SIFT-Rank: Ordinal Description for Invariant Feature Correspondence

Matthew Toews and William Wells III
Harvard Medical School, Brigham and Women's Hospital
{mt,sw}@bwh.harvard.edu

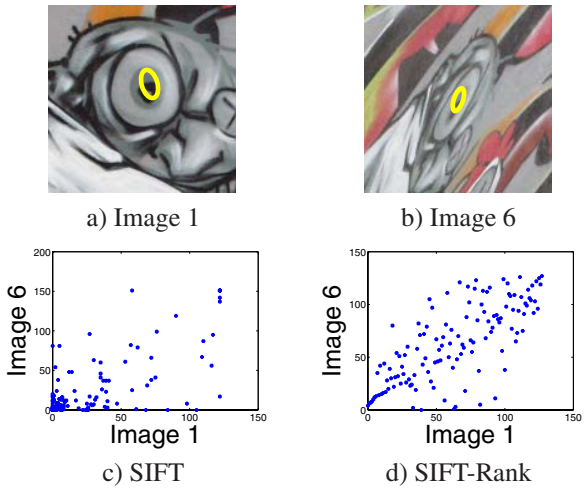
Abstract

This paper investigates ordinal image description for invariant feature correspondence. Ordinal description is a meta-technique which considers image measurements in terms of their ranks in a sorted array, instead of the measurement values themselves. Rank-ordering normalizes descriptors in a manner invariant under monotonic deformations of the underlying image measurements, and therefore serves as a simple, non-parametric substitute for ad hoc scaling and thresholding techniques currently used. Ordinal description is particularly well-suited for invariant features, as the high dimensionality of state-of-the-art descriptors permits a large number of unique rank-orderings, and the computationally complex step of sorting is only required once after geometrical normalization. Correspondence trials based on a benchmark data set show that in general, rank-ordered SIFT (SIFT-Rank) descriptors outperform other state-of-the-art descriptors in terms of precision-recall, including standard SIFT and GLOH.

1. Introduction

Automatically establishing correspondences in large, unordered sets of images is a central challenge in many computer vision applications, including wide-baseline stereo [23], content-based image retrieval [17], object class modeling [3] and medical image analysis [21]. Local invariant image features are widely used for correspondence [11, 14], as they can be efficiently extracted and matched between images without an explicit search over the parameters of geometrical deformations relating different images. Correspondence involves measuring the similarity of image descriptors associated with extracted features, typically using computationally efficient, element-wise measures such as the Euclidean or chi-squared distances. These measures assume a linear relationship between descriptors, and it is thus crucial to normalize descriptors in the presence of imaging non-linearities, including illumination changes, sensor saturation, etc. Currently, data normalization is often achieved via ad hoc scaling and thresholding strategies [11].

This paper investigates ordinal description as a means of normalizing invariant feature image measurements for correspondence. Ordinal description is a general strategy which compares sets of data measurements in terms of their rankings in an ordered array [20, 9, 28, 12], instead of the raw measurement values themselves. A rank-ordering is a normalized representation which is invariant under arbitrary monotonic deformations of the underlying data, and therefore offers a simple, non-parametric solution for coping with imaging non-linearities. Ordinal description is particularly well-suited for high-dimensional invariant feature descriptors, where it incurs no additional online computational cost, yet significantly improves the precision-recall the popular scale-invariant transform (SIFT) descriptor [11] as shown in Figure 1. While ordinal techniques are well-studied [20, 9, 28, 1, 12], their suitability in the context invariant feature description has been overlooked to date.



a) Image 1 b) Image 6

c) SIFT d) SIFT-Rank

Figure 1. The effect of ordinal descriptor normalization. Images a) and b) show a pair of corresponding affine-invariant features (yellow elliptical regions) extracted in images acquired from different viewpoints [15]. Standard and rank-ordered SIFT descriptor values associated with this feature pair are plotted in c) and d), respectively. Note the more prominent linear trend in d) due to rank-ordering. Correspondence for this pair is incorrect using standard SIFT but correct using SIFT-Rank descriptors.

The remainder of this paper is organized as follows. Section 2 reviews related work in image similarity measurement and invariant feature correspondence. Section 3 reviews theory and techniques relating to ordinal description. Section 4 presents experiments in which several popular descriptors and their ordinal forms are compared on standard benchmark database, where rank-ordered SIFT descriptors generally outperform all others tested in terms of precision-recall, including standard SIFT.

2. Related Work

Correspondence involves identifying regions in different images which arise from the same underlying surface, object or class of objects in the real world, based on the similarity of their image content. Similarity measures can be generally described in terms of assumptions made regarding the relationship between image measurements (i.e. intensities or features thereof) in different images [19]. General measures such as mutual information [24] make few assumptions regarding the relationship between measurements, however are computationally expensive. Measures assuming linear relationships such as the Euclidean distance, correlation and chi-squared distance [29] are often used in computer vision applications. Linear measures are computationally efficient, operating in $O(N)$ time complexity where N is the number of image measurements. They suffer the drawback that the assumption of linearity is often invalid, as image measurements arising from the same 3D surface can vary in a non-linear manner, due to the interaction between surface reflectance, changes in sensor geometry and illumination, or sensor saturation, etc. Measurements must therefore be normalized, and ad hoc techniques are often used. For instance, gradient-based SIFT features are first scaled to unit length, thresholded to an empirically-determined level, then re-scaled [11], providing invariance to affine illumination deformations and resistance to changes in gradient magnitude.

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 [20] or the Kendall coefficient [9]. In the context of image similarity, several authors propose comparing image windows in terms of the rankings of sorted image intensities [28, 1]. In a similar vein, Luo et al. use the Kemeny-Snell measure [8] as an image similarity metric [12], 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) [22] which advocate dimensionality reduction. Second, data sorting is required to establish rank-orderings, 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 [14, 15] 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 [11, 13], image phase [2] or entropy [7]. Once identified, invariant regions are geometrically normalized according to similarity [13, 11] or affine transforms [14] 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 [10], steerable filters [5], principal components [22]. 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 [11]. Detailed comparisons have shown these representations to be highly effective in terms of distinctiveness when matching images of both planar surfaces [15] and 3D objects [16]. Concerted efforts have been made to improve on gradient orientation-based descriptors, using alternative spatial sampling strategies [15], PCA encoding [26], learning methods [25, 6], alternative similarity measures such as the earth mover's distance [18], and alternative image representations prior to SIFT computation [4].

Ordinal descriptors are particularly well-suited for local invariant features. The high dimensionality of state-of-the-art descriptors (i.e. SIFT, 128 dimensions) is sufficient to support a large pool of unique rankings and discriminative correspondence. Furthermore, as invariant features are geometrically normalized only once during feature extraction,

sorting is only required once. While ordinal description offers a computationally efficient, non-parametric means of normalizing invariant feature measurements and achieving invariance to non-linear deformations, it has nonetheless been overlooked in the computer vision literature. In a comprehensive evaluation of invariant feature descriptors [15], ordinal techniques are mentioned but were ultimately untested.

3. Ordinal Description

Ordinal description considers descriptor vector elements in terms of their ranks in an array sorted according measurement values. Let $\bar{x} = \{x_1, \ldots, x_N\}$ denote a vector of N unique, scalar-valued image measurements. Ordinal description begins by transforming $\bar{x}$ to its rank-order form $\bar{r} = \{r_1, \ldots, r_N\}$. The rank-order value r_i corresponding to measurement value x_i is defined as:

$$r_i = |\{x_k : x_k \leq x_i\}|. \tag{1}$$

The rank-order transform can be achieved in $O(N \log N)$ time complexity by sorting $\bar{x}$, then setting element r_i according to the index of x_i in the sorted vector. Each rank-ordered vector $\bar{r}$ is a permutation of the set of integers $\{1, \ldots, N\}$. There are a total of $N!$ unique rank-ordered vectors, each of normalized squared length $||\bar{r}||^2 = \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$.

Standard element-wise measures can be used to evaluate the similarity of different rank-ordered vectors $\bar{r}$ and $\bar{r}'$ in $O(N)$ time complexity. The squared Euclidean distance between $\bar{r}$ and $\bar{r}'$ lies on the range $[0, \frac{N(N^2-1)}{3}]$. The Spearman correlation coefficient ρ [20] is defined as:

$$\rho(\bar{r}, \bar{r}') = 1 - \frac{6 \sum_{i=1}^{N} (r_i - r_i')^2}{N(N^2 - 1)}, \tag{2}$$

mapping the squared Euclidean distance to the range $-1 \leq \rho \leq 1$. Alternatively, similarity can be evaluated based on relative orderings of rank-ordered elements. The Kendall coefficient τ [9] is defined as follows:

$$\tau(\bar{r}, \bar{r}') = \frac{2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} s(r_i - r_j, r_i' - r_j')}{N(N-1)}, \tag{3}$$

$$s(a, b) = \begin{cases} 1 & \text{if } sign(a) = sign(b) \\ -1 & \text{otherwise} \end{cases}$$

where $-1 \leq \tau \leq 1$. Considering relative orderings is generally of $O(N^2)$ time complexity, although a method was proposed to reduce the time complexity [12]. Note that $\tau(\bar{r}, \bar{r}') = \tau(\bar{x}, \bar{x}')$, as the relative orderings of ranks and measurements are equivalent.

Ordinal description effectively imposes an equivalence relation between any two measurement vectors $\bar{x}$ and $\bar{x}'$ whose elements differ by an arbitrary monotonic function of x. As the number of unique rank-ordered vectors is $N!$, ordinal description is less effective for descriptors with small numbers of elements, e.g. differential invariants [10] or steerable filters [5] as used in [15], as the number of unique rank-ordered vectors is too small to allow distinctive correspondence. It is well-suited for high dimensional representations such as the SIFT [11] and GLOH (gradient location and orientation histogram) [15] descriptors, where the number of unique orderings is large.

An important practical issue is accounting non-unique measurement values, i.e. $x_i = x_j, i \neq j$, which result in tied rankings. In representations such as SIFT, for example, several histogram bins may be zero and thus tied in rank. Such ties in ranking can be broken by ordering tied elements according to their rankings in a vector of expected measurement values. Such a vector can be generated by averaging a large database of descriptors off-line.

4. Experiments

The goal of experimentation is to compare standard and ordinal descriptors in the context of local invariant feature correspondence. The hypothesis is that ordinal descriptors will result in improved correspondence, as they offer invariance to non-linear (monotonic) measurement deformations that may be present. Trials involve three different feature descriptors based on image gradient orientations: SIFT [11], PCA-SIFT [26] and GLOH [15]. The SIFT and GLOH descriptors have be shown to be highly effective for correspondence [15, 16], and all three descriptors are of relatively high dimensionality, making them suitable candidates for ordinal description. The SIFT descriptor [11] consists of histograms of image gradient orientations, where 8-dimensional orientation gradient histograms are sampled in a 4x4 spatial grid centered on the image feature, resulting in a 128-valued feature vector. The GLOH descriptor is also based on histograms of gradient orientation, however a log-polar spatial sampling scheme of 8 angular and 3 radial increments is used, along with 16-dimensional orientation gradient histograms. This results in a 272-valued vector, which is projected onto a 128-dimensional basis of principal components, resulting in a 128-valued feature vector of coefficients. The PCA-SIFT descriptor projects a 3,042-valued vector of image gradients within a 39x39 image window onto a 36-dimensional basis of principal components, resulting in a 36-valued feature vector of coefficients.

Experiments consist of correspondence trials in which invariant features are extracted and matched between pairs of images, using the experimental methodology, public data set and software of Mikolaczyk et al [15]. The data consists of images of eight different scenes, where each scene is imaged six times, with one reference image I_1 and five images $I_2 \ldots I_6$ acquired over successive increments of a particular

image deformation, including zoom/rotation, viewpoint, illumination, blur and JPEG compression noise. Each image pair is related geometrically via a known planar homography, and correspondence correctness can be evaluated according to the overlap between automatically identified correspondences and ground truths [15]. Correctly corresponding features are defined here as those whose regions in the image overlap by more than 50% with ground truth.

Trials are performed between I_1 and I_6, the case of most extreme deformation, to compare descriptors in the most challenging correspondence scenarios. Correspondence for each image pair I_1 and I_6 is performed as follows: features are first extracted in both images, using an affine invariant detector which identifies elliptical image regions in the presence of affine deformations. Note that a variety of detectors could be used, however descriptor performance has been found to be relatively independent of the particular detector used [15], the affine-invariant Harris detector [14] is used here. Then for each feature in I_1, the 1st and 2nd nearest neighbor features in I_6 are identified, based on the descriptor and similarity measure used. Descriptor similarity is measured according to the Euclidean distance between 1) standard gradient-based descriptors as is most common in the literature [11, 15, 16] and 2) rank-ordered descriptors, in a manner similar to the Spearman correlation coefficient ρ [20]. Correspondences are sorted according the ratio of distances between the 1st and 2nd nearest neighbor. The distance ratio serves as a measure of match distinctiveness, and is commonly used as a heuristic to identify the correspondences most likely to be correct [11, 15, 16, 27].

Figures 2 and 3 illustrate image pairs and precision-recall curves for descriptors tested. In general, rank-ordered SIFT descriptors (SIFT-Rank) outperform all other descriptors in terms of precision-recall, including standard (SIFT). In 7 of 8 pairs (all except Figure 3 b)), SIFT-Rank results in the highest absolute recall. In 7 of 8 pairs (all except Figure 3 a)), SIFT-Rank results in the highest precision at virtually all values of recall. Conversely, rank-ordered PCA-SIFT (PCA-Rank) is significantly worse than standard PCA-SIFT (PCA). This may be due to the relative low-dimensionality of the PCA-SIFT descriptor, or to the fact that the rank-ordering of PCA-SIFT features is highly correlated with the magnitudes of principal component variances, and thus the rank-orderings of different features may be significantly correlated. Although GLOH features are also encoded in terms of principal components, the performances of both GLOH and GLOH-Rank descriptors are similar, possibly because the larger size of the GLOH descriptor offers improved discrimination of ranks.

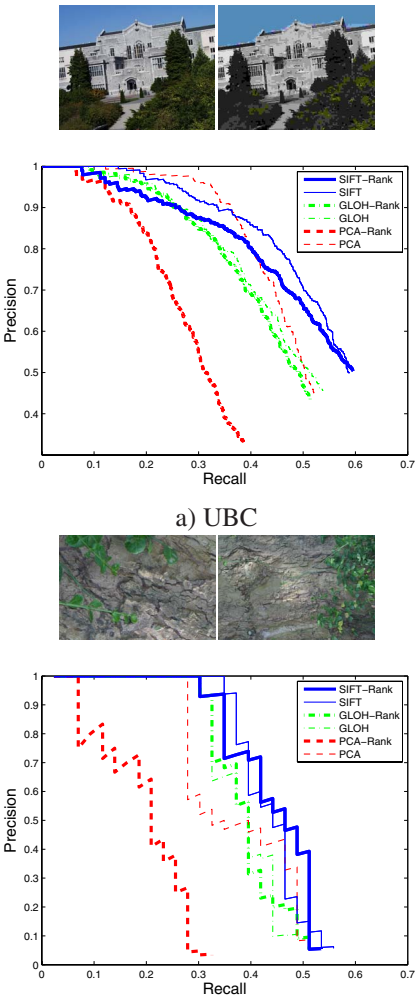


Figure 3. The two of eight image pairs for which all descriptors except PCA-Rank result in similar performance in terms of precision-recall. Images in pair UBC a) differ by JPEG compression noise and in Bark b) differ by a rotation and zoom.

5. Discussion

This paper presents a first investigation into ordinal descriptors for invariant feature correspondence. Ordinal description is a meta-technique which encodes measurements in terms of their rankings in an ordered array rather than their raw values, and thus provides invariance under arbitrary monotonic measurement deformations. Ordinal descriptors are particularly well-suited for local invariant feature correspondence, as the high-dimensionality of popular gradient orientation-based descriptors supports a large number unique rankings, and the computation of descriptor similarity is fast, since the computationally expensive step of descriptor sorting is only required once off-line after geometric feature normalization.

Experimental trials on a diverse set of public images show that rank-ordered SIFT descriptors generally result in a precision-recall characteristic superior to any other descriptor tested, including standard SIFT. Similar results not