Annals of the BMVA Vol. 2008, No. 3, pp 1–18 (2008)

SIFT Keypoint Descriptors for Range Image Analysis

Tsz-Wai Rachel Lo and J. Paul Siebert

Department of Computing Science,
University of Glasgow, Sir Alwyn Williams Building, Lilybank Gardens, Glasgow, G12 8RZ, UK
⟨{rachel|psiebert}@dcs.gla.ac.uk⟩

Abstract

This paper presents work in progress to extend the two-dimensional (2D) Scale Invariant Feature Transform (SIFT) into the 2.5 dimensional (2.5D) domain. Feature descriptors are extracted from range images of human faces and the form of these descriptors is analogous to the structure of Lowe's 2D SIFT. Lowe's descriptors are derived from the histogram of the image gradient orientations, computed over a Gaussian weighted local support region centred on each sampling (keypoint) location. We adapt this concept into the 2.5D domain by extracting the relative frequencies of the [-1,1] bounded range surface shape index and the relative frequencies of the range surface in-plane orientations simultaneously at each sampled keypoint location. Nine Gaussian weighted sub-regions, overlapped by one standard deviation, are used to sample each keypoint location and thereby construct a keypoint descriptor. Since these overlapped Gaussian sub-regions are spatially correlated, this sampling configuration reduces both the spatial aliasing and the sensitivity to small keypoint location errors in the extracted descriptors. Each histogram pair, extracted from each Gaussian weighted sub-region, is normalised and concatenated to form a feature descriptor that is tolerant to a degree of viewpoint rotational change. We have validated the current 2.5D SIFT formulation using synthetically rotated human face data over the range ±30° out-of-plane rotation, and demonstrate that our 2.5D SIFT maintains a similar matching performance to 2D SIFT applied to the (comparatively richer) intensity images of the same face.

1 Introduction

This paper presents the ongoing work to extend the SIFT algorithm [Lowe, 2004] into the 2.5D range image domain, by adapting Lowe's concept of a *keypoint descriptor* to include information about range surface local topology. We have based this new keypoint descriptor on a combination of the local surface orientation histogram and a histogram of the local surface *shape index* [Koenderink and van Doorn, 1992]. By employing range images, we propose to achieve a greater degree of invariance to illumination and 3D pose changes than could otherwise be achieved using 2D images alone.

A range image comprises a 2D matrix in which each element encodes not the intensity of the light focused on an optical imaging sensor, but the distance (or range/depth) of the nearest world surface to each element in the imaging plane [Besl, 1998]. Due to the availability of measurements in the third dimension, this imaging modality supports machine interpretation of imaged surface shape directly via differential geometry. Moreover, range images are partially invariant to lighting, pose and viewpoint changes [Gordon, 1992], which confers a number of added advantages over an analysis based on 2D images alone [Bowyer et al., 2006]. Accordingly, range images have the potential to capture surface shape variation, irrespective of illumination variations [Hesher et al., 2003]. These properties have therefore been the predominant motivation for our goal of machine classification of objects, including human faces, captured in 2.5D images as reported here.

Machine interpretation based on 3D image sensing has become more popular over the last decade, and in particular using the range images that these devices typically generate. In order to perform machine interpretation using range images, a *feature descriptor* is extracted from an appropriate sampling window, i.e. *measurement aperture*. This feature descriptor should be capable of encapsulating the predominant "shape signature" of the underlying surface to provide sufficient descriptive richness to discriminate between different types of descriptors, while retaining rotational invariance to viewpoint changes.

Differential geometry can be used to categorise surface topology and has been widely employed in the object recognition community since the 1980s; examples include work by Ittner and Jain [1985], Fan et al. [1986], Cartoux et al. [1989]. The signs of the mean (H) and Gaussian (K) curvatures have been used to segment smooth and differentiable surfaces into eight surface types [Besl and Jain, 1985]; e.g. Lee and Milios [1990] segment range images of the human face and match convex regions of different individuals, instead of using the entire face for recognition. In early 1990s, Gordon [1992] proposed the use of the principal curvatures in order to segment a facial range image using ridge and valley lines and used a local feature histogram to achieve face recognition.

Following from the above ideas, the use of a local feature histogram for face recognition, extracted using surface curvature from 3D images, has become more popular since the 1990s; examples include work by Mustafa et al. [1999], Hetzel et al. [2001], Moreno et al. [2003], Lee et al. [2005], Huang et al. [2006], Chen and Bhanu [2007]. However, little work has been conducted on the effects of viewpoint rotation until recently [Pansang et al., 2005, Norman et al., 2006, Akagündüz and Ulusoy, 2007, Li and Guskov, 2007, Lo et al., 2007]. This is an important issue to address since, to be of general utility, a feature descriptor should be invariant to perspective rotations. Moreover, most of the methods mentioned above are limited to a selection of single surface types only and typically require a user-defined threshold in order to segment the object with respect to the H, K and the principal curvatures ($k1$ and $k2$), which could lead to a different threshold being required for different types of images.

Further examples of local 3D keypoint descriptors include *point signatures* [Chua and Jarvis, 1997] in which "signatures" are extracted from arbitrary points and these signatures are used to vote for models with similar signatures. For a given point, a contour on the surface is defined around the point of interest. Each point on the contour may be characterised by the signed distance from the point of interest to a point on the contour and a clockwise rotation from the reference vector about the normal. However, local representations of 3D surfaces can be sensitive to noise, which can affect the features derived from differential quantities such as curvatures and surface normals. As a result, many new recognition systems have adopted geometric representations which combine local and global representations together; examples include *spin images* [Johnson, 1997, Johnson and Hebert, 1999] and *COSMOS* [Dorai and Jain, 1997].

In the approach taken here, we focus on *local* representations and we expect the measurement aperture, or the support region, to contain a *mixture* of surface types. Consider the surface of the human face for example: if the descriptor is extracted from the *pronasale* facial landmark, it will be dominated by a single surface type, whereas the descriptor extracted from the *exocanthion* landmark location will be expected to contain a wider mix of surface types. This concept is illustrated in Figure 1, showing the unique mixtures of surface types (shown as surface patches) and their relative frequencies (shown as bar graphs) taken at three different landmark locations of the face. Therefore, instead of segmenting surfaces into a piecewise patchwork of single surface types, we attempt to extract the distributions of the underlying surface using the *shape index, s,* (Equation 1) [Koenderink and van Doorn, 1992], derived from the $k1$ and $k2$ curvatures (Equations 3a and 3b respectively). The surface types derived from the shape index focus on the values ranging from cup (concave) to cap (convex), along with a wider intermediate values, as shown in Figure 2.

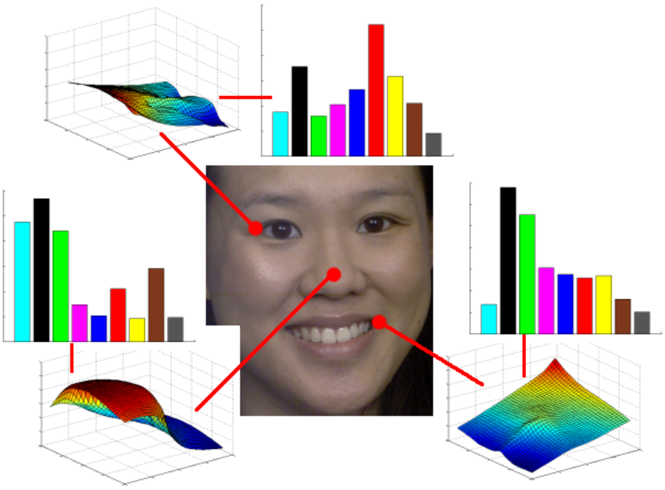


Figure 1: Surface types (as surface patches) and their histograms (as bar graphs) extracted from three keypoint locations on the face range data. Each bar represents a different surface type, where the colouring scheme corresponds to Figure 2.

$$s = \frac{2}{\pi} tan^{-1} \left[\frac{k2 + k1}{k2 - k1} \right] \qquad (1)$$

where $k1 > k2$.

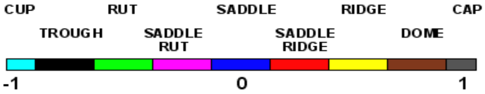


Figure 2: Shape Index scale ranges from -1 to 1.

The feature descriptor we extract in this work comprises a composite structure: it encapsulates the *mixture* of surface types present within the measurement aperture, derived using the shape index and weighted by the degree of curvedness, and also encodes their orientations, weighted by the gradient magnitude. This feature descriptor can characterise the local surface shape of a 2.5D range map, while affording a useful degree of invariance to Euler's out-of-plane rotations in viewpoint, thereby providing a means to range-map matching under pose changes in six-degrees-of-freedom.

The remainder of the document is organised as follows: Section 2 gives an overview of the concepts and methodology involved in the extraction of the 2.5D SIFT feature descriptors and the subsequent matching process. Section 3 presents validation results and finally Section 4 concludes this document and gives the details of proposed future research.

2 Methodology

There are four key stages of our 2.5D SIFT implementation:

1. The **keypoint localisation** stage, where the (x, y) positions of stable keypoints, along with their appropriate scales σ are detected in range images using scale-space, as proposed by Mikolajczyk and Schmid [2004].

2. The **canonical orientation(s) assignment** stage, where a consistent canonical orientation, θ, is assigned to each keypoint.

3. The **feature extraction** stage that computes stable feature descriptors located at the (x, y, σ, θ) coordinates of each keypoint.

4. The local feature **matching stage**.

Figure 3 illustrates the steps taken for the a full 2.5D SIFT to be accomplished.

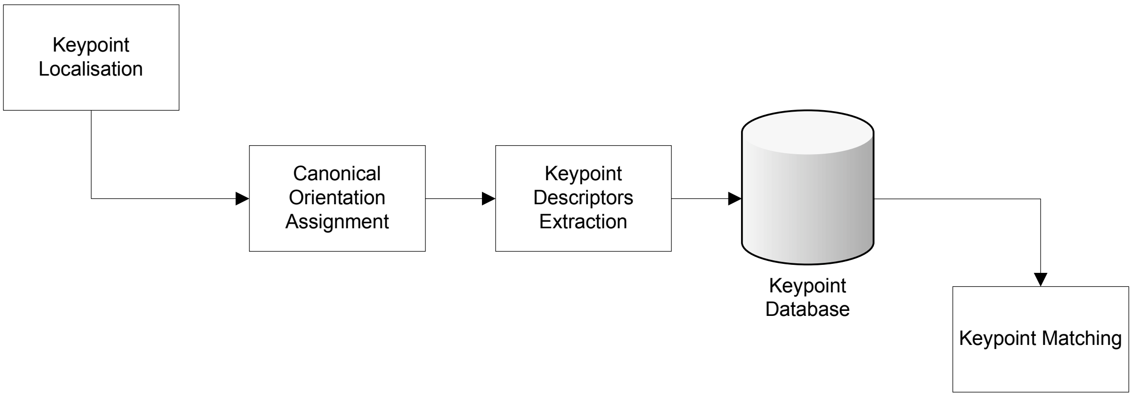

Keypoint Localisation
Canonical Orientation Assignment
Keypoint Descriptors Extraction
Keypoint Database
Keypoint Matching

Figure 3: Flowchart illustrating the stages involved in 2.5D SIFT.

2.1 Keypoint Localisation

A Gaussian-tapered segmentation mask is applied to the range image to isolate the area of interest while avoiding sharp boundaries that would create spurious keypoints. The resulting image is z-normalised to have global statistics of $(\mu = 0, \sigma = 1.0)$ in order to standardise the dynamic range of vales in the captured range image. Lowe's approach to keypoint detection is then applied thereafter: the z-normalised range image is blurred with a standard deviation of 0.5 to reduce aliasing and is up-sampled by a factor of two using linear interpolation. A scale-space representation [Lindeberg, 1994] is created by generating the Gaussian and the Difference-of-Gaussian (DOG) pyramid with sub-interval layers. Local extrema (maxima and minima) are localised in each layer by comparing each pixel within each layer to all the pixels in space and in scale (i.e. adjacent layers). Those keypoints falling below a user-defined contrast threshold T are rejected. In the work reported here, a T value of 0.003 was determined by experiment to extract satisfactory numbers of keypoints from range images of human faces (this T value contrasts with Lowe's threshold of 0.3 which was tuned for use on intensity images).

The H, K (Equations 2a and 2b respectively [Jain et al., 1995]), $k1$ and $k2$ curvatures (Equations 3a and 3b respectively) are then computed for each sub-level of the Gaussian pyramid: the $k1$ and $k2$ curvatures are derived from the H, K curvatures, which are in turn computed from the Gaussian derivatives. In order to reduce the effect of range image noise in the calculation of these derivatives [Marr, 1982], the Gaussian smoothing σ is set to that of the scale-space. Spatially compact feature locations are selected by comparing the ratio of the principal curvatures $\frac{k1}{k2}$, to a curvature threshold $r = 5$ tuned for use with the z-normalised range images (Lowe proposed $r = 10$ for intensity image data). The above procedure enables a set of (x, y, σ) values to be recovered for each keypoint detected.

$$H(i,j) = \frac{(1 + f_y^2(i,j))f_{xx}(i,j) + (1 + f_x^2(i,j))f_{yy}(i,j) - 2f_x(i,j)f_y(i,j)f_{xy}(i,j)}{2\left(\sqrt{1 + f_x^2(i,j) + f_y^2(i,j)}\right)^3} \tag{2a}$$

$$K(i,j) = \frac{f_{xx}(i,j)f_{yy}(i,j) - f_{xy}^2(i,j)}{\left(1 + f_x^2(i,j) + f_y^2(i,j)\right)^2} \tag{2b}$$

where (i,j) is the i^{th} and j^{th} pixel of the range image. f_x, f_{xx}, f_{xy}, f_y and f_{yy} denote the first and second Gaussian derivatives at (i,j) position.

$$k1 = H + \sqrt{H^2 - K} \tag{3a}$$

$$k2 = H - \sqrt{H^2 - K} \tag{3b}$$

2.2 Orientation Assignment

A consistent *canonical orientation* can be assigned to each keypoint location based on the local image gradient orientation properties. Multiple canonical orientations can be assigned to a keypoint, resulting in assorted descriptors for the keypoint. A modified version of Lowe's orientation assignment algorithm has been used for this work. The steps involved in the orientation assignment for each keypoint location over a measurement aperture is as follows:

1. Following Lowe's methodology, a circular Gaussian mask, set to the detected measurement aperture scale, is used to sample the image and the Gaussian is centred

on the keypoint location with sub-pixel accuracy. The equation for a symmetric two-dimensional Gaussian square kernel with scale σ_i used to place Gaussian support regions on an image with sub-pixel accuracy is as follows:

$$G = \frac{1}{2\pi\sigma^2} e^{-\dfrac{(x_i - round(x_i) - \text{offset}_x)^2 + (y_i - round(y_i) - \text{offset}_y)^2}{2\sigma^2}} \tag{4}$$

2. The local image gradient magnitudes and orientations within the sampling mask are computed using the Gaussian first derivatives of the image. A histogram is formulated that comprises 360 bins, each bin containing a relative frequency entry for each of the 360° potentially detectable orientations. Each detected orientation entry is weighted by its corresponding Gaussian derivative magnitude value prior to being accumulated in the appropriate histogram bin. In Lowe's SIFT implementation, only 36 histogram entries are employed (i.e. 10° bin intervals). Our experimental observations reveal that this level of orientation quantisation appears to result in instability of the orientation estimates recovery. We have determined empirically that by increasing the orientation histogram resolution to 1°, and then applying a wide filter to each histogram entry followed by orientation peak interpolation, it is possible to recover the full range of in-plane canonical orientations reliably, as described in the remaining steps.

3. The values of the orientation histogram are stabilised, in terms of orientation continuity, by applying a wide 1D symmetric Gaussian convolution kernel of size$= 17$ and of $\sigma = 17$ to the histogram three times. This step anti-aliases the orientation histogram and stabilises the keypoint canonical orientation allocation process by providing estimates of orientation that change smoothly as the input visual stimulus changes in orientation.

4. The orientation peaks in the filtered histogram are located and each peak within 80% of the magnitude of the largest peaks is deemed to represent a keypoint. Sub-bin orientation precision is obtained by interpolation. A quadratic polynomial is fitted to the three histogram values closest to the peak, as shown in Equation 5 below:

$$\begin{pmatrix} (\theta_{peak} - \Delta\theta)^2 & \theta_{peak} - \Delta\theta & 1 \\ \theta_{peak}^2 & \theta_{peak} & 1 \\ (\theta_{peak} + \Delta\theta)^2 & \theta_{peak} + \Delta\theta & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} H\left(\theta_{peak} - \Delta\theta\right) \\ H\left(\theta_{peak}\right) \\ H\left(\theta_{peak} + \Delta\theta\right) \end{pmatrix} \tag{5}$$

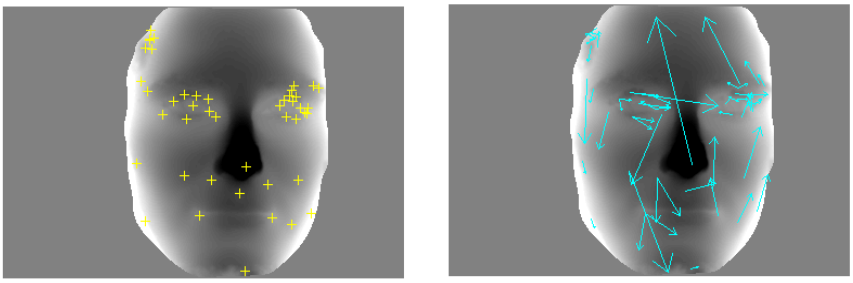
   where $H(\theta_{peak})$ is the histogram value at θ_{peak}.

5. The canonical orientation of the descriptor is given by:

$$\theta_{canonical} = -\frac{b}{2a} \tag{6}$$

2.3 Feature Descriptor Extraction

A feature descriptor is extracted at each keypoint landmark location, over a Gaussian weighted measurement aperture scaled to the current keypoint σ and orientated to the keypoint canonical orientation, in the following steps:



(a) Keypoints located

(b) Keypoints, scale and orientations

Figure 4: (a) Keypoint locations (shown as +) extracted using the modified SIFT keypoint localisation algorithm. (b) The scale (demonstrated by the magnitude of the arrows) and the canonical orientation(s) (the directions of the arrows) for each keypoint location.

1. The image patch sampling each keypoint location is rotated to the keypoint canonical orientation in order to achieve viewpoint rotation invariance. Differential geometry is used to compute the H, K, $k1$ and $k2$ curvatures (Equations 2a, 2b, 3a and 3b respectively) using the first and second Gaussian derivatives, as opposed to the raw derivatives used by Lowe, in order to obtain stable curvature estimates. Thereby it is possible to categorise the underlying distribution of the surface types present at the keypoint, using the bounded [-1,1] local shape index (Equation 1). The degree of local curvedness (Equation 7), along with the local image gradient orientation and corresponding local magnitude estimate can also be computed from the first and second Gaussian derivatives.
$$curvedness = \sqrt{2H^2 - K} \tag{7}$$

2. Lowe constructs a *keypoint descriptor* by subdividing a 16×16 pixel patch he defines in order to sample the keypoint location into a 4×4 matrix of non-overlapping sub-patches and then extracts orientation histograms from these sub patches. Here we have adopted a slightly different patch sub-sampling strategy. Nine Gaussian weighted sub-regions [Balasuriya, 2005, Balasuriya and Siebert, 2006], overlapped by one standard deviation, are placed over the sampling patch, as shown in Figure 5. Since overlapping the Gaussian sub-regions results in the feature descriptors extracted from adjacent sub-regions being correlated, this reduces spatial aliasing and also enforces spatial continuity that occurs during sampling. For example, small shifts in the location of the keypoint will now result in small (continuous) changes in the magnitude of the composite keypoint descriptor (and its component vectors). The choice of how many sub-regions to utilise is a trade-off between excessive dimensionality and feature discriminability, particularly to symmetric patterns. A sampling configuration comprising 3×3 overlapped matrix had been found by Balasuriya to achieve a good working compromise.

3. For each of the nine sub-regions placed over the sampling patch, a local histogram is computed to characterise the relative frequencies of the nine surface types, weighted by the degree of curvedness, within the sampled keypoint region. Similarly, an eight-element histogram, covering the 360° range of orientations, is formulated, weighted by

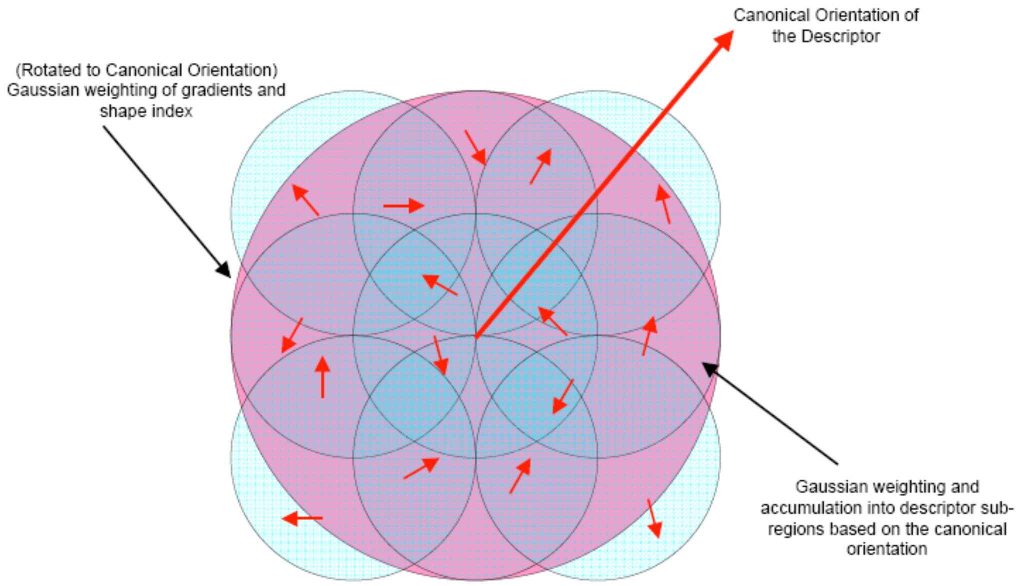

Canonical Orientation of the Descriptor
(Rotated to Canonical Orientation) Gaussian weighting of gradients and shape index
Gaussian weighting and accumulation into descriptor sub-regions based on the canonical orientation

Figure 5: Placement of the nine sub-regions, with the spatial support at one standard deviation, over the keypoint landmark location.

the gradient magnitude. Therefore this feature descriptor is designed to represent keypoint locations in terms of their mixes of surface types and characteristic orientation directions, i.e. surface shape signature and its dominant directions. Each histogram is normalised to unity magnitude (i.e. to a unit vector) to endow the feature descriptor with signal magnitude invariance. The influence of large histogram values (i.e. outliers caused by spurious range data, such as noise spikes) in each normalised histogram is reduced by clipping the value at a threshold of $\frac{1}{\sqrt{a}}$, where a is the number of bins in the histogram. The histograms are then concatenated to form H_i, which is then normalised to unity magnitude.

$$\widehat{LocalHist_i} = \left(\widehat{H_{surface}}\right)\left(\widehat{H_{orientation}}\right) \tag{8}$$

4. The nine normalised histograms $LocalHist_i$ are juxtaposed to form the final feature descriptor:

$$Descriptor_{\theta_{canonical}} = \left(\widehat{LocalHist_1}, \widehat{LocalHist_2}, ..., \widehat{LocalHist_9}\right) \tag{9}$$

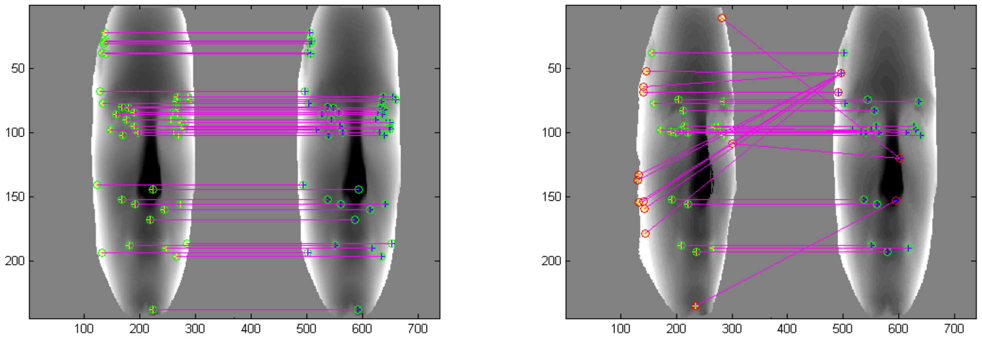
2.4 Matching Feature Descriptors

The discriminability of the extracted feature descriptors against viewpoint rotational changes can be determined by matching the feature descriptors extracted from different set of images, captured at different angles. Following the methodology proposed by Lowe [2004], a candidate match is located by computing and ranking (in ascending order) the angle between the descriptors using the vector dot product. False matches can be initially rejected using the log likelihood ratio test if the ratio between the potentially best matched descriptor (val_1) to its next best matched descriptor (val_2) is above a `distRatio` threshold of 0.8 (Equation 10).

$$\text{Match} = \begin{cases} 1, & \text{if } \dfrac{\text{val}_1}{\text{val}_2} < \text{distRatio} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

In order to verify matches between two different range images (captured at different angles), a similarity transform is computed between the two sets of descriptors by means of the Hough Transform. Clusters of matching features with a consistent interpretation (i.e. matches between features exhibiting the same relative shift in orientation, translation and scale) are identified. In other words, a similarity transform between a test set of descriptors and an image in the database is computed using the Hough Transform. If three or more entries are located in each cluster, it is possible to apply a robust affine transform fitting procedure to the cluster in order to recover the affine pose between the matched features and also identify outliers. This process matches reliably a set of extracted feature descriptors to sets of feature descriptors contained in a database and extracted from range images captured at different angles.

Figure 6(a) shows a self-matching range image at the same scale while Figure 6(b) shows the baseline range image being matched to a range image rotated to a different angle (30° in the yaw axis).



(a) Self-matching

(b) Matching onto different orientation

Figure 6: Examples of point-to-point matching of the range images where (a) shows a self-matching range image and (b) shows the baseline range image being matched to a rotated range image.

2.5 Hough Transform

The Hough Transform [Duda and Hart, 1972, Ballard, 1981] is used to identify clusters of features that have a consistent interpretation of an object hypothesis by a voting procedure, where the object hypothesis contains not only the object label but also its position, scaling and rotation (in this work). The Hough Transform is especially useful when there are a high proportion of outliers in the matched feature descriptors.

The Hough Transform maps descriptor matches from spatial coordinates in the visual scene to a hypothesis voting accumulator-space to eliminate outlying object, position or pose hypotheses which accumulate fewer votes. Feature descriptor matches vote into the Hough accumulator space, which is parameterised by the underlying degrees of freedom considered within the problem domain: translation (in plane), rotation (in plane) and scale in size.

2.6 Affine Transformation

Lowe's methodology is applied directly for estimation refinement here [Lowe, 2004]. Once the Hough Transform has identified three or more entries in each cluster, the affine pose between the matched features can be recovered, thereby allowing outliers to be located, using affine transformation.

If $f(x,y)$ and $f'(x',y')$ are the feature descriptors from training and test respectively, the transformation of the object from the training image to the test image may be accurately given as follows:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{11}$$

where m_1, m_2, m_3, m_4 and t_x, t_y are the parameters of the affine transformation of the object from the training appearance view to the test scene. These may be determined by solving the following the least squares system where a single match $f(x,y)$ and $f'(x',y')$ is indicated. Since there are six unknowns, at least three match pairs (six equations) will be needed to determine transformation parameters.

$$\begin{bmatrix} x' \\ y' \\ \vdots \end{bmatrix} = \begin{bmatrix} x & y & 0 & 0 & 1 & 0 \\ 0 & 0 & x & y & 0 & 1 \\ & & \cdots & & & \\ & & \cdots & & & \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_x \\ t_y \end{bmatrix} \tag{12}$$

The principal differences between Lowe's 2D SIFT and our version of 2.5D SIFT are shown in Table 1.

3 Validation Results and Analysis

In this work, we compare the performance between standard 2D SIFT on 2D images and our version of 2.5D SIFT on 2.5D images, against rotational changes. We employed a set of range images of human faces generated from stereo-pair images captured using 13.5 Megapixel digital cameras and processed using a stereo-photogrammetry package, C3D [Ju et al., 2003], along with their corresponding 2D images. A total of 28 2D and 2.5D images have been used for this validation. Rotational changes with respect to the out-of-plane (yaw and pitch) axes have been simulated by re-projecting the data, in increments of 10° up to ±30°, into a new image. A selection of synthetically rotated range images are shown in Figure 7 and Figure 8 respectively. This enables us to test the discriminability of the extracted descriptors to viewpoint rotational changes. Furthermore, the original size of the range images, and their corresponding 2D images, is 1498×2249 pixels and at this scale, it is computationally

Stages	Process	2D SIFT	2.5D SIFT
Keypoint Localisation	Contrast threshold	0.3	0.003
	Ratio threshold	10	5
	Curvature extraction method	Hessian Matrix	1st & 2nd Gaussian derivatives
Canonical Orientation Assignment	Number of bins used in the orientation histogram	36	360
	Smoothing of histogram	Gaussian-weighted circular window with $\sigma = 1.5\times$ scale of keypoint	Gaussian convolution kernel of size $= 17$,$\sigma = 17$ (applied three times)
Keypoint Descriptor Extraction	Information captured	Orientation	Surface types + orientation
	Subsampling	4×4 descriptors computed from a 16×16 sample array	9 overlapped Gaussian sub-regions (by 1 s.d)
	Length of descriptor	128-element	153-element

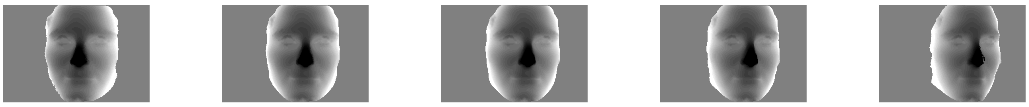
Table 1: Differences between Lowe's 2D SIFT and our version of 2.5D SIFT.

expensive to extract any feature descriptors (taking approximately 5 hours on a 64-bit machine to extract one set of feature descriptors). As a result, we have downsized the images to 244×369 pixels using a half-octave Gaussian pyramid, thereby reducing the time taken to extract feature descriptors to approximately four minutes. These images are also similar in size to the examples Lowe employed in his work Lowe [2004].

In order to deduce the stability of our 2.5D keypoint descriptors, we match descriptors extracted from each image captured (in the set of rotated images) against all other keypoint descriptor sets extracted from the remaining images in the set. Therefore all combinations of rotated viewpoints are matched against each other. Furthermore, we use the Hough Transform to eliminate matching outliers in Hough space. While the Hough Transform itself is not infallible in removing outliers, from visual inspection of the outliers, it does appear to be effective. Although the positive matches detected cannot be guaranteed to be all correct, using this filtered data we can compute a match-matrix that represents more reliably the percentage of correctly labelled keypoints. We can therefore estimate more accurately the discriminability of our 2.5D keypoint descriptors. Tables 2 and 3 show the number of keypoints found for each view.

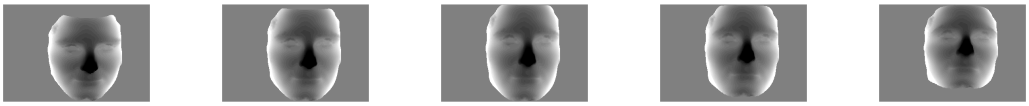
The validation conducted in this work is separated into three parts: we first investigate the performance of 2D SIFT on 2D images, against rotational changes. Secondly, we apply the standard 2D SIFT to range images and measure its performances against rotational changes. Finally, we apply our version of 2.5D SIFT to range images and test the discrimination of the keypoint descriptors against rotational changes. The results are presented here.

Figure 9 shows the match-matrices obtained by comparing all the combinations of the



(a) -30°   (b) -10°   (c) 0°   (d) 10°   (e) 30°

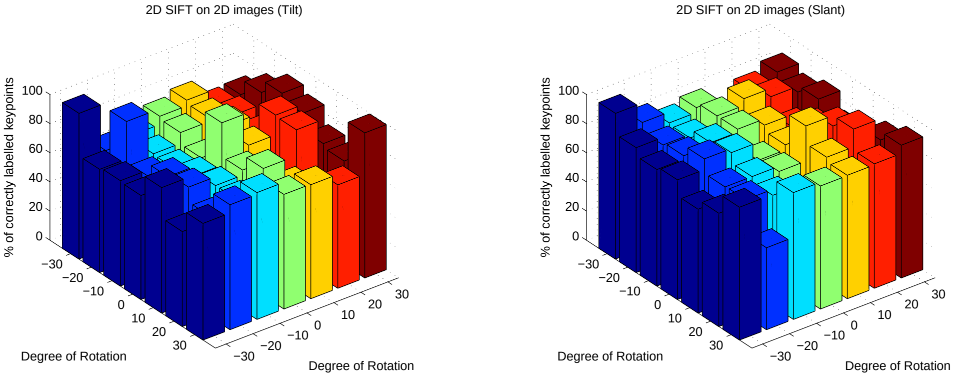
Figure 7: A selection of synthetically rotated range images about the yaw axis, generated at (a) -30°, (b) -10°, (c) 0°, (d) 10° and (e) 30°.



(a) -30°   (b) -10°   (c) 0°   (d) 10°   (e) 30°

Figure 8: A selection of synthetically rotated range images about the pitch axis, generated at (a) -30°, (b) -10°, (c) 0°, (d) 10° and (e) 30°.

keypoint descriptors extracted from 2D intensity images, using the standard 2D SIFT algorithm, captured at different angles (from -30° to 30° in both clockwise and anticlockwise directions about the (a) yaw and (b) pitch axis). The lower axes of these match-matrices show the viewpoint angles of each pair of the compared range images, while the heights of the columns show the percentages of matched keypoints. Closer examination of these matrices shows that 76.2% of the test images captured about the yaw axis are matched while 82.3% of the keypoints captured about the pitch axis are matched.



(a) Yaw   (b) Pitch

Figure 9: Match-matrix results for the discriminability of the feature descriptors extracted from 2D images using standard 2D SIFT against out-of-plane (about the (a) yaw (b) pitch axis) viewpoint rotational changes.

Figure 10 presents the match-matrices obtained by exploring all the combinations of the feature descriptors extracted from 2.5D range images by applying the standard 2D SIFT algorithm, captured at different angles from ±30° in both clockwise and anticlockwise di-

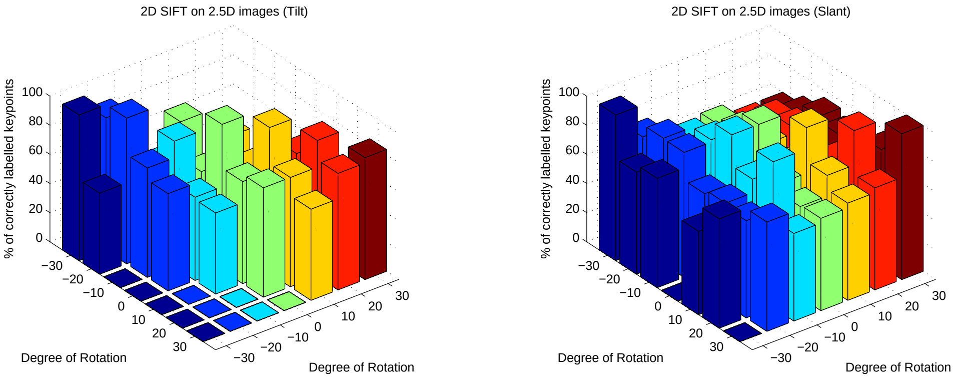
	2D SIFT on 2D images	**2D SIFT on 2.5D images**	**2.5D SIFT on 2.5D images**
-30°	86	54	106
-20°	79	43	92
-10°	74	39	107
0°	79	40	84
10°	78	46	106
20°	69	45	102
30°	65	63	114

Table 2: Number of keypoints found for each view (rotation about **yaw** axis).

	2D SIFT on 2D images	**2D SIFT on 2.5D images**	**2.5D SIFT on 2.5D images**
-30°	74	44	102
-20°	70	46	104
-10°	72	42	92
0°	79	40	84
10°	78	49	88
20°	79	45	87
30°	82	53	91

Table 3: Number of keypoints found for each view (rotation about **pitch** axis).

rections about the (a) yaw and (b) pitch axis. Examination of these matrices shows that approximately 41.3% of the keypoints captured about the yaw axis are matched while 69.5% of the keypoints captured about the pitch axis are matched. Note that there are instances in the matrices where 0% of matched keypoints are returned, indicating the performance of standard 2D SIFT on range images is relatively poor, compared to using standard 2D SIFT on 2D images.

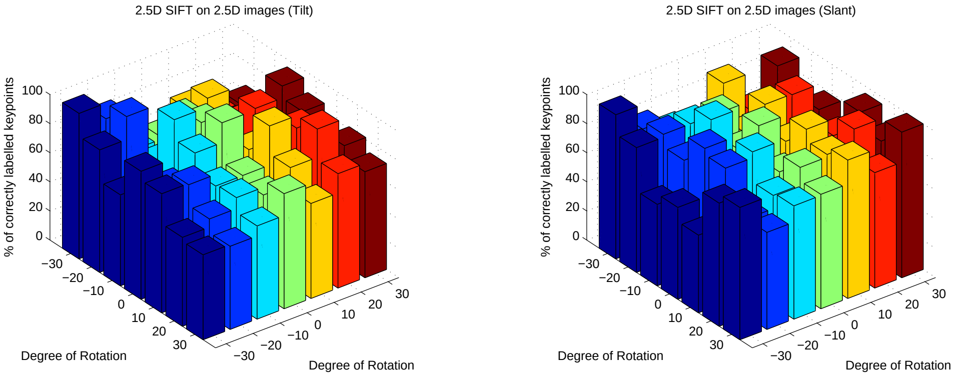


(a) Yaw

(b) Pitch

Figure 10: Match-matrix results for the discriminability of the feature descriptors extracted from 2.5D range images using standard 2D SIFT against out-of-plane (about the (a) yaw (b) pitch axis) viewpoint rotational changes.

Figure 11 illustrates the match-matrices obtained by comparing all the combinations of the feature descriptors extracted from 2.5D range images, using our version of the 2.5D SIFT. Figure 11(a) shows the results obtained from the images rotated about the yaw axis while Figure 11(b) presents the results obtained from the images rotated about the pitch axis. On closer examination of these match-matrices, 77.5% of the keypoints captured about the yaw axis are matched and 81.5% of the keypoints captured at the pitch axis are matched. These results illustrate that the performance of our 2.5D SIFT on range images is comparable to the standard 2D SIFT on 2D images and outperforms the standard 2D SIFT on range images.



(a) Yaw

(b) Pitch

Figure 11: Match-matrix results for the discriminability of the feature descriptors extracted from 2.5D range images using our version of 2.5D SIFT against out-of-plane (about the (a) yaw (b) pitch axis) viewpoint rotational changes.

4 Conclusions and Future Work

This paper describes a 2.5D formulation of Lowe's SIFT keypoint descriptor and compares the performance of standard SIFT applied to 2D intensity images with our version of 2.5D SIFT on 2.5D range images. In our 2.5D implementation, by adopting statistical normalisation of the input range images, it becomes possible to set a consistent set of parameters appropriate to detecting stable keypoint locations and their appropriate scales (independently of the dynamic range of the input range maps or their content). By increasing the precision with which the canonical orientation at each keypoint sampling location is extracted (to the nearest degree), and by rotating the keypoint sampling patch to this canonical orientation, stable viewpoint rotation can be achieved. In order to capture a perceptually meaningful description of the surface patches sampled at keypoint locations, our 2.5D keypoint descriptors are formulated to sample the underlying relative frequencies of surface types present. Since this form of feature descriptor is based on the local shape index distribution, it is capable of capturing a statistical description of the local surface topology. Potential sampling effects caused by spatial aliasing within the standard SIFT keypoint descriptor formulation have been minimised by placing nine overlapping Gaussian circular sub-regions, with spatial support of one standard deviation, over each sampled keypoint location at the detected feature scale. This sub-sampling strategy enables our 2.5D keypoint formulation to be capable of differentiating between mirror-image features in range images, thereby disambiguating such feature-pairs present on symmetric objects such as faces. As the shape index measure is theoretically invariant to viewing directions, the aim of this feature descriptor formulation is to increase the invariance properties of the feature descriptor to Euler's out-of-plane rotations.

Validation has been performed using a set of range data of human faces and their corresponding 2D images, synthetically rotated in both clockwise and anticlockwise directions ($\pm 30^\circ$) about the yaw and pitch axes. We applied the standard 2D SIFT algorithm to 2.5D range images and the results showed the performance was inferior to that of standard 2D SIFT applied to 2D images. We then applied our version of 2.5D SIFT to range images and results showed that the performance was comparable to applying standard SIFT to 2D intensity images. This shows that our version of 2.5D SIFT provides feature descriptors that maintain good invariance to rotational changes over the range of angles investigated. In order to investigate orientation invariance beyond the angular range reported here, either real data or projected full 3D data is necessary to avoid the errors that result when rotating range data. Availability of such data would allow a more extreme comparison between standard 2D SIFT and the 2.5D SIFT formulation reported here.

At a fundamental level, the use of range data allows additional information (over that available to 2D SIFT) to be captured that not only describes the shape of the range surface manifold but also the local direction of this surface with respect to the imaging sensor. Therefore, in order to fully exploit the potential information offered by 2.5D range data, in the next stages of our work we propose to investigate the extraction of surface *canonical normals* at each keypoint location. This process would allow 3D pose information to be recovered and then used to adapt the current circular sampling support region to an elliptical window in order to sample more accurately the underlying surface. Moreover, it was outwith the scope of this work at this stage to address the sensitivity of the feature descriptors extracted, using our version of 2.5D SIFT, w.r.t. noise; therefore investigation of this issue will be required in the future. We also plan to explore an architecture that combines the 2D and 2.5D analyses by grouping keypoints from each modality within the Hough Transform. Finally, we intend to investigate the representation of local and global biological variability by performing Principal Component Analysis (PCA) on keypoint descriptors and their spatial locations in the Hough Transform, allowing range image recognition to be performed on different biological forms.

References

E. Akagündüz and I. Ulusoy. 3D object representation using transform and scale invariant 3D features. In *Proceedings of ICCV'07 Workshop on 3D Representation for Recognition (3dRR-07)*, pages 1–8, 2007.

L. S. Balasuriya and J. P. Siebert. Hierarchical feature extraction using a self-organised retinal receptive field sampling tessellation. *Neural Information Processing - Letters & Reviews*, 10(4-6):83–95, 2006.

S. Balasuriya. *A Computational Model of Space-Variant Vision Based on a Self-Organised Artificial Retina Tessellation*. PhD thesis, University of Glasgow, 2005.

D. H. Ballard. Generalizing the Hough transform to detect arbitrary patterns. *Pattern Recognition*, 13(2):111–122, 1981.

P. J. Besl. *Surfaces in Range Image Understanding*. Springer-Verlag, 1998.

P. J. Besl and R. C. Jain. Three-dimensional object recognition. *ACM Computing Surveys*, 17(1):75–145, 1985.

K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.

J. Y. Cartoux, J. T. LaPreste, and M. Richetin. Face authentication or recognition by profile extraction from range images. In *Proceedings of the Workshop on Interpretation of 3D Scenes*, pages 194–199, 1989.

H. Chen and B. Bhanu. 3D free-form object recognition in range images using local surface patches. *Pattern Recognition Letters*, 28(10):1252–1262, 2007.

C. S. Chua and R. Jarvis. Point signatures: a new representation for 3D object recognition. *International Journal of Computer Vision*, 25(1):63–85, 1997.

C. Dorai and A. K. Jain. COSMOS – A representation scheme for 3D free-form objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10):1115–1130, October 1997.

R. O. Duda and P. E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, January 1972.

T. G. Fan, G. Medioni, and R. Nevatia. Description of surfaces from range data using curvature properties. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 86–91, 1986.

G. G. Gordon. Face recognition based on depth and curvature features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 808–810, 1992.

C. Hesher, A. Srivastava, and G. Erlebacher. A novel technique for face recognition using range imaging. In *Seventh International Symposium on Signal Processing and Its Applications*, pages 201–204, 2003.

G. Hetzel, B. Leibe, P. Levi, and B. Schiele. 3D object recognition from range images using local feature histograms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 394–399, 2001.

Y. Huang, Y. Wang, and T. Tan. Combining statistics of geometrical and correlative features for 3D face recognition. In *Proceedings of British Machine Vision Conference*, volume 3, pages 879–888, 2006.

D. J. Ittner and A. K. Jain. 3-D surface discrimination from local curvature measures. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 119–123, 1985.

R. Jain, R. Kasturi, and B. G. Schunck. *Machine Vision*. MIT Press and McGraw-Hill Inc, 1995.

A. E. Johnson. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, August 1997.

A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.

X. Ju, T. Boyling, and J. P. Siebert. A high resolution stereo imaging system. In *Proceedings of 3D Modelling 2003*, 2003.

J. J. Koenderink and A. J. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8): 557–565, 1992.

J. C. Lee and E. Milios. Matching range images of human faces. In *Proceedings of International Conference on Computer Vision*, pages 722–726, 1990.

Y. Lee, H. Song, U. Yang, H. Shin, and K. Sohn. Local feature based 3D face recognition. In *Proceedings of International Conference on Audio- and Video-based Biometric Person Authentication*, volume 3546, pages 909–918, 2005.

X. J. Li and I. Guskov. 3D object recognition from range images using pyramid matching. In *Proceedings of ICCV'07 Workshop on 3D Representation for Recognition (3dRR-07)*, pages 1–6, 2007.

T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

T. W. R. Lo, J. P. Siebert, and A. F. Ayoub. An implementation of the scale invariant feature transform in the 2.5D domain. In *Proceedings of MICCAI 2007 Workshop on Content-based Image Retrieval for Biomedical Image Archives: Achievements, Problems, and Prospects*, pages 73–82, 2007.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

D. Marr. *Vision*. W. H. Freeman and Co., 1982.

K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

A. B. Moreno, Á. Sánchez, J. F. Vélez, and F. J. Díaz. Face recognition using 3D surface-extracted descriptors. In *Proceedings of Irish Machine Vision and Image Processing Conference*, 2003.

A. A. Y. Mustafa, L. G. Shaprio, and M. A. Ganter. 3D object identification with color and curvature signatures. *Pattern Recognition*, 32(3):339–355, March 1999.

J. F. Norman, J. T. Todd, H. F. Norman, A. M. Clayton, and T. R. McBride. Visual discrimination of local surface structure: Slant, tilt and curvedness. *Vision Research*, 46(6–7):1057–1069, 2006.

S. Pansang, B. Attachoo, C. Kimpan, and M. Sato. Invariant range image multi-pose face recognition using gradient face, membership matching score and 3-layer matching search. *IEICE Transactions on Information and Systems*, E88-D(2):268–277, 2005.