Spectral-Spatial Scale Invariant Feature Transform for Hyperspectral Images

Suhad Lateef Al-khafaji, Student Member, IEEE, Jun Zhou, Senior Member, IEEE, Ali Zia and Alan Wee-Chung Liew, Senior Member, IEEE

Abstract—Spectral-spatial feature extraction is an important task in hyperspectral image processing. In this paper we propose a novel method to extract distinctive invariant features from hyperspectral images for registration of hyperspectral images with different spectral conditions. Spectral condition means images are captured with different incident lights, viewing angles, or using different hyperspectral cameras. In addition, spectral condition includes images of objects with the same shape but different materials. This method, which is named Spectral-Spatial Scale Invariant Feature Transform (SS-SIFT), explores both spectral and spatial dimensions simultaneously to extract spectral and geometric transformation invariant features. Similar to the classic SIFT algorithm, SS-SIFT consists of keypoint detection and descriptor construction steps. Keypoints are extracted from spectral-spatial scale space and are detected from extrema after 3D difference of Gaussian is applied to the data cube. Two descriptors are proposed for each keypoint by exploring the distribution of spectral-spatial gradient magnitude in its local 3D neighborhood. The effectiveness of the SS-SIFT approach is validated on images collected in different light conditions, different geometric projections, and using two hyperspectral cameras with different spectral wavelength ranges and resolutions. The experimental results show that our method generates robust invariant features for spectral-spatial image matching.

Keywords—Hyperspectral image, spectral-spatial feature extraction, SIFT, keypoint detection.

I. INTRODUCTION

Hyperspectral image (HSI) is more informative than grayscale and RGB images thanks to its high spectral resolution and wider sensing spectrum. HSI first appeared in the 1980s, when NASA’s Jet Propulsion Laboratory developed a remote sensing instrument called AVIRIS (Airborne, Visible Infra-Red imaging Spectrometer) [1]. Since then, HSI has been utilised in remote sensing for surveillance, agriculture, environmental and monitoring applications [2]. Recently, due to the cost reduction of imaging devices, HSI has also been introduced into computer vision applications, for example, health analysis, food security, and close-range object detection and recognition [3].

An HSI is represented as a three dimensional data cube, with each pixel being a vector whose dimension equals the number of wavelength indexed bands. The pixel-wise spectral responses characterise the object reflectance at different wavelengths, which are related to the material property of objects [4]. Therefore, pixel-wise spectral information is significant in discriminating different types of materials. Unfortunately, pixel-based spectral analysis cannot distinguish objects that have the same material but different structures [5], [6]. On the contrary, many spatial and structural features have been proposed in the computer vision domain for scene description. When applied to HSI, such features can be extracted on a single spectral band or on a grayscale image produced by reducing the dimension of HSI, e.g., using principal component analysis [5], but they miss the advantage of abundant spectral information.

A natural solution to this problem is to develop spectral-spatial features which describe data distribution in both spectral and spatial domains simultaneously [7], [8]. These features can serve as the basic tools for a number of HSI processing tasks, such as image registration, image segmentation, retrieval and classification. Many efforts have been made to explore this option. Typical spectral-spatial features include 3D Gabor filters [9], 3D discrete wavelet transform (3D DWT) [10], 3D scattering wavelet [11], 3D local derivative pattern [5], and 3D gray level co-occurrence matrix [12]. Furthermore, Liang et al. proposed a 3D high-order texture pattern descriptor for hyperspectral face recognition [13]. In more recent, deep learning method is also adopted for HSI image representation and classification [6], [14]. All these methods show their effectiveness in exploring different properties of the spectral and spatial features of HSI.

In practice, HSIs with various spectral resolutions can be captured using different hyperspectral sensors and cameras. For example, in remote sensing, Landsat and EO-1 satellites capture multispectral and hyperspectral images with the same 30 meter spatial resolution but 7 spectral bands for multispectral image and 220 spectral bands for HSI. In computer vision, hyperspectral cameras with different spectral properties can be used to capture complimentary information of objects in the same scene. Furthermore, due to the varying sensitivity functions of complementary metal-oxide semiconductor (CMOS) or charge-coupled device (CCD) sensors adopted by different camera producers, images may have different intensity levels but similar trends of spectral responses for the same material. All these make fundamental image processing tasks, such as registration and matching, difficult to complete. There still lacks a robust keypoint detection and description method that can deal with such spectral-spatial scale invariant problem.

The main objective of this paper is to address this issue by extracting spectral-spatial scale invariant features from
HSI. These features are used for matching different HSIs of an object or scene with different spatial views and spectral resolutions. We propose a new algorithm, Spectral-Spatial Scale Invariant Feature Transform (SS-SIFT), that exploits both spectral and spatial dimensions of the 3D data at the same time. Distinctive features are identified and extracted from different spectral-spatial scales, so they are invariant to spectral and spatial dimension changes. Our method is inspired from Scale-Invariant Feature Transform (SIFT) algorithm [15] which has been modified for 3D images, multispectral and hyperspectral images in recent research [16], [17]. Another related work is 3D SIFT which was proposed for video or 3D medical images [18], [19], [20], but it has not been applied to HSI cube.

Existing extension of SIFT have been applied to multispectral image and HSI, respectively. On multispectral images, researchers have proposed modified SIFT algorithm to extract spectral invariant features in addition to geometric transformation invariant features [16], [21], [22]. The main objective of these research is to solve the problem of significantly different spectral appearance of objects in multispectral band images as well as difference viewpoints and resolutions. However, all these methods have been implemented on individual bands of multispectral image without exploiting the abundant spectral information. On hyperspectral images, attempts have been made to use SIFT algorithms for scale, rotation and affine transformation invariant feature extraction [17], [23], [24], [25]. These works were applied on either spectral or spatial domain only. Whereas, our method explores the joint spectral and spatial data distribution simultaneously. This is the main novel contribution of our work, i.e., utilizing whole hyperspectral data cube to extract spectral-spatial scale invariant features.

Another contribution of this work is the spectral-spatial keypoint detection method. Spectral-spatial keypoint detection method is normally based on detecting local extremum in a spectral-spatial image patch after applying 3D Difference of Gaussian (DoG). Spatially, an extremum can represent an edge, a corner, a curvature or a blob, while spectrally it represents the peak and valley that formulate the shape of a spectral curve. Consequently, extrema are the most stable points that can provide distinctive spectral-spatial information simultaneously. After detecting the keypoints, two novel spectral-spatial descriptors are constructed depending on the orientation of the spectral-spatial gradient magnitude of each keypoint in its neighborhood. The first descriptor \( D_{\theta,\phi} \) is a 3D histogram constructed from two angle orientations \((\theta, \phi)\) and gradient magnitudes which are calculated in \( x, y, \lambda \) dimensions. The second descriptor \( D_{\phi} \) is created by concatenating three 2D histograms of gradient magnitude and orientation. These three 2D histograms are calculated in \((x, y), (x, \lambda),\) and \((y, \lambda)\) dimensions separately. Both proposed \( D_{\theta,\phi} \) and \( D_{\phi} \) descriptors are considered as spectral-spatial scale invariant features that provide robust HSI matching.

In the experiments, keypoints matching can be determined by firstly calculating the minimum Euclidean distance between keypoint descriptors, and then refined by calculating the spectral angle distance in Eq. (15). If two images were captured by the same camera, they have the same wavelength range and bands, so the spectral angle distance is calculated on all bands. Otherwise, two images may have different wavelength ranges. In this case, only the overlapping bands are used.

The rest of this paper is organized as follows. Section II presents a review on 3D SIFT, SIFT algorithm for multispectral image and SIFT algorithm for hyperspectral images. Section III describes the proposed 3D SS-SIFT method. Hyperspectral data and keypoints matching process are presented in sections IV and V respectively. The experimental results are presented in section VI. Finally, comparison and conclusions are given in sections VII and VIII respectively.

II. BACKGROUND

2D SIFT features are identified as the robust invariant feature to rotation, translation, scaling and linear intensity changes for greyscale image matching and registration in computer vision [26]. In recent years, SIFT has been modified to extract features from 3D images, such as video, medical image, multispectral and hyperspectral images. In this section, we will briefly describe the related work on SIFT for 3D images, multispectral images and hyperspectral images.

A. 3D SIFT

The original SIFT algorithm has been extended from 2D space to 3D space [18], [19], [20], [27], [28]. Scovanner et al. [18] proposed a 3D SIFT descriptor that was applied to video data. This work showed that 3D SIFT provides superior results in video action recognition based on extracting repeatable keypoint features. Ni et al. [27] applied 3D SIFT for registering global ultrasound volumes acquired by dedicated ultrasound probe. Flitton et al. [28] presented 3D SIFT with a volumetric image produced by dual energy Computed Tomography (CT) scanners. In this research, 3D SIFT operator was used to recognize rigid objects in complex CT volumes which had significant metal artefact and poor resolution. In general, 3D SIFT operators demonstrate their capability in extracting significant keypoints with feature descriptors for robust matching and recognition process [18], [19], [20].

Since HSI is considered as a 3D data cube, 3D SIFT can be used to extract features from HSI. However, there is a dearth of research using 3D SIFT with HSI since the third dimension of HSI has different physical interpretation from other types of 3D data, i.e., it corresponds to continuous reflectance change across wavelengths. Whereas, in videos or medical images, the third dimension indexes time or spatial location. Such differences imply that different consideration shall be taken when developing 3D SIFT for HSI to explore the spectral changes at the same spatial location for each band.

Each spectral band in HSI can provide different information for the same scene according to the reflectance of material in each wavelength. Based on this fact, we propose the SS-SIFT method to extract spectral-spatial features. These features are invariant to geometric transformation, which include rotation, translation and scaling, as well as partially invariant to illumination changes. More importantly, these features are invariant to spectral changes which means HSI can be captured with
different incident lighting condition and viewing angle, or using different hyperspectral cameras with distinctive sensitivity functions from their imaging sensors.

**B. SIFT for Multispectral Image**

Multispectral images and HSIs provide both spatial and spectral information that is useful for many applications such as object recognition and registration [1]. Although SIFT descriptor is superior for greyscale image matching and registration, its performance declines on multispectral image and HSI. This is because the number of incorrect matches of keypoints due to significant differences of spectral image intensities [16].

Yi et al. [16] proposed a scale restriction criteria to reduce the number of outliers and improve the matching performance. Scale difference ratio is calculated for each keypoints matching. Correct matching has much smaller scale difference ratio than that of incorrect matching. An orientation restricted SIFT approach was proposed by Vura et al. [21]. A SIFT descriptor is represented as a combination of histogram bins that have 180° distance in orientation between each other. The results show that this modification provides better performance than the method by Yi et al. [16]. Saleem and Sablatnig [22] pointed out that high spectral variations still affects SIFT performance even though the above modifications have been made. Therefore, they used local contrast and differential excitation functions to construct SIFT descriptor rather than calculating gradient magnitude. These new descriptors provided spectral invariant features that improved the performance for multispectral registration [22].

However, all previous methods applied SIFT for cross band matching of multispectral images without exploiting the whole spectral dimensions simultaneously.

**C. SIFT for Hyperspectral Image**

For remote sensing HSI, Xu et al. [17] proposed a 1D SIFT algorithm for image classification. In this work, spectral keypoints were extracted from pixels (spectral signatures) to achieve invariant spectral features. However, such spectral feature cannot distinguish different object structures with the same material.

Mukherjee et al. [23] extracted distinctive features from HSI by applying SIFT algorithm to perform image matching. In this method, principal component analysis (PCA) was first used to reduce the dimensionality of HSI and to combine spectral information. For each PC, scale-space representation was performed and DoG was applied. A nonlinear function was then used to combine DoG for each PC to produce a single scalar image. After that, they followed same Low’s keypoint detection method to extract the extrema. However, using PCA for dimensionality reduction and combining DoG for each PC make spectral information distorted or lost.

Dorado-Munoz et al. [24] proposed a vector SIFT method for HSI. This work considered the vectorial nature of HSI pixels, i.e., the spectral response at a pixel is a vector of wavelength indexed values. In this work, vector nonlinear diffusion function was used to construct the scale-space representation. After that, keypoints were detected by determining local maximum and minimum pixel-vector by comparing each pixel-vector with its neighbours according to their spectral values. The performance of this method was evaluated using repeatability criteria and image registration and the results showed that the detected keypoints are invariant to insignificant scale change. However, no keypoint descriptor was developed in this work, so the matching of detected keypoints could only be determined based on their spatial location [24]. As a consequence, it is difficult to apply this method to general scenes captured with free camera setting.

For real-life hyperspectral image registration, Mainali et al. [25] presented a scale-invariant feature detection algorithm (D-SIFER). In this work, authors proposed 10th-order Gaussian derivative (GDO-10) filter with the scale-space representation to obtain more robust keypoints for band image registration.

In summary, existing work on HSI were either applied to spectral or spatial dimension only to detect keypoints. In most cases, keypoint descriptor was not developed. Our method, on the contrary, fully explores both spectral and spatial data distribution to extract keypoints and to build two different types of distinctive descriptors.

**III. SS-SIFT for HSI**

In this section, we describe the details of the proposed SS-SIFT method which includes two main steps, 3D keypoint detection and 3D local descriptor construction.

**A. Spectral-spatial Keypoint Detector for HSI**

Keypoint detection is a pre-processing step of feature extraction method [29], [30]. Keypoint detector searches the whole image for the most significant and distinguishing points in image data. Spatially, these points can be at edges, corners, curvatures or complex objects [31], [32]. In HSI, 3D keypoint operator aims to exploit both spatial and spectral dimensions simultaneously for detecting distinctive points that are invariant to geometric transformation and spectral changes. More formally, the proposed operator detects keypoints in a 3D data cube by searching points in spatial location (x, y) at spectral band indexed by λ. These keypoints and their neighbours formulate structural features, which are distributed over all spectral bands at different spatial locations. Spectral-spatial keypoint detector includes the following main steps.

1) **Spectral-Spatial Scale Representation:** Smoothing an input image by Gaussian filter is the initial step for spectral-spatial scale representation. HSI is often affected by noises from sensors and the imaging process. This makes the spectral signature vary in different regions of object made of the same material [33]. In our work, a 3D Gaussian filter $G(x, y, \lambda, \sigma_1, \sigma_2)$ is adopted to smooth both spectral and spatial domains simultaneously

$$G(x, y, \lambda, \sigma_1, \sigma_2) = \frac{1}{2\pi\sigma_1^2\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x^2+y^2)}{2\sigma_1^2} + \frac{\lambda^2}{2\sigma_2^2}}$$  \hspace{1cm} (1)$$

where $\sigma_1 = 1.6$ is for the spatial dimension and $\sigma_2 = 1.8$ is for the spectral dimension. The value of $\sigma_1 = 1.6$ is based on
a special property: the Difference of Gaussian (DoG) approximates well the LoG (Laplacian of Gaussian) at $\sigma_1 \approx 1.6$. This value is the same as used in traditional SIFT. For $\sigma_2 = 1.8$, this value is selected based on the fact that the HSI is more noisy in the spectral domain. Therefore using a little bit higher $\sigma_2$ value leads to stronger smoothing effect and a better result for keypoints detection. However, we don’t want to over-smooth the image to affect the number of accurate detected keypoints. Therefore, $\sigma_2 \in [1.6, 2.0]$ is a good range.

The hyperspectral cube $I(x,y, \lambda)$ is convolved with the Gaussian filter to construct spectral-spatial scale $S(x,y,\lambda,\sigma_1,\sigma_2)$ for the image

$$S(x,y,\lambda,\sigma_1,\sigma_2) = I(x,y,\lambda) \otimes G(x,y,\lambda,\sigma_1,\sigma_2) \quad (2)$$

where $\otimes$ is a convolution operation in the $x$, $y$ and $\lambda$ space.

A pyramid of three levels (Octave=1, 2, 3) is constructed by sub-sampling HSI spectrally and spatially in each level. In other words, we down-sampling HSI spectrally by reducing the number of spectral bands $M$ in each octave repeatedly (i.e. Octave 1: $M$, Octave 2: $M/2$, Octave 3: $M/4$) while down-sampling the spatial image size simultaneously. In each octave, multiple spectral-spatial scale DoGs are calculated between each two contiguous scales separated by $k$ which is a multiplicative scaling factor:

$$D(x,y,\lambda,\sigma_1,\sigma_2) = I(x,y,\lambda) \otimes G(x,y,\lambda,k\sigma_1,k\sigma_2) - I(x,y,\lambda) \otimes G(x,y,\lambda,\sigma_1,\sigma_2) \quad (3)$$

As illustrated in Figure 1, the scales in each octave are blurred incrementally using spectral-spatial Gaussian convolution with $\sigma_1$ and $\sigma_2$ values multiplied by incremented $k$ factor. The $\sigma_1$ and $\sigma_2$ values for the next octave are then squared. We construct a pyramid of three octave by repeating this process. Subsequently, keypoints are detected by scanning the scale spectral-spatial $D(x,y,\lambda,\sigma_1,\sigma_2)$ image for local minima or maxima, i.e. the smallest or largest point value, in the DoG image. Note that Figure 1 only demonstrates two octaves.

2) Locating the Extrema of DoG: Our algorithm searches for keypoints in the spectral-spatial dimension simultaneously. We name these keypoints SS-point which is the abbreviation for spectral-spatial keypoint whose values are the radiance of object material at location $(x, y)$ and wavelength ($\lambda$).

In this work, each SS-point is represented as $P(x,y,\lambda)$, where $x = 1, 2, 3, ..., M, y = 1, 2, 3, ..., N$ and $\lambda = 1, 2, 3, ..., B$. $M$ and $N$ defines the spatial size of the image and $B$ is the number of spectral bands. For each $3 \times 3 \times 3$ overlapped window of SS-points in HSI, the center SS-point is compared with 26 neighboring SS-points in the same scale and with 27 SS-points in scales above and below respectively. The selected point will be the extremum ($Ex$) only if it is the largest or the smallest point of the 80 compared SS-points. If this point is not the largest or the smallest point, then in the same window we move to check if $P(x,y,\lambda-1)$ or $P(x,y,\lambda+1)$ is the largest or the smallest point of the 80 SS-points. Extrema detection can be summarize in algorithm 1.

Algorithm 1 Extrema Detection

1: for Each Octave do
2:     for Each $3 \times 3 \times 3$ overlapped window in scale$_i$ do
3:         if Central SS-point $P(x,y,\lambda) = \min | \max$(80 SS-points in scale$_i$, scale$_{i-1}$ and scale$_{i+1}$) then
4:             Extremum = $P(x,y,\lambda)$
5:         else if Central SS-point $P(x,y,\lambda-1) = \min | \max$(80 SS-points in scale$_i$, scale$_{i-1}$ and scale$_{i+1}$) then
6:             Extremum = $P(x,y,\lambda-1)$
7:         else if Central SS-point $P(x,y,\lambda+1) = \min | \max$(80 SS-points in scale$_i$, scale$_{i-1}$ and scale$_{i+1}$) then
8:             Extremum = $P(x,y,\lambda+1)$
9:     end if
10:     end for
11: end for

Essentially, each spectral band in HSI can provide different information according to the radiance from material in each wavelength. For this reason, some spatial structure may occur in some bands but not in others. Therefore, this comparison is used to increase the number of extracted keypoints and to obtain keypoints in most of spectral bands at the same spatial location. In the end, each keypoint has five parameters: $x$, $y$, $\lambda$, $\sigma_1$ and $\sigma_2$.

3) Keypoints Filtering: Keypoint detection operator can produce a vast number of extrema due to high HSI dimensionality. These SS-points need to be filtered to obtain accurate feature point localization. Filtering process includes eliminating low contrast points according to the extended quadratic Taylor expansion function that was developed by Brown and Lowe [34]. These points may produce unstable descriptors that lead to inaccurate matching. Rejection of the candidate extrema also applies to those that are poorly localized along edges. Responses of edges in 3D space can be calculated by a $3 \times 3 \times 3$ Hessian matrix as implemented with 3D SIFT in [18], [19], [20], [27], [28]. We extend the Hessian matrix as a spectral-spatial matrix $H = (D_{ij})_{3 \times 3 \times 3}$ that describes the local structural-spectral information at the candidate point in each 3D sub-
Figure 2: The detected keypoints in 16 spectral bands for Monkey HSI.

\( H = \begin{bmatrix} D_{xx} & D_{yx} & D_{x\lambda} \\ D_{xy} & D_{yy} & D_{y\lambda} \\ D_{x\lambda} & D_{y\lambda} & D_{\lambda\lambda} \end{bmatrix} \)  

(4)

where \( D_{ij} \) is the second derivative calculated at SS-point in the DoG scale space. Similar to the method in [19] and [20], trace and determinant of \( H \) are derived to facilitate the rejection process, where

\[
\text{Trace}(H) = D_{xx} + D_{yy} + D_{\lambda\lambda}
\]

(5)

\[
\text{Det}(H) = D_{xx}D_{yy}D_{\lambda\lambda} + 2D_{xy}D_{y\lambda}D_{x\lambda} - D_{xx}D_{y\lambda}^2 - D_{yy}D_{x\lambda}^2 - D_{\lambda\lambda}D_{xy}^2
\]

(6)

Finally, poorly localized SS-points can be removed if the following rule is verified

\[
\frac{\text{Trace}(H)^3}{\text{Det}(H)} < \frac{(2r + 1)^3}{r^2}
\]

(7)

We filtered out the detected keypoints that have greater than 20 ratio between the principal curvatures because \( r = 20 \) is the optimal value determined experimentally. We tried different values for \( r \) (\( r = 10, 30, 40, 50 \)). When \( r = 10 \), we still obtain poor edge points. When \( r \geq 30 \), we may lose some accurate detected points. Figure 2 illustrates an example of detected keypoints on a toy monkey image.

B. 3D Local Descriptor

After keypoints detection, two different spectral-spatial descriptors are constructed. Constructing these descriptors depends on the orientation of the spectral-spatial gradient magnitude of each keypoint neighbours.

The first descriptor \( D_{\theta,\phi} \) is a 3D histogram build upon two angle orientations \( \theta \) and \( \phi \), and the gradient magnitude are calculated in joint spectral spatial dimensions. The angle \( \theta \) indicates the orientation in the spatial dimension and the angle \( \phi \) is defined for the orientation in the spectral dimension, as shown in Figure 3. 3D spectral-spatial gradient magnitude \( \mathcal{M} \) and two orientations for each keypoint are computed as follows:

\[
\mathcal{M} = \sqrt{G_x^2 + G_y^2 + G_{\lambda}^2}
\]

(8)

\[
\theta(x, y, \lambda) = \tan^{-1}\left( \frac{G_y}{G_x} \right)
\]

(9)

\[
\phi(x, y, \lambda) = \tan^{-1}\left( \frac{G_{\lambda}}{\sqrt{G_x^2 + G_y^2}} \right)
\]

(10)

where \( G_x, G_y \) and \( G_{\lambda} \) are calculated using difference approximation along the \( x, y \) and \( \lambda \) dimensions respectively as
acousto-optical tunable filter (AOTF) and a highly sensitive two cameras support different spectral and spatial resolutions in the outdoor environment using two hyperspectral cameras. These spatial invariance. In the first step, for each SS-point, 16 spectral-spatial scale of the image in order to achieve spectral-vector for each keypoint. Consequently, we have 4 SS-points. Spectral-spatial gradient magnitude and two orientation angles \( \theta \) and \( \phi \) are calculated for each SS-point in the window. For each patch, the gradient magnitude, which is weighted by a Gaussian window centred at the keypoint, is accumulated at the corresponding bin for gradient orientation, where 8 bins covering 360 degrees are used for \( \theta \) and 4 bins covering 180 degrees are used for \( \phi \). Therefore, vector feature of \( 32 \times 4 \times 4 \times 2 = 1024 \) dimensions has been obtained for each keypoint. Finally, to reduce the influence of illumination condition, the feature vector is normalized to unit length. After that, as in Lowe’s method each value in feature vector larger than 0.2 will be set to 0.2 and then the feature vector is re-normalized. The threshold equal to 0.2 for the normalized feature vector is selected experimentally by using images with different illumination conditions for the same scene. This normalization is used to reduce the effect of illumination changes.

The second descriptor \( D_0 \) is created by concatenating three 2D histograms of gradient magnitude over orientation \( \psi \) on each dimension, i.e., in \((x, y), (x, \lambda), \) and \((y, \lambda)\) dimensions separately as shown in Figure 4. For each SS-point, gradient magnitude and orientation are calculated in a \( 16 \times 16 \) window of surrounding neighbours in each plane. For each \( 4 \times 4 \) sub-region, 2D histogram of 8 orientation bins is calculated. The orientation in each histogram bin is accumulated by gradient magnitude which is weighted by Gaussian window. Consequently, we have \( 4 \times 4 \times 8 \times 3 = 384 \) dimensional feature vector for each keypoint.

IV. HYPERSONTICAL DATA

The HSIs in our experiments were collected in the lab and outdoor environment using two hyperspectral cameras. These two cameras support different spectral and spatial resolutions and wavelength ranges. The first camera consists of a Brimrose acousto-optical tunable filter (AOTF) and a highly sensitive visible to infrared camera. It generates 61 spectral bands HSI with wavelengths range from 400nm to 1000nm at 10nm spectral resolution. Another camera is an XIMEA camera which uses IMEC CMOS sensor, providing HSI with 16 spectral bands starting at 467nm and ending at 641nm.

Many factors affect HSI quality such as incident light, camera focusing and distance between the camera and the objects. Furthermore, HSI suffers from noises due to narrow bands of light that can be sensed by the imaging sensor within a short period of time. Although our cameras are highly sensitive, the signal to noise ratio is still low in some bands, especially on the short wavelength range image collected by the AOTF camera. Therefore, we removed these bands from the images to reduce their impact on experiments. As a result, 40 spectral bands from 590nm to 980nm with 10nm spectral resolution is selected for HSIs from AOTF camera.

All images in our dataset are real word hyperspectral image captured inside or outside the lab environment. Unlike remote sensing images in which light shall travel long distance from objects to the imaging sensor, the distances between the objects and the camera in close range setting are not too far in most scenarios so the atmospheric effect is negligible. In addition, we performed a white calibration using a white balance reflectance target to reduce the influence of lighting condition. Converting radiance to reflectance requires further image processing and parameter estimation in a close range imaging setting, e.g., using the method in [35]. This is beyond the scope of this paper.

We collected HSIs inside the Spectral Imaging Lab of Griffith University for different types of plants and objects made of different materials. Furthermore, we collected HSIs with different views in the outdoor environment. Some of the collected images are shown in Figure 5 and Figure 6.

V. KEYPOINTS MATCHING PROCESS

In this work, the performance of the SS-SIFT is evaluated through image registration by keypoint matching. Euclidean distances of descriptors are calculated between all keypoints in two images \( I_1 \) and \( I_2 \). A match has been detected if the minimum Euclidean distance between descriptor of one SS-point in \( I_1 \) and its nearest neighbour in \( I_2 \) is less than a threshold \( \xi \). The value \( \xi = 0.5 \) is experimentally determined, which is used to choose the nearest match for each keypoint detected on images captured by different cameras. To match images captured by the same camera, we used \( \xi = 0.2 \). In reality, images captured by different cameras is a more
common case. Matching such HSIs requires higher threshold value. In our experiments, the best matching result was obtained using $\xi = 0.5$. For more effective matching, SS-points that have more than one matches with the same distance are discarded to reduce the number of incorrect matches and eliminate ambiguous matching. Matching percentage between two corresponding images is calculated as follow:

$$\text{Matches\%} = \frac{\text{MatchNum}}{\min(Np_1, Np_2)} \times 100 \quad (14)$$

where $Np_1$ is the number of detected SS-points in image $I_1$, $Np_2$ is the number of SS-points in image $I_2$, and $\text{MatchNum}$ is the number of matched points.

In SS-SIFT algorithm, for both descriptors, spectral-spatial features are extracted locally depending on the surrounding neighbours of SS-points which is 16 spatial neighbours and 8 spectral neighbours. Accordingly, objects with the same shape but different materials can be matched spatially but may not spectrally. For example, in Figure 7 we have an HSI of two minions toy with the same spatial shape but they are made of different materials, one from plastic and the other from cloths. It can be seen that the spectral signatures of specific spatial $(x, y)$ location in HSIs captured by the same camera are different spectrally. Furthermore, the spectral signatures of the same $(x, y)$ location in the same HSI that are captured using different cameras are not identical. This difference is caused by different wavelength ranges and spectral resolutions between two cameras. Matching such HSIs is the main challenge for our work. Therefore, spectral similarity measures such as spectral angle mapper (SAM) and spectral correlation function (SCF) [36] are calculated between spectral signatures of each matched pair of SS-points $x_i$ and $y_i$ to verify high spectral matching:

$$\text{SAM} = \arccos\left(\frac{\sum_{i=1}^{B} x_i y_i}{\sqrt{\sum_{i=1}^{B} x_i^2} \sqrt{\sum_{i=1}^{B} y_i^2}}\right) \quad (15)$$
Figure 7: Spectral signatures of specific location pixels for minions objects. (a) RGB image. (b) Band image at 780nm wavelength of HSI captured using AOTF camera. (c) Band image at 548nm wavelength of HSI captured using XIMEA camera.

\[ SCF = \frac{B \sum_{i}^{B} x_i y_i - \sum_{i}^{B} x_i \sum_{i}^{B} y_i}{\sqrt{[B \sum_{i}^{B} x_i^2 - (\sum_{i}^{B} x_i)^2][B \sum_{i}^{B} y_i^2 - (\sum_{i}^{B} y_i)^2]}} \]  

(16)

where \( i = 1, 2, 3, ..., B \) and \( B \) is the number of overlapping spectral bands.

We adopted RANSAC [37] algorithm to remove the incorrect matching (outlier) and estimate the transformation between two HSIs. RANSAC algorithm is applied only spatially on single spectral 2D band since there is no straightforward structural information in the spectral dimension. After that, we investigate if each two matched points (inlier) belongs to similar spectral band and then verify spectral-spatial matching \( M_{ss} \) as follows

\[ M_{ss} = \frac{M_{SPEC}}{NumInlier} \times 100 \]  

(17)

where \( M_{SPEC} \) is the number of spectrally matched points and NumInlier is the number of spatially matched points (inliers).

VI. EXPERIMENTAL RESULTS AND EVALUATION

The experiments are divided into two parts to evaluate the performance of SS-SIFT descriptors (\( D_{\theta, \phi} \) and \( D_{\theta} \)). We first compare and register HSIs captured using the AOTF camera (40 spectral bands) and the XIMEA camera (16 spectral bands). Then we match each HSI pair captured using the same camera (AOTF or XIMEA camera). The experiments lead to both quantitative and qualitative results for the matching process between any two hyperspectral cubes. Due to the limitation of space, we display the matching results on central band of each HSI.

A. Keypoints matching for HSIs Captured Using Different Cameras

Different hyperspectral cameras produce HSIs with different spectral dimensions and resolutions. Matching such HSIs is a challenging task and it is the focus of our work. In the following experiments, we test AOTF and XIMEA HSIs which include both spectral and geometric transformation (scale, rotation and translation) changes. The spectral-spatial matching \( M_{ss} \) is calculated on the overlapped spectral range of both cameras which is 590nm to 650nm for the AOTF camera and 580nm to 641nm for the XIMEA camera.

Cross-Spectral Camera Matching

Figure 8 shows the matching results for HSIs of two toy minions made from different materials. The left image was captured by the AOTF camera and the right image was captured using the XIMEA camera. The number of matched SS-points using the first descriptor is 710, consisting of 137 inliers that include 30 spectral-spatial matches. The number of matches for the second descriptor is 659, consisting of 124 inliers that includes 35 spectral-spatial matches. The number of matched SS-points are distributed over all spectral bands of each HSI. The result shows that both descriptors provide a high performance of matching for HSIs with different spectral dimensions and resolutions. Furthermore, our method is robust for matching objects with the same material and spatial shape, as shown in Figure 8.

Another experiment for cross-spectral camera matching is conducted on outdoor HSIs. Figure 9 shows the matching result for recycle bin HSIs. In this experiment, outdoor HSIs have been captured using both AOTF and XIMEA cameras. The number of SS-points using descriptor \( D_{\theta, \phi} \) is 931, consisting of 138 inliers which includes 46 spectral-spatial matches. For descriptor \( D_{\psi} \), the number of matched SS-point is 942, consisting of 136 inliers with 33 spectral-spatial matches.

Cross-Spectral Camera Matching and Rotation Invariance

The purpose of the second experiment is to show the performance of matching HSIs with both spectral scale changes and rotation changes. In Figure 10 The left HSI is from XIMEA camera and the right HSI is from AOTF camera with 10° rotation in the spatial domain. Descriptor \( D_{\theta, \phi} \) produces 530 matched SS-points with 129 inliers that includes 32 spectral-spatial matches whereas descriptor \( D_{\psi} \) produces 277 matched SS-points with 95 inliers including 37 spectral-spatial matches.
In this experiment, we select two HSIs for coffee can as shown in Figure 11. The left HSI was captured using AOTF camera and the right image was captured by XIMEA camera with spatial scale value 0.7. We obtained 492 matched SS-points which included 140 inliers with 42 spectral-spatial matches using $D_{\theta,\phi}$ descriptor. For $D_{\psi}$ descriptor, we obtained 463 matched SS-points consisting of 148 inliers with 37 spectral-spatial matches. We can observe that both descriptors $D_{\theta,\phi}$ and $D_{\psi}$ provide large number of matches in spite of the different spectral and spatial scales between the HSIs.

**B. Matching HSIs Captured by the Same Camera**

For HSIs from the same camera, spectral changes occur due to influence of light condition (amount of incident light and angle), scale, rotation or viewing angle. Both cameras were used to capture HSIs with different spectral and spatial changes.

**Rotation Invariant**

Figure 12 illustrates the results of matching two strawberry plant HSIs captured by the AOTF camera with $15^\circ$ difference in viewing angles. For the AOTF HSI, the number of matched SS-points is very high given 40 bands for both HSIs. Consequently, the number of spectral-spatial matches is high as well.
In total, we obtained 4601 matched SS-points, consisting of 1202 inliers using descriptor $D_{\theta_{\psi}}$. 1058 of these inliers are spectral-spatial matches. Whereas, matching using descriptor $D_{\psi}$ produces 4156 matched SS-points that include 1175 inliers consisting of 1068 spectral-spatial matches.

Scale Invariant

Figure 13 illustrates the matching result with different scales using the two proposed descriptors. Both HSIs were captured using the AOTF camera. The number of matched SS-points using descriptor $D_{\theta_{\psi}}$ is 426, consisting of 187 inliers with 186 spectral-spatial matches. Whereas, using descriptor $D_{\psi}$ produces 422 matches which consists of 188 inliers with 182 spectral-spatial matches.

Blurrness Invariant

Figure 14 illustrates the matching results of two outdoor HSIs with different blurriness settings caused by focusing. Both HSIs were captured using the XIMEA camera. For the first descriptor, the number of matched SS-points is 816, consisting of 645 inliers with 506 spectral-spatial matches. For the second descriptor, we obtained 758 SS-points matches which consists of 530 inliers that includes 406 spectral-spatial matches.

Illumination partially Invariant

Our extracted features are partially invariant to change of illumination which affects the radiance. As in normal SIFT [15], adding a constant intensity value to each pixel value will not affect the gradient values, since the gradient is calculated from pixel differences. Increasing the contrast of image will affect the gradients but not the detection of the extrema which is directly related to the location of keypoints. Non-linear illumination change may affect some gradient magnitudes but will not affect the gradient orientation, so it will not make significant impact on the detection of keypoints at local minimum or maximum, which though may be slightly shifted to neighboring bands. For this reason, the impact of large gradient magnitude is decreased by thresholding the feature vector values to be no larger than 0.2 as did in SIFT.

To experimentally evaluate the influence of changing illumination, we have done experiments with or without white calibration step (which normalizes the image so as to remove the influence of changing illumination in order to recover more accurate reflectance). The results show that there is not much difference between these two settings. Figure 15 shows the results of applying our method on a normalized radiance image and its calibrated reflectance image. Most detected keypoints from these two images are similar. To verify the similarity of the detected points and to show that both radiance and reflectance can provide a good result, we...
match both images using the proposed SS-SIFT descriptors and show the results in Figure 16. For the first descriptor, the number of inliers is 151 with 146 spectral-spatial matches. For the second descriptor, we obtained 146 inliers that includes 143 spectral-spatial matches. The matching seems to be very accurate, which proves that our method is partially invariant to illumination conditions.

C. Spectral-Spatial Object Detection

SS-SIFT can also provide a high degree of accuracy for spectral-spatial object detection in HSI. Figure 17 illustrates an example of detecting objects using two different HSIs. The left image contains a monkey toy that has been captured using the XIMEA camera with 16 spectral bands with the band image size of 659×602 pixels. The second image contains several objects including the same monkey toy, a coffee-can, a toy moose and a Rubik’s cube. This image was captured using the AOTF camera with 40 spectral bands and 956×1227 spatial dimension. The result shows a high performance of our method for object detection. Another experiment for spectral-spatial object detection in HSIs is shown in Figure 18. In this experiment, two HSIs were captured using the AOTF camera with 40 spectral bands and 956×1227 spatial dimension. The left image is for two minions toy and the right one is for objects including one of minions which is made of cloths. Although there are two minions in the first HSI, with SS-SIFT first descriptor \(D_{\theta,\phi}\), the matching is obtained just between minions that are made of cloths in both HSIs. For the second descriptor of SS-SIFT, although it provides high spectral-spatial matching and corresponding features are matched spectrally, it still get some incorrect spatial matches. For example, we can observe in Figure 18(b) that there is a wrong matching between SS-points in plastic minion and paper coffee can.

VII. METHOD COMPARISON AND EVALUATION

A. SS-SIFT Versus 3D SIFT

In order to evaluate our SS-SIFT method, its performance is first compared with 3D SIFT that has been proposed for 3D image data [28], [27]. With this method, features are extracted from the whole HSI cube simultaneously. Table I shows the number of SS-points detected by SS-SIFT and 3D SIFT on images used in experiments of section VI-A and section VI-B. From Figure 8 to Figure 14 we mark left image as ‘a’ and the right image as ‘b’. As shown in Table I, our method produces much higher number of SS-points than 3D SIFT can detect. More extracted keypoints allows higher number of detected matches and inliers. The superiority of our method is based on the searching technique that explores all spectral bands to detect accurate spectral-spatial keypoint. According to the characteristic of HSI, each spectral band has different information based on object reflectance at specific wavelength. Consequently, using our method, we can obtain from HSI all the significant object information, which is distributed over all spectral bands. In addition, many of the detected keypoints are at the same spatial location for many spectral bands. This detection emphasizes that these points represent significant information and can produce a robust feature. On the contrary, 3D SIFT can not produce spatially overlapped keypoints because the extrema are searched just in one spectral band.

For visual comparison, Figure 19 illustrates matching between two HSIs using 3D SIFT. These HSIs are the same that generated the result in Figure 8. The number of matched keypoints is 73, consisting of 16 inliers with 2 spectral-spatial matches only.

Visually and quantitatively, it is verified that our method can
obtain much better matchings than 3D SIFT for two HSIs that are spectrally different. 3D SIFT features has lower accuracy in dealing with spectral variance. In Figure 19, it can be seen that some matched points are incorrect because they are found between two minions with different materials.

B. SS-SIFT versus 2D SIFT

Most methods in the literature [16], [21], [22], [23], [25] modified 2D SIFT to overcome the spectral change in multispectral or hyperspectral images to obtain a robust spatial matching. These methods normally extract features band-by-band or by converting an HSI to a grayscale image before keypoint extraction. Processing HSI band-by-band focuses on geometric structure more than spectral properties of objects. In addition, converting HSI to single band or reducing the number of HSI bands causes loss of spectral information.

On the contrary, our SS-SIFT extracts spectral and spatial features from HSI cube simultaneously, which is a comprehensive solution to explore complete HSI information. Therefore, comparing or matching two HSIs means not only comparing the features spatially (shapes and structure) but also spectrally
Figure 18: Object detection using SS-SIFT. Left image is a band image of two minions toy and right image is a band image of one minion and coffee can. Both of them were captured using the XIMEA camera. (a) matching results using descriptor \( \hat{D}_{\theta,\phi} \). (b) matching results using descriptor \( D_\psi \).

<table>
<thead>
<tr>
<th>Method</th>
<th>Figure 8 a</th>
<th>Figure 9 a</th>
<th>Figure 10 a</th>
<th>Figure 11 a</th>
<th>Figure 12 a</th>
<th>Figure 13 a</th>
<th>Figure 14 a</th>
<th>Figure 8 b</th>
<th>Figure 9 b</th>
<th>Figure 10 b</th>
<th>Figure 11 b</th>
<th>Figure 12 b</th>
<th>Figure 13 b</th>
<th>Figure 14 b</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS-SIFT</td>
<td>4748</td>
<td>5390</td>
<td>2189</td>
<td>14490</td>
<td>2917</td>
<td>11997</td>
<td>1014</td>
<td>1712</td>
<td>11283</td>
<td>12014</td>
<td>1750</td>
<td>11283</td>
<td>12014</td>
<td></td>
</tr>
<tr>
<td>3D SIFT</td>
<td>575</td>
<td>374</td>
<td>831</td>
<td>3740</td>
<td>2762</td>
<td>735</td>
<td>393</td>
<td>1559</td>
<td>1529</td>
<td>522</td>
<td>578</td>
<td>1559</td>
<td>1529</td>
<td></td>
</tr>
</tbody>
</table>

Table I: Number of detected keypoints using SS-SIFT and 3D SIFT. ‘a’ for the left image and ‘b’ for the right image in these figures.

<table>
<thead>
<tr>
<th>Method</th>
<th>Figure 8</th>
<th>Figure 9</th>
<th>Figure 10</th>
<th>Figure 11</th>
<th>Figure 12</th>
<th>Figure 13</th>
<th>Figure 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS-SIFT</td>
<td>0.345</td>
<td>0.245</td>
<td>0.213</td>
<td>0.333</td>
<td>0.103</td>
<td>0.215</td>
<td>0.057</td>
</tr>
<tr>
<td>SS-SIFT ( D_\theta )</td>
<td>0.344</td>
<td>0.265</td>
<td>0.305</td>
<td>0.305</td>
<td>0.145</td>
<td>0.263</td>
<td>0.098</td>
</tr>
<tr>
<td>2D SIFT</td>
<td>0.367</td>
<td>0.495</td>
<td>0.357</td>
<td>0.442</td>
<td>0.137</td>
<td>0.275</td>
<td>0.069</td>
</tr>
</tbody>
</table>

Table II: Matching error using SS-SIFT and 2D SIFT for HSI.

Figure 19: Matching results using 3D SIFT. Left image was captured by the AOTF camera. The right was captured by the XIMEA camera.

Figure 20: Matching results using 2D SIFT.

In this experiment, we show the effect of the spectral information on spatial matching performance by comparing SS-SIFT method with 2D SIFT. The HSI are converted to single band grayscale image using PCA. After that, traditional 2D SIFT is applied on each image to extract spatial features which are then utilized for matching. In order to compare the results of 2D SIFT with those of our method, spatial keypoints matching error is calculated. For all correct matched points (inliers), the spatial coordinates of the points in geometrically transformed image \( I \) are \((x_{I1}, y_{I1}), (x_{I2}, y_{I2}), ..., (x_{In}, y_{In})\) where the coordinates for the corresponding points in image \( J \) are, respectively, \((x_{J1}, y_{J1}), (x_{J2}, y_{J2}), ..., (x_{Jn}, y_{Jn})\). Where \( n \) is the number of matched points. The root mean square error \( R_s \) for spatial matching is calculated as follow:

\[
R_s = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (x_{Ij} - x_{Jj})^2 + (y_{Ij} - y_{Jj})^2}
\]  

This error indicates the performance of the extracted features for the spatial matching between two images. The smaller error means the higher spatial matching. Table II shows the results of calculating spatial matching error \( R_s \) for all matched images in previous experiments from Figure 8 to Figure 14. We can
observe that for experiments in Figure 8 to Figure 11, SS-SIFT descriptors, esp. $D_{\phi,\psi}$, provide significantly smaller keypoints matching errors than 2D SIFT. Converting HSIs captured by different cameras to grayscale images leads to inconsistent intensities and loss of spectral distinguishing capability. As a consequence, many of the detected keypoints can not be matched accurately. If the same camera is used, i.e., for HSIs in Figure 12, Figure 13 and Figure 14, the errors from 2D SIFT are reduced and its performance is closer to our method.

However, matching two HSIs spatially without exploiting the spectral information leads to high matching error. For example, two objects have same shape but different materials can lead to incorrect matching if we depend only on spatial features. Figure 20 illustrates the matching for two HSIs, which are the same HSIs used in Figure 18 for SS-SIFT. 2D SIFT produced incorrect spatial matching on this image. This implies that extracting spectral-spatial features is more effective for matching HSIs.

C. Efficiency Comparison

The running time for the three methods (SS-SIFT, 3D SIFT and 2D SIFT) on several images is calculated. The results are shown in Table III. The complexity of our method increases linearly with respect to the number of spectral bands. The time complexity of applying 2D SIFT band by band is less than ours, this is mainly due to the less key point detected by 2D SIFT so the time for descriptor generation is much saved. Nevertheless, our method produces much accurate spectral-spatial features for image matching as illustrated in section VII-B. Compared with 3D SIFT which is applied on the whole HSI, our method has similar running time, but our method produces much higher number of detected keypoints than 3D SIFT as illustrated in Table I.

<table>
<thead>
<tr>
<th>HSI</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS-SIFT</td>
</tr>
<tr>
<td>Minions AOTF in Figure 8</td>
<td>294</td>
</tr>
<tr>
<td>Minions XIMEA in Figure 8</td>
<td>108</td>
</tr>
<tr>
<td>Recycle bin AOTF in Figure 9</td>
<td>232</td>
</tr>
<tr>
<td>Recycle bin XIMEA in Figure 9</td>
<td>85</td>
</tr>
<tr>
<td>Coffeecan AOTF in Figure 11</td>
<td>109</td>
</tr>
<tr>
<td>Coffeecan XIMEA in Figure 11</td>
<td>31</td>
</tr>
</tbody>
</table>

Table III: Time cost of three methods under comparison.

VIII. Conclusion

In this work, we have presented a new method to extract 3D spectral-spatial scale invariant features from hyperspectral image. It explores both spectral and spatial dimension simultaneously, and allows keypoints be detected from the spectral-spatial scale space. Then descriptors are constructed from the spectral-spatial neighborhood of keypoints. This method is particularly useful to register hyperspectral images captured by different cameras with different number of spectral bands and wavelength ranges. We have evaluated this method on several sets of images captured in the lab and outdoor environment. The experimental results show that our method provides robust matching between these images and outperforms the traditional 2D SIFT and 3D SIFT methods.


