Unsupervised Multimodal Image Registration by Polynomial Warping over Correlation-Maximizing Shifts

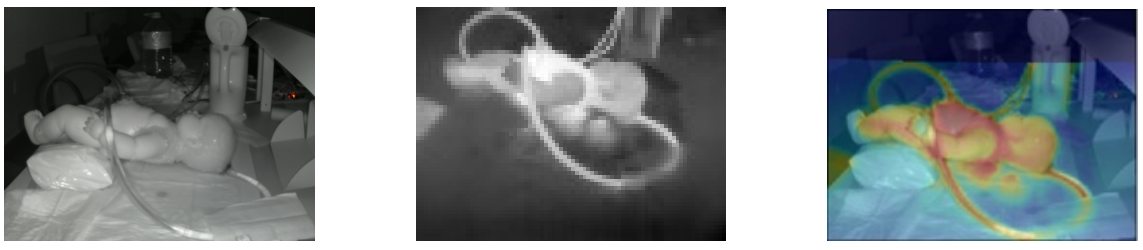
M. Eren Akbiyik
Software Developer, IBM Deutschland GmbH, Schönaicher Str. 220,71032 Böblingen, DE
E-mail: murtaza.akbiyik@ibm.com

Abstract. Alignment of images from multiple modalities is a very important procedure for many medical and industrial applications. Often times it is not possible to utilize a supervised method due to the lack of labeled data for any specific sensor architecture. In this study, a new unsupervised approach is proposed for a sensor-camera system aligned in one axis, that warps the image-like frames onto each other with a second-degree polynomial sampled from the cross-correlation maximizing segment shifts. This methodology will allow the registration process to adjust for focal differences and varying image modalities between the sensors. Thus, novel architectures utilizing seldom-used sensors will more easily adapt to industrial and medical work environments.

1. Introduction

For many multiple-sensor architectures employed in industrial or medical environments, the initial task is to align and fuse the frames from varying modalities. Medical diagnosis often requires the combination of the information from different sources, such in the cases of brain function analysis or radiotherapy planning, that is decided upon the aggregation of CT, MR and/or PET images [1]. For the industrial purposes, use of thermal and RGB cameras have been one of the main applications of image registration research, as complete and accurate fusion is necessary to acquire valid information from these sensors. A registration example for a thermal-RGB monitoring system can be seen at Figure 1.

Image registration has also been well-studied in the area of computer vision as a stand-alone problem. The standard methods for this task [2] can be exemplified by the methods of Scale-Invariant Feature-Transform (SIFT) [3] and ORB [4]. However, as in these methods and their many other counterparts



(a) RGB image (b) Thermal data (c) Overlay

Figure 1. Example of thermal and RGB output from a monitoring system, along with a possible overlay. Created by the proposed method. Notice the slight curve of thermal image in the overlay, sourced from registration by a quadratic transformation.

in the field, the results are not very successful in images from different sources that have varying modalities [5]. Furthermore, the focal differences of the sensors are often overlooked, by applying projective or perspective transformations to register the images that require polynomial warps. In this study, we are proposing a novel method to counteract these issues for a sensor-sensor architecture that satisfies the following assumptions:

- The position of these two sources are known (whether they are vertically or horizontally aligned) and their locations are sufficiently close to ensure focal overlap,
- The edge profiles of the sensors are sufficiently informative, i.e. there exist features that correspond to the edge profiles of the image-like data provided by both sensors.

For a system described above, we first extract the edge features from the image-like data of the sensors using a proposed variation of Canny edge detector [6] whose hysteresis thresholds are optimized by Powell's method [7]. Then, assuming an underlying second-degree polynomial that maps one edge to the other, we sample the shifts of fixed-height image regions iteratively by calculating the FFT-upsampled cross-correlation maximizing vectors [8]. Finally, we estimate the second-degree polynomial by a weighted linear regression with weights proportional to the size of each sampled region. This pipeline will provide a general solution for many variety of sensor architectures, and also allow easier modification of the currently utilized systems.

This study is conducted to solve a registration problem in an RGB-thermal sensor system, therefore the chosen examples will mostly cover such cases. However, the pipeline itself is suitable to any multimodal architecture by nature.

2. Previous Work

Image registration for monomodal and multimodal sources have been studied extensively in the fields of computer vision, medical imaging, and remote sensing. The current array of techniques can be divided into two categories of supervised and unsupervised methods, for which the former frequently employs deep learning based approaches [9] while the latter can be divided into different classes. These classes of algorithms either use pixel values directly by estimating the correlation [10] [11] or mutual information [12], use low-level features like edges and corners to acquire the ideal registration [13], apply fast Fourier transform to work on frequency domain [8], or key-points and invariant descriptors [3] [4].

Despite their successes, deep learning based methods are not applicable to many systems with varying sensors due to the lack of labeled data. Also, a large part of the approaches have inherent problems and/or assumptions that make them not viable for many industrial applications: inability to handle multimodality, registering images with solely affine, similarity-based or projection based transformations that are unable to take the focal perturbations into account, or failing to register the images with insufficient overlaps, are examples for the inadequacies of these techniques. Our purpose in this study is to find a method that is able to address the above-mentioned problems while providing a new approach to the concept of unsupervised edge detection.

3. Edge Extraction by Optimization

Images from different modalities may have varying edge profiles that do not overlap for small variations. In order to reduce such irregularities from the outputs of both sensors, we regard them as noisy values and apply edge-preserving noise reduction methods to remove such perturbations from the images. Next, we extract the edges of both images by posing the process as an optimization problem that results in edge profiles that cover approximately the same areas for both sensors, thus preventing the need of hyper-parameter tuning for different contrast and lighting conditions.

3.1. Total Variation Denoising

The concept of minimizing total variation represents the reduction of a signal's absolute gradient, thus removing the variations due to noise [14]. Parametrization of this optimization problem allows us to

target small perturbations caused by the inherent responses of the sensors, and preserve the edge features that are shared between all components in a system.

Minimization of total variation in an image, as expressed in [15], involves an observed image $g = (g_{i,j})_{1\leq i,j\leq N}$ as the addition of a piece-wise smooth image $u = (u_{i,j})_{1\leq i,j\leq N}$ and a random Gaussian noise. The minimization to recover the original, smooth image can be stated as

$$\min_{u\in X} \frac{||u-g||^2}{2\lambda} + J(u) \tag{1}$$

where $\lambda > 0$ is the weight of the denoising process, and $J(u)$ is the total variation of u defined as

$$J(u) = \sum_{1\leq i,j\leq N} |(\nabla u_{i,j})| \tag{2}$$

with ∇ being the discrete gradient operator introduced also in [15]. Solving this equation with the algorithm proposed by Chambolle in the aforementioned study, we can acquire the smooth sensor images without modality-related irregularities. An exemplary application of this technique can be seen at Figure 2. The parameters of the denoising process for each sensor, λ_1 and λ_2, can be empirically set to match with their comparative responses.

3.2. Canny Edge Detection as an Optimization Problem

After removing the exclusive edge-like features from the sensor outputs, we need the extract the edges from the images. Edge detection algorithm proposed by Canny [6] is implemented in this study with two variations from the original pipeline:

(i) The initial Gaussian smoothing is removed, as total variation denoising handles this step, and
(ii) The hysteresis thresholds are chosen by an optimization process that ensures a predetermined percentage of coverage on the binary output.

The full structure of the original algorithm will not be explained here and is left for the interested reader. The hysteresis thresholding is done after the extraction of edges, and the expression to optimize for the contour i's pointwise gradient set $v_i = \{g_j \mid g \in \mathbb{R}, j \in \mathbb{N}\}$, low and high thresholds T_l, T_h is

$$\min_{(T_l,T_h)\in\mathbb{R}} \left| \left| \bigcup_{i\in L} \{v_i \mid \exists g.g > T_h \wedge \nexists g.g < T_l, g \in v_i\} \right| - \rho N \right| \tag{3}$$

representing the minimization of the absolute difference between cardinality of contours that have at least one element that exceeds T_h and no element that is less than T_l, and the desired coverage of the binary

(a) Original (b) $\lambda = 0.2$ (c) $\lambda = 0.8$

Figure 2. Different levels of smoothing by TV-deniosing. Unlike a Gaussian filter, the output of TV-denoising is "cartoon-like", i.e. the edges are very well-preserved.

image. Here, parameter ρ is the percentage coverage of edges on the binary output, while L and N are the fully-connected edge count and the total pixel count, respectively.

Since Equation 3 has no well-defined gradient with respect to T_l and T_h, we can minimize it by using an approach that does not involve the calculation of derivatives. Powell's method is preferred in this study as it ensures efficient convergence from a bad approximation to a minimum [7]. Initial points for T_l and T_h are recommended to have a ratio between two and three, as advised by Canny [6].

4. Quadratic Approximation to Cross-Correlation Maximizing Samples

The extracted and processed edges can be used to create a transformation that registers the output of one sensor to the other. However, such registrations are generally projective and therefore cannot account for the focal differences between the sensors. To counteract, we will assume an underlying quadratic polynomial $P(x)$, and try to approximate to it by sampling the shifts that maximize the cross-correlation between the edges.

4.1. Shift Sampling by FFT-upsampled Cross-Correlation

Let there be a quadratic polynomial $P(x)$ for a system of two sensors which are aligned on the vertical axis that maps the output of one sensor on top of the other. This polynomial can be sampled using windows of fixed size over the images, by extracting the cross-correlation maximizing affine registration vectors. These vectors register the necessary shift (i.e. required one-dimensional deviation of one region onto the other) y_i to align the regions with horizontal center x_i in the Cartesian coordinates.

The calculation of the cross-correlation peaks between the edges inside the respective windows is done in the frequency domain by applying fast Fourier transforms to both images and first estimating an initial point for the peak location with an up-sampling factor of $\kappa_0 = 2$. Then, using a method referred as two-step DFT, the cross-correlation peak is calculated with a complexity of $\mathcal{O}(MN\kappa^{1/2})$, κ representing the up-sampling factor and MN being the size of the region [8]. A Python implementation of this algorithm from scikit-image library is utilized for the purposes of this study [16].

To determine the ideal sizes and locations of the windows, an iterative logic can be applied: Initial full-sized (i.e. $W \times H$, W and H representing the width and height of the sensor image, respectively) window is capable of sampling the intercept c of the polynomial at point $x_0 = \frac{W}{2}$. The behavior of the function before and after x_0 can be sampled by two non-overlapping windows of size $\frac{W \times H}{2}$, which can effectively construct a quadratic equation in an ideal case. However, as the ideal case is often not possible in different modalities, this iteration can be continued to sample as many shifts as possible to ensure the acquisition of the correct quadratic equation, with windows of size $\frac{W \times H}{2^n}$ at nth iteration, for $n \in \mathbb{N}$.

For visualization purposes, the starting point for each sampled registration vector is chosen to be the center of their respective region, as seen in Figure 3. The quadratic polynomial is estimated from the end points of the vectors.

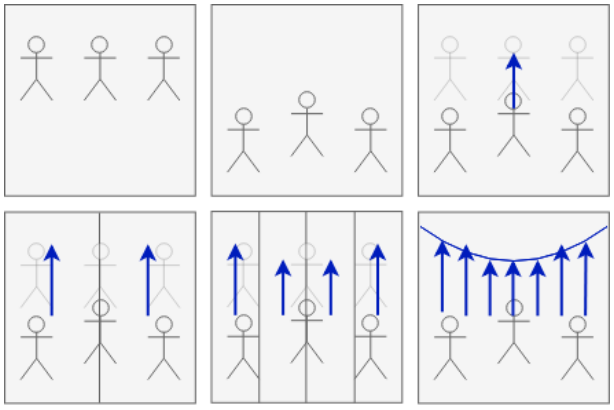


Figure 3. Estimation of the second-degree transformation polynomial. Left to right, top to bottom: Edges of the first sensor, edges of the second sensor, first iteration of the estimation (1 region), second iteration of the estimation (2 regions), third iteration of the estimation (4 regions), and the estimated polynomial.

It is important to note that $P(x)$ need not be quadratic for this algorithm to approximate it over the calculated shifts: If the modalities of the sensors are sufficiently close, higher degree polynomials can be used instead of the current model. However, this is in many cases not necessary, as focal differences tend to cause only quadratic warps on the edges of the images.

4.2. Constructing a Quadratic Polynomial via Weighted Linear Regression

The set of sampled points that represents the shifts

$$S = \{(x_i, y_i) \mid y_i \in [-\tfrac{H}{2}, \tfrac{H}{2}], x_i \in [0..W]\} \tag{4}$$

can be regressed over to estimate the quadratic equation $P(x)$. It must be noted that the amount of information used to calculate each shift is different, the size of the window for each calculation is negatively correlated with the variance of each shift y_i from their true values $P(x_i)$. Therefore, each point can be weighted for the linear regression by the ratio of the size of the windows that they are calculated from to the total frame, thus creating the weight vector V and allowing the analytic solution of the weighted linear regression. This approach also makes the iteration count N a trivial parameter, as the weights of each shift decreases rapidly by each iteration.

For further robustness, the outliers can be filtered according to their distances from the median value, i.e. values of ideal shifts for each region can be ordered and the ones that are above or below the predetermined region can be removed.

This filtering can be conducted by calculating the first quartile Q_1, third quartile Q_3 and interquartile range R of all shifts from the original set S, and then applying standard statistical procedure to remove outliers

$$S' = \{(x_i, y_i) \mid Q_1 - \alpha R < y_i < Q_3 + \alpha R, (x_i, y_i) \in S\} \tag{5}$$

for α denoting the strictness of the filter. Parameter α is taken as 1.5 according to the guidelines of statistical analysis [17].

Using the appropriate regularizations described above, the weighted linear regression for quadratic features can be analytically solved as

$$\hat{\beta} = (X^T V X)^{-1} X^T V y \tag{6}$$

where X,y and $\hat{\beta}$ denote the quadratic features matrix of input points x_i, matrix of polynomial predictions y_i, and OLS-minimizing coefficients matrix for $(x_i, y_i) \in S'$, respectively.

5. Demonstration and Results

The results of this study are investigated in two sections. First, the results of the variations made on Canny edge detector are quantitatively analyzed. Then, the main body of the algorithm is compared with its counterparts in the literature.

5.1. Analysis of the Novel Edge Detector

Starting from the second proposed change, the conversion of hysteresis thresholding into an optimization problem is tested on its convergence capabilities with the Powell's minimizer. The final output of the error defined in Equation 3 is plotted against the exhaustive searches for the parameters of Gaussian smoothing and Total-Variation denoising, proposed in this study. The results are acquired using the OSU Color-Thermal Database [18] which includes 1054 registered thermal and RGB images with sizes (240,320), averaging over a subset ($n = 50$) of shuffled image pairs.

The results have shown that the convergence is highly likely for both methods when the smoothing is applied in a conservative manner: Even with low levels of smoothing, the algorithm is able to converge to the desired level of coverage. The patterns of increase in Figure 4 should also be noted, as the divergence from the optima is resulted from the lack of the necessary amount of edges to attain desired percentage

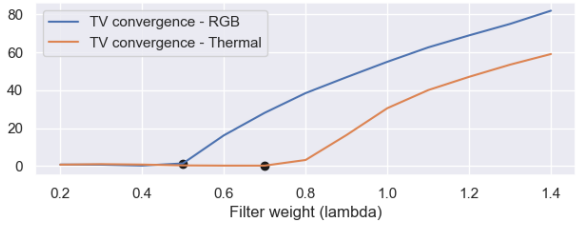


Figure 4. Average difference from desired coverage (i.e. the result of the minimized function in Equation 3) of TV-denoising for filter weight λ [15], for 50 thermal and RGB image pairs. Black dots represent the maximum smoothing that ensures convergence. Percentage coverage parameter $\rho = 3$.

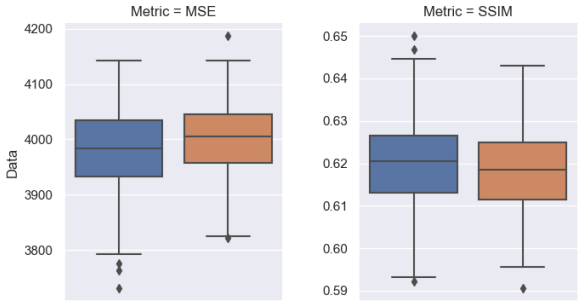


Figure 5. Boxplots of the performances of methods for different metrics. Lower is better for the MSE, higher is better for the SSIM [19] (Structural Similarity).

coverage. Consequently, the Powell's method can be said to ensure the convergence to the minimum on the Equation 3.

The strength of Powell's minimizer provides another advantage to the pipeline, as the smoothing parameters λ_1, λ_2 and coverage parameter ρ appears to be stabilizing each other against a possible choice of an inappropriate value. The designated coverage, as seen in Figure 4, can be attained by a wide set of filter weights, thus alleviating the need of parameter fine-tuning for the pipeline. A value between 2 to 4 appears to be the best choice for ρ in any case, and 3 is used throughout this study.

The choice of the TV-denoising over Gaussian smoothing is investigated by comparing their performances on the aforementioned dataset, OSU Color-Thermal Database. The parameters for the both algorithms are chosen from the Figure 4, at the hand-picked elbow points of the curves which represent the maximum values allowing the convergence of hysteresis optimization. The similarity metrics between the already-registered edges of thermal and RGB images are recorded.

The metrics used in Figure 5 are taken from binary edge images, therefore even the slightest difference on the medians is meaningful; the differences between MSE medians adds $\approx 1\%$ more overlap of the thermal and RGB images' edges on average, for the given coverage parameter. Therefore, the choice of TV-denoising instead of Gaussian smoothing can be justified in the context of multimodal edge extraction.

5.2. Comparative Performances of the Registration Algorithms

Two different datasets are used to test and compare the performance of the proposed algorithm. The first is the OSU Color-Thermal Database, also used in Section 5.1, while the second one is a subset of Visible-Infrared Database [20] with 1000 pair of images from various scenes, rescaled to the same shape with the former. The datasets have been warped to simulate vertical shifts ($P(x) = c$), quadratic distortions ($P(x) = ax^2$) and their combinations ($P(x) = ax^2 + c$), and the resulting images have been tested over the methods at hand. The warps are sampled from uniform distribution, where $a \in [-0.005, 0.005]$ and $c \in [-240, 240]$. 16 homogeneously distributed ground truth points have been used to estimate the root mean squared deviation for each transform.

To compare with the proposed method, ORB, Harris corners and CENSURE [21] feature extraction algorithms are chosen. The final two algorithms are combined with BRIEF descriptor [22] to match the extracted features.

The results displayed at Table 1 indicates the superiority of the proposed method, especially in the existence of quadratic warps. The focal distortions of multimodal images can be accounted for using the given methodology. The quadratic distortions are best handled by CENSURE detector after our approach, as it looks for local similarities that consequently allow better handling of slightly warped features.

Table 1. Average RMSE of 16 ground truth points for each method.

Dataset	Method	Vertical Shift	Quadratic Distortion	Shift and Distortion
OSU Color-Thermal Database	Corr. Max. Shifts	**15.55**	**3.41**	**17.67**
	ORB	38.50	10.94	36.39
	Harris + BRIEF	54.64	6.49	119.20
	CENSURE + BRIEF	37.96	4.72	94.92
Visible-Infrared Database	Corr. Max. Shifts	**16.90**	**6.91**	**24.47**
	ORB	37.59	26.17	927.21
	Harris + BRIEF	362.94	146.60	169.20
	CENSURE + BRIEF	38.71	8.52	35.15

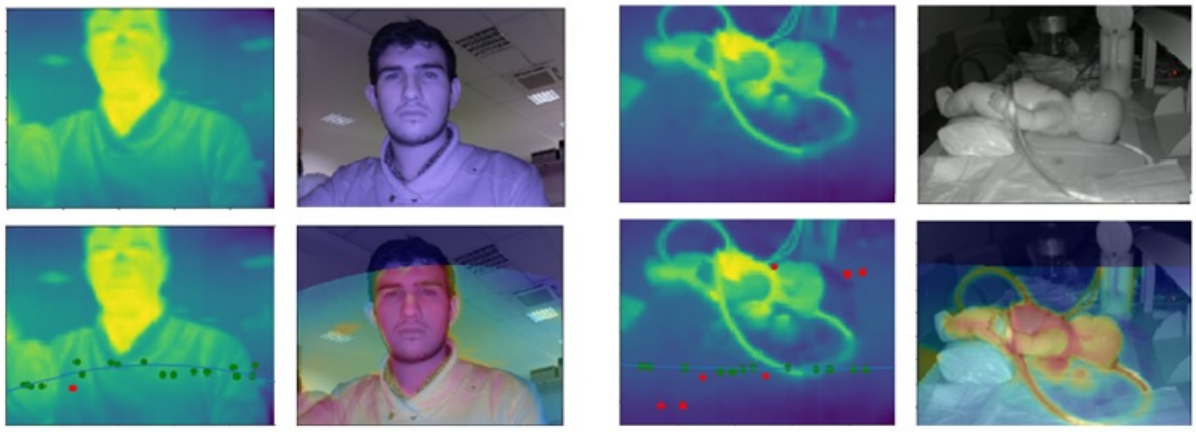
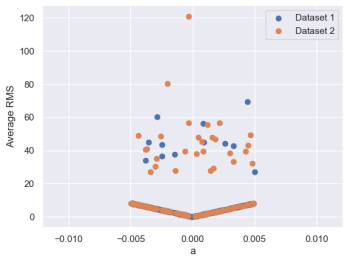


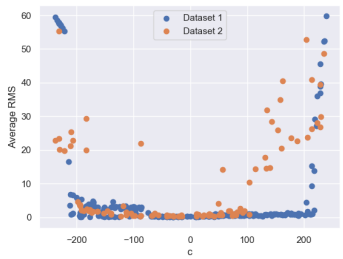
Figure 6. Two exemplary usages of proposed algorithm. Left to right, top to bottom: Thermal frame, RGB frame, thermal frame with polynomial approximation, and overlaid output. Green dots are end points of the valid shift vectors and red dots represent the detected outliers.

Example usages of the proposed algorithm can be seen at the Figure 6. The images are taken from a vertically aligned RGB-thermal monitoring system, featuring second-degree perturbations along the horizontal axis.

To investigate the behavior of the method on different vertical shifts and distortions, the RMS errors with respect to the given transform parameters can be observed in Figure 7. For the images with height 240, the proposed algorithm is able to successfully register the given images shifted up to $\approx 90\%$ of the vertical length.



(a) Average RMSE w.r.t. a (quadratic distortion)



(b) Average RMSE w.r.t. c (vertical shift)

Figure 7. Same datasets are used with the comparative experiments. As long as there are sufficient amounts of overlapped features, the proposed method is able to handle the shift-based registration. Beware that error values below 10 pixel is less than 4%.

6. Conclusion

A method is proposed in this study to register images from different modalities. Although the utilized techniques in this pipeline may appear dated, we believe that the presented approach was not fully developed or tested in the literature despite the convenience of its implementation by common image processing toolkits. Furthermore, as an improvement to previous work, this algorithm is able to take into account the focal distortions of the source sensors, and register the provided images with a second-degree polynomial transformation. The experiments have shown success on different datasets of thermal-RGB architectures, while also showing great promise in terms of registering the images with a very small overlap.

The current methodology, which is designed for stationary sensors with known positions, can be extended to handle non-aligned sensors by iterative application of the algorithm over horizontal and vertical axes. Then, the resulting set of second-degree polynomial transformation matrices can be approximated by one second-or-higher degree transformation matrix to increase registration speed. This will be regarded as a future work for the study.

References

[1] Maes F, Collignon A, Vandermeulen D, Marchal G and Suetens P 1997 *IEEE Transactions on Medical Imaging* **16** 187–198 ISSN 0278-0062
[2] Işık Ş 2014 *International Journal of Applied Mathematics, Electronics and Computers* **3**
[3] Lowe D 2004 *International Journal of Computer Vision* **60** 91–
[4] Rublee E, Rabaud V, Konolige K and Bradski G 2011 *2011 International Conference on Computer Vision* pp 2564–2571 ISSN 2380-7504
[5] Ofir N, Silberstein S, Rozenbaum D, Keller Y and Bar S D 2018 *2018 25th IEEE International Conference on Image Processing (ICIP)* pp 1857–1861 ISSN 2381-8549
[6] Canny J 1986 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-8** 679–698 ISSN 0162-8828
[7] Powell M J D 1964 *The Computer Journal* **7** 155–162 ISSN 0010-4620 (*Preprint* http://oup.prod.sis.lan/comjnl/article-pdf/7/2/155/959784/070155.pdf) URL https://doi.org/10.1093/comjnl/7.2.155
[8] Guizar-Sicairos M, Thurman S and Fienup J 2008 *Optics letters* **33** 156–8
[9] Ma K, Wang J, Singh V, Tamersoy B, Chang Y J, Wimmer A and Chen T 2017 *Medical Image Computing and Computer Assisted Intervention MICCAI 2017* ed Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins D L and Duchesne S (Cham: Springer International Publishing) pp 240–248 ISBN 978-3-319-66182-7
[10] Barnea D I and Silverman H F 1972 *IEEE Transactions on Computers* **C-21** 179–186 ISSN 0018-9340
[11] Pratt W K 1974 *IEEE Transactions on Aerospace and Electronic Systems* **AES-10** 353–358 ISSN 0018-9251
[12] Pluim J P W, Maintz J B A and Viergever M A 2003 *IEEE Transactions on Medical Imaging* **22** 986–1004 ISSN 0278-0062
[13] Lin H, Du P, Zhao W, Zhang L and Sun H 2010 *2010 3rd International Congress on Image and Signal Processing* vol 5 pp 2184–2188
[14] Rudin L I, Osher S and Fatemi E 1992 *Physica D: Nonlinear Phenomena* **60** 259 – 268 ISSN 0167-2789 URL http://www.sciencedirect.com/science/article/pii/016727899290242F
[15] Chambolle A 2004 *Journal of Mathematical Imaging and Vision* **20** 89–97 ISSN 1573-7683 URL https://doi.org/10.1023/B:JMIV.0000011325.36760.1e
[16] van der Walt S, Schönberger J L, Nunez-Iglesias J, Boulogne F, Warner J D, Yager N, Gouillart E, Yu T and the scikit-image contributors 2014 *PeerJ* **2** e453 ISSN 2167-8359 URL https://doi.org/10.7717/peerj.453
[17] Devore J L 2017 *Probability and statistics for engineering and the sciences* (Cengage Learning)
[18] Davis J W and Sharma V 2007 *Computer Vision and Image Understanding* **106** 162 – 182 ISSN 1077-3142 special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum URL http://www.sciencedirect.com/science/article/pii/S1077314206001834
[19] Wang Z, Bovik A C, Sheikh H R and Simoncelli E P 2004 *IEEE Transactions on Image Processing* **13** 600–612 ISSN 1057-7149
[20] Ellmauthaler A, Pagliari C L, da Silva E A B, Gois J N and Neves S R 2019 *Multidimensional Systems and Signal Processing* **30** 119–143 ISSN 1573-0824 URL https://doi.org/10.1007/s11045-017-0548-y
[21] Agrawal M, Konolige K and Blas M R 2008 *Computer Vision – ECCV 2008* (Berlin, Heidelberg: Springer Berlin Heidelberg) pp 102–115 ISBN 978-3-540-88693-8
[22] Calonder M, Lepetit V, Strecha C and Fua P 2010 *Computer Vision – ECCV 2010* (Berlin, Heidelberg: Springer Berlin Heidelberg) pp 778–792 ISBN 978-3-642-15561-1