Machine Vision and Applications (2015) 26:975–990
DOI 10.1007/s00138-015-0714-x

ORIGINAL PAPER

Descriptor evaluation and feature regression for multimodal image analysis

Xuanzi Yong[1] · **Michael Ying Yang**[2] · **Yanpeng Cao**[3] · **Bodo Rosenhahn**[4]

Received: 13 September 2014 / Revised: 25 July 2015 / Accepted: 24 August 2015 / Published online: 15 September 2015

Abstract In this paper, we present feature descriptor evaluation and feature regression for multimodal image analysis. First, we compare the performances of several popular interest point detectors and feature descriptors from multimodal images with focus on visual and infrared images. The performances of detectors are evaluated mainly by the score of repeatability and accuracy and the descriptors are assessed by using the rate of precision and recall. Secondly, we analyze the relationship between the corresponding descriptors computed from multimodal images. The descriptors are regressed by means of linear regression as well as Gaussian process. Then the features on infrared images are predicted by mapping the descriptors from visual images to the infrared modality through the regression results. Predictions are assessed in two ways: the statistics of absolute error between true values and actual values, and the precision score of matching the predicted descriptors to the original infrared descriptors. We believe that this evaluating information will be useful when selecting an appropriate detector and descriptor for multimodal image analysis. Also the experimental results show that regression methods achieve a well-assessed relationship between corresponding descriptors from multiple modalities.

✉ Michael Ying Yang
ying.yang1@tu-dresden.de

[1] Institute for Geodesy, TU Darmstadt, Darmstadt, Germany

[2] Computer Vision Lab (CVLD), TU Dresden, Dresden, Germany

[3] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

[4] Institute for Image Processing (TNT), Leibniz University Hannover, Hanover, Germany

Keywords Multimodal images · Feature regression · Feature performance

1 Introduction

Recent advances in imaging technique have resulted in an explosion in the use of multimodal images in a variety of fields, such as scene reconstruction, pose estimation and video surveillance. The popular multimodal sensors include visual RGB and infrared sensors, consumer RGB-D cameras, and Time-of-Flight sensors. The integration of images from these multiple sensors can provide complementary information and therefore increase the accuracy with an observed and characterized quantity [28].

Image feature detection and description play key roles in computer vision. In most applications, such as image matching and scene registration, key points detection or feature detection is regarded as the basic starting step in the procedure of processing. The behaviour of detectors decides the performance or even the success of the research largely. In order to choose the proper detector and descriptor for different scenarios, the evaluation is necessary for all the related work.

In the domain of computer vision, the detected points can be represented by some descriptors. In this way a point with its surroundings is described by a vector. Therefore, we represent an image with a set of vectors, which get rid of the noises and some unnecessary information. Moreover the computational costs are also reduced as well as the memory costs, which shows to be more efficient.

Given multi-modal images, a series of applications are provided such as matching objects and scene registration. It is easy to obtain the interest points from the given images and reform them by some feature descriptors. However, it

comes to a question if there is some relationship between the two corresponding feature vectors. Or can we get the feature descriptor of a point in infrared image by the corresponding point in visual image. Moreover, with a mapping function, a descriptor is mapped to an infrared image as a new vector. So how would this new vector looks like, where might this new vector locate in the infrared images. In this work, we address the aforementioned problems using regression methods. To the best of our knowledge, there is no existed work on analyzing the relationship among descriptors in multimodal images.

In this paper, we analyse the behaviour of detectors and feature descriptors in multimodal images with focus on visual and infrared images. In order to get a reliable result, we construct the datasets with different types of images from different categories. We present a comprehensive evaluation of the performance of interest point detectors and feature descriptors in multimodal images. The performances of interest point detectors are evaluated by the scores of repeatability and accuracy as well as quantity analysis. And in the case of feature descriptors, we analyse the rate of precision and recall to assess their performances.

Moreover we present an idea of feature regression from visual images to infrared modality, which indicates the existence of the relations between the descriptors. The regression is worked in two ways: linear regression and Gaussian process for regression (GPR). The former has a computational advantage that it runs faster and costs lower than other common regression methods. And the latter can obliquely represent the underlying regression function without claiming, but rigorously. As a result, the descriptors of points detected in visual images are mapped as the descriptors from infrared images. We evaluate the performances of linear regression mainly by the value of coefficient of determination and the results are evaluated by the mean and variance of error between the descriptor vectors. In order to assess the performance of Gaussian process regression, we apply the regression result to the application of matching. The results are evaluated by the precision of matching. Moreover the regression error is considered as the criterion as well. From the results, we can find that, based on specific covariance functions and inference methods, the regression process performs well that the predicted descriptors are similar to the actual descriptor vectors. Example results of GPR are shown in Fig. 1.

The rest of the paper is organized as follows: Section 2 presents the related work of this paper and Sect. 3 details the evaluation strategy. In Sect. 4, we provide a short review of related detectors and descriptors and the experimental results are analysed in Sect. 5. Section 6 presents the other part of work, namely our two regression methods for multimodal image analysis: linear regression and Gaussian process for regression. Section 7 shows the results of two regression methods for the feature descriptors in multimodal images. Finally, this work is concluded in Sect. 8.

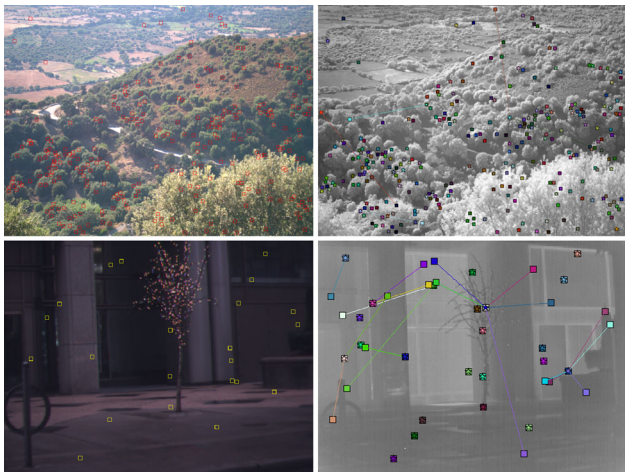


Fig. 1 Example results of GPR using SIFT. The *left column* are original visual images with detected *points*, which are also displayed in the form of *star* on the *right side images*. And the *squares* of the infrared images on the *right column* represent the relocated descriptors that are achieved by Gaussian process regression. The *lines* here connect the corresponding points between the multimodal images

2 Related work

On account of the development of sensor fusion technique, many applications were developed under multiple modality. Morris et al. [22] explored a statistic research on analyzing the significant characters on infrared images. Compared to corresponding visual images, the infrared images had noticeably less texture indoors because of the homogeneous temperature. Further, the joint wavelet statistics presented strong correlation between object boundaries in visual and infrared images, which could be used in vision applications with the combined statistical model. Moreover an overview of registering different types of sensors was provided by Zitova and Flusser [32]. Hrkać et al. [11] studied an approach to multimodal image registration based on corners and Hausdorff distance. The approaches using the mutual information as the matching criterion are the state-of-the-art technique in multispectral matching [12,17]. Due to the points and the contours of infrared images are different enough relative to visual images of the same scene, this region-based technique performs relatively well. Han et al. [9] implemented a line-based global transformation using the edge properties for the image registration between visual images and infrared images. Besides, Firmenichy et al. [6] provided a feature based matching and multimodal RGB to NIR registration with multispectral interest points. Furthermore an experiment for multimodal 2D and 3D face recognition was presented by Chang et al. [4]. Bansal and Daniilidis [1] pursued on the problem of matching images with disparate appearance arising from factors such as dramatic illumination (day vs. night), time period (historic vs. new) and

rendering style differences. By using the eigen-spectrum of the joint image graph, the persistent features were detected and matched into pairs.

Additionally, several researches are focusing on the features in some other area of multimodality. For instance Lu et al. [16] presents a novel binary range-sample feature in depth for action recognition, which has a great advantage in running speed and preserves the invariance against scale, viewpoint and background. It samples the pixel pairs from activity layer as well as the pairs from background and activity layer respectively and then generates the two bits feature by τ tests, by which the sign and amplitude of depth difference between pixel pairs are encoded. Besides, Xia and Aggarwal [30] presented an algorithm to extract spatio-temporal interest points (STIPs) from depth videos named DSTIP, whose response function is calculated by combining a 2D Gaussian smoothing filter on spatial dimensions and a temporal filter. Meanwhile a novel depth cuboid similarity feature (DCSF) is proposed, which described the spatio-temporal shape of the 3D cuboid around DSTIP by self-similarity and showed good performance on action recognition. And Ni et al. [23] introduced two multi-modality feature representations for activity recognition. One is a Depth-Layered Multi-Channel STIPs (DLMC-STIPs), which extended the STIPs as a multi-depth channel histogram representation. And the other is a Three-Dimensional Motion History Images (3D-MHIs) approach. Comparing to the original MHI, this fusion scheme used an additional depth sensor and therefore can contain forward as well as backward motion histories.

Furthermore, numbers of works studied on the performance of detectors and descriptors. In [26], the repeatability rate and information content of interest points were introduced in evaluating different interest point detectors. Mikolajczyk and Schmid [18] presented a comparative evaluation of affine invariant interest point detectors and Mikolajczyk et al. [20] evaluated the performance of affine region detectors under varying imaging conditions. Besides Moreels and Perona [21] explored the performance of a number of popular feature detectors and descriptors in matching 3D object features. The relationship between the corresponding descriptors computed from visual and infrared images has been analyzed in [31]. In [19], a set of local descriptors such as shape context, steerable filters, PCA-SIFT, SIFT are evaluated by using criterion recall with respect to precision. They proposed an extension of SIFT that shows the better performance comparing to the original method.

Moreover Sedai et al. [27] presented a comparative evaluation of appearance descriptors as Discrete Cosine Transform and the Histogram of Shape Context, and shape descriptors such as several variants of the Histogram of Oriented Gradients (HOG) descriptor for 3D human pose estimation using the Relevance Vector Machine regression and K-nearest neighbour regression methods. Specific to visual SLAM, [8] compared the behaviour of different interest point detectors and local descriptors. Besides, [7] evaluated the interest point detectors and feature descriptors for real-time visual tracking comprehensively by combinations with detectors and descriptors. And the method turned out to be appropriate for relevant factors such as performance measures, testbed, etc.

3 Evaluation of detectors and descriptors in multi-modal images

3.1 Evaluation of interest point detectors

Since the performance of computer vision applications depends on the robustness of detection, the repeatability score is taken as the evaluation criterion in this paper. It is defined as the ratio of the number of detected keypoints in both RGB and infrared images and the amount of the in RGB image detected points for each image pair, which is formed as

$$\text{Repeatability} = \frac{\#\{p|p \in IR \cap RGB\}}{\#\{p|p \in RGB\}}, \tag{1}$$

where IR and RGB refer to the sets of detected points in infrared (or near-infrared) images and corresponding visual images perspectively.

The feature detectors with a high score of repeatability present a robust performance between the image pairs: the keypoints detected in visual images by these detectors can be likely detected in the corresponding infrared images as well. In practice, it makes sense in the applications such as matching in multi modal images. The points detected in both modal images show the very important information of the images, especially the similarity and the relationship between the correspondences.

Another criterion is accuracy of the detectors in the form of

$$\text{Accuracy} = \frac{\#\{p|p \in IR \cap RGB\}}{\#\{p|p \in IR\}}. \tag{2}$$

It represents the proportion of the interest points from infrared images that can be also detected from visual images. A detector with high accuracy means that most interest points detected from infrared images can be found in visual images, namely there is no many points that appear only in infrared images.

In a word, a detector with high repeatability and accuracy can easily catch the commonalities between visual images and infrared images. Therefore they can be used in multi modal matching and perform very well theoretically.

After the procedure of detection, data are stored with the location of the points in two sets, IR and RGB. Since the two images in a pair are in the same position condition, the two nearly close points indicate an interest point pair. Assuming the third set, which is the joint of the two sets mentioned before, the interest point pairs are put into it. Based on the Eqs. 1 and 2, it is easy to get the evaluation results.

3.2 Evaluation of feature descriptors

The criteria of the evaluation on descriptors are precision and recall, which are expressed as follows:

$$\text{Precision} = \frac{\#correct\_matches\_retrieved}{\#matches\_retrieved}, \tag{3}$$

$$\text{Recall} = \frac{\#correct\_matches\_retrieved}{\#correct\_matches}. \tag{4}$$

In the Expression 3, *precision* expresses capability to obtain the correct matches. A feature descriptor with low precision indicates the poor competence to get the potential important information of the interest points. And *recall* represents the ability of finding all the correct matches as in the Expression 4. The factor $\#correct\_matches\_retrieved$ represents the number of correct correspondences obtained, the variable $\#matches\_retrieved$ is the number of matches given by the matching methods with a threshold and the variable $correct\_matches$ refers to the total number of correct correspondences.

The criteria for a correct matches depend on the location of the correspondence. In this paper, the threshold of the distance different is within 5 pixels in Euclidean Distance. Namely, if the two components in a correspondence are at a distance from 5 pixel, then they are considered as a correct match. While the two images in a pair are in the similar scene, this judgement is reasonable to draw a conclusion.

Considering SIFT [14] and SURF [2], which are all common local descriptors in computer vision applications, the keypoints are detected by DoG and Hessian respectively. And then the descriptor vectors are computed to represent the information in the range within the neighbourhoods, especially the interest points.

Given the feature descriptors in two datasets, the vectors are matched by the algorithm represented in [15], which can efficiently reject the matches that are too ambiguous. A threshold is specified in this method, and a descriptor is matched to another descriptor only if their distance times the threshold is not greater than the distance of first descriptor to any other descriptors. In our experiment, the threshold is set as 1.5.

Aiming at feature descriptor LBP and HOG, the process is different. The descriptors are applied to the whole image, i.e., we get a LBP and a HOG vector for each image. Then the matches are obtained by the descriptors describing the whole images. As above, the criterion approach is to compare the precision and recall of the matches among images. Obviously, the correct matches should be the image pairs with the same scene. In this way, LBP and HOG are more like global descriptors rather than local descriptors.

4 Detectors and descriptors

In this section, several state-of-art detectors and descriptors are introduced respectively.

4.1 Interest point detectors

Interest point detectors are used for extracting features from images. And an interest point refers to the point that differs in properties, such as brightness comparing to its surroundings. It may be an independent point, a corner or on edges. There are various of detectors to extract these kinds of features.

In this subsection, five detectors, i.e., Harris corner detector [10], Difference of Gaussian [14], Harris Laplace detector [13], Hessian detector and Hessian Laplace detector [18], will be introduced in detail.

4.1.1 Harris corner detector

The algorithm of this detection focuses on each pixel of images. It calculates the gradient for each pixel and looks for the pixel with maximal gradient. A matrix with respect to every pixel (x, y) is needed, which is an approximation to the local auto-correlation function of image I:

$$M = \sum w_{u,v} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix},$$

where I_x and I_y denote the derivatives of image I, and $w_{u,v}$ specifies the window: $w_{u,v} = exp\{-(u^2 + v^2)/2\sigma^2\}$. A region is classified as a corner, if the eigenvalues λ_1, λ_2 of M are both large.

4.1.2 Difference of Gaussian (DoG)

DoG [14] was first introduced as a part of SIFT. Given a single image I, a DoG pyramid is build with several images as output, each being a unique difference of Gaussian. The input image is blurred by different scales σ. And one "Octave" is no other than the difference between consecutive blur amounts:

$$\begin{aligned} DoG_{k,\sigma} &= G(x, y, k\sigma) - G(x, y, \sigma) \\ &= \frac{1}{2\pi (k\sigma)^2} e^{-x^2+y^2/2(k\sigma)^2} - \frac{1}{2\pi (\sigma)^2} e^{-x^2+y^2/2(\sigma)^2}. \end{aligned}$$

The maximum and minimum are determined as interest points by comparing each pixel to its twenty-six neighbours including the 8 neighbours on the same octave and each 9 neighbours in the upper and lower octave.

4.1.3 Harris Laplace detector

In order to create a scale-invariant detector, the traditional 2D Harris corner detector are combined with a Gaussian scale space representation, that is the idea of Harris Laplace detector [13]. Since the points detected by Harris corner detector are not scale invariant, $M = \mu(\mathbf{x}, \sigma_I, \sigma_D)$ is denoted as the scale adapted second-moment matrix used in the Harris Laplace detector:

$$M = \mu(\mathbf{x}, \sigma_I, \sigma_D) = \sigma_D^2 g(\sigma_I) \otimes \begin{bmatrix} L_x^2(\mathbf{x}, \sigma_D) & L_x L_y(\mathbf{x}, \sigma_D) \\ L_x L_y(\mathbf{x}, \sigma_D) & L_y^2(\mathbf{x}, \sigma_D) \end{bmatrix},$$

where $g(\sigma_I)$ is the Gaussian kernel of scale σ_I and $\mathbf{x} = (x, y)^{\mathrm{T}}$. Here is $L(\mathbf{x}, \sigma_D)$ the Gaussian-smoothed image by Gaussian kernel with scale σ_D. The Laplacian scale selection procedure is applied to find the characteristic scales at the locations of interest points detected by Harris corner detector.

4.1.4 Hessian detector

The approach of Hessian detector searches the points in given image I, which have strong change in gradient along both orthogonal directions. Hessian Matrix is calculated issued from Taylor expansion:

$$H = \begin{bmatrix} I_{xx}(\mathbf{x}) & I_{xy}(\mathbf{x}) \\ I_{xy}(\mathbf{x}) & I_{yy}(\mathbf{x}) \end{bmatrix}.$$

The points are considered as interest points, which are the maximum by the determinant of H within a 3×3 window. This method responses mainly on corners and strongly textured areas.

4.1.5 Hessian Laplace detector

Hessian Laplace detector [18] is a scale invariant detector used in blob detection. Similar to Harris–Laplace detector, it uses the Hessian matrix to locate points in space and the Laplacian function to compute their scales.

4.2 Feature descriptors

Descriptors are used to represent the image structure in spatial neighbourhoods at a set of feature points. There are various kinds of descriptors, and we can choose an appropriate one based on the application. In this subsection, we will present four descriptors used in this paper, namely SIFT, SURF, LBP and HOG.

4.2.1 Scale-invariant feature transform (SIFT)

SIFT is based on the interest points detected by Difference of Gaussian [14]. The descriptor records the direction for each interest point, thus it has good scale and rotational invariance. A key point is characterized with location, scale and direction. The orientations of 16 × 16 neighbors of each keypoint are calculated and then projected into one of eight directions with 4 × 4 region. Subsequently, a histogram is built with 8 bins, which indicate 8 directions. As a result, the descriptor is in the form of vector with 128 dimensions. With the help of this descriptor, we can match key points between images.

4.2.2 Speeded up robust feature (SURF)

SURF is an improvement of SIFT, which is first presented by [2]. It is claimed that it performs excellent on repeatability, distinctiveness and robustness. The interest points are detected using Hessian matrix, that is named as Fast Hessian detector, which is calculated for each point. To solve it, SURF makes efficient use of integral images. Then by comparing each point with its 26 neighbours on the same octave and the octave above and below, the points with maximum or minimum responses are considered as interest points after filtered by given threshold. The descriptor is based on sum of Haar wavelet responses within the region in the size of 4 × 4, instead of histogram in SIFT, which is in the form as:

$$\sum d_x \sum d_y \sum |d_x| \sum |d_y|,$$

d_x and d_y are the filter responses to the Haar wavelets. Thus the output of SURF is a feature vector with 64 dimensions.

4.2.3 Local binary pattern (LBP)

Local binary pattern (LBP) is a type of texture spectrum model proposed in [29] and first described by Ojala et al. [24]. In this approach, an examined window is first divided into 16 × 16 cells. And then for each pixel in a cell, comparing the gray-value with other eight neighbors. It is assigned as 1, when the neighbor is greater than center pixel. Thus, an 8 bit binary pattern comes, i.e LBP. Compute the histogram of the frequency of each binary number occurring over the cell and normalize. The feature vector for the window should be the concatenate normalized histograms of all cells.

4.2.4 Histogram of oriented gradients (HOG)

HOG is first represented by Dalal and Triggs [5], which focuses on pedestrian detection at that time. And the essential

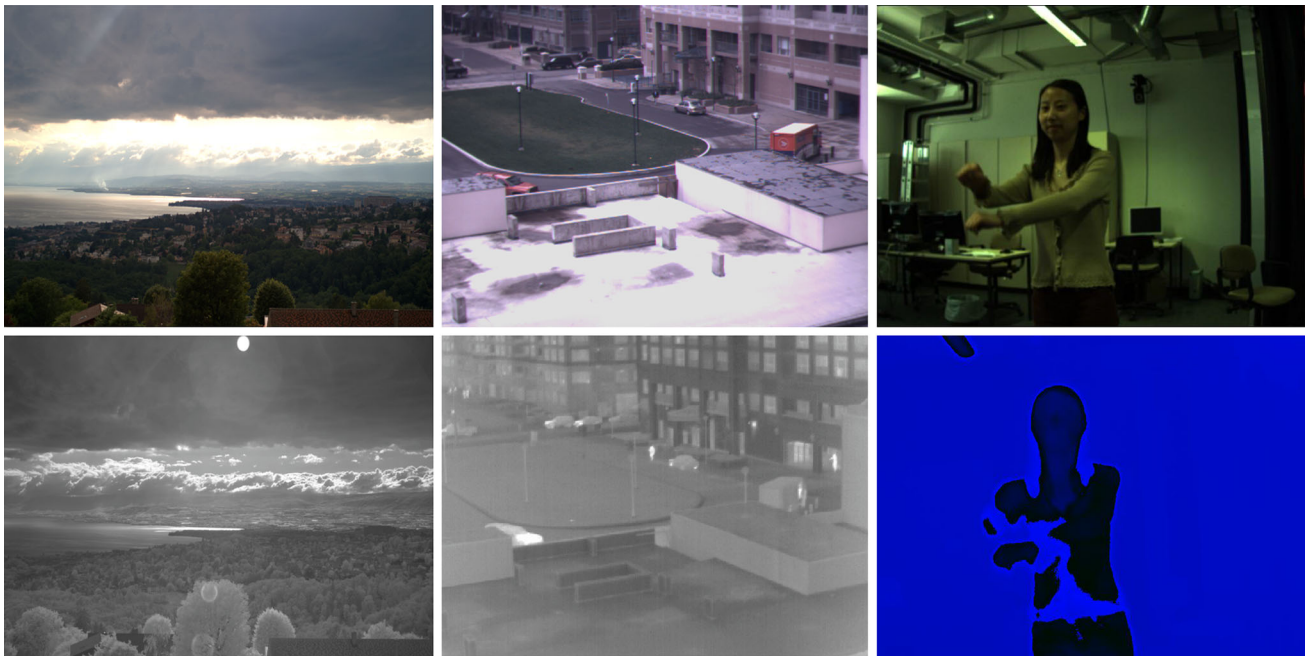


Fig. 2 Sample images from the datasets RGB-NIR, OutdoorUrban and MoCap respectively. The images in the *first row* are visual RGB images and the images in the *second row* are the corresponding infrared (or near-infrared) images with respect to the images above

idea behind the Histogram of oriented gradient descriptors is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. To implement it, the image need to be divided into small connected regions, called cells. And then compute the gradient for each pixel in the region of a cell. The histogram of gradient in each cell is the descriptor for the cell and the combination of these histograms presents the descriptor. In some advanced process, the cells are grouped into larger spatial blocks and these blocks are normalized separately. As a result, the final descriptor is exact the vector composed of all the components of the normalized cells by the blocks in the detection window.

5 Experiments of evaluation

5.1 Datasets

We use three datasets for evaluating interest point detectors and feature descriptors, namely RGB-NIR scene dataset from Image and Visual Representation Group in EPFL, OutdoorUrban by Dynamic Graphics Project in University of Toronto and MoCap from Institute for Information Processing in Leibniz University Hannover.

These three datasets contain different image types and are about different views. The dataset RGB-NIR consists of numbers of images captured in RGB and Near-infrared (NIR) by visible and NIR filters using separate exposures from modified SLR cameras. There are totally 9 categories such as field, forest, mountain and water. And the images in the dataset OutdoorUrban are fully about the views around the city, such as cars, buildings and some other city scenes. Compared to the natural scenery, there are greater thermal variations in urban environments [22]. In this dataset, there are total 290 outdoor urban daytime image pairs as well as about 30 urban night-time image pairs. The data are captured by a single axis, multiparameter camera which combines an infrared camera and a visible light camera. And the content of dataset MoCap is the human motions such as waving, boxing and jogging indoors. Since the two cameras are set with a small baseline, the taken images are not identical in view, but nor too far away from each other. Some samples of the datasets are shown in Fig. 2 and the basic information is summarized in Table 1.

Table 1 Information of datasets

Dataset	Type	Data number	Image size	Contents
RGB-NIR	Near-infrared	370	640×480	Nature views
OutdoorUrban	Infrared	330	384×288	City views
MoCap	Infrared	1300	640×480	Human motions

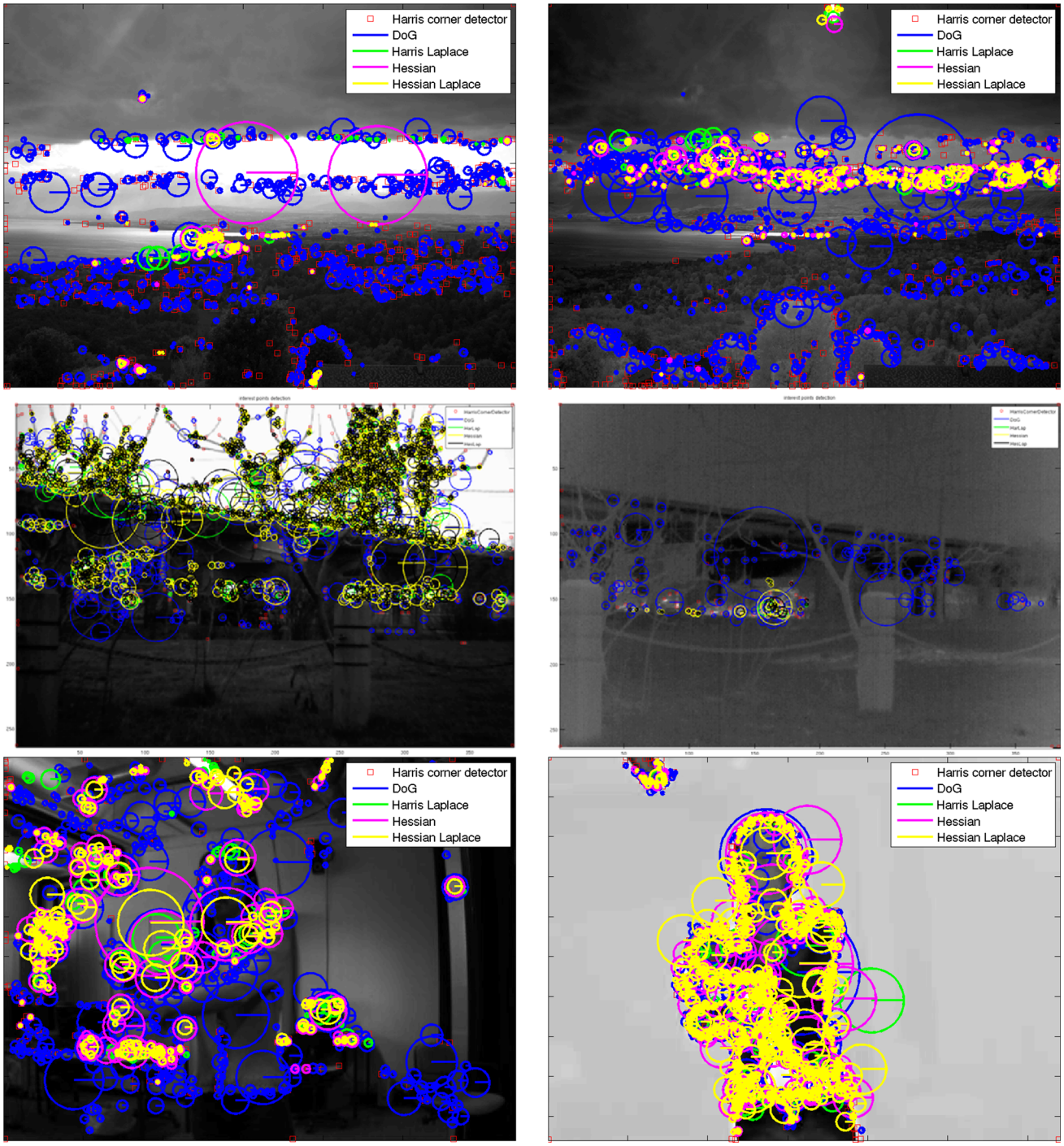


Fig. 3 The example images with detected points by different detectors, where each feature refers to a detected point. In details, *red* is Harris corner detector, *blue circles* refer to DoG, the *green* ones are detected by Harris Laplace detector, *magenta* refers to Hessian detector and the *yellows* are Hessian Laplace detector (color figure online)

5.2 Results of detectors evaluation

For a fair comparison, 100 image pairs are firstly random selected from each dataset. The interest points are extracted by the detectors introduced in Sect. 4.1 and the example results are shown in Fig. 3. The results are presented as illustrated in Figs. 4, 5 and 6. From the figures, it is obvious that Harris corner detector and Difference of Gaussian detection stand out among the five evaluating objects on both repeatability and accuracy.

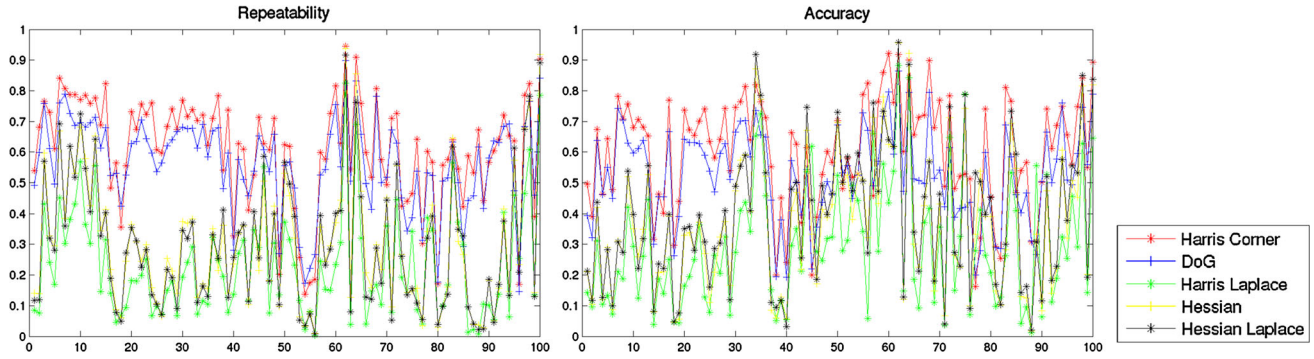


Fig. 4 The repeatability (*left*) and accuracy (*right*) performance of detectors on dataset RGB-NIR. The detectors are distinguished by *colours*, where *red line* refers to Harris corner detector, *blue line* is DoG detector, *green line* refers to Harris Laplace detector, *yellow lines* refer to Hessian detector and last the *black lines* refer to Hessian–Laplacian detector (color figure online)

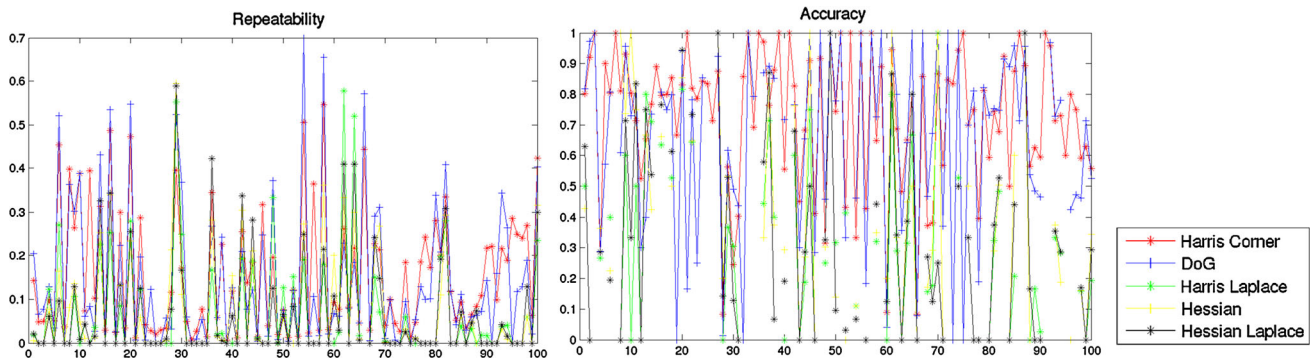


Fig. 5 The repeatability (*left*) and accuracy (*right*) performance of detectors on dataset OutdoorUrban

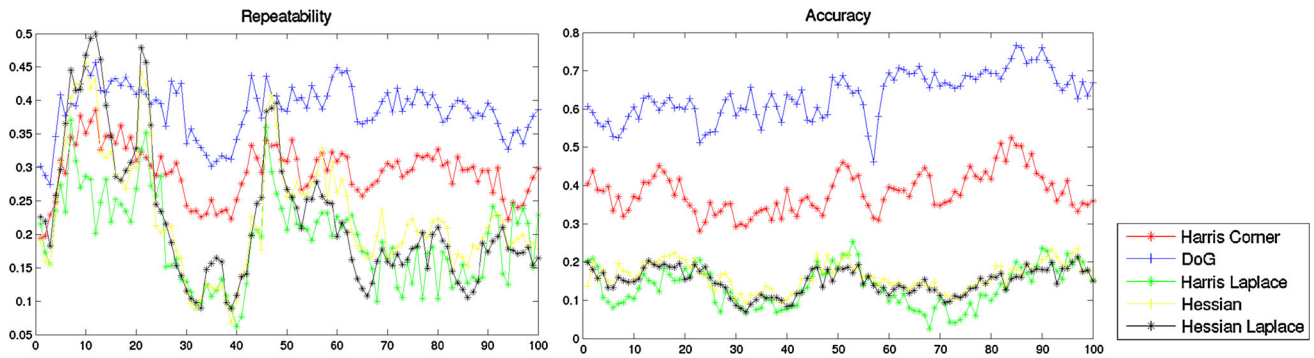


Fig. 6 The repeatability (*left*) and accuracy (*right*) performance of detectors on dataset MoCap

Firstly, we analyse the quantity of the detectors simply. The Harris detector and DoG detector explore much more keypoints than the others in the three datasets. However Harris Laplace detector performs the worst among them, which can detect only about 25 % of the amounts compared to Harris detector. Especially in dataset RGB-NIR, Harris detector extracts 3019 points in average, and DoG follows with 2153 detected keypoints. Besides, there is a distinct difference among the datasets drawn from our results. It shows that more visual detected points are explored than the infrared keypoints in dataset RGB-NIR as well as OutdoorUrban, however there are more keypoints detected from infrared images than them from visual images in the dataset MoCap. Since the MoCap camera is a kind of thermal camera that is sensible to thermal, the variations of body temperature are shown clearly, which however can not be seen in visual images.

On the perspective of repeatability, the red line referring to Harris corner detector (Har) and the blue line referring to Difference of Gaussian detector (DoG) lie higher than any

Table 2 Repeatability and accuracy of detectors

Dataset	Har	DoG	HarLap	Hes	HesLap
Repeatability					
RGB-NIR	0.8913	0.9166	0.6815	0.6622	0.6080
OutdoorUrban	0.3298	0.4314	0.4322	0.2881	0.3913
MoCap	0.2123	0.3124	0.1356	0.1494	0.1433
Accuracy					
RGB-NIR	0.9996	0.9949	0.9504	0.9574	0.6080
OutdoorUrban	0.7909	0.7724	0.6981	0.7885	0.8571
MoCap	0.3057	0.6377	0.1230	0.1264	0.1186

other detectors in dataset RGB-NIR as illustrated in Fig. 4. The score of repeatability of Harris corner detector is over 60 % and the DoG's is about 56 %. The distance between them is not so great, but the other detectors are standing quite far away from them. The repeatability scores of Harris Laplace detector (HarLap), Hessian detector (Hes) and Hessian Laplace detector (HesLap) are around 25 %, which are only about half of the scores on Harris corner detector and DoG. The situation of dataset OutdoorUrban is the same as in dataset RGB-NIR, but the degree falls down heavily in this dataset. The bad performance mainly dues to the low sharpness of the infrared images. In most images, there are only a few and even no key points detected from the IR image. Hence by Expression 1 the score is much less than expectation. Furthermore we can find that in all the three dataset, the green line referring to Harris-Laplace and the black line referring to Hessian Laplace and the yellow line referring to Hessian detection always accompany with each other.

The trend of accuracy is similar to the repeatability. Harris corner detector and Difference of Gaussian detection are better than the other three detectors.

Moreover, we extend the process on all the images from each dataset. And the result is shown in Table 2 in detail. The repeatability scores of Harris Laplace detector, Hessian detector and Hessian Laplace detector increase particularly twice than before. Considering accuracy, the first two positions belong to Harris corner detector and DoG detection as before. And the accuracy scores are all located in a reasonable range, which means that most key points detected from infrared image can be detected from the corresponding visual RGB image.

For dataset MoCap, the motion is classified with jogging and waving. The result shows that the performance of waving are overall greater than jogging. It is notable that the accuracy of detectors from dataset MoCap is lower than them on other datasets. The reason is, as mentioned before, that more interest points are detected from the body of humans in infrared images.

Notice that another reason is due to the content of the images. An unexpected noise arises from the appearance of a second person. While this person is standing behind the participant, he has not been captured by the visual camera. But because of a small angle between cameras, the shape of the person is shown in infrared images.

As a result, DoG stands in a dominant position comparing to any other detectors not only in repeatability but also in accuracy. This detector is strongly recommended in multi modal detections.

Table 3 Precision and recall of feature descriptor SIFT, SURF, HOG and LBP

	RGB-NIR		OutdoorUrban		MoCap	
	Precision	Recall	Precision	Recall	Precision	Recall
SIFT	0.8674	0.0200	0.0297	0.0000	0.0019	0.0000
SURF	0.7850	0.3400	0.0000	0.0000	0.0000	0.0000
HOG	1.0000	0.7000	0.4696	0.3333	0.7036	0.4538
LBP	0.9315	0.3676	0.3978	0.2126	0.3453	0.3440

5.3 Results of descriptors evaluation

As shown in Table 3, both SIFT and SURF perform well in the dataset RGB-NIR. And the average of precision of SIFT is 86.7 % that is about 8 % greater than it of SURF. However, the precision in other datasets are so terrible that there is few correct matches being found between a pair of images. The recall of SIFT compared to SURF in the dataset RGB-NIR is very low, mainly due to the less matches it has found.

From the Table 3, it is also clear that HOG and LBP both have a good precision score. Especially the precision of HOG is 100 % with recall being equal to 70 %. However the recall score of LBP is a little less with about 35 %.

The performance of HOG and LBP is worse in dataset OutdoorUrban than the other datasets. HOG still outperforms than LBP in this situation, but not significantly. In dataset MoCap, the precision score of HOG is over twice greater than the score of LBP. Nevertheless, there is no big difference of recall.

6 Feature regression analysis

Regression analysis is a statistical method to explore the dependent relationship between variables. A regression model consists of unknown parameter β, independent variables X as well as the dependent variable Y. It relates Y to a function of X and β as $Y \sim f(X, \beta)$, which describes the relationship between X and Y. In this way, Y can be predicted, given X.

Based on this principle, we propose an idea named *Feature Regression* analysis, in which the feature descriptor vectors

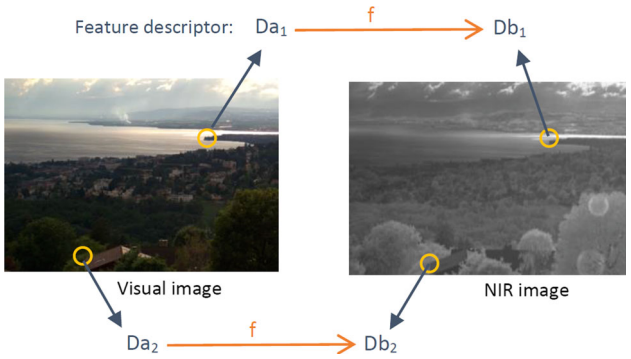


Fig. 7 A diagram of the general illustration of Feature Regression analysis

are regarded as variables of relation function. In our case, the feature descriptors from visual images are treated as source variables and the target variables are the ones from infrared images. The detailed introduction is as following:

For a scene I, $DA = \{Da\}_i$ and $DB = \{Db\}_i$ are sets of feature descriptors representing the interest points of I in modality A and modality B respectively, where the index i indicates the interest point. With the observations DA and DB and using the technique of regression analysis, an implicit or explicit mapping function is trained as $DB = f(DA)$, which relates the dependency between the feature vectors from modality A to B. An illustration of Feature Regression analysis is given in Fig. 7.

Once the mapping function is determined, the feature descriptor Db_j of an interest point j about another scene on modality B can be calculated, given the feature descriptor Da_j of the point on modality A, as $Db_j = f(Da_j)$. As in this work, the infrared features can be obtained through its correspondences from visual images.

Since linear regression is a common and light method used in statistical problems, it is considered in this work at first. Then Gaussian process is dealt as an advanced method, which can especially solve non-linear problems.

6.1 Linear regression

In statistics, linear regression is an approach to model the relationship between a scalar dependent variable and one or more explanatory variables, in which data are modeled by linear functions and unknown model parameters are estimated from the data [3].

Upon to the principle of linear regression, a descriptor Da with n dimensions from visual image is mapped to a descriptor as Db in the same dimension in infrared image through a matrix as linear transformation, which is given as:

$$Db = Da \times H \tag{5}$$

where H is a $n \times n$ matrix. H is firstly calculated by training and then used to predict the new input Da forward to a Db. The process of linear regression in our work is implemented by the method of Least squares, which is a technique for mathematical optimizing that the sum of the squares of the errors is minimized by equating its gradient to zero and then the regressors are obtained through the mean value.

6.1.1 Regression estimating

The regression procedure is estimated by a set of statistics, such as R^2, F, p and the estimate of the error variance $err.var$. R^2 is the *coefficient of determination* defined as

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \tag{6}$$

and SS_{tot} is the total sum of squares in the model that depends only on the dependent variables, namely Db here. And SS_{res} is the sum of squared errors in the linear model. It is a very important indicator to state if the regression is efficient while it informs the *goodness of fit* of a model. In regression, R^2 represents the percent of the data that is the closest to the line of best fit, in other words, it informs how well the regression line approximates the real data points. The F statistic is the test statistic of the F-test on the regression model, for a significant linear regression relationship between the response variable and the predictor variables. P value p is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. When the p value is less than the given significant level, in usual case as 0.05, the null hypothesis will be rejected. By using these arguments, the performance of linear regression is evaluated.

6.1.2 Predictions evaluating

After the regression procedure, the linear transformation matrix H is obtained and then used to predict the new descriptor Da forward to a Da'. For sake of predictions evaluating, we compare the prediction Da' and the true descriptor vector Db by the absolute difference between corresponding components in vectors as $\varepsilon = |Da' - Db|$. Two parameters are used to evaluate, that is $mean$, which is the average value of ε and the variance of ε.

However, the method of linear regression can not solve non-linear problems. Hence we use Gaussian process for regression as an advanced method, in which a specific model need not to be claimed at first.

6.2 Gaussian process for regression

Given some noisy observations of a dependent variable, the estimate of a new value x comes out easily by using a function

$f(x)$, which can describe the distribution of the observations. Rather than a specific model which the claimed function $f(x)$ relates to, a Gaussian process can represent $f(x)$ obliquely, but rigorously [25]. That is so-called Gaussian Process Regression (GPR).

Taking account of the noise on the observed target values from measurement errors and so on, which are given by

$$t_n = y_n + \epsilon_n \tag{7}$$

where $y_n = f(x_n)$, and ϵ_n is a random noise variable whose value is chosen independently for each observation n.

The conditional distribution of t_{N+1} given target values $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$ is itself Gaussian-distributed as the form:

$$t_{N+1}|\mathbf{t} \sim \mathcal{N}(\mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{t}, c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k}). \tag{8}$$

The mean, $\mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{t}$, is known as the *matrix of regression coefficients*, and the variance, $c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k}$, is the *Schur complement* of $\mathbf{C}_N$ in $\mathbf{C}_{N+1}$. These are the key results that define Gaussian process regression. While the vector $\mathbf{k}$ is a function with respect to the test input value $\mathbf{x}_{N+1}$, the predictive distribution is a Gaussian depended on $\mathbf{x}_{N+1}$.

As a crucial component of a Gaussian process predictor, covariance function controls how much the data are smoothed in estimating the unknown function [25]. Two functions are considered: the *squared exponential* (SE) covariance function has the form

$$k_{SE}(r) = \exp\left(-\frac{r^2}{2\ell^2}\right), \tag{9}$$

with parameter ℓ defined as *characteristic length-scale*. This covariance function has sample functions with infinitely many derivatives and thus is very smooth. Another is rational quadratic (RQ) covariance function

$$k_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{10}$$

with $\alpha, \ell > 0$, which can be regarded as a *scale mixture* (an infinite sum) of squared exponential (SE) covariance functions with different characteristic length-scales (sums of covariance functions are also a valid covariance).

The descriptors are treated in two ways:

6.2.1 Global descriptors

Assuming a particular structure, where the covariance function is set as the *squared exponential* function and the mean of Gaussian process is defined as zero like the assumptions in most cases. In addition, the Expectation propagation (EP) is applied as the inference function and the likelihood function is in the form of Laplace. The parameters of covariance function are initialized with zero at first and later they are optimized by minimizing their negative log marginal likelihood.

6.2.2 Local descriptors

The goal of this part is to check the potential location relationship of the descriptors. Based on the training data, the descriptor vectors **DA** from visual images are mapped to the infrared images as $\mathbf{DA}'$. For each descriptor Da in the set **DA**, we are looking for the most similar descriptors among all the descriptors **DB** in infrared images. In other words, a vector is predicted by a descriptor from visual image. And the task is to check the location of this vector in the infrared image.

The processing procedure is as following: first the interest points are detected from infrared and visual images and then represented by feature descriptors. Hence we obtain a set of vector pairs. Each vector consists of two parts, the descriptor of the interest points and its location. The next step is to obtain the predictions by Gaussian process. The initial hyperparameter of covariance function is set with 0.7. And then for one prediction vector, find the closest vector among all the original descriptor vectors in infrared images by using the Euclidean distance between two vectors. Moreover min-pooling approach is used to avoid too many incorrect matchings. In practice, assuming the infrared image and visual image display completely the same scene. Five candidates are chosen with most similar vectors. And then the prediction is determined to locate in the position of its nearest candidate.

7 Experiments of feature regression

We construct the experiments to regress the descriptors by using linear regression and Gaussian process. And the results are assessed by the criteria of *error* and the precision of matching. Three datasets are used in this paper the same as shown in Sect. 5.1.

7.1 Results of linear regression

Considering the regression, 300 images in RGB-NIR are selected in the training set and the rest 37 images are kept for testing. For dataset OutdoorUrban, the size of training data is 100 and it is 27 of testing data. In dataset MoCap, there are 800 images in the training dataset and 500 images are regarded as testing data.

The regression is assessed by the results in Tables 4 and 5. In the tables, the value of R^2 is around 0.9, which means

Table 4 The statistics of linear regression on HOG

HOG	Size	R^2	F	p	Err.var
RGB-NIR	300	0.9206	35.1212	0	0.0011
Urban	100	0.9072	2.5281	0.0670	0.0055
MoCap	1300	0.8482	101.2218	1.9102e−276	2.4714e−04

Table 5 The statistics of linear regression on LBP

LBP	Size	R^2	F	p	Err.var
RGB-NIR	300	0.9735	8.3986	6.8146e-06	5.5635e+05
Urban	100	1	NaN	NaN	NaN
MoCap	1300	0.8842	49.4688	5.4552e−16	1.0459e+06

Table 6 The test result of linear regression on HOG

HOG	Size	Mean	Var
RGB-NIR	70	0.0356	0.1111
Urban	27	0.2710	0.0656
MoCap	500	0.0243	0.0005

Table 7 The test result of linear regression on LBP

LBP	Size	Mean	Var
RGB-NIR	70	0.0246	0.0010
Urban	27	0.0277	0.0006
MoCap	500	0.0169	0.0002

that the regression function is much closer to the true values, and it understands the information of the data very well. Also most of the p-value in two tables are greater less than 0.05, so the null hypothesis is rejected, namely the linear model is correct for the data. But HOG in dataset OutdoorUrban with the value 0.0670 is an exception. In a word, the two descriptors are both regressed well with the training data, and they draw linear lines perfectly fitting to the points.

For testing, the *mean* and *var* in Tables 6 and 7 refer to the average value and variance of the error between the actual value and the true value. That the mean error of HOG feature is only about 0.03 on RGB-NIR and MoCap indicates an excellent result. In order to compare with HOG feature we normalize the LBP feature. And the low average values of mean provides the possibility of linearity.

7.2 Results of GPR

7.2.1 GPR for HOG and LBP

By using these hyperparameters, the new feature descriptors are predicted. First set with 10 test data, the two criteria, *mean* and *variance* are computed as shown in Table 8. On datasets RGB-NIR and MoCap, the averages of absolute error are both under 0.05. Meanwhile the variances on the two dataset are in a great level as well.

Table 8 The result of GP regression for HOG

HOG	RGB-NIR	OutdoorUrban	MoCap
Mean	0.0342	0.1376	0.0146
Variance	5.8757e−04	0.0070	8.3017e−06

Table 9 The result of GP regression for HOG with 100 training and 10 testing data

HOG	RGB-NIR	OutdoorUrban	MoCap
Mean	0.0310	0.1416	0.0042
Variance	4.0629e−4	0.0036	4.5617e−06

Table 10 The result of GP regression for HOG with 100 training and 50 testing data

HOG	RGB-NIR	OutdoorUrban	MoCap
mean_err	0.0238	0.0280	0.0502
var_err	0.0004	0.0010	0.0024
fron_actual	21.1325	21.2427	20.8977
fron_true	21.4242	21.4240	21.2131

Table 11 The result of GP regression with exact inference method and Gaussian likelihood function

HOG	RGB-NIR	OutdoorUrban	MoCap
mean_err	0.0251	0.0287	0.0500
corr2	0.8909	0.9609	0.8983
frob_actual	21.1551	21.2685	20.8614
frob_true	21.4242	21.4240	21.2131

Further, for the sake of analyzing the effect on training and testing dataset, we enlarge the size of training data to 100 images and the size of testing data is amplified to 50 respectively. Comparison the data in Tables 8, 9 and 10 in vertical direction, the result indicates that the size of neither training data nor testing dataset can effect the performance of Gaussian process heavily. Therefore, Gaussian process is robust and efficient with fewer training data.

Applying exact inference method and Gaussian as likelihood function, the sizes of training data and testing data are set with 100 and 50 respectively. Based on this setting, the process runs much faster than using EP inference method. From Table 11, we can see that the performance of evaluation is excellent, the average value of error is less than 0.03.

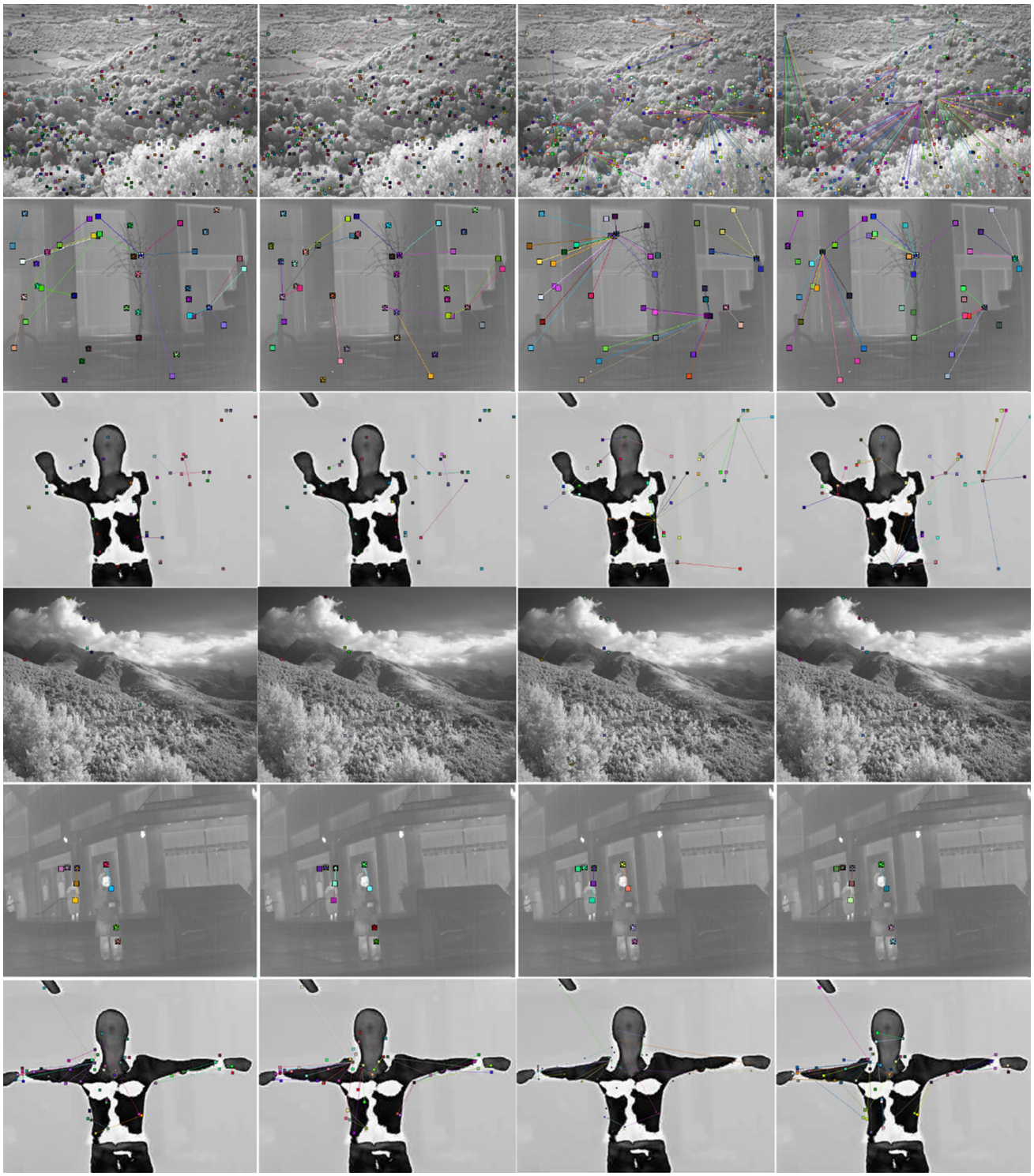


Fig. 8 Example of descriptor matching. The descriptors are regressed by Gaussian process with exact inference method. First two *columns* refer to the results with RQ covariance function with mean value zero and 100, and the last two *columns* show the results with SE covariance function with mean value zero and 100. The *squares* refer to the detected points in visual images and the *stars* refer to the relocated descriptors in infrared images

According to Frobenius norm, the ratio between the actual value and true value is over 99 %, which shows the similarity completely.

Also we consider the role of the initial value of the hyper-parameters of the covariance functions. The values are set to change with step of 0.1. Since the procedure of parameter

optimization is applied, the initial values make no sense to effect the result and performance.

For the purpose of LBP , the same process contents have been executed as HOG. However, the result turns that we can not obtain an answer to the regression for LBP.

7.2.2 GPR for SIFT and SURF

We consider squared exponential function (SE) and rational quadratic function as the covariance function for Gaussian process regression. And also the two inference methods: EP inference method and exact inference method, are applied in this work. In addition the mean value of Gaussian process is set as zero and 100. Thus, based on these three conditions, where each has two values, there are totally 2^3 combinations.

From the results of SIFT, the process with RQ covariance function by exact inference method outperforms than the others with an optimal result, especially in RGB-NIR with a precision over 90 %. And the precisions on other sets are also acceptable with about 50 %. But the SE covariance function is not fitted in this model. In addition, we can find that the value of mean has little effect on the results. On the aspect of SURF, both RQ and SE covariance functions perform well in this work. Based on the value of precision as well as the illustration of the example result images in Fig. 8, the regression results from EP method are totally failed in each situation despite of higher computational cost. In a word, the best performance is the process by exact inference method with RQ covariance function.

7.3 Linear regression versus Gaussian process regression

For descriptor HOG, both linear regression and Gaussian process can predict reasonable mapping models from visual images to infrared images. Comparing the two methods, the results are shown in Table 12 depending on the condition of *error* introduced before. Both approaches perform well with a low error. And it is obvious that Gaussian process performs better than linear regression, where *error* by Gaussian process is extraordinarily small. Another advantage of Gaussian process appears when the training set is small. It means that in practice, the requisite prior knowledge is much less for GP than linear regression. However, for this instance, linear regression also performs well, so it is a good choice as well because the complexity and cost of linear regression is much lower than GP. Notice that actually for the dataset OutdoorUrban, it does not fit into a linear model. But Gaussian process can deal with linear model and also non-linear model problems. We can see that comparing to the error in OutdoorUrban by linear regression, Gaussian process is much better than it in this case.

Table 12 Comparison of linear regression and Gaussian process for HOG

mean_error	Linear regression	Gaussian process
NIR	0.1074	0.0553
OutdoorUrban	0.7019	0.0505
MoCap	0.0243	0.0475

8 Conclusion

In this paper, we have first presented a comprehensive evaluation of detectors as well as descriptors among multimodal images based on three extensive datasets of infrared (or near-infrared) and visual image pairs, which include different kinds of images with various contents that can present a global view of all the possibilities of the multimodal images.

We have extracted the interest points with five different detectors, that is Harris corner detector, DoG, Harris Laplace detector, Hessian detector and Hessian Laplace detector. We used the repeatability score in order to analyse the percentage of points found in a visual image that can be found in the corresponding infrared image. On the other hand, the evaluation was also performed using the accuracy rate for purpose of analysing the percentage of points detected in infrared image that are detected in visual image as well. A simple quantity analysis is taken into accounts. In the view of results, Harris corner detector and DoG outperform than the other three detectors in quantity, repeatability and accuracy rate among all the datasets we have considered. But the difference between Harris corner detector and DoG are not significant except in dataset MoCap. The good performance of accuracy indicates the limit of infrared images that usually visual images can present the scenes more significantly while most infrared images happen to reduce some information than RGB images such as brightness.

In the case of descriptors, i.e., SIFT, SURF, LBP and HOG, we have analysed their performance based on two criteria: precision and recall. The result of SIFT provided a performance as great as SURF. The feature descriptors with high precision make sense in computer vision community, especially to the applications of matching, while these descriptors contain more information. And HOG outperformed than LBP among all the datasets. It means that comparing to HOG, by the feature descriptors LBP we can obtain correct matches, but many others have been missed. And the low recall score of LBP represents the poor ability to find the common information among multi modal images.

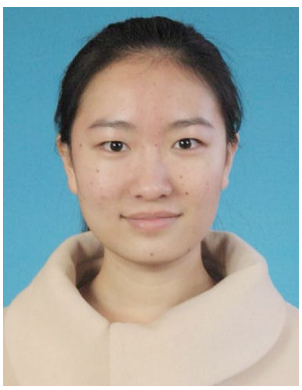
Besides, we have focused on the relationships among descriptors from multimodal image pairs. Between corresponding HOG and LBP, linear relations have been provided by least squares method with good regression qualities. This

indicated the possibility to map a descriptor from visual image to infrared modality by a linear transformation. Furthermore, we have used Gaussian process for regression on HOG and LBP. The optimal regression results have been shown with small error by using squared exponential covariance function. The GPR results of SIFT and SURF have been evaluated by the application of matching. The process of SIFT with rational quadratic function as covariance function has a good performance by evaluating the precision score of matching. The results have presented not only the relationships of SIFT and SURF corresponding descriptors, but also the possibility of obtaining the relationship of descriptors in multi-modal images by means of Gaussian process. In addition, comparing the results of linear regression and GPR of HOG, Gaussian process performs better than linear regression but with a higher computational costs. For the future work, we will perform some regression analysis for other multimodal data, such as visual and depth images.

Acknowledgments The work is partially funded by DFG (German Research Foundation) YA 351/2-1. The authors gratefully acknowledge the support.

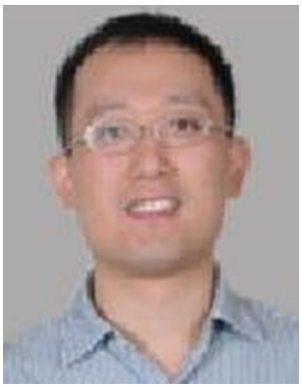
References

1. Bansal, M., Daniilidis, K.: Joint spectral correspondence for disparate image matching. In: CVPR, pp. 2802–2809 (2013)
2. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Computer Vision: ECCV 2006, Springer, Graz, pp 404–417 (2006)
3. Bingham, N.H., Bingham, N., Fry, J.M.: Regression: Linear models in statistics. Springer, New York (2010)
4. Chang, K.I., Bowyer, K.W., Flynn, P.J.: An evaluation of multimodal 2d+ 3d face biometrics. PAMI **27**(4), 619–624 (2005)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp 886–893 (2005)
6. Firmenichy, D., Brown, M., Susstrunk, S.: Multispectral interest points for rgb-nir image registration. In: ICIP, pp 181–184 (2011)
7. Gauglitz, S., Höllerer, T., Turk, M.: Evaluation of interest point detectors and feature descriptors for visual tracking. Int. J. Comput. Vis. **94**(3), 335–360 (2011)
8. Gil, A., Mozos, Ó.M., Ballesta, M., Reinoso, Ó.: A comparative evaluation of interest point detectors and local descriptors for visual slam. Mach. Vis. Appl. **21**(6), 905–920 (2010)
9. Han, J., Pauwels, E.J., de Zeeuw, P.M.: Visible and infrared image registration in man-made environments employing hybrid visual features. Pattern Recognit. Lett. **34**(1), 42–51 (2013)
10. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, Manchester, UK, vol 15, p. 50 (1988)
11. Hrkać, T., Kalafatić, Z., Krapac, J.: Infrared-visual image registration based on corners and hausdorff distance. In: Scandinavian Conference on Image Analysis, pp 383–392 (2007)
12. Kern, J.P., Pattichis, M.S.: Robust multispectral image registration using mutual-information models. IEEE Trans. Geosci. Remote Sens. **45**(5), 1494–1505 (2007)
13. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 79–116 (1998)
14. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV, pp 1150–1157 (1999)
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
16. Lu, C., Jia, J., Tang, C.K.: Range-sample depth feature for action recognition. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE, pp 772–779 (2014)
17. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imaging **16**, 187–198 (1997)
18. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. Int. J. Comput. Vis. **60**(1), 63–86 (2004)
19. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transa. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
20. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. Int. J. Comput. Vis. **65**(1–2), 43–72 (2005)
21. Moreels, P., Perona, P.: Evaluation of features detectors and descriptors based on 3d objects. Int. J. Comput. Vis. **73**(3), 263–284 (2007)
22. Morris, N.J.W., Avidan, S., Matusik, W., Pfister, H.: Statistics of infrared images. In: CVPR, pp 1–7 (2007)
23. Ni, B., Wang, G., Moulin, P.: Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: Consumer Depth Cameras for Computer Vision, Springer, New York, pp 193–208 (2013)
24. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognit. **29**(1), 51–59 (1996)
25. Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
26. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. Int. J. Comput. Vis. **37**(2), 151–172 (2000)
27. Sedai, S., Bennamoun, M. & Huynh, D.: Evaluating shape and appearance descriptors for 3d human pose estimation. In: Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on, IEEE, pp 293–298 (2011)
28. Varshney, P.K.: Multisensor data fusion. Electron. Commun. Eng. J. **9**(6), 245–253 (1997)
29. Wang, L., He, D.C.: Texture classification using texture spectrum. Pattern Recognit. **23**(8), 905–910 (1990)
30. Xia, L. & Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, pp 2834–2841 (2013)
31. Yang, M.Y., Yong, X., Rosenhahn, B.: Feature regression for multimodal image analysis. In: CVPR Workshop on Multi-Sensor Fusion for Outdoor Dynamic Scene Understanding, pp 756–763 (2014)
32. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**(11), 977–1000 (2003)



Xuanzi Yong is currently a PhD student with the Institute for Geodesy, TU Darmstadt, Germany. She graduated with MSc from Leibniz University Hannover. Her research interests are computer vision and remote sensing image classification.

Michael Ying Yang is currently a Senior Researcher at Computer Vision Lab Dresden (CVLD), TU Dresden, Germany. From 2012 to 2015, he was a postdoctoral research associate at the Institute for Information processing (TNT), Leibniz University Hannover. He received his Ph.D. (summa cum laude) from University of Bonn in 2011. His research interests are in the areas of computer vision and photogrammetry, with focuses on probabilistic graphical models, multisensor fusion, and scene understanding.



Yanpeng Cao is a Research Scientist (II) in the department of Visual Computing of the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. He graduated with MSc in Control Engineering (2005) and PhD in Computer Vision (2008), both from the University of Manchester, UK. Before joining A*STAR, he worked in a number of institutes such as Mtech Imaging Ptd Ltd (Singapore), National University of Singapore (Singapore), and National University of Ireland Maynooth (Ireland). His major research interests include infrared imaging, sensor fusion, image processing, and 3D reconstruction.

Bodo Rosenhahn gained his Ph.D. in 2003 at the University of Kiel, Germany. From 2003 to 2005 he was (DFG) Post-Doc at the University of Auckland, New Zealand. From 2005 to 2008 he worked as senior researcher at the Max-Planck Insitute for Computer Science in Saarbrucken, Germany. Since 2008 he is a full professor at the Leibniz University Hannover. His research focus is on markerless motion capture, human model generation and animation, pose estimation, and image segmentation. His works received several awards.