Multi-scale Phase-based Local Features

Gustavo Carneiro and Allan D. Jepson
Department of Computer Science
University of Toronto ,Toronto, ON, Canada.
{carneiro,jepson}@cs.utoronto.ca

Abstract

Local feature methods suitable for image feature based object recognition and for the estimation of motion and structure are composed of two steps, namely the 'where' and 'what' steps. The 'where' step (e.g., interest point detector) must select image points that are robustly localizable under common image deformations and whose neighborhoods are relatively informative. The 'what' step (e.g., local feature extractor) then provides a representation of the image neighborhood that is semi-invariant to image deformations, but distinctive enough to provide model identification. We present a quantitative evaluation of both the 'where' and the 'what' steps for three recent local feature methods: a) phase-based local features [2], b) differential invariants [14], and c) the scale invariant feature transform (SIFT) [9]. Moreover, in order to make the phase-based approach more comparable to the other two approaches, we also introduce a new form of multi-scale interest point detector to be used for its 'where' step. The results show that the phase-based local features lead to better performance than the other two approaches when dealing with common illumination changes, 2D rotation, and sub-pixel translation. On the other hand, the phase-based local features are somewhat more sensitive to scale and large shear changes than the other two methods. Finally, we demonstrate the viability of the phase-based local feature in a simple object recognition system.

1 Introduction

Local feature matching methods for view-based object recognition and structure and motion estimation have received a great deal of attention lately. The extraction of local features is performed in two steps: a) a 'where' step involving an interest point detector, and b) a 'what' step consisting of a local feature extractor. The interest point detector must select image locations that contain a high degree of information content, while being robust to common image deformations. The local feature extractor must provide a representation of such image neighborhoods that is semi-invariant to typical image deformations, yet highly distinctive to afford identity information.

We are particularly interested in methods that match robustly detectable, highly informative and relatively sparse features. Rao and Ballard [12] explore the use of such local features for recognizing human faces by using principal component analysis (PCA) to reduce the dimensionality of localized natural image patches at multiple scales. In [11], Nelson presented a technique to automatically extract a geometric description of an object by detecting semi-invariants at localized points. A new concept was presented by Schmid and Mohr [14], where the authors use a set of differential invariants extracted from interest points. In [9] Lowe presents a novel method based on local scale-invariant features detected at interest points. Shoukoufandeh et al. [16] present a multi-scale view-based representation for 3D objects such that the local characteristic scale is used to build a graph that serves to measure similarity between test image and model. In [13] the features are based on receptive field histograms, which are robust to scale, translation, 2D rotation, minor occlusions, and common brightness changes. Recently, some authors have proposed affine invariant local features [1, 8], where the region around an interest point is iteratively transformed to an affine invariant space. A brief overview of feature based methods for structure and motion estimation is provided in [17], where the authors advocate that local features should be used for image matching relations/camera geometry initialization and then followed by dense reconstruction methods

Even though there is an extensive literature in the area, there has been a lack of quantitative comparison of the recent approaches suggested for local features. In [15], a quantitative comparison of interest point detectors in terms of robustness to image deformations and distinctiveness takes place, but there is no comparison of different local feature extractors. The discriminance of interest points is also explored in [6]. Here, we propose a quantitative evaluation of the 2 steps (i.e., the 'where' and 'what' steps) separately, and we investigate the performance of the following three recent approaches: a) phase-based local features [2], b) differential invariants [14], and c) scale invariant feature transform (SIFT) [9]. Moreover, we propose a new form of multi-scale interest point detector to be used in the 'where' step of the phase-based local features in order to make the

approach robust to scale changes and, thus, more comparable to the other two approaches. The experiments conducted in this work also investigate the use of brightness renormalization for the local differential invariants, as in [13], in order to reduce the brightness sensitivity of the differential invariant approach and provide a fairer comparison.

Our results show that the phase-based local features perform better than the other two approaches when dealing with common illumination changes, 2D rotation, and sub-pixel translation. For scale and large shear changes, both SIFT and the differential invariant features lead to better results than ours. Finally, a simple system that performs object recognition is provided to demonstrate the general viability of the phase-based approach.

2 Image Deformations Studied

The image deformations considered here are: a) two types of global brightness changes, b) non-uniform local brightness variations, c) additive noise, d) scale changes, e) 2D rotation, f) shear and g) sub-pixel translation. The non-uniform global brightness changes are implemented by adding a constant to the brightness value, taking into account the gamma correction non-linearity:

$$I_h(\vec{x}) = 255 * \left[\max\left(0, \left(\frac{I(\vec{x})}{255}\right)^{\gamma} + k\right)\right]^{\frac{1}{\gamma}}, \quad (1)$$

where $\gamma = 2.2$, and $k \in [-.5, .5]$ controls the changes in brightness. The resulting image is linearly mapped to values between 0 and 255, and then quantized. The uniform brightness change is simply based on the division of gray values by a constant $c \in [1, 3]$.

For the non-uniform local brightness variations, a highlight at a specific location of the image is simulated by adding a Gaussian blob as follows: $I_h(\vec{x}) = I(\vec{x}) + 255 * G(\vec{x} - \vec{x}_0; \sigma)$, where $\sigma = 10$, $\vec{x}_0$ is a specific position in the image, and $G(\vec{x}; \sigma) = \exp(-x^2/(2\sigma^2))$. Again, the resulting image is mapped to values between 0 and 255, and then quantized. For noise deformations, we simply add Gaussian noise with varying standard deviation ($\sigma = 255 * [10^{-3}, 10^{-1}]$), followed by normalization and quantization, as above.

The geometric deformations are 2D rotations (from -90^o to $+90^o$ in intervals of 15^o), uniform scale changes (with expansion factors in the range $[0.25, 1]$), shear in the horizontal direction (so that a vertical line is perturbed by $\pm 26^o$), and sub-pixel translation (in the range [0,1]) pixel. The geometrically deformed images are quantized to $[0, 255]$ without normalization.

3 Where: The Interest Points

An interest point detector must select highly informative image locations that are robustly localizable given common image deformations. Here we extend the Harris corner detector (see [7]) which is known to be robust common illumination changes and rotation, but it is also known to be sensitive to scale changes. Therefore, we must seek a way to make the Harris corner detector robust to scale changes. We use an approach similar to the one suggested in [3], in which we check local spatial information to determine whether the current scale is appropriate.

The Harris corner detector is based on a matrix that averages the products of the first derivatives of the signal in a window, which is built as follows:

$$\mathbf{C}(\vec{x}) = \exp -\frac{x^2+y^2}{2\sigma_h^2} * \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \quad (2)$$

with $\sigma_h \geq 2.0$, and $*$ is the convolution operation. Here $I_x(\vec{x}) = G_x(\vec{x}, \sigma_c) * I(\vec{x})$, where G_x is the x-derivative of a Gaussian with mean $\vec{x}$ and standard deviation $\sigma_c = \sigma_h/2$, and similarly for I_y. The eigenvalues of this matrix, $\mu_1(\vec{x})$ and $\mu_2(\vec{x})$, represent edge strength, so a corner is an image location at which $\mu_1(\vec{x}) \geq \mu_2(\vec{x}) \geq t$, with t as a threshold. Here, the function described in [7] is substituted by

$$T(\vec{x}) = \frac{\mu_2(\vec{x})}{c + (1/2)(\mu_1(\vec{x}) + \mu_2(\vec{x}))}, \quad (3)$$

where $c = 1$ is set based on the histogram of $T(\vec{x})$ of various types of images, and $T \in [0, 1]$. The initial set of interest points is defined as $\text{In}(I_k, \sigma_c) = \{\vec{x}_i | T(\vec{x}_i) \geq 0.5\}$, where $\vec{x}_i$ is a point in image I_k.

In order to filter the initial set of interest points computed by the Harris corner detector, we utilize the procedure described in [4], where the quadrature pair filters specified in [5] are used, tuned to a specific orientation θ and scale σ_c (as mentioned before $\sigma_c = \sigma_h/2$). More specifically, let

$$R(\vec{x}, \sigma_c, \theta) = (G_2 + iH_2) * I(\vec{x}), \quad (4)$$

where $G_2(\vec{x}, \sigma_c, \theta)$ is the second directional derivative of a Gaussian, $H_2(\vec{x}, \sigma_c, \theta)$ is the approximation of Hilbert transform of G_2, and σ_c is the standard deviation of the Gaussian kernel used to derive G_2 and H_2 [1]. The complex polar representation of (4) can be written as $R(\vec{x}, \sigma_c, \theta) = \rho(\vec{x}, \sigma_c, \theta) e^{i\phi(\vec{x}, \sigma_c, \theta)}$, where $\rho(\vec{x}, \sigma_c, \theta)$ is the local amplitude information and $\phi(\vec{x}, \sigma_c, \theta)$ is the local phase information. The complex valued filter $(G_2 + iH_2)$ is bandpass with peak frequency response at $\omega_c = 2\pi/(3.918\sigma_c)$, which corresponds to a wavelength of $\lambda_c = 3.918\sigma_c$. Below we use σ_c and λ_c interchangeably to refer to the local scale.

The local frequency of the response R is defined as the spatial derivative of the phase signal [4]. In particular,

$$\phi_x(\vec{x}, \lambda_c, \theta) = \frac{\text{Im}[R^*(\vec{x}, \sigma_c, \theta) R_x(\vec{x}, \sigma_c, \theta)]}{|R(\vec{x}, \sigma_c, \theta)|^2}, \quad (5)$$

[1] These responses for all orientations θ can be computed from just seven basis images at each scale, which can be computed using a total of 14 1D convolutions.

where $R^*(\vec{x},\sigma_c,\theta)$ is the complex conjugate of $R(\vec{x},\sigma_c,\theta)$. Similarly, we can compute $\phi_y(\vec{x},\sigma_c,\theta)$, and the local frequency at the maximum energy orientation θ_M (computed as described in [5]) is

$$\omega(\vec{x},\lambda_c) = ||(\phi_x(\vec{x},\lambda_c,\theta_M),\phi_y(\vec{x},\lambda_c,\theta_M))||, \quad (6)$$

where $||.||$ denotes *Euclidean* norm. This gives a local wavelength of

$$\lambda(\vec{x},\lambda_c) = 2\pi/\omega(\vec{x},\lambda_c). \quad (7)$$

Using equations (5-7), we can compute the local wavelength $\lambda(\vec{x},\lambda_c)$ of each $\vec{x}_i \in \text{In}(I_k,\sigma_c)$. Finally, the mean wavelength at interest points at the scale λ_c, namely $\lambda_m(\lambda_c)$, is then defined as the average of all observed wavelengths $\lambda(\vec{x}_i,\lambda_c)$, for which $\lambda(\vec{x}_i,\lambda_c) \in [\lambda_c/2, 2\lambda_c]$, where $\vec{x}_i$ ranges over a large collection of interest points.

We are particularly interested in interest points that are stable in terms of local phase information. Following [4] these points can be identified as those whose local frequency is sufficiently close to the filter tuning. That is, we filter the initial set $\text{In}(I_k,\sigma_c)$ as follows

$$\text{In}_f^{\lambda_c}(I_k) = \Big\{\vec{x}_i | \vec{x}_i \in \text{In}(I_k,\sigma_c), \text{and} \\ \lambda(\vec{x}_i,\lambda_c) \in \Big[\frac{\lambda_m(\lambda_c)}{\sqrt{2}}, \sqrt{2}\lambda_m(\lambda_c)\Big]\Big\}, \quad (8)$$

where $\lambda_c = 3.918\sigma_c$, as above.

3.1 Comparison of Interest Points

Here we study the stability of multi-scale interest points with respect to the deformations described in section 2. We denote the deformed test image by $\tilde{I}_{k,d}$, where $\{I_k\}_{k=1}^{100}$ is a database of images, and d denotes the deformation applied to I_k to form $\tilde{I}_{k,d}$. Thus the set of interest points of deformed test image $\tilde{I}_{k,d}$ is $\text{In}_f^{\lambda_d}(\tilde{I}_{k,d})$, where we use $\lambda_d = 8$. The closest transformed scale for the undeformed test image I_k is then defined as

$$\lambda_c = argmin_{\lambda_c \in \Lambda_o}\{|\lambda_c - \frac{\lambda_d}{\kappa(d)}|\}, \quad (9)$$

where $\Lambda_o = \{4, 4\sqrt{(}2), 8, 8\sqrt{(}2), 16, 16\sqrt{(}2), 32\}$, and the expansion factor $\kappa(d) \in [0.25, 1]$ for scale changes, and otherwise $\kappa(d) = 1$. Finally, the spatial warp for the deformation d is denoted by

$$\vec{x}_j = M(d)\vec{x}_i + \vec{b}(d), \quad (10)$$

where $\vec{x}_i$ are the original image coordinates and $\vec{x}_j$ the deformed image coordinates.

In order to assess the interest point detector performance, two measures are computed, namely the precision and recall rates. The precision rate measures the probability that an interest point detected in a deformed test image $\tilde{I}_{k,d}$ is actually an interest point in the corresponding database image, I_k. That is

$$P_{rate}(d) = \frac{TP(d)}{TP(d)+FP(d)}, \quad (11)$$

where

$$TP(d) = \Big|\{(\vec{x}_j,d,k) \mid \vec{x}_j \in \text{In}_f^{\lambda_d}(\tilde{I}_{k,d}) \text{ and } \\ \exists \vec{x}_i \in \text{In}_f^{\lambda_c}(I_k) \text{ s.t. } \\ ||M(d)\vec{x}_i + \vec{b}(d) - \vec{x}_j|| < \epsilon\}\Big|,$$

$$FP(d) = \Big|\{(\vec{x}_j,d,k) \mid \vec{x}_j \in \text{In}_f^{\lambda_d}(\tilde{I}_{k,d}) \text{ and } \\ \neg\exists \vec{x}_i \in \text{In}_f^{\lambda_c}(I_k) \text{ s.t. } \\ ||M(d)\vec{x}_i + \vec{b}(d) - \vec{x}_j|| < \epsilon\}\Big|,$$

where I_k for $k = 1,\ldots,100$ ranges over the image database and $\epsilon = 2.0$ pixels. On the other hand, the recall rate measures the probability of finding an interest point in a deformed image $\tilde{I}_{k,d}$, given that it is detected in the corresponding database image I_k. That is,

$$R_{rate}(d) = \frac{TP(d)}{TP(d)+FN(d)}, \quad (12)$$

where

$$FN(d) = \Big|\{(\vec{x}_i,d,k) \mid \vec{x}_i \in \text{In}_f^{\lambda_d}(I_k) \text{ and } \\ \neg\exists \vec{x}_j \in \text{In}_f^{\lambda_c}(\tilde{I}_{k,d}) \text{ s.t. } \\ ||M(d)\vec{x}_i + \vec{b}(d) - \vec{x}_j|| < \epsilon\}\Big|.$$

Using P_{rate} and R_{rate}, we provide a comparison in Fig. 1 between our interest point detector (solid curves) and the ones described in [10] (Harris-Laplacian, dotted curve), and in [9] (difference-of-Gaussian, dashed curve). The Harris-Laplacian is based on the interest points detected using the Harris corner detector at several scales and the scale selection is done using local maxima of the normalized Laplacian in image and scale spaces. On the other hand, the difference-of-Gaussian uses difference of images convolved with the Gaussian filter at neighboring scales and local maxima and minima in image and scale spaces are selected as interest points. Another important observation is that we respect the image space sampling originally described in the papers [10] (no subsampling in coarser scales) and [9] (subsampling = $\lfloor\lambda_c/4\rfloor$). Our interest point detector also uses subsampling = $\lfloor\lambda_c/4\rfloor$. In general, our detector performs better than the other two considered in terms of common illumination changes, 2D rotation, and sub-pixel translation while giving comparable results for large shear changes.

We note that the frequency of interest point detection at $\lambda_c = 8$ (i.e., number of interest points detected divided

by the number of original pixels in the image) on the deformed images was rather different, namely our detector selected 1.48% of the image points as interest points, while the Harris-Laplacian detected 0.20%, and the difference-of-Gaussian detected 0.20%. A significant observation here is that we do not apply non-maximum suppression on the filtered interest point map $\mathrm{In}_f^{\lambda_c}(I_k)$, so we are likely to have small clusters of interest points which may provide a bias in our favor in terms of the precision and recall curves. However, this is not a problem as long as we still have high precision and recall rates, and the regions detected are still distinctive (see section 4). The tradeoff is a highly populated database of features that will demand a more efficient search scheme and a larger amount of memory space.

4 What: The Local Features

Ideally, suitable local features must extract a representation for local image data with the following two properties: a) be complex enough to provide strong information about a specific location of an image; and b) be relatively stable to changes in the object configuration, so that small transformations do not significantly affect the identification process. In this section we consider the problem of finding good candidates for such local features.

4.1 Phase and Amplitude Information

We use the phase-based local feature (described in detail in [2]), which is a complex representation of local image data that is obtained through the use of the quadrature pair filters described in section 3. In order to make the system robust to brightness changes, any sufficiently large amplitude is saturated according to

$$\tilde{\rho}(\vec{x},\sigma,\theta) = 1 - e^{\frac{-\rho(\vec{x},\sigma,\theta)^2}{2\sigma_p^2}}, \qquad (13)$$

where $\sigma_p = 1.0$. Therefore, whenever the local amplitude is high enough, the saturated amplitude is roughly constant.

4.2 Local Image Description

Since a single pixel does not provide a distinctive response we consider several sample points, say $\{\vec{x}_{i,m}\}_{m=1}^{M}$, taken from a region around each interest point, $\vec{x}_i$. We use the sampling pattern depicted in Fig. 2, with the center point $\vec{x}_{i,1}$ denoting the specific interest point $\vec{x}_i$ (the reasons for selecting this particular sampling pattern are discussed further below). It is worth seeing that the fixed radius is in terms of the subsampling grid for the scale at which the feature is being evaluated. In terms of the original image pixels, this radius increases proportionally to the scale σ_c (i.e. image subsampling = $\lfloor \lambda_c/4 \rfloor$). At each spatial sample point $\vec{x}_{i,m}$ the filters are steered to N equally spaced orientations,

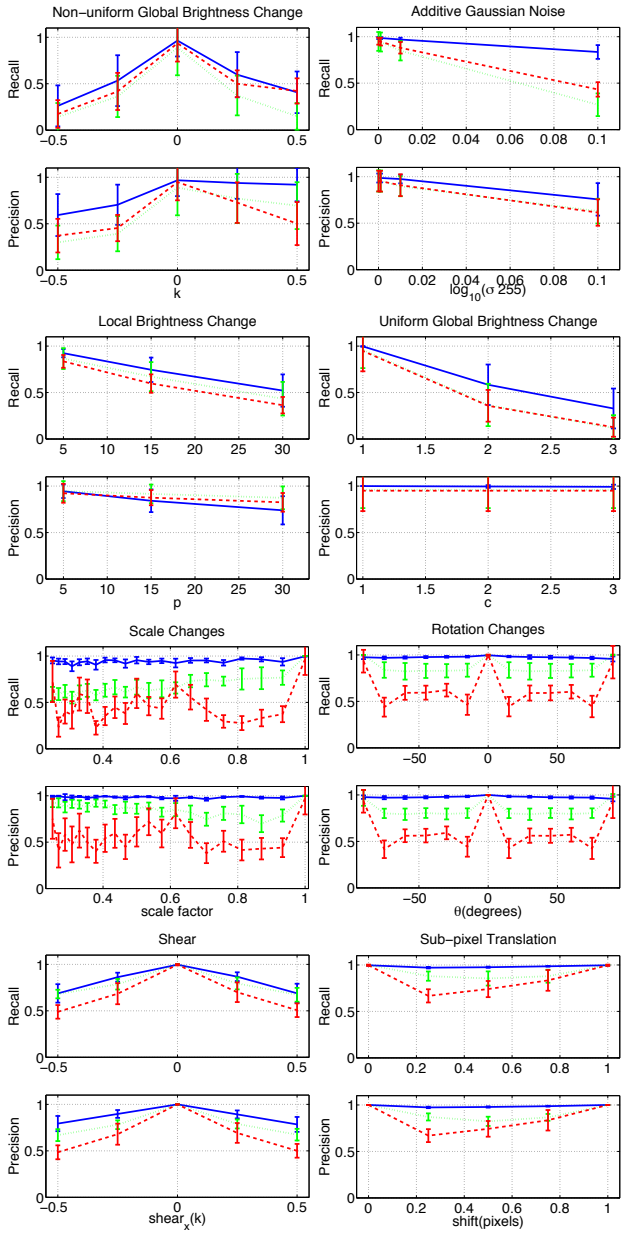

Figure 1: Comparison between the scale robust interest point detector described above (solid line), and the interest point detectors Harris-Laplacian (dotted curve) and difference-of-Gaussian (dashed curve).

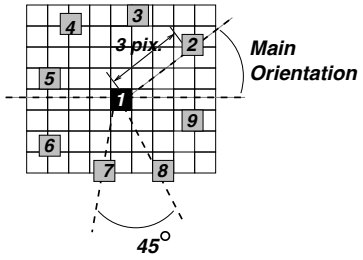

Figure 2: Configuration of local descriptor for $M = 9$.

namely

$$\theta_n(\vec{x}_i) = \theta_M(\vec{x}_i) + (n-1)\frac{180^o}{N}, \text{ for } n = 1, \ldots, N. \quad (14)$$

Here $\theta_M(\vec{x}_i)$ is the main orientation of the pixel computed as described in [5], except we use the sign of the imaginary response of the filter steered to this orientation to resolve a particular direction (i.e. mod 360^o) from this orientation. Notice that this main orientation $\theta_M(\vec{x}_i)$ determines both the orientations that the filters are steered to and the positions of the sample points along the circle centered on the interest point $\vec{x}_i$ (see Fig. 2).

The feature vector $\vec{F}(\vec{x}_i)$ at scale σ_c has individual components specified by the saturated complex filter responses. We use $\tilde{R}_{inm} = \tilde{\rho}(\vec{x}_{i,m}, \sigma_c, \theta_n(\vec{x}_i))e^{i\phi(\vec{x}_{i,m}, \sigma_c, \theta_n(\vec{x}_i))}$ to denote the filter response evaluated at $\vec{x}_{i,m}$ and steered to orientation $\theta_n(\vec{x}_i)$, for $n = 1, \ldots, N$, and $m = 1, \ldots, M$. Together these responses form the NM-dimensional complex feature vector $\vec{F}(\vec{x}_i)$. An empirical study was used to select the following feature vector configuration: a) number of sample points, $M = 9$; b) number of steering directions, $N = 4$; c) radius of the circle, $l = 3$. This configuration was found to provide a reasonable tradeoff between expressiveness and stability.

4.3 Phase Correlation

The similarity between local features is computed using phase correlation since this is known to provide some stability to typical image deformations such as brightness changes and near identity image warps [4]. The similarity measure for our feature vector is the normalized phase correlation

$$S(\vec{F}(\vec{x}_i), \vec{F}(\vec{x}_j)) = \frac{\left|\sum_{n,m} \tilde{R}_{inm} \tilde{R}^*_{jnm}\right|}{1 + \sum_{n,m} \left|\tilde{R}_{inm}\right| \left|\tilde{R}_{jnm}\right|}, \quad (15)$$

where $\tilde{R}^*_{jnm}$ is the complex conjugate of $\tilde{R}_{jnm}$ (see section 4.2). The reason for adding the 1 in the denominator above is to provide a low-amplitude cut-off for the normalization. This results in similarity values $S(\vec{F}(\vec{x}_i), \vec{F}(\vec{x}_j)) \in [0, 1]$.

4.4 Scale and Rotation Semi-Invariant Feature Vectors

While the feature vector $\vec{F}(\vec{x}_i)$ described above is semi-invariant to common illumination changes, image translations and rotations, it is only locally robust to scale changes [2]. In order to achieve semi-invariance to scale changes, we consider sampling the scale specific features at a discrete set of scales, say Λ_o as defined in section 3.1. That is, to form a feature database we compute the feature vectors $\{\vec{F}(\vec{x}_i) | \vec{x}_i \in \text{In}_f^{\Lambda_o}(I_k)\}$, for each model image I_k. Given a feature vector from a test image, we search the entire database for similar features, irrespective of the specific scales at which they were observed in the test and model images. The specific scales of matching features then provides some information about the relative scales of the target in the test and the database images.

4.5 Alternative Local Features

Here we are evaluating two things: a) the feature robustness in terms of image deformation, and b) whether the interest points selected by the 'where' step provides highly informative image locations, as encoded by local features. In order to assess the effectiveness of the multi-scale phase-based local features, a comparison with the feature vectors used in [14] and in [9] is provided below.

The differential invariant [14] characterizes the neighborhood of an interest point by a set of its derivatives which is theoretically proven to be invariant to rotation. SIFT features [9] are based on image gradient histogramming processed at several orientations.

A normalized version of the differential invariant feature vector [14] is used due to its high sensitivity to common illumination changes of the unnormalized version [2]. Two types of normalization were applied, namely: a) energy normalization, as described in [13]; and b) dividing the local jets by the norm of the Gaussian filtered image $N(\vec{x}) = G(\vec{x}, \sigma_c) * I(\vec{x})$. Therefore, the normalization is achieved by dividing every term $L_u(\vec{x}) = (\sigma_c)^u G_u(\vec{x}, \sigma_c) * I(\vec{x})$ by $N(\vec{x})$, where $u \in [0, 3]$ is the differentiation order. We report the results for only the latter normalization since it provided the best results. After this normalization, the first component of the feature vector is just $\tilde{L}(\vec{x}) = L(\vec{x})/N(\vec{x}) = 1$. This component is uninformative, and is therefore deleted. An important remark is that, for the differential invariant features, we only consider the case where no subsampling is applied at coarser scales; as opposed to the phase-based and SIFT features where subsampling = $\lfloor \lambda_c/4 \rfloor$. Moreover, we found it important to add Gaussian noise during the training phase of the differential invariants (i.e. in the computation of the covariance matrix for the Mahalanobis distance) in order to reduce its sensitivity to noise.

Another critical consideration is the dimensionality of each local feature vector, which is an issue that directly affects the database search process. The phase-based local feature has 36 dimensions in the complex domain, but the saturated amplitude could be compressed to a few bits. The differential invariant feature vector [14] has 8 dimensions, and the SIFT local feature [9] has 160 dimensions. We expect the search time for larger dimensional features to be higher. Finally, the interest point detector described in [10] is used for the differential invariants; the SIFT features were extracted using the interest point detector in [9]; while the phase-based local features were extracted from interest

points detected as explained in section 3.

4.6 Comparison of Local Features

The comparison tests utilize the database of 100 test images along with one model image database consisting of 12 images. None of the test images were included in database.

For the experiments below, we build a database of random features extracted from interest points detected in each of the 12 images at the scales $\lambda = \{4, 8, 16\}$, and we have, on average, 1000 features stored in the database, which are used in the false positive rate calculation. For the true positive rate, we compute the multi-scale features over all the undeformed test images I_k, $1 \leq k \leq 100$, providing the feature database $\{\vec{F}(\vec{x}_i) | \vec{x}_i \in \text{In}_f^{\Lambda_o}(I_k)\}$, with Λ_o as in (9). For each test image k and each image deformation d (as described in section 2) we obtain a deformed test image, say $\tilde{I}_{k,d}$. From this deformed image $\tilde{I}_{k,d}$ we compute the set of features $\{\vec{F}(\vec{x}_j) | \vec{x}_j \in \text{In}_f^{\lambda_d}(\tilde{I}_{k,d})\}$, with $\lambda_d = 8$.

Given the 8 types of image deformations studied, the comparison is based on the Receiver Operating Characteristics (ROC) curves where the detection rate vs false positive rate is computed for each of the local feature types. In order to define these rates, let $\vec{x}_i$ be an interest point in an undeformed test image. Suppose $\vec{x}_i^d = M(d)\vec{x}_i + \vec{b}(d)$ denotes the transformed position of this interest point in the deformed test image, according to the spatial deformation d. The detection rate (DT) is then defined as

$$DT(d) = \frac{SM(d)}{IM(d)} \tag{16}$$

where

$$SM(d) = \Big| \{(\vec{x}_i, d, k) \mid \vec{x}_i \in \text{In}_f^{\lambda_c}(I_k) \text{ and } \exists \vec{x}_j \in \text{In}_f^{\lambda_d}(\tilde{I}_{k,d}) \text{ s.t. } \|\vec{x}_j - \vec{x}_i^d\| < \epsilon \text{ and } S(\vec{F}(\vec{x}_i), \vec{F}(\vec{x}_j)) > \tau\} \Big|$$

and

$$IM(d) = \Big| \{(\vec{x}_i, d, k) \mid \vec{x}_i \in \text{In}_f^{\lambda_c}(I_k) \text{ and } \exists \vec{x}_j \in \text{In}_f^{\lambda_d}(\tilde{I}_{k,d}) \text{ s.t. } \|\vec{x}_j - \vec{x}_i^d\| < \epsilon\} \Big|.$$

Here ϵ was fixed at 2.0 pixels, while τ was varied to generate the ROC curves. Similarly, given a feature vector $\vec{F}(\vec{x}_j)$ in the deformed test image, a false positive is defined by the presence of a similar feature vector $\vec{F}(\vec{x}_t)$ in the database (i.e. $S(F(\vec{x}_j), F(\vec{x}_t)) > \tau$). The false positive rate (FP) is defined to be the number of these false positives divided by the product of the number of deformed test image features evaluated and the number of features in the database.

In Fig. 3 we show the detection rate for thresholds τ at which the false positive rate is held fixed at 0.01. Note that

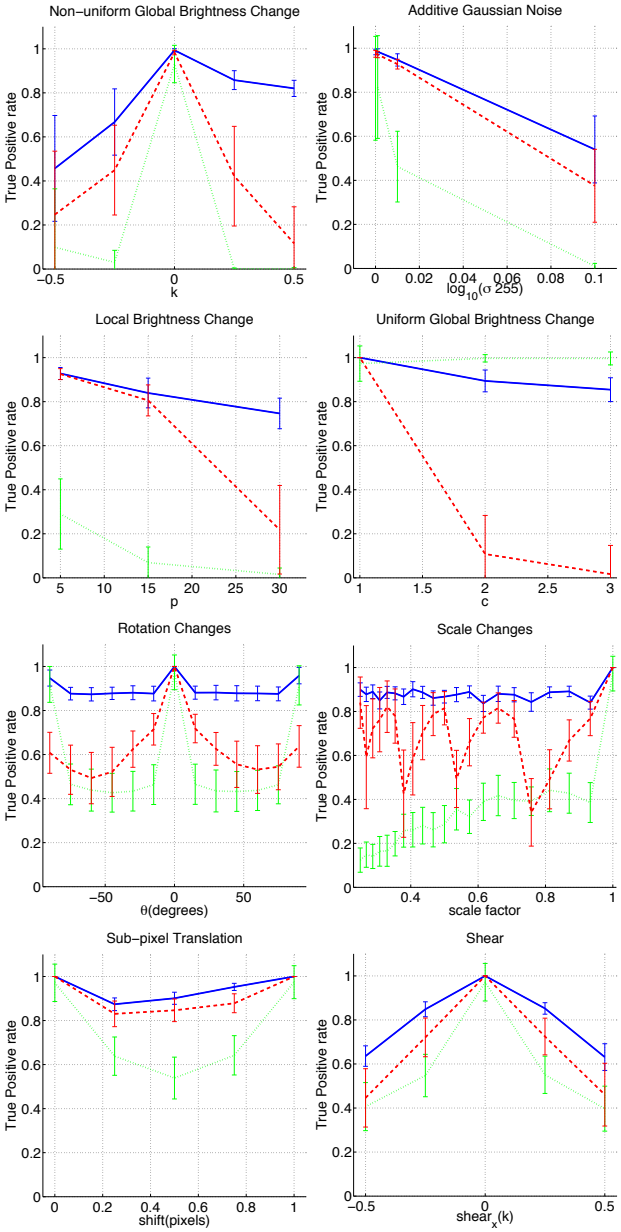

Figure 3: All image deformations with a false positive rate fixed at 0.01 and computing the detection rate for varying amount of change. Here, the phase-based, differential invariant and SIFT features are represented by the solid, dotted and dashed lines, respectively. The vertical axis represents detection rate and the horizontal axis shows the amount of variation

Figure 4: Top left image: segment of the image selected by the user to define the "boy's left eye" model. Other images: recognizing the model over a sequence of 100 images (only 5 are shown). The light points inside the distorted rectangles represent the interest points used for the best similarity match.

Figure 5: Top left image: segment of the image selected by the user to define the "tetley box" model. Remaining images: the light points inside the rectangles represent the interest points used for the best similarity match.

the phase-based feature (solid line) gives more robust and more distinctive results than the differential feature (dotted line) and SIFT (dashed line) in terms of illumination changes, 2D rotation and sub-pixel translation; while for scale and large shear changes, we observe that the other two methods produce better results than the phase-based feature. In order to observe a comparable result for scale change, the phase-based local feature would need to use a denser sampling in scale space.

We also conducted the same experiments, but omitted the graphs, holding the false positive rate fixed at 0.001. The comparison results are similar, but the phase-based features show a higher sensitivity to scale changes. Also, the differential invariants produced a significantly worse result than the one displayed in Fig. 3. This is likely due to its lower dimensionality (8 dimensions) compared to the other methods.

5 Object Recognition System

In order to demonstrate the general viability of the phase-based local feature and its robustness under real image deformations, a simple object recognition system was implemented. The system can be divided into the learning and recognition subsystems. The learning subsystem accepts an input image and requests the user to select a region from that image (see top left image of Fig. 4), and to give an identity ID to it. The features, extracted from the multi-scale interest points inside the region, are computed at 7 different scales (see Λ_o in section 4.4). Along with each feature vector, we store the model identity ID, the wavelength λ that the filter was tuned to when it was extracted, and the main orientation θ_M.

The recognition subsystem receives a test image, the wavelength to use as the filter tuning (default wavelength is 8), and the object to look for within that image. The first step is to find the nearest neighbor $\vec{F}(\vec{x}_i)$, using phase correlation described in (15), in the model database for each local feature $\vec{F}(\vec{x}_j)$ at an interest point $\vec{x}_j$ detected in the test image. Next, using random sampling and robust M-estimation from the matches where $S(\vec{F}(\vec{x}_i), \vec{F}(\vec{x}_j)) > \tau$ (here we consider $\tau = 0.6$), we compute the similarity transform which is locally optimal with respect to an error measure that takes into account the phase correlation and the Euclidean distance between matched points.

The system returns only the best possible match in terms of this error measure. It is worth mentioning that the system is not doing any tracking of the features whatsoever. Instead it is only trying to find the given model in every frame separately. The point here is to test both the 'where' (i.e., the interest point detector) and the 'what' (i.e., the local feature extractor) steps together when dealing with real image deformations.

Figure 6: Top left image: user selected "baking soda box" model. Sequence: searching the model over a series of cluttered images containing the model at different poses and partially occluded. The light points inside the distorted rectangles represent the interest points used for the best similarity match.

The first test consists of a sequence of 100 images where we are looking for the boy's left eye (see Fig. 4), and even though the sequence presents 3D rotation, lighting variations, scale changes, and non-rigid transformations, this simple object recognition system was able to correctly find the model in almost 60% of the images. The second test consists of a rotating object (see Fig. 5), and we obtained 100% true positive rate. Finally, the last test comprises 16 different images where we try to find a baking soda box (see top left image of Fig. 6). The first set of four images (second row of Fig. 6 shows 2 out of those 4 images) has the box at the same position, but with clutter (varying background); we obtain 100% true positive rate. The second set presents the same box with a small rotation in depth (third row of Fig. 6), scale change, and with clutter and occlusion; we still obtain 100% true positive rate. A larger rotation in depth in the third set of four images (Fig. 6, top right), and we obtained 50% true positive rate, and 50% false positive rate. Finally, the last set does not show the front part of the box, and we got 100% for the true negative rate.

6 Conclusions

We have presented a quantitative comparison between three approaches suggested for local features . We investigate the 'where' step (i.e., the interest point detector), and the 'what' step (i.e., the local feature extractor) separately. Furthermore, the phase-based local feature in [2] uses a new form of multi-scale interest point detector in its 'where' step so that we have a fair comparison between this approach and others based on local differential invariants [14] and SIFT [9]. The differential invariant feature is modified to make it robust to illumination changes due to its sensitiveness to those types of deformations [2]. The comparison shows that the phase-based local feature performs better than the differential invariant and SIFT features in terms of common illumination changes, 2D rotation, and sub-pixel translation. For scale and large shear changes we see that both SIFT and the differential invariants produce better results. However it is relevant to say that the robustness of the phase-based feature to scale changes can be improved by using a denser sampling in the scale space. Finally, the viability of the phase-based approach is demonstrated in a simple object recognition system.

References

[1] A. Baumberg. Reliable feature matching across widely separated views. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, South Carolina, USA, June 2000.

[2] G. Carneiro and A. Jepson. Phase-based local features. In *European Conference on Computer Vision*, Copenhagen, Denmark, May 2002.

[3] J.H. Elder and S.W. Zucker. Local scale control for edge detection and blur estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):699–716, 1998.

[4] D. Fleet. *Measurement of Image Velocity*. Kluwer Academic Publishers, 1992.

[5] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[6] D. Hall, B. Leibe, and Bernt Schiele. Saliency of interest points under scale changes. In *BMVC*, pages 646–655, Cardiff, UK, 2002.

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, Manchester, 1988.

[8] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d depth cues from affine distortions of local 2-d brightness structure. In *European Conference on Computer Vision*, pages 415–434, Stockholm, Sweden, May 1994.

[9] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, September 1999.

[10] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest point. In *The Eighth IEEE International Conference on Computer Vision*, pages 525–531, Vancouver, BC, Canada., July 2001.

[11] R. C. Nelson. Memory-based recognition for 3-d objects. In *ARPA Image Understanding Workshop*, pages 1305–1310, Palm Springs, USA, February 1996.

[12] R.P.N. Rao and D.H. Ballard. Natural basis functions and topographic memory for face recognition. In *International Joint Conference on Artificial Intelligence*, pages 10–17, 1995.

[13] B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms. In *4th European Conference on Computer Vision*, volume 1, pages 610–619, April 1996.

[14] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[15] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[16] A. Shokoufandeh, I. Marsic, and S. Dickinson. View-based object recognition using saliency maps. *Image and Vision Computing*, 17:445–460, 1999.

[17] P.H.S. Torr and A. Zisserman. Feature based methods for structure and motion estimation. In *Vision Algorithms: Theory and Practice*, pages 278–294, Corfu, Greece., September 1999.