Journal of Intelligent & Robotic Systems
https://doi.org/10.1007/s10846-017-0755-7

Combining Multiple Image Descriptions for Loop Closure Detection

Xiaolong Wang[1] · **Guohua Peng**[1] · **Hong Zhang**[2]

Received: 25 May 2017 / Accepted: 1 December 2017

Abstract
The success of visual loop closure detection depends on the discrimination ability of the image descriptions. Different sources of image descriptions may carry complementary information as well as redundant information. Though integrating them properly can be beneficial, a main obstacle is the lack of analytical quality indicators to weigh different descriptions jointly. Inspired by the linear discriminant analysis, we propose an efficacy index to evaluate the weighted linear combinations of multiple image descriptions for loop closure detection. When a collection of image descriptions is given, the optimal weights maximizing the efficacy index are deduced analytically. As negative weights may negatively affect the performance of detection, a gradient descent algorithm is further proposed to jointly optimize the nonnegative weights. We use the proposed weighting strategies to combine the image descriptions extracted from multiple local image patches by multiple descriptor extractors. It is experimentally demonstrated that our weighted combinations of image descriptions can greatly improve the performance of loop closure detection by emphasizing informative components and de-emphasizing redundant components.

Keywords Visual loop closure detection · Feature selection · Weighted linear combination · Linear discriminant analysis · Nonnegative optimization

1 Introduction

Visual loop closure detection (LCD) is a key step in vision-based simultaneous localization and mapping (SLAM) algorithms in robotics research. It efficiently detects the loop closures, i.e., the places a robot or vehicle revisits, by matching image descriptions. Using the knowledge of matching places, the SLAM system can optimize the map and reduce mapping and localization errors. In general, a visual LCD method includes two parts. First, each keyframe the robot perceives is represented by one or a collection of image descriptions, such as SURF [1] or SIFT [2] descriptors. Then the detection of the loop closures of an image is realized by searching the images with high similarity scores in the robot map. A retrieval method, such as the bag of words [3] or the locality-sensitive hashing [4], is often applied to improve the search efficiency [5, 6].

Two established and widely applied ideas of the extraction of image descriptions can be combined to broaden the sources of image representations for loop closure detection. The first idea is to partition an image into local patches and represent the image by the collection of local descriptors extracted from individual patches. The collection of local patch descriptors carries more discriminative information than a single global descriptor as the former captures local details of the image as well as the spatial arrangement among local patches. The second idea is to extract several types of image descriptions from each local patch using different descriptor extractors. As different image descriptors may detect complementary information, combining them can benefit the recognition of places.

However, the discrimination abilities of the image descriptions depend on the datasets, descriptor extractors

✉ Xiaolong Wang
wangxiaolongnwpu@163.com

Guohua Peng
penggh@nwpu.edu.cn

Hong Zhang
hzhang@ualberta.ca

[1] Department of Applied Mathematics, School of Natural and Applied Sciences, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China

[2] Department of Computing Science, University of Alberta, Edmonton, AB, Canada

Published online: 11 December 2017

as well as the regions of support. When various image descriptions are combined, some of them may be redundant or suboptimal. Therefore a careful weighting strategy is required to integrate different sources of image descriptions in order to assemble a stronger image description. To carry out such a strategy, a capable efficacy index is crucial to weigh the different image descriptions jointly.

In this paper, we propose a general statistical framework to evaluate and optimize the linear combinations of image descriptions for visual LCD. To begin with, we treat the loop closure detection as a binary classification problem where the two classes are the true and false matches of image pairs. The distances between the corresponding image descriptions in two images are interpreted as features of classifiers. Hence different image descriptions are converted into real-valued classifiers, which can be jointly evaluated by statistical tools.

Next, an efficacy index is proposed to evaluate the weighted combinations of different sources of image descriptions. Unlike the other well-known criteria, such as the F-score or the highest recall at 100% precision, the proposed efficacy index is differentiable with respect to the weights of the combinations. The differentiability enables us to optimize the weights jointly in order to assemble effective combinations. We show that the weights maximizing the index are close to the weights by the linear regression.

The weights maximizing the efficacy index may contain negative components. It can negatively affect the performance of LCD since the image pairs from the different places with dissimilar descriptions can result in low weighted distances and be misclassified as true matches. To overcome the problem, a gradient descent algorithm is proposed to maximize the efficacy index with respect to nonnegative weights. We show that the algorithm indeed finds nonnegative local optimal weights. Experiments verify that the proposed weighted combinations of image descriptions outperform the equally weighted concatenation significantly by relying on informative image descriptions and ignoring redundant image descriptions.

The remaining part of this paper is organized in the following way. Section 2 reviews the related research. Section 3 contains the proposed framework for combining image descriptions. Section 4 applies the results in Section 3 to solve the loop closure detection problem. Section 5 presents extended experiments. The work is summarized in Section 6.

2 Related Work

Visual loop closure detection is technically similar to image retrieval but with the temporal information when the images are captured by a robot sequentially as it travels. Successful visual LCD algorithms often use contextual information in images to enhance the detection. For example, the FAB-MAP [7] algorithm models the co-occurrence frequency of visual words in the images from map places. SeqSLAM [8] measures the similarity between image sequences instead of individual images. The bag of words based visual LCD has been used in several visual SLAM systems such as LSD-SLAM [9] and ORB-SLAM [10].

One key decision of visual LCD is the selection of image descriptions. The SIFT and SURF descriptors have been widely applied as two reliable local image descriptors [11]. GIST [12], BRIEF [13], LBP [14] and ORB [15] descriptors have also been widely adopted for place recognition and loop closure detection [16–19]. The Fisher vector [20] and the vector of locally aggregated descriptors (VLAD) [21, 22] of these hand-crafted descriptors take advantage of vocabulary learning, enhancing the performance of recognition [23]. Recently, the deep features based on convolutional neural networks have achieved outstanding performance [23, 24]. Though numerous image descriptors are available, the research on combining them for loop closure detection is limited.

Combining multiple image descriptions is beneficial to scene recognition as different sources of image descriptions may supply complementary information of the scenes. However, these descriptions may also provide redundant information which needs to be properly handled. A probabilistic framework for combining classifiers using distinct pattern representations was proposed in [25]. Several classifier combination strategies were discussed, such as the sum rule and product rule, which were adopted in [26] to fuse local features for visual localization. However, these simple rules can not decide the relative weights among local features.

Many efficacy indices, such as mutual information adopted by MRMR [27], logistic regression [28] and Fisher criterion [29], have been applied to evaluate the usefulness of individual or pairwise image features in feature selection researches. The subset of image features with low redundancy and high relevance is often found in a heuristic way by incrementally including the features with large efficacy scores. These incremental styles of feature selection often has two drawbacks. Firstly, the combination of those features with relatively low individual scores may have high joint scores and vice versa [30, 31]. Therefore these methods often find suboptimal solutions. Secondly, in a feature selection process, the state of a feature is binary, i.e., being present or absent, whereas the suitable continuous weights may improve the recognition further. The most relevant technique to our method is the linear discriminant analysis (LDA) [32, 33], which finds the linear weights of the classifier combination jointly. LDA minimizes the

intra-class variance and maximizes the inter-class variance, or equivalently, maximizes the Fisher criterion of the linear combination. The Fisher criterion can also be used to jointly select a predefined number of features [31]. Though using the idea behind the Fisher criterion, we define a slightly different efficacy index tailored for LCD. We explain and experimentally verify that our definition outperforms the Fisher criterion. Besides, LDA allows negative weights, which may cause negative impacts on the performance of LCD. Instead, we develop an algorithm to optimize the nonnegatively weighted combination.

Li et al. [34] explored the use of different importance weights on the local patches within an image for place recognition. They partitioned each image with a 4×4 grid and used a Gaussian distribution to weigh the dissimilarity scores of image descriptions between local patches. A key difference of our work from theirs is that they modeled the statistics of each image pair separately. Instead, our model is based on the statistics of the whole image sequences, which are more statistically significant since there are much more samples than considering individual image pairs. We also present a thorough analysis of our method, making it applicable for various types of image descriptions, including but not limited to a fixed grid of local patch descriptors.

3 Weighting Image Descriptions

In this section, we describe a framework to evaluate the image descriptions for visual loop closure detection. The goal is to express the quality of the linear combination of image descriptions as a differentiable function of the linear weights such that the optimization of the weights can be carried out. The framework is summarized in Fig. 1.

In Section 3.1 we model the loop closure detection as a binary classification problem and treat the distances between image descriptions as features of classifiers. An efficacy index is proposed in Section 3.2 to evaluate individual classifiers as well as classifier combinations. Two weighting strategies are proposed in Sections 3.3 and 3.4 to maximize the efficacy index of the linear combination of classifiers.

3.1 A Classification Perspective of Visual LCD

We suppose that there are two image sequences A and B. A label $l(I, J)$ is assigned to a pair of images (I, J) where $I \in A$ and $J \in B$ such that

$$l(I, J) = \begin{cases} 1, & \text{if } I \text{ and } J \text{ come from the same place;} \\ 0, & \text{otherwise.} \end{cases}$$

Hence the majority of image pairs from the two sequences can be separated into two classes, i.e., the class of true matches

$$T = \{(I, J) | l(I, J) = 1\} \tag{1}$$

and the class of false matches

$$F = \{(I, J) | l(I, J) = 0\}. \tag{2}$$

A small fraction of image pairs from the adjacent but not identical places are not assigned to T or F since their labels are ambiguous. The sets T and F can be determined for the

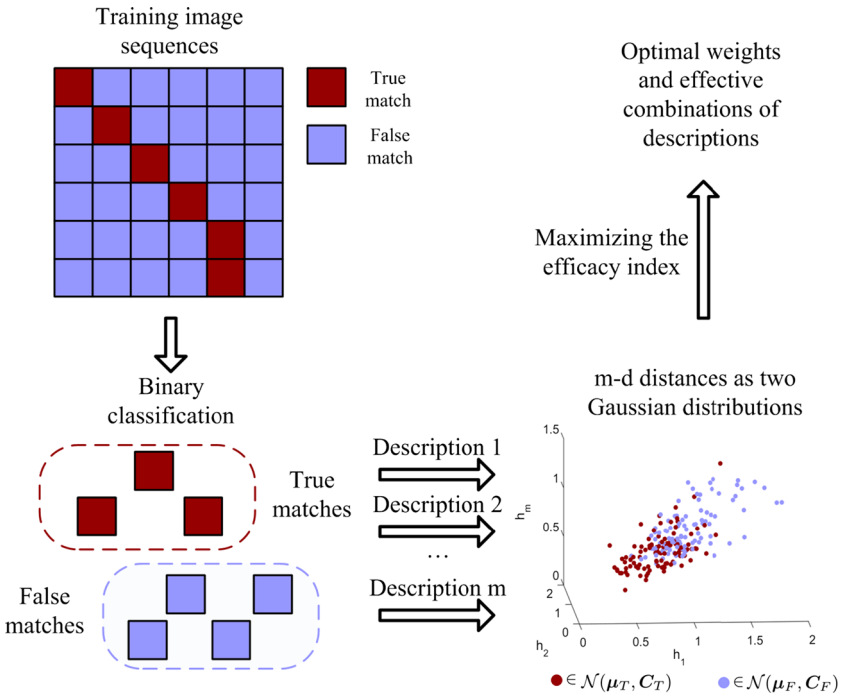

Fig. 1 A summary of the proposed work

image sequences where the ground truths of loop closures are obtained by manual labeling or GPS.

A real-valued classifier h for solving the binary classification is defined as a function

$$h : T \cup F \to \mathbb{R} \tag{3}$$

which is capable of separating T from F. The linear combination of classifiers is still a classifier in the sense that

$$(w_1 h_1 + w_2 h_2)(I, J) = w_1 h_1(I, J) + w_2 h_2(I, J) \tag{4}$$

where w_1 and w_2 are real numbers, and h_1 and h_2 are classifiers. In the context of visual LCD, $h(I, J)$ is the distance between the image descriptions representing I and J. Hence h can be analyzed by evaluating the statistics of the measurements of h on T and F.

We adopt Gaussian distributions to model the measurements a classifier h operating on true matches and false matches, i.e.,

$$\begin{aligned} h(x) &\sim \mathcal{N}(\mathrm{E}_T[h], \mathrm{Var}_T[h]) \ \text{ for } x \in T, \\ h(x) &\sim \mathcal{N}(\mathrm{E}_F[h], \mathrm{Var}_F[h]) \ \text{ for } x \in F \end{aligned} \tag{5}$$

where $\mathrm{E}_T[\cdot]$, $\mathrm{Var}_T[\cdot]$, $\mathrm{E}_F[\cdot]$ and $\mathrm{Var}_F[\cdot]$ are the expectations and variances of the measurements of h on T and F, respectively. When the sizes of the samples are large enough, the central limit theorem suggests that the Gaussian distributions are good approximations of the real distributions. Our preliminary experiments have also verified it.

In general, the image descriptions from the same places are more similar than from the different places. Therefore on average, the distance of the descriptions of true matches is lower than that of false matches, i.e.,

$$\mathrm{E}_T[h] \leq \mathrm{E}_F[h] \tag{6}$$

always holds.

3.2 The Novel Efficacy Index

Based on the intuition that a strong classifier should separate true matches from false matches as much as possible, we define a nonnegative efficacy index γ to evaluate the classifier h

$$\gamma[h] = \frac{\mathrm{E}_F[h] - \mathrm{E}_T[h]}{\sqrt{\mathrm{Var}_F[h]}}. \tag{7}$$

Intuitively, suppose that we are given a true match $x \in T$ as well as two classifiers a and b. If classifier a is better than b to distinguish x from F, the probability of the distance of x in the distribution of false matches should be lower by a than that by b. By the assumption of Gaussian models, this observation is described as

$$\frac{a(x) - \mathrm{E}_F[a]}{\sqrt{\mathrm{Var}_F[a]}} < \frac{b(x) - \mathrm{E}_F[b]}{\sqrt{\mathrm{Var}_F[b]}}. \tag{8}$$

If a is stronger than b, the inequality should hold on average for all true matches, i.e.,

$$\gamma[a] > \gamma[b]. \tag{9}$$

Therefore we conclude that an effective classifier should have a high γ value.

In Appendix A we prove that for any two classifiers a and b, if $\gamma[a] < \gamma[b]$, then $\gamma[a] < \gamma[a + b]$. This implies that the addition of two classifiers is always stronger than the weaker of the two. However, there is no guarantee whether the addition is stronger than either of the individual classifiers. A simple contradiction can be constructed if we suppose that a is the distance of any effective image descriptions while b is the distance of uniform random vectors representing the images. The discrimination ability of the addition $a + b$ is better than b but not as good as a since random noises are contained in $a + b$. Therefore, a careful weighting strategy is crucial, which we will detail in the next section.

The γ index is related to the Fisher criterion in the linear discriminant analysis. The latter is defined as $\phi[h]$ where

$$\phi[h] = \frac{\mathrm{E}_F[h] - \mathrm{E}_T[h]}{\sqrt{\mathrm{Var}_T[h] + \mathrm{Var}_F[h]}}. \tag{10}$$

The numerator of ϕ^2 is the between-class separation of the measurements and the denominator is the within-class separation. The key distinction between our definition γ and the Fisher criterion ϕ is that we omit the variance of the true matches $\mathrm{Var}_T[h]$ in the denominator. The reason is that $\mathrm{Var}_T[h]$ cannot be measured as accurately as $\mathrm{Var}_F[h]$ since the number of true matches is often much less than the number of false matches. Consider two image sequences, each of which has n images. The number of true matches is at most $O(n)$ while the number of false matches is $O(n^2)$. Besides, the images from the same places are often poorly aligned, contributing biased measurements to $\mathrm{Var}_T[h]$. As a result, we only use the variance of false matches in our definition of γ. The experiments in Section 5 show that the performance of loop closure detection using γ is significantly better than using ϕ.

3.3 Linear Combination of Classifiers

Given a collection of classifiers $\Omega = [h_1, h_2, \ldots, h_m]^T$, we assume the distributions of $\Omega(x) = [h_1(x), \ldots, h_m(x)]^T$ for an image pair x in T or F are m dimensional Gaussian distributions, namely,

$$\begin{aligned} \Omega(x) &\sim \mathcal{N}(\mu_T, C_T) \ \text{ for } x \in T, \\ \Omega(x) &\sim \mathcal{N}(\mu_F, C_F) \ \text{ for } x \in F \end{aligned} \tag{11}$$

where

$$\begin{aligned} \mu_T &= [\mathrm{E}_T[h_1], \ldots, \mathrm{E}_T[h_m]]^T, \\ \mu_F &= [\mathrm{E}_F[h_1], \ldots, \mathrm{E}_F[h_m]]^T \end{aligned} \tag{12}$$

and the ij-entries of the covariances matrices C_T and C_F are

$$\begin{aligned} \mathrm{Cov}_T[h_i, h_j] &= \mathrm{E}_T\left[(h_i - \mathrm{E}_T[h_i])(h_j - \mathrm{E}_T\left[h_j\right])\right], \\ \mathrm{Cov}_F[h_i, h_j] &= \mathrm{E}_F\left[(h_i - \mathrm{E}_F[h_i])(h_j - \mathrm{E}_F\left[h_j\right])\right]. \end{aligned} \tag{13}$$

When a weight vector $w = [w_1, w_2, ..., w_m]^T$ is available, a weighted linear combination h can be constructed as

$$h = \Omega^T w = \sum_{i=1}^{m} w_i h_i. \tag{14}$$

Measured by γ, the efficacy of h is a function of w, i.e.,

$$\gamma[h] = \frac{s^T w}{\sqrt{w^T C_F w}} = R(w) \tag{15}$$

where s is

$$s = \mu_F - \mu_T. \tag{16}$$

The observation (6) ensures that s is nonnegative.

The generalization coincides with the general belief that combining more classifiers can improve the efficacy of recognition. To see this, we set the weight w_m of the last classifier h_m to 0. Then the maximal efficacy that the first $m-1$ classifiers can achieve is nonincreasing, since

$$\max_{\{w|w_m=0\}} R(w) \leq \max_{w} R(w). \tag{17}$$

The optimal weights maximizing $\gamma[h]$ can be obtained if we maximize $R(w)$ with respect to w. Setting $x = C_F^{1/2} w$ and $t = C_F^{-1/2} s$, we have

$$\max_{w} R(w) = \max_{x} \frac{t^T x}{\|x\|} = \max_{\{y|\|y\|=1\}} t^T y = \|t\| \tag{18}$$

where $\|\cdot\|$ is the Euclidean norm. Hence when $x = t$ or

$$w^* = C_F^{-1} s, \tag{19}$$

the maximum is attained [33]. In addition, we note that $R(w)$ is positively homogeneous, i.e., for any weight vector w and any positive real number λ, $R(w) = R(\lambda w)$. Thus λw^* also maximizes $R(w)$ for each $\lambda > 0$. This is desirable since $\Omega^T w$ and $\Omega^T(\lambda w)$ function identically. It also can be shown that λw^* is the only solution making the Jacobian $J(w)$ of $R(w)$ vanishes, where

$$J(w) = \frac{(w^T C_F w)s - (s^T w)C_F w}{(w^T C_F w)^{3/2}}. \tag{20}$$

To gain insight into the optimal weight vector w^*, we assume here that the off-diagonal entries of C_F are 0. When the classifiers in Ω measure the non-overlapping regions of the images, the correlation between each two classifiers is weak and thus the assumption is approximately fulfilled. Therefore C_F can be simplified to

$$\mathrm{Cov}_F[h_i, h_j] = \begin{cases} \sigma_i^2 & \text{if } i = j; \\ 0 & \text{otherwise} \end{cases}$$

where $\sigma_i = \sqrt{\mathrm{Var}_F[h_i]}$. In this special case, $w^* \propto w^{DC}$, where we define the decoupled weight vector w^{DC} as

$$w^{DC} = [\frac{s_1}{\sigma_1^2}, \ldots, \frac{s_m}{\sigma_m^2}]^T = [\frac{\gamma(a_1)}{\sigma_1}, \ldots, \frac{\gamma(a_m)}{\sigma_m}]^T \tag{21}$$

and use $x \propto y$ to denote that $x = cy$ for some positive real value c. Eq. 21 shows that the classifier component h_i with a relatively large between-class separation and a small variance on the false matches should have a large weight. If all the classifiers possess the same γ value, the classifiers with smaller variations contribute more to the combination.

When the classifiers are strongly correlated, the combination weighted by w^{DC} is often inferior to the one weighted by w^*. As the correlations among classifiers are ignored by w^{DC}, w^{DC} is unable to deal with the redundancy among classifiers. The comparisons of w^{DC} and w^* are detailed in Sections 5.3.2 and 5.4.

The optimal weight vector w^* is closely related to the linear regression. Suppose p true matches and q false matches are sampled from T and F and measured by the m classifiers in Ω, resulting in p observations $\{x_i\}_{i=1}^p$ and q observations $\{y_j\}_{j=1}^q$, respectively. Each observation is a vector of length m. When the two different real-valued aims u and v of the true matches and false matches are given, the linear regression seeks an optimal weight vector w^{RG} minimizing the sum of squared residuals

$$\min_{w} \|Aw - b\|^2, \tag{22}$$

where

$$\begin{aligned} A &= [x_1 - \bar{\mu}, \ldots, x_p - \bar{\mu}, y_1 - \bar{\mu}, \ldots, y_q - \bar{\mu}]^T, \\ b &= [u, \ldots, u, v, \ldots, v]^T. \end{aligned} \tag{23}$$

The sample mean $\bar{\mu}$ is

$$\bar{\mu} = \frac{1}{p+q}\left(\sum_{i=1}^{p} x_i + \sum_{j=1}^{q} y_j\right) = \alpha\bar{\mu}_T + \beta\bar{\mu}_F \tag{24}$$

where $\alpha = p/(p+q)$, $\beta = q/(p+q)$ and the sample means of true matches and false matches are

$$\bar{\mu}_T = \frac{1}{p}\sum_{i=1}^{p} x_i, \bar{\mu}_F = \frac{1}{q}\sum_{j=1}^{q} y_j, \tag{25}$$

respectively.

The well-known regression solution of Eq. 22 is

$$w^{RG} = (A^T A)^{-1} A^T b \tag{26}$$

A direct calculation shows that

$$\begin{aligned} \frac{1}{p+q}A^T A = &\frac{\alpha}{p}\sum_{i=1}^{p}(x_i - \bar{\mu}_T)(x_i - \bar{\mu}_T)^T + \frac{\beta}{q}\sum_{j=1}^{q}(y_j - \bar{\mu}_F)(y_j - \bar{\mu}_F)^T \\ &+ \alpha\beta(\bar{\mu}_F - \bar{\mu}_T)(\bar{\mu}_F - \bar{\mu}_T)^T \end{aligned} \tag{27}$$

and

$$\frac{1}{p+q}A^T b = \alpha\beta(v-u)(\bar{\mu}_F - \bar{\mu}_T). \tag{28}$$

Therefore when the number of the samples increases, w^{RG} converges to

$$\begin{aligned} w^{RG} &\to \left[\alpha C_T + \beta C_F + \alpha\beta(\mu_F-\mu_T)(\mu_F-\mu_T)^T\right]^{-1}\alpha\beta(v-u)(\mu_F-\mu_T) \\ &\propto \left(\alpha C_T + \beta C_F + \alpha\beta s s^T\right)^{-1} s. \end{aligned} \tag{29}$$

When the number of false matches are much larger than the number of true matches in the applications of visual LCD, $\alpha \approx 0$ and $\beta \approx 1$. Therefore the optimal weight vector w^* maximizing $\gamma[h]$ approximates to the regression solution w^{RG}. The experiments in Section 5.3.3 show that the combinations using w^* and w^{RG} are almost identical, in accord with the present analysis.

3.4 Combination of Nonnegatively Weighted Classifiers

As there is no constraint on w when we maximize $R(w)$ in Section 3.3, the components of w^* corresponding to the classifiers of inferior quality can be negative. To see this, we consider a stronger classifier h_1 and a weaker classifier h_2 and denote

$$s = [s_1, s_2]^T, C_F = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{30}$$

where σ_1 and σ_2 are the standard deviations of h_1 and h_2 on F, respectively, and ρ is the correlation coefficient between h_1 and h_2 on F. The optimal weight vector maximizing (15) is

$$w^* = \begin{bmatrix} w_1^* \\ w_2^* \end{bmatrix} = C_F^{-1} s \propto \begin{bmatrix} \sigma_2^2 s_1 - \rho\sigma_1\sigma_2 s_2 \\ -\rho\sigma_1\sigma_2 s_1 + \sigma_1^2 s_2 \end{bmatrix} = \begin{bmatrix} (1-\rho\lambda)/\sigma_1 \\ (\lambda-\rho)/\sigma_2 \end{bmatrix} \tag{31}$$

where $\lambda = \gamma[h_2]/\gamma[h_1]$, $\gamma[h_1] = s_1/\sigma_1$ and $\gamma[h_2] = s_2/\sigma_2$.

As h_1 is stronger than h_2, $\lambda < 1$. Since $\rho \leq 1$ always holds, $\rho\lambda < 1$. As a result, $w_1^* > 0$, i.e., the weight of the stronger classifier is always positive. In contrast, the sign of w_2^* depends on λ and ρ. w_2^* is negative when

$$\lambda = \frac{\gamma[h_2]}{\gamma[h_1]} < \rho, \tag{32}$$

i.e., h_2 is considerably inferior to h_1. The negative weights may not occur when the correlations among classifiers are weak so that ρ is too small to satisfy the inequality (32). They also barely happen when the number of classifiers is limited so that a small enough ratio λ cannot be assembled. When there are a large amount of classifiers, w^* may include negative components.

The negative weights may lead to problems. According to our definition, all classifiers are the distances of image descriptions. If negative weights are included, dissimilar image pairs from the different places can have low weighted distances when they have dissimilar enough descriptions indexed by the negative components of w^*. As a result, false matches will be recognized as true matches.

To avoid negative weights, the non-negativity of the weight vector is required and the maximization problem (18) becomes to

$$\max_{w \geq 0} R(w). \tag{33}$$

We propose a gradient decent algorithm to solve (33), as shown in Algorithm 1. When the algorithm converges, we obtain the weight vector

$$w^+ = [w_1^+, w_2^+, \ldots, w_m^+]^T \tag{34}$$

consisting of positive or zero-valued components. Suppose that u is the set of indices i such that $w_i^+ > 0$, v is the set of indices i such that $w_i^+ = 0$ and J is the Jacobian matrix of $R(w)$ at w^+, the following theorem is proved in Appendix B.

Algorithm 1 The gradient descent algorithm solving (33)

Require: the difference of means s, the covariance matrix C_F, the number of iterations K, the step size $\delta > 0$

1: $w_{(0)} = \frac{C_F^{-1}s}{\|C_F^{-1}s\|}$
2: **for** i = 1:K **do**
3: $q_{(i)} = w_{(i-1)} + \delta \cdot J(w_{(i-1)})$, where the Jacobian J is defined in Eq. 20
4: Set the negative components of $q_{(i)}$ to 0
5: $w_{(i)} = \frac{q_{(i)}}{\|q_{(i)}\|}$
6: **end for**
7: Return $w^+ = w_{(K)}$

Theorem 1 *w^+ locally maximizes (33). w^+ maximizes*

$$\max_{\{w|w_v=0\}} R(w) \tag{35}$$

where w_v is the subvector of w, consisting of the entries with the indices in v. In addition, $J_u = 0$ and $J_v \leq 0$.

Theorem 1 shows that the nonnegative solution w^+ is at least locally optimal. In addition, w^+ is the direct sum of two subvectors, the positive vector w_u^+ and the zero vector w_v^+. All the classifiers in Ω are separated accordingly into two groups, indexed by u and v. If we only consider the classifiers whose indices are in u, the unconstrained optimal weight vector maximizing (15) is exactly w_u^+. These classifiers are informative and mutually supported. The remaining classifiers, whose indices are in v, are inferior as they can further increase the combined γ

only when negative weights are allowable. Thus they are excluded by the algorithm. Though w^+ is locally optimal, we find that Algorithm 1 always converges to the same w^+ when it is randomly initialized.

The deductions of the optimal weight vector w^* and the nonnegative weight vector w^+ are not only applicable for γ but also for the other similar efficacy indices, such as the Fisher criterion ϕ. For example, the analogues of Eqs. 15 and 19 using ϕ as the efficacy index are

$$\phi[h] = \frac{s^T w}{\sqrt{w^T (C_T + C_F) w}} \tag{36}$$

and

$$w^* = (C_T + C_F)^{-1} s. \tag{37}$$

We use the subscript γ or ϕ to distinguish the weight vectors derived by maximizing $\gamma[h]$ or $\phi[h]$, respectively.

4 Weighting Strategies for LCD

We now apply the weighting strategies to solve the loop closure detection problem by combining several image descriptions. We partition each image into M fixed patches and use N descriptor extractors to extract N descriptions in each patch. Therefore each image is represented by $m = MN$ image descriptions. Though we apply a regular grid of non-overlapping patches to cover the image throughout the experiments for its popularity [34], other overlapping or nonoverlapping patches patterns, such as [35, 36], can also be applicable. The types of descriptor extractors are not restricted either. In the experiments, we show that the color-based, shape-based and clustering-based descriptors can be combined to improve the discrimination ability.

Suppose each of the two images I and J is represented by m image descriptions $\{f_i\}_{i=1}^m$ and $\{g_i\}_{i=1}^m$, respectively. Specifically, f_i and g_i are the descriptions extracted from the i_1th image patch in I and J by the i_2th descriptor extractor such that $i = (i_2 - 1)M + i_1$. The classifier h_i is defined as the L1-normalized distance between the corresponding image descriptions in the two images,

$$h_i(I, J) = \left\| \frac{f_i}{\|f_i\|_1} - \frac{g_i}{\|g_i\|_1} \right\|_1. \tag{38}$$

Typically in visual LCD, the images as well as local patches from the same places are assumed to be roughly aligned. Therefore the distances between the corresponding image descriptions from the the same patch locations are capable of measuring the similarity of the scenes.

Various linear combinations of the m classifiers can be constructed by substituting different weight vectors into Eq. 14. The baseline is the equally weighted concatenation $h^{EQ} = \Omega^T w^{EQ}$ weighted by the equal weight vector w^{EQ}. The optimal combination $h_\gamma^* = \Omega^T w_\gamma^*$ using Eq. 19 maximizes $\gamma[h]$. The nonnegative combination $h_\gamma^+ = \Omega^T w_\gamma^+$ from Algorithm 1 locally maximizes $\gamma[h]$. For the purpose of comparison, we also adopt the Fisher criterion ϕ, i.e., Eq. 10, as the efficacy index. $h_\phi^* = \Omega^T w_\phi^*$ and $h_\phi^+ = \Omega^T w_\phi^+$ are the combination maximizing $\phi[h]$ and the nonnegative combination from Algorithm 1 using ϕ, respectively. A training step is required to calculate the mean values and covariance matrices of the classifiers for the calculation of these weight vectors.

The time consumption of the training stage is mainly for the calculation of the distances and the covariance matrix C_F. We assume that K_1 and K_2 images in the image sequences A and B are selected into the training set. The average length of the N types of image descriptors is L. It spends $O(K_1 K_2 L m)$ to calculate the distances between the image descriptions in the training sequences and $O(K_1 K_2 m^2)$ to calculate C_F. Hence the time complexity of the training step is $O(K_1 K_2 m^2 + K_1 K_2 L m)$.

In the testing stage, $O(mL)$ multiplications are required to calculate the m distances between the m corresponding image descriptions for each image pair. To calculate the novel weighted distances, it only takes m additional multiplications, which are relatively negligible.

The weighted combinations can be easily integrated into the bag of words or any other image retrieval framework, though we do not present experiments here. In general, for any weights $\{w_i\}_{i=1}^m$, the weighted combination of the m distances can be formulated as a single distance,

$$h(I, J) = \sum_{i=1}^{m} w_i h_i(I, J) = \|H - G\|_1 \tag{39}$$

where H and G are the concatenations of the image descriptions of I and J weighted by the respective weights, i.e.,

$$\begin{aligned} H &= \left[w_1 \frac{f_1}{\|f_1\|_1}, \ldots, w_m \frac{f_m}{\|f_m\|_1} \right] \\ G &= \left[w_1 \frac{g_1}{\|g_1\|_1}, \ldots, w_m \frac{g_m}{\|g_m\|_1} \right] \end{aligned} \tag{40}$$

The length of both of H and G is mL. The number of multiplications to integrate the weights into the concatenation is mL for each image, which is time-efficient.

5 Evaluations

In this section, we evaluate the proposed weighted linear combinations of image descriptions for loop closure detection. The datasets are shown in Section 5.1. The descriptor extractors, the dissimilarity measure as well as the implementation details are introduced in Section 5.2. The comparison of several linear combinations are presented in

Section 5.3. The weights of the proposed combinations are detailed in Section 5.4.

5.1 Datasets

Three visual SLAM datasets listed below are adopted in this paper. Some sample images from the datasets are shown in Fig. 2.

- The UA Campus dataset includes several image sequences taken in the campus of University of Alberta. The two sequences taken at 16:40 and 22:15 are used and denoted as UA1640 and UA2215, respectively. Strong shadows are presented in UA1640 while the lighting condition of UA2215 is considerably insufficient. Each sequence has 630 images.
- The Nordlandsbanen dataset [37] includes four videos of a long train ride in different seasons. The videos of summer and winter are selected and denoted as SUMMER and WINTER. In each video, 1/500 of the frames are equally sampled for our experiments. All the unrecognizable frames taken in tunnels are manually deleted. Each image sequence results in 1449 images.
- The St. Lucia dataset [38] includes various image sequences taken in several days with severe illumination variations. The two sequences taken in the morning (100909-0845) and in the afternoon (110909-1545) are used and denoted as SL0845 and SL1545, respectively. We uniformly sample 1/4 of the frames in each sequence. The two sequences result in 5283 and 5031 images, respectively.

Fig. 2 Some sample images of the image sequences we use in the experiments

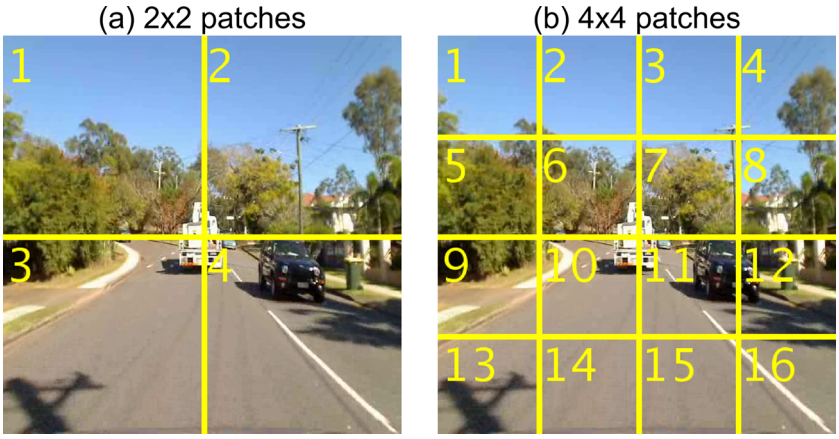

Fig. 3 The regions of support of the local patch descriptors when **a** 2×2 or **b** 4×4 patches are employed. The patch indices are drawn on the images

In each dataset, 10% images are randomly chosen in the training stage and the remaining 90% images are used in the testing stage.

5.2 Experimental Settings

In the testing stage, the recall-precision (RP) curve [39] is employed to evaluate the performance of loop closure detection of a classifier h. In detail, given two image sequences A and B for testing, we calculate the similarity matrix consisting of $h(I, J)$ for each image pair (I, J), $I \in A, J \in B$. The recall-precision curve is plotted by comparing the similarity matrix with the ground truths of loop closures.

Throughout all the experiments, each image is resized to the size of 160×160. It is treated as a single patch ($M = 1$) or partitioned into 2×2 ($M = 4$) or 4×4 ($M = 16$) identical patches. The patterns of the patches are shown in Fig. 3.

Without loss of generality, we adopt five representative descriptor extractors to describe a local image patch, i.e., SURF [1], SIFT [2], GIST [12], local edgel chordiogram (LEC) [40] and VLAD [21]. SIFT and SURF are widely applied local image descriptors which are histograms of gradient orientations. GIST is a popular global scene descriptor which considers image orientations. LEC adopts chordiogram [41], a shape descriptor, to represent the edgels of the image. VLAD concatenates the differences of the

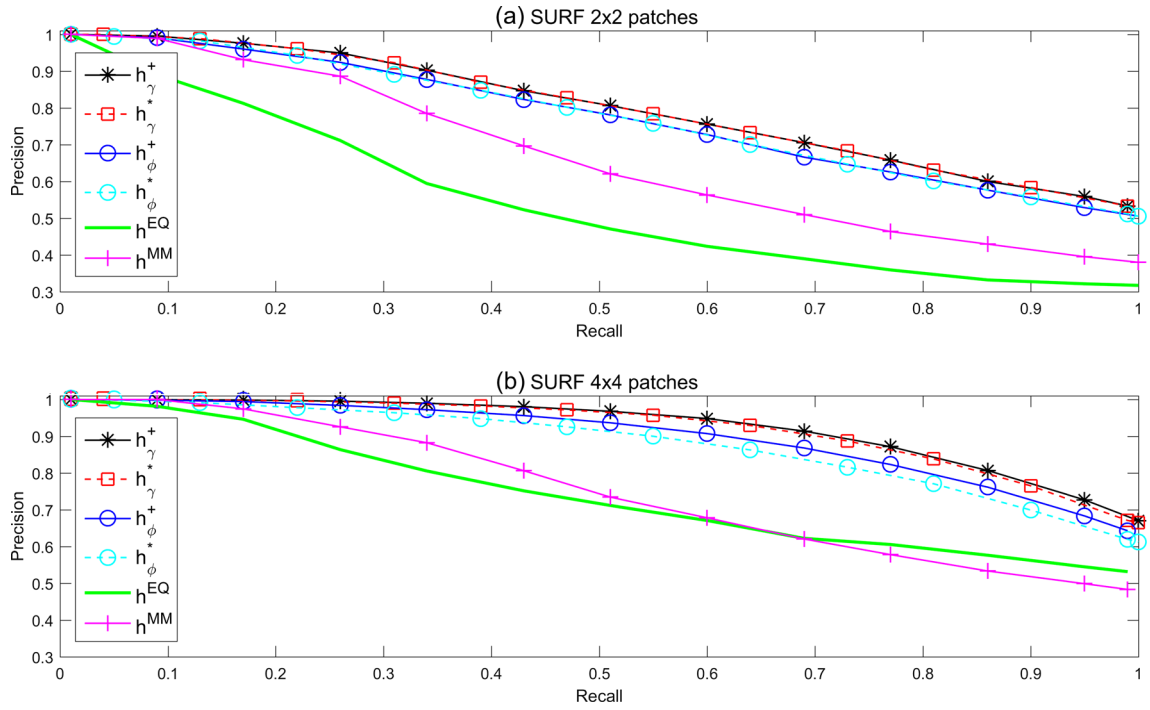

Fig. 4 The performance of loop closure detection using **a** 2×2 or **b** 4×4 grid of SURF on SL0845 vs. SL1545

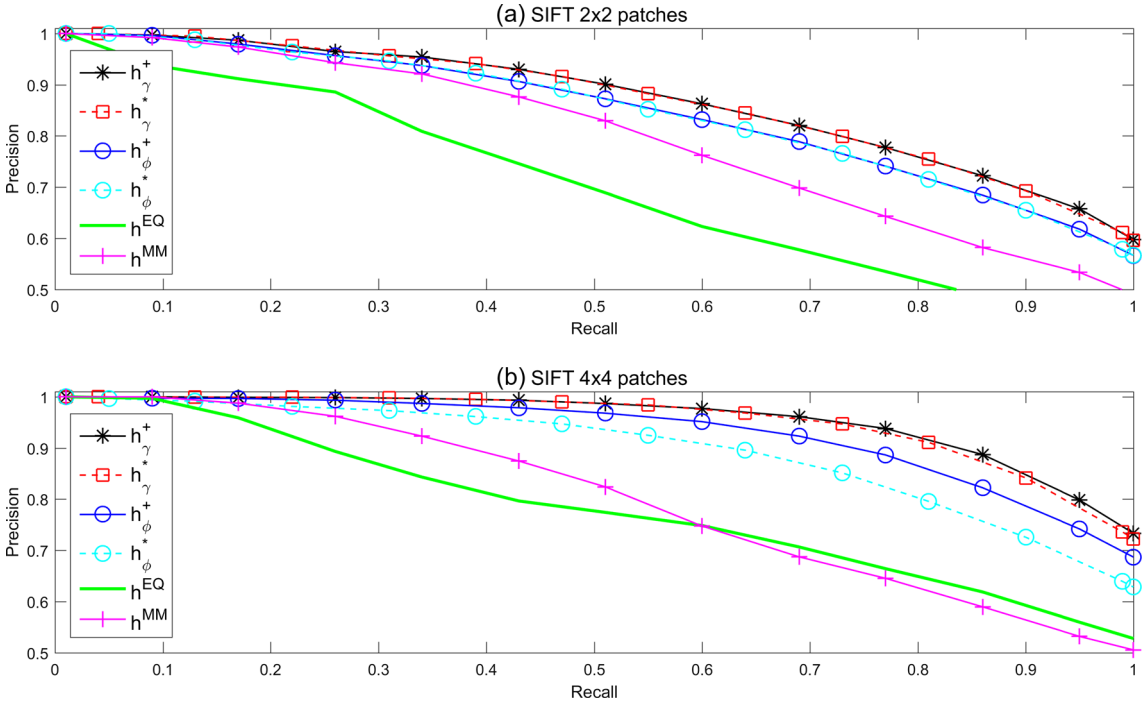

Fig. 5 The performance of loop closure detection using **a** 2 × 2 or **b** 4 × 4 grid of SIFT on SL0845 vs. SL1545

local features in the patch to the centers of the visual words. We use dense SIFT [42] as the local feature of VLAD and train a dictionary of 40 visual words for each dataset.

When a collection of classifiers is obtained, several combinations are considered, including h_γ^* by maximizing the proposed efficacy index γ, h_ϕ^* by maximizing the Fisher criterion ϕ and the corresponding nonnegative combinations h_γ^+ and h_ϕ^+ by Algorithm 1. These combinations are effective only when they outperform the equally weighted combination h^{EQ}. To ensure the convergence of Algorithm 1,

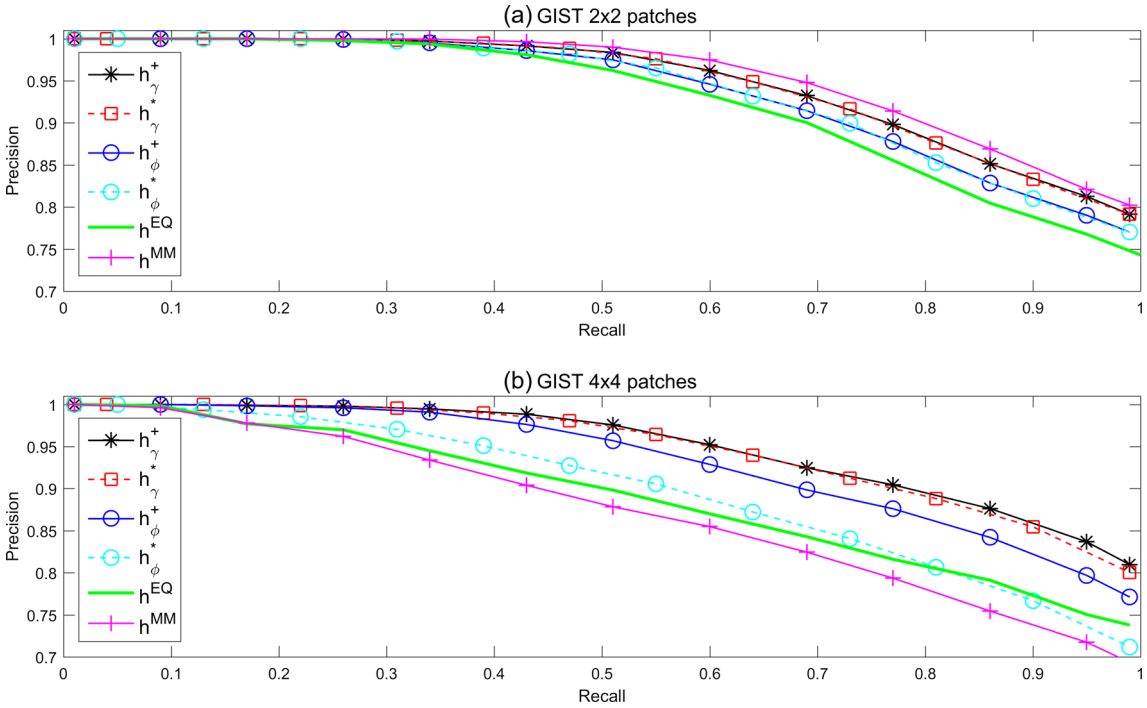

Fig. 6 The performance of loop closure detection using **a** 2 × 2 or **b** 4 × 4 grid of GIST on SL0845 vs. SL1545

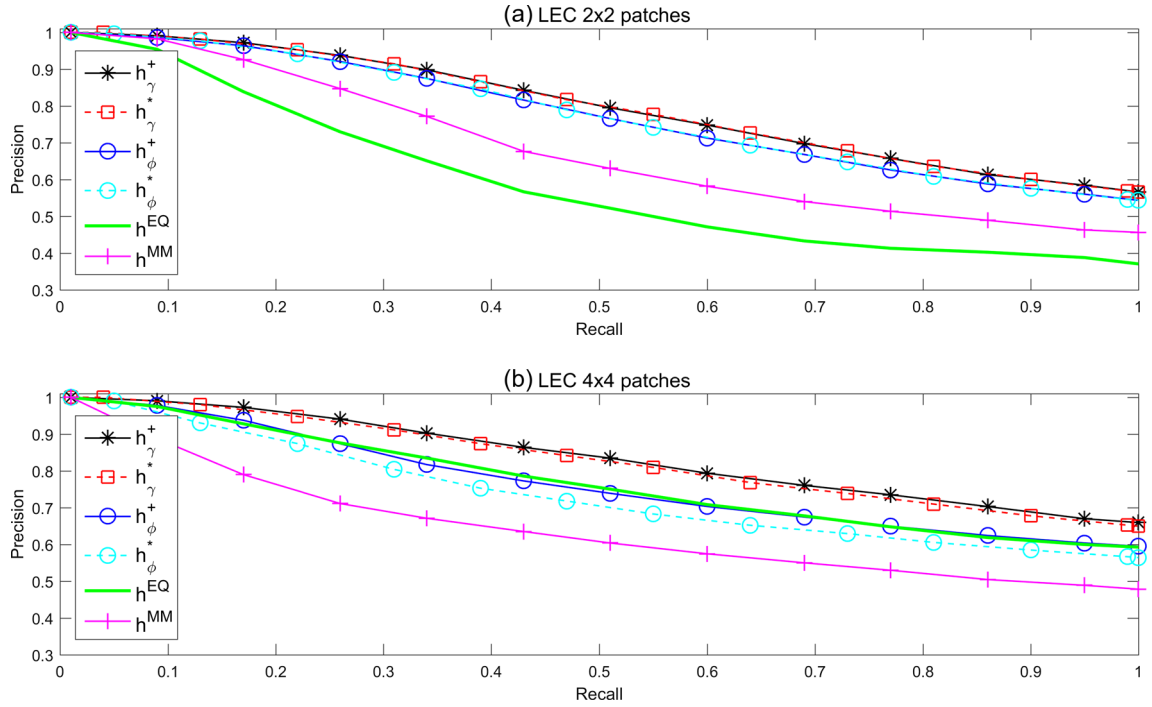

Fig. 7 The performance of loop closure detection using **a** 2 × 2 or **b** 4 × 4 grid of LEC on SL0845 vs. SL1545

we set the number of iterations K to 500 and the step size δ to 0.01. The decoupled weighted combination $h^{DC} = \Omega^T w^{DC}$, which ignores the correlations among classifiers, is compared with h^*_γ as well. Besides, MRMR and the regression solution (26) are utilized for comparison. The resulting classifiers are denoted as h^{MM} and $h^{RG} = \Omega^T w^{RG}$, respectively. As a feature selection method, MRMR requires the number of remaining classifiers, which we set to 75% of the size of the collection of classifiers.

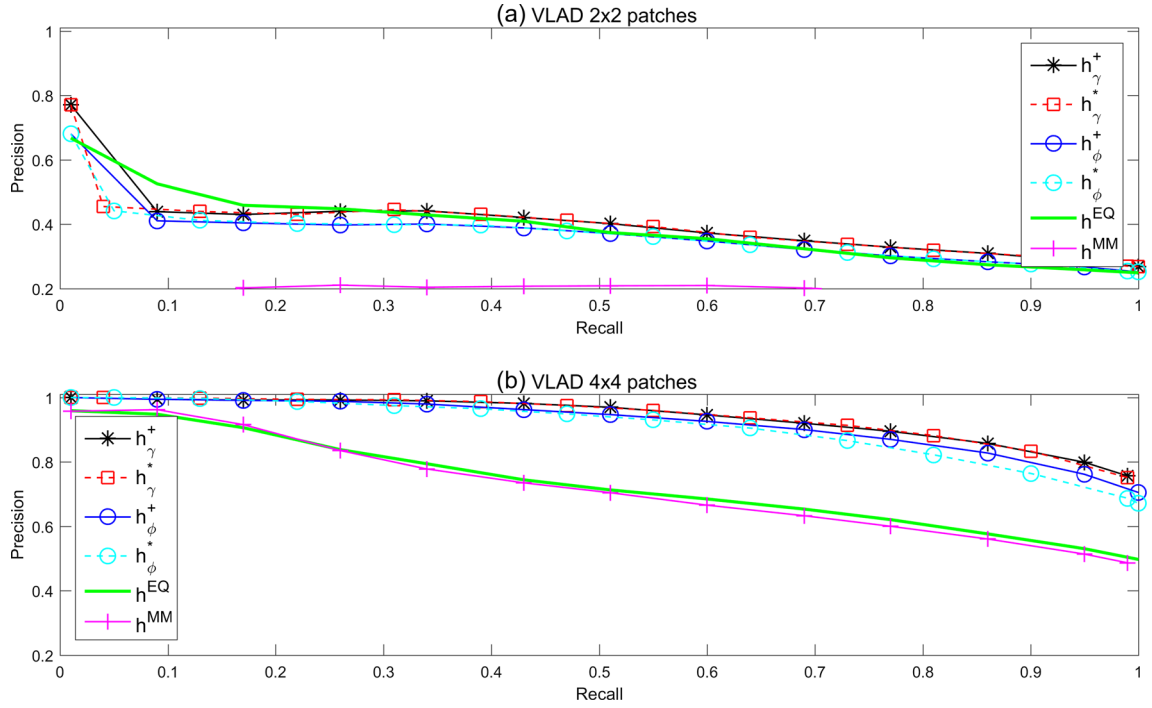

Fig. 8 The performance of loop closure detection using **a** 2 × 2 or **b** 4 × 4 grid of VLAD on SL0845 vs. SL1545

5.3 Comparison of the Combinations

In this section, three group of experiments are presented to evaluate the proposed combinations of image descriptions. As these combinations require a train stage, every test is repeated for 10 times and we only show the average curve of the 10 RP curves.

5.3.1 Single Descriptor

In the first group of the experiments, we utilize each of the five descriptors individually, i.e., only one descriptor is adopted in each local patch. When 2×2 or 4×4 patches are utilized, there are 4 or 16 classifiers in total, respectively. The RP curves of several combinations are

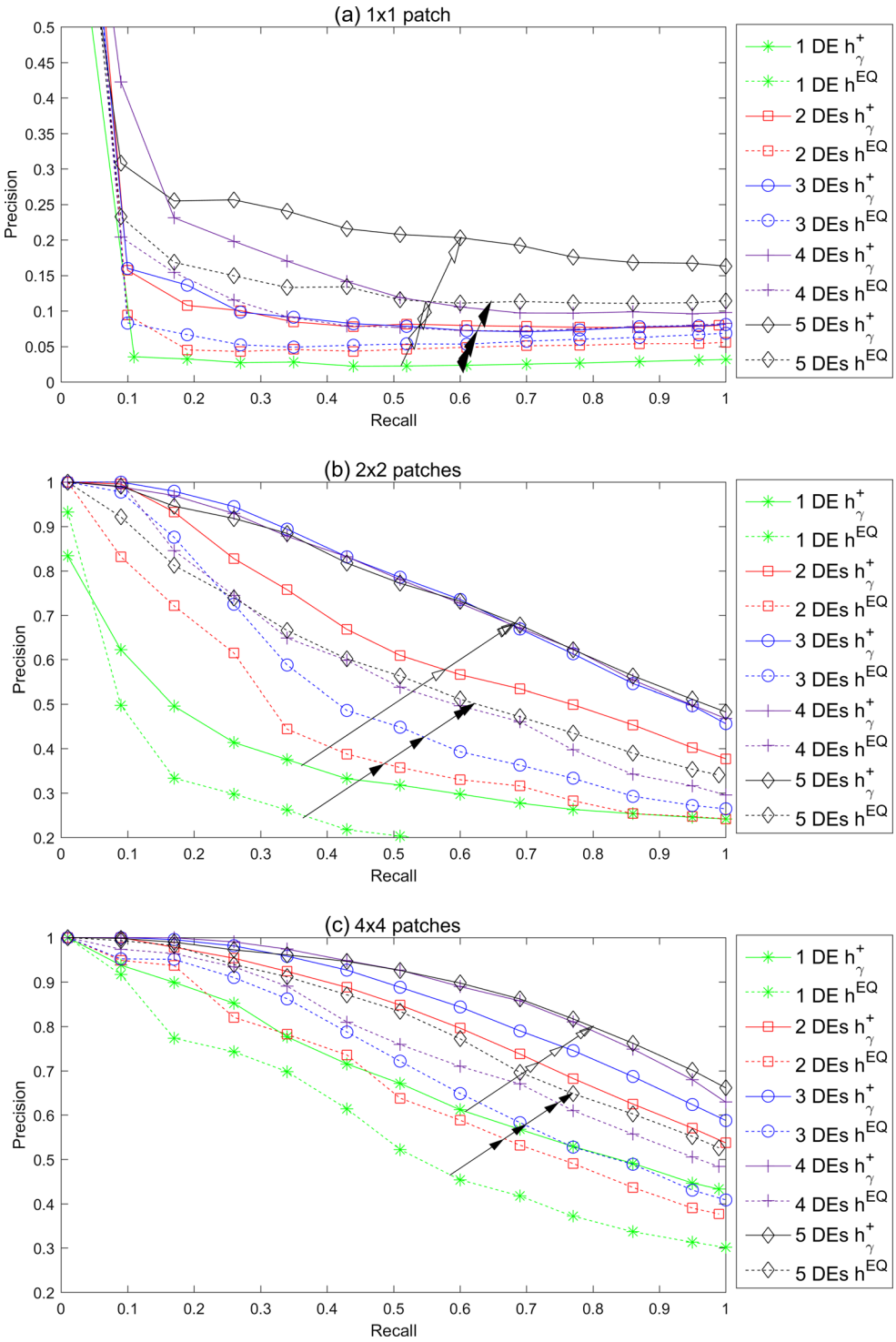

Fig. 9 The RP curves of h^{EQ} and h^+_γ using **a** 1×1, **b** 2×2 or **c** 4×4 patches of descriptions on the SUMMER and WINTER sequences. The descriptor extractors are SURF (1DE), SURF and SIFT (2DEs), SURF, SIFT and GIST (3DEs), SURF, SIFT, GIST and LEC (4DEs), and SURF, SIFT, GIST, LEC and VLAD (5DEs). The five curves of h^{EQ} and h^+_γ are connected by the arrows with white and black heads, respectively

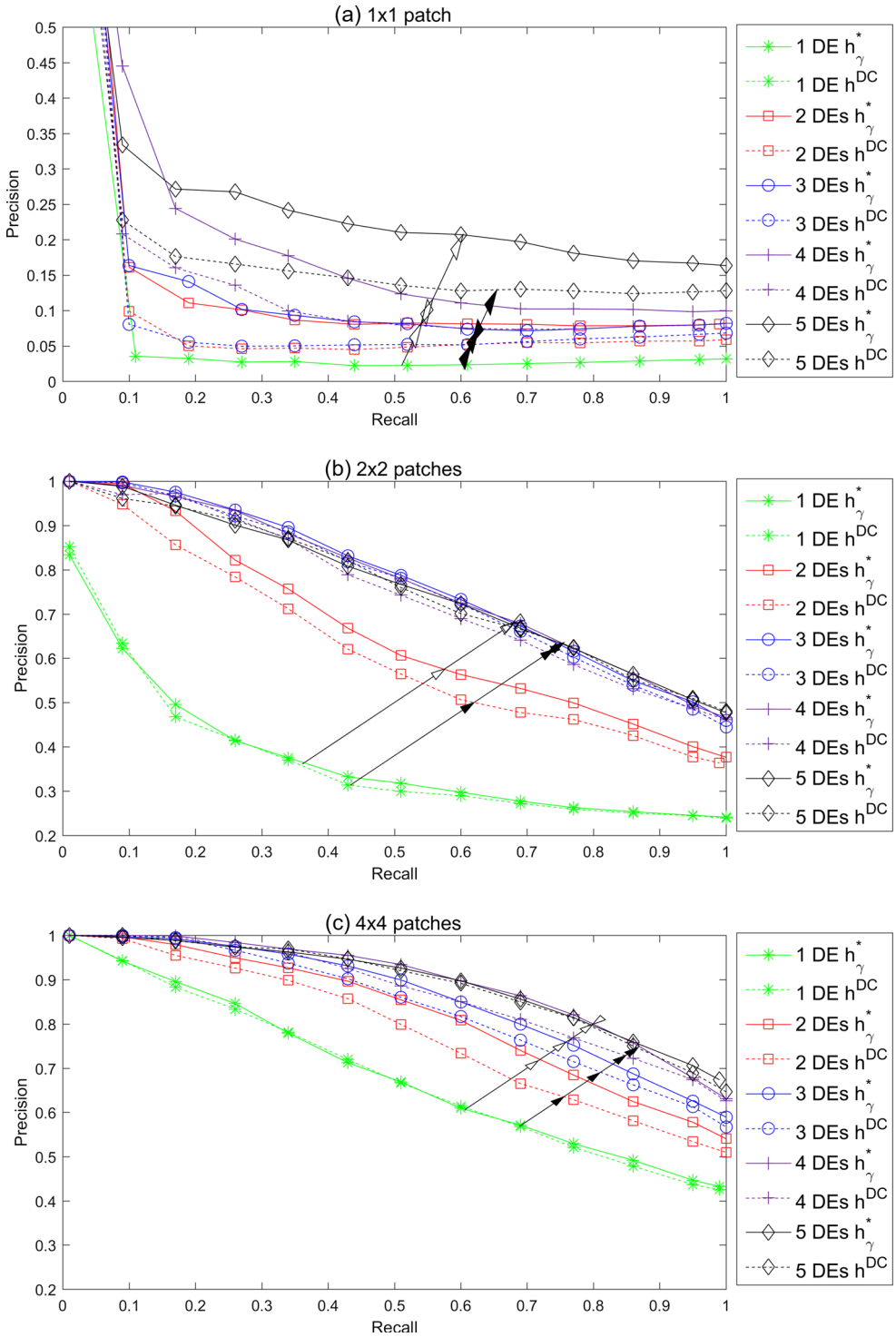

Fig. 10 The RP curves of h^*_γ and h^{DC} on the SUMMER and WINTER sequences. See Fig. 9 for more details

shown in Figs. 4, 5, 6, 7 and 8, from which several common trends are demonstrated. It can be seen that no matter which descriptor is adopted, the performance of each combination using 4×4 patches (Figs. 4–8b) is better than that using 2×2 patches (Figs. 4–8a), respectively, since the number of descriptions of the former is 4 times as many as the latter.

The proposed nonnegative combination h^+_γ outperforms all the other combinations in the most cases. Only when GIST is the descriptor in Fig. 6a, the combination h^{MM} based on MRMR is a slightly better than h^+_γ. However, in the other cases, h^{MM} performs poorer than h^+_γ, h^*_γ, h^+_ϕ and h^*_ϕ, showing that it is not a competitive combination.

In Figs. 4–8, the RP curves of h_γ^+ are almost identical to those of h_γ^*. It happens when the optimal weight vector w_γ^* is almost nonnegative. The similar uniformity of h_ϕ^+ and h_ϕ^* can be observed when 2×2 patches are adopted. On the contrary, a clear improvement of h_ϕ^+ over h_ϕ^* can be seen in Figs. 4–8b when 4×4 patches are adopted. As we have pointed out in Section 3.4, combining more classifiers may increase the occurrence of negative components of w_γ^* and w_ϕ^*. Therefore the improvement of h_γ^+ over h_γ^* and the improvement of h_ϕ^+ over h_ϕ^* are more considerable when 4×4 patches are utilized.

5.3.2 Increasing Descriptors

In the second group of the experiments, starting by using SURF solely, we include more descriptor extractors one by one in the order SURF, SIFT, GIST, LEC and VLAD. When m descriptors in 1×1, 2×2 or 4×4 patches are used, there are m, $4m$ or $16m$ classifiers, respectively. The RP curves of h^{EQ} and h_γ^+ are drawn in Fig. 9. The RP curves of h_γ^* and h^{DC} are plotted in Fig. 10. On one hand, comparing (b) to (a) and (c) to (b) in Figs. 9 and 10, one can see that increasing the number of patches can greatly improve the performance of all the combinations. On the other hand, from each diagrams in Figs. 9 and 10 one can infer that increasing the number of utilized descriptor extractors can also increase the performance of all the combinations.

Figure 9 shows that no matter which collection of descriptors is considered, h_γ^+ outperforms h^{EQ} considerably. When 2×2 or 4×4 patches are used, h_γ^+ of SURF and SIFT outperforms the equally weighted combination of all the five descriptors. In other words, 40% descriptions are able to outperform the concatenation of all descriptions if the proposed nonnegative combination is utilized. Including more descriptors, i.e., GIST, LEC and VLAD, makes the improvement more significant.

Figure 10 shows that the decoupled combination h^{DC} using the weight vector w^{DC} is inferior to h_γ^*. As we have shown in Section 3.3, w^{DC} ignores the correlations among the classifiers. Hence the redundant components may be included in h^{DC} while the informative components may not be properly strengthened.

5.3.3 All Descriptors

In the third group of the experiments, several combinations jointly using all the five descriptors on the three datasets are shown in Figs. 11, 12 and 13. The numbers of classifiers are 20 for 2×2 patches and 80 for 4×4 patches, respectively.

For the UA1640 and UA2215 sequences shown in Fig. 11, when 2×2 patches are used, h_γ^+ achieves the

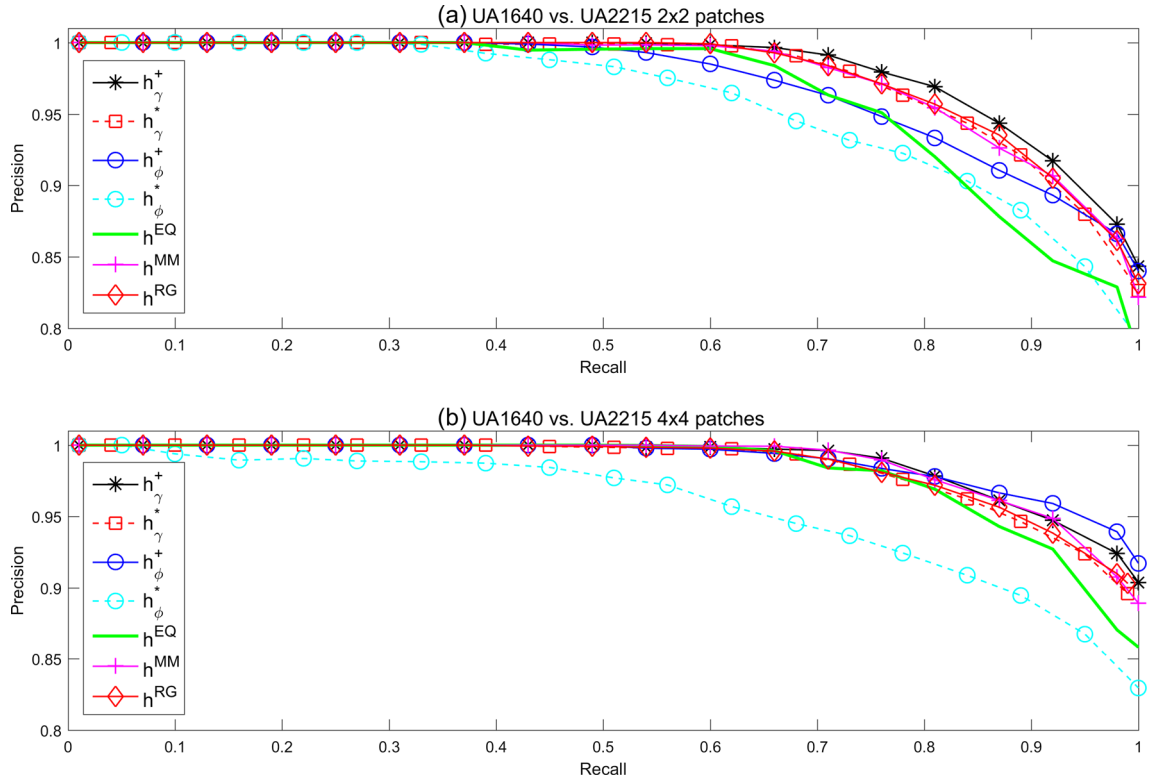

Fig. 11 The performance of loop closure detection using five descriptors jointly on UA1640 vs. UA2215. **a** 2×2 patches or **b** 4×4 patches are utilized

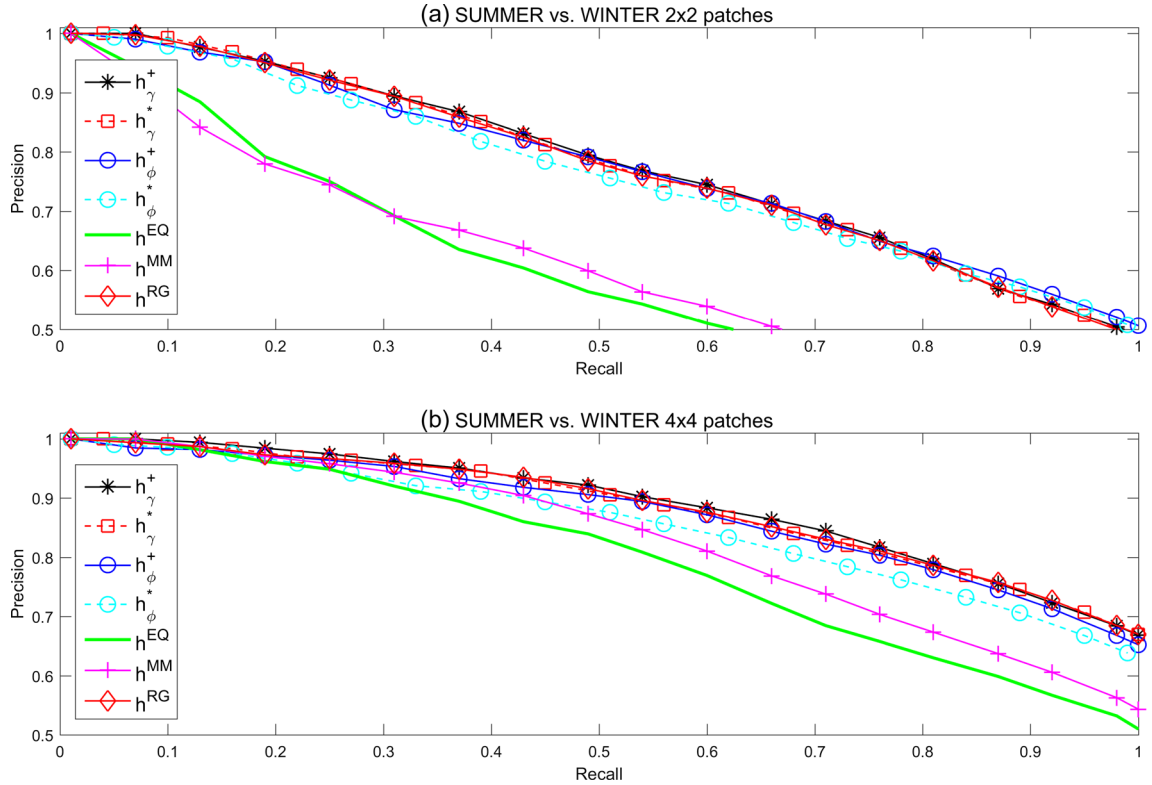

Fig. 12 The performance of loop closure detection using five descriptors jointly on SUMMER vs. WINTER. **a** 2×2 patches or **b** 4×4 patches are utilized

best performance. h_γ^+, h_γ^* and h^{MM} achieve overall better performances than the equally weighted combination h^{EQ}. The optimal combination h_ϕ^* using Fisher criterion as the efficacy index performs poorer than h^{EQ}. Though the nonnegative combination h_ϕ^+ performs better, its precision drops more quickly than h^{EQ} when the recall is around 60%. When 4×4 patches are used, h_γ^+, h_γ^*, h_ϕ^+ and h^{MM} achieve better performance than h^{EQ} and the differences among them are not significant. h_ϕ^* performs significantly worse than h^{EQ}.

For the SUMMER and WINTER sequences shown in Fig. 12, when 2×2 patches are used, h_γ^+, h_γ^*, h_ϕ^+ and h_ϕ^* perform equally well and significantly better than h^{EQ}. The combination by MRMR is not effective since it is similar to h^{EQ}. When 4×4 patches are used, h_γ^+ performs the best. h_γ^* ranks the second, slightly better than h_ϕ^+. The MRMR solution h^{MM} is better than h^{EQ} but worse than h_ϕ^*.

For the SL0845 and SL1545 sequences shown in Fig. 13, h_γ^+ always performs the best while h^{MM} performs the poorest. When 4×4 patches are used, the improvement of h_γ^+ over h_γ^* and the improvement of h_ϕ^+ over h_ϕ^* are remarkable. The precisions of h_γ^* and h_ϕ^* cannot reach 100% when the recalls are small. It shows that some false matches have very small weighted distances and therefore exert negative impact on the precision of loop closure detection. In contrast, the nonnegative solutions h_γ^+ and h_ϕ^+ greatly improve the performance.

The experiments on the three datasets show that the performance of the regression solution h^{RG} is close to h_γ^*, as we have analyzed in Section 3.3.

5.3.4 Summary and Discussion

From the comparison we conclude that:

1. Integrating more image descriptions, i.e., using more descriptor extractors in more local patches, can increase the performance of loop closure detection.
2. The proposed nonnegative combination h_γ^+ achieves the best performance. In particular, it outperforms the equally weighted combination h^{EQ} significantly and performs better than the decoupled weighted combination h^{DC}.
3. The nonnegative solution h_γ^+ outperforms h_γ^* and the nonnegative solution h_ϕ^+ outperforms h_ϕ^*. Therefore the nonnegative combinations obtained by Algorithm 1 can significantly improve the performance of loop closure detection.
4. h_γ^* outperforms h_ϕ^* and h_γ^+ outperforms h_ϕ^+ considerably. As the only difference between the proposed index γ and the Fisher criterion ϕ is that we omit the

variance of the true matches. The improvement justifies our definition.

5. The combination h^{RG} obtained by linear regression performs almost identical to h^*_γ.
6. The concatenation h^{MM} of the 75% classifiers selected by MRMR is not better than the equally weighted concatenation h^{EQ} using all classifiers. Though other proportions may result better performance, more dedicated algorithms are required to decide the optimal proportions. On the contrary, the proposed combination h^+_γ has no decisive parameters and achieves much better performance.

The experiments demonstrate that the RP curves of the nonnegative combination h^+_γ outperform the curves of the combination h^*_γ. As $\gamma[h^+_\gamma] \leq \gamma[h^*_\gamma]$, this observation shows that the proposed efficacy index γ is not in perfect accord with the recall-precision curve. γ measures the separation between the whole distributions of the distances of the true matches and false matches. The combinations with large separations or γ values may lead to the false matches with rather low weighted distances, resulting in inferior RP curves with low precisions. In contrast, the nonnegative combination h^+_γ trades off the separation for wiping out those false matches with low distances.

5.4 The Weights

In this section, we detail three groups of the weights w^*_γ, w^+_γ, w^{DC} and w^{DC} when the adopted five descriptors are utilized individually or jointly. The results using 2×2 and 4×4 patches on SL0845 and SL1545 sequences are shown in Figs. 14 and 15, respectively. The results using 4×4 patches on SUMMER and WINTER sequences are shown in Fig. 16. When 2×2 or 4×4 patches are used, there are 4 or 16 classifiers when the descriptors are used individually, shown in the left column of the three figures. When the descriptors are used jointly, there are 20 or 80 classifiers for 2×2 or 4×4 patches. The much longer weight vectors are split into five components and shown in the right column of the three figures. The weight vectors are normalized suitably for demonstration purposes.

When 2×2 patches are used, Fig. 14a1-e1 show that no matter which descriptor is adopted to represent the 4 patches, the weighting components of w^*_γ and w^{DC} reveal similar patterns. The weights of the top left and top right patches are larger than the weights of the bottom left and bottom right patches. This is because the upper half of the image contains most of the informative objects of the scene while the lower half of the image contains mainly roads and grass, which are less informative. When the five descriptors are combined, as shown in Fig. 14a2-e2, the

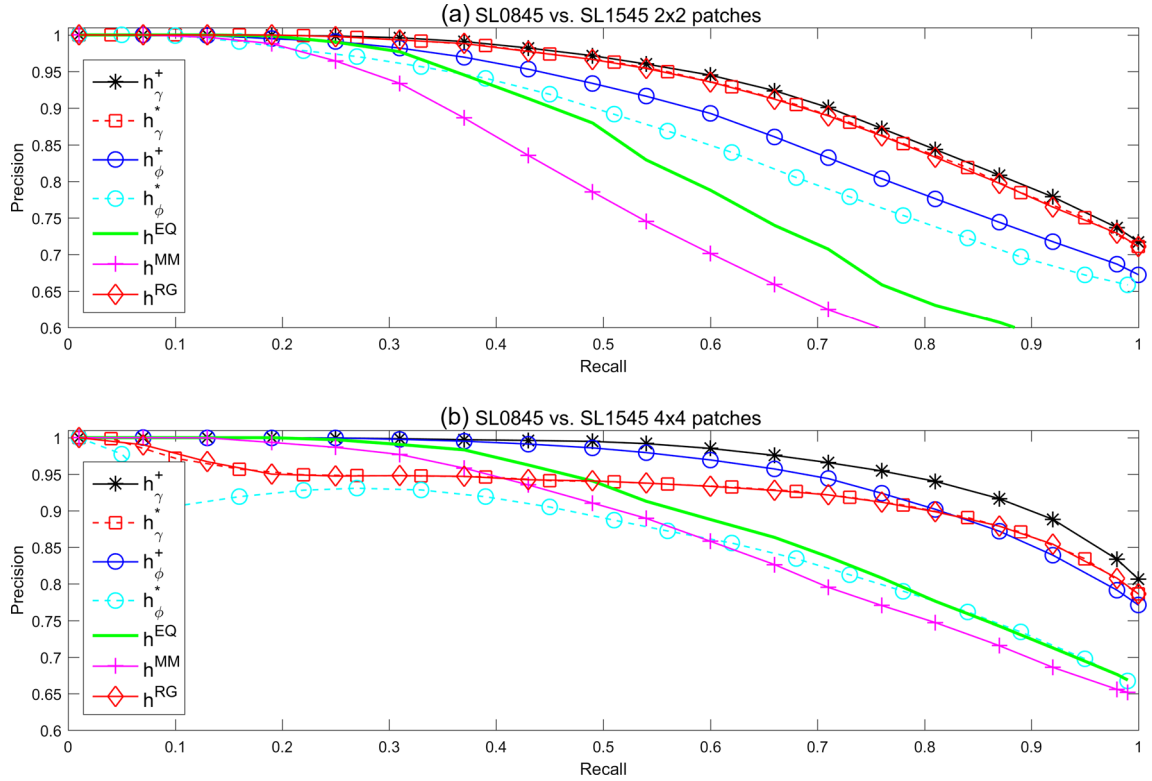

Fig. 13 The performance of loop closure detection using five descriptors jointly on SL0845 vs. SL1545. **a** 2×2 patches or **b** 4×4 patches are utilized

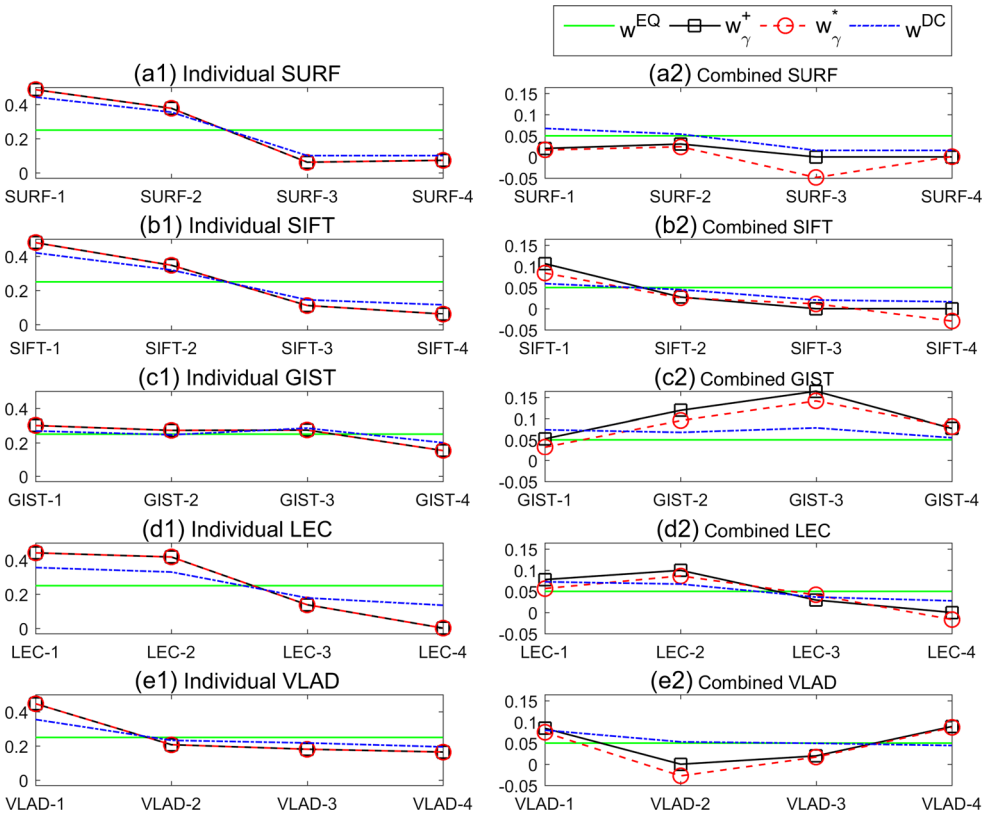

Fig. 14 Various weight vectors on SL0845 vs. SL1545 when 2×2 patches are used. The five descriptors are utilized individually (**a1-e1**) or jointly (**a2-e2**). The indices of the patches are attached to the name of the descriptors

decoupled weight vector w^{DC} is exactly the concatenation of the decoupled weights in Fig. 14a1-e1. In contrast, the weights of the five descriptors in w_γ^* and w_γ^+ are dramatically different. For instance, the top left patch is mainly represented by SIFT (SIFT-1) while the bottom left patch is mainly represented by GIST (GIST-3). The four classifiers by SURF are weakened, whereas the classifiers by GIST are strengthened. SURF is suppressed by SIFT because SIFT is highly correlated with SURF and performs better than SURF. Since GIST (GIST-3) is the most informative descriptor representing the bottom-left patch, it is strengthened sharply. As a result, w_γ^* and w_γ^+ are capable of strengthening the informative classifiers and weakening the redundant or inferior classifiers. The difference between w_γ^* and w_γ^+ is limited.

When each descriptor is used in the 4×4 patches individually, the classifiers of the 6th and 7th patches covering the central area of the image are strengthened for SL0845 and SL1545 in Fig. 15a1-e1, while the classifiers of the top eight patches covering the upper half of the image are emphasized mostly for SUMMER and WINTER in Fig. 16a1-e1. Therefore the informative areas of the images depend on the datasets since different descriptors tend to strengthen the classifiers of the same informative patches.

When the five descriptors in the 4×4 patches are combined, w_γ^+ and w_γ^* are significantly different. Among the 80 components of w_γ^+, only 23 components are nonzero for SL0845 and SL1545 and 41 components are nonzero for SUMMER and WINTER, as shown in Figs. 15a2-e2 and 16a2-e2, respectively. It shows that Algorithm 1 serves as a feature selection and weighting process. For SL0845 and SL1545 shown in Fig. 15a2-e2, the combination h_γ^+ mainly consists of the classifiers using VLAD, while for SUMMER and WINTER shown in Fig. 16a2-e2, h_γ^+ mainly consists of the classifiers covering the central area of the image by SIFT and VLAD and the classifiers covering the top area of the image by GIST and LEC.

In Fig. 17, we plot an example of the Jacobian matrix of $R(w)$ at w_γ^+ as well as the corresponding weight vector w_γ^+, which is the repetition of the one shown in Fig. 16a2-e2. One can see that the components of $J(w_\gamma^+)$ are zero if the corresponding components of w_γ^+ are positive and the components of w_γ^+ are zero if the corresponding components of $J(w_\gamma^+)$ are negative. Therefore the statements of Theorem 1 are verified and w_γ^+ indeed maximizes $R(w)$ locally.

In summary, the patterns of the weight vectors w^{DC}, w_γ^* and w_γ^+ depend on the datasets as well as the descriptors. w^{DC} is merely the concatenation of the individual weights of the classifiers. Informative classifiers are strengthened by w_γ^* and w_γ^*+ while redundant and uninformative classifiers are weakened. When the number of classifiers are limited, the optimal weight vector w_γ^* is likely to be nonnegative and hence the nonnegative weight vector w_γ^+ approximates

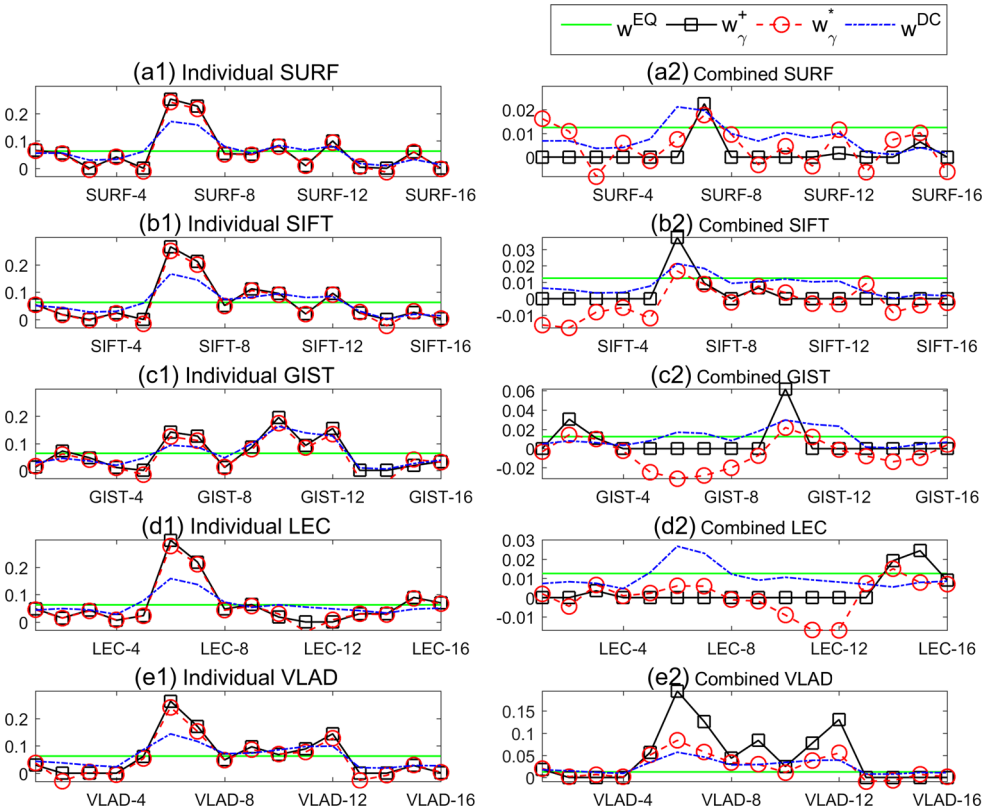

Fig. 15 Various weight vectors on SL0845 vs. SL1545 when 4×4 patches are used. The five descriptors are utilized individually (**a1-e1**) or jointly (**a2-e2**)

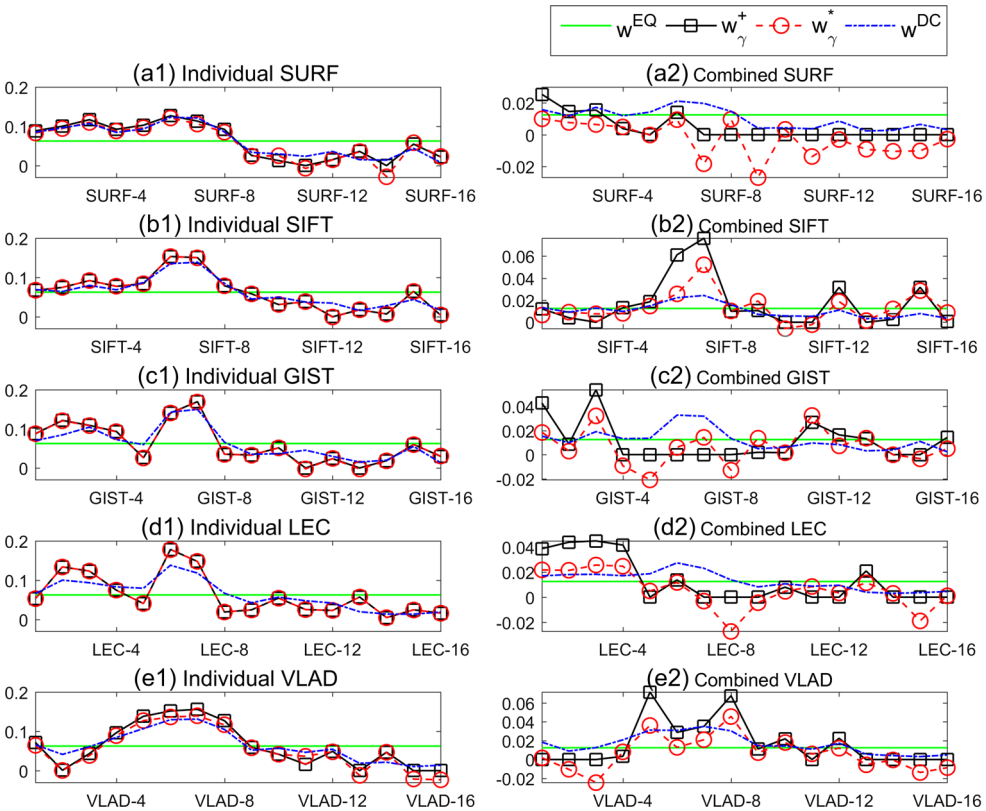

Fig. 16 Various weight vectors on SUMMER vs. WINTER when 4×4 patches are used. The five descriptors are utilized individually (**a1-e1**) or jointly (**a2-e2**)

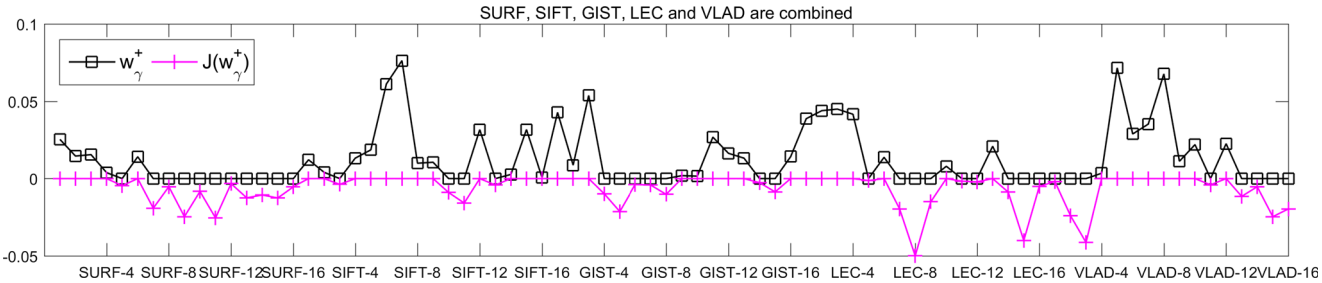

Fig. 17 The weight vector w_γ^+ and the Jacobian matrix $J(w_\gamma^+)$ on SUMMER vs. WINTER when 4×4 patches of the five descriptors are utilized jointly

to w_γ^*. As a result, h_γ^* and w_γ^+ achieve similar performance. When the number of classifiers are large, some components of w_γ^* are likely to be negative. The nonnegative weight vector w_γ^+ may include many zero components, acting as a feature selection process. The results in Section 5.3.3 show that the improvement of h_γ^+ over h_γ^* becomes significant.

6 Conclusion

The goal of this paper was to build a framework to evaluate and jointly optimize the combinations of image descriptions for visual loop closure detection. To tackle the problem, we described the visual LCD as a binary classification problem and converted the image descriptions into classifiers. A real valued efficacy index was proposed to evaluate the weighted combinations of image descriptions. As the index is a differentiable function of the weights, various optimization strategies can be carries out conveniently.

Two weighted combinations of image descriptions were proposed. The first one maximizes the efficacy index. We have shown analytically and experimentally that it approximates to the solution of linear regression. The weights of the second one are nonnegative and produced by a gradient descent algorithm. We proved that the nonnegative combination is locally optimal. Experiments have show that the two weighted combinations outperform the equally weighted combination significantly since informative image descriptions are emphasized while redundant image descriptions are de-emphasized.

The proposed efficacy index is similar to the Fisher criterion. It excludes the unreliable covariance matrix of true matches, which is adopted by the Fisher criterion. Experiments have shown that the proposed efficacy index works considerably better than the Fisher criterion for loop closure detection.

This work will be extended further in three directions. Firstly, as the proposed evaluation framework is not limited to loop closure detection, we will utilize it to develop combination methods to solve the other image comparison problems. Secondly, as the proposed framework is a convenient tool to evaluate more combinations of image descriptions for loop closure detection, we will apply it to deep features as well as features of point clouds. Thirdly, we will explore other forms of efficacy indices.

Appendix A: Addition of Two Classifiers

Here we prove that for two classifiers a and b, if $\gamma[a] < \gamma[b]$, then $\gamma[a] < \gamma[a+b]$, which we have mentioned in Section 3.2. The proof is also applicable for the Fisher criterion ϕ with minor modifications.

According to the definition (7), $\gamma[a] < \gamma[b]$ is equivalent to

$$L_1 < \frac{E_F[b] - E_T[b]}{E_F[a] - E_T[a]} \tag{41}$$

where

$$L_1 = \sqrt{\frac{\mathrm{Var}_F[b]}{\mathrm{Var}_F[a]}}. \tag{42}$$

$\gamma[a] < \gamma[a+b]$ is equivalent to

$$L_2 < 1 + \frac{E_F[b] - E_T[b]}{E_F[a] - E_T[a]} \tag{43}$$

where

$$L_2 = \sqrt{\frac{\mathrm{Var}_F[a+b]}{\mathrm{Var}_F[a]}}. \tag{44}$$

Squaring (44) and expanding it we have

$$L_2^2 = 1 + L_1^2 + 2\frac{\mathrm{Cov}_F[a,b]}{\mathrm{Var}_F[a]}. \tag{45}$$

Since $\mathrm{Cov}_F[a,b] \le \sqrt{\mathrm{Var}_F[a]\mathrm{Var}_F[b]}$ always holds,

$$\frac{\mathrm{Cov}_F[a,b]}{\mathrm{Var}_F[a]} \le L_1 \tag{46}$$

holds too. Thus $L_2 < 1 + L_1$ always holds. Therefore (41) implies (43), which proves the assertion.

Appendix B: The Proof of Theorem 1

Proof By suitable reordering the variables, we can partition the indices of a vector of length m into $u = \{1, \dots r\}$ and $v = \{r+1, \dots m\}$, such that w^+ can be expressed as $w^+ = [w_1^+, \dots, w_m^+]^T = [w_u^{+T}, w_v^{+T}]^T$ where $w_u^+ > 0$ and $w_v^+ = 0$. Similarly, we denote $J(w^+)$ as $J = [j_1, \dots, j_m]^T = [J_u^T, J_v^T]^T$. Thus we have

$$w^+ + \delta J = [w_u^{+T} + \delta J_u^T, w_v^{+T} + \delta J_v^T]^T \tag{47}$$

We claim that $J_v \leq 0$. Otherwise if $j_i > 0$ for some $i \in v$, the algorithm will update w_i^+ to a positive value, contradicting to the assumption of convergence. Then, since $w_v^+ = 0$, $w_v^+ + \delta J_v$, the second subvector of Eq. 47, is nonpositive. The first subvector $w_u^+ + \delta J_u$ is nonnegative if a small enough δ is chosen. Indeed, for each $i \in u$ two situations exist: (1) $w_i^+ > 0$ and $j_i \geq 0$; (2) $w_i^+ > 0$ and $j_i < 0$. In the first situation the non-negativity of $w_i^+ + \delta j_i$ is ensured. In the second situation, if we choose $\delta \leq -w_i^+ / j_i$, the non-negativity is also guaranteed.

According to Algorithm 1, setting the negative components of Eq. 47 to 0 and normalizing the result, we have

$$w^+ = [w_u^{+T}, w_v^{+T}]^T = \left[\frac{(w_u^+ + \delta J_u)^T}{\|w_u^+ + \delta J_u\|}, 0^T \right]^T \tag{48}$$

Thus w_u^+ parallels to $w_u^+ + \delta J_u$, which implies that J_u parallels to w_u^+ or $J_u = 0$. The former case is impossible. Otherwise we can assume that $\lambda J_u = w_u^+$ for some value λ. Since

$$s = [s_u^T, s_v^T]^T, C_F = \begin{bmatrix} C_{uu} & C_{uv} \\ C_{vu} & C_{vv} \end{bmatrix} \tag{49}$$

considering that $w_v^+ = 0$ and using (20), we have

$$J_u = \eta[(w_u^{+T} C_{uu} w_u^+) s_u - (s_u^T w_u^+) C_{uu} w_u^+] \tag{50}$$

where η is a nonzero normalization factor. Thus

$$\begin{aligned} \|w_u^+\|^2 &= w_u^{+T} w_u^+ = \lambda w_u^{+T} J_u \\ &= \lambda \eta w_u^{+T} [(w_u^{+T} C_{uu} w_u^+) s_u - (s_u^T w_u^+) C_{uu} w_u^+] = 0, \end{aligned}$$

contradicting to the assumption that w^+ is a nonzero vector. Therefore we conclude that $J_u = 0$ and $J_v \leq 0$ at w^+. An example is shown in Section 5.4. As these conditions are exactly the Karush-Kuhn-Tucker conditions [43] of Eqs. 33, 33 attains a local maximum at w^+. □

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J Comput. Vis. **60**(2), 91–110 (2004)
3. Sivic, J., Zisserman, A., et al.: Video google: A text retrieval approach to object matching in videos. Iccv **2**, 1470–1477 (2003)
4. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Twentieth Annual Symposium on Computational Geometry, pp. 253–262. ACM (2004)
5. Angeli, A., Filliat, D., Doncieux, S., Meyer, J.-A.: Fast and incremental method for loop-closure detection using bags of visual words. IEEE Trans. Robot. **24**(5), 1027–1037 (2008)
6. Shahbazi, H., Zhang, H.: Application of locality sensitive hashing to realtime loop closure detection. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1228–1233. IEEE (2011)
7. Cummins, M., Newman, P.: Fab-map: Probabilistic localization and mapping in the space of appearance. Int. J. Robot. Res. **27**(6), 647–665 (2008)
8. Milford, M.J., Wyeth, G.F.: Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 1643–1649. IEEE (2012)
9. Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: European Conference on Computer Vision, pp. 834–849. Springer (2014)
10. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: A versatile and accurate monocular slam system. IEEE Trans. Robot. **31**(5), 1147–1163 (2015)
11. Valgren, C., Lilienthal, A.J.: Sift, surf and seasons: Long-term outdoor localization using local features. In: European Conference on Mobile Robots (ECMR), pp. 253–258 (2007)
12. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
13. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: Brief: Computing a local binary descriptor very fast. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1281–1298 (2012)
14. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
15. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 IEEE International conference on computer vision (ICCV), pp. 2564–2571. IEEE (2011)
16. Liu, Y., Zhang, H.: Visual loop closure detection with a compact image descriptor. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1051–1056. IEEE (2012)
17. Sünderhauf, N., Protzel, P.: Brief-gist: Closing the loop by simple means. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1234–1241. IEEE (2011)
18. Campos, F.M., Correia, L., Calado, J.M.F.: Loop closure detection with a holistic image feature. In: Portuguese Conference on Artificial Intelligence, pp. 247–258. Springer (2013)
19. Arroyo, R., Alcantarilla, P.F., Bergasa, L.M., Javier Yebes, J., Gámez, S.: Bidirectional loop closure detection on panoramas for visual navigation. In: 2014 Intelligent Vehicles Symposium Proceedings IEEE, pp. 1378–1383. IEEE (2014)
20. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. Comput Vis. Pattern Recogn. 1–8 (2007)
21. Jegou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. Comput. Vis. Pattern Recog.,3304–3311 (2010)
22. Arandjelovic, R., Zisserman, A.: All about vlad. Comput. Vis. Pattern Recog., 1578–1585 (2013)
23. Yi, H., Zhang, H., Zhou, S.: Convolutional neural network-based image representation for visual loop closure detection. In: 2015 IEEE International Conference on Information and Automation, pp. 2238–2245. IEEE (2015)

24. Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4297–4304. IEEE (2015)
25. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
26. Campos, F.M., Correia, L., Calado, J.M.F.: Robot visual localization through local feature fusion: An evaluation of multiple classifiers combination approaches. J. Intell. Robot. Syst. **77**(2), 377–390 (2015)
27. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
28. Levner, I., Zhang, H.: Classification-driven watershed segmentation. IEEE Trans. Image Process. **16**(5), 1437–1445 (2007)
29. Li, W., Mao, K., Zhang, H., Chai, T.: Selection of gabor filters for improved texture feature extraction. In: 2010 17th IEEE International conference on Image Processing (ICIP), pp. 361–364. IEEE (2010)
30. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
31. Quanquan, G., Li, Z., Han, J.: Generalized fisher score for feature selection. arXiv:1202.3725 (2012)
32. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics **7**(2), 179–188 (1936)
33. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley (2012)
34. Li, Q., Ke, L., You, X., Shuhui, B., Liu, Z.: Place recognition based on deep feature and adaptive weighting of similarity matrix. Neurocomputing **199**, 114–127 (2016)
35. Tola, E., Lepetit, V., Fua, P.: Daisy: An efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 815–830 (2010)
36. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE (2006)
37. The Norwegian Broadcasting Corporation: The Nordlandsbanen Dataset. http://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/ (2013)
38. Glover, A.J., Maddern, W.P., Milford, M.J., Wyeth, G.F.: Fab-map+ ratslam: Appearance-based slam for multiple times of day. In: 2010 IEEE International Conference on Robotics and Automation (ICRA), pp. 3507–3512. IEEE (2010)
39. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240. ACM (2006)
40. Wang, X., Zhang, H., Peng, G.: A chordiogram image descriptor using local edgels. J. Vis. Commun. Represent. **49**, 129–140 (2017)
41. Toshev, A., Taskar, B., Daniilidis, K.: Shape-based object detection via boundary structure segmentation. Int. J. Comput. Vis. **99**(2), 123–146 (2012)
42. Ce, L., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 978–994 (2011)
43. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)

Xiaolong Wang received his B.Sc. degree in Applied Mathematics and M.Sc. degree in Computational Mathematics, both from Northwestern Polytechnical University (NWPU), P.R. China. He is currently pursuing his Ph.D. in Applied Mathematics in NWPU under the supervision of Guohua Peng. His current research is focusing on appearance-based place recognition and visual-SLAM system.

Guohua Peng received his Ph.D. degree in Applied Mathematics from Northwestern Polytechnical University, P.R. China. Currently, he is a professor in the School of Natural and Applied Sciences of NWPU. His major research interests are CAGD, computer graphics, and image processing.

Hong Zhang received his B.Sc. degree from Northeastern University, Boston, USA, and his Ph.D. degree from Purdue University, West Lafayette (IN), USA, both in Electrical and Computer Engineering. He is currently a Professor in the Department of Computing Science, University of Alberta. He is a Fellow of IEEE and a Fellow of Canadian Academy of Engineering. His current research interests include robotics, computer vision and image processing.