Multi-modal Image Registration Based on Modified-SURF and Consensus Inliers Recovery

Yanjia Chen, Xiuwei Zhang(✉), Fei Li, and Yanning Zhang

Northwestern Polytechnical University, 127 West Youyi Road,
Xi'an 710072, Shaanxi, People's Republic of China
xwzhang@nwpu.edu.cn

Abstract. Multi-modal image registration has been received significant research attention in past decades. In this paper, we proposed a solution for rigid multi-modal image registration, which focus on handling gradient reversal and region reversal problems happened in multimodal images. We also consider the common property of multi-modal images in geometric structure for feature matching. Besides the improvements in features extraction and matching step, we use a correspondences recovery step to obtain more matches, thus improving the robustness and accuracy of registration. Experiments show that the proposed method is effective.

1 Introduction

Image registration is a fundamental task in computer vision. It is an application of feature detection, feature description, feature matching, image transformation and interpolation. Each step is a classic problem and there exist many solutions to it. Recently, multi-sensor technology achieves huge progress benefit from physics researches. Traditional single-modal image registration enlarges the view of visible modality, while multi-modal image registration makes the view much deeper and expose essential characteristic of targets.

Solutions to single-modal image registration have been proposed in literatures, most of them utilize the common properties in intensity and describe local features with gradient information. Thus, the intensity-based registration methods cannot be used in multi-modal image registration, and these gradient feature-based methods cannot handle multi-modal image registration as the intensities and gradient usually show inconsistency in multi-modal images, which is as point A illustrated in Fig. 1. To solve this problem, some modified variation of classic feature descriptors have been proposed. Chen and Tian proposed a Symmetric Scale Invariant Feature Transform (symmetric-SIFT) descriptor [3], which is symmetric to contrast, thus suitable to multi-modal images. Hossian [6] improve symmetric-SIFT in the process of descriptor merging. Dong Zhao proposed a variance of the SURF [2] named Multimodal-SURF (MM-SURF)

Y. Zhao et al. (Eds.): ICIG 2017, Part II, LNCS 10667, pp. 612–622, 2017.
https://doi.org/10.1007/978-3-319-71589-6_54

[12], inherits the advantages of the SURF and is able to generate a large number of keypoints. It is superior to symmetric-SIFT and CS-LBP [5], which is a modified version of the well-known local binary pattern (LBP) [9]. However, the adaptive ability of MM-SURF is obtained by changing the way of dominant orientation assignment, and limiting the gradient direction in $[0, \pi)$. This kind of revise decreases the distinguishability of descriptors. Thus resulting in a consequence of more but wrong matches, which cannot been removed by Random sample consensus (RANSAC) [4].

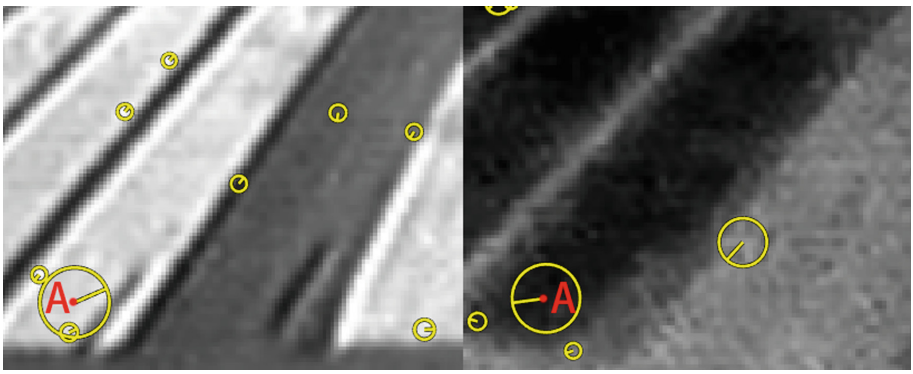

Fig. 1. Gradient reversal in multi-modal images

Another problem of multi-modal image registration is that existing feature-based methods cannot retain adequate accurate correspondences between different modal images. Lack of correspondences or inaccurate correspondences will result in bad transformation and errors. It is usually because of strict matching and outlier remove algorithms. Aguilar [1] proposed a simple and highly robust point-matching method named Graph Transformation Matching (GTM), it finds a consensus nearest-neighbor graph emerging from candidate matches and eliminates dubious matches to obtain the consensus graph. GTM shows superior to RANSAC for high outlier rates. However, it cannot handle some contradictory circumstances, for instance, two falsely matches points have the same neighbors. Then Izadi [7] proposed a weighted graph transformation matching (WGTM) method to overcome the limitations with a more strict matching rules. They are all end with a few matches, and the result is vulnerable even there only one pair of wrong match points. Zhao [13] proposed a dual-graph-based matching method, it generates Delaunay graphs for outlier removal, and recover inliers located in the corresponding graph of Voronoi cells, the inliers recovery make the result to be more robust and stable.

In this paper, we aim to solve the problems above mentioned in multi-modal image registration. First, we propose the modified-SURF (M-SURF) to describe keypoints, and match them refer to the ratio of nearest neighbor and second-closest neighbor. The raw matches set contains many outliers, then we eliminate

them through a graph-based method. The graph-based outlier remove method uses geometry consistency between different modal images, which is believed to be survived in a wide range of geometric and photometric transformation. Second, in order to bring back inliers eliminated former and delete persistent outliers, we create a correspondences recovery step in a reverse way of RANSAC.

The rest of the paper is organized as follows. Section 2 explains the proposed method. Section 3 analyzes the performances of the proposed method in some realworld datasets. Section 4 states conclusions and outlines future work.

2 Our Proposed Method

The overall diagram of the proposed method is shown in Fig. 2. It is obvious that our method includes three step. Firstly is to find a raw matches set utilizes the M-SURF. Secondly, a graph-based matching step is used to remove outliers and retain correct matches as many as possible. Finally, a consensus correspondences recovery step is applied. The results of each step are all matches set.

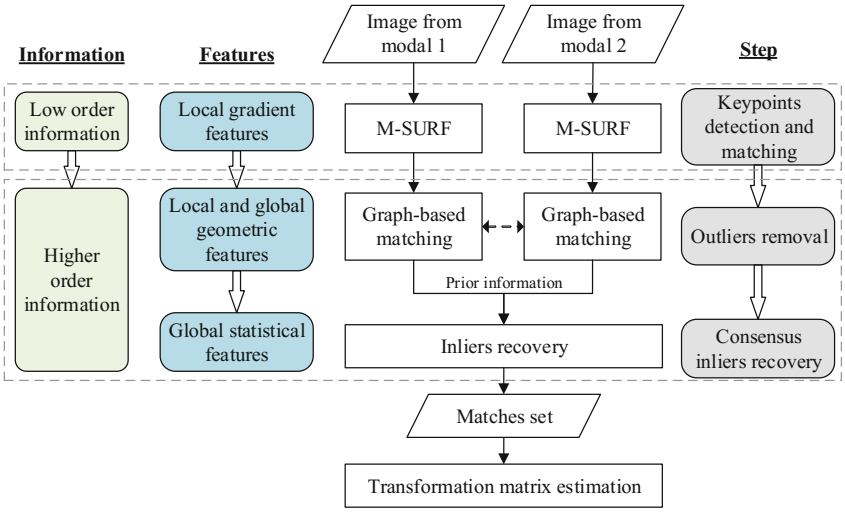

Fig. 2. The overall diagram of the proposed method

2.1 Modified-SURF

Review of SURF: The SURF is much fast than the SIFT and also can ensure the repeatability, distinctiveness and robustness. The SURF is a three stage procedure: (1) keypoints detection; (2) local feature description; (3) keypoints matching. In keypoints detection, the integral image is employed to reduce computation time, Gaussian scale-space and Hessian matrix is employed for keypoints location. In feature description, the dominant orientation of a keypoint is

the orientation of summed haar wavelet responses within a circular neighborhood of radius 6 scale around it. The SURF descriptor for a keypoint is generated in a 20 scale square region centered the keypoint and oriented along its dominant orientation, then the 20 scale square region which is divided into $4 * 4$ subregions, each subregion contains $5 * 5$ sample points. For each subregion, the SURF calculate its haar wavelet responses and weighted with a Gaussian distribution, then obtain a 4 length's vector $(\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$. d_x and d_y are the haar wavelet responses in horizontal direction and vertical direction, $\sum |d_x|$ and $\sum |d_y|$ are their absolute values. Finally, the SURF descriptor is composed of all feature vectors of 16 subregions. After obtain the SURF descriptor, it is usually employ distance ratio between the closest neighbor and second-closest neighbor.

M-SURF: In the SURF, the dominant orientation assignment is based on the horizontal and vertical haar wavelet responses within radius 6 scale around the keypoint. However, haar wavelet responses are related to gradient, which is unstable in multi-modal images. Thus, the SURF cannot obtain desirable results in multi-modal image registration. Inspired by the gradient reversal phenomenon, we modified the dominant orientation assignment in the SURF and limited it in $[0, \pi)$. For the dominant orientation θ calculated in SURF, the modified orientation θ_m defined below.

$$\theta_m = \begin{cases} \theta, & \theta \in [0°, 180°] \\ \theta - 180°, & \theta \in (180°, 360°) \end{cases} \tag{1}$$

Except for the revise in dominant orientation, we then limited the direction of haar wavelet responses to the interval $[0, \pi)$ according to equation below.

$$(d_x, d_y) = sgn(d_y)(dx, dy) \tag{2}$$

where

$$sgn(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \tag{3}$$

The modification of dominant orientation assignment and haar wavelet responses' direction are a kind of relaxation, it handle the problem of gradient reversal in multi-modal images but also decreases the distinctiveness of descriptor for wrong matches. Therefore, we employ a graph-based matching algorithm to remove these outliers.

2.2 Outliers Removal

After applying the M-SURF, we obtain two sets of corresponding keypoints $P = \{p_i\}$ and $P' = \{p_i'\}$ where p_i matches p_i'. Outliers removal is to delete wrong matches in these two sets using certain rules and remain correct matches as accuracy as possible. Recently, graph has been utilized for establishing a higher level geometrical or spatial relationship between feature points. No matter what transformation relationship is between the two images, the spatial relationship between feature points can be maintained.

Many graph-based matching algorithms have been proposed recently. They used adjacency matrix to describe the spatial relationship between feature points and their adjacent feature points. The weighted graph transformation matching (WGTM) algorithm is inspired by GTM algorithm to remove outliers using K-nearest-neighbor (K-NN) graph. It takes the angular distance as a criterion to judge the outliers (false matches).

WGTM starts with creating median K-NN directed graph G for each image, a directed edge $e(i,j)$ exists when p_j is one of the closest neighbors of p_i and also $\|p_i - p_j\| \leq \eta$, and all directed edges formed a edge set E. η is defined by:

$$\eta = \underset{(l,m)\in P\times P}{median} \|p_l - p_m\| \tag{4}$$

A adjacency matrix $\boldsymbol{A}$ is defined by:

$$A(i,j) = \begin{cases} 1 & e(i,j) \in E \\ 0 & otherwise \end{cases} \tag{5}$$

In addition, points without any neighbors are removed as we cannot identify their spatial relationship with other feature points.

Next, a weight matrix $\boldsymbol{W}$ is generated for each point p_i using graph G_p. For another point p_m and their correspondences $p_i^{'}$ and $p_m^{'}$, the weight value is defined by:

$$W(i,m) = \left| \arccos\left(\frac{(p_m - p_i)((p_m^{'} - p_i^{'}) Rot(\theta(k_{min}, i)))}{\|p_m - p_i\| \|p_m^{'} - p_i^{'}\|} \right) \right| \tag{6}$$

where

$$Rot(\theta(k_{min}, i)) = \begin{bmatrix} cos(\theta(k_{min}, i)) & sin(\theta(k_{min}, i)) \\ -sin(\theta(k_{min}, i)) & cos(\theta(k_{min}, i)) \end{bmatrix} \tag{7}$$

Here k_{min} represents the optimal rotation angle between each pair of matches. The optimal rotation angle is defines as the angle that minimizes the sum of angular distances between p_i and $p_m^{'}$. For more information about WGTM, please refer to [7], its performances proved superior to that of GTM and RANSAC. However, there are still problems when applied it to multi-modal image registration.

WGTM uses angular distance as the criterion to find outliers, it is invariant to scale, rotation and sensitive to noise. However, its sensitivity shows more obvious in multi-modal images as the attributes in heterologous modals are quite different, these differences are easy to be identified as noise and removed finally.

2.3 Consensus Inliers Recovery

After outliers removal, the least square method is usually used in literatures to estimate transformation matrix. However, due to the strict rules of graph-based outliers removal and massive noise, there are few correspondences remained after WGTM. It will make the registration result inaccurate if the remained keypoints

are not extracted accurate enough or there still exist one pair of false match points. It is found that some true matches are eliminated in outliers removal because of the strict rule of WGTM. Thus, we focus on how to recover these true matches.

Random sample consensus (RANSAC) is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers, when outliers are to be accorded no influence on the values of estimates. It is usually used to fine correspondences. However, RANSAC is not suitable for multi-modal image registration as there exist too many false matches and it would fail to find a satisfied consensus set. In this case, inspired by RANSAC, we design a consensus inlier recovery method, which use inliers identified by WGTM as prior. Its steps are as follows.

(1) Assume that the correspondences sets are P_i and P_i^*, which are remained after WGTM. We estimate the transformation relationship H_0 between them using the method of least squares.
(2) Use H_0 to check all keypoints with a threshold ε. For a keypoint v_k and its corresponding keypoint v_k^*, the transformed point of v_k is $v_{k2} = H_0 \cdot v_k$, if $\|v_k^* - v_{k2}\| \leq \varepsilon$, then the keypoint is viewed as the consensus inliers, and its corresponding point are also inliers and recovered.
(3) Update the correspondences set P_i and P_i^* with recovered inliers. if there has no point recovered or the sum error reach the top value, stop iteration, otherwise, re-computing the transformation matrix H_0 and back to step (2) and continue the iteration.

3 Experiments

We applied the proposed method to three datasets: (1) The dataset released by Shen [11], which is composed of different exposures images, flash and noflash images, RGB images and Depth images, RGB images and NIR images; (2) The dataset released by Palmero [10], which is composed of RGB images, Depth images and infrared images; (3) Our own dataset, which contains visible/infrared image pairs and visible/hyperspectral (band 66) image pairs. Figure 3 shows some typical examples of datasets. The development environment of experiments is Intel Core i5-4570 CPU @3.20 GHz, 32 GB RAM. The operating system is 64 bit Windows 10. The development platform is Visual Studio 2013 with OpenCV 2.4.9 and Matlab 2016b.

3.1 Evaluation Measures

The accuracy of a registration technique is highly depended on the match sets. The more correct matches, the better registration result. Therefore, we evaluate our results in two ways. One is the final correct matches, another is the target registration error (TRE) [8]. They are defined as follows.

The final correct matches means the number of final correct matches, they are used to estimate the transformation matrix. As long as enough correct

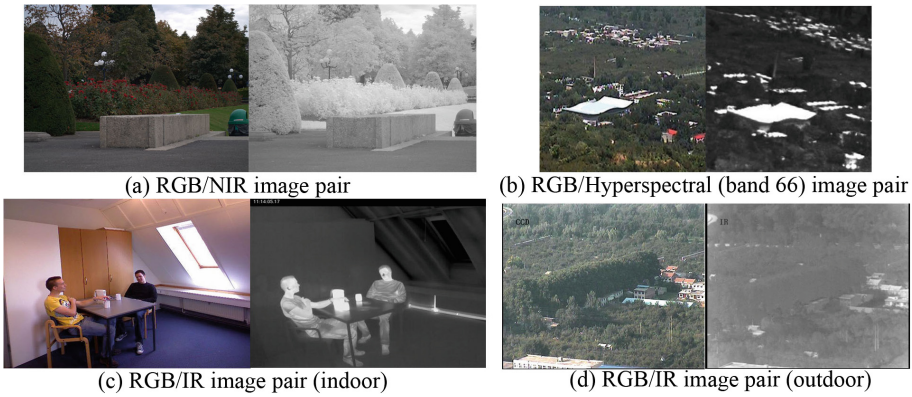

Fig. 3. Example image pairs of datasets

matches are retained, the final correspondences and transformation matrix can be obtained by RANSAC algorithm. The final correct matches is obtained in this way. Due to the transformation matrix is estimated by the method of least square, the more true matches, the little influence of false match and inaccurate feature point extraction, and the better result.

For the TRE, assume that the transformation relationship is $T_1 = \begin{bmatrix} R_1 & t_1 \\ 0 & 1 \end{bmatrix}$ and the ground truth is $T_2 = \begin{bmatrix} R_2 & t_2 \\ 0 & 1 \end{bmatrix}$, where R_1, R_2 are 2×2 rotation matrices and t_1, t_2 are translation vectors. For a point $p = (x, y)^T$ in the reference image, thus

$$p_1 = T_1(p) = R_1 p + t_1 \tag{8}$$

$$p_2 = T_2(p) = R_2 p + t_2 \tag{9}$$

On eliminating p, it follows that,

$$p_2 = R_2 R_1^{-1} p_1 + t_2 - R_2 R^{-1} t_1 \tag{10}$$

The TRE Δp is, thus

$$\Delta p = p_2 - p_1 = (R_2 R_1^{-1} - I) p_1 + t_2 - R_2 R^{-1} t_1 \tag{11}$$

The TRE is a measurement of image registration in a way of reprojection. The value of TRE means the distance between reference image and transformed image in pixel level.

3.2 Matching Comparisons

The matching comparisons is conducted between initial matches identified by M-SURF, matches before recovery and matches after recovery. Figures 4, 5, 6

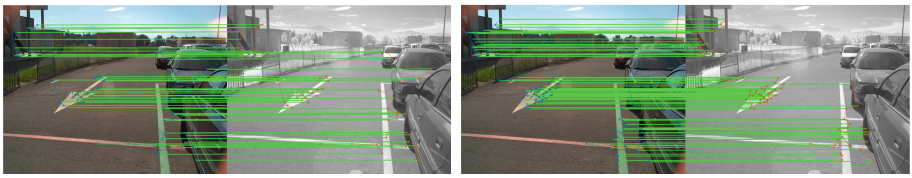

Fig. 4. Matching comparison between RGB/NIR image pair

Fig. 5. Matching comparison between RGB/Hyperspectral (band 66) image pair

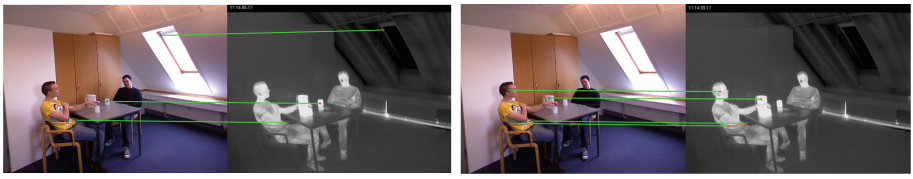

Fig. 6. Matching comparison between RGB/IR image pair (indoor)

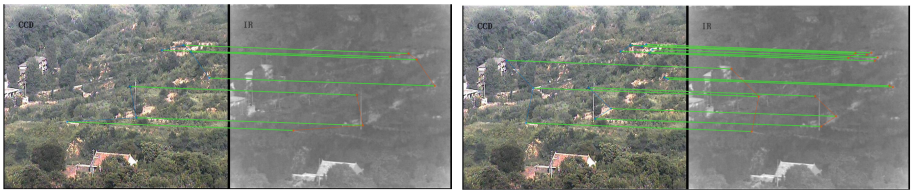

Fig. 7. Matching comparison between RGB/IR image pair (outdoor)

and 7 show the experimental results. The k in WGTM used to create K-NN graph is set to be 5 in our experiments.

From the comparisons, it is obvious that the consensus inliers recovery is worked effectively. In RGB/NIR image pair, although the initial matches obtained by M-SURF and WGTM is enough, we still recovered more matches. Because the NIR image is similar with RGB image in gradient and texture, M-SURF is enough to describe the correspondences. However, in RGB/Hyperspectral (band 66) image pair and RGB/IR image pairs, the initial matches are just exactly enough to estimate the transformation. Any one

of false match or inaccurate feature point extraction can result in a failure registration. For example, there are only three matches in the initial matches of Fig. 6, but the points around the window in the upright of the image are not match. The consensus inliers recovery step not only recover more matches, but also eliminated the false match.

3.3 The TRE Comparisons

The goal of image registration is to align the two images exactly in pixel. Despite comparing the matching results, we evaluate the proposed method with the TRE described before in the final fusion of images. The ground truth is obtained by selecting more than twenty matches per image manually, these points are distributed evenly. To compute the average TRE, we randomly choose 70% pixels of each image as sample points.

We divide the results into two part for considering the TRE results. One is that the input images (set1) are aligned and we cannot distinguish which one is better from the fusion image ($TRE < 5$), Table 1 shows the TRE results of these images. Another one is that the input images (set2) are hard to be aligned or traditional method cannot perform well ($TRE > 5$), Table 2 shows the TRE results of these images.

Table 1. The TRE comparison of set1

Image	1	2	3	4	5	6	7
Before recovery	17.1075	8.5796	49.7602	24.3963	17.521	6.3292	7.7719
After recovery	0.8769	5.1742	2.0491	0.5429	2.0013	1.5387	2.9594
Image	8	9	10	11	12	13	
Before recovery	11.5286	12.6282	17.1823	10.5873	8.12535	12.535	
After recovery	2.241	5.3173	2.5131	2.8637	5.4977	4.7348	

Table 2. The TRE comparison of set2

Image	1	2	3	4	5	6	7
Before recovery	5.6674	2.11	3.6151	4.748	3.2283	5.1124	2.0531
After recovery	2.1828	1.8065	1.9424	2.776	1.9247	2.9118	1.8819
Image	8	9	10	11	12	13	
Before recovery	4.9546	4.4336	1.8301	1.3644	3.4728	4.9187	
After recovery	3.1551	0.8652	1.7125	0.8249	4.4033	1.5867	

From the comparisons of the TRE, we can conclude that the proposed method is effective and robust to multimodal image registration. M-SURF and WGTM

filter most outliers, the inliers recovery find matches with more accurate feature points. Moreover, the consensus inliers recovery step also can eliminate the stubborn outliers that graph-based outliers removal cannot identify. Therefore, From the comparisons of the TRE, for those images (set1) that traditional method cannot align, the proposed method performs well. For those images (set2) that traditional method can align with ordinary results, the proposed method performs better.

4 Conclusions

In this paper, we proposed a novel multimodal image registration method. It is based a modified SURF to extract feature points and create the poor correspondences. By introducing the spatial relationship of matching points, a graph-based outliers removal method (WGTM) is applied then to eliminate false matches. By considering too few inliers were reserved and some stubborn outliers still existed in the residual matches set, the results of the previous two steps are viewed as a prior to recover the consensus inliers. The matching and registration results in the experiments have indicated the effectiveness and robustness of the proposed method. Image registration is a foundation work of image processing, our future work will include incorporating multimodal information to improve the performances in other computer vision tasks.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (Nos. 61231016, 61303123, 61273265), the Natural Science Foundation of Shaanxi Province (No. 2015JQ6256), the Fundamental Research Funds for the Central Universities (No. 3102015JSJ0008), the NPU Foundation for Fundamental Research (No. JCT20130108).

References

1. Aguilar, W., Frauel, Y., Escolano, F., Martinez-Perez, M.E., Espinosa-Romero, A., Lozano, M.A.: A robust graph transformation matching for non-rigid registration. Image Vis. Comput. **27**(7), 897–910 (2009)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features. Comput. Vis. Image Underst. **110**(3), 404–417 (2008)
3. Chen, J., Tian, J.: Real-time multi-modal rigid registration based on a novel symmetric-SIFT descriptor. Prog. Nat. Sci.: Mater. Int. **19**(5), 643–651 (2009)
4. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. ACM (1981)
5. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recogn. **42**(3), 425–436 (2009)
6. Hossain, M.T., Lv, G., Teng, S.W., Lu, G., Lackmann, M.: Improved symmetric-sift for multi-modal image registration. In: International Conference on Digital Image Computing Techniques and Applications, pp. 197–202 (2011)
7. Izadi, M., Saeedi, P.: Robust weighted graph transformation matching for rigid and nonrigid image registration. IEEE Trans. Image Process. **21**(10), 4369–4382 (2012)

8. Maurer, C., Maciunas, R.J., Fitzpatrick, J.M.: Registration of head CT images to physical space using a weighted combination of points and surfaces [image-guided surgery]. IEEE Trans. Med. Imaging **17**(5), 753–761 (1998)
9. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. In: European Conference on Computer Vision, pp. 404–420 (2000)
10. Palmero, C., Claps, A., Bahnsen, C., MØgelmose, A., Moeslund, T.B., Escalera, S.: Multi-modal rgbcdepthcthermal human body segmentation. Int. J. Comput. Vis. **118**(2), 217–239 (2016)
11. Shen, X., Xu, L., Zhang, Q., Jia, J.: Multi-modal and multi-spectral registration for natural images. In: European Conference on Computer Vision, pp. 309–324 (2014)
12. Zhao, D., Yang, Y., Ji, Z., Hu, X.: Rapid multimodality registration based on MM-SURF. Neurocomputing **131**(131), 87–97 (2014)
13. Zhao, M., An, B., Wu, Y., Chen, B., Sun, S.: A robust delaunay triangulation matching for multispectral/multidate remote sensing image registration. IEEE Geosci. Remote Sens. Lett. **12**(4), 711–715 (2014)