Abstract

Finding corresponding image points is a challenging computer vision problem, especially for confusing scenes with surfaces of low textures or repeated patterns. Despite the well-known challenges of extracting conceptually meaningful high-level matching primitives, many recent works describe high-level image features such as edge groups, lines and regions, which are more distinctive than traditional local appearance based features, to tackle such difficult scenes.

In this paper, we propose a different and more general approach, which treats the image matching problem as a recognition problem of spatially related image patch sets. We construct augmented semi-global descriptors (ordinal codes) based on subsets of scale and orientation invariant local keypoint descriptors. Tied ranking problem of ordinal codes is handled by increasingly keypoint sampling around image patch sets. Finally, similarities of augmented features are measured using Spearman correlation coefficient. Our proposed method is compatible with a large range of existing local image descriptors. Experimental results based on standard benchmark datasets and SURF descriptors have demonstrated its distinctiveness and effectiveness.

1. Introduction and related works

Correspondence is one of the fundamental and intensively studied problems in computer vision. Approaches based on texture analysis around local interest points [1, 2, 3, 4, 5, 6] have gained much attention recently due to their robustness to distortion and occlusion, and avoidance of difficult problems such as edge detection and segmentation. However, when used alone, the locality property of such features frequently leads to noisy results. Geometric constraints and consistency check generally need to be applied as additional layer to refine the result. Furthermore, although experimentally proven to be remarkably robust and invariant to viewpoint changes and distortions usually at the cost of heavily computational demands, local features are also typically not very distinctive basically because of the lack of built-in geometry information. Therefore, they tend to face great difficulty for scenes without high-textured surfaces or with repeated patterns, easily confuse texture-based methods.

To tackle the problem, many recent works in matching and recognition domains try to directly extract and describe high-level matching primitives such as constellation of edges [7, 8] and regions [9, 10, 25] to enhance the descriptors distinctiveness. However, the accurate and clean acquisition of those primitives remains a challenging open problem in computer vision. The acquired curves and shapes are often insufficient stable between images due to viewpoint and lighting changes.

This paper presents our novel approach in a general framework to generate more distinctive features based on invariant local feature descriptors, without the need of high-level feature extraction. We propose to treat the classic image matching problem as a collection of image recognition problems, which naturally integrates geometry information into our augmented features.

Fig. 1 shows the major steps involved when generating augmented features. First, we detect stable interest points in input images and extract corresponding image patches based on positions, orientations and scales of those points. Next, the semi-global descriptor for each sub-image is computed by first accumulating offsets between the sub-image center and member patches’ descriptors, which produces what we call "augmented distinctive features." The acquired curves and shapes are often insufficient stable between images due to viewpoint and lighting changes.
called relative features, containing geometry information of the neighborhood. The augmented features (ordinal codes) are generated by considering the relative rankings of features’ components instead of their original values, which can be regarded as a feature normalization process.

From certain point of view, our feature augmentation process is analogous to the coding process in the general framework of image recognition. One difference is what we are coding here are sub-images, not local patches (e.g. [19]) or the whole image (e.g. [24]). Additionally, we propose increasingly keypoint sampling around sub-images to handle tied ranking problem of ordinal codes. Finally, Spearman correlation is used to measure the similarity between augmented features and establish the point-to-point correspondences for image pairs.

Extensive experimental results based on SURF features have demonstrated the effectiveness of our augmented features. We conduct experiences using benchmark datasets plus precision-recall analysis, together with supplemental real-word challenging image pairs, both indicating higher distinctiveness and lower outlier level of our augmented features compared with base features. The additional computational cost of the feature augmentation process is nominal, which makes building interactive or even real-time computer vision applications possible when combined with fast base features. Furthermore, our feature augmentation framework is general enough to be compatible with a wide range of existing invariant local features, providing important performance gain to image representation methods or upgrading existing image feature databases at a minimum cost. We also provide theoretical analysis about compatibility issues of our method working with different kinds of local features.

Our proposed approach is related to the following recent works: In [11], Boureau, et al. provided a systematic evaluation about combinations of various coding and pooling techniques for recognition, and concluded that large performance increase can be obtained by merely representing neighboring descriptors jointly. Similar idea was explored in [12] which also proposed a novel way to optimize dimension reduction and approximated nearest neighbor searching at the same time.

Concerning the ordinal description method we used, the early work of ordinal measurement tracks back to M. Kendall [15]. Recent works applying ordinal description on image correspondence and recognition problems include R. Zabih and J. Woodfill [16], J. Luo, et al. [17] and M. Toews and W. Wells [18], which built upon SIFT descriptors and reported superior results in terms of precision-recall when compared with many widely-used descriptors such as original SIFT, PCA-SIFT and GLOH.

Our sub-image concept is also enlightened by the following classic techniques: first of all, the shape context [20] takes sampled contour points as inputs and constructs shape descriptors in log-polar histograms using relative positions (direction and distance) between contour points. Two fundamental differences are: our sub-images are described by counting offsets in invariant feature space, not in image space. And the construction of sub-images is directly based on stable interest points, therefore avoiding the need to acquire clean shapes and contour points from images, which is a challenging fundamental vision problem by itself. Second, our sub-image description is also related to Fisher kernel, which provides the offset directions in parameter space into which the learnt distribution should be modified in order to better fit the newly observed data. Recent work of Perronnin et al. [21] uses fisher kernel to obtain compact and efficient image representations for image classification. Last, the famous bag-of-features approach [22] uses K-mean to generate code-books based on invariant local features, assigns each query descriptor to one item in the code-book, producing a histogram of codes representing the whole image, while our approach proposes sub-image as matching primitives and uses accumulated offsets instead of the cluster centers in the feature space. Each base descriptor is converted into exactly one augmented descriptor, facilitating dense point-to-point correspondences.

2. Feature augmentation process

This section presents the reasoning and details of our feature augmentation process. Based on keypoint locations and descriptions, we construct sub-images representing interested neighborhood in image space, then compute relative features for each sub-image integrating geometry information in feature space, which are normalized using ordinal description and produce final augmented features.

2.1. Sub-images

Many existing image matching methods use local image patches around interested keypoints as matching primitives. The image patches can be extracted either from scale and rotation-adaptive neighborhoods, where transformation parameters are determined through searching in scale space and orientation histograms (e.g. [2], [3], [4] and [5]), or from regular regions of fixed sizes, which achieve viewpoint invariance through separated multiple-view training process (e.g. [6], [19] and [23]). The produced local features can be very robust against viewpoint changes and distortions, but usually insufficient distinctive due to the lack of global geometry information. On the other hand, image classification and recognition works focus more on distinctiveness among different images rather than robustness, generally producing one unified descriptor for each image and can by no means provide robust point-to-point correspondences.

We propose to use sub-images as matching primitives, aiming to fill the gap between the above two extremes. Sub-images are sets of local image patches that are close
to each other in image space (Fig. 2). As relative local structures, they are robust to factors such distortions and occlusions. They also integrate semi-global geometry information in order to improve feature distinctiveness.

Figure 2: Illustrations of sub-image concept ($k = 5$ cases). Red and blue dots are leaders and members of each sub-image respectively. Grey crosses are other interest points.

To construct sub-images, we first detect stable interest points (represented by $P_i$) and extract fixed or invariantly adaptive image patches around each interest point depending on which detector is used. After the position of each $P_i$ is stored and properly indexed (hierarchy tree structure preferred especially when the total number of keypoints, represented as num, is large), we construct one sub-image structure for each $P_i$ (represented by $S(P_i)$) with $P_i$ as its leader and its $k$ nearest neighbors in image space as members.

$$S(P_i) = \{P_i\} \cup \{k-NN(P_i)\}, \quad 1 \leq i \leq \text{num}$$

Each $S(P_i)$ is an abstract type of “image” containing information for the neighborhood of $P_i$. Both leaders and members of sub-images are centrally organized by keypoints. As a result, the keypoint detector should be sufficient stable so that the constitution of the same sub-image will remain consistent to certain degree between different input images under various viewing conditions. In our current experiments, we use the efficient keypoint detector in [3] based on integral images, which is very fast to apply and provides us with sufficient repeatability.

After the sub-image construction, finding matching points in the input image pair is equivalent to classifying a large set of sub-images generated from both images.

2.2. Relative features

Given sub-images, our next task is to form distinctive descriptors, which will be used later to measure the similarities between sub-images efficiently. This step is analogous to the coding process in image recognition.

Suppose we choose an invariant local feature descriptor (with descriptor dimensionality represented by $\text{dim}$) and $D(P_i)$ is the feature vector of $P_i$ and $D_j(P_i)$ represents its j-th vector component. Notice that there is no need to adopt the keypoint detector and descriptor all belonging to the same image matching system, as long as the detector selected is stable and the descriptor invariant, meeting general requirement of those components. This provides extra flexibility for our augmented feature to work well with various combinations of existing keypoint detection and image representation techniques.

If we seek to combine the sub-image’s descriptors directly, for example, simply concatenating $D_j(P_i)$ for all the $P_i$ belonging to the same sub-image, at least two drawbacks will immediately follow. One is the rather high dimensionality of the resulting descriptors (when $k = 5$ and standard SURF [3] descriptor is used, each sub-image will be associated with a 384-D vector), which raises practical obstacles to later indexing and searching steps due to dimensionality curse. The other is such combination will highly likely compromise the invariant properties of the base descriptors because of the different orders when performing the concatenation.

Encouraged by successful classic techniques such as Fisher kernel [21] and shape context [22], we believe that generally speaking, features based on relative and aggregated values (either in image space like shape context or in feature space like “macro features” in [11]) usually demonstrate more robust performance than features directly based on raw values of base descriptors.

Based on the above reasoning and analysis, we compute relative features for $S(P_i)$ (denoted as $R(P_i)$) by accumulating $k$ offsets between leader and member descriptors of $S(P_i)$ for each dimension. Formally, we can define:

$$\delta_{i,n} = \begin{cases} 1 & \text{if } P_n \text{ belongs to the sub-image leaded by } P_i \\ 0 & \text{otherwise} \end{cases}$$

Then the j-th dimension of the relative feature is computed as:
The computed relative features have a constant low dimensionality regardless of $k$, the number of members in each sub-image. Moreover, assuming the bases descriptors are scale and rotation-invariant, and the keypoint detector is stable, it can be easily proved that the relative features produced will also be invariant, since the sub-image members are considered in an order-independent way. This is a considerable advantage for image matching applications, achieved by computing relative features in invariant feature space and maintaining invariance during the computation process. As a comparison, features computed by aggregating offsets in image space, such as shape context, are fundamentally not invariant to orientations and typically obtain limited degree of scale invariance through normalization techniques.

In our current experiments, we use SURF descriptor [3] with $\text{dim} = 64$. We visualize the generated relative features by first transforming their components into 0–255 intensity range and then rendering each descriptor as a 2D (8 by 8) histogram. Fig. 3 shows some of our relative features generated using a modified version of the famous testing image pair from [2] (extra lighting changes and distortions are added).

2.3. Normalization

After relative features in each image are computed, we can directly measure their similarities using standard distance metrics such as Euclidean distance and establish the initial correspondences between two feature sets.

For simple image matching pairs, e.g. one image is the in-plane rotated or uniformly scaled version of the other, accurate results could be expected. However, when it comes to challenging image pairs involving complicated viewpoints and lighting changes (e.g. Fig. 3), we found such initial matching results are generally very noisy.

The primary reason lies in the linear relationship assumption of today’s widely-used computational efficient metrics such as Euclidean or chi-squared distances. To handle the violation of linear assumption (e.g. from illumination changes), original feature vectors are usually normalized to remove non-linearity before entering the matching process. Many existing methods use a set of experiences- or experiments-determined parameters to threshold, translate and rescale, fitting the original data into a fixed range, which is naturally difficult to generalize to other viewing conditions or combinations of situations.

As can be observed from figure 3, in terms of raw values’ magnitudes (the intensities of the 2D histogram cells), sub-images of similar physical locations could produce rather different relative features due to viewing condition changes, which is exactly the reason of noisy results when directly measuring the similarities of relative features. However, it can also be observed that although those bins’ absolute values are not identical across different views, their relative orders are overall consistent.

Based on the above analysis and observations, we propose to use ordinal description as a parameter-free and computational efficient approach to normalize our relative features. First, for each relative feature, we locally sort its vector components and produce a sorted array. The augmented feature (represented by $A(P_i)$) is generated by replacing the original values with its corresponding ranks in the sorted array for every dimension. The overall computational complexity of normalizing each relative feature is only $\text{dim} \times \log(\text{dim})$ due to sorting.

$$A_j(P) = |R_m(P)| \quad \text{where} \quad R_m(P) \le R_j(P) \quad 1 \le j \le \text{dim}$$

Our final augmented features consider the relative ranking of features’ each dimension instead of their original values. As a result, each augmented feature is represented as an integer vector of $\text{dim}$ dimensions, more specifically, a permutation of the set $\{1, 2, ..., \text{dim}\}$. The ordinal description we use generates normalized image representations invariant to monotonic deformations and also robust against certain degree of challenging non-linear and non-uniform factors such as partial lightings.

Since each ordinal codes is a permutation of integer set $\{1, 2, ..., \text{dim}\}$, we treat those integers as intensity values and simply visualize each ordinal codes as a barcodes-like 1D pattern. Figure 4 shows the “barcodes” associated with the same sub-images as in Fig.3 in the two input images. It indicates that after the ordinal description, our feature similarity is notably improved. Under different viewing conditions, features belongs to the same physical locations are more similar while others remain distinctive.

Figure 4: Augmented features visualization for similar physical locations in the two input images. The left column is extracted from the left image while the right column from the right image.

One crucial issue for ordinal description is how to handle tied ranking. Given a relative feature vector, when its two components with different indices contain the same original value, the produced ordinal codes will contain tied identical rankings for the two indices, resulting in
unexpected outputs. The simplest way to handle tied ranking is to use the different indices as fail safe reference to break the tied situation. This method is equivalent to using a vector of original indices as a reference. In [18], such reference vector is obtained by averaging a large set of descriptors, e.g. the descriptors of the whole image, offline, which is appropriate for coding the whole image for recognition purpose. However, we argue that for image matching, the reference vector used by semi-local ordinal description should reflect the local ordering trend, which is usually not consistent with the global trend.

### Algorithm: Tied Ranking Breaker

1. **Assume:** \( R_j = R_j; j \neq j \)
2. **Initialize:** set \( S = \{ \text{all keypoints}\} - S(P_i); \)
3. **do iteration** \( t = 1, 2, ..., M; \)
4. \( P_x = NN(P_x, S); \)
5. \( S = S - P_x; \)
6. \( S(P_x) = S(P_x) + P_x; \)
7. **compute** new \( R(P_x, S(P_x)); \)
8. if \( (R_j \neq R_j) \) break;
9. **end iteration**;
10. if \( (R_j \neq R_j) \) use \( (R_j, R_j) \) to handle tied ranking;
11. else use \( (j, j) \) instead;

Therefore, we propose increasing keypoint sampling around sub-images to break tied ranking. The idea is to iteratively include additional nearest neighbors of the current sub-image one by one and compute new relative features during each iteration until the tied ranking is broken, or, for efficiency consideration, until a maximum number of iteration (denoted by \( M \)) is reached, in which case, we use the original indices as reference vector instead. Pseudo codes for tied ranking handling are provided above. \( NN(P_x, S) \) is a function returning the nearest neighbor of \( P_x \) in \( S \).

### 3. Establishing correspondences

Although standard distance metrics e.g. Euclidean distance can be used to measure the similarities, better comparison results can be obtained more efficiently using special metrics based on characteristics of the ordinal codes, for example, given \( dim \), there are a fixed total number of unique codes (\( dim! \)) and are all of the same length.

Ordinal description and measurement have been studied for nearly a century and many specially designed distance metrics were proposed. In this paper, we studied three measurements compatible well with our features and problem domain. First is the measurement of element-wise consistency counting. Assume \( A \) and \( A' \) are two augmented feature vectors for two sub-images, and \( A_j \) is the \( j \)-th vector component. Their similarity can be efficiently computed as:

\[
\text{Dist}(A, A') = \left | \{ A_j | A_j \neq A_j' \} \right | \quad 1 \leq j \leq \text{dim}
\]

The second metric is based on relative ordering of vector component pairs measured by \( \text{sign()} \) function, which returns 1 if the input parameters have the same sign and -1 otherwise. Formally, the Kendall coefficient [15] is defined as the following:

\[
\text{Dist}(A, A') = 2 \sum_{[j \leq \text{dim}], [j', j \leq \text{dim}]} \text{sign}(A_j - A_{m}, A_j' - A_{m}') / \text{dim}(\text{dim} - 1)
\]

Last, based on the ordinal codes’ property of equal length, our modified version Spearman correlation coefficient [14] returns the similarity of two ordinal codes in the range of 0~1.

\[
\text{Dist}(A, A') = 1 - \frac{3 \sum (A_j - A_j')^2}{\text{dim}(\text{dim}^2 - 1)}
\]

Our own experiments indicate that the Spearman correlation coefficient generally provides better matching results than the other two. As additional but rather practical considerations, its conventional unit range also works well with many existing image matching components and frameworks, such as distance ratio (used in [2] and [5]) and consistent check methods like RanSac. Therefore, we select the Spearman correlation coefficient as our distance metric in the next section’s evaluation.

### 4. Experimental results

Our proposed augmented feature has been intensively tested using standard benchmark datasets [5] with known ground truth. We primarily focus on outdoor scenes according to our project’s interest. Besides the standard test data, we also conducted experiments on other real-world image pairs of outdoor nature scenes with man-made objects such as buildings and signs. Many of those additional testing image pairs contain interest objects with surfaces of low-textures or repeated patterns, particularly challenging for various existing local image features.

We experiment different kinds of invariant image description methods and finally select SURF (Speeded Up Robust Feature) [3] as our base descriptor, because it provides the best tradeoff between matching speed and feature robustness. SURF is built upon other experiments-verified and successful detectors and descriptors, such as famous SIFT [2], but simplify those steps to the essential. It proposes the use of integral images to drastically reduce the number of operations for simple box convolutions, independent of the chosen scale. The feature description is based on sums of Haar wavelet components, which can be constructed and matched more efficiently compared with other state-of-the-art methods. In our evaluation, the widely-used 64 dimension of SURF descriptor is used. We set \( k \), the only parameter of the whole feature
augmentation process, to 5 which we experimentally find to provide best results.

Overall, our experimental results demonstrate that compared with base descriptors, the proposed augmented features achieve remarkably higher level of distinctiveness without loss of robustness, at a nominal additional computational cost. Specifically, besides outperforming base features in the standard precision-recall curves, the new augmented feature integrating geometry information generally produces larger number of correspondences under the same distance ratio requirement and less number of noisy matchings especially for challenging real-world scenes. The proposed method is integrated into our image matching system, which generates augmented features of one target image offline then matches the features of query images online and frame by frame. The total processing time each frame including keypoint detection, description, feature augmentation process and finally establishing correspondences takes only 100~150ms with a peak memory occupation of around 15M, making the system an attractive solution for image matching problems of interactive/real-time applications such as outdoor navigation and augmented reality on small mobile devices.

4.1. Standard tests

The standard test sets we experimented consist of image sets of various scenes (K. Mikolajczyk and C. Schmid [5]), with each containing six images (five pairs when the first image is fixed as reference image) of successive increments of one type of image deformation including image blur and compression, lighting and viewpoint changes. One image transformation matrix is associated with each image pair, so that for any keypoint position in the reference image, we are able to compute the ground truth corresponding position in any of the other five images, verifying matchings returned by various methods.

During the evaluation process, for each image pair, we construct base descriptors and two versions of augmented descriptors using local (L) and global (G) reference vector respectively, for the same set of interest points and then use Euclidean distance and Spearman correlation coefficient to measure their similarities and establish initial matchings. Within one test, the same distance ratio is applied to both base and augmented tracks to obtain final reported matchings. Distance ratios are varied in different tests to generate the figure data points. Next, for each keypoint position in the reference image, by comparing its reported matching positions with ground truth matching positions, we are able to distinguish true and false matchings. This crucial information, combined with the total number of reported and possible correspondences, allow us to compute precision and recall for one image pair, under one distance ratio.

According to our project interests, we focus our experiments on challenging outdoor scenes with non-planar objects and low-textured surfaces, handling complicated viewpoint changes. To generate figure 5, we choose 13 image pairs meeting the above guidelines from three image sets (the boat set, the graffiti set, and the bike set), which also represents challenging factors in outdoor mobile applications such as rotations of the device and blurred input images. Each image pair is tested by using base and augmented features and ten different distance ratios. Figure 5 shows the precision and recall results summary of over 300 image matching tests, demonstrating the superior performance of our proposed method.

![Figure 5: Recall-precision for our standard test results summary, evaluated on 64-D SURF as base descriptors. Results averaged over 13 images pairs of K. Mikolajczyk’s testing data set.](image)

4.2. Dense matching from distinctive features

In this subsection, we demonstrate that our augmented features are more distinctive than base features, resulting denser and cleaner correspondences. Towards this goal, we compute base and augmented features using the five image pairs of boat image set (with increasing level of zoom and rotation), apply the same distance ratio and record the number of reported matchings for the two feature tracks respectively.

Since distance ratio is defined as the ratio of the best matching’s similarity over the second best, we believe that from some aspects of view, the number of remaining matchings after applying a fixed distance ratio can also be used as a simple indicator of feature distinctiveness. Assume one method generates similar, clustered thus less distinctive features while the other produces features well scattered and distributed in the feature space. After applying the same distance ratio filter, the first method will have much less matchings remaining because the difference between its different matching’s similarities is much limited.
Figure 6 shows that our augmented features lead to a significantly larger number of reported matchings than the base features, under the same distance ratio. This result can also be considered jointly with section 4.1, which indicates better precision and recall tradeoff of the denser matchings, together with section 4.3, which demonstrates consistent results for images outside the standard datasets as well.

4.3. Matching visualization and analysis

This subsection provides some line by line matching visualization results for both standard test images and other real-world scenes. We apply the same distance ratio and for each of remaining final matching, draw a blue line from its keypoint in reference image to the matched keypoint in the current query image.

As can be observed from figure 7 and consistent with the quantitative evaluation results from the above two subsections, through augmented features, we are able to obtain denser correspondences and less matching outliers than using base features. The primary reason lies in the fact that our augmented features enhance distinctiveness by integrating semi-global geometry information based on sub-images concepts and offset aggregation in feature space, while maintaining and even improving feature invariance and robustness through ordinal description.

Last, regarding the compatibility issue of our feature augmentation, theoretically the proposed method should work well with a large range of existing invariant local features including but not limited to SIFT [2], SURF [3], and GLOH [5]. However, there are a few issues worth considering beforehand.

First of all, it goes without saying that the base feature needs to have descriptors in vector form. Otherwise the computation of relative features and ordinal description can not be applied directly. So special approaches based on direct learning on local image patches without vector representations such as [13] and [23] couldn’t benefit from the feature augmentation process.

Figure 7: Matching visualization and comparison of base (top) and augmented (bottom) features. Standard dataset: (a) Graffiti; (b) Boat; Extra testing dataset: (c) building with similar or repeated patterns; (d) street scene with large viewpoint changes.
remain repeated and consistent across different viewing conditions. Finally, concerning the normalization component in our method, ordinal description generally works better when the feature vectors it applies to have a high dimensionality, providing a large pool of unique codes. Image matching methods advocating dimension reduction (e.g. by using PCA) of their feature vectors, such as [4] and [19] are ill-suited for our normalization component. It is reported in [18] that directly applying ordinal description onto PCA-SIFT [4] can even lead to inferior performance.

5. Conclusion
This paper presents our feature augmentation process, producing more distinctive features for efficiently image matching. The proposed method is built upon the concept of sub-images, which connects close small image patches in image space, converting one image matching problem into a collection of image recognition problems. Our relative features aggregate descriptor offsets in invariant feature space and within sub-images, in order to integrate geometry information and produce semi-global features. Based on visualization and analysis of corresponding relative features, we propose to use ordinal description to normalize and generate augmented features, invariant to monotonic deformations and beyond. Finally, similarities are measured by Spearman correlation coefficient and correspondences are established. Experimental results using standard and supplemental datasets verified the augmented features containing additional geometry information are more distinctive, leading to denser and cleaner correspondences.

References