with key frame n-2. In other words, T_{n1} means that the image a matches the feature number with the previous key frame, T_{n2} denotes a matching feature number with the before previous frame.

Fig. 5 shows an example of key frame selection. 13 images are chosen as key frames among 1600 images. The interval for key frames is set to approximately 100 images, which proves to be ten times faster for full searches after implementing key frame and non-key frame search.

3.3.2. Image search based on image matching

To detect which image is the most similar image to the query image in a set of environmental images, the feature points of the images in the environmental map are matched to the feature points of the query image. The environmental image that is nearest to the query image is the one that has the largest number of positive matches in the environmental map. Provided that v_p^c is the descriptor of the p feature point extracted through the current image c and $v_{p'}^e$ is the descriptor of the p' feature point extracted through the e image of the environmental map.

$$(\bar{c}, \bar{p}) = \underset{(c,p')}{\arg\min} \left(\left\| v_p^c - v_{p'}^e \right\|^2 \right) \tag{6}$$

If the $(\bar{c}, \bar{p})$ which decide all of the feature points of the current image is c, the search of the Nearest Environmental Map (NEM) Ψ is

Please cite this article as: J. Son et al., A multi-vision sensor-based fast localization system with image matching for challenging outdoor environments, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.035

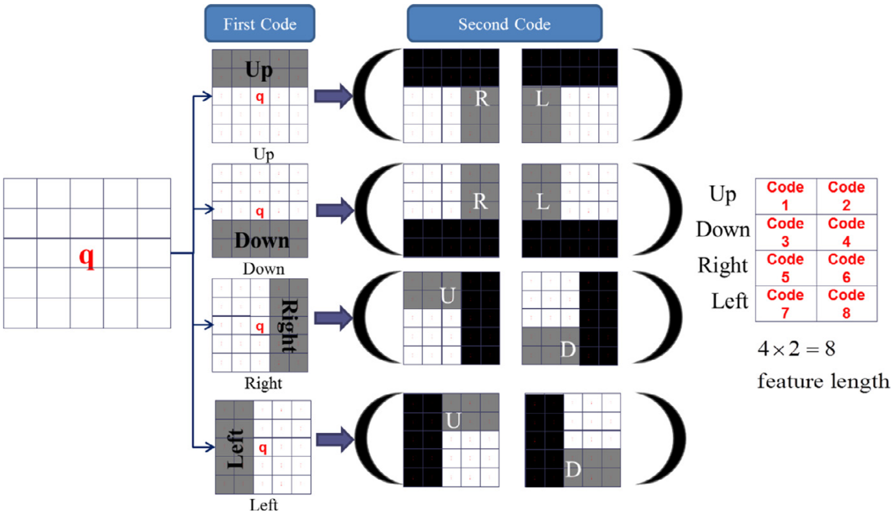

Fig. 3. Four regions of a 5 × 5 patch and modified binary pattern.

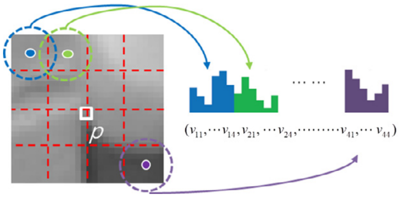
p
$(v_{11},\cdots v_{i1}, v_{21},\cdots v_{2i},\cdots\cdots\cdots v_{41},\cdots v_{4i})$
……

Fig. 4. Compact descriptor generation.

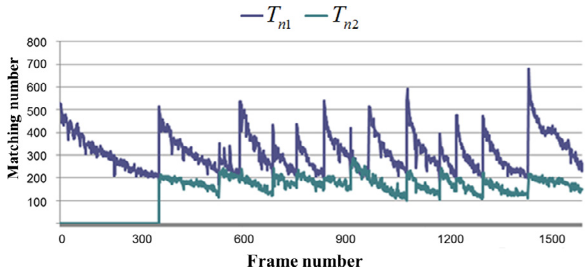

Fig. 5. The application result of the key frame selection.

as follows:

$$\Psi = \max_{c}\left(\sum_{p} a_{(\bar{c},\bar{p})}\right),$$

$$a_{(\bar{c},\bar{p})} = \begin{cases} 1 & (\bar{c},\bar{p}) \in c \\ 0 & (\bar{c},\bar{p}) \notin c \end{cases} \tag{7}$$

It takes a great deal of time to implement a full search that, includes all of the environmental map images and query images, as well as feature point matching. We employ a vocabulary tree(KD-tree method) used from a former localization method. A KD-tree is a useful data structure for several applications, such as searches involving multidimensional search keys, range searches and nearest neighbor searches (Choi et al., 2011). The problem with a KD-tree scheme, however, is that it requires a high degree of computational complexity since it must be configured for every input image. As Shown in Fig. 6(a), the data-saving method can show GPS data and image information. This data structure is comprized of GPS data for latitude, longitude. Also, the flag and number of entire feature points, which indicate whether it is a key frame or not, are recorded in this data structure. Next, the bottom of the data structure consists of 128-dimensional descriptors and the station coordinates (x, y) of each feature point. The virtual coordinates change with the changing view point into real coordinate values and are then saved.

Table 1
Time search feature points from DB.

Number of frames	KD-tree (ms)	Proposed search method (ms)
1000	4.199	2.42
2000	12.393	2.52
5000	32.1	3.12

To overcome these limitations, we design an effective image searching method by the integration of the key frame selection and search range tree. We determine the the key frame with the highest matching value by matching the key frames of an environmental map to the query images. We fix the range from key frame-1/2 to key frame+1/2 in terms of the best key frame matching the most feature points. At this time, we individually change from the current key frame to key frame+1, and from key frame-1 to key frame if the best key frame is the first or the last. We then build the binary tree matching of each frame in a given search range. Finally, we detect the image that, has the most matching values in the last built tree, by constantly matching the only line that has a relatively higher matching value after we have designated the various matching values for each rank.

For example, assume that an experimental map has 1000 frames with key frames = (1,180,372,510,640,780,890,1000) and non key frames = (all frames except key frames). If the best key frame number is 510, we fix our range from 441 to 665. And then we generate the tree using this search range.

Fig. 6(b) shows the framework of image retrieval and matching. Table 1 shows the processing time comparison of the KD-tree and proposed search method using fusing key-frame selection and a search tree. A KD-tree is a space-partitioning data structure for organizing points in a k-dimensional space. In contrast to the proposed system combining key-frame selection and the search range tree, the processing time is reduced dramatically through the optimized configuration. In addition, the results show that the proposed method proves faster than the existing method, and has satisfactory performance.

3.3.3. Location correction

Localization based on an image is performed by the estimation of the relative location between two images. We assume the third dimensional coordinate of the other images as (x', y', z') by calculating Eq. (8) from the third dimensional coordinates (x, y, z) of a certain

Please cite this article as: J. Son et al., A multi-vision sensor-based fast localization system with image matching for challenging outdoor environments, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.035

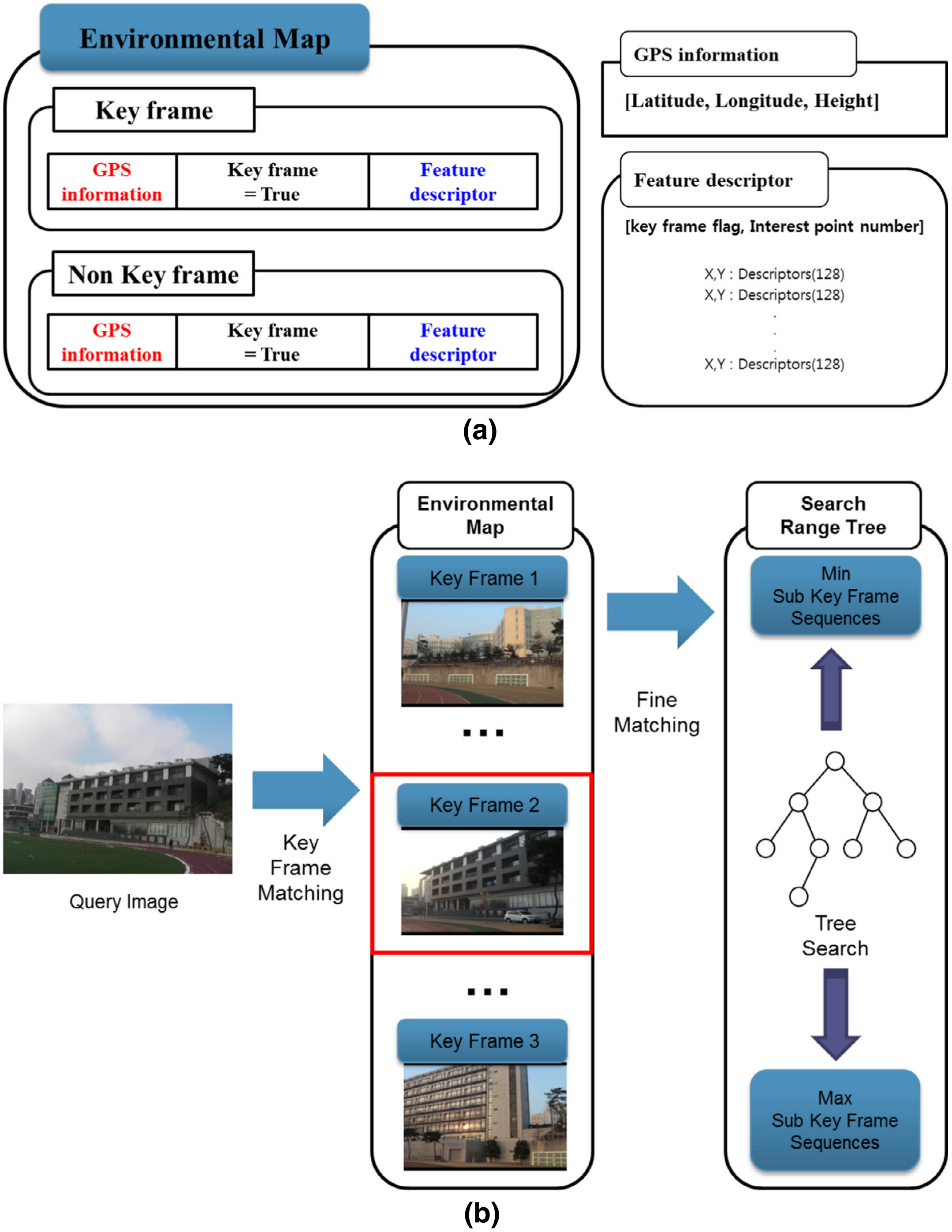

Fig. 6. Sample of location estimation. (a) The structure of environmental map and (b) the process of image retrieval and matching.

image when we already know the rotational transformation matrix R and horizontal transformation vector T, which indicates the geometrical relationship of the two images.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = R \begin{bmatrix} x \\ y \\ z \end{bmatrix} + T, \tag{8}$$

where T and R can be gained by restoring the fundamental matrix F from the corresponding points between the two images. F is a 3×3 matrix, explaining the geometrical attributes received by the two cameras. Its attributes are as follows:

$$x_1 F x_2 = 0, \tag{9}$$

where x_1 and x_2 are the correspondence between the two images. Because F is a 3×3 matrix, we can estimate linearly it by Direct Linear Transform (DLT) when knowing the 8 correspondences. However, it is very susceptible to correspondence matching errors. The algorithm affected by this flaw can't be applied to real system. We must need the unvarying the estimate method of fundamental matrix. Specifically, we ought to exclude wrong outliers and use the method to identify well-informed correspondences only in order not to change its original intent. Random Sample Consensus (RANSAC) is fit for this procedure (Choi, Kim, & Yu, 2009). For example, select random 8 correspondences for each attributes in order to make fundamental matrix, and then calculate the fundamental matrix by using them. To confirm

Table 2
The difference between the conventional system and the proposed system.

System	Camera		Conventional system (Choi's method) Stereo	Proposed system Multiple
Method	Image matching	Feature detector	SURF	Virtual view feature extraction (VVFE)
		Feature descriptor		Scale Invariant Haar-like feature local binary pattern(SIH-LBP)
	Image search		Vocabulary tree (KD-tree)	Key frame selection, Search range tree
	Pose estimation		3-point(P3P) algorithms	8-point algorithms with RANSAC

that the made fundamental matrix can accord with Eq. (9), we should look for error e_i by substituting individual correspondences into Eq. (10).

$$e_i = x_{1i} F x_{2i} \quad (i = 1, 2, \ldots, n) \tag{10}$$

x_{1i} or x_{2i} is the correspondence of two image. We classify these as wrong correspondences when e_i, the calculated error, is above a certain threshold value. After determining which correspondences are wrong, we record the number of wrong correspondences. We then select another 8 random correspondences and calculate the number of wrong correspondences, repeating the process outlined above. Continues until only the most exact correspondences remain. In this way, we calculate the fundamental matrix with minimal errors. Also, the essential matrix is extracted from the fundamental matrices to get both R and T from F. The essential matrix is the matrix without the intrinsic parameters from the fundamental matrix. The essential matrix and fundamental matrix are related as follows:

$$E = K^T F K \tag{11}$$

K is the internal variable matrix of the cameras. We should know this value so that we can look for E, which is one of the fundamental matrices. This resembles Eq. (11), and includes the focal distance of cameras, the ratio of length and width, the form of the CCD and other related information. We can find the relation between the coordinates of image and the real coordinate system using:

$$K = \begin{bmatrix} \alpha_x & s & c_x \\ 0 & \alpha_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

We use the general camera calibration method by calibration patterns. The relative location between the two cameras can be found by substituting K and F into Eq. (12), and dividing this by the known rotational transformation matrix R and horizontal transformation vector T as follows:

$$E = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} R \tag{13}$$

We calculate its relative location after finding an image on the environmental map that is most similar to the input image. Next, we can use GPS to make an estimation of the input image by moving the GPS data of the environmental map into the calculated, rotated, and transformed data since the environmental map contains GPS data.

3.4. Comparison with conventional system

Choi et al. (2011) presented a conventional localization system based on image matching which is a similar to the proposed system. Table 2 compares and analyzes the differences between the two systems. While the conventional system used a stereo camera, the proposed system uses three multi-cameras, meaning that the proposed system has better viewing angles. Moreover, the conventional system uses SURF to create descriptors and extractions from the feature points, potentially decreasing performance if illumination and/or viewpoints change. The proposed system uses the Virtual View Feature Extraction (VVFE) and Scale Invariant Haar-like feature-Local Binary Pattern (SIH-LBP) method, which is a feature point detector and descriptor that is able to account for illumination and viewpoint changes. Moreover, the conventional system, in terms of the localization, uses a vocabulary tree to search images, establishing a larger database with faster performance. To create and even faster search method, the proposed system employs a key frame selective method and binary tree method fixing search range. In addition, the proposed method uses a virtual viewpoint that can extract feature points and carry out 8-point algorithms, allowing for similarities to be determined between 8 images simultaneously.

4. Experimental result

4.1. Experimental environments

To evaluate the performance of the proposed algorithm, we set up a database including one set (743 frames) of internal images with illumination changes, one set (660 frames) of images with viewpoint changes, and six sets (6800 frames, 6713 frames, 6812 frames, 7332 frames, 7221 frames, and 7312 frames) of environmental images of a 6-km drive in the outdoors, that include GPS data. During experimentation, we used three cameras (Point Gray Research's Flea) with a specific type of lens (Avenir's 2.8-mm lens). The size of the experiment's image and data set is 640 × 480. The viewpoint change image set is made by rotating objects leftward, center-ward, and upward, while the illumination change image set is made by changing the brightness three times and through five illuminations. The outside drive image set is composed of two illumination change data sets, one data set according to the environmental map, one illumination, and three viewpoint change data sets. Table 3 shows the information of the test data set in detail.

In order to evaluate the performance of the proposed image matching, we divided true detection and false detection into matching feature points manually. The rate of the matching feature is calculated as the true ratio number of matching points and the total number of matching points (Kim et al., 2012). In addition, the GPS discrepancy was calculated by measuring the difference in distance between the real GPS coordinates of the query image and the measured location coordinates of the query image (Alex, Denis, John, & Eyal, 2008). The proposed system has three thresholds, T_{nd}, T_{n1} and T_{n2}. T_{nd} , which means that the neighboring distance of SIH-LBP is set to 2. Based on key frame selection, the feature matching number T_{n1} is set to 150 and we use $T_{n2} = T_{n1} \times 0.7$. Also, by matching the distance of the proposed and conventional methods, we were able to optimize the values experimentally.

4.2. Comparison of image matching performance

To make objective observations, SURF (Bay et al., 2008), Affine-SIFT (Guoshen & Morel, 2009), FAST (Rosten et al., 2010), and the extraction from the feature point of the proposed virtual view feature extraction are evaluated for a data set having both illumination and viewpoint changes. These are combined into SURF (Bay et al., 2008), CS-LBP (Kim et al., 2012), BRIEF (Calonder, 2011), sGLOH (Ballavia et al., 2014), and the proposed generation method of an illumination

Table 3
The information of experimental dataset.

Dataset		Illumination change (time)	Viewpoint change (degrees)	Number of frames
Indoor	Input image-1	3 lights/10–100 lux	–	743
	Input image-2	–	0–160	660
Outdoor	Experimental map	12:00	0	6800
	Input image-1	9:00	0	6713
	Input image-2	18:00	0	6812
	Input image-3	12:00	30–60	7332
	Input image-4	12:00	90–115	7221
	Input image-5	18:00	60–90	7312

Table 4
The matching result of both the conventional method and the proposed method in response to illumination changes.

Detector	Descriptor	Number of average matching features	Rate of average matching feature (%)	Processing time(s)
SURF	SURF	21	43	0.26
SURF	CS-LBP	13	77	0.22
Affine-SIFT	SURF	321	45	3.3
Affine-SIFT	CS-LBP	311	21	4.3
FAST	BRIEF	98	86	0.05
Harris	sGLOH	92	72	0.92
Propose method VVFE	propose method SIH-LBP	176	92	0.18

Table 5
The matching result of both the conventional method and the proposed method in response to viewpoint changes.

Detector	Descriptor	Number of average matching features	Rate of average matching feature (%)	Processing time(sec)
SURF	SURF	12	40	0.21
SURF	CS-LBP	14	32	0.19
Affine-SIFT	SURF	421	73	3.3
Affine-SIFT	CS-LBP	387	77	3.5
FAST	BRIEF	32	26	0.04
Harris	sGLOH	41	72	0.9
Propose method VVFE	Propose method SIH-LBP	226	90	0.19

Table 6
An estimate of the conventional localization system and the GPS, function, and performance time of the proposed system.

	Conventional localization system		Proposed system	
Condition	GPS discrepancy(m)	Processing time (s)	GPS discrepancy(m)	Processing time (sec)
Illumination change (9:00)	32.1	2.3	4	0.51
Illumination change (18:00)	34.7	2.1	4.1	0.53
Viewpoint change (30–60°)	62.3	2.5	6.2	0.59
Viewpoint change (90–115°)	71.1	2.6	7.1	0.6
Illumination and viewpoint change (18:00, 60–90°)	103.2	2.4	11.2	0.56

robust descriptor. The parameter value of the algorithm used is the most ideal value, and the estimate of the algorithm is implemented by average matching rate and processing time.

Table 4 shows that the proposed method has the highest matching rate in response to illumination changes, outperforming FAST+BRIEF by 6% in terms of matching rate. However, we expect it to perform better in terms of the calibration technology, which analyzes the differences between two images, since the proposed method has 1.7 times more matching capabilities. Moreover, its processing time is found to have the second place score compared to FAST +BRIEF. It is slightly slower by 0.13 s.

Table 5 shows that the proposed method has the best matching rate in response to viewpoint changes, performing 13% better than the second place ASIFT + CS-LBP. Additionally, the proposed method has a higher matching rate while its performance time is just slightly slower by 0.15 s, than FAST + BRIEF, which is 64% more than FAST + BRIEF. In other words, the proposed method has the most reliable matching rate in terms of illumination and viewpoint changes, and is the best in terms of performance time, making it the most suitable localization system.

4.3. Comparison of image localization performance

Fig. 7 shows the comparison results for real GPS data with a localized GPS value according to five conditions. Fig. 7(b and c) are the images obtained with changing illumination; Fig. 7(d and e) are the images obtained after changing the viewpoint; Fig. 7(f) is the image acquired after changing both illumination and viewpoint. Table 6 shows the GPS discrepancy and processing time of the proposed system and localization system. The proposed system decreases the GPS discrepancy by six to ten times and the processing time is decreased by a factor of four compared to the conventional localization system. GPS discrepancy is seen as normal, which is what makes it possible for a robot today to locate its position to within 10 m even in a non-GPS receiving area if the road width is 3 m. Most images, including obvious buildings and characteristics, can be resolved to within 3 m, while error tends to swell in the open fields that have no obvious characteristics. To reduce this weakness of the localization system, the proposed system utilizes a virtual viewpoint change technique, creating the descriptor based on the light from the images, which improves the system's ability to

Please cite this article as: J. Son et al., A multi-vision sensor-based fast localization system with image matching for challenging outdoor environments, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.035

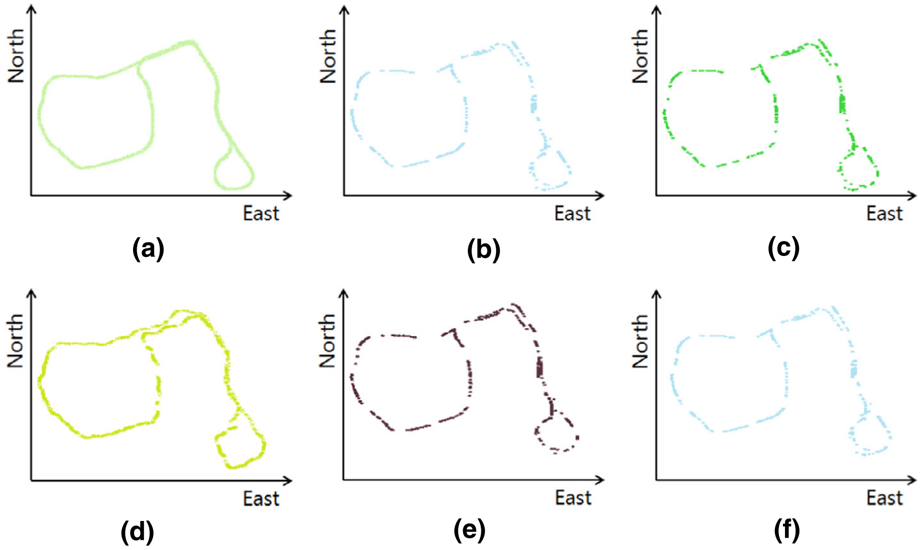

Fig. 7. The localization result. (a) GPS environmental map (12:00), (b) illumination change (9:00), (c) illumination change (18:00), (d) viewpoint change (30–60°), (e) viewpoint change (90–115°), and (f) illumination and viewpoint change (18:00, 60–90°).

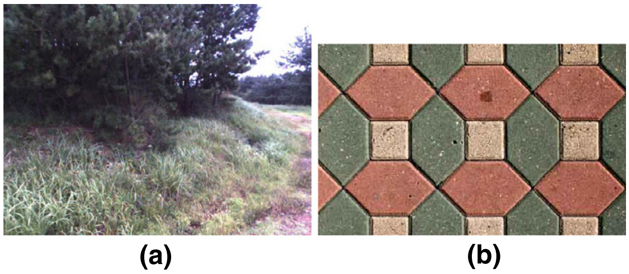

Fig. 8. Examples of limitation images. (a) Bush, (b) Repeat patterns.

handle illumination and viewpoint changes. Its multi-camera improves on the narrow viewing angle of single cameras, minimizing the usage of necessary memory, decreasing processing time and increasing performance. Demonstration simulations of the localization system can be found at http://diml.yonsei.ac.kr/jison/localization.

4.4. Limitations of the proposed system

The failed cases of the proposed method can be divided into two categories. The first case occurs when the image does not have enough features as shown in Fig. 8. The second case occurs due to repeating patterns. These limitations are generally due to local matching methods. In order to combat these limitations, we need to combine local and global matching. Also, we must analyze experimental results. If the matching number is not more than 20, the localization system is deemed to have has poor performance, otherwise, the GPS discrepancy is reduced significantly and system reliability is increased, as seen in Fig. 9.

5. Conclusion

Conventional localization systems based on image matching cannot be successfully used in outdoor environments where illumination and viewpoint variation occur frequently. To help alleviate these concerns, we developed a robust localization system, which is resistant to challenging environments and works with changing illumination and viewpoints.

The contribution of this paper is two-fold; first, we propose a robust feature point extraction method that works using virtual viewpoint changes and is robust to illumination changes. This method features descriptors that improve performance. Second, localization systems based on multiple images have massive amounts of information but perform too slowly due to image sizes and complex algorithms. To combat this, we have designed and optimized these complex algorithms to reduce the processing time of the localization system.

Experimental results show that our proposed localization system improves localization performance in challenging outdoor environments. Moreover, the reduction in computational complexity means that it is four times faster than conventional systems. Also,

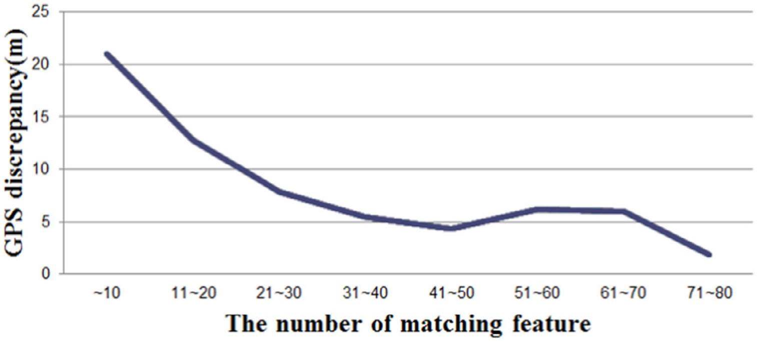

Fig. 9. Average GPS with matching features.

GPS discrepancies have been improved compared with conventional systems.

Even though the proposed system was invariant against various illumination and viewpoint changes, it is still difficult to handle several extreme conditions such as repeated patterns and field environments. These extreme conditions are a common problem for local feature extraction based localization systems. Limited local feature matching is caused by having small features or by having many error features in an image. To overcome limitations of local feature extraction-based localization systems, including those inherent to the proposed system, we will study efficient combinations of using the global matching method. Thus, a combined method using proposed system will be developed in the near future.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2013R1A2A2A01068338).

References

Abdel-hafez, M. F., Kim, D. J., Lee, E. S., Chun, S. B., Lee, Y. J., & Kang, T. S. (2008). Performance improvement of the wald test for GPS RTK with the assistance of INS. *International Journal of Control, Automation, and Systems, 6*, 534–543.

Ahmed, M., Dailey, M., Landabaso, J., & Herrero, N. (2010). Robust key frame extraction for 3D reconstruction from video streams. In *Proceedings of the international conference on computer vision theory and applications (VISAPP)* (pp. 231–236).

Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). FREAK : Fast retina keypoint. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Alex, V., Denis, P., John, K., & Eyal, L. (2008). Calibree: Calibration-free localization using relative distance estimations. In *Pervasive computing*. In *Lecture notes on computer science: Vol. 5013* (pp. 146–161).

Ballavia, F., Tegolo, D., & Valenti, C. (2014). Keypoint descriptor matching with context-based orientation estimation. *Image and Vision Computing, 32*, 559–567.

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. V. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding, 110*(3), 346–359.

Calonder, M. (2011). BRIEF : Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(7), 1281–1298.

Chen, M., Shao, Z., Li, D., & Liu, J. (2013). Invariant matching method for different viewpoint angle images. In *Applied Optics: 52* (pp. 96–104).

Choi, J. H., Park, Y. W., Kim, J., & Choe, T. S. (2012). Federated-filter-based unmanned ground vehicle localization using 3D range registration with digital elevation model in outdoor environments. *Journal of Field Robotics, 29*(2), 298–314.

Choi, J. H., Park, Y. W., Song, J. B., & Kweon, I. S. (2011). Localization using GPS and VISION aided INS with an image database and a network of a ground-based reference station in outdoor environments. *International Journal of Control, Automation and Systems, 9*(4), 716–725.

Choi, S. L., Kim, T. M., & Yu, W. P. (2009). Performance evaluation of RANSAC family. In *Proceedings of the British machine vision conference (BMVC)*.

Fan, B., Wu, F., & Hu, Z. (2012). Rotationally invariant descriptors using intensity order pooling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 2031–2045.

Guo, Z., Zhang, L., & Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. *Image Processing, IEEE Transactions on, 19*(6), 1657–1663.

Guoshen, Y., & Morel, J. M. (2009). ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences, 2*(2), 438–469.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the 4th conference on alvey vision*.

Hays, J., & Efros, A. (2008). Im2gps: Estimating geographic information from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Heinly, J., Dunn, E., & Frahm, J. M. (2012). Comparative evaluation of binary features. In *Proceedings of the IEEE European conference on computer vision*.

Juan, L., & Gwun, O. (2009). A comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing, 3*(4), 143–152.

Kim, B., Choi, J., Joo, S., & Sohn, K. (2012). Haar-like compact local binary pattern for illumination robust feature matching. *Journal of Electronic Imaging, 21*(4) 043014-1–043014-8.

Kim, S., Min, D., Ham, B., Ryu, S., Do, M. N., & Sohn, K. (2015). DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Kim, S., Ryu, S., Ham, B., Kim, J., & Sohn, K. (2014). Local Self-Similarity Frequency Descriptor for Multispectral Feature Matching. In *Proceedings of the IEEE international conference on image processing*.

Kim, S., Yoo, H., Ryu, S., Ham, B., & Sohn, K. (2013). ABFT : Anisotropic binary feature transform based on structure tensor space. In *Proceedings of the IEEE international conference on image processing*.

Knopp, J., Sivic, J., & Pajdla, T. (2010). Avoding confusing features in place recognition. In *Proceedings of the european conference on computer vision*.

Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). BRISK : Binary robust invariant scalable keypoints. In *Proceedings of the IEEE international conference on computer vision*.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints,. *International Journal of Computer Vision, 60*(2), 91–110.

Matas, J., Chum, O., Urban, M., & Stereo, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British machine vision conference (BMVC)* (pp. 414–432).

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision, 60*(1), 63–86.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors.. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27*(10), 1615–1630.

Rosten, E., Porter, R., & Drummond, T. (2010). Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(1), 105–119.

Royer, E., Lhuillier, M., Dhome, M., & Lavest, J. (2008). Monocular vision for mobile robot localization and autonomous navigation. *International Journal of Computer Vision, 74*(3), 237–260.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB : An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE international conference on computer vision*.

Sattler, T., Leibe, B., & Kobbelt, L. (2010). Fast image-based localization using direct 2d-to-3d matching. In *International conference on computer vision*.

Sattler, T., Leibe, B., & Kobbelt, L. (2012). Improving image-based localization by active correspondence search. In *Proceedings of the European conference on computer vision*.

Schindler, G., Brown, M., & Szeliski, R. (2007). City-scale location recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Shechtman, E., & Irani, M. (2007). Matching local self-similarities across images and videos,. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Wang, J., Zha, H., & Copolla, R. (2006). Coarse-to-Fine vision-based localization by indexing scale-invariant features. *IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics, 36*(2), 413–422.

Wang, Z., Fan, B., & Wu, F. (2011). Local intensity order pattern for feature description. In *Proceedings of the international conference on computer vision, ICCV*.

Yu, Y., Huang, K., Chen, W., & Tan, T. (2012). A novel algorithm for view and illumination invariant image matching,. *Proceedings of the IEEE transactions on image processing, 21*, 229–240.

Zamir, A. R., & Shah, M. (2010). Accurate image localization based on google maps street view. In *Proceedings of the European conference on computer vision*.