IEICE TRANS. INF. & SYST., VOL.E101–D, NO.12 DECEMBER 2018

PAPER

Local Feature Reliability Measure Consistent with Match Conditions for Mobile Visual Search

Kohei MATSUZAKI[†a)], **Kazuyuki TASAKA**[†b)], ***and*** **Hiromasa YANAGIHARA**[†c)], ***Members***

SUMMARY We propose a feature design method for a mobile visual search based on binary features and a bag-of-visual words framework. In mobile visual search, detection error and quantization error are unavoidable due to viewpoint changes and cause performance degradation. Typical approaches to visual search extract features from a single view of reference images, though such features are insufficient to manage detection and quantization errors. In this paper, we extract features from multiview synthetic images. These features are selected according to our novel reliability measure which enables robust recognition against various viewpoint changes. We regard feature selection as a maximum coverage problem. That is, we find a finite set of features maximizing an objective function under certain constraints. As this problem is NP-hard and thus computationally infeasible, we explore approximate solutions based on a greedy algorithm. For this purpose, we propose novel constraint functions which are designed to be consistent with the match conditions in the visual search method. Experiments show that the proposed method improves retrieval accuracy by 12.7 percentage points without increasing the database size or changing the search procedure. In other words, the proposed method enables more accurate search without adversely affecting the database size, computational cost, and memory requirement.

key words: *mobile visual search, binary feature, feature selection, maximum coverage problem*

1. Introduction

With the advancement of mobile devices equipped with high-resolution cameras, mobile visual search (MVS) has become one of the major applications of image retrieval and recognition technology. These applications can be used for recognizing specific objects captured by the mobile device's camera from a pre-constructed database, e.g., retail catalog, real consumer products, and so forth. While there are some studies on MVS as a server-client system, this paper focuses on a stand-alone MVS system [1], [2] which will even work without a network connection. We assume that the stand-alone MVS system is implemented as a mobile application containing the database and installed in mobile devices. In such an application, smaller application size is preferred because users are disinclined to install large-data applications. The most straightforward way to achieve a compact application is to limit the database size because it is usually difficult to reduce significantly the data size of other components of the application (e.g., search function). In this paper, we aim to maximize the retrieval accuracy under the limitation of database size. However, it is challenging to achieve both high accuracy and small database size.

Local features have been successfully employed in many computer vision tasks such as image retrieval and recognition [3], [35], object detection [5], and image-based localization [6]. In a typical MVS framework, local features are also employed as the first step in both search and indexing. The most widely known local features are SIFT [7] and SURF [8]. Their keypoint detector and feature descriptor are invariant to scaling, rotation, illumination variation, and noise. However, they incur a high computational cost for real-time applications running on mobile devices with low computing power and memory capacity. Apart from real-valued features like SIFT, binary local features such as BRISK [9], ORB [10], and FREAK [11] have attracted much attention due to their efficiency. The performance of these binary features is comparable to SIFT and SURF while being one or two orders of magnitude faster. Therefore, binary features are better suited to facilitating the incorporation of visual search applications into mobile devices.

In many image retrieval and recognition frameworks, local features are encoded into an image representation, such as bag-of-visual words (BoVW) [3], vector of locally aggregated descriptors (VLAD) [4], and Fisher vector (FV) [12], [13]. Similarity scores between a query image captured by a mobile device and reference images in the database are calculated by comparing their representations. Finally, geometric verification is performed using reference images with top similarities. VLAD and FV achieve both small database size and high accuracy by aggregating local descriptors into a compact image representation. However, they cannot achieve feature-level matching although they realize image-level matching since they aggregate all local features from an image. Therefore, they require additional data for feature-level matching if we are to apply them to a visual search framework including geometric verification. On the other hand, BoVW needs less data as a result since it can perform feature-level matching based on visual word matching.

Therefore, we focus on an MVS system based on binary local features and a BoVW framework. BoVW quantizes feature descriptors into representative vectors called visual words (VWs) with a visual codebook and represents an image as a histogram of VWs. Then, image similarity is calculated as a cosine similarity between histograms of the

Manuscript received March 23, 2018.
Manuscript revised August 10, 2018.
Manuscript publicized September 12, 2018.

[†]The authors are with the KDDI Research, Inc., Fujimino-shi, 356–8502 Japan.
a) E-mail: ko-matsuzaki@kddi-research.jp
b) E-mail: ka-tasaka@kddi-research.jp
c) E-mail: yanap@kddi-research.jp

DOI: 10.1587/transinf.2018EDP7107

reference image and the query image. Consequently, only local features which satisfy the following match conditions between the reference image and the query image contribute to the search: 1) corresponding keypoints are detected from the same location of an object, and 2) corresponding feature descriptors are quantized into the same VW. These conditions are not satisfied if the following two errors corresponding to them occur due to viewpoint changes: 1) detection error, and 2) quantization error. Local features with either of these errors decrease the cosine similarity between correct image pairs, and adversely affect retrieval accuracy.

Another property of the BoVW framework is that database size depends on the number of features per image. Therefore, feature selection according to certain reliability measure is an effective strategy for reducing database size. In this strategy, features that have a high probability of contributing to a search should be selected. In other words, the feature should not be registered in the database if either its detection repeatability or its quantization repeatability is low.

In this paper, we propose a feature design method for image indexing which improves the probability of satisfying the match conditions. We use synthetic images that simulate viewpoint changes like Affine-SIFT (ASIFT) [14] instead of actual image collection. We target the recognition of planar objects since this approach is able to generate synthetic images by warping a single reference image. That is, the proposed method has a limitation in that it can only be applied to images of planar objects. However, the proposed method can be applied to 3D objects using computer graphics rendering techniques provided 3D models of target objects exist.

Figure 1 presents an overview of the proposed method. First, we generate various synthetic images from a reference image and extract features from them individually. Then we project keypoints detected from every synthetic image onto the coordinate system of the reference image. We associate keypoints lying on the same location of an object. We measure how frequently features associated with the keypoints are quantized to the same VW and use it as a guide to feature reliability. For instance, the reliability scores for green features become one since their repeatability is low in regard to both keypoint detection and feature descriptor quantization. The reliability scores for blue features become two since their repeatability in terms of keypoint detection is high but their repeatability in terms of feature descriptor quantization is low. The reliability scores for orange features become three (i.e., the most reliable) since their repeatability is high in terms of both keypoint detection and feature descriptor quantization. With this feature reliability measure, we design features to index the reference image while preventing any increase in database size by selecting features.

This paper is an extended version of the paper [15] that appeared in ACPR 2015. Specifically, we formulate feature selection as a maximum coverage problem and introduce a solution to find a finite set of local features maximizing an objective function under certain constraints. For this purpose, the proposed method uses novel constraint functions designed to be consistent with the match conditions.

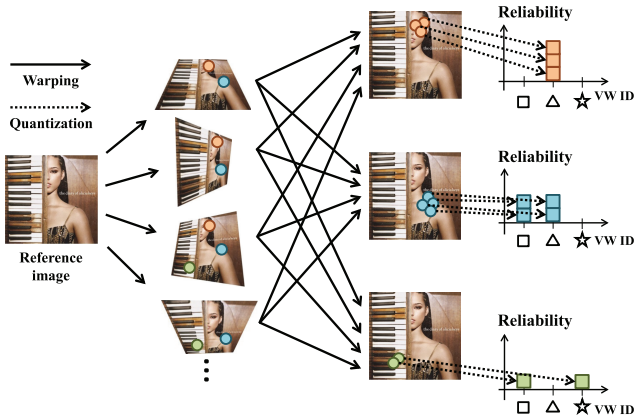


Fig. 1 Overview of the proposed method. Circles represent keypoints. Squares represent quantization results.

The rest of this paper is organized as follows. In Sect. 2, we describe related work on feature design in the context of image retrieval and image matching. In Sect. 3, we introduce a visual search method to apply our proposal. In Sect. 4, we propose a novel feature design method suitable for MVS. In Sect. 5, we evaluate the effectiveness of our proposal. In Sect. 6, we conclude this paper.

2. Related Work

There has been research on feature design in the context of image retrieval and image matching. In this section, we discuss some general approaches to feature design and their practical issues for MVS.

Wang et al. [16] selected informative features which are robust and distinctive against viewpoint changes. This published work is the closest to ours in terms of considering both detection error and quantization error. This approach first connects keypoints across multiview images by performing geometric verification with random sample consensus (RANSAC) [17]. Isolated keypoints which are not connected to others are discarded. With only the remaining keypoints, corresponding features are quantized to VWs. Then, features are ranked according to the score based on an idea similar to term frequency-inverse document frequency (tf-idf) scoring. Finally, top-ranked features are selected to index a reference image for image retrieval. In this way, this approach can reduce both detection error and quantization error. However, this approach is not realistic in practical terms since it requires several images to be taken of the same object from different viewpoints.

Multiple assignment (MA) [18] assigns a single feature to multiple VWs that are the k-nearest neighbor. This is often used in to reduce the quantization error. Mikulík et al. [19] achieved a more accurate MA than this. Their approach trained alternative k VWs using features extracted from different viewpoint images. The alternative k VWs are based on the probability $P(W_b|W_a)$ of observing a VW W_b

from a matching reference feature when a VW W_a was observed from the query feature. However, these methods increase the database size k times. Furthermore, the method of Mikulík et al. is not a practical option for actual use for the same reasons as the method of Wang et al.

Chen et al. [20] projectively warp a reference image to simulate a view expected in the query image and extract local features from the warped image. They generate synthetic images for five different types of views, namely, front, top, right, bottom, and left view from a reference image. Then, they construct five individual sets of data from these images. Their approach should reduce both detection error and quantization error since local features are extracted from every warped image. However, it has limited effectiveness against any views except for the above-mentioned types of views, while it causes a five-fold increase in the overall database size.

Affine-SIFT (ASIFT) [14] achieves a robust image matching result. This approach simulates various viewpoint changes by warping images. It should reduce both detection error and quantization error for the same reasons as the approach of Chen et al. although original ASIFT does not include the quantization process. However, if we apply this strategy in the context of image retrieval, we have to register all synthetic images to the database, which results in massive increases in the size of the database.

In this paper, we also focus on VWs observed from multiview images. We propose a practical method to local feature reliability measure based on image warping. In this method, it is not necessary to collect several images taken from different viewpoints. The proposed approach achieves more effective MA to improve the probability of satisfying the match conditions. It makes it possible to adjust database size by selecting features according to the reliability measure. In our experiments, it is shown that the proposed method outperforms regular MA when the database size is the same.

3. Visual Search Methods

Although many methods have been proposed to improve the BoVW framework for real-valued features [21], [22], [36], only a few studies have applied binary features to BoVW framework for the purpose of image retrieval. In [23], binary features are applied to a BoVW framework, referred as bag-of-binary words (BoBW). In [24], a variant of the Hamming embedding method [21] for binarized features is proposed to improve retrieval accuracy while suppressing memory cost. This method takes out the first a-bit binary string of the feature as VW, and the next b-bit binary string is stored in an inverted index. In [25], a visual search method based on binary features is proposed, which achieves a real-time MVS on a smartphone. It is shown that this method can improve retrieval accuracy compared to conventional methods without increasing the database size. In contrast to the conventional stand-alone MVS methods which require tens of megabytes (MBs) of storage [1], [2], this method has the advantage of needing only a few MBs of storage to recognize several hundred objects. Therefore, we regard this method [25] as a baseline method in this study, i.e., our feature design method for image indexing builds on the framework. We also add a weak geometric consistency method [21] in order to boost the search performance. Figure 2 shows the framework of the baseline method. This framework consists of a search procedure (left side) and an indexing procedure (right side). We describe this framework in detail below.

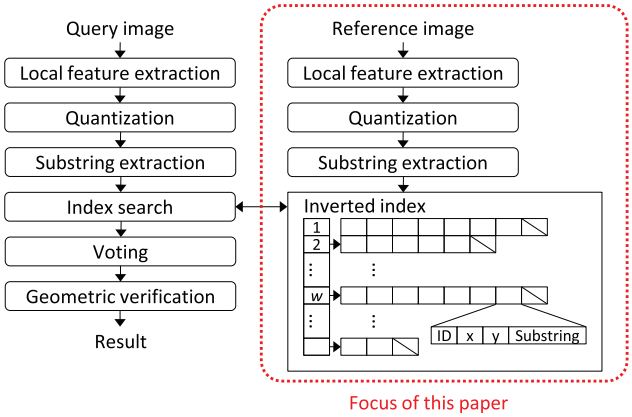


Fig. 2 Framework of the baseline method.

In accordance with the standard BoVW framework, local features extracted from an image are quantized to representative vectors called visual words. Initial ranking is obtained according to image similarity, which is calculated by each visual word of a query image voting scores on reference images. After the voting process, spatial re-ranking is performed by geometric verification to improve the initial ranking [36]. In addition, this framework includes the following extension methods.

Adaptive substring extraction Substring (SS) [25] converts the feature descriptors into compact codes and stores them in the inverted index, similar to Hamming embedding (HE) [21]. SS generates a short binary string by extracting bits of specific positions of the binary feature descriptor. These positions are adaptively changed for each visual word to extract distinctive bits. As with HE, SS measures the Hamming distance between each short binary string.

Modified local NBNN scoring. Given short binary strings extracted from a query feature q and a reference feature r assigned to the same visual word, the scoring function will be $w(d_k) = (d_K/d_k)^2 - 1$, where d_x represents the Hamming distance between short binary strings of the q and the x-th nearest neighbor r. K is an adjustable parameter that specifies the number of the nearest neighbors used in the scoring. In this paper, we use $K = 2$ as in the original paper [25]. This is a modified version of the local NBNN scoring scheme [26].

Weak geometric consistency. Weak geometric consistency (WGC) provides constraints based on angle and scale information [21]. It filters feature correspondences that are not consistent in terms of orientation difference and scale ratio

in the voting process of Fig. 2. We use only angle information because scale information is not effective as shown in [27].

4. Proposed Method

In this section, we propose a feature design method for image indexing which selects features according to a certain measure, referred as a feature reliability measure. This measure is the probability of satisfying the match conditions when the query image is actually given. However, the actual query image is unknown when indexing a reference image. We therefore use synthetic images emulating the actual query images instead. In order to make this practical, we generate synthetic images by warping a reference image. It allows us to index a reference image that achieves robust recognition if the actual query image is similar to one of the synthetic images. Our goal is to satisfy the match conditions for as many query features as possible with a small number of reference features. To accomplish this, we formulate and solve the feature selection problem as a maximum coverage problem. In the indexing procedure, we register optimized features to the database in order to improve SS and WGC performance. Furthermore, we reduce the burstiness of the VW directly and database size simultaneously. This method changes only the indexing procedure of the visual search framework shown on the right side of Fig. 2.

4.1 Problem Formulation

Let $X = \{x_1, x_2, \ldots, x_N\}$ be N sets of unknown local features and $Q = \{q_1, q_2, \ldots, q_M\}$ be $M(\geq N)$ sets of local features extracted from synthetic images. We formulate feature selection as the maximization of an objective function under certain constraints:

$$max. \quad \sum_{i \in X} \sum_{j \in Q} z_j f(x_i, q_j), \tag{1}$$

$$s.t. \quad \sum_{i \in X} f(x_i, q_j) \leq N, \ \forall j \in Q, \tag{2}$$

$$f(x_i, q_j) \in \{0, 1\}, \ \forall i \in X, \ \forall j \in Q, \tag{3}$$

$$z_j \in \{0, 1\}, \ \forall j \in Q, \tag{4}$$

where f is a constraint function with regard to feature matching, and z_j is a variable which is 1 if q_j has not satisfied the constraint for any x_i, otherwise 0. We use the variable z to eliminate the redundancy of X. This induces the finite number of x_i to match as many q_j as possible. In other words, X covers the maximum number of features of Q that are effective for the visual search. This is a maximum coverage problem known to be NP-hard [28], thus computationally infeasible. In order to find a solution within a feasible computational time, we propose an approximate solution.

Specifically, the proposed method is based on a greedy algorithm which is referred to as the polynomial time approximation approach to the maximum coverage problem [29], [30]. Such an approximation approach to the maximum coverage problem itself is not novel, and it is widely used in the context of designing wireless sensor networks [31], text summarization [32], and so forth. However, to the best of our knowledge, this is the first attempt to formulate feature selection in the context of image retrieval as a maximum coverage problem. For this purpose, we propose novel constraint functions in Sect. 4.2 and describe our algorithm in detail in Sect. 4.3.

4.2 Constraint Functions

The constraint functions are designed to be consistent with the match conditions in the baseline method described in Sect. 3. That is, they follow the constraints based on keypoint detection and feature quantization between a query feature q and a reference feature r. Furthermore, they include a constraint based on WGC.

Each local feature l has a spatial coordinate $c(l)$, quantized feature descriptor, namely visual word $v(l)$, and an orientation $\theta(l)$. The keypoint-based constraint (KBC) function is as follows:

$$f_c(r, q) = \begin{cases} 1 & \text{if } \|c(r) - Pc(q)\| \leq \epsilon_c \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where ϵ_c is a threshold value of the re-projection error and P is a known projection matrix (e.g., homography) to the coordinate system of the reference image. Then, the visual word-based constraint (VBC) function (i.e., whether feature descriptors are assigned to the same visual word or not) is represented by:

$$f_v(r, q) = \begin{cases} 1 & \text{if } v(r) = v(q) \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

The orientation-based constraint (OBC) function for the WGC is given by:

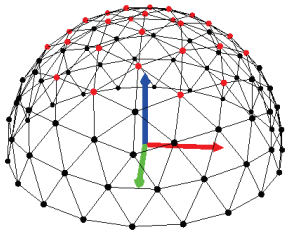
$$f_\theta(r, q) = \begin{cases} 1 & \text{if } |\theta(r) - \theta(q)| \leq \epsilon_\theta \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where ϵ_θ is a threshold value of the orientation difference and thus is consistent with the resolution of the quantized angle [21]. This OBC function mimics the filtering of feature correspondences by WGC. That is, the OBC function is a constraint function designed to be consistent with a match condition based on WGC.

In the baseline method, voting is performed only when all of the above constraints are satisfied. The constraint functions are summarized by the following equation:

$$f(r, q) = \begin{cases} 1 & \text{if } f_c(r, q) \wedge f_v(r, q) \wedge f_\theta(r, q) \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

In Eq. (6), we assume that a feature descriptor is assigned to a VW. However, we can also use this constraint function in the case where a feature descriptor is assigned to multiple VWs by MA, by regarding a local feature to be multiple



(a) Illustration of virtual viewpoints (same scaling factor). Red dots represent viewpoints where elevation exceeds 45 degrees.

(b) Example of synthetic images (same scaling factor).

Fig. 3 Synthetic image generation.

local features with different $v(l)$ and the same $c(l)$ and $\theta(l)$. For instance, if a feature descriptor of l is assigned to $v_1(l)$ and $v_2(l)$, the l is considered to be two local features with $\{c(l), v_1(l), \theta(l)\}$ and $\{c(l), v_2(l), \theta(l)\}$ respectively. Then, we input these local features into the constraint function individually.

4.3 Algorithm of Feature Design for Image Indexing

The proposed feature design method for image indexing consists of the four steps listed below.

Step 1: Generate synthetic images to emulate the actual query images captured by mobile devices (Sect. 4.3.1).

Step 2: Measure the feature reliability using local features extracted from synthetic images (Sect. 4.3.2). This step finds a finite set of local features maximizing an objective function under certain constraints.

Step 3: Average feature descriptors and orientations extracted from the synthetic images to calculate the optimal feature for SS and WGC (Sect. 4.3.3).

Step 4: Select features according to the reliability while avoiding selecting the same VW multiple times (Sect. 4.3.4). The purpose is to reduce the burstiness of the VW directly and the database size simultaneously.

4.3.1 Synthetic Image Generation

We generate synthetic images from a virtual viewpoint V_i by warping the reference image in order to simulate various query images captured by mobile devices. We perform a uniform sampling of the virtual viewpoints positions as done in [33]. In this paper, we empirically use 26 viewpoints where elevation exceeds 45 degrees in 71 viewpoints [34] because a finer sampling interval results in more calculation (see Fig. 3a). For each viewpoint, we also generate multi-scale images to simulate scale changes using scaling factors of 1, $1/\sqrt{2}$, and $1/2$. We calculate the homography projection matrix P_i corresponding to V_i as done in [33], and generate synthetic images as shown in Fig. 3b.

Algorithm 1 Feature design method for image indexing

Input: $Q = \{q_1, q_2, ..., q_M\}$
Output: $X = \{x_1, x_2, ..., x_N\}$

1: Set $X \leftarrow \emptyset, Q' \leftarrow Q$
2: **for** $i = 1 \rightarrow M$ **do**
3: Set $C_i \leftarrow \emptyset$
4: **for** $j = 1 \rightarrow M$ **do**
5: **if** $f(q'_i, q_j) = 1$ **then**
6: Vote a reliability score to q'_i
7: Insert q_j to C_i
8: **end if**
9: **end for**
10: **end for**
11: **for** $i = 1 \rightarrow M$ **do**
12: Select q' with the highest reliability score: q'_h
13: Average q'_h using C_h
14: **if** a feature with the same VW as q'_h is not included in X **then**
15: Insert the averaged q'_h to X as x
16: **else**
17: Store C_h as discarded features
18: **end if**
19: Set each z corresponding $q \in C_h$ to 0
20: Re-calculate reliability scores for each $q' \in Q'$
21: Remove each $q \in C_h$ from all C_i
22: **if** $|X| \geq N$ **then**
23: **break**
24: **end if**
25: **end for**
26: **while** $|X| < N$ **do**
27: Repeat processing of lines 11-25 using only the discarded features
28: **end while**

4.3.2 Feature Reliability Measure

In order to solve the problem described in Sect. 4.1 within a feasible computational time, we explore approximate solutions based on a greedy algorithm. That is, we explore N sets of features from Q by iteratively selecting a feature that matches the most features in each iteration round. Let $Q' = \{q'_1, q'_2, \ldots, q'_M\}$ be the copy of Q. We vote a score from Q to Q' and then i-th q' obtains the following value, referred as the feature reliability:

$$s_i = \sum_{j \in Q} z_j f(q'_i, q_j), \tag{9}$$

where f is the constraint function of Eq. (8). We set all z to 1 initially and then sequentially select q' with the highest score. This is a combinatorial optimization problem and thus it can be solved in polynomial time. The details of this approach are described in Algorithm 1.

If a q_j votes a reliability score for q' (i.e., $f(q', q_j) = 1$), the q_j is associated with the q'. Let C_i be the set of q associated with i-th q' and $C = \{C_1, C_2, \ldots, C_M\}$ be a set of C_i. The C_i means a set of q covered by i-th q' in the feature space. After measuring the reliability scores for all features, we select a q' with the highest score. Let it be h-th q', namely q'_h. At this time, we average q'_h using C_h. The

details of this process are explained later in Sect. 4.3.3. We insert the averaged q'_h to X as x instead to the original q'_h. We then set each z corresponding to q of C_h to 0. We recalculate scores for q_i according to Eq. (9) and remove the q of C_h from all C_i ($i = 1, 2, \ldots, M$). We repeat the above process until N sets of x are obtained.

The following effects are expected when we select a specified number of features according to the reliability:

Useful feature selection. In the proposed method, features with a high probability of satisfying the match conditions among the synthetic images emulating actual query images are selected. It is expected that these features tend to be matched with the correct query features in an actual search and significantly improve the final result.

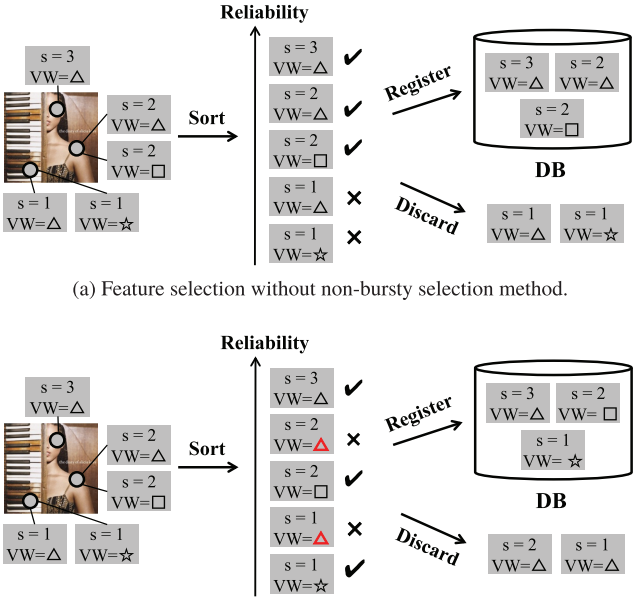
Adaptive multiple assignment. As a result of the proposed method, multiple (non-fixed n) VWs could be assigned to the single keypoint since we associate features that simultaneously satisfy multiple constraints. Figure 1 illustrates this MA. In the middle right part of this figure, we can regard blue keypoints lying on the same location as a single keypoint according to the KBC. Then, we associate two VWs, "□" and "△" with this keypoint based on VBC. It resembles the result of MA, but the number of VWs associated with a keypoint n is adaptively determined. The n becomes large for a feature with high probability of satisfying the KBC and low probability of satisfying other constraints. In contrast, the n becomes small for a feature with high probability of satisfying all constraints. Therefore, we can realize a more efficient MA with respect to memory capacity.

4.3.3 Feature Averaging

In order to improve the performance of SS and WGC described in Sect. 3, we output the average of features associated with the feature q'_h in Algorithm 1. Usually, the features extracted from a front-view reference image are used in SS and WGC. However, in the proposed method, multiple features extracted from various synthetic images are associated with the feature q'_h, namely C_h. Therefore, we average the feature quantities (i.e., feature descriptors and orientations) of C_h, and then replace the feature quantities of q'_h with the averaged values. It is expected to increase the robustness of search because this is the maximum likelihood estimation of feature quantities among various views. That is, we improve the SS and WGC by using the maximum likelihood feature descriptors and orientations under the constraints described in Sect. 4.2. Regarding the feature descriptor, we average each dimension, and binarize the average value. Regarding the orientation, we average unit vectors with the angles of features and use the angle of the averaged vector in WGC.

4.3.4 Non-Bursty Selection

Jégou et al. argued that burstiness of the visual element corrupts the visual similarity measure [35]. They proposed that this burstiness phenomenon could be reduced by scoring, but it is impossible to reduce the database size using their approach. Therefore, we propose a feature selection method to reduce the burstiness of the VW directly and database size simultaneously. In our selection method, if multiple features are quantized to the same VW, only the feature with the highest reliability score is registered and the others are discarded.



(a) Feature selection without non-bursty selection method.

(b) Feature selection with non-bursty selection method.

Fig. 4 Overview of non-bursty selection.

Figure 4 shows an overview of our non-bursty selection method. In Fig. 4b, s and VW represent the reliability score and the visual word of a feature, respectively. Figures 4a and 4b correspond to the feature selection without and with our non-bursty selection method, respectively. In Fig. 4a, features are simply selected in descending order of their reliability scores. In Fig. 4b, the feature with reliability score $s = 3$ and VW = △ is selected first. According to the score s, the feature with $s = 2$ and VW = △ should be selected second. However, unlike Fig. 4a, the feature where $s = 2$ and VW = △ is discarded because the feature with the same VW "△" is already registered. By selecting a specified number of features using this method, we can suppress the burstiness of the VW directly and reduce the database size simultaneously. If the number of registered features did not reach the specified number, we repeat the same process to select new features from the discarded features in the previous process. This method is described in Algorithm 1, lines 14-18 and lines 26-28.

5. Experimental Evaluation

We conduct experiments to evaluate the effectiveness of the proposed feature design method. In the experiments, we use the Stanford mobile visual search dataset[†]. The dataset consists of eight classes, namely book covers, business cards,

[†]https://purl.stanford.edu/rb470rw0983/

Fig. 5 Examples of images in Stanford mobile visual search dataset. The top row shows reference images of a front view. The bottom row shows query images of various views taken on mobile devices.

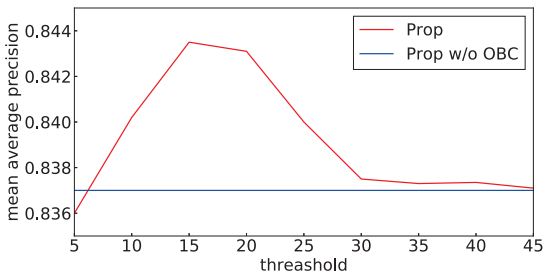


Fig. 6 Average MAP scores of eight classes as a function of the threshold ϵ_θ.

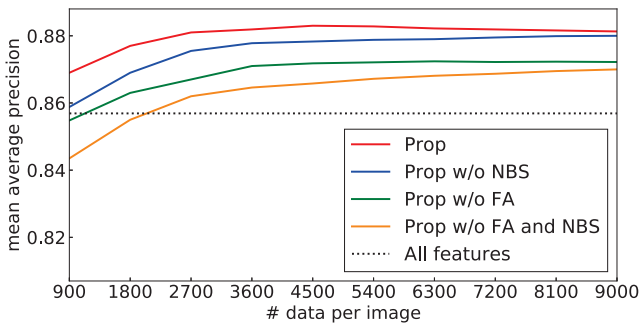


Fig. 7 Evaluation of feature averaging and non-bursty selection. Prop with vs. without FA represents the effect of the feature averaging. Prop with vs. without NBS represents the effect of the non-bursty selection. For reference, the retrieval accuracy in the case of registering all local features is also shown.

CD covers, DVD covers, landmarks, museum paintings, print documents, and video frames. In the dataset, there are in total 1,200 clean reference images and 3,300 query images which are taken from mobile devices. Because the dataset contains images of different sizes, we reconfigure both the query images and the reference images in the VGA format. Figure 5 shows examples of images in this dataset.

As an indicator of retrieval performance, we use the mean average precision (MAP) as in [36]. We adopt the ORB feature [10] since it is the most robust for several general deformations among well-known binary features [37], where 900 features are extracted from four scales on average. The number of VWs and the length of SS are fixed to 1,024 and 64 bits, respectively. We set $\epsilon_c = 3$ in Eq. (5) in imitation of a typical value used as reprojection error threshold for the RANSAC algorithm (e.g., OpenCV library[†]).

Let "KBC" (keypoint-based constraint), "VBC" (visual word-based constraint), and "OBC" (orientation-based constraint) denote a constraint in Eqs. (5), (6), and (7), respectively. Let "FA" (feature averaging) and "NBS" (non-bursty selection) denote the processing described in Sect. 4.3.3 and Sect. 4.3.4, respectively.

5.1 Effect of the Orientation-Based Constraint

We evaluate the effectiveness of the proposed OBC in Eq. (7). In order to focus on the evaluation of the OBC, we exclude FA and NBS processing from the proposed method. We compare the proposed method (Prop) where the OBC (Prop w/o OBC) is excluded. We measure the MAP scores of eight classes while varying the threshold ϵ_θ in Eq. (7).

Figure 6 shows the experimental results. We can see that by comparing the Prop and the Prop w/o OBC, the OBC improves retrieval accuracy. This is because the proposed method gives a high reliability to features with high repeatability of orientation. That is, the stable features whose orientation hardly varies due to viewpoint changes tend to be included in the database. In contrast, unstable features are not included in the database. The accuracy is improved since the WGC filters features that are inconsistent with the orientation difference. The peak value is found in the case of $\epsilon_\theta = 15°$. A threshold that is too small excludes features that may satisfy the match conditions in the actual search. On the other hand, an overly large threshold reduces the effect of the OBC. Therefore, we use the parameter $\epsilon_\theta = 15°$ in subsequent experiments.

[†]https://opencv.org/

5.2 Evaluation of the Feature Averaging and Non-Bursty Selection in Relation to Increasing Database Size

We evaluate FA and NBS when database size increases. We compare the proposed method (Prop) with the cases where FA (Prop w/o FA), NBS (Prop w/o NBS), and both FA and NBS (Prop w/o FA and NBS) are excluded. For reference, we also measure retrieval accuracy in the case of registering all local features extracted from synthetic images in the database (All features).

Figure 7 shows the average MAP of the eight classes of each method as a function of the number of data per image. We can see that feature averaging makes a steady contribution regardless of database size when comparing with or without FA. This is because the averaged features boost the performance of the SS, namely, improve the scores voted from the query features among various views according to the scoring method described in Sect. 3. The averaged orientation also boosts the performance of the WGC by reducing the difference from orientations of query features among various views. Thus the average of features extracted from synthetic images is more reliable than the one of them. By comparing with or without NBS, it is clear that NBS makes

Table 1 Comparison of the proposed method with conventional methods and evaluation of each element of the proposed method in terms of mean average precision. KBC = keypoint-based constraint, VBC = visual word-based constraint, OBC = orientation-based constraint: see Sect. 4.2. FA = feature averaging: see Sect. 4.3.3. NBS = non-bursty selection: see Sect. 4.3.4. In all the methods, 900 features per image are registered in the database.

	KBC	VBC	OBC	FA	NBS	book	cards	cd	dvd	landmarks	paintings	print	video	average
BoBW [23]						0.610	0.173	0.427	0.465	0.080	0.486	0.125	0.584	0.369
[24]						0.874	0.463	0.752	0.811	0.197	0.671	0.423	0.824	0.627
Baseline						0.935	0.616	0.882	0.945	0.292	0.743	0.609	0.923	0.742
Prop	x	x				0.944	0.839	0.949	0.977	0.362	0.861	0.780	0.987	0.838
Prop	x	x	x			0.949	0.848	0.946	0.976	0.382	0.877	0.785	0.981	0.843
Prop	x	x		x		0.956	0.867	0.965	0.980	0.407	0.890	0.801	0.983	0.856
Prop	x	x	x	x		0.951	0.871	0.959	0.982	0.419	0.895	0.811	0.986	0.859
Prop	x	x			x	0.960	0.854	0.951	0.975	0.379	0.849	0.805	0.992	0.846
Prop	x	x	x		x	0.957	0.882	0.953	0.973	0.380	0.866	0.825	0.999	0.855
Prop	x	x		x	x	0.961	0.900	0.961	0.983	0.417	0.874	0.850	0.989	0.867
Prop	x	x	x	x	x	0.964	0.901	0.961	0.984	0.419	0.876	0.855	0.989	**0.869**

a significant contribution, especially when the database size is small. The result suggests that it is better to impose a non-burstiness constraint rather than registering features in order of reliability without that constraint. The effect becomes negligible as the database size becomes larger since the database includes more unreliable features as the number of registered features increases. Therefore, NBS is suitable for constructing a compact database.

In the case of All features, the number of data per image is $78 \times 900 = 70{,}200$. Although it uses all features, it is inferior to the proposed method using fewer features. This is because increasing database size causes retrieval accuracy to deteriorate. In other words, this is because increased reference features behave as distractor features for a certain query feature in the search procedure. For a similar reason, the retrieval accuracy of the proposed method saturated to its upper bound level in Fig. 7. However, the proposed method provides a good trade-off between retrieval accuracy and database size.

5.3 Comparison of the Proposed Method with Conventional Visual Search Methods

We compare the proposed method with conventional visual search methods: BoBW [23], the method [24], and the baseline method [25]. The conventional methods use local features extracted from a front-view reference image, while the proposed method indexes an image using local features extracted from synthetic multiview images. In this experiment, we further investigate the impact of each element of the proposed method. For this purpose, we measure the accuracy when excluding each element from the proposed method. The elements to consider are the constraints of KBC, VBC, and OBC, and the feature design processing of FA and NBS. However, the KBC and the VBC are always used since they are the minimum configuration of the proposed method. For all methods, we registered 900 features per image in the database.

Table 1 summarizes all the results. The baseline method outperforms the other conventional methods due to the extension methods described in Sect. 3. The proposed method raises the accuracy by 9.6 percentage points compared to the baseline method, even in the case of the minimum configuration (i.e., KBC and VBC only). This is because the reference images are indexed using selected reliable features which have a high probability of satisfying the match conditions on various query features. The difference between the baseline method and the proposed method is only the indexing procedure, and the search procedure is common to both methods. Therefore, the proposed method enables more accurate search without adversely affecting the database size, computational cost, and memory requirement as compared with the baseline method. Furthermore, the proposed method does not require additional data such as images taken from different viewpoints of reference objects since it exploits the synthetic images. These facts emphasize the practicality of the proposed method.

In the proposed method, we can see that each element improves accuracy compared to the minimum configuration. This is because OBC selects features with high repeatability of orientation, FA provides the optimal feature descriptors and orientations for SS and WGC respectively, and NBS reduces the burstiness of the visual elements on the database. The effect can be enhanced by combining each element since they complement to each other. When all elements are included, the proposed method can improve the accuracy by 12.7 percentage points compared to the baseline method. However, the NBS was not useful for the painting class. This is because this class includes fewer textured images, so the repeatability of keypoint detection is especially low. The NBS adversely affected this class since it may discard features with high repeatability of keypoint detection.

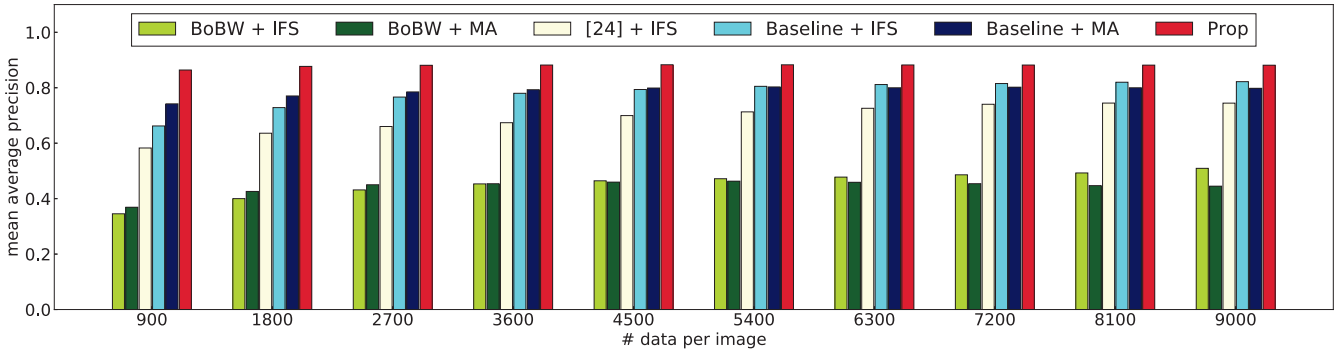


Fig. 8 Comparison of the proposed method with the conventional feature design methods, namely informative feature selection (IFS) and multiple assignment (MA). For MA, 900 features are extracted per image and each feature assigned to k-nearest neighbor VWs (k = 1,...,10). IFS and the proposed method select features so that database size is equal to that of MA.

5.4 Comparison of the Proposed Method with Conventional Feature Design Methods

We compare the proposed method with conventional feature design methods: informative feature selection (IFS) [16] and reference side MA [18]. We apply these feature design methods to the conventional visual search methods. In [16], IFS requires real multiview images and uses the RANSAC algorithm to estimate the geometric relationships among them. In this experiment, we apply the strategy of our synthetic images and ground-truth geometric relationships to IFS for fair comparison with the proposed method. MA that assigns a feature descriptor to k-nearest neighbor VWs cannot be applied to [24] because this method directly uses a subset of the feature as a VW. Therefore, we do not include [24] with MA in the comparison. MA assigns a feature descriptor to k VWs, which increases the database size k times (k = 1,...,10). IFS and the proposed method select features so that database size is equal to that of MA. Although the Hessian affine covariant detector [38] and SIFT descriptor, which are robust to viewpoint changes but computationally expensive, are used in the original papers of IFS [16] and MA [18], we adopt ORB features in this experiment.

Figure 8 shows all the experimental results. As shown, the proposed method achieves the highest accuracy regardless of the database size. Although IFS gradually improves accuracy as database size increases, its accuracy is always inferior to that of the proposed method. This is because the IFS does not always select features with high repeatability of keypoint detection and feature descriptor quantization. That is, although IFS discards non-repeatable keypoints, it does not preferentially select features with high repeatability of keypoint detection. Furthermore, IFS selects features based on the idea of tf-idf but does not consider the repeatability of feature descriptor quantization associated with keypoints lying on the same location. The MA improves the accuracy of conventional visual search methods by alleviating quantization error. However, it does not exceed the accuracy of the proposed method for the smallest database size, even with sufficiently large k. This is because the proposed method alleviates keypoint detection error and orientation description error as well as quantization error. Therefore, the proposed method can achieve more effective feature design than MA for image indexing.

In comparing IFS and MA, IFS is inferior to MA when the database size is small, but superior to MA when the database size is large. This is because IFS does not register many of the features with high repeatability of keypoint detection when the database size is small. However, IFS gradually registers many of such features as the database size increases, resulting in improved accuracy. In the original paper, although IFS improves accuracy even when the database size is small, this seems to be because the Hessian affine covariant detector can detect keypoints robustly with respect to viewpoint changes. On the other hand, MA has a relatively low dependency on database size since it uses common keypoints.

Table 2 Computation times [milliseconds] of the proposed method.

	M (# features extracted from synthetic images)				
	9,000	18,000	27,000	36,000	45,000
Synthetic	38.53	78.40	117.6	152.4	186.1
Extract	100.9	202.4	300.8	396.8	483.7
Quantize	24.20	48.07	72.43	95.81	120.3
Design	162.6	569.7	1,249	2,296	3,671
FA	22.81	42.07	61.67	79.97	108.9
NBS	0.391	0.598	0.822	1.043	1.270
Total	349.5	941.2	1,803	3,022	4,571

5.5 Computation Times for Various Input Sizes

We measure the computation times of the proposed method in different M, namely number of features extracted from synthetic images. We generate 10, 20, 30, 40, and 50 synthetic images and then extract 900 features from each image. As a result, the Ms become 9,000, 18,000, 27,000, 36,000, and 45,000, respectively. Table 2 shows computa-

tion times for synthetic image generation (Synthetic), local feature extraction from synthetic images (Extract), feature descriptor quantization (Quantize), feature design described in Algorithm 1 (Design), feature averaging (FA), and non-bursty selection (NBS). Note that we independently measured the FA and NBS so that the Design does not include them. These computation times are measured on a desktop PC with 3.4 gigahertz Intel Core i7-6800K CPU and 32 gigabytes of RAM.

Synthetic and Extract increased linearly with respect to the number of synthetic images. Quantize also increased linearly with respect to M. These tendencies are obvious considering their algorithms. On the other hand, Design increased quadratically with respect to M. As shown in Algorithm 1, the computational complexity of Design is $O(M^2)$ since the procedure includes double iterations with respect to M. In our implementation, the processing of line 12 in Algorithm 1 is based on linear search and its computational complexity is $O(M)$. The computational cost of lines 19-21 in Algorithm 1 is negligible. FA increased linearly with respect to M because the size of C_h described in Sect. 4.3.2 tended to increase linearly. NBS increased slightly as M increased. This is because burstiness of VW is more likely to occur as M is larger. Therefore, the proposed method can be processed in polynomial time, namely $O(M^2)$.

6. Conclusion

In this paper, we proposed a feature design method for mobile visual search based on binary features. We formulated feature selection as a maximum coverage problem and provided an approximate solution to solve the problem in feasible time. The objective function is designed to be consistent with the match conditions in the search method. Experimental results on a public dataset with viewpoint changes showed the effectiveness of the proposed method. The proposed method improved retrieval accuracy without increasing the database size and changing the search procedure. The next task is to extend our procedure to non-planar objects in order to improve the accuracy of 3D object search.

References

[1] J. Panda, M.S. Brown, and C.V. Jawahar, "Offline mobile instance retrieval with a small memory footprint," Proc. International Conference on Computer Vision, pp.1257–1264, 2013.
[2] D.M. Chen and B. Girod, "Memory-efficient image databases for mobile visual search," IEEE Multimedia Mag. Mag., vol.21, no.1, pp.14–23, 2014.
[3] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," Proc. International Conference on Computer Vision, pp.1470–1477, 2003.
[4] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," Proc. Computer Vision and Pattern Recognition, pp.3384–3391, 2010.
[5] P. Piccinini, A. Prati, and R. Cucchiara, "Real-time object detection and localization with SIFT-based clustering," Image. Vision. Comput., vol.30, no.8, pp.573–587, 2012.
[6] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," Proc. Computer Vision and Pattern Recognition, pp.1582–1590, 2016.
[7] D.G. Lowe, "Distinctive image feature from scale-invariant keypoints," Int. J. Comput. Vision., vol.60, no.2, pp.91–110, 2004.
[8] H. Bay, T. Tuytelaars, and L.V. Gool, "SURF: Speeded up robust features," Proc. European Conference on Computer Vision, pp.404–417, 2006.
[9] S. Leutenegger, M. Chli, and R.Y. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," Proc. International Conference on Computer Vision, pp.2548–2555, 2011.
[10] E. Rubee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," Proc. International Conference on Computer Vision, pp.2564–2571, 2011.
[11] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," Proc. International Conference on Computer Vision, pp.510–517, 2012.
[12] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," Proc. Computer Vision and Pattern Recognition, pp.3384–3391, 2010.
[13] Y. Uchida, S. Sakazawa, and S. Satoh, "Image retrieval with fisher vectors of binary features," ITE Trans. Media Technology and Applications, vol.4, no.4, pp.326–336, 2016.
[14] J.-M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," SIAM J. Imaging Sci., vol.2, no.2, pp.438–469, 2009.
[15] K. Matsuzaki, Y. Uchida, S. Sakazawa, and S. Satoh, "Local feature reliability measure using multiview synthetic images for mobile visual search," Proc. Asian Conference on Pattern Recognition, pp.156–160, 2015.
[16] Z. Wang, Q. Zhao, D. Chu, F. Zhao, and L.J. Guibas, "Select Informative Features for Recognition," Proc. International Conference on Image Processing, pp.2477–2480, 2011.
[17] M.A. Fischler and R.C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM, vol.24, no.6, pp.381–395, 1981.
[18] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," Proc. Computer Vision and Pattern Recognition, pp.1–8, 2008.
[19] A. Mikulík, M. Perdoch, O. Chum, and J. Matas, "Learning a fine vocabulary," Proc. European Conference on Computer Vision, pp.1–14, 2010.
[20] D. Chen, M. Rabbani, R.L. Stevenson, S.S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod, "Robust image retrieval using multiview scalable vocabulary trees," Proc. Visual Communications and Image Processing, pp.72570V, 2009.
[21] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," Int. J. Comput. Vision., vol.87, no.3, pp.316–336, 2010.
[22] M. Shi, Y. Avrithis, and H. Jégou, "Early burst detection for memory-efficient image retrieval," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.605–613, 2015.
[23] D. Gálvez-López and J.D. Tardós, "Real-time loop detection with bags of binary words," Proc. International Conference on Intelligent Robots and Systems, pp.51–58, 2011.
[24] W. Zhou, Y. Lu, H. Li, and Q. Tian, "Scalar quantization for large scale image search," Proc. 20th ACM international conference on Multimedia, pp.169–178, 2012.
[25] Y. Uchida, S. Sakazawa, and S. Satoh, "Adaptive Substring Extraction and Modified Local NBNN Scoring for Binary Feature-based Local Mobile Visual Search without False Positives," ITE Trans. Media Technology and Applications, vol.5, no.1, pp.24–34, 2017.
[26] S. McCann and D.G. Lowe, "Local naive bayes nearest neighbor for image classification," Proc. International Conference on Computer Vision, pp.3650–3656, 2012.
[27] S.S. Tsai, D.M. Chen, G. Takacs, V. Chandrasekhar, R. Vedantham, R. Grzeszczuk, and B. Girod, "Fast geometric re-ranking for image

based retrieval," Proc. International Conference on Image Processing, pp.1029–1032, 2010.
[28] U. Feige, "A threshold of ln n for approximating set cover," J. ACM, vol.45, no.4, pp.634–652, 1998.
[29] D.P. Willamson and D.B. Shmoys, "The design of approximation algorithms," Cambridge University Press, 2011.
[30] D.S. Hochbaum, "Approximating covering and packing problems: Set cover, vertex cover, independent set, and related problems," Approximation Algorithms for NP-Hard Problem, pp.94–143, 1997.
[31] M.A. Guvensan and A.G. Yavuz, "On coverage issues in directional sensor networks: A survey," Ad Hoc Networks, vol.9, no.7, pp.1238–1255, 2011.
[32] E. Filatova and V. Hatzivassiloglou, "A formal model for information selection in multi-sentence text extraction," Proc. 20th International Conference on Computational Linguistics, pp.397–403, 2004.
[33] S. Hinterstoisser, V. Lepetit, S. Benhimane, P. Fua, and N. Navab, "Learning real-time perspective patch rectification," Int. J. Comput. Vision., vol.91, no.1, pp.107–130, 2011.
[34] D. Kurz, T. Olszamowski, and S. Benhimane, "Representative feature descriptor sets for robust handheld camera localization," Proc. International Symposium on Mixed and Augmented Reality, pp.65–70, 2012.
[35] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," Proc. Computer Vision and Pattern Recognition, pp.1169–1176, 2009.
[36] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," Proc. Computer Vision and Pattern Recognition, pp.1–8, 2007.
[37] I. Şahin, "A comparative evaluation of well-known feature detectors and descriptors," International Journal of Applied Mathematics, Electronics and Computers, vol.3, no.1, pp.1–6, 2014.
[38] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," Int. J. Comput. Vision., vol.60, no.1, pp.63–86, 2004.

Kohei Matsuzaki received his B.E. and M.E. from Tohoku University, Miyagi, Japan, in 2010 and 2012 respectively. In 2012, he joined KDDI Co. Ltd. and has been engaged in research and development in the field of content-based multimedia retrieval and image recognition. He is an associate research engineer of the Media Recognition Laboratory at KDDI Research, Inc.

Kazuyuki Tasaka received his B.E. degree from Niihama National College of Technology in 2002. and his M.E. and Ph.D. degree from Nara Institute of Science and Technology in 2004 and 2010, respectively. Since joining KDDI Research, Inc. in 2004, he has worked in the field of network architecture, communication protocols and context recognition. Now, he is a R&D manager in the Media Recognition Laboratory and is a member of IEICE and IPSJ.

Hiromasa Yanagihara received his B.E., M.E., and Ph.D degrees from Nagoya University in electronics engineering, electrical engineering and computer science, in 1988, 1990 and 2014, respectively. In 1990, he joined KDD Co. Ltd. Since 1997, he has been with its R&D Division. His current research interests include video coding, video transmission, image processing, and related multimedia systems. He is currently an executive director in the Media ICT Division in KDDI Research, Inc.