基于某某方法的改进视觉测距系统：
1.基本信息了解：背景简述，现有局限，改进局限— 三步走：①路口（特征）检测并在线建立landmark库；②任意方向检测到landmark进行机器人方向校正；③“ 全景指南针”……
2.
3.英文表述：单词，介词句式……
4.语言结构与表达；
5.

Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

改进传统VO的增强版VO，通过在线构建和匹配“ 压缩全景地图”的方法来实现，标题的另一种表达形式“ 基于在线构建和匹配…… 的增强视觉测距的实现”

Design of an enhanced visual odometry by building and matching compressive panoramic landmarks online*

Wei LU[†1,2], Zhi-yu XIANG[†‡1,2], Ji-lin LIU[1,2]

([1]*Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China*)

([2]*Zhejiang Provincial Key Laboratory of Information Network Technology, Hangzhou 310027, China*)

[†]E-mail: lwhfh01@zju.edu.cn; xiangzy@zju.edu.cn

Received Apr. 20, 2014; Revision accepted Nov. 24, 2014; Crosschecked Jan. 6, 2015

背景简述：高效精准的定位是移动机器人智能导航的先决条件

现有局限说明：1.累积误差引起的漂移问题；2.光线变化与非正常移动引起的错误运动估计

局限改进方法：实时建库、库匹配与方向校正，位置丢失后作为“ 全景指南针”，压缩感测（？？？）与自适应压缩

Abstract: Efficient and precise localization is a prerequisite for the intelligent navigation of mobile robots. Traditional visual localization systems, such as visual odometry (VO) and simultaneous localization and mapping (SLAM), suffer from two shortcomings: a drift problem caused by accumulated localization error, and erroneous motion estimation due to illumination variation and moving objects. In this paper, we propose an enhanced VO by introducing a panoramic camera into the traditional stereo-only VO system. Benefiting from the 360° field of view, the panoramic camera is responsible for three tasks: (1) detecting road junctions and building a landmark library online; (2) correcting the robot's position when the landmarks are revisited with any orientation; (3) working as a panoramic compass when the stereo VO cannot provide reliable positioning results. To use the large-sized panoramic images efficiently, the concept of compressed sensing is introduced into the solution and an adaptive compressive feature is presented. Combined with our previous two-stage local binocular bundle adjustment (TLBBA) stereo VO, the new system can obtain reliable positioning results in quasi-real time. Experimental results of challenging long-range tests show that our enhanced VO is much more accurate and robust than the traditional VO, thanks to the compressive panoramic landmarks built online.

Key words: Visual odometry, Panoramic landmark, Landmark matching, Compressed sensing, Adaptive compressive feature
doi:10.1631/FITEE.1400139 **Document code:** A **CLC number:** TP391

1 Introduction

Visual localization is a basic and important module for autonomous land vehicles (ALVs), mobile robots, and other intelligent navigation systems. Much progress has been made in this area of science in recent years. Visual odometry (VO) (Nistér *et al.*, 2004; Konolige *et al.*, 2011; Geiger *et al.*, 2012; Lu *et al.*, 2013) and simultaneous localization and mapping (SLAM) (Durrant-Whyte and Bailey, 2006; Sünderhauf and Protzel, 2011) are the two most important methods in the area. In general, VO uses local inter-frame information to estimate motion parameters, while SLAM uses whole sequence information to construct and maintain a global map. As a result, usually SLAM is able to achieve better precision with the cost of much higher computational complexity (Scaramuzza and Fraundorfer, 2011). Munguia and Grau (2007) built a local SLAM to reduce the computational cost, resulting in a system very close to VO, as only the 3D points and their projections within a local short time window are involved. Since only temporary information is kept, these systems can meet the real-time requirement, but typically cannot recover if they suffer from large errors or lose track of the robot's position.

In practice, due to the existence of illumination variation, weak texture, and moving objects, visual localization in some situations is prone to error. Even

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (Nos. 61071219 and 90820306) and the Fundamental Research Funds for the Central Universities, China
ORCID: Wei LU, http://orcid.org/0000-0002-7456-1834; Zhi-yu XIANG, http://orcid.org/0000-0002-3329-7037

if every motion is estimated with high precision, small inter-frame localization errors will accumulate with time, eventually leading to considerable drift. To deal with this problem, a global position correction algorithm (or loop constraints) was presented and researched (Fraundorfer and Scaramuzza, 2012). Its basic idea is to use places of known position to correct the robot's position when the places are revisited.

To recognize the visited scenes, various image matching algorithms have been invented. For a high success rate of matching, the bag-of-words (BoW) (Sivic and Zisserman, 2003; Galvez-López and Tardos, 2012) method is often applied to solve this problem. However, this method usually needs prior knowledge and offline training, which is impractical for a robot exploring unknown areas. In contrast, approaches based on image appearance generally do not need prior knowledge of environments. Se *et al.* (2002) recognized visited places using scale invariant feature transform (SIFT) features, which can also be used to reconstruct 3D maps in SLAM. But this system cannot run in real time due to the high computational cost. Singh and Košechá (2010) used Gabor-Gist descriptors to detect visual loop closure. Liu and Zhang (2012) employed principal component analysis (PCA) to transform high-dimensional Gabor-Gist descriptors into a lower dimensional form to improve computational efficiency and discriminative power. Sünderhauf and Protzel (2011) proposed the use of BRIEF-Gist descriptors to perform place recognition, which could be computed very quickly. In their realization, the image was resized and divided into several patches of fixed size (e.g., 60×60). Then compact binary codes were generated for each patch through simply comparing the intensity of randomly selected point pairs. However, the resultant BRIEF-Gist descriptor did not retain sufficient information of the image, which weakened the power to distinguish each landmark. Moreover, limited by the matching strategy, in all these systems, only images with a similar orientation can be matched, seriously restricting their application.

Our solution is to introduce a panoramic camera into the traditional stereo-only VO system to build online panoramic landmarks. Compared to binocular images, panoramic images have a 360° field of view (FOV), making it possible to match landmarks in any orientation. Unlike in previous approaches, our landmark library considers mainly images of road junctions, which are the places most likely to be revisited. A compact library will reduce unnecessary processing of matching. Unlike the SLAM algorithm whose efficiency decreases gradually from constantly maintaining a global map, our method selects and records only relatively sparse landmark images.

To match the current image with landmarks precisely and accelerate the panoramic image processing, we introduce the concept of compressed sensing (CS) (Donoho, 2006) to the visual localization field. Taking advantage of the signal's sparseness or compressibility in some domain, CS allows the entire signal to be determined from relatively few measurements, while the original information is kept with high probability. By modeling image patches with the compressive features (CF), the computational cost can be reduced greatly. In consideration of the discrimination of different features, we present an adaptive CF to describe the image more efficiently.

While playing the key role in global position correction, panoramic images have another important function in local motion estimation. For binocular VO, there are some situations in which the image contains insufficient consistent features due to weak texture or moving objects, leading to erroneous motion estimation. Limited by the FOV, little can be done for further processing of the binocular images. In contrast, with a 360° FOV, panoramic images have plentiful visual information and are much less affected by this problem. Through tracking the motion of a pair of small rectangular patches in the panoramic image, the change in the robot's azimuth angle can be easily obtained, with relatively low precision but high robustness.

In summary, our contribution lies in four aspects:

1. A new enhanced VO system is proposed, combining a panoramic camera with a binocular camera. The binocular visual system is responsible for major local motion estimation, while the panoramic visual system is in charge of global position correction, as well as providing robust local azimuth angle for reference.

2. A quasi-real time global position correction algorithm using the online-built panoramic landmarks is proposed. The algorithm helps keep localization

error below a certain level in a long range with an acceptable computing complexity. Compared with SLAM, our method does not need to maintain and update a global map constantly, and allows quasi-real time localization.

3. An adaptive compressive feature (ACF) is proposed, to describe the panoramic landmark images. Using this feature, the panoramic images are described more adaptively and processed more efficiently.

4. A panoramic compass for robust local azimuth angle estimation is presented.

2 Overview of the enhanced visual odometry

Gathering the data from panoramic and binocular cameras, our enhanced VO system can be divided into two parts: global position correction and local motion estimation (Fig. 1). There are four modules in total: three are based on panoramic images and the ACF, and one is based on binocular images. Brief descriptions of the four modules are given in the following paragraphs:

1. Building landmarks online: Detecting road junctions from panoramic images as the robot moves, and extracting multi-level features to build landmarks. The multi-level features include intensity histograms, and large and small scale compressive features. In particular, for efficiency and accuracy, the compressive features are constructed adaptively according to the specific content of landmarks.

2. Global position correction: The current panoramic image will be matched with the landmark library through multi-level feature matching. Using a grid-based searching method, arbitrary directional landmarks can be matched when they are revisited. The robot's position can therefore be corrected, and the accumulated error greatly reduced.

3. Panoramic compass: By tracking patches in panoramic images, the changes in the azimuth angle can be estimated. This value is used to improve orientation estimation when stereo VO is not able to produce reliable estimations due to illumination variation, weak texture, or moving objects.

4. High-performance binocular VO (Lu *et al.*, 2013): With the optimization of two-stage local binocular bundle adjustment (TLBBA) and the acceleration of the GPU, this module can estimate the inter-frame motion precisely at about 25 Hz.

3 Adaptive features based on compressed sensing

3.1 Compressive intensity feature

CS is a signal processing technique for efficiently acquiring and reconstructing signals. Originally, CS was used to recover high-dimensional original signal from low-dimensional observation signal (Donoho, 2006; Cai *et al.*, 2014). Recently, it has been introduced to the field of pattern recognition, in areas such as face recognition (Wright *et al.*, 2009) and object tracking (Zhang *et al.*, 2012).

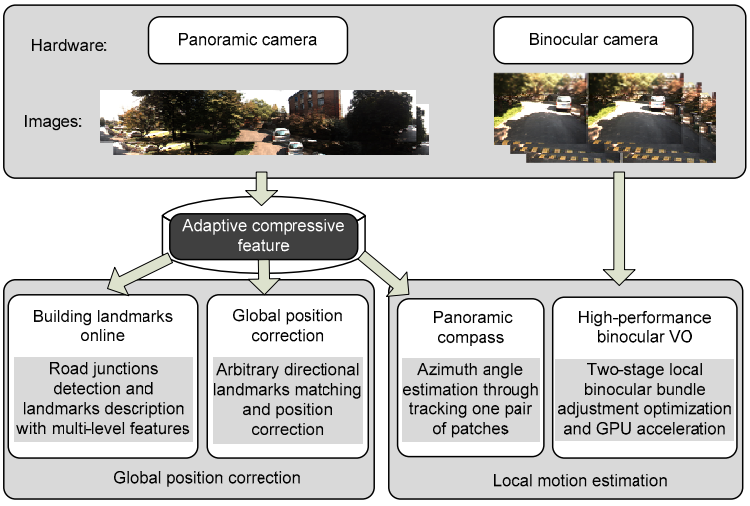


Fig. 1 Framework of the enhanced visual odometry

CS theory holds that a K-sparse high-dimensional signal $\boldsymbol{x}_{m\times1}$ can be compressed by a measurement matrix $\boldsymbol{R}$, which satisfies the restricted isometry property (RIP) (Candes and Tao, 2005). The resulting low-dimensional signal $\boldsymbol{v}_{n\times1}$ preserves the distance between all pairs of original signals with high probability. The signal is formulated as follows:

$$\boldsymbol{v}_{n\times1}=\boldsymbol{R}_{n\times m}\cdot\boldsymbol{x}_{m\times1}, \tag{1}$$

where $K<n<<m$.

Zhang *et al.* (2012) presented a CF for object tracking and achieved very good performance. To extract a CF, a tracked image patch $\boldsymbol{Z}_{h\times w}$ is convolved with a set of rectangle filters at multiple scales $\{f_{i,j}\}$, defined as

$$f_{i,j}(x,y)=\begin{cases}1, & 1\le x\le i, 1\le y\le j,\\ 0, & \text{otherwise.}\end{cases} \tag{2}$$

The resultant rectangular intensity features are concatenated as a very high-dimensional multi-scale feature $\boldsymbol{x}_{m\times1}$ to describe $\boldsymbol{Z}$, where $m=(h\times w)^2$. Subsequently, $\boldsymbol{x}_{m\times1}$ is compressed by a very sparse random measurement matrix $\boldsymbol{R}$, whose elements are

$$r_{ij}=\begin{cases}\sqrt{s}, & p=1/(2s),\\ 0, & p=1-1/s,\\ -\sqrt{s}, & p=1/(2s),\end{cases} \tag{3}$$

where p is the probability of assigning r_{ij} with the specific value, and $s=m/4$ according to Zhang *et al.* (2012).

3.2 Adaptive compressive feature

In image processing, intensity is one of the most basic features. It is easy to compute, but sensitive to variation in illumination and weak in describing image details. In actual applications, cloudy weather may cause sudden and frequent illumination variation, leading to unstable intensity features. On the other hand, the speeded up robust features (SURF) (Bay *et al.*, 2006) descriptor is gradient-based and more robust to illumination variation, and can be combined with intensity features to describe the image. SURF is complementary to intensity, and the image details can be better modeled. However, unlike intensity, one patch's SURF feature includes four component statistics: $\sum \mathrm{d}x$, $\sum|\mathrm{d}x|$, $\sum \mathrm{d}y$, and $\sum|\mathrm{d}y|$, where $\mathrm{d}x$ and $\mathrm{d}y$ are the horizontal and vertical gradients, respectively. To accommodate this requirement, we design a modified random measurement matrix $\boldsymbol{R}$, whose elements are four-dimensional vectors:

$$\dot{\boldsymbol{r}}_{ij}=\begin{cases}\sqrt{s}\bar{\boldsymbol{r}}_{1\times4}, & p=1/(2s),\\ \boldsymbol{0}_{1\times4}, & p=1-1/s,\\ -\sqrt{s}\bar{\boldsymbol{r}}_{1\times4}, & p=1/(2s),\end{cases} \tag{4}$$

where $\boldsymbol{r}$ is a small random measurement matrix, compressing the four elements of one patch's SURF into one:

$$\bar{r}_k=\begin{cases}1, & p=1/4,\\ 0, & p=3/4,\end{cases}\quad k=0,1,2,3. \tag{5}$$

In practice, some patches are not as good as others for matching purposes. For example, a patch from a solid color image is difficult to track, due to multiple matches with similar appearance. In other words, the rectangular sample patches have different discrimination abilities, although their locations are randomly generated. In general, the sample patches located in weak texture or repetitive texture regions have less discrimination. Moreover, some patches are more distinctive in terms of intensity, while others are better featured using SURF descriptors. We propose an adaptive description method to solve this problem, through analyzing the feature's discrimination.

Given an image $\boldsymbol{I}$, several rectangular patches are randomly sampled through the measurement matrix $\boldsymbol{R}$. Fig. 2a illustrates only some patches of the same size for clarity. Denote one patch as $\boldsymbol{z}_0$, and select two sets of round patches $\boldsymbol{z}_0$ of the same size. One set is called positive samples, whose location $l(\boldsymbol{z})$ satisfies $\{\boldsymbol{z}|\,\|l(\boldsymbol{z})-l(\boldsymbol{z}_0)\|<\alpha\}$, and the other set is negative samples whose location satisfies $\{\boldsymbol{z}|\beta<\|l(\boldsymbol{z})-l(\boldsymbol{z}_0)\|<\gamma\}$, where $\alpha<\beta<\gamma$. The intensity and SURF features extracted from both sets are fitted into four Gaussian functions separately. If the variances of the four Gaussian functions are all too small, $\boldsymbol{z}_0$ will be removed because it may be located on a background with similar texture (Fig. 2b).

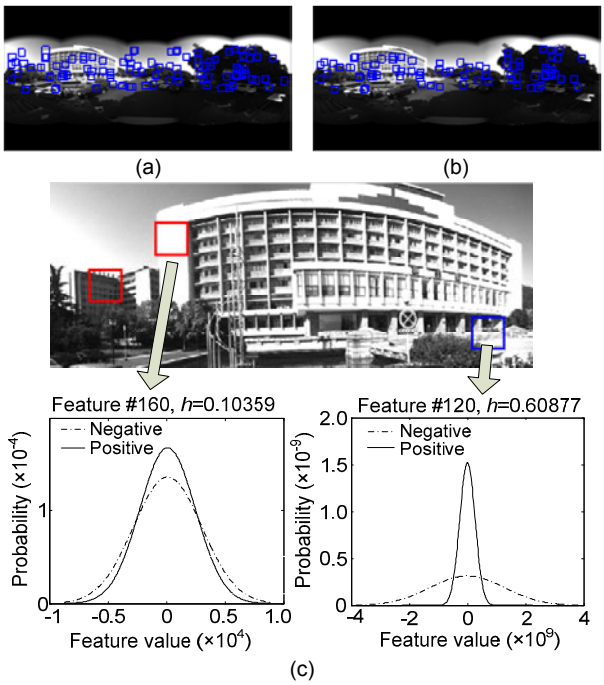


Fig. 2 Illustration of the adaptive compressive feature
(a) Randomly located rectangles in image $\boldsymbol{I}$; (b) Remove the features whose variances are too small; (c) Weight the features with the Hellinger distance

Taking the intensity feature as an example, its discrimination is measured by the Hellinger distance of the two Gaussian functions. Denoting the positive sample as y=1 and the negative sample as y=0, the conditional probabilities of compressive feature v_i are Gaussian:

$$\begin{cases} p^1 = p(v_i \mid y=1) \sim N(\mu_{1i}, \sigma_{1i}), \\ p^0 = p(v_i \mid y=0) \sim N(\mu_{0i}, \sigma_{0i}). \end{cases} \tag{6}$$

The square Hellinger distance of p^1 and p^0 is

$$h_i^2 = 1 - \sqrt{\frac{2\sigma_{1i}\sigma_{0i}}{\sigma_{1i}^2 + \sigma_{0i}^2}} \exp\left(-\frac{1}{4}\frac{(\mu_{1i}-\mu_{0i})^2}{\sigma_{1i}^2 + \sigma_{0i}^2}\right). \tag{7}$$

As shown in Fig. 2c, some patches are more discriminating when described by their intensities than with SURF (i.e., the feature on the right), and others are the reverse (such as the feature on the left). All the intensity and SURF features are sorted together according to h in descending order, and only the first n' weighted features are selected. Image $\boldsymbol{I}$ can now be represented by the n'-dimensional ACF, with the following form:

$$\boldsymbol{I} \Leftrightarrow \{h_1 v_1, h_2 v_2, ..., h_{n'} v_{n'}\}. \tag{8}$$

4 Building compressive panoramic landmarks online

4.1 Determination of landmark images

The selection of landmarks directly affects the following matching and position correction performance. If the landmarks are selected too densely, the current image has to match many similar landmarks, which is time-consuming and error-prone. On the other hand, if the landmarks are too sparse, some visited places may be missed and the valuable experience cannot be used for position correction. The ideal situation is to gain the most travel experience from visiting as few landmarks as possible. To this end, we select panoramic images with scenes of road junctions as landmarks, because they are the places most likely to be revisited. We build the landmarks with panoramic images, whose 360° FOV provides the possibility of arbitrary directional landmark matching. This means that only one landmark image is enough for a road junction. An efficient landmark matching algorithm will be given in the next section.

Detection of road junctions is a difficult problem for traditional stereo-only vision systems due to their limited FOV providing insufficient road information. Fortunately, with panoramic images this is less of a problem. The algorithm of Wang (2013) is adopted to detect the road junctions. First, the panoramic image is projected to the ground according to calibration parameters. Then the road area in the image can be obtained using Gaussian mixture model segmentation, followed by morphology based passable region identification. Finally, if the number of local maximum values in the radon field of the road binary image is greater than one, the current visiting place is considered as a road junction.

Building global landmarks only on road junctions may not be enough in all cases. For example, the robot may walk on a circular track or move back and forth between two adjacent road junctions, where no landmarks can be built. Therefore, a supplementary rule is presented: if the robot has moved more than a certain distance from the last landmark, a panoramic image of current place will be selected as a landmark. The positions of landmarks are decided

by the VO positioning results when they are visited for the first time.

4.2 Landmark description

In our application, the descriptor should meet two requirements: (1) It should robustly recognize and lock the landmarks that the vehicle is approaching; (2) From incoming images, it should find the best matching frame to the locked landmark.

In general, the global feature is not sensitive to local variation, such as moving objects and small-range appearance changes. This characteristic, in turn, makes it difficult to analyze the motion precisely using global features. Analogously, a large-scale feature is more robust to small-scale changes, while a small-scale feature can better reflect slight information variation. To trade-off the recognition requirements and computing efficiency, we construct multi-level features, finding the best match from the library step by step.

An intensity histogram is used as the first level feature, removing the obviously impossible landmarks. Since the histogram is a global feature, we use the proposed ACF to describe a more detailed level of the image. In consideration of both efficiency and accuracy, two scales of ACFs are included in the multi-level features. The features are extracted within the middle region of the panoramic images, because little useful information can be obtained from the cloudy sky on the top and the car body on the bottom. The width of the image is much larger than the height, and thus the two scale filters differ mainly in the widths of the rectangles. Denoting the large-scale rectangle filters as $\{f_{h_l,w_l}\}$ and the small-scale rectangle filters as $\{f_{h_s,w_s}\}$, according to our experience, the thresholds are set as $h/10\le h_s<h/8\le h_l<h$ and $w/40\le w_s<w/20\le w_l<w$. The distribution of the four most discriminative ACFs is illustrated in Fig. 3.

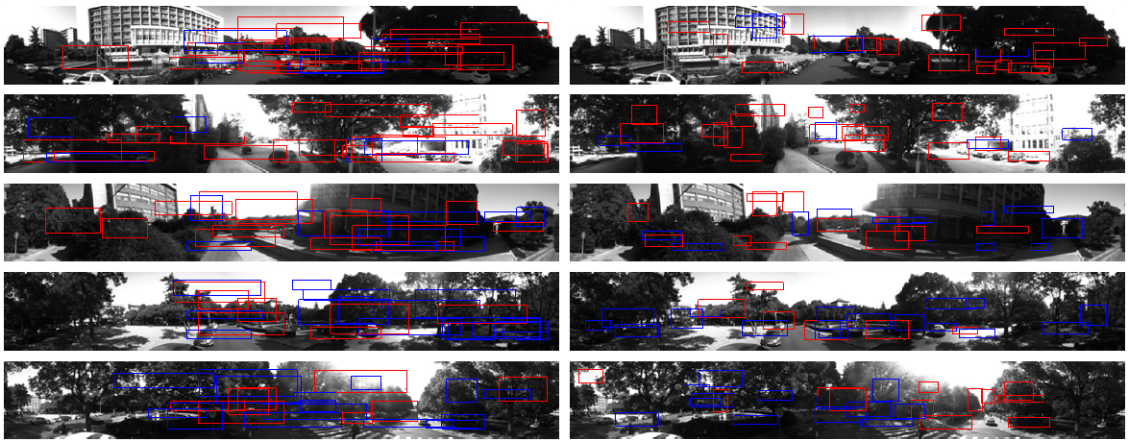
When the best match to the current image is found within the landmark library, a small position difference may still exist between them. SIFT features are then induced for estimating the residual position error. Note that here the complete SIFT feature is adopted for its excellent performance in accurate feature localization and description, while in ACF the concept of a SURF descriptor (i.e., without a detector) is used for efficiently expressing the intensity gradient within the image block.

Finally, the landmarks contain the positions $\boldsymbol{P}$, intensity histograms $\boldsymbol{H}$, large-scale ACF $\boldsymbol{F}_l$, small-scale ACF $\boldsymbol{F}_s$, and SIFT features $\boldsymbol{S}$ for final position correction.

5 Global position correction using landmarks

5.1 Matching landmarks with arbitrary orientation using multi-level features

Given the landmark library $L\{\boldsymbol{P}, \boldsymbol{H}, \boldsymbol{F}_l, \boldsymbol{F}_s, \boldsymbol{S}\}$, the following task is to find the landmark closest to the current image $\boldsymbol{I}_n$. First, the vehicle's approximate



(a) (b)

Fig. 3 Distribution of the four most discriminative adaptive compressive features

(a) Large-scale features; (b) Small-scale features. Red rectangles: intensity feature; blue rectangles: SURF feature. References to color refer to the online version of this figure

position $\boldsymbol{P}_{\mathrm{w}}$ is obtained from the binocular VO, and a landmark subset $L_{\mathrm{N}}\{\boldsymbol{H}_{\mathrm{N}}\}$ near to $\boldsymbol{P}_{\mathrm{w}}$ is selected from L (Fig. 4). This allows the visual landmark matching to be limited to a relatively small range, even if library L is very big.

The intensity histogram $\boldsymbol{h}$ of the current image $\boldsymbol{I}_n$ is computed and compared with $\boldsymbol{H}_{\mathrm{N}}$, and those landmarks whose distances to $\boldsymbol{h}$ are less than the threshold T_{h} are reserved for further matching. Since the histogram is a kind of global features and invariant

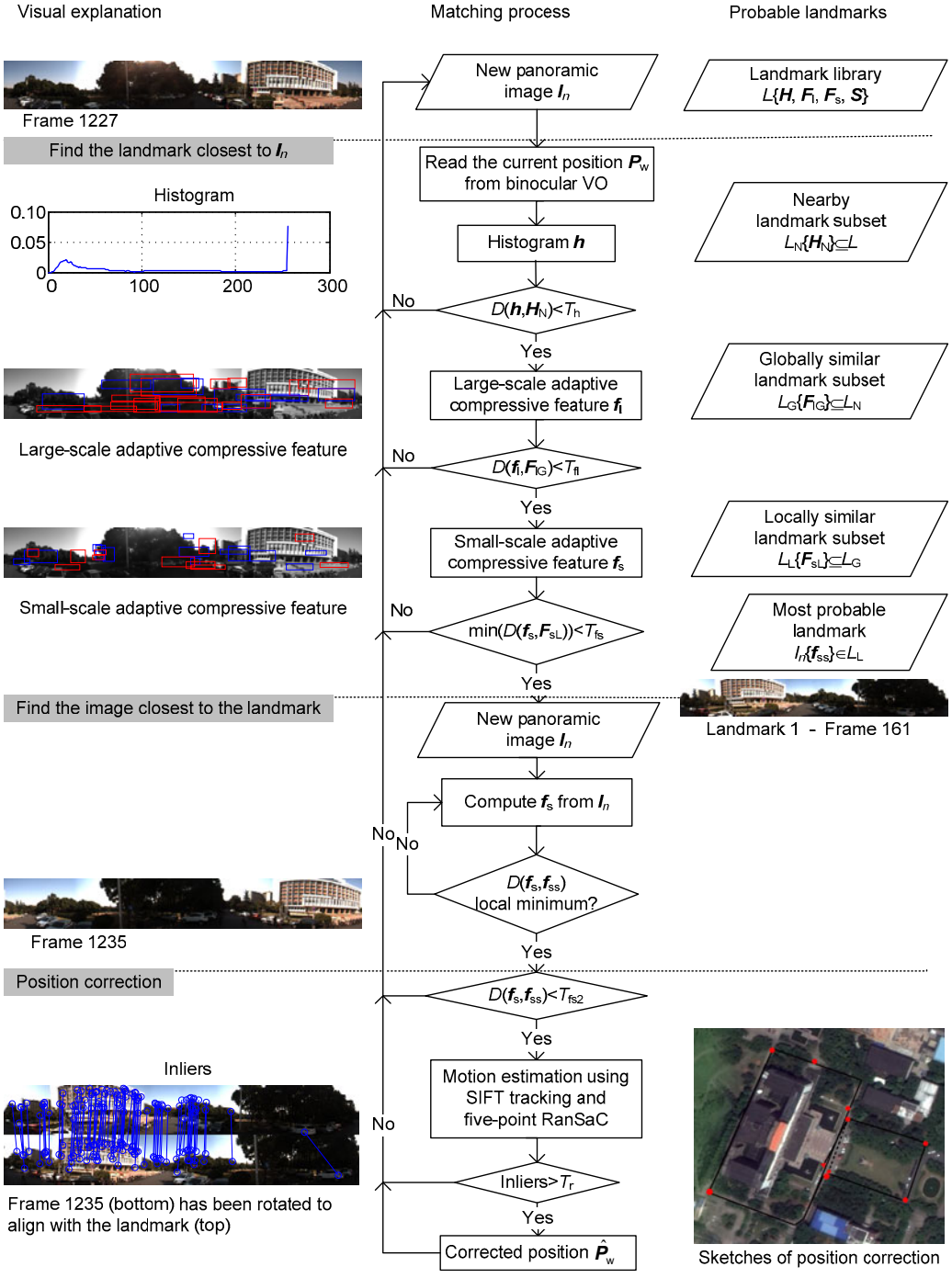


Fig. 4 Global position correction using panoramic landmarks

to directions, this step of matching can remove obvious impossible landmarks robustly.

Next, large-scale ACFs $\boldsymbol{f}_\mathrm{l}$ are used to match with landmark subset $\boldsymbol{L}_\mathrm{G}\{\boldsymbol{F}_\mathrm{lG}\}$, which is globally similar to $\boldsymbol{I}_n$. Since the randomly generated locations of the patches are related to the view angle, the ACFs are not invariant to directions. In fact, features computed based on the appearance of a partial image patch are all dependent on directions. We solve this problem by extracting multiple features to approximate features of all directions: rotating the 360° panoramic image, extract features at every 30° and match the obtained features with $\boldsymbol{L}_\mathrm{G}$. Although this matching method needs to extract features in $\boldsymbol{I}_n$ several times, the optimal matching can be found quickly, thanks to the efficient ACF.

If the minimum of $D(\boldsymbol{f}_\mathrm{l}, \boldsymbol{F}_\mathrm{lG})$ is smaller than T_fl, the landmark subset $\boldsymbol{L}_\mathrm{L}\{\boldsymbol{F}_\mathrm{sL}\}$, which is locally similar to $\boldsymbol{I}_n$, can be selected from $\boldsymbol{L}_\mathrm{G}$. Since the small-scale features $\boldsymbol{f}_\mathrm{s}$ are more sensitive to motion, they are used for matching with $\boldsymbol{L}_\mathrm{L}$ in more detail. The matching of $\boldsymbol{f}_\mathrm{s}$ can be controlled within the 30° region on both sides of the matching orientation of $\boldsymbol{f}_\mathrm{l}$. A grid-based searching method (Algorithm 1) is used to gradually find the optimal matching orientation, where 'grid_size' means the number of the grids in the searching region.

If the minimum of $D(\boldsymbol{f}_\mathrm{s}, \boldsymbol{F}_\mathrm{sL})$ is smaller than the threshold T_fs, the corresponding landmark $l_n\{\boldsymbol{f}_\mathrm{ss}\}$ is considered as the most probable landmark the vehicle is approaching, and is locked. Then each of the incoming images is matched to the locked landmark to decide the exact time the vehicle passed the landmark position. The image closest to l_n is found by searching the local minimum of $D(\boldsymbol{f}_\mathrm{s}, \boldsymbol{f}_\mathrm{ss})$ in a time window. If the final minimum is less than threshold T_fs2, the corresponding panoramic image $\boldsymbol{I}_n$ is matched with l_n as the result.

Algorithm 1 Grid-based searching method

f_s_orientation=f_l_orientation;
$\boldsymbol{f}_\mathrm{s}$**_opt**=small-scale ACF at f_s_orientation;
grid_size=2;
step=30°/(grid_size+1);
while step>0.5°
  start_orientation=f_s_orientation;
  for i=−grid_size:1:grid_size
    temp_orientation=start_orientation+i*step;
    temp_$\boldsymbol{f}_\mathrm{s}$=small-scale ACF at temp_orientation;
    if temp_$\boldsymbol{f}_\mathrm{s}$ is closer to $\boldsymbol{F}_\mathrm{sL}$ than $\boldsymbol{f}_\mathrm{s}$**_opt then**
      f_s_orientation=temp_orientation;
      $\boldsymbol{f}_\mathrm{s}$**_opt=temp_**$\boldsymbol{f}_\mathrm{s}$;
    end if
  end for
  step=step/(grid_size+1);
end while
return f_s_orientation

5.2 Position correction

We match SIFT features of $\boldsymbol{I}_n$ and l_n, and the inter-frame motion is estimated with the five-point algorithm (Stewénius *et al.*, 2006). The RanSaC method (Fischler and Bolles, 1981) is adopted to remove the outliers. Since the images are roughly aligned after landmark matching, the rotation angle should be constrained in a relatively small range, and most false estimation results can be filtered out by this constraint. In some cases, there are only a few matched features and final inliers due to the large-scale translation or occasional false landmark matching. In such cases, the robot's position will not be corrected and refined. From a result-oriented perspective, position correction could also be viewed as the final verification of landmark matching.

The correction result does not contain scale information and only the orientation correction can be directly used. We then check the curve of $D(\boldsymbol{f}_\mathrm{s}, \boldsymbol{f}_\mathrm{ss})$: if the curve is steep, which means the displacement between $\boldsymbol{I}_n$ and l_n is relatively small, then the matching is regarded as perfect and the position of the landmark is directly used to update the current VO estimation; if the curve of $D(\boldsymbol{f}_\mathrm{s}, \boldsymbol{f}_\mathrm{ss})$ is flat, the translation error may still be large and the robot's position will not be corrected except for its rotation (Fig. 7).

6 Panoramic compass

The inter-frame motion is estimated mainly by our previous binocular VO (Lu *et al.*, 2013), which runs in parallel with two threads: feature tracking and motion estimation. The first thread detects and describes features using SIFT_GPU (Wu, 2007), and then matches the features and reconstructs 3D points. The second thread estimates the motion using the RanSaC and quaternion methods (Horn, 1987), and the result is optimized with the TLBBA algorithm.

In general, the binocular VO can achieve good motion estimation results. However, in some cases,

the feature tracking may be corrupted by abnormal scene contents (e.g., moving objects covering a large part of the FOV or weak texture on the ground). Since the binocular camera looks down and forward to track the close features for accuracy, this problem becomes more serious, especially at road junctions, where most of the FOV of binocular images is occupied by weak-texture asphalt and few features can be extracted and tracked, making the binocular VO suffer from larger errors. Another difficult situation is the appearance of close moving objects in front of the camera, especially with a weak-texture background. The panoramic camera can capture more than 80% of environmental information of the 360° sphere (except for the very bottom). So, the panoramic images record nearly all the useful scenes, and significantly reduce the probability of a weak texture. Thanks to the large FOV, the influence of the same moving object is much less serious in a panoramic image than in a binocular one. With the help of the panoramic image, a special panoramic compass algorithm is designed to improve motion estimation.

In most cases, the road is locally flat, and the robot's motion is approximately planar. So, the rotation is determined mainly by the change of the azimuth angle. The translation can be roughly estimated with the assumption that the displacement is proportional to the offsets of the left and right patches. By tracking the front and the rear patches in panoramic images, we can easily compute the azimuth angle. The elevation angles of the patches are about 30° above the horizon in the image, and the influence of moving objects can be removed.

The spherical coordinate system is used to construct the panoramic image, and the image is expanded according to the longitude and latitude. For a 3D point $\boldsymbol{P}(X, Y, Z)$, the direction angles (θ, φ) are

$$\begin{cases}\theta = \operatorname{atan2}(X, Z), \\ \varphi = \arcsin(Y / r),\end{cases} \tag{9}$$

where $r=(X^2+Z^2)^{1/2}$ is the radius of the cylinder. Given the image resolution of nCols×nRows, the relationship between (θ, φ) and the pixel coordinates of the 3D point's image projection $P'(u, v)$ is

$$\begin{cases}\theta = \dfrac{2\pi u}{\text{nCols}} - \pi, \\ \varphi = \dfrac{\pi v}{\text{nRows}} - \dfrac{\pi}{2}.\end{cases} \tag{10}$$

To track the patch, n'-dimensional ACFs are extracted within the patch, and the distance D between the patch in the current image $\boldsymbol{I}_n$ and the matched patch in the next image $\boldsymbol{I}_{n+1}$ can be measured as

$$D = \sum_{i=1}^{n'} w_i d(v_{i,n}(u, v), v_{i,n+1}(u + \Delta u, v + \Delta v)). \tag{11}$$

Both the front and rear patches are tracked, but only the patch having the smaller distance D is used to estimate the azimuth angle $\Delta\theta$. For example, if the tracking results of the front patch are better, then $\Delta\theta$ can be calculated using the offset of the front patch $(\Delta u^{\mathrm{f}}, \Delta v^{\mathrm{f}})$:

$$\Delta\theta = \frac{2\pi\Delta u^{\mathrm{f}}}{\text{nCols}}. \tag{12}$$

7 Experments

We used an electric vehicle (Fig. 5a) as the mobile platform to test the proposed enhanced VO, capturing binocular images with BumbleBee-2 (Fig. 5b) and panoramic images with Ladybug-3

(a)

(b)

(c)

Fig. 5 Experimental equipment
(a) Electric vehicle; (b) BumbleBee-2 binocular camera; (c) Ladybug-3 omni-directional camera

(Fig. 5c). Ladybug-3 is an omni-vision system manufactured by Point Grey (Canada). The system is composed of six Sony color cameras, each having a maximum resolution of 1600×1200. It can output six independent images, and one stitched panoramic image. We directly used the panoramic images with a resolution of 1080×540, and the binocular image resolution was 640×480. The experiments were conducted on a computer equipped with an Intel i5 2.8 GHz CPU, 4 GB RAM, and NVIDIA GTS450 graphics card.

7.1 Evaluation of ACFs in landmark matching

For the purpose of landmark matching, the ideal feature should not only obtain minimum distance with the true positive (TP), but also discriminate itself with the true negative (TN) with a large gap. We matched each landmark with other landmarks in the library. Since they are different places, the results of minimum distance matching are considered as negative samples. The panoramic images of revisited landmarks were also matched with the library. The correct matching results are positive samples and the false results are false negatives. To better quantify the power of separating true positive and negative samples, we define

$$V_{sc}=d_N-d_P, \tag{13}$$

where d_N and d_P are the average distance of negative and positive samples, respectively. The BRIEF-Gist feature (Sünderhauf and Protzel, 2011) was selected as the state-of-the-art method to compare with our proposed ACFs in this application of landmark matching. To show the comparison clearly, the sample distances D of each kind of feature were normalized with feature scaling:

$$D'=\frac{D-D_{min}}{D_{max}-D_{min}}. \tag{14}$$

The library contains 36 landmarks, and seven revisits should have been detected. Although the BRIEF-Gist successfully matched six of the seven revisited landmarks, it did not discriminate well between the positive and negative samples. If we wanted to reserve all six TPs, at least eight negative samples would be included as false positives (FP). In contrast, our large-scale ACF performed better (Fig. 6). All seven revisited places were correctly detected and the positive and negative samples were separated well (Table 1).

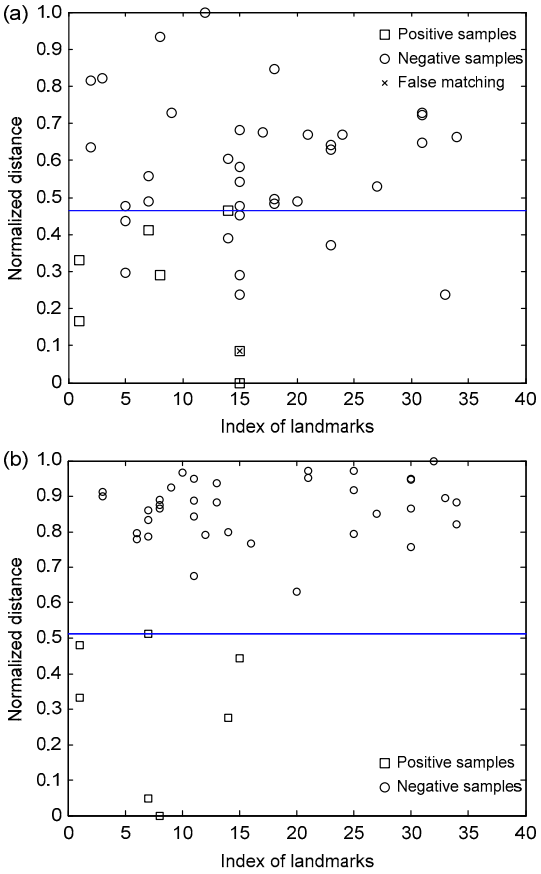


Fig. 6 Global property of the features
(a) The matching result of BRIEF-Gist; (b) Our large-scale ACF

Table 1 Comparison of the features' global properties

Method	V_{sc}	TP	FP	Recall	Precision
BRIEF-Gist	0.33	6	8	0.86	0.43
Large-scale CF	0.54	7	0	1.00	1.00
Large-scale ACF	0.56	7	0	1.00	1.00
Small-scale CF	0.47	7	0	1.00	1.00
Small-scale ACF	0.37	7	14	1.00	0.33

According to the results, the large-scale ACF had the best global matching property. Next, we analyzed the local matching property of these features. The features of locked landmarks were compared with the successive incoming images. To minimize

the position correction error, the most similar frame to the landmark had to be found, which corresponds to the feature's local distance curve with the most obvious local minimum. The small-scale ACF was the best choice, because it is the most sensitive to small motion changes (Fig. 7). Since the randomly generated CFs contained some fewer discriminative features, their curves were flatter than those of the ACFs.

Moreover, using the ACFs reduced the computational cost greatly. In landmark matching, for either large- or small-scale features, 200 randomly located CFs were generated, but only 100 of them were selected and weighted to construct ACFs. In the panoramic compass, the corresponding numbers of features were cut in half.

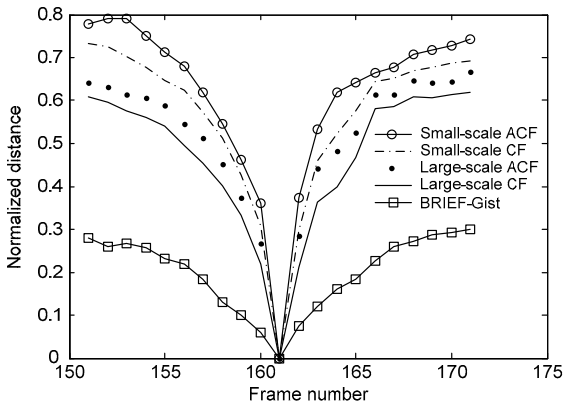


Fig. 7 Local feature properties of landmark 1

7.2 Performance of the panoramic compass

Due to a sudden illumination change (Fig. 8a), moving objects (Fig. 8d) or large rotation at weakly textured corners (Figs. 8b and 8c), binocular feature tracking becomes erroneous. In contrast, panoramic images contain much more consistent information. In these cases, the panoramic compass is activated to produce better azimuth angle estimation. We show one result of the panoramic compass in Fig. 9, which corresponds to the case of Fig. 8d, where the robot saw some moving objects at a corner. In Fig. 9b, the estimation results from stereo-only VO are with an obviously erroneous azimuth angle, and the corrected estimation results use the panoramic compass, which eliminates the effect of moving objects and provides much more precise azimuth angle estimation.

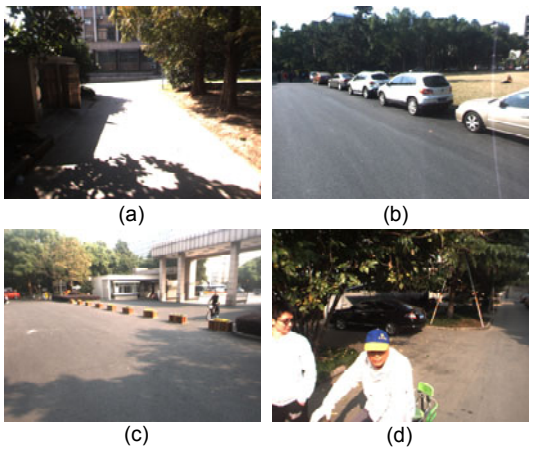


Fig. 8 Binocular image samples where the feature tracking failed

(a) Dataset-I, frame 1634; (b) Dataset-I, frame 5225; (c) Dataset-II, frame 5631; (d) Dataset-II, frame 8931

Fig. 9 Panoramic compass results

(a) Patch tracking results; (b) Azimuth angle correction. Red line: estimation results from stereo-only VO; blue line: corrected estimation results using the panoramic compass. References to color refer to the online version of this figure

7.3 Performance of the enhanced VO

The proposed enhanced VO system was tested in our campus environment, where there are weak-texture asphalt pavements, tall trees, buildings, grass, and frequent moving objects, such as pedestrians and vehicles. Dataset-I includes 4700 panoramic images and 17600 binocular images. Assuming the starting point of the vehicle as the origin of the world coordinate system, the vehicle's position was continuously estimated based purely on our proposed vision methods. The resulting estimated paths from binocular VO and enhanced VO are illustrated in Fig. 10.

In this experiment, 36 panoramic images were successfully selected as landmarks, among which six revisits were used to correct the VO positioning results (Table 2). The fourth revisit was successfully detected while the location was not corrected, due to insufficient inliers in RanSaC to produce a stable motion estimate. The detection of the eighth revisit was ignored, because the position was just corrected at the seventh revisit. It is not necessary to correct the position so frequently.

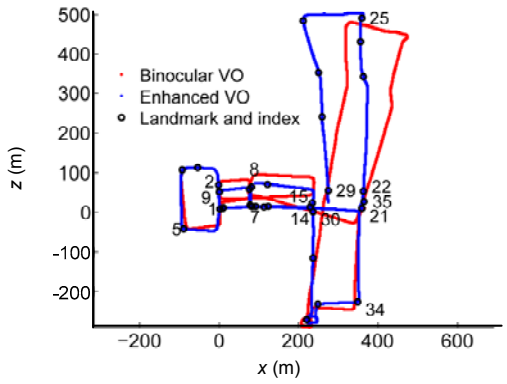


Fig. 10 Localization results of dataset-I

Red line: binocular VO; blue line: enhanced VO. References to color refer to the online version of this figure

In Table 2, we can see the entrance orientation varied greatly when landmarks were revisited. The smallest angle was 8° while the largest was 168°. All the landmarks with an arbitrary directional entrance were successfully matched. At each landmark, the GPS data were collected as ground truth for comparison purposes. From Table 2 we can see that the robot's location was effectively corrected after revisiting the landmarks, and the 2D root mean square accumulated error was reduced.

The capturing frequencies of the panoramic and binocular cameras were 5 Hz and 25 Hz, respectively. The processing time of the enhanced VO is shown in Table 3. The first thread built the landmarks online: the multi-level features were extracted from 1080×150 panoramic image patches, and the landmarks were detected on the down-sampling images. The second thread matched the current image with the landmark library and corrected the robot's position. The binocular VO used two threads for parallel processing. The panoramic compass ran in one of the two threads when it was activated. Given the locked approaching landmark, the best incoming frame most similar to the landmark could be found within 20 ms. Together with the following position correction stage, the whole global position correction process could be finished within 200 ms. Each module could run faster than the respective camera frequency, but to find the image locally closest to the landmark, the position correction has to be performed with at least one frame delay, and the whole system could run only in quasi-real time.

Dataset-II includes 4800 panoramic images and 17600 binocular images. This time the robot could

Table 2 Performance of the enhanced VO on dataset-I

Route number	Revisited landmark	Position correction	Route description	Entrance angle (°)	2D root mean square (m)		Traveling distance (m)
					Binocular VO	Enhanced VO	
1	1	Corrected	1-2-5-1	−8	13.5	1.5	440
2	1	Corrected	1-7-8-9-1	−138	23.9	2.1	592
3	7	Corrected	7-8-9-1-7	−50	22.5	3.5	662
4	8	Uncorrected	8-9-1-14-8	119			
5	7	Corrected	7-14-15-8-7	−107	38.3	4.4	1082
6	14	Corrected	14-15-8-7-14	−62	10.5	4.2	1223
7	15	Corrected	15-21-25-29-15	−168	29.7	5.3	1570
8	14	Given up	14-21-25-29-14				

benefit from the traveling experience in the first experiment, and no longer needed to travel circuitously to increase the probability of revisiting landmarks. The positioning results from the enhanced VO and the binocular VO are illustrated in Fig. 11. This time 14 landmarks were revisited, among which 13 were successfully used to correct the robot's position (Table 4). The tenth revisit was missed because the revisiting route was relatively far away from the landmark we detected. The starting point of this experiment was set to the position of the first landmark, where there was an initial localization error of 1.5 m compared with the ground truth from GPS.

In summary, among the 22 revisited landmarks (dataset-I and dataset-II), 20 revisits were successfully detected and matched, of which 19 were used to improve the positioning results. We conclude that the landmark library constructed with road junctions is very efficient, and high repeatability is one of the most important advantages of our algorithm. The global position correction algorithm and the panoramic compass can be very useful for visual localization in long distance journeys.

Table 3 Processing time of the enhanced VO

Item	Time (ms)
Building landmarks online	
Landmark detection	70
Landmark description	
Multi-level feature	80
SIFT	40
Global position correction	
Landmark matching	
$\boldsymbol{h}$	3
f_{l}	8
f_{s}	7
Position correction	
SIFT matching	50
Five-point RanSaC	100
Panoramic compass	100
Binocular VO	35

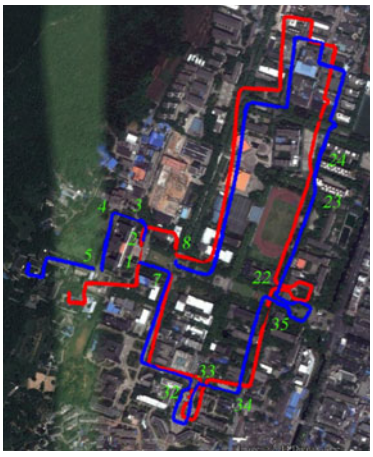


Fig. 11 Localization results of dataset-II
Red line: binocular VO; blue line: enhanced VO. References to color refer to the online version of this figure

Table 4 Performance of the enhanced VO on dataset-II

Route number	Revisited landmark	Position correction	Entrance angle (°)	2D root mean square (m)		Traveling distance (m)
				Binocular VO	Enhanced VO	
1	1	Corrected	−16	1.5	1.5	0
2	7	Corrected	−53	4.9	3.6	73
3	32	Corrected	−1	29.3	2.5	386
4	33	Corrected	7	30.4	3.8	494
5	34	Corrected	8	42.6	2.0	550
6	35	Corrected	−7	38.9	7.0	753
7	22	Corrected	5	41.4	5.3	862
8	23	Corrected	14	48.3	9.8	1091
9	24	Corrected	0	56.5	15.1	1180
10	8	Miss				
11	2	Corrected	12	83.7	4.2	1672
12	3	Corrected	2	80.8	4.5	1687
13	4	Corrected	32	85.4	3.1	1692
14	5	Corrected	−102	108.2	3.3	1724

8 Conclusions

We proposed an enhanced VO solution composed of a panoramic camera and a stereo camera. Through building online panoramic landmarks, landmark views from arbitrary orientations can be matched efficiently using the multi-level features, and the robot's position can be corrected globally when the landmarks are revisited. A panoramic compass can be activated to produce more robust azimuth angle estimation, when the binocular VO cannot give reliable results. To accelerate the landmark filtering and matching process, a multi-level adaptive compressive feature is presented to use the panoramic information more effectively. Experimental results show that our enhanced VO is able to relieve the drift problem that plagues the traditional VO, and obtain better localization results in challenging environments.

References

Bay, H., Tuytelaars, T., van Gool, L., 2006. SURF: speeded up robust features. Proc. 9th European Conf. on Computer Vision, p.404-417. [doi:10.1007/11744023_32]

Cai, X., Zhang, Z., Zhang, H., *et al.*, 2014. Soft consistency reconstruction: a robust 1-bit compressive sensing algorithm. arXiv:1402.5475 (preprint).

Candes, E.J., Tao, T., 2005. Decoding by linear programming. *IEEE Trans. Inform. Theory*, **51**(12):4203-4215. [doi:10.1109/TIT.2005.858979]

Donoho, D.L., 2006. Compressed sensing. *IEEE Trans. Inform. Theory*, **52**(4):1289-1306. [doi:10.1109/TIT.2006.871582]

Durrant-Whyte, H., Bailey, T., 2006. Simultaneous localization and mapping: part I. *IEEE Robot. Autom. Mag.*, **13**(2):99-110. [doi:10.1109/MRA.2006.1638022]

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, **24**(6):381-395. [doi:10.1145/358669.358692]

Fraundorfer, F., Scaramuzza, D., 2012. Visual odometry: part II. Matching, robustness, optimization, and applications. *IEEE Robot. Autom. Mag.*, **19**(2):78-90. [doi:10.1109/MRA.2012.2182810]

Galvez-López, D., Tardos, J.D., 2012. Bags of binary words for fast place recognition in image sequences. *IEEE Trans. Robot.*, **28**(5):1188-1197. [doi:10.1109/TRO.2012.2197158]

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.3354-3361. [doi:10.1109/CVPR.2012.6248074]

Horn, B.K.P., 1987. Closed-form solution of absolute orientation using unit quaternions. *JOSA A*, **4**(4):629-642. [doi:10.1364/JOSAA.4.000629]

Konolige, K., Agrawal, M., Solà, J., 2011. Large-scale visual odometry for rough terrain. Proc. 13th Int. Symp. on Robotics Research, p.201-212. [doi:10.1007/978-3-642-14743-2_18]

Liu, Y., Zhang, H., 2012. Visual loop closure detection with a compact image descriptor. Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p.1051-1056. [doi:10.1109/IROS.2012.6386145]

Lu, W., Xiang, Z., Liu, J., 2013. High-performance visual odometry with two-stage local binocular BA and GPU. Proc. IEEE Intelligent Vehicles Symp., p.1107-1112. [doi:10.1109/IVS.2013.6629614]

Munguia, R., Grau, A., 2007. Monocular SLAM for visual odometry. Proc. IEEE Int. Symp. on Intelligent Signal Processing, p.1-6. [doi:10.1109/WISP.2007.4447564]

Nistér, D., Naroditsky, O., Bergen, J., 2004. Visual odometry. Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, p.652-659. [doi:10.1109/CVPR.2004.1315094]

Scaramuzza, D., Fraundorfer, F., 2011. Visual odometry (tutorial). *IEEE Robot. Autom. Mag.*, **18**(4):80-92. [doi:10.1109/MRA.2011.943233]

Se, S., Lowe, D., Little, J., 2002. Global localization using distinctive visual features. Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p.226-231. [doi:10.1109/IRDS.2002.1041393]

Singh, G., Košechá, J., 2010. Visual loop closing using gist descriptors in Manhattan world. ICRA Omnidirectional Vision Workshop.

Sivic, J., Zisserman, A., 2003. Video Google: a text retrieval approach to object matching in videos. Proc. 9th IEEE Int. Conf. on Computer Vision, p.1470-1477. [doi:10.1109/ICCV.2003.1238663]

Stewénius, H., Engels, C., Nistér, D., 2006. Recent developments on direct relative orientation. *ISPRS J. Photogr. Remote Sens.*, **60**(4):284-294. [doi:10.1016/j.isprsjprs.2006.03.005]

Sünderhauf, N., Protzel, P., 2011. BRIEF-Gist—closing the loop by simple means. Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p.1234-1241. [doi:10.1109/IROS.2011.6094921]

Wang, Y., 2013. Navigational Road Modeling Based on Omnidirectional Multi-camera System. PhD Thesis, Zhejiang University, Hangzhou, China (in Chinese).

Wright, J., Yang, A.Y., Ganesh, A., *et al.*, 2009. Robust face recognition via sparse representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, **31**(2):210-227. [doi:10.1109/TPAMI.2008.79]

Wu, C., 2007. SiftGPU: a GPU Implementation of Scale Invariant Feature Transform (SIFT). Available from http://cs.unc.edu/~ccwu/siftgpu/

Zhang, K., Zhang, L., Yang, M.H., 2012. Real-time compressive tracking. Proc. 12th European Conf. on Computer Vision, p.864-877. [doi:10.1007/978-3-642-33712-3_62]