IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 10, 2017, accepted September 26, 2017, date of publication September 28, 2017,
date of current version October 25, 2017.

*Digital Object Identifier 10.1109/ACCESS.2017.2757765*

# Robust Topological Navigation via Convolutional Neural Network Feature and Sharpness Measure

**JIAYI MA[1,2] AND JI ZHAO[3]**

[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]Beijing Advanced Innovation Center for Intelligent Robots and Systems, Beijing Institute of Technology, Beijing 10081, China
[3]ReadSense Ltd., Shanghai 200040, China

Corresponding author: Ji Zhao (zhaoji84@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773295 and Grant 61503288, in part by the Joint Advanced Research Foundation of Departments of Equipment and Education under Grant 6141A02022303, in part by the Beijing Advanced Innovation Center for Intelligent Robots and Systems under Grant 2016IRS15, and in part by the China Postdoctoral Science Foundation under Grant 2016T90725.

**ABSTRACT** Visual navigation for mobile robots has emerged in recent years. Among the various methods, topological navigation using visual information provides a scalable map representation for large-scale mapping and navigation. A topological map is essentially a graph with keyframes as its nodes and adjacency relations as its edges. Previous topological mapping uses local feature descriptors, such as scale-invariant feature transform or Speeded-Up Robust Features, to select keyframes in mapping, localization, and estimate relative pose. In practice, local features are not robust for severe motion blur or large illumination change. In this paper, we improve topological mapping to make it more efficient and robust. First, we use a convolutional neural network (CNN) feature as the holistic image representation. The CNN feature can be used to effectively retrieve keyframes that have similar appearance from a topological map, and it is robust to motion blur and illumination change. Thus, it improves the performance for place recognition and robot relocalization. Second, we use sharpness measure to select high-quality keyframes and avoid selecting blurry ones. Third, an efficient and robust non-rigid matching method, vector field consensus, is used for efficient geometric verification and to retrieve the most similar keyframe. The qualitative and quantitative experimental results demonstrate that our method is satisfactory.

**INDEX TERMS** Topological navigation, visual navigation, convolutional neural network, sharpness measure.

## I. INTRODUCTION

Various methods for visual mapping and navigation have emerged in recent years. The map representative plays a key role in these tasks. It produces a representation of the environment built based on the visual information obtained through attached cameras, which is then used in subsequential tasks such as path-planning, localization, and obstacle avoidance [1]. There are a variety of kinds of map representations, such as metric map (feature map, occupied grid map), topological map, hybrid map, semantic map, *etc.* [2]. It has been mentioned in many previous works that, for the purpose of navigation a global consistent metric map is not a necessity. Rather, topological connections with rough metric information for navigation and planning is good enough [3].

Topological map sometimes is called image path, visual graph, appearance map, or appearance-based topological map. A topological map is a graph-based representation of the environment. Each node corresponds to a characteristic feature or zone of the environment. Each edge encodes the adjacency relations, or it is associated with an action, such as turning, crossing a door, stopping, or going straight ahead. Compared to metric maps, the topological maps are simple and compact, which take up less computer memory, and consequently are able to speed up computational navigation processes.

The topological mapping has been successfully applied for visual navigation. Topological navigation using omnidirectional camera is proposed in [4]. The invariant column segments is utilized as wide baseline features. In [5], an efficient image retrieval by bag-of-words feature representation and vocabulary tree is used in topological mapping, which enables mapping of very large environments. An incremental topological mapping method and loop closure detection is proposed in a similar way [6]. In [7], topological mapping by

multiple heterogenous mobile robots is proposed, which can build and share the image defined map for indoor navigation. In [8], autonomous quadrotor navigation is proposed through an image-defined path, where a visual map is first build by using collect images, and then quadrotor follows the desired visual path for navigation.

Among the above mentioned topological navigation methods, local features are used for selecting keyframes, re-localization, as well as estimating relative pose. The adopted features typically include invariant column segments [4], scale-invariant feature transform (SIFT) [5]–[7], [9], and fast retina keypoint (FREAK) [8], [10]. To further accelerate the calculation, sometime bag-of-words (BoW) [11]–[13] based on local features are used. In such feature representation, it first extracts local features from a large amount of images, and clusters these features to obtain a visual vocabulary. Then it quantizes local features of an image by visual vocabulary, and represents images by a histogram of visual words. Sometimes each bin of the histogram is further weighted by term frequency-inverse document frequency (TF-IDF), which is a numerical statistic that reflects how important a visual word is to an image in an image set.

Local features provide certain degree of invariance for illumination change and motion blur. Still it is fragile in challenging indoor navigation, because of the large illustration variance and motion blur. In addition, there are mismatches according to local features due to the large repetitive patterns or occlusions in the environment, and hence some robust estimator such as random sample consensus (RANSAC) [14] is used to achieve an accurate feature matching. Nevertheless, if the mismatch ratio is high, it takes RANSAC too much time to find a good estimation of relative pose. In this paper, we aim to remedy these problems, and make topological navigation more efficient and robust.

More specifically, we use the convolutional neural network (CNN) feature as the holistic image representation to effectively retrieval keyframes that have similar appearance from a topological map. The CNN feature is robust to motion blur and illumination change, and hence improves the performance for place recognition and robot relocalization. We subsequently use geometric verification to refine the retrieval results and seek the most similar keyframe based on an efficient and robust non-rigid matching method. In addition, the sharpness measure is adopted to select high-quality keyframes and avoid selecting blurry ones. Experiments on a publicly available dataset and on a real robot platform demonstrate the effectiveness of our method for both outdoor and indoor navigation.

### A. CONVOLUTIONAL NEURAL NETWORK SCHEME

Recently, one important breakthrough in machine learning is known as deep learning. By exploring deep architectures to learn features at multiple level of abstracts from data in a supervised way, deep learning methods allow a system to learn complex functions that directly map raw sensory input data to the output, without relying on human-crafted features using domain knowledge. Many recent studies have reported encouraging results for applying deep learning techniques to a variety of applications.

As the breakthrough of deep learning in computer vision [15], it significantly raises the interest in learning feature representations. Features generated by convolutional neural networks in a supervised way has been demonstrated powerful ability in various tasks including image classification and object detection. It also has been generalized to other tasks by fine-tuning of parameters. In robotics community, CNN features have been used for place recognition [16], [17], and also have been used to regress the motion direction for drone navigation to go through forest trails [18], and many other tasks from perception to control [19].

### B. ROBUST FEATURE MATCHING

Establishing reliable correspondences of feature points is a fundamental problem in computer vision [20]–[22], pattern recognition [23]–[25], image restoration [26]–[29], remote sensing [30], medical image analysis [31], and it also plays an important role in robot navigation during retrieving the most similar keyframe from a topological map. A popular strategy to establish reliable point correspondences involves the following two steps [32]: (i) computing a set of putative correspondences, and (ii) then removing the mismatches that utilize geometric verification. Putative correspondences are obtained in the first step by pruning the set of all possible point correspondence and removing the matches whose descriptors [9], [33] are excessively dissimilar.

To remove false matches from the putative set in the second step, different methods have been proposed in the past decades. One of the most widely used method is RANSAC [14], which adopts a hypothesize-and-verify approach and attempt to obtain the smallest possible outlier-free subset to estimate a provided parametric model by resampling. The geometric verification used in RANSAC relies on a predefined parametric model, which become less efficient when the underlying image transformation is non-rigid; it also tends to severely degrade if the mismatch proportion becomes large [34]. Several non-parametric interpolation methods [32], [35], [36] have recently been introduced to address these issues. These methods commonly interpolate a non-parametric function by applying the prior condition, in which the motion field associated with the feature correspondence is slow-and-smooth. In addition, a sparse approximation for solving the spatial transformation has been introduced to reduce the computational complexity of these methods [37].

## II. TOPOLOGICAL MAPPING AND LOCALIZATION

Our topological mapping and navigation is similar to that in [4], [5], and [8]. In the mapping stage, it finds keyframes by bag-of-words feature representation and then performs geometric verification by RANSAC scheme [14]. In the localization stage, it finds the most similar keyframes with current observed image from the topological map. Then it

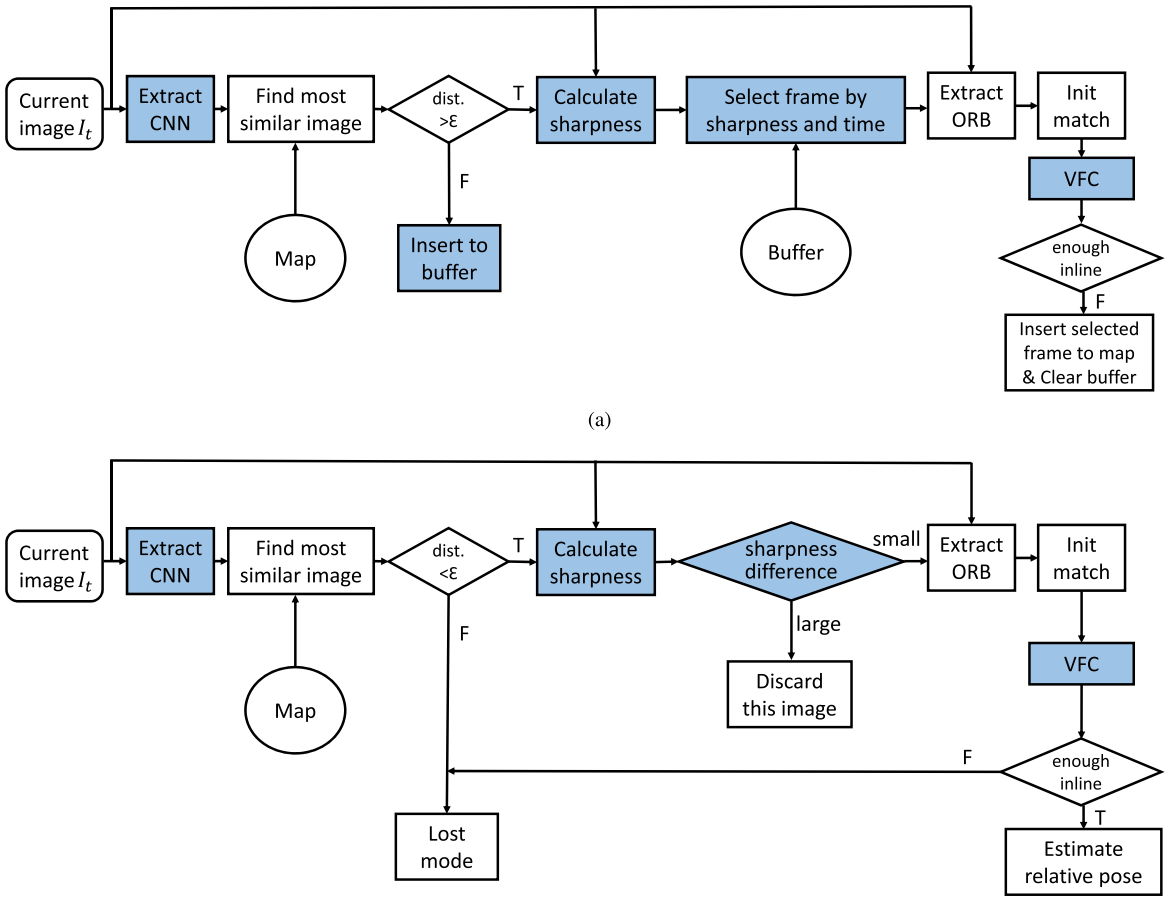



**FIGURE 1.** Flowcharts of topological mapping and localization. Blue blocks are added by this paper. In previous literature, it is RANSAC in VFC module. CNN: convolutional neural network [15]; ORB: oriented FAST and rotated BRIEF [38]; VFC: vector field consensus [32]. (a) Flowchart of topological mapping. (b) Flowchart of topological localization.

estimates the relative pose between current image and the most similar keyframe. Path planning is realized by traversing a sequence of way-point images. Navigation is achieved by iteratively (i) determining the desired motion to the corresponding keyframe; (ii) motion control; and (iii) appropriately switching to a new keyframe in the planned path. The flowcharts of improved mapping and localization are shown in Fig. 1, where the blue blocks are our newly added modules.

In the mapping stage, we aim to find some representative frames (keyframes) for the environment. The ideal keyframes should satisfy the following criteria: (i) can cover the whole scene; (ii) have less redundancy, i.e., little overlap between keyframes; (iii) sharp and clear images are preferred. In the localization stage, the target is to find the keyframe that is most similar with the current image, and estimate their relative pose.

Though hand-crafted local features performs well in well-conditioned situations, they are fragile for large motion blur and illumination change, see Fig. 2 for an example. In this example, the ORB feature is used and the vocabulary tree is used to find feature matches [13]. Due to the large

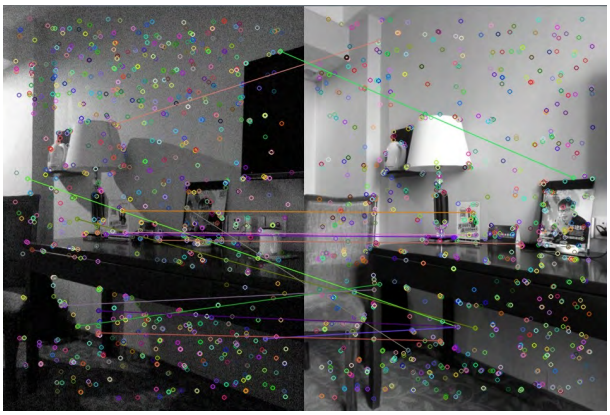


**FIGURE 2.** Two images that share common content do not have sufficient feature matching due to the large illumination change. The circle means ORB feature point, and line means correspondence.

illumination change, two images do not have sufficient feature matching even they share a large amount of common content. To make the topological mapping and localization more robust, we make the following improvements:

- CNN feature is used during mapping and localization to compare the similarity between two images.
- Image sharpness measure is used to exclude blurred images from being keyframes, and to omit blurred images during localization.
- The local feature extractor we adopted is oriented FAST and rotated BRIEF (ORB) [38], which is much more efficient and requires less memory than SIFT and Speeded-Up Robust Features (SURF) features that are adopted in previous literature.
- A non-rigid matching method, vector field consensus [32], is used to remove most of the mismatches instead of RANSAC. This procedure can improve both the matching accuracy and the matching efficiency.

In the following text of this section, we will introduce CNN feature, sharpness measure, and ORB feature extractor respectively. Subsequently, we will introduce the vector field consensus (VFC) method for geometric verification.

### A. IMAGE COMPARISON BY CNN FEATURE

The convolutional networks usually contains several convolutional layers and full-connected networks (FCN) in the last few layers. Besides the linear convolutional layers, there are some non-linear layers such as rectified linear unit (RELU), pooling and dropout. The last layer of the convolutional network is a fixed-size feature vectors. Then such features are fed to classifiers for specific tasks. It has been verified that the features has good generalization ability. Recent studies shown that the state-of-the-art performance can be achieved by networks trained on general dataset. For example, the network trained on ImageNet ILSVRC for image classification has achieved excellent performance on various visual tasks, such as object detection, scene recognition, semantic segmentation.

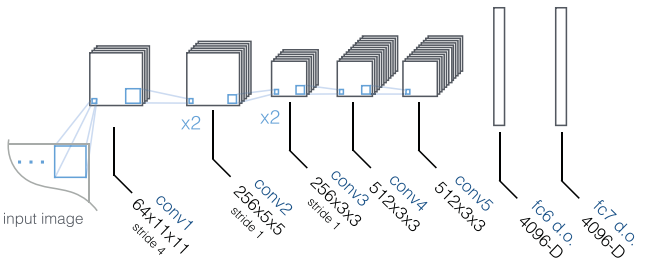



**FIGURE 3.** An illustration of the CNN-F architecture proposed in [39].

The network used in our method is CNN-F architecture [39], which is similar to AlexNet [15]. This network was pre-trained on the ImageNet ILSVRC dataset for image classification. The CNN-F architecture is shown in Fig. 3. It comprises 8 learnable layers, 5 of which are convolutional, and the last 3 layers are fully-connected. The output of last layer was fed to a soft-max layer for classification. The output of each layer can be extracted from the network and used as a holistic image feature representation.

Since the third fully connected layer and soft-max layer are adopted specifically to the ILSVRC task, we use the second FCN layer as feature representation as many other works. The output of CNN layer is a fixed-size feature vectors. In this work the feature size is 4, 096, and is normalized by $L_2$-norm. Due to the good representation of CNN feature, we simply use cosine distance as the dissimilarity between two images. In experimental section, we will demonstrate the effectiveness of CNN feature by finding similar images from image set.

### B. IMAGE SHARPNESS MEASURE

Due to the fast motion of hand-held cameras or cameras amounted in a robot, the captured images usually have severe motion blur. When finding the keyframes during mapping, we want the keyframes to be sharp images. If blurred images are selected as keyframes, it will cause the topological maps containing little informative keyframes and degrading the localization accuracy.

In this paper, we take advantage of a sharpness measure [40]. The measure captures whether the edge slope changes quickly or not. It uses difference of $\Delta$(DoM) (differences of a median-filtered image) as an indicator of edge sharpness values. Since the width of noticeable blur decreases as contrast increases, it normalizes the quantities by the contrast at the edge. This sharpness measure is general, and has been successfully used for natural scene images and text images.

During topological mapping, only frames whose sharpness measure is above a threshold can be selected as a keyframe. In this paper, the threshold is set as 0.5.

### C. ORB FEATURE EXTRACTION

Oriented FAST and rotated BRIEF, ORB for short, is a fast robust local feature detector [38], which provides an efficient alternative to SIFT and SURF. It is based on the keypoint detector features from accelerated segment test (FAST) and the visual descriptor binary robust independent elementary features (BRIEF), in which many modifications have been made to enhance the performance.

To make the ORB features spread evenly on the image plane, we divide the image plane into several non-overlapped cells, and extract nearly fixed feature points in each cell using adaptive thresholds, as that in [41]. The FAST corners are extracted at 8 scale levels with a scale factor 1.2. To speedup image retrieval and feature point matching, the extracted ORB features are further converted to bag-of-words representation using vocabulary tree [12].

## III. GEOMETRIC VERIFICATION BASED ON NON-RIGID FEATURE MATCHING

The matches introduced by feature descriptors unavoidably have mismatches. To remove the mismatches, usually the classic RANSAC is used and all the correct matches are used to estimate geometric parameters [14]. However, when the mismatch percentage is high, it takes much time to find a correct solution. In this paper, rather than using RANSAC, we take advantage of an efficient non-rigid matching, *i.e.*, vector field consensus [32], to remove the mismatches.

### A. VECTOR FIELD INTRODUCED BY IMAGE PAIRS

Assume a match is comprised by a pair $(\mathbf{u}_i, \mathbf{u}_i')$, where $\mathbf{u}_i$ and $\mathbf{u}_i'$ are positions of two feature points in two images. We convert the match into a vector field sample by a transformation $(\mathbf{u}_i, \mathbf{u}_i') \rightarrow (\mathbf{x}, \mathbf{y})$, where $\mathbf{x} = \mathbf{u}_i$, $\mathbf{y} = \mathbf{u}_i' - \mathbf{u}_i$.

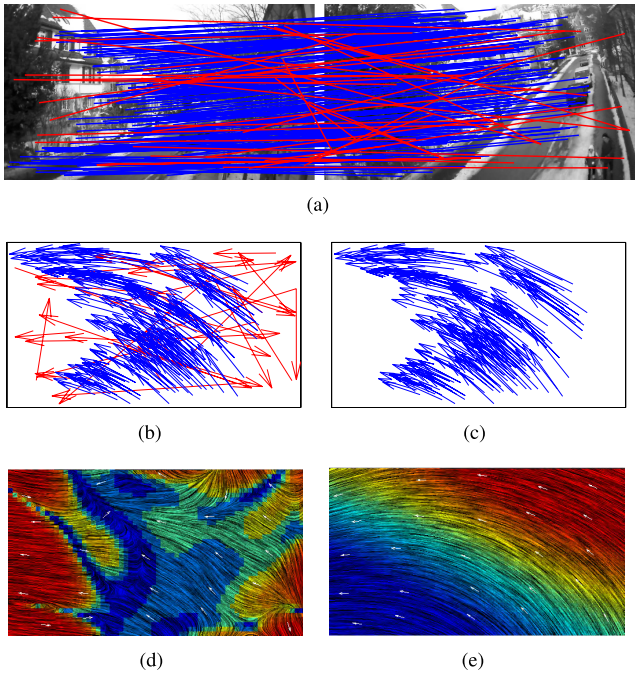



**FIGURE 4. Mismatch removal and robust vector filed interpolation. (a): An image pair and its putative correspondences. Blue and red lines represent inliers and outliers respectively. (b)(c): Motion field samples generated by all putative correspondences and only inliers, respectively. The head and tail of each arrow correspond to the positions of feature points in two images. (d)(e): The interpolated vector field using samples from (b) and (c), respectively. The visualization method is line integral convolution (LIC) [42], and color indicates the magnitude of the displacement at each point.**

Figure 4 demonstrates the relationship between mismatch removal and robust vector filed interpolation. As shown in Fig. 4(a), blue and read lines represent correct matches and mismatches, respectively. We first convert the matches into vector field training set as shown in Fig. 4(b). In the context of vector filed interpolation, we will call correct matches as inliers, and mismatches as outliers. The inlier set is shown in Fig. 4(c). Using the traditional vector field interpolation, we obtain the vector fields in Figs. 4(d) and (e) from the training sets with and without outliers, respectively. Obviously, the vector field in Fig. 4(d) is too complex since its training set contains outliers. Therefore, the problem is how to use the training set with outliers such as in Figs. 4(b) to estimate the vector field in Fig. 4(e), and distinguish outliers automatically.

### B. VECTOR FIELD CONSENSUS FORMULATION

The vector field consensus [32] uses the Bayesian framework to estimate the vector filed and distinguish outliers automatically. Denote the observed samples of vector field as $S = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^2$ and $\mathbf{y}_n \in \mathbb{R}^2$ are spatial positions of two feature points constructed from matches as described before. Our purpose is to distinguish outliers from inliers and learn a mapping $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ to fit the inliers well, where $\mathbf{f} \in \mathcal{H}$, and we assume $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) [43].

In the following we make the assumption, without loss of generality, that for inliers, the noise is Gaussian with zero mean and uniform standard deviation $\sigma$; and for outliers, the observations of output occur within a bounded region of the image plane, so the distribution is assumed to be uniform $\frac{1}{a}$, where $a$ is just a constant (*e.g.,* the volume of this region). Let $\gamma$ be the percentage of inliers which we do not know in advance. Thus the likelihood is a mixture model of distributions for inliers and outliers:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) &= \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\theta}) \\ &= \prod_{n=1}^{N} \left( \frac{\gamma}{2\pi\sigma^2} e^{-\frac{\|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2}{2\sigma^2}} + \frac{1-\gamma}{a} \right), \end{aligned} \quad (1)$$

where $\boldsymbol{\theta} = \{\mathbf{f}, \sigma^2, \gamma\}$ is the set of unknown parameters, $\mathbf{X}_{N\times 2} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^{\mathrm{T}}$, $\mathbf{Y}_{N\times 2} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^{\mathrm{T}}$. Note that the uniform distribution function is nonzero only in a bounder region, here we omit the indicator function in it for clarity.

Considering the smoothness constraint, the prior of $\mathbf{f}$ can be written as:

$$p(\mathbf{f}) \propto e^{-\frac{\lambda}{2}\|\mathbf{f}\|_{\mathcal{H}}^2}, \quad (2)$$

where $\lambda > 0$ is the regularization parameter, $\|\cdot\|_{\mathcal{H}}$ is the norm in the RKHS $\mathcal{H}$.

Given the likelihood (1) and prior (2), the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y})$ can be estimated by applying Bayes rule: $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{f})$. In order to optimally estimate $\boldsymbol{\theta}$, a MAP solution, $\boldsymbol{\theta}^*$, is obtained according to

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{f}), \quad (3)$$

with $\boldsymbol{\theta}^*$ corresponding to the estimate of the true $\boldsymbol{\theta}$. Thus the vector field $\mathbf{f}$ will be obtained.

The VFC algorithm solves with this problem under an expectationÍCmaximization (EM) framework. Whether a sample is an inlier or outlier can be determined by a latent variable after the convergence of the algorithm. In addition, a sparse approximation similar to the subset of regressors method can be adopted to $\mathbf{f}$ to accelerate the mapping estimation. This procedure significantly reduces the computational complexity from cubic to linear without sacrifice the matching accuracy [37]. An example is shown in Fig. 5. At first, all samples are treated as inliers. After a few iterations, some samples are treated as outliers. And all the outliers are identified after the algorithm converges. The code is publicly available from https://sites.google.com/site/jiayima2013/home.

## IV. EXPERIMENTAL RESULTS

To validate the image similarity by CNN feature extraction, coarse geometric verification by VFC, and sharpness

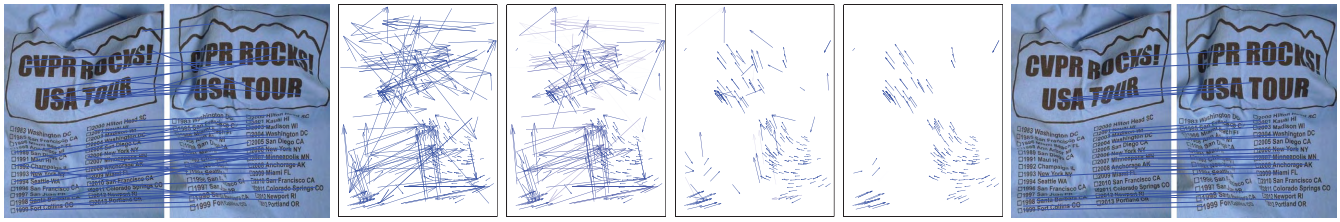



**FIGURE 5.** Iterative mismatch removal progress of VFC. The blueness indicates to what degree a sample belongs to inlier. For visibility, only 50 randomly selected matches are presented in the first image.


**FIGURE 6.** CNN features validation by image retrieval on two samples. Inquiry image is the top-left image. From left to right and from top to bottom, image retrieval results are sorted by similarity with the inquiry image. The number in bottom-left is the frame index in the video.

measure, we perform experiments on the air-ground matching dataset [44]. The air images in this dataset is collected by a micro aerial vehicle (MAV). The experiments are performed on a laptop with Intel i7-5500U CPU @ 2.40GHz. All the code is implemented in C++. In addition, we also test the whole topological mapping and navigation method using a ground robot for indoor navigation.

### A. IMAGE SIMILARITY COMPARISON BY CNN FEATURE

For the CNN-F architecture, the input image is resized as $224 \times 224$. Fast processing is ensured by the 4 pixel stride in the first convolutional layer. It takes about 4 ms for PC.

We randomly select a few frames from air-ground matching dataset [44] as query images, and retrieve the most similar images before this frame as retrieval images, see Fig. 6. It can be seen from the results that the appearance similar images are among retrieved images. These candidate images need further geometric verification by VFC.

### B. SHARPNESS MEASURE

To avoid blurred images being selected as keyframes, we calculate the sharpness for each captured frame during mapping. Figure 7 provides an sharpness comparison for two images for a same scene. Figure 7(a) is a clear image, and Fig. 7(b) has a little motion blur. The sharpness for these two images are 0.605 and 0.548, respectively. It can be seen that the relative value of sharpness is consistent with image quality.


**FIGURE 7.** Two frames share common content in [44]. (a) a clear image with sharpness measure 0.605; (b) A little blurred image with sharpness measure 0.548. The clear image (a) is preferred as representative frame.

We also perform a synthetic experiment. Given a clear image in Fig. 7(a), we add motion blur with different length $\ell$ of horizontal blur kernel. Larger length $\ell$ in motion blur kernel means severer blur is occurred. We plot the relationship between sharpness and length of blur kernel, as shown in Fig. 8. We can see that the sharpness decreases nearly monotonically with respect to the length of blur kernel. This experimental results demonstrate that the sharpness measure is consistent with the image's blurriness.

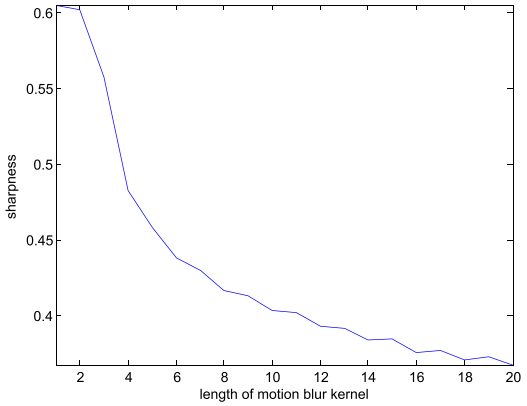



**FIGURE 8.** Sharpness with respect to length of motion blur. Horizontal axis is the length of motion blur kernel, and vertical axis is the sharpness measure.

### C. EFFICIENCY OF ORB FEATURE EXTRACTION

To validate the efficiency of ORB feature extraction, we report the average running time on this dataset. The performance comparison of SIFT, SURF and ORB is shown in Table 1. Obviously, ORB is much efficient than SIFT and SURF.

**TABLE 1.** Average running time comparison of SIFT, SURF and ORB. (unit: millisecond).

| | SIFT | SURF | ORB |
|---|---|---|---|
| feature extraction | 318 | 36 | **15** |
| feature match | 95 | 1 | **1** |

### D. GEOMETRIC VERIFICATION BY VECTOR FIELD CONSENSUS

We test the geometric verification performance of VFC. Given an inquiry image, first we use CNN feature to obtain some candidates, then further check them for geometric verification. In this paper, the top 4 images retrieved by CNN features are used for geometric verification. For each image that need geometric verification, we first build its putative matches with query image. Then we run VFC to remove mismatches. If the preserved matches is more than 30 and the inlier percentage is above 25%, it pass the geometric verification. If several images pass the verification, we select the image which has the largest number of inliers as the best match.

Figure 9 demonstrates the results for geometric verification. From Fig. 6(b), the top 4 similar candidates for frame 95 are frames 94, 93, 24 and 23. Then we verify these 4 frames by the VFC algorithm. From Fig. 9, frames 94 and 93 have enough inliers (correct matches) and high inlier percentage, and they can pass the geometric verification. While frames 24 and 23 fail to pass the geometric verification because of insufficient inliers and low inlier percentage. The detailed results are provided in Table 2.

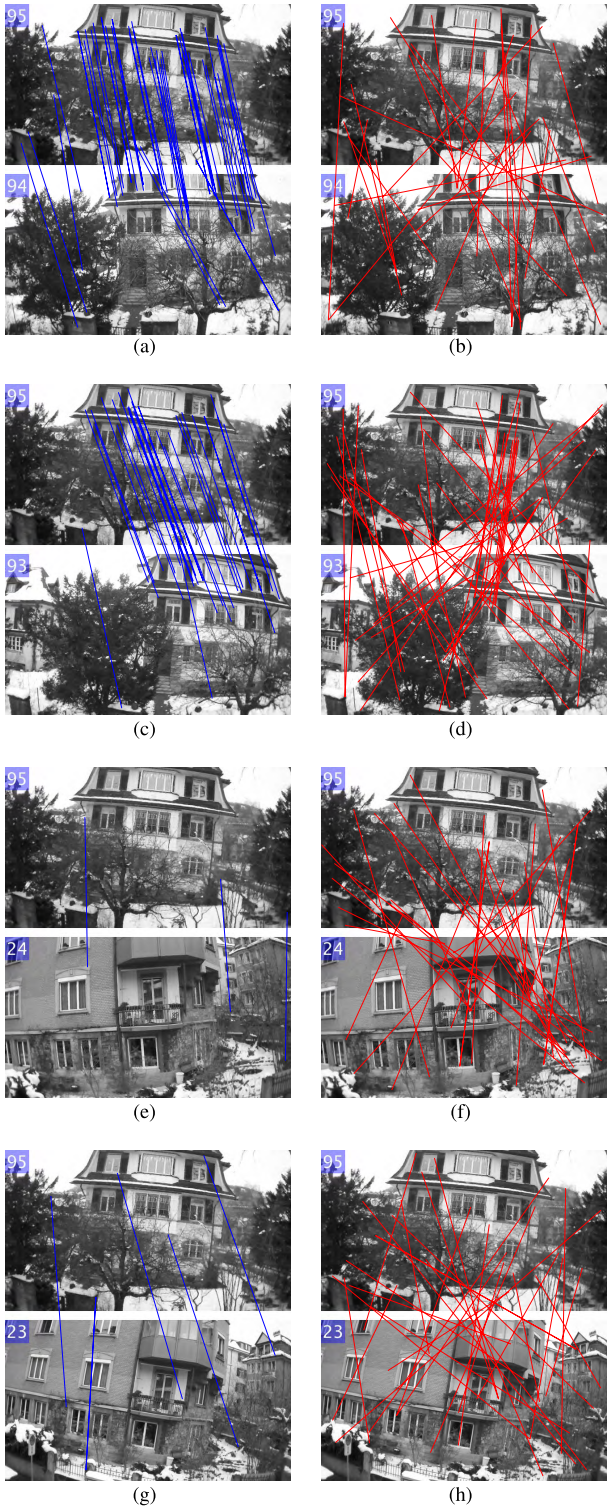



**FIGURE 9.** Geometric verification for frame 95 by VFC. The number in top-left is the frame index in the video. Subfigures (a) (c) (e) (g) in the first column demonstrate inliers; and subfigures (b) (d) (f) (h) in the second column demonstrate outliers.

Compared with RANSAC method, one additional advantage of VFC is that we do not need to know the camera intrinsic parameters. In fact, the intrinsic parameters are not provided by air-ground matching dataset.

**TABLE 2. Geometric verification results for frame 95.**

| | # inlier | # outlier | inlier percentage | pass verification |
|---|---|---|---|---|
| frame 94 | 48 | 31 | 60.8% | Yes |
| frame 93 | 31 | 44 | 41.3% | Yes |
| frame 24 | 3 | 33 | 8.3% | No |
| frame 23 | 5 | 31 | 13.9% | No |

### E. TOPOLOGICAL NAVIGATION

To validate the efficiency and robustness of our whole method, we use a ground robot to test the performance. The sequence is taken in a coworking space. It is challenging for navigation, since it contains image saturation caused by ceiling lights and reflection caused by glass walls. The size of experimental environment is about $15 \times 10$ m. The robot can successfully performs mapping and navigation in this environment. The built topological maps usually contain $50 \sim 70$ keyframes, and there is no severe blurred keyframes in the map.

**TABLE 3. Time consumption. (unit: millisecond).**

| | calculate time |
|---|---|
| CNN feature extraction | 4 |
| sharpness measure | 1 |
| ORB feature extraction | 15 |
| ORB feature match | 1 |
| vector field consensus | $2 \sim 4$ |
| place recognition / relocalization | 1 |

We list the time consumption for the main modules in Table 3. It is notable that the introduce of VFC can make RANSAC one-order faster (from a few hundred milliseconds to donzens of milliseconds).

## V. CONCLUSIONS

In this paper, we improve the efficiency and robustness for topological mapping and navigation. To achieve this goal, we take advantage of a powerful CNN feature, non-rigid feature matching, ORB feature extraction, and sharpness measure. First, we use CNN feature as the holistic image representation. It provides good metric to measure the common content between images, and it is robust to severe motion blur and large illumination change. Second, a non-rigid matching method is used for efficient geometric verification. Third, we use sharpness measure to select good keyframes. Experimental results demonstrate its effectiveness.

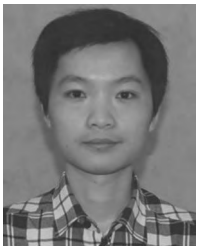
## REFERENCES

[1] E. Garcia-Fidalgo and A. Ortiz, "Hierarchical place recognition for topological mapping," *IEEE Trans. Robot.*, to be published, doi: 10.1109/TRO.2017.2704598.
[2] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Jun. 2016.
[3] J. Boal, A. Sánchez-Miralles, and A. Arranz, "Topological simultaneous localization and mapping: A survey," *Robotica*, vol. 32, no. 5, pp. 803–821, 2014.
[4] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. van Gool, "Omnidirectional vision based topological navigation," *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 219–236, 2007.
[5] F. Fraundorfer, C. Engels, and D. Nistér, "Topological mapping, localization and navigation using image collections," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2007, pp. 3872–3877.
[6] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Incremental vision-based topological SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2008, pp. 1031–1036.
[7] G. Erinc and S. Carpin, "Image-based mapping and navigation with heterogenous robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2009, pp. 5807–5814.
[8] T. Do, L. C. Carrillo-Arce, and S. I. Roumeliotis, "Autonomous flights through image-defined paths," in *Proc. Robot. Res.*, 2018, pp. 39–55.
[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[10] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 510–517.
[11] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1–8.
[12] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2161–2168.
[13] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.
[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[16] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *Proc. Austral. Conf. Robot. Autom.*, 2014, p. 1.
[17] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2015, pp. 4297–4304.
[18] A. Giusti *et al.*, "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, pp. 661–667, Feb. 2016.
[19] L. Tai and M. Liu. (2016). "Deep-learning in mobile robotics-from perception to control systems: A survey on why and why not." [Online]. Available: https://arxiv.org/abs/1612.07139
[20] J. Ma, J. Zhao, J. Tian, Z. Tu, and A. Yuille, "Robust estimation of nonrigid transformation for point set registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2147–2154.
[21] G. Wang, Z. Wang, Y. Chen, Q. Zhou, and W. Zhao, "Context-aware Gaussian fields for non-rigid point set registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5811–5819.
[22] G. Wang, Q. Zhou, and Y. Chen, "Robust non-rigid point set registration using spatially constrained Gaussian fields," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1759–1769, Apr. 2017.
[23] J. Jiang, R. Hu, Z. Wang, and Z. Cai, "CDMMA: Coupled discriminant multi-manifold analysis for matching low-resolution face images," *Signal Process.*, vol. 124, pp. 162–172, Jul. 2016.
[24] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.
[25] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.
[26] J. Jiang, C. Chen, J. Ma, Z. Wang, Z. Wang, and R. Hu, "SRLSP: A face image super-resolution algorithm using smooth regression with local structure prior," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 27–40, Jan. 2017.
[27] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug. 2014.
[28] J. Jiang, J. Ma, C. Chen, X. Jiang, and Z. Wang, "Noise robust face image super-resolution through smooth sparse representation," *IEEE Trans. Cybern.*, to be published.
[29] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4220–4231, Oct. 2014.

[30] K. Yang, A. Pan, Y. Yang, S. Zhang, S. H. Ong, and H. Tang, "Remote sensing image registration using multiple image features," *Remote Sens.*, vol. 9, no. 6, p. 581, 2017.
[31] J. Ma, J. Jiang, C. Liu, and Y. Li, "Feature guided Gaussian mixture model with semi-supervised em and local geometric constraint for retinal image registration," *Inf. Sci.*, vol. 417, pp. 128–142, Nov. 2017.
[32] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, Apr. 2014.
[33] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.
[34] J. Ma, J. Zhao, H. Guo, J. Jiang, H. Zhou, and Y. Gao, "Locality preserving matching," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 4492–4498.
[35] Y. Yang, S. H. Ong, and K. W. C. Foong, "A robust global and local mixture distance based non-rigid point set registration," *Pattern Recognit.*, vol. 48, no. 1, pp. 156–173, 2015.
[36] J. Ma, J. Zhao, J. Jiang, and H. Zhou, "Non-rigid point set registration with robust transformation estimation under manifold regularization," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4218–4224.
[37] J. Ma, J. Zhao, J. Tian, X. Bai, and Z. Tu, "Regularized vector field learning with sparse approximation for mismatch removal," *Pattern Recognit.*, vol. 46, no. 12, pp. 3519–3532, 2013.
[38] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2011, pp. 2564–2571.
[39] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.
[40] J. Kumar, F. Chen, and D. Doermann, "Sharpness estimation for document and scene images," in *Proc. Int. Conf. Pattern Recognit.*, 2012, pp. 3292–3295.
[41] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, May 2015.
[42] B. Cabral and L. C. Leedom, "Imaging vector fields using line integral convolution," *Comput. Graph.*, vol. 27, pp. 263–270, Aug. 1993.
[43] C. A. Micchelli and M. A. Pontil, "On learning vector-valued functions," *Neural Comput.*, vol. 17, no. 1, pp. 177–204, 2005.
[44] A. L. Majdik, Y. Albers-Schoenberg, and D. Scaramuzza, "MAV urban localization from Google street view data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 3979–3986.

**JIAYI MA** received the B.S. degree from the Department of Mathematics and the Ph.D. degree from the School of Automation, Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively.

From 2012 to 2013, he was an Exchange Student with the Department of Statistics, University of California at Los Angeles, Los Angeles, CA, USA. He is currently an Associate Professor with the Electronic Information School, Wuhan University, Wuhan, where he holds a post-doctoral position from 2014 to 2015. His current research interests include computer vision, machine learning, and pattern recognition.



**JI ZHAO** received the B.S. degree in automation from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2005, and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2012.

From 2012 to 2014, he was a Post-Doctoral Research Associate with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently with ReadSense Ltd., Shanghai, China. His research interests include computer vision and machine learning.