Scale Invariant Feature Transform (SIFT) in Image Processing: A Review on Different Techniques

Zeinab Sam Daliri
Department of Basic Science, Islamic Azad University, Chalous Branch, Iran

Abstract
For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. To perform reliable recognition, it is important that the features extracted from the training image be detectable even under changes in image scale, noise and illumination. Such points usually lie on high-contrast regions of the image, such as object edges. Another important characteristic of these features is that the relative positions between them in the original scene shouldn’t change from one image to another. Similarly, features located in articulated or flexible objects would typically not work if any change in their internal geometry happens between two images in the set being processed. However, in practice SIFT detects and uses a much larger number of features from the images, which reduces the contribution of the errors caused by these local variations in the average error of all feature matching errors. The review gives a wide scope of such SIFT applications. Keywords-SIFT, Image processing, Descriptor, Ordinal description, Real Time Description, 3D applications

Keywords- SIFT, Image Processing, Ordinal description, Real Time Description

Introduction
The reliable matching of image features is a basic problem in computer vision applications, like 3D reconstruction from stereo images, structure-and-motion estimation, panorama generation, or object recognition. Especially, if the change in 3D viewpoint between the images is large, the matching of the image features must be invariant to image transformations and illumination changes. Usually, the matching process can be divided into two steps. The first step is the detection of feature points (also called keypoints). In this step descriptive image regions are selected and their exact image position is determined. The second step is the keypoint correspondence analysis, where pairwise assignments of keypoints are determined based on local region descriptors (also called keypoint descriptors). A well-established keypoint detector and descriptor is the Scale Invariant Feature Tranform (SIFT), which was published in 2004 by Lowe [1]. After detection and localization of keypoints in different scale space images, an orientation is assigned to each keypoint using local image gradients. Then a keypoint descriptor is assembled from the local gradient values around each keypoint using orientation histograms. In 2005, Mikolajczyk and Schmid [2] carried out a performance evaluation of local descriptors and concluded that the SIFT-based descriptor performs best.

SIFT Descriptor: Local descriptors [3] are commonly employed in a number of real-world applications such as object recognition and image retrieval because they can be computed efficiently, are resistant to partial occlusion, and are relatively insensitive to changes in viewpoint. There are two considerations to using local descriptors in these applications. First, we
must localize the interest point in position and scale. Typically, interest points are placed at local peaks in a scale-space search, and filtered to preserve only those that are likely to remain stable over transformations. Second, we must build a description of the interest point; ideally, this description should be distinctive (reliably differentiating one interest point from others), concise, and invariant over transformations caused by changes in camera pose and lighting. While the localization and description aspects of interest point algorithms are often designed together, the solutions to these two problems are independent. Mikolajczyk and Schmid presented a comparative study of several local descriptors including steerable filters, differential invariants [9], moment invariants, complex filters, SIFT, and cross-correlation of different types of interest points. Their experiments showed that the ranking of accuracy for the different algorithms was relatively insensitive to the method employed to find interest points in the image but was dependent on the representation used to model the image patch around the interest point. In the last few years we have witnessed an explosion of object recognition methods based on the detection of local key-points and construction of local photometric descriptors around these key-points and to construct from the region or its surrounding a discriminative description which is used for matching. The requirement is that the structures can be re-detected with high reliability and that the descriptor is robust and possesses certain invariance properties. The big advantage of these approaches is that they do not require a segmentation of the image and due to the local nature they are robust to occlusions. Local approaches have demonstrated considerable success in a variety of applications, like recognition of objects, wide-base line stereo, robot navigation, image retrieval, building of panoramas, etc. Probably the most popular and widely used local approach is the DoG detector with the SIFT descriptor as proposed by Lowe. SIFT has been used with success in all of the above mentioned application areas. Evaluations and comparison demonstrate the excellent performance of the method compared to other approaches. It has been shown that SIFT descriptors can be used to achieve fairly robust object detection in still images ([4]Lowe 1999). SIFT has furthermore been used in several other related applications such as metric robot localization [5](Se, Lowe, and Little 2001) and medical imaging [6](Moradi, Abolmaesoumi, & Mousavi 2006); and it was shown to be one of the best currently available descriptors in a comparative study (Mikolajczyk & Schmid.)

**SIFT Flow:** Image alignment, registration and correspondence are central topics in computer vision. There are several levels of scenarios in which image alignment dwells. The simplest level, aligning different views of the same scene, has been studied for the purpose of image stitching and stereo matching. The image alignment problem becomes more complicated for dynamic scenes in video sequences, e.g. optical flow estimation [7]. The correspondence between two adjacent frames in a video is often formulated as an estimation of a 2D flow field. The extra degree of freedom transitioning from 1D in stereo to 2D in optical flow introduces an additional level of complexity. Typical assumptions in optical flow algorithms include brightness constancy and piecewise smoothness of the pixel displacement field [8]. Image alignment becomes even more difficult in the object recognition scenario, where the goal is to align different instances of the same object category. Sophisticated object representations [9] have been developed to cope with the variations of object shapes and appearances. However, these methods still typically require objects to be salient, similar, with limited background clutter. In image alignment we must first define the features based on which image correspondence will be established: an image measurement that does not change from one image to another. In stereo
and optical flow, the brightness constancy assumption was often made for building the correspondence between two images. But soon researchers came to realize that pixel values are not reliable for image matching due to changes of lighting, perspective and noise. Features such as phase, filter banks, mutual information and gradient are used to match images since they are more reliable than pixel values across frames, but they still fail to deal with drastic changes. Middle-level representations such as scale-invariant feature transform (SIFT), shape context histogram of oriented gradients (HOG) [10] have been introduced to account for stronger appearance changes, and are proven to be effective in a variety of applications such as visual tracking, optical flow estimation and object recognition. Nevertheless, little has been investigated for exploring features to establish correspondences at the scene level. The SIFT flow algorithm in more depth and will demonstrate a wide array of applications for SIFT flow

**Vector Quantization (VQ)-SIFT:** The performances of several typical local descriptors, such as steerable filters, differential invariants [11], moment invariants and SIFT, etc. have been investigated. It was found the accuracies of algorithms were relatively insensitive in key-point detection stage, but that were much different in description stage. SIFT algorithm which proposed by Lowe obtained the best performance in matching experiments. Previously, Kotani et al. [12] have proposed a very simple yet highly reliable VQ-based face recognition method called VQ histogram method by using a systematically organized Codebook for 4x4 blocks with 33 code vectors having monotonic intensity variation without DC component. VQ algorithm [13] is well known in the field of image coding and schematically. Input image is first divided into small blocks, which are taken as input vectors in VQ operation. Each input vector is then matched with code vectors in a codebook by calculating distances between them. The code vector having the maximum similarity to the input vector is selected by searching the minimum distance and the index number of the selected code vector is output. This index number information is used for represent facial feature. It was found that a code vector histogram, which is obtained by counting the matching frequency of individual code vector, contains very effective facial feature information. By utilizing this technique, a novel face recognition algorithm called VQ histogram method has been developed. The essence of VQ histogram method can be considered that the operation detects and quantizes the direction and the amount of intensity variation in the image block of the face. Hence VQ histogram also can effectively represent image feature information.

**PCA-SIFT:** The problem of dimensionality reduction for feature descriptor has been addressed by several researchers. A first attempt to reduce dimension for local features was PCA-SIFT proposed by Ke and Sukthankar [14]. The key of PCA-SIFT is to apply the standard Principal Components Analysis (PCA) technique to the normalized gradient patches extracted around local features. PCA is able to linearly project high-dimensional descriptors onto a low-dimensional feature space and reduce the high frequency noise in the descriptors. PCA is also employed in the descriptors GLOH by Mikolajczyk and Schmid and Daisy [15]. However, the PCA based feature descriptors need an offline stage to train and estimate the covariance matrix used for PCA projection. Thus the application of PCA based features is limited. Herbert et al. [16] proposed SURF that approximates SIFT by using Haar wavelet response and integral images to compute the histogram bins. SURF has been shown to have similar performance to SIFT, while at the same time being much faster. Recently, Tasi et al. [17] proposed a compact feature descriptor (CDIKP) by combining the scaleinvariant feature detection method and the
kernel projection technique. However, CDIKP only considers the first order derivatives of pixel intensity along horizontal and vertical directions, and in most cases, PCA-SIFT outperforms CDIKP.

**Binary SIFT:** SIFT features are widely used in content-based image retrieval. Typically, a few thousand key points are extracted from each image. Image matching involves distance computations across all pairs of SIFT feature vectors from both images, which is quite costly. We show that SIFT features perform surprisingly well even after quantizing each component to binary, when the medians are used as the quantization thresholds. Quantized features preserve both distinctiveness and matching properties. In the SIFT approach, the salient points of an image are found as the extreme of a multi-resolution image computed using a difference of Gaussian function. Thus, the key points can be detected in different scales. Each key point is described by a 128 dimensional vector which is essentially a histogram of gradient directions for an image patch around the detected key point. The dominant gradient direction(s) are selected as the reference direction, hence providing rotation invariance. Each of the 128 components takes integer values between 0 and 255. SIFT features are widely used in many applications from stereo to object detection, and found to be robust against scale and orientation changes, and quite discriminative even in large databases of features [18]. When searching an image in a database, the key points of the query image are compared to each key point of the each target image. Usually, a few thousand key points are detected per image. Comparison between two images involves vector distance computations in the order of square of number of key points, which is quite costly. A number of methods have been suggested to speed up the matching process. Grauman et al. [19] proposes pyramid match, an approximate but fast matching method between sets of features. In the work by Crandall et al., all the features from all images in a database are clustered and a reduced set of representative vectors are selected, thus providing a more scalable approach. SIFT feature vectors perform quite well even after each component of the vector is quantized to binary. The median value of each component can be used as the quantization threshold for that component.

**Shape Index SIFTs:** There are two recent studies that use shape index within the context of local description. In the first study [20], the authors identify keypoints, where the shape index values are extremum. Around these keypoints, 2D histogramming is formed, in which one of the dimensions belongs to the shape index values of the neighboring pixels, whereas the other one denotes the angles between the normal of the keypoint and the normal of the neighboring pixels (SI-Normal-Hist). In their work, scale invariance problem is not addressed. In a different recent effort, namely 2.5D Scale Invariant Feature Transform (2.5D SIFT) [21], keypoint detection is achieved on 2D range images similar to the Lowes work. However, an extra preprocessing step is required. After this step, using the histogram of the shape index values and the range gradient orientations around these keypoints, a description is obtained. Both of the methods use isolated objects in which problematic boundary regions can be eliminated by simple thresholding however it is not applicable in range scenes due to occlusion. Yet another similar histogramming technique is 3D shape contexts [22]. In that approach, spatial distributions of the surface points are accumulated utilizing their spherical coordinates.

**Color- (C) SIFT:** SIFT was mainly developed for gray images which limits its performance with some colored objects. However, there are some attempts in the literature which have been introduced to make use of the color information inside the SIFT descriptors. For example, in
[23], the normalized RGB model has been used in combination with SIFT to achieve partial illumination invariance besides its geometrical invariance. The color invariance of this approach is still limited because of the primitive color model used. In [24], a multi-stages recognition approach has been developed in order to achieve both color and geometrical invariance. In the first stage, a color classifier is used to label the different image regions. Then, the SIFT descriptors are augmented by adding the color labels. In spite of the good performance of this approach, its need for colored learning instances limits its performance in several applications.

**Ordinal Description:** Ordinal description can be viewed as a meta-technique which considers data samples not in terms of their raw measurement values per se, but rather in terms of their indices or ranks in an array of sorted measurements. Similarity of ordinal or rank-ordered data has been studied for at least a century, classic examples include the Spearman correlation coefficient [25] or the Kendall coefficient [26]. In the context of image similarity, several authors propose comparing image windows in terms of the rankings of sorted image intensities [27]. In a similar vein, Luo et al. use the Kemeny-Snell measure [28] as an image similarity metric [29], based on the relative rankings of image intensities. Ordinal description assumes only a monotonic functional relationship between measurements in different images, and thereby offers a principled, non-parametric means of coping with nonlinearities in the imaging process and avoiding ad hoc normalization schemes to enforce linearity. Despite these strengths, ordinal description is not widely used in the current computer vision literature, for two main reasons. First, descriptors of relatively high-dimensionality are required in order to establish a pool of distinct rankings for discriminative correspondence, as the number of unique rankings is a function of the number of data measurements (N!). As a result, rank-ordering is ill-suited for popular compact coding strategies such as principle component analysis (PCA) [30] which advocate dimensionality reduction. Second, data sorting is required to establish rankorderings, which must be performed each time the geometry of the underlying image lattice is modified. This imposes an O (N log N) computational complexity requirement on similarity measurement, which is unattractive, particularly for correspondence techniques based on searching the space of geometrical transforms relating images. Invariant feature correspondence is a widely used, computationally efficient approach designed specifically to avoid an explicit geometrical search. Invariant feature correspondence involves first detecting image regions of interest, typically by identifying extrema in an image scale-space. The scale-space can generally be defined according to a variety of image measurements, such as Gaussian derivatives, image phase or entropy [31]. Once identified, invariant regions are geometrically normalized according to similarity or affine transforms and encoded by image descriptors, after which they can be matched between images without requiring an explicit geometrical search. A wide range of techniques can be used to represent or describe invariant feature image measurements, e.g. differential invariants, steerable filters, principal components. While earlier research focused on compact, dimensionality-reduced descriptors such as principal components, recent interest has turned to high-dimensional representations based on histograms of localized gradient orientations, e.g. the SIFT descriptor. Detailed comparisons have shown these representations to be highly effective in terms of distinctiveness when matching images of both planar surfaces and 3D objects [32].
REFERENCES


