Uncooperative Spacecraft Pose Estimation Using Monocular Monochromatic Images

Jian-Feng Shi*
MDA, Brampton, Ontario L6S 4 J3, Canada
and
Steve Ulrich†
Carleton University, Ottawa, Ontario K1S 5B6, Canada

https://doi.org/10.2514/1.A34775

Imaging cameras are cost-effective sensors for spacecraft navigation. Image-driven techniques to extract the target spacecraft from its background are efficient and do not require pretraining. In this paper, we introduce several image-driven foreground extraction methods, including combining the difference of Gaussian-based scene detection and graph manifold ranking-based foreground saliency generation. We successfully apply our foreground extraction method on infrared images from the STS-135 flight mission captured by the space shuttle's Triangulation and LIDAR Automated Rendezvous and Docking System (TriDAR) thermal camera. Our saliency approach demonstrates state-of-the-art performance and provides an order of magnitude reduction in processing speed from the traditional methods. Furthermore, we develop a new uncooperative spacecraft pose estimation method by combining our foreground extraction technique with the level-set region-based pose estimation with novel initialization and gradient descent enhancements. Our method is validated using synthetically generated Envisat, Radarsat model, and International Space Station motion sequences. The proposed process is also validated with real rendezvous flight images of the International Space Station.

Nomenclature

$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$	=	pose quaternion matrices
$\tilde{\boldsymbol{A}}$	=	unnormalized optimal affinity matrix
$\mathcal{B}(\boldsymbol{x})$	=	box filter operator
C	=	image curve entity
$\mathcal{CG}(\boldsymbol{x}), \mathcal{CL}(\boldsymbol{x})$	=	image converting functions from color to gray and red green blue to International Commission on Illumination L*a*b, respectively
$\boldsymbol{D}$	=	degree matrix
$\hat{D}, \check{D}$	=	maximum and minimum feature vector differences, respectively
$\bar{\mathcal{D}}(\boldsymbol{x})$	=	normalized difference of Gaussian function
$d^{(ij)}$	=	ith row, jth column element in the degree matrix
E	=	image energy function
$\mathcal{F}(\boldsymbol{x})$	=	fast explicit diffusion function
$\boldsymbol{F}$	=	coordinate frame vectrix
f	=	camera focal length
$\boldsymbol{f}, \boldsymbol{f}^*$	=	superpixel and optimized graph manifold ranking function respectively
$\hat{\boldsymbol{f}}_i$	=	unit vector of the i vector gradient
$\boldsymbol{G}$	=	11×11 Gaussian filter
$\mathcal{G}(\boldsymbol{x})$	=	Gaussian distribution map generation function
$H(\bar{z}), \delta(\bar{z})$	=	Heaviside step and smoothed Dirac-delta function, respectively
h_i	=	step size in the i direction, where i is x, y
$\boldsymbol{h}^{(i)}$	=	feature vector for the ith superpixel
$\boldsymbol{I}, \bar{\boldsymbol{I}}$	=	input color image and normalized image, respectively
$\boldsymbol{I}_{\text{Lab}}$	=	image in CIE L*a*b color space
$\boldsymbol{K}$	=	intrinsic camera matrix
$\boldsymbol{K}_B, K_B$	=	box filter kernel and kernel size, respectively
$\boldsymbol{K}_i, K_i$	=	elliptical kernel and kernel size, respectively, where i is dilation d or erosion e
K_L	=	Laplace filter kernel size
$\boldsymbol{K}_s$	=	3×3 Scharr kernel
$\mathcal{L}(\boldsymbol{x})$	=	Laplacian filter
$\mathcal{M}(\boldsymbol{x})$	=	mean value binary threshold function
$\mathcal{MB}(\boldsymbol{x})$	=	minimum boundary distance saliency response generation function
$\mathcal{MC}(\boldsymbol{x})$	=	image moment centroid generation function
$\mathcal{MD}(\boldsymbol{x}), \mathcal{ME}(\boldsymbol{x})$	=	morphological dilation and erosion function, respectively
M_i	=	appearance model, where i is foreground f or background b
$\mathcal{N}(\boldsymbol{x})$	=	image normalization function
$\mathcal{O}(\boldsymbol{x})$	=	Otsu thresholding function
o_i	=	image center position in the i direction, where i is x, y
P_i	=	probability density function, where i is foreground f or background b
$\boldsymbol{q}, q_i$	=	quaternion represented target orientation and ith parameter quaternion, respectively
$\tilde{\boldsymbol{q}}$	=	nonnormalized quaternion
$\boldsymbol{R}$	=	rotation from target body frame to camera frame
$\mathcal{R}(\boldsymbol{x})$	=	apply contours and keep only maximum perimeter region
$\mathcal{RC}(\boldsymbol{x})$	=	regional contrast saliency response generation function
$r_i(\boldsymbol{x})$	=	likelihood of the pixel property, where i is foreground f or background b
r_{tol}	=	foreground/background classification ratio tolerance
$\boldsymbol{S}_{bg}, \boldsymbol{S}_{fg}$	=	background and foreground masks, respectively
$\boldsymbol{S}_d$	=	simplified difference of Gaussian normalized response map
$\boldsymbol{S}_e, \boldsymbol{S}_{fe}$	=	Scharr filtered edge response map and edge contour response, respectively

Presented as 2018-5281 at the 2018 AIAA SPACE and Astronautics Forum and Exposition, Orlando FL, September 17–19, 2018; received 17 February 2020; revision received 26 June 2020; accepted for publication 23 September 2020; published online 8 February 2021. Copyright © 2021 by Jian-Feng Shi and Steve Ulrich. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. All requests for copying and permission to reprint should be submitted to CCC at www.copyright.com; employ the eISSN 1533-6794 to initiate your request. See also AIAA Rights and Permissions www.aiaa.org/randp.

*Senior GNC Engineer, 9445 Airport Road. Senior Member AIAA.

†Associate Professor, Department of Mechanical and Aerospace Engineering, 1125 Colonel By Drive. Senior Member AIAA.

$\bar{S}_{\text{fed}}$ = mean value fast explicit diffusion response map

$\boldsymbol{S}_G$ = Gaussian centroid response map

$\boldsymbol{S}_{\text{hsf}}$ = high-frequency saliency feature response map

S_i = camera image scaling in image i direction where i is x, y, θ; θ is the skew scaling

$\boldsymbol{S}_i$ = normalized image intersected with i is foreground f or background b mask

$\mathcal{SL}(\boldsymbol{x})$ = simple linear iterative clustering superpixel generation function

$\mathcal{SR}(\boldsymbol{x})$, $\boldsymbol{S}_{SR}$ = spectral residual response function and response map, respectively

$\boldsymbol{S}_L$ = Laplace algorithm foreground response map

$\boldsymbol{S}_m$, $\boldsymbol{S}_M$ = mean value binary threshold contours and response map, respectively

$\boldsymbol{S}_{\text{mask}}$ = processed foreground mask

$\boldsymbol{S}_{\text{mbd}}$ = minimum boundary distance response map

$\boldsymbol{S}_o$, $\boldsymbol{S}_O$ = contour region Otsu threshold and Otsu thresholded response map, respectively

$\boldsymbol{S}_p$ = saliency enhanced response map

$\boldsymbol{S}_{rc}$ = regional contrast response map

$\boldsymbol{S}_u$ = main region contour mask

$\boldsymbol{T}_i$ = transformation matrix for the i axis where i is x, y, z

$\boldsymbol{t}$ = translation vector from camera to target body expressed in camera frame

$\tilde{\boldsymbol{t}}$ = noncentralized initial translation vector

$\boldsymbol{W}$, $w^{(ij)}$ = affinity matrix; and ith row, jth column parameter in the affinity matrix

X, Y, Z = x-, y-, z-axis directions

$\boldsymbol{X}_i$ = target body coordinate where i is the body b or camera c frame

x, y, z = translational x, y, and z or for x, y used as image pixel position

$\boldsymbol{x}$, $\boldsymbol{x}_k$ = input image or pose state vector at the kth step

$\boldsymbol{x}_{\gamma_i}$ = image coordinate partial derivative with respect to γ_i direction

$\bar{x}$, $\bar{y}$ = x, y-pixel location from image center

$\boldsymbol{y}$ = pixel intensity or color feature vector

$\boldsymbol{z}$ = manifold ranking indication vector

α = graph manifold ranking scaling parameter

$\boldsymbol{\epsilon}$, η = $[x \quad y \quad z]^T$ and w quaternion, respectively

μ_i, σ_i = pixel mean and standard deviation, where i is foreground f or background b

Φ = level-set embedding function

Ω_i = image domain where i is foreground f or background b

$\nabla(\boldsymbol{x})$ = gradient function

$\boldsymbol{0}$, $\boldsymbol{1}$ = zero and identity matrices, respectively

I. Introduction

VISION- and laser imaging, detection, and ranging-based spacecraft rendezvous and docking are topics of interest in the spacecraft guidance, navigation, and control (GNC) community [1–4]. Precise visual pose estimation is typically performed using prepared targets with fiducial markers [5–8]. Other research has used image features to determine unprepared spacecraft pose [9–13]; however, when the visual scene is cluttered with the Earth as background, extracting useful image features becomes more difficult [14]. Some have used shadow prediction and regions for simple shapes [15,16]; complex shapes, on the other hand, requires innovative ways to extract the foreground from a cluttered background [17].

Spacecraft pose computed from monochromatic cameras is known to be a key requirement for autonomous GNC systems; however, image patterns from the background scene may complicate the pose analysis. Although recent techniques in deep learning offer precise foreground extraction [18,19], they nevertheless require pretraining and quality training data that are time-consuming to generate. To this end, recent work by Kisantal et al. [20] developed a photorealistic spacecraft pose estimation dataset for machine learning algorithm training and inference. Our approach developed in this paper limits image clutter by extracting spacecraft foreground using saliency detection. Specifically, we use an image-driven approach to detect and extract the spacecraft target in real time and reduce pose estimation processing time and prediction errors.

Object image pose estimation is a core problem in computer vision. Rosenhahn et al. [21] and Lepetit and Fua [22] summarized methods using feature-based structures such as points, lines, circles, and chains to freeform objects such as active contours and implicit surfaces. The state-of-the-art methods in spacecraft pose estimation are based on image features and registration. Both Sharma et al. [12] and Capuano et al. [23] present similar pipelines using region of interest estimation, feature extraction, and point or shape registration. As Capuano et al. [23] points out, "the assumption of a number of pixels in the region of interest must be large enough to allow the extraction of a minimum number of point features." For the extracted features to be useful, the region of interest must be estimated precisely. Unsupervised image processing methods [12] for region estimation can be quickly generated; it may, however, be confused by the intense geographical features of the Earth background. Alternatively, regions of interest can be achieved with good confidence using machine learning or deep learning [24,25] techniques; it is, however, a highly time-consuming task [23]. The difficulty in the latter feature of extraction assumption is often caused by intense space lighting and shadowing.

Key points based on image corners [26,27] and extrema image blobs [28,29] can be difficult to classify due to their quantity and quality under extreme lighting or nondistinctive visual targets. The feature descriptors used for key-point matching can be time-consuming to compute and can be erroneous, especially under affine or viewpoint transformations and illumination distortions. Edge-based pose estimation [30] represents stable features to track with many invariant properties, but extracting relevant edge features and discarding line clutters requires robust decision tools. On the other hand, the region-based pose estimation uses invariant region boundaries with better recognition properties. In this context, we develop a real-time region-based method for monocular infrared (IR) image spacecraft pose estimation.

II. Related Work

We organize foreground object extraction into three branches of computer vision research: techniques in using sequential images for background subtraction [31,32], techniques in machine learning or deep learning neural networks for semantic segmentation [33,34], and image-driven saliency detection [35,36]. Many background subtraction methods require multiple image sequences to train a model before the foreground extraction. Similarly, machine learning methods also demand training a classifier before inference. Particularly in the case of some convolutional neural networks, whereas hundreds of labeled images can be used to perform transfer learning on an established network, millions of labeled images are required to develop weights in a noninitialized network [37]. Saliency methods, on the other hand, can either be top–down [35], bottom–up [36], or a mixture of both [38]. Top–down methods refer to using training images to establish a foreground feature database, and then using a classifier to extract specific image regions; this database can be used to retrieve the desired foreground object from the image. The bottom–up approach is image driven; it is usually faster and easier to implement since it does not require external training as a separate process.

Three recent real-time capable saliency detection methods are regional contrast (RC) [39], minimum barrier distance (MBD) [40], and graph manifold ranking (GMR) [41]. Regional contrast was proposed by Cheng et al. [39] to extract saliency from local regions as the weighted sum of color contrast. Regional contrast can be computed efficiently and can produce precise foreground from color images. Zhang et al. [40] proposed a fast raster-scanning algorithm to approximate the MBD transform [42]. An extended version, MB+, was proposed by adding an *image boundary contrast* map using border pixels as color contrast seeds in the whitened color space. Finally, Yang et al. [41] proposed GMR saliency using document ranking; GMR optimizes a graph-based ranking cost function while replacing documents with image superpixels. The aforementioned saliency detection methods have better precision working with colored images; however, their performance degrades if the image is monochromatic. Grayscale image saliency includes using gradient orientation [43] and texture [44]. The first original contribution of this paper is to propose several saliency models; our most effective saliency model uses an enhanced GMR [41] in combination with MBD [40], RC [39], and difference of Gaussian (DOG)-based background identification. Our method is designed for grayscale images to improve the prior mask used in the level-set regional pose estimation.

Some of the newly developed methods in level-set-based segmentation and pose estimation are as follows: Dambreville et al. [45] projected three-dimensional (3-D) models on to the two-dimensional (2-D) image and used region-based active contours for shape matching; his method showed robustness to occlusions. Prisacariu and Reid used posterior membership probabilities for foreground and background pixel in support of real-time tracking in the Pixel-Wise Posterior 3-Dimensional (PWP3-D) model [46]. Perez-Yus et al. [47] used a combination of depth and color for robustness to occlusions in the level-set segmentation and pose estimation. Hexner and Hagege [48] proposed local templates to enhance the PWP3-D global framework. Tjaden et al. [49] proposed a pixelwise optimization strategy based on a Gaub–Newton-like algorithm using linearized twist for pose parametrization. Tjaden et al. [50] also proposed a temporally consistent, local color histogram as an object descriptor for the level-set pose estimation template. The general trend for level-set pose estimation research is enhancing model compatibility. We take a different approach to improve the input image to achieve better performance.

To summarize, the main contribution of this paper is the development of novel unsupervised scene recognition and foreground extraction using image saliency principles. We provide a spacecraft image dataset complete with ground truth foreground masks for model experimentation and evaluation. We contribute to the noncooperative region-based pose estimation technique by combining the level-set segmentation approach with our scene recognition and foreground extraction method. Furthermore, this paper introduces innovative pose initialization and gradient descent approaches to speed up and stabilize the pose convergence process.

This paper is organized as follows: Sec. III provides original saliency model techniques for foreground extraction and scene classification. Section IV provides the enhanced region-based level-set pose estimation method. Section V provides details on benchmark methods, dataset, metrics, computing platform, and test image details. Section VI provides the results and discussions for the proposed saliency detection and pose estimation models. Finally, Sec. VII concludes this work.

III. Image Processing Method

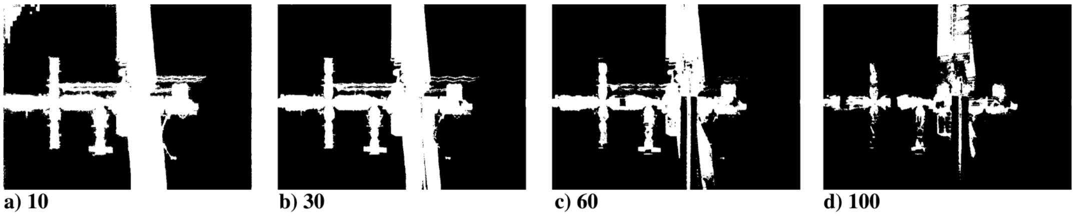
We define two scenarios in the spacecraft rendezvous imagery: the *nadir* (i.e., Earth pointing) and *nonnadir* pointing by the chaser spacecraft camera. In the latter case, the target spacecraft can easily be extracted using thresholding since the background is generally black with some minor polluting light source from stars or camera distortions. Strong light sources from the moon or the sun are localized and rarely occur. The nadir-pointing scenario is much more difficult to resolve, especially for monochromatic images. The Earth background can clutter the input image with clouds, land patterns, and brightly reflected sunlight from the oceans. Operationally, a nadir view during rendezvous operation is unavoidable. For example, there is only one zenithwise corridor for logistical vehicles to approach the International Space Station (ISS); therefore, the view of the target spacecraft from the ISS will always have the Earth backdrop. Another example is in a geostationary servicing mission; the most logical docking face is on the anti-Earth deck, which houses the launch adaptor ring. Therefore, the servicing vehicle must approach the target satellite from the nadir direction facing the Earth.

The computation resource required for both scenarios is vastly different; the nonnadir view can be computed using fast adaptive thresholding, whereas the nadir view requires more software intelligence to extract the foreground target. In the first part of this paper, we developed a reliable unsupervised foreground extraction technique for the nonnadir scenario. At the same time, this technique detects possible Earth background so the image processing can be transfer to a more complex algorithm to separate the foreground vehicle from the Earth. It is possible to use the results from attitude motion tracking in combination with inertial navigation sensors to determine the spacecraft pointing direction. Our approach purely relies on the image sensor to avoid dependencies on spacecraft attitude determination systems. The image-only approach provides more flexibility when integrated into the spacecraft pose estimation pipeline. Furthermore, when the camera system is pointing forward or aft, the proposed foreground extraction methods are useful tools for system designers. Although we have not explored sensor fusion and combined systems, they are also valid ways of solving the same background scene classification problem and can be revisited in the future.

A. Scene Classification

During the nonnadir-pointing scenario, only the backdrop of black space is behind the target vehicle. One may consider the input image already as a form of the foreground saliency map without any additional processing. A closer examination, however, shows there can be various defects in the image. Figure 1 shows increasing threshold level of an ISS IR image, where values under each figure represent the intensity threshold value. Figure 1a shows hardware heating regions can be detected in the upper left corner when the threshold level is 10; this is due to thermal sensor bias that is continuously changing over time. On the other hand, nearly 30% of the vehicle is nonvisible when the threshold is 100. The threshold level of 60 removes the space station remote manipulator system (SSRMS) and the lower solar panel. Out of the four images, a threshold of 30 is the closest match to the ground truth while some of the border sensor noise remains visible.

Our method for region detection and background identification is provided in Algorithm 1 and Fig. 2, where the foreground (FG) and background (BG) histograms are red and blue lines in Fig. 2l, respectively. First, we resize the input image $\boldsymbol{I}$ to 120 rows and normalize the resized image between 0 to 255; $\bar{\boldsymbol{I}} = \mathcal{N}(\boldsymbol{I})$. Using the experimental datasets described in Sec. V, resizing the input image to 120 rows provides optimal precision and speed compared to resizing the input image to 60 and 240 rows. The size reduction of the input image is the main reason for computation speed increase; however, precision will reduce when the reduced image is used by conventional methods. Our formulation ensures the output precision from the reduced input image remains equal to or better than the conventional methods on the original image with a significant increase in computation speed. Next, we apply mean value binary thresholding $\mathcal{M}(\boldsymbol{x})$ on the image to remove low-intensity defects such as image sensor local bias and



a) 10 b) 30 c) 60 d) 100

Fig. 1 ISS IR image threshold level variation increasing from 10 to 100. Pixel detectability is reduced with increasing threshold level.

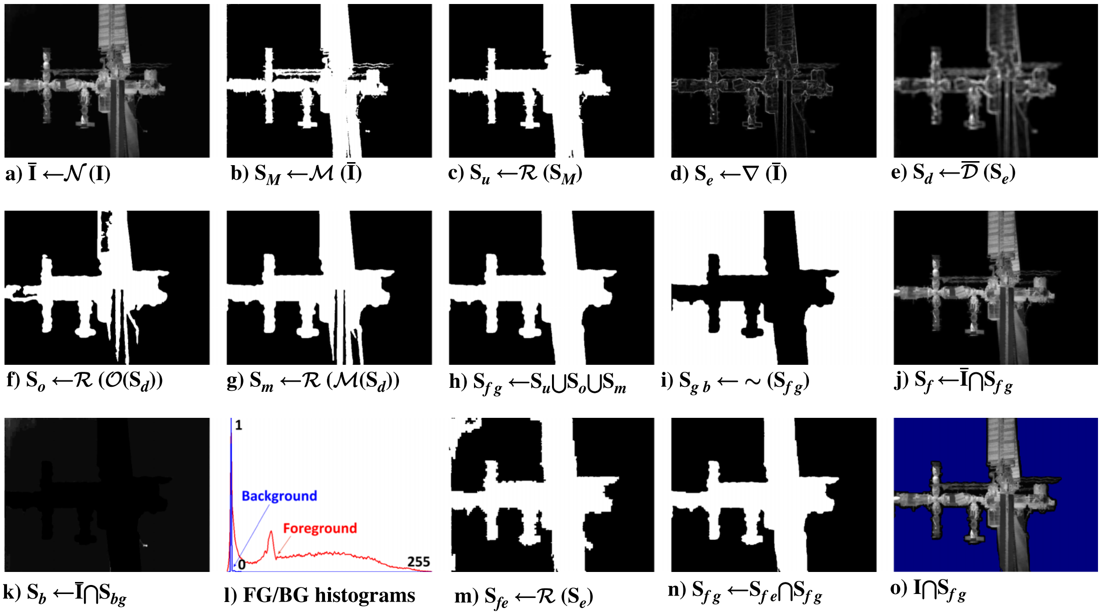
background starlight. If the detector is operating during the nadir phase, the bright Earth background will remain unaffected. A drawback of thresholding is the creation of holes when removing shadowed regions; it will also keep dust particles from image noise. We extract contours from the threshold mask and keep only the maximum perimeter region using the operator $\mathcal{R}(\boldsymbol{x})$. The main region contour will fill in holes and remove all dust particles from it; we find this to be more effective than using an *open* morphological operation. The main region contour mask $\boldsymbol{S}_u$ has fine border resolution, but it is only the result of illumination intensity; it will not exclude the Earth background if the camera is nadir pointing.

To distinguish which direction the camera is pointing to, we use the intensity on either side of the high-frequency response to decide the region class. High-frequency edge features are the result of spacecraft boundaries, internal vehicle connections, Earth horizon, and sharp Earth textures such as coasts and other geological boundaries. During the Earth passage, illuminated regions will occur on both sides of the border feature; whereas if the imager is nonnadir pointing, the most brightly lit region will only occur on the inside of the outer spacecraft border. We use this crucial observation as the decision rule to classify the pointing direction. To extract the high-frequency content, we first approximate the gradient of the normalized input image $[\boldsymbol{S}_e = \nabla(\bar{\boldsymbol{I}})]$ by convoluting $\bar{\boldsymbol{I}}$ with the 3×3 Scharr kernel [17] $\boldsymbol{K}_s$. To extract the local region, we increase the edge responses by taking the DOG. The DOG is a faster numerical approximation of the Laplacian of the Gaussian from the heat equation, and it will diffuse the edge response uniformly in all directions surrounding the high-frequency edges. The normalized response map $\boldsymbol{S}_d$ is a faster and simplified form of the DOG [28] by convoluting the edge response with the difference of two Gaussian kernels of varying scales. We denote the normalized DOG as $\bar{\mathcal{D}}(\boldsymbol{x})$.

We also compare using the linear diffusion heat equation with the nonlinear diffusion equation. We compute the nonlinear diffusion filtering using fast explicit diffusion (FED) [29]. We find similar results in precision using both approaches; the DOG approach takes nearly half the FED computation time using one FED iteration step. Figure 3 shows the DOG and FED response maps and the respective final foreground masks, where $\bar{S}_{\text{fed}}$ is the mean value of $\mathcal{F}(\bar{\boldsymbol{I}}, \boldsymbol{S}_e)$, and the FED results are slightly more accurate by preserving the end effector on the SSRMS. Next, we compute the foreground mask $\boldsymbol{S}_{fg}$ by applying both the Otsu [51] and the mean binary thresholds to $\boldsymbol{S}_d$ and union with $\boldsymbol{S}_u$ to fill large gaps and holes. We can compute the background mask $\boldsymbol{S}_{bg}$ by taking the bitwise *NOT* operator of the foreground map. The foreground and background pixel maps ($\boldsymbol{S}_f$ and $\boldsymbol{S}_b$, respectively) are the intersect of $\bar{\boldsymbol{I}}$ and the respective masks. After computing the mean and standard deviation of the foreground and background intensity histograms, we formulate our decision rule for the pointing phase classification as

$$\left(\frac{\mu_b + \sigma_b}{\mu_f + \sigma_f} < r_{\text{tol}}\right) = \begin{cases} 1 & \text{nonnadir phase} \\ 0 & \text{nadir phase} \end{cases} \tag{1}$$

where μ_i and σ_i, $i \in \{f, b\}$, are the mean and standard deviation of the foreground and background intensity histograms, respectively. During Earth passage, the background will become brighter and the mean and standard deviation for the background histogram will shift higher. Based on this observation, Eq. (1) separates two highly distinctive pointing phase described by the intensity histograms; an example is shown in Fig. 21. We select the ratio tolerance to $r_{\text{tol}} = 0.2$ in our implementation. If the image class is nonnadir pointing, we confirm the foreground mask by intersecting $\boldsymbol{S}_{fg}$ with the main region from the edge response $\boldsymbol{S}_{fe}$. The benefits of the intersection can be observed in Fig. 4, where the image computed with $\boldsymbol{S}_{fe}$ has a more refined foreground map beneath the right solar panel and around the solar wing connections. Figure 4a is the original image; and Figs. 4b and 4c are the foreground mask intersection without and



a) $\bar{\mathbf{I}} \leftarrow \mathcal{N}(\mathbf{I})$ b) $\mathbf{S}_M \leftarrow \mathcal{M}(\bar{\mathbf{I}})$ c) $\mathbf{S}_u \leftarrow \mathcal{R}(\mathbf{S}_M)$ d) $\mathbf{S}_e \leftarrow \nabla(\bar{\mathbf{I}})$ e) $\mathbf{S}_d \leftarrow \bar{\mathcal{D}}(\mathbf{S}_e)$

f) $\mathbf{S}_o \leftarrow \mathcal{R}(\mathcal{O}(\mathbf{S}_d))$ g) $\mathbf{S}_m \leftarrow \mathcal{R}(\mathcal{M}(\mathbf{S}_d))$ h) $\mathbf{S}_{fg} \leftarrow \mathbf{S}_u \bigcup \mathbf{S}_o \bigcup \mathbf{S}_m$ i) $\mathbf{S}_{gb} \leftarrow \sim (\mathbf{S}_{fg})$ j) $\mathbf{S}_f \leftarrow \bar{\mathbf{I}} \bigcap \mathbf{S}_{fg}$

k) $\mathbf{S}_b \leftarrow \bar{\mathbf{I}} \bigcap \mathbf{S}_{bg}$ l) FG/BG histograms m) $\mathbf{S}_{fe} \leftarrow \mathcal{R}(\mathbf{S}_e)$ n) $\mathbf{S}_{fg} \leftarrow \mathbf{S}_{fe} \bigcap \mathbf{S}_{fg}$ o) $\mathbf{I} \bigcap \mathbf{S}_{fg}$

Fig. 2 Pointing phase identification and foreground mask generation; refer to Algorithm 1 for details.

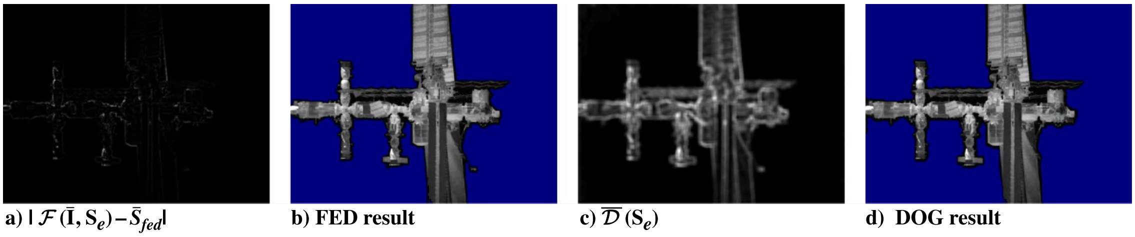
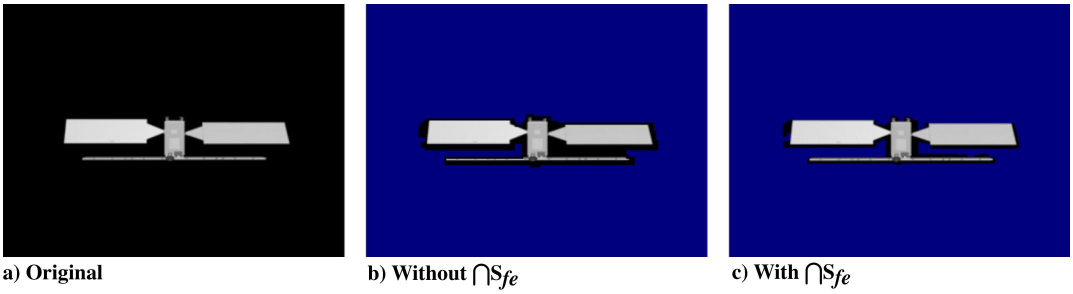

a) $|\mathcal{F}(\bar{\mathbf{I}}, \mathbf{S}_e) - \bar{S}_{fed}|$ b) FED result c) $\bar{\mathcal{D}}(\mathbf{S}_e)$ d) DOG result

Fig. 3 FED vs DOG comparisons: a,c) FED and DOG response maps; and b,d) FED and DOG segmentation results, respectively.

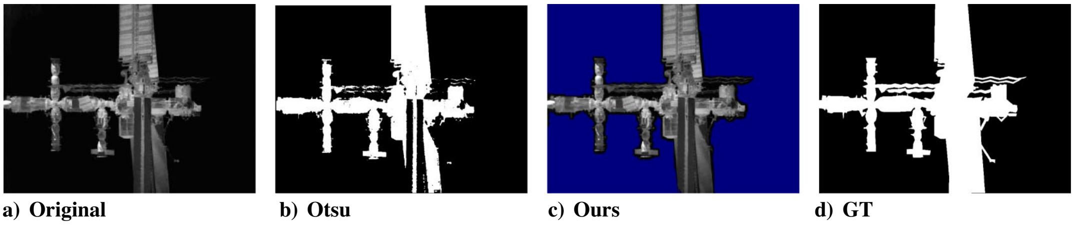


a) Original

b) Without $\bigcap \mathbf{S}_{fe}$

c) With $\bigcap \mathbf{S}_{fe}$

Fig. 4 Radarsat 3-D model test image. Original, without, and with intersection of the main region to the edge response.



a) Original

b) Otsu

c) Ours

d) GT

Fig. 5 ISS IR image comparison of contrast thresholding and region detection, comparing the original image, the Otsu response, our method, and the ground truth.

with $\boldsymbol{S}_{fe}$, respectively. If the image class is nadir pointing, more sophisticated saliency detection is needed to extract the foreground spacecraft; refer to Sec. III.B for details. Finally, we resize the foreground mask to the input image size if required.

An automatic thresholding technique by Otsu [51] selects the optimal separation in the image histogram as the threshold value. Figure 5 provides comparisons between thresholding and region detection, where Fig. 5a is the original image, Fig. 5b is the Otsu thresholding [51], Fig. 5c is our region detection method in Algorithm 1 overlaid on top of the original image (the nonblue region represents the foreground mask), and Fig. 5d is the ground truth (GT) mask. Our method slightly overpredicts the border region but does not exclude any spacecraft regions.

B. Nadir Pointing

The nadir-pointing Earth passage view is more difficult to process than the nonnadir-pointing images. In addition to a cluttered background, part of the challenge comes from the single intensity channel of the IR image. In most cases, the foreground pixels are none distinctive from the background, and therefore harder to classify. We develop two nadir-pointing foreground extraction procedures with the cluttered Earth background; these methods are purely image-driven and do not require preinference training. The first method uses the Laplace operator response to extract the foreground, and the second method is an accelerated manifold ranking technique.

Algorithm 1: The *REGION_DETECT* foreground mask algorithm

```
procedure REGION_DETECT($\boldsymbol{I}$)
    if image larger than 120 rows, then
        Resize input image to 120 rows
        Resized = TRUE
    $\bar{\boldsymbol{I}} \leftarrow \mathcal{N}(\boldsymbol{I})$
    $\boldsymbol{S}_u \leftarrow \mathcal{R}(\mathcal{M}(\bar{\boldsymbol{I}}))$
    $\boldsymbol{S}_e \leftarrow \boldsymbol{K}_s^{(3\times3)} * \bar{\boldsymbol{I}}$
    $\boldsymbol{S}_d \leftarrow \mathcal{N}((\boldsymbol{G}(\sqrt{2^7}) - \boldsymbol{G}(\sqrt{2^{-7}}))^{(11\times11)} * \boldsymbol{S}_e)$
    $\boldsymbol{S}_o \leftarrow \mathcal{R}(\mathcal{O}(\boldsymbol{S}_d))$
    $\boldsymbol{S}_m \leftarrow \mathcal{R}(\mathcal{M}(\boldsymbol{S}_d))$
    $\boldsymbol{S}_{fg} \leftarrow (\boldsymbol{S}_u \bigcup \boldsymbol{S}_o \bigcup \boldsymbol{S}_m)$
    $\boldsymbol{S}_{bg} \leftarrow \sim (\boldsymbol{S}_{fg})$
    $\boldsymbol{S}_f \leftarrow \bar{\boldsymbol{I}} \bigcap \boldsymbol{S}_{fg}$
    $\boldsymbol{S}_b \leftarrow \bar{\boldsymbol{I}} \bigcap \boldsymbol{S}_{bg}$
    Form intensity histograms from $\boldsymbol{S}_f$ and $\boldsymbol{S}_b$.
    Compute $\mu_i$ and $\sigma_i$, $i \in \{f, b\}$, from histograms.
    Compute decision rule per Eq. (1).
    if nonnadir phase, then
        Foreground mask: $\boldsymbol{S}_{\text{mask}} \leftarrow \mathcal{R}(\boldsymbol{S}_e) \bigcap \boldsymbol{S}_{fg}$
    else
        compute fst+ per Algorithm 3
    if resized, then
        Resize foreground mask to the original input image size
```

1. *Laplace Operator Response*

We develop the *LAPLACE* algorithm as an efficient way to extract the high-frequency response by using the Laplacian and morphological operators. Let us define $\mathcal{CG}(\boldsymbol{x})$ as a function that converts a color image into a standard grayscale image, and $\mathcal{B}(\boldsymbol{x})$ as the box filter operator. We also define $\mathcal{MD}(\boldsymbol{x})$ and $\mathcal{ME}(\boldsymbol{x})$ as the morphological dilation and erosion operators using an elliptical kernel respectively. The box filter and morphological operators are used to limit image noise and enlarge the high-frequency response regions. The *LAPLACE* algorithm is given in Algorithm 2, kernel sizes are selected based on sensitivity studies on dataset images from Sec. V.

2. *Background on Graph Manifold Ranking*

Before discussing our accelerated manifold ranking technique, we review background definitions of the manifold ranking or GMR method. The GMR method is a rating algorithm that spreads the seeding query scores to neighboring nodes via a weighted network [52]. GMR has been widely adopted for document [53] and image retrievals [54]. The details of the original GMR algorithm $\overline{\text{GMR}(\boldsymbol{x})}$ are provided by Yang et al. [41]. We will briefly describe key parameters and concepts of the baseline GMR algorithm to support discussions of our enhancements.

Algorithm 2: The *LAPLACE* foreground mask algorithm

```
procedure LAPLACE ($\boldsymbol{I}$)
    $\boldsymbol{S}_L \leftarrow \mathcal{O}(\mathcal{N}(|\mathcal{B}(\mathcal{L}(\mathcal{CG}(\boldsymbol{I}), K_L = 3), K_B = 5)|))$
    for $i = 1 : (K = 3)$, do
        $\boldsymbol{S}_L \leftarrow \mathcal{MD}(\boldsymbol{S}_L, K_d = 10)$
    for $i = 1 : (K = 2)$, do
        $\boldsymbol{S}_L \leftarrow \mathcal{ME}(\boldsymbol{S}_L, K_e = 10)$
    Foreground mask: $\boldsymbol{S}_L \leftarrow \mathcal{R}(\boldsymbol{S}_L)$
```

The input image is converted from *RGB* into the CIE L*a*b color space using $\mathcal{CL}(\boldsymbol{x})$. We compute N number of simple linear iterative clustering (SLIC) [55] superpixels using the function $\mathcal{SL}(\boldsymbol{x})$. Each superpixel has a feature vector $\boldsymbol{h}^{(i)} \in \mathbb{R}^m$. We denote $\boldsymbol{f}$ as the ranking function for the superpixels. Note that $\boldsymbol{z}$ is an indication vector: if $z^{(i)}$ is one, we perform a query on $\boldsymbol{h}^{(i)}$; else, if $z^{(i)}$ is zero, then the query is not performed. We can use an *affinity matrix* of $\boldsymbol{W} = [w^{(ij)}]$, $\boldsymbol{W} \in \mathbb{R}^{N \times N}$, and a *degree matrix* of $\boldsymbol{D} = \mathrm{diag}\{d^{(11)}, \ldots, d^{(NN)}\}$ to rank the queried superpixels. For any given superpixel, its adjacent neighbors and the closed-loop border boundaries are used for the feature-distance differencing. The GMR [41] ranking function, which represents optimal estimation of the foreground, is

$$\boldsymbol{f}^* = \boldsymbol{D}^{1/2}\tilde{\boldsymbol{A}}\boldsymbol{D}^{1/2}\boldsymbol{z} = \boldsymbol{D}^{1/2}(\boldsymbol{D} - \alpha\boldsymbol{W})^{-1}\boldsymbol{D}^{1/2}\boldsymbol{z} \tag{2}$$

where $\tilde{\boldsymbol{A}}$ is the unnormalized optimal affinity matrix (OAM), and α is manually tuned to 0.9. We take the normalized complementary vector for the foreground ranking score to be

$$\overline{\boldsymbol{f}}^* = \boldsymbol{1} - \frac{\boldsymbol{f}^*}{\arg\max\limits_{f^*} f^{*(i)}} \tag{3}$$

where $f^{*(i)}$ is the ranking score at the ith node. This ranking score is used to label individual superpixels in the final saliency map. The weighting is computed as $w^{(ij)} = \exp((\check{D} - D^{(ij)})/\delta(\hat{D} - \check{D}))$, where $i, j \in \overline{\boldsymbol{V}}$, $\overline{\boldsymbol{V}}$ contains all feature vectors from superpixels of the image, and δ is manually tuned to 0.1. We use the L1 norm instead of the L2 norm for speed [56]

$$D^{(ij)} = \sum_{k \in m} \left| h_k^{(i)} - h_k^{(j)} \right|$$

where the maximum and minimum feature vector differences are

$$\hat{D} = \arg\max_{i,j \in \overline{V}} D^{(ij)}$$

and

$$\check{D} = \arg\min_{i,j \in \overline{V}} D^{(ij)}$$

respectively. The background seeds taken from each of the four outer border superpixels are ranked by Eqs. (2) and (3) to form the foreground saliency map and are piecewise multiplied. A mean value binary threshold is applied, and the resulting normalized ranking scores are fed into Eq. (2) again. The ranking scores are transformed back to the saliency image space by $S_f^{(i)} = \bar{f}^{*(i)}$, where i is the node label index $i \in \{1 \ldots N\}$.

3. Accelerated Manifold Ranking

Our accelerated manifold ranking increases the speed of GMR by more than 11 times and improves its precision performance. We use the best-estimate foreground and background seedings rather than border seeding. The final GMR ranking computed by Eq. (2) is a function of the OAM $\tilde{\boldsymbol{A}}$ and the estimated nodes of the foreground $\boldsymbol{w}$ as queries. We replace the border background seeds with nodes under the best-estimate background and foreground maps. Since the target object in spacecraft applications typically contains strong artificial edges compared to the softly blended background, the high-frequency edges are extracted by applying a 3×3 Laplacian filter $\mathcal{L}(\boldsymbol{x})$ on the grayscale input image $\boldsymbol{I}$ and blurred by a square box filter of size K_B to limit noise. We focus the foreground region from eye fixation attention cues computed by the spectral residual (SR) [57] method. The attention cue is combined with the high-frequency region by using a Gaussian distribution map $\mathcal{G}(\boldsymbol{x})$ centered at the SR moment centroid $\mathcal{MC}(\mathcal{SR}(\boldsymbol{I}))$. The high-frequency saliency feature (HISAFE) $\boldsymbol{S}_{\mathrm{hsf}}$ is computed by normalizing the combined SR response and the Laplace filtered response. We also leverage the saliency map generated using a modified color RC [39] $\mathcal{RC}(\boldsymbol{x})$ and MBD [40] $\mathcal{MB}(\boldsymbol{x})$. In addition, for a faster implementation of RC, we replace the region segmentation with the SLIC superpixels. Our experiments suggest the best estimate of the foreground map is the Otsu [51] threshold of the Laplace response map $\boldsymbol{S}_L$ unioned with the mean value binary threshold on the piecewise multiplication of the Gaussian, HISAFE, MBD, and RC response maps as shown in step 14 of Algorithm 3, where lines 2–11, 17–23, and 26 are the *fst* algorithm such that fst_flag $==$ TRUE. We compute both the foreground and background GMR responses and combine them into one final saliency map $\boldsymbol{S}_{fg}$; we then apply a sigmoid function to enhance contrast [40]. The foreground mask can be computed by applying the mean binary threshold operation on $\boldsymbol{S}_{fg}$.

Algorithm 3: The *fst* and *fst*+ saliency algorithm

```
 1: procedure FST_FST_PLUS($\boldsymbol{I}$, fst_flag)
 2:     $\boldsymbol{I}_{\mathrm{Lab}} \leftarrow \mathcal{CL}(\boldsymbol{I})$
 3:     $\boldsymbol{S}_{SR} \leftarrow \mathcal{N}(\mathcal{SR}(\boldsymbol{I}, \sigma = 2))$
 4:     $\boldsymbol{S}_G \leftarrow \mathcal{G}(\mathcal{MC}(\boldsymbol{S}_{SR}))$
 5:     $\boldsymbol{S}_L \leftarrow \mathcal{N}(|\mathcal{B}(|\mathcal{L}(\boldsymbol{I}, K_L = 3)|, K_B = 5)|)$
 6:     $\boldsymbol{S}_{\mathrm{hsf}} \leftarrow \mathcal{N}(\boldsymbol{S}_{SR} + \boldsymbol{S}_L)$
 7:     $\boldsymbol{S}_{\mathrm{hsf}} \leftarrow \boldsymbol{S}_{\mathrm{hsf}} + \mathrm{Mean}(\boldsymbol{S}_{\mathrm{hsf}})$
 8:     $\boldsymbol{S}_{\mathrm{mbd}} \leftarrow \mathcal{N}(\mathcal{MB}(\boldsymbol{I}_{\mathrm{Lab}}))$
 9:     $\boldsymbol{S}_{rc} \leftarrow \mathcal{RC}(\boldsymbol{I}_{\mathrm{Lab}})$
10:     if fst_flag $==$ TRUE, then
11:         $\boldsymbol{S}_{fg} \leftarrow \mathcal{O}(\boldsymbol{S}_L) \bigcup \mathcal{M}(\boldsymbol{S}_G \circ \boldsymbol{S}_{\mathrm{hsf}} \circ \boldsymbol{S}_{\mathrm{mbd}} \circ \boldsymbol{S}_{rc})$
12:     else
13:         $\boldsymbol{S}_{fg} \leftarrow \mathcal{O}(\boldsymbol{S}_L)/255$
14:         $\boldsymbol{S}_{fg} \leftarrow \boldsymbol{S}_{fg} + \mathcal{M}(\boldsymbol{S}_G \circ \boldsymbol{S}_{\mathrm{hsf}} \circ \boldsymbol{S}_{\mathrm{mbd}} \circ \boldsymbol{S}_{rc})/255$
15:         $\boldsymbol{S}_{fg} \leftarrow \boldsymbol{S}_{fg} + \boldsymbol{S}_o/255 + \boldsymbol{S}_m/255$[a]
16:         $\boldsymbol{S}_p \leftarrow (\boldsymbol{S}_{fg} > 1)$
17:     $\boldsymbol{S}_{fg} \leftarrow \mathcal{N}(\overline{\mathcal{GMR}}(\boldsymbol{S}_{fg}) \circ (255 - \overline{\mathcal{GMR}}(255 - \boldsymbol{S}_{fg})))$
18:     for $i = 1{:}\, \boldsymbol{S}_{fg}$height, do
19:         for $j = 1{:}\, \boldsymbol{S}_{fg}$width, do
20:             $\boldsymbol{S}_{fg}(i, j) \leftarrow \frac{1}{1 + e^{-10(\boldsymbol{S}_{fg}(i,j) - 0.5)}}$
21:     $\boldsymbol{S}_{fg} \leftarrow \mathcal{N}(\boldsymbol{S}_{fg})$
22:     if fst_flag $==$ TRUE, then
23:         Foreground saliency map: $\boldsymbol{S}_{fg}$
24:     else
25:         Foreground saliency map: $\boldsymbol{S}_{fg} \leftarrow \boldsymbol{S}_{fg} + \boldsymbol{S}_p$
26:     Foreground mask: $\boldsymbol{S}_{\mathrm{mask}} \leftarrow \mathcal{M}(\boldsymbol{S}_{fg})$
```

[a] Note that $\boldsymbol{S}_o$ and $\boldsymbol{S}_m$ are from Algorithm 1.

We designate the aforementioned procedure as the *fst* algorithm. It demonstrates a net increase in performance and significant speed improvements from the original GMR when evaluated on color and grayscale images. It, however, still lacks the desired precision when applied to the IR images from the STS-135 ISS flight mission. An extended version of the *fst* algorithm is therefore developed to improve performance on spacecraft IR images. We designate this extended version as *fst*+ in Algorithm 3, where lines 1, 12–16, 24, and 25 contain the extended *fst*+ algorithm such that fst_flag $==$ FALSE. The main addition to the extended *fst* model is by reusing the edge response maps computed from Algorithm 1. The DOG threshold responses add more confidence to the *fst* foreground prediction without requiring additional computational time. Finally, the best estimate foreground response is added to the GMR saliency response to adjust the coarse resolution output from the SLIC superpixels.

IV. Pose Estimation

The best-estimate foreground mask described in the previous sections can be used as a prior, or to enhance a known prior input, or to simplify the input image for the region-based pose estimation.

Our region-based pose estimation combines level-set segmentation and 3-D model registration by minimizing an energy function using 3-D projection feedback and foreground–background pixel likelihood estimation. In this section, we describe how the previously computed foreground mask can be used with our modified level-set function for pose estimation purposes.

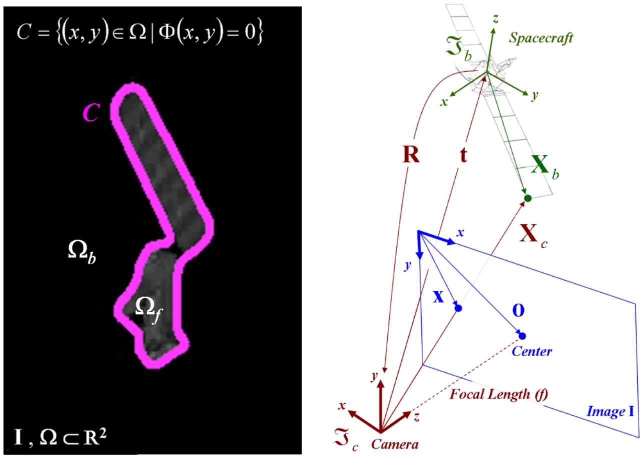
A. Notation Definitions and Rotation Transformations

Given an input image $\boldsymbol{I}$ and the image domain $\Omega \subset \mathbb{R}^2$, the image pixel $\boldsymbol{x}$ with coordinates (x, y) has a corresponding feature $\boldsymbol{y}$. This feature could be the pixel intensity or the color vector (e.g., RGB or CIE Lab). Let us define C as the contour around the object of interest. The foreground region segmented by C is Ω_f, and the background is Ω_b; an example is provided in Fig. 6a, where the contour line of the Envisat image includes definitions for its foreground and background regions. The foreground and background regions have their own statistical appearance model, $P(\boldsymbol{y}|M_i)$ for $i \in \{f, b\}$, where P is the probability density function and Φ is the level-set embedding function. More details of Φ shall be provided in Sec. IV.B. Finally, let $H(z)$ and $\delta(z)$ denote the smoothed Heaviside step function and the smoothed Dirac-delta function, respectively.

A 3-D point $\boldsymbol{X}_c \in \mathbb{R}^3$ with coordinates $(X_c, Y_c, Z_c)^T$ expressed in the camera frame $\boldsymbol{F}_c$ can be a transformation of the object point $\boldsymbol{X}_b \in \mathbb{R}^3$ expressed in the object body frame $\boldsymbol{F}_b$ with coordinates $(X_b, Y_b, Z_b)^T$, using a translation $\boldsymbol{t}$ from $\boldsymbol{F}_c$ to $\boldsymbol{F}_b$ expressed in $\boldsymbol{F}_c$ and a rotation $\boldsymbol{R}$ from $\boldsymbol{F}_b$ to $\boldsymbol{F}_c$ that is parametrized by the quaternion $\boldsymbol{q} = (\boldsymbol{\epsilon}^T, \eta)^T = (q_x, q_y, q_z, q_w)^T$, such that

$$\boldsymbol{R} = (\eta^2 - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon})\mathbf{1} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T - 2\eta\boldsymbol{\epsilon}^{\times} \tag{4}$$

where $\mathbf{1} \in \mathbb{R}^{3\times3}$ is the identity matrix. Note the sign direction of q_w from Eq. (4) in our implementation. The individual coordinates of $(\boldsymbol{t}^T, \boldsymbol{q}^T)^T$ are represented by λ_i, where $i \in \{1 \ldots 7\}$.



a) Region definition b) Coordinate systems

Fig. 6 Spacecraft image contour, level-set function, and coordinate systems.

The camera is precalibrated by the intrinsic matrix

$$\boldsymbol{K} = \begin{bmatrix} fS_x & fS_\theta & o_x \\ 0 & fS_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

where f is the focal length; S_θ is the pixel skew scaling; and S_i and o_i are the image scale and center coordinate for $i \in \{x, y\}$, respectively. The translation vector $\boldsymbol{t}$, the rotation matrix $\boldsymbol{R}$, the projection image $\boldsymbol{I}$, the spacecraft body coordinate $\mathcal{F}_b$, and the camera coordinate system $\mathcal{F}_c$ are shown in Fig. 6b.

B. Level-Set Pose Estimation

The level-set formulation [58] provides a simple mathematical framework for the implicit description of contour evolution. The merging, splitting, appearing, and disappearing of contours can be easily described by a higher-dimensional entity Φ than by the explicit formulation of the curve entity C. The contour can be expressed explicitly as the zeroth level in the level-set function Φ. For example, a contour in a 2-D image is defined by the zero level in a Lipschitz continuous function Φ in a 3-D surface. Formally, $C = \{(x, y) \in \Omega | \Phi(x, y) = 0\}$, and the level-set function Φ is evolved rather than directly evolving C. An illustration of the level-set function evolution based on ISS motion is provided in Fig. 7, where the top row provides synthetic images of ISS motion with 3-D model mesh projection overlay. The bottom row provides the corresponding level-set functions. The zero level and zero crossing are indicated by gradient lines and magenta lines, respectively. The zero boundary is also projected above the zero level as blue lines for clear illustration.

1. *Segmentation Energy and Pixel Likelihood*

The level-set formulation of the piecewise constant Mumford–Shah functional [59–61], will produce the two-phase segmentation of an image $I: \Omega \to \mathbb{R}$ by minimizing an energy function [62]. Bibby

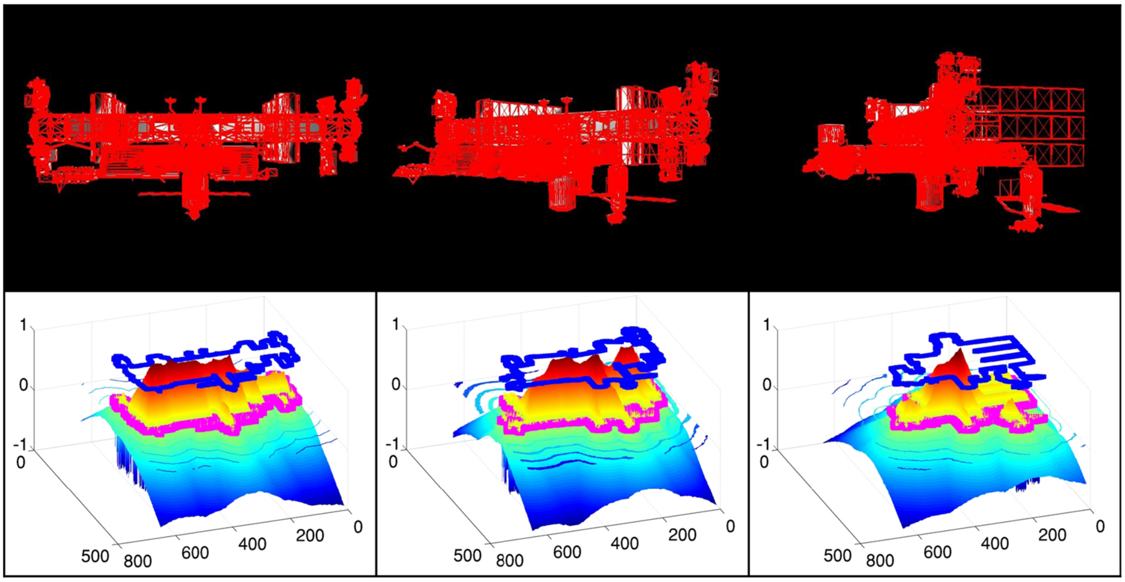


Fig. 7 ISS Pose estimation level-set functions: image registration overlay based on computed pose (top), and level-set function zero crossings in magenta and projected to blue lines for visualization (bottom).

and Reid [63] modified the energy formulation to use the likelihood of the pixel property $r_i(\boldsymbol{x}) = P(\boldsymbol{y}|M_i)$ for $i \in \{f, b\}$. The energy function is

$$E = \int_{\Omega_f} r_f(\boldsymbol{I}(\boldsymbol{x}), C)\,\mathrm{d}\Omega + \int_{\Omega_b} r_b(\boldsymbol{I}(\boldsymbol{x}), C)\,\mathrm{d}\Omega = -\sum_{\boldsymbol{x}\in\Omega} \log(H(\Phi)P_f + (1 - H(\Phi))P_b) \tag{6}$$

where the respective probabilities are as follows, and where $i \in \{f, b\}$:

$$P_i = \frac{P(\boldsymbol{y}|M_i)}{P(\boldsymbol{y}|M_f)\sum_{\boldsymbol{x}\in\Omega} H(\Phi(\boldsymbol{x})) + P(\boldsymbol{y}|M_b)\sum_{\boldsymbol{x}\in\Omega}(1 - H(\Phi(\boldsymbol{x})))} \tag{7}$$

2. *Three-Dimensional Model Projection and Pose Estimation*

The target object pose can be estimated using the energy functional as described in Eq. (6) by taking the partial derivative with respect to the individual pose parameters γ_i [46]; this allows the evolution of the target boundary with respect to its pose rather than time. Let us define $\partial(a)/\partial\gamma_i = a_{\gamma_i}$, $\nabla_t(a) = (a_{t_x}, a_{t_y}, a_{t_z})^T$, and $\nabla_q(a) = (a_{q_x}, a_{q_y}, a_{q_z}, a_{q_w})^T$. The energy partial derivative is

$$E_{\gamma_i} = -\sum_{\boldsymbol{x}\in\Omega} \frac{P_f - P_b}{(P_f - P_b)H(\Phi) + P_b} \delta(\Phi)(\nabla\Phi)^T \boldsymbol{x}_{\gamma_i} \tag{8}$$

where ∇ is the image gradient over $\boldsymbol{x}$. The camera projection model can be used to relate the 3-D model to the 2-D image as follows:

$$\begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix} = Z_c^{-1}\boldsymbol{K}[\boldsymbol{1} \quad \boldsymbol{0}]\begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} X_b = \boldsymbol{K}\frac{\boldsymbol{X}_c}{Z_c} \tag{9}$$

where $\boldsymbol{0} \in \mathbb{R}^3$, $\boldsymbol{K}$ is the intrinsic camera matrix, $\boldsymbol{X}_c/Z_c$ is the depth normalized object point observed and expressed from the camera frame, f is the focal length of the camera, S_θ is the pixel skew scaling, S_i and o_i are the pixel scaling and image origin to center distance, respectively, where $i \in \{x, y\}$. Equation (9) can be used to derive an expression for $\boldsymbol{x}_{\gamma_i}$ such that

$$\boldsymbol{x}_{\gamma_i} = \frac{f}{Z_c^2}\begin{bmatrix} \boldsymbol{X}_c^T(S_x\boldsymbol{T}_x + S_\theta\boldsymbol{T}_y) \\ \boldsymbol{X}_c^T(S_y\boldsymbol{T}_y) \end{bmatrix}(\boldsymbol{X}_c)_{\gamma_i} \tag{10}$$

where $\boldsymbol{T}_i \in \mathbb{R}^{3\times3}$, $i \in \{x, y\}$, has its elements equal to zero, with the exceptions of $\boldsymbol{T}_x(3,1) = \boldsymbol{T}_y(3,2) = 1$ and $\boldsymbol{T}_x(1,3) = \boldsymbol{T}_y(2,3) = -1$. The partial derivative of $\boldsymbol{X}_c$ with respect to the pose parameters γ_i is derived from the extrinsic translation and rotation of the body coordinates to the camera coordinates through $\boldsymbol{X}_c = \boldsymbol{R}\boldsymbol{X}_b + \boldsymbol{t}$, and the partial derivative results are as follows:

$$\nabla_t \boldsymbol{X}_c^T = \boldsymbol{1}, \quad \nabla_q \boldsymbol{X}_c^T = 2[\boldsymbol{A}\boldsymbol{X}_b \quad \boldsymbol{B}\boldsymbol{X}_b \quad \boldsymbol{C}\boldsymbol{X}_b] \tag{11}$$

where $\boldsymbol{1}$ is the identity matrix, and

$$\boldsymbol{A} = \begin{bmatrix} 0 & q_y & q_z \\ -2q_y & q_x & -q_w \\ -2q_z & q_w & q_x \\ 0 & q_z & -q_y \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} q_y & -2q_x & q_w \\ q_x & 0 & q_z \\ -q_w & -2q_x & q_y \\ -q_z & 0 & q_x \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} q_z & -q_w & -2q_x \\ q_w & q_z & -2q_y \\ q_x & q_y & 0 \\ q_y & -q_x & 0 \end{bmatrix} \tag{12}$$

3. *Center Initialization*

Section IV.B.2 provides the connection between the model projection and level-set function using pixel probability. The gradient landscape surrounding the final pose minimum depends on the foreground and background pixel intensity variations. If the initial condition pose is specified in a region with a black space background far from the target object, the gradient descent process can be highly sluggish. In this context, we develop a novel and simple initialization scheme to avoid black space projection and reduce the number of steps that are required to reach the final pose potential minimum. Let us define the unaltered initial condition pose translation as $\tilde{\boldsymbol{t}}$ and our altered approach as $\boldsymbol{t}$. Figure 8 shows the original initial condition and our altered approach by using centralization. The top figure in Fig. 8 shows the raw initial condition pose, captured image, and the actual target in gray. The bottom figure shows the centralized initial condition pose using the saliency mask geometric centroid (blue dot) to set the 3-D model body frame projection. We use the computed saliency mask to generate a geometric center of the region of interest. The pose translation image coordinate is set to the geometric center $(\bar{x}, \bar{y})$, which is computed using image areal moments. While there is no guarantee that the body frame, which is normally located at the center of mass, is the saliency mask geometric center. It is more likely for the geometric center to center of mass approximation to result in a projected region overlapping the region of interest than some arbitrary chosen initial pose. This center shift allows the gradient descent method to initiate in a region where the pose potential is more pronounced than if projecting the initial pose into a black space region with even gradients. The initial pose is computed as

$$\boldsymbol{t} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{1}{fS_x}(Z_c(\bar{x} - o_x) - fS_\theta y) \\ \frac{Z_c}{fS_y}(\bar{y} - o_y) \\ Z_c \end{bmatrix} \tag{13}$$

4. *Gradient Descent*

The PWP3 D [46] gradient descent method increments the pose parameter by manually adjusting the step size. Let us define the pose vector as $\boldsymbol{x} = [\boldsymbol{t}^T \quad \boldsymbol{q}^T]^T$. The baseline PWP3-D gradient descent is

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{h} \circ \boldsymbol{f}(\boldsymbol{x}_k) \tag{14}$$

where $\boldsymbol{h}$ is the step size for the individual pose axis, $\circ$ is the element-wise multiplication Hadamard product operator, $\boldsymbol{f}(\boldsymbol{x})$ is the gradient from Eq. (10), and q_{k+1} is L2 normalized after computing Eq. (14). The baseline gradient descent procedure is unstable and fails to produce the correct pose in several test cases. Specifically, we tested various gradient descent procedures including the Nelder–Mead multidimensional simplex method [64] and the combined Polack–Ribiere and Fletcher–Reeves method [46,65]; however, neither method produced satisfactory timing and accuracy. Finally, we developed an enhanced gradient descent procedure with superior estimation results. The gradient magnitude variation directs the Nelder–Mead simplex [64] state change. The Polack–Ribiere and Fletcher–Reeves [46] step direction includes the current and previous step gradients. Our tests show the current gradient direction outperforms both methods and is computationally more efficient and straightforward to implement. Unlike the baseline PWP3-D, we use an alternative magnitude based on the inverse of the translation distance to modify the step size because closer distance results in a larger image projection, and therefore requires smaller gradient descent movement, and vice versa. Our improved gradient descent formulation is as follows:

$$\boldsymbol{t}_{k+1} = \boldsymbol{t}_k + \frac{h_t}{\|\boldsymbol{t}\|}\hat{\boldsymbol{f}}_t(\boldsymbol{x}_k) \tag{15}$$

where $\hat{\boldsymbol{f}}_t(\boldsymbol{x}_k)$ is the unit direction of the translational gradient $\boldsymbol{f}_t/\|\boldsymbol{f}_t\|$, and h_t is the translational step size. The rotational gradient descent formulation is

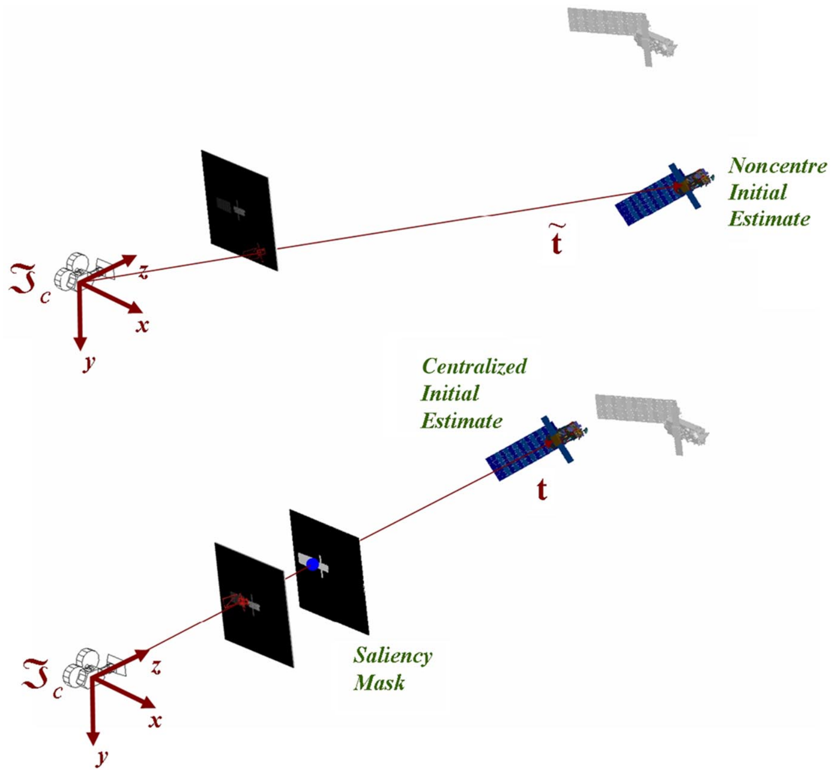


Fig. 8 Center initialization of the target object: initial noncentered pose (top), and applying centered formulation of Eq. (13) for faster gradient descent (bottom).

$$\tilde{\boldsymbol{q}}_{k+1} = \boldsymbol{q}_k + \frac{h_q}{\|\boldsymbol{t}\|}\hat{\boldsymbol{f}}_q(\boldsymbol{x}_k) \tag{16}$$

where $\tilde{\boldsymbol{q}}$ is an nonnormalized quaternion, and $\hat{\boldsymbol{f}}_t(\boldsymbol{x}_k)$ is the unit vector of the quaternion gradient. The final state vector is $\boldsymbol{x}_{k+1} = [\boldsymbol{t}_{k+1}^T \quad \tilde{\boldsymbol{q}}_{k+1}^T/\|\tilde{\boldsymbol{q}}_{k+1}\|\,]^T$. Figure 9 provides the difference of using the basic gradient descent formulation and our enhanced method, where the first and second rows provide the Envisat pose overlay for frames 0, 150, and 300; the third and fourth rows provide Radarsat pose overlay for frames 200, 250, and 300; the first and third rows use the PWP3-D gradient descent implementation [46]; and the second and fourth rows use the enhanced gradient descent method. The original version destabilized when the projection silhouette transitioned into the minimum region, whereas the enhanced method remained stable throughout the entire estimation process.

V. Experimental Images and Datasets

In this section, we benchmark our saliency method against 14 traditional and recent saliency detection methods shown in Table 1 [39–41,51,57,66–73]. These methods were selected based on their real-time and near-real-time running time performances. There are many popular image datasets available for saliency and the segmentation benchmark [39,41,74]; however, there is no image dataset developed explicitly for spacecraft navigation applications. To facilitate our experiment and testing, we developed the satellite segmentation (SATSEG) dataset composed of 100 color and grayscale spacecraft images captured by photographic and thermal cameras that are representative of the various mission scenarios ranging from flight to laboratory tests. Ground truth data were produced manually and are freely available to download from our project website.[‡]

We use the receiver operating characteristics (ROCs) [75] to evaluate our saliency method. For the computed saliency map, it is converted to an 8 bit binary mask image by applying a constant threshold. The precision and recall of a single image can be computed from the ground truth by performing piecewise operation to the saliency maps [69]. A threshold ranging from 0 to 255 is used to control recall. We compute the standard units of measure for saliency evaluations, including the average and maximum F measures [70]. The area under the curve (AUC) is from the ROC true positive rate versus false positive rate plot [69]. We used two computer platforms in the saliency evaluation: a Linux platform (AMD with eight cores, 4.0 GHz, and 32 GB of RAM) and a Windows platform (AMD with four cores, 1.5 GHz, and 6 GB of RAM). Two platforms and operating systems were used because some of the benchmark methods are provided in Windows compiled executables.

We use a combination of synthetic CAD images and space flight IR images for evaluation. The CAD images include 3-D models of the Envisat, the Radarsat, and the ISS model. The 3-D models are also the pose estimation software internal models. We use 3-D Studio Max® to generate synthetic videos from the 3-D models, which include some lighting and shadowing effects. We also use flight images from the STS-135 ISS undocking phases recorded by a 813 × 604 MacDonald, Dettwiler and Associates Triangulation and LIDAR Automated Rendezvous and Docking System (*TriDAR*) thermal camera. For faster processing, the IR image resolution was reduced to 320 × 240 with estimated camera calibration properties of $fS_x = 752.517$, $fS_y = 752.517$, and $fS_\theta = 0$.

VI. Results and Discussions

The contribution of this work includes the creation of a novel saliency detection method to extract the target spacecraft from its background and uses the spacecraft image to calculate its pose estimation. This section will discuss the saliency detection and pose estimation results separately.

A. Saliency Performance

Figure 10 provides the saliency map comparisons of selected images from the SatSeg Dataset. The images from the left to the right columns are: the input image, Otsu thresholding (OTSU-75) [51]; spectral residual (SR-07) [57]; graph manifold ranking (GMR-13) [41]; minimum barrier distance (MB + −15) [40]; regional contrast (RC-15) [39], our *fst*+ method from Algorithm 3 (in green box), and

[‡] Data available online at http://ai-automata.ca/research/hisafe2.html [retrieved 29 January 2021].

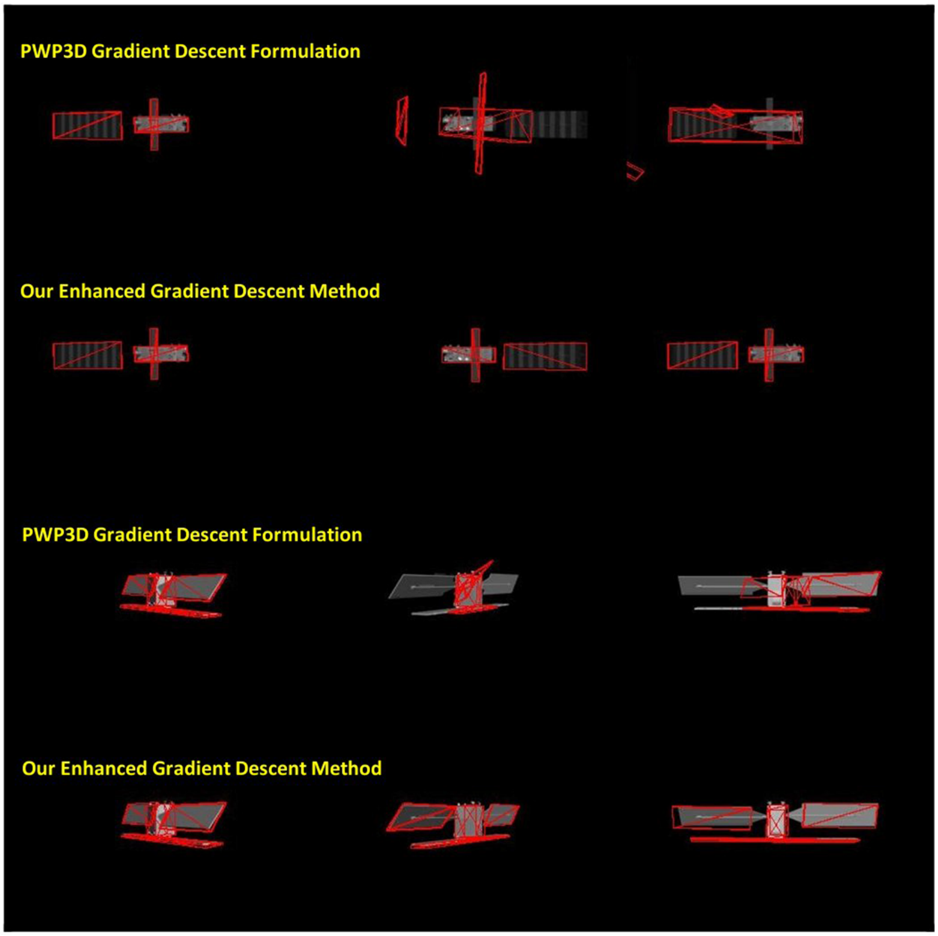

PWP3D Gradient Descent Formulation
Our Enhanced Gradient Descent Method
PWP3D Gradient Descent Formulation
Our Enhanced Gradient Descent Method

Fig. 9 Enhanced gradient descent results for Envisat and Radarsat synthetic image pose estimation.

the ground truth. The Otsu thresholding [51] is not typically considered as a saliency detection model. However, it is nevertheless included in the comparison to demonstrate that although it can generate a precise foreground map in the nonnadir-pointing phase, it has the worst error in the nadir-pointing phase when there is a cluttered background. The SR [57] model provides useful attention location cues but does not produce sufficient detail of the target object. GMR [41], RC [39], and MB+ [40] are state-of-the-art saliency methods that have good performance. However, both RC and GMR runtimes are in the half-second range, and MB+ requires additional segmentation to precisely extract the foreground mask. Figure 10 shows our method having the least background error while maintaining the acceptable resolution of the foreground region.

Quantitative performance plots are provided in Fig. 11, where Figs. 11a and 11b provide precision versus recall, the mean F measure, the maximum F measure, and the AUC for 14 traditional and state-of-the-art saliency detection methods, plus our *fst*+ method. In Fig. 11, the compared methods are (Method-Year): Otsu thresholding (OTSU-75) [51], watershed (WS-92) [76], graph cut (GCUT-04) [77], local contrast raster scan (LC-06) [39,66], spectral residual (SR-07) [57], phase spectrum quaternion Fourier transform (QFT-08) [67], local contrast raster scan (AC-08) [68,69], frequency tuned (FT-09) [70], maximum symmetric surround (MSS-10) [71], histogram-based contrast (HC-11) [72], graph manifold ranking (GMR-13) [41], global cues (GC-13) [73], geodesic (GD-15) [40], minimum barrier distance (MBD-15) [40], MBD extended (MB+-15) [40], regional contrast (RC-15), and ours. Figure 11c provides the timing analysis for single image average run time of the SATSEG dataset; cross and diamond marker runs are executed using the C++/Linux platform; circle marker runs are executed using built executables on the Windows platform; and the red line represents the design timing requirement for the saliency algorithm. Our *fst*+ model shows the highest precision versus recall and maximum F-measure performance compared to all the methods. Our mean F measure is second to the RC, and our AUC is comparable to the RC and GMR. Speedwise, our *fst*+ model average computation time for the SATSEG dataset is 44.223 ± 0.309 ms; this is 11 times faster than the original GMR, 10 times faster than RC, and 2.2 times faster than MB+. Faster methods such as Minimum Barrier Distance (MBD), Geodesic (GD), Frequency Tuned (FT), Phase Spectrum Quaternion Fourier Transform (QFT), Spectral Residual (SR), and Colour Contrast (LC) have much lower precision than our *fst* + method. We also compare with nonsaliency detection methods such as watershed (WS) [76], graph cut (GCUT) [77], and Otsu thresholding (OTSU) [51]. In summary, Figs. 10 and 11 show our *fst*+ model to have the best performance overall.

Table 1 Benchmark methods

Code	Description	Reference(s)
OTSU-75	OTSU thresholding	[51]
LC-06	Color contrast	[39,66]
SR-07	Spectral residual	[57]
QFT-08	Phase spectrum quaternion Fourier transform	[67]
AC-08	Local contrast raster scan	[68,69]
FT-09	Frequency tuned	[70]
MSS-10	Maximum symmetric surround	[71]
HC-11	Histogram-based contrast	[72]
GMR-13	Graph manifold ranking	[41]
GC-13	Global cues	[73]
GD-15	Geodesic	[40]
MBD-15	Minimum barrier distance	[40]
MB+-15	MBD extended	[40]
RC-15	Regional contrast	[39]

In addition, we applied our saliency algorithms to the TriDAR thermal camera video of the STS-135 mission where the space shuttle performed an undock and flyby of the ISS. Figure 12 shows the input video and the results of the various methods, where from left to right, the first column is the original ISS IR video; the second column is the

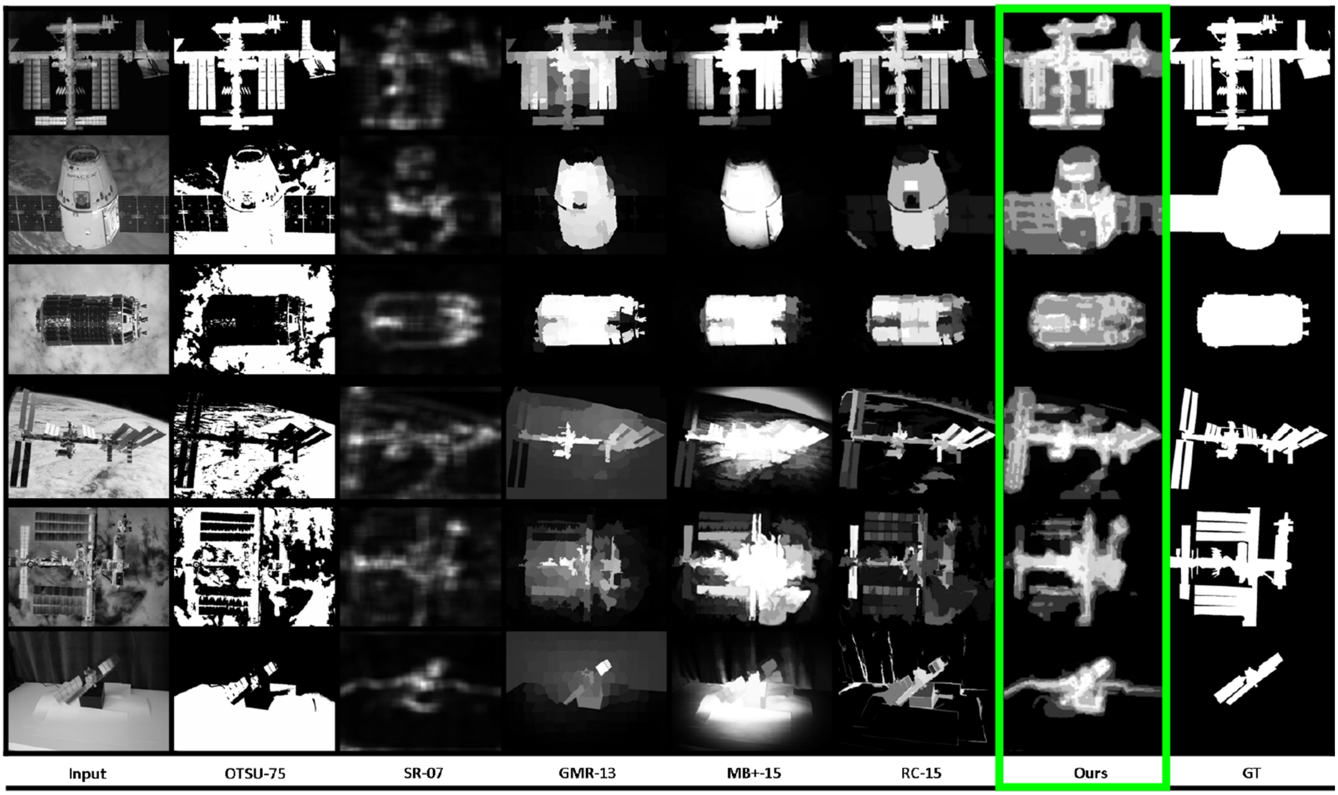
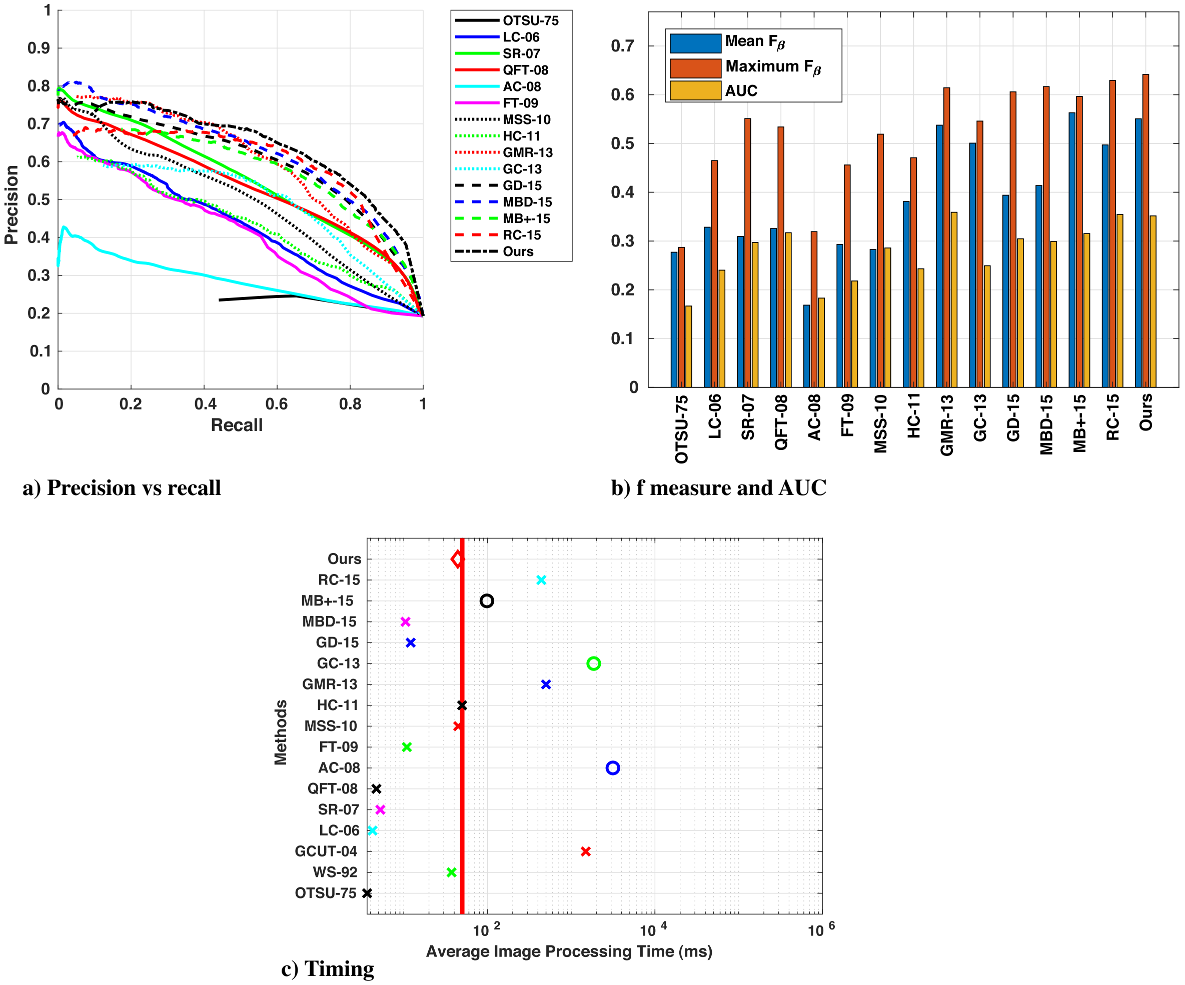


Fig. 10 Saliency map comparison of selected images from the grayscale SATSEG dataset.



a) Precision vs recall

b) f measure and AUC

c) Timing

Fig. 11 Performance comparison plots for image saliency using the grayscale SATSEG dataset: saliency precision performance (top), and saliency speed performance (bottom). Our method is *fst+* (Algorithm 3).

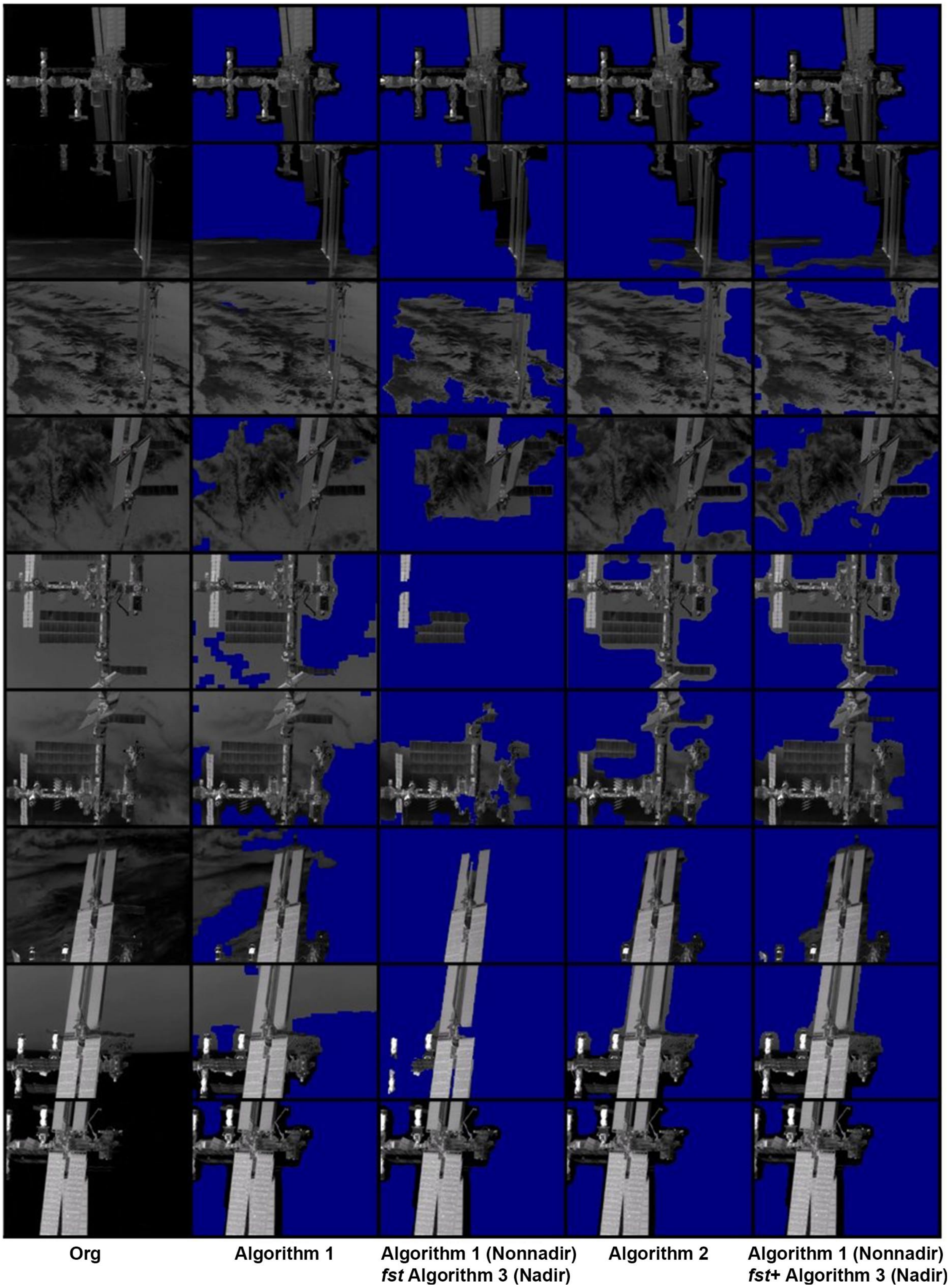


Fig. 12 ISS saliency method comparisons.

foreground extraction using Algorithm 1 on the entire video; the third column is using Algorithm 2 during the nonnadir phase and using *fst* in Algorithm 3 during the nadir phase; the fourth column is using Algorithm 2 for the entire video; and the fifth column is using Algorithm 1 during the nonnadir phase and using *fst*+ in Algorithm 3 during the nadir phase. For the ISS IR video test, *fst*+ in Algorithm 3 provides the best foreground extraction out of all techniques. The REGION_DETECT method overpredicts, whereas the *fst* method underpredicts the foreground region. The LAPLACE method performed better than the more complex and relatively more expensive REGION_DETECT and *fst*. However, the LAPLACE method can develop isolated holes or blocks as a result of the morphological operator kernel. The *fst*+ method has the best foreground extraction performance overall; it can handle the majority of the IR video by accurately extracting only the ISS image. The only exception is in a section of the Earth passage (third and fourth rows in Fig. 12), where the background developed sharp edges from the cloud regions; this section of the video is problematic for all the methods, where the ISS is mostly outside the viewable frame with only the solar panel being partially visible. The *fst*+ method has the best prediction of the solar panels but overpredicted the foreground region by including the cloud regions; this is due to a purely image-driven saliency detection, which is incapable of classifying useful and nonuseful high-frequency contents. Future work may include temporal data with some top–down guidance to increase precision.

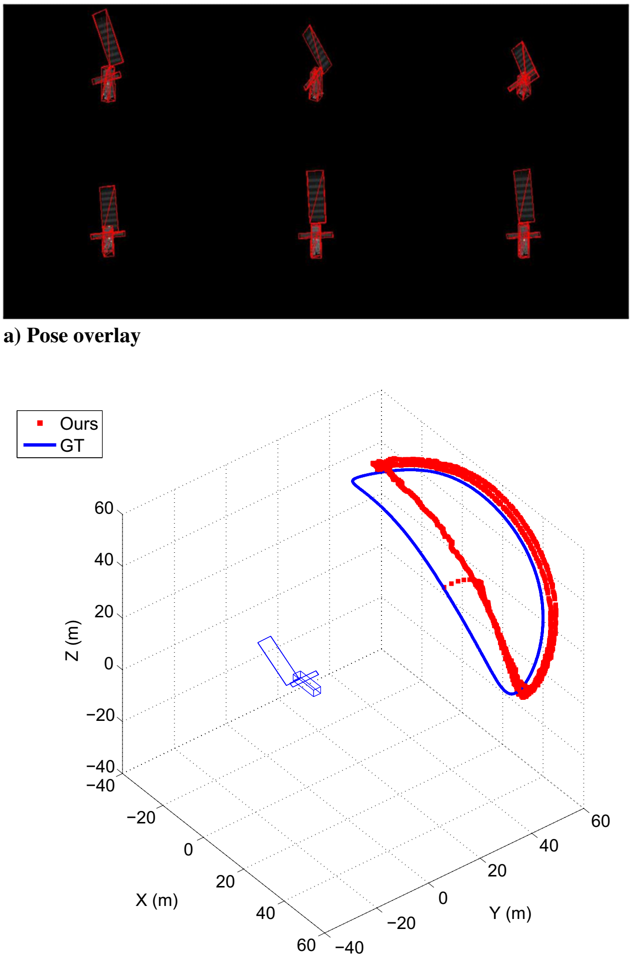
The *fst*+ effectiveness is demonstrated when it is combined with the region-based pose estimation. Figure 13 provides a single frame example of pose estimation during the nadir phase, where the first row is the pose estimation using direct IR image input without any saliency detection. The second row is the normalized foreground and background histograms associated with the first row's images. The first image in the first row is the first input image with the projected initial misalignment. The first image in the second row is the histogram template based on the prior mask. The third row is the pose estimation after applying the *fst*+ saliency map with mean value binary threshold preprocessing. The fourth row is the foreground to background histograms associated with the third row. The gradient descent iteration steps associated with the second to fifth columns are 0, 10, 30, and 60, respectively. The red lines in the histograms represent the foreground intensities, and the blue lines represent the background intensities. The initial translation and rotation misalignments are 17.3 m and 17.3 deg root sum squared, respectively. The saliency detection removes barriers from the gradient descent path to the true minimum solution. The background histogram in the second row has dimmer pixels than the background histogram in the fourth row; this is because both foreground and background histograms are normalized within its class. Indeed, the converged solution histogram for the fourth row is much closer to its template than the second row where the image without saliency detection is trapped in a local minimum.

B. Envisat Synthetic Image Pose Estimation

Next, we applied the region-based pose estimation technique on various spacecraft proximity operations scenarios including simulated Envisat tumbling and STS-135 ISS depart flight segments.

1. *Envisat Synthetic Image Pose Estimation*

The simulated Envisat motion includes rotation in the roll and yaw axis over two cycles. The Envisat is 80 m away from the servicing spacecraft camera with a tumbling rate of 10 deg/s. The camera resolution is 320 × 240 with calibrated intrinsic properties of $fS_x = 439.967$, $fS_y = 432.427$, and $fS_\theta = -0.0699286$. Figures 14 and



a) Pose overlay

b) 3-D pose

Fig. 14 Envisat pose estimation and image overlay: estimated pose image registration overlays (top), and camera pose estimation 3-D plot relative to the target Envisat (bottom).

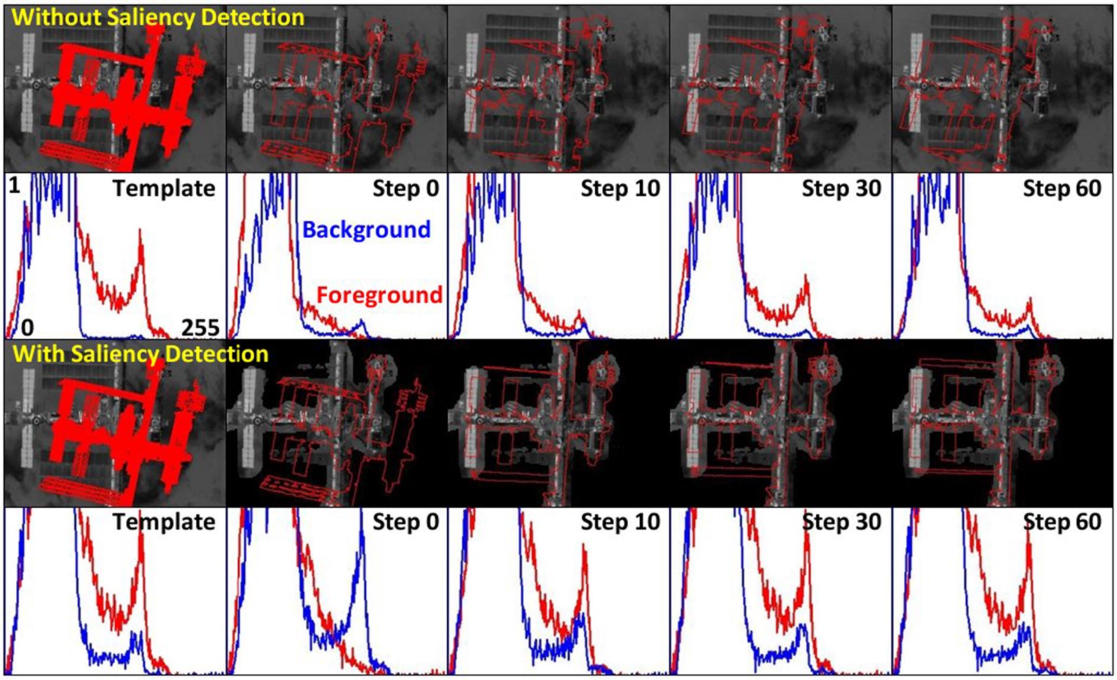
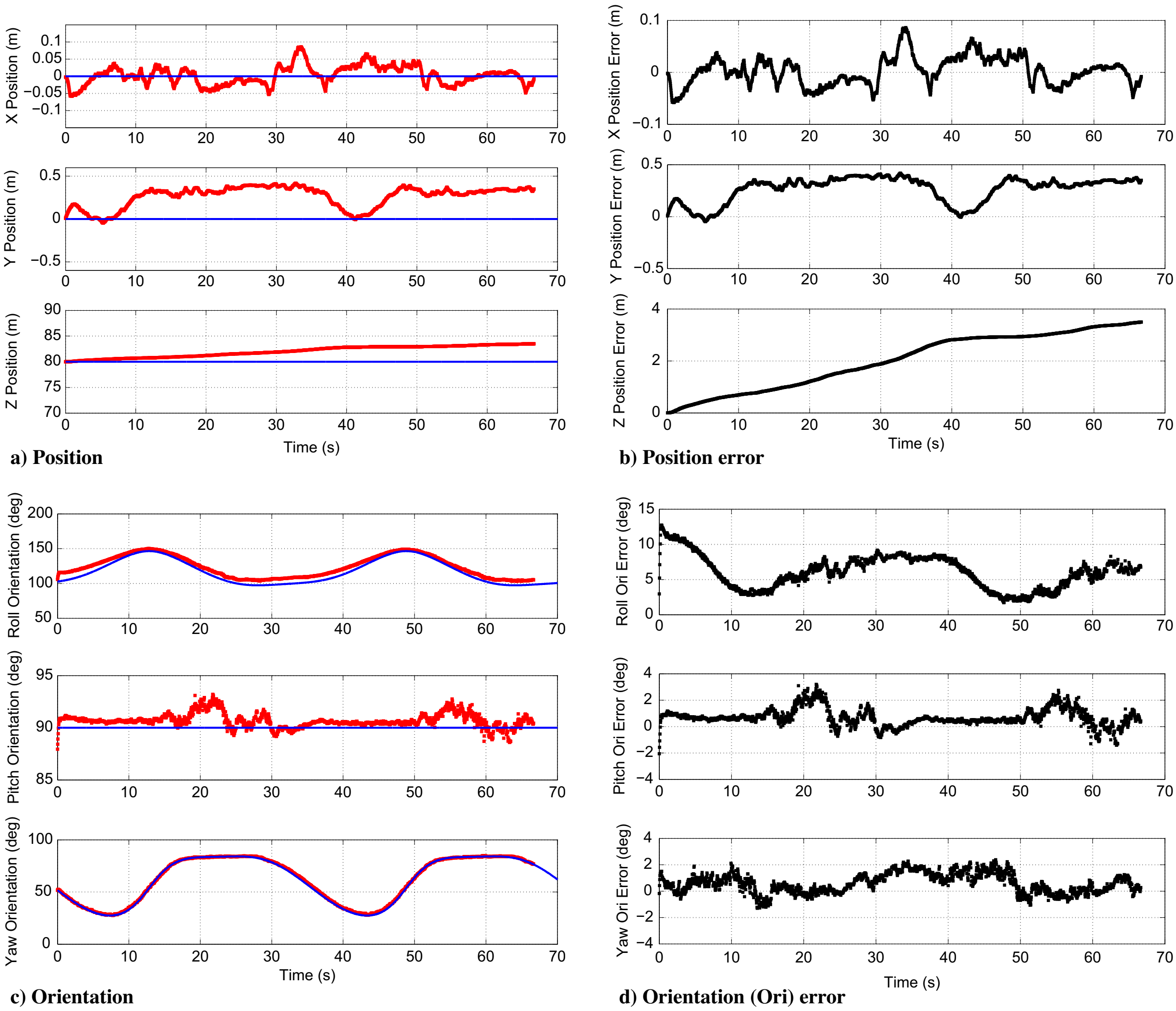


Fig. 13 STS-135 ISS undocking and flyby pose estimation during Earth passage. Blue and red lines are background and foreground normalized histograms, respectively.



a) Position

b) Position error

c) Orientation

d) Orientation (Ori) error

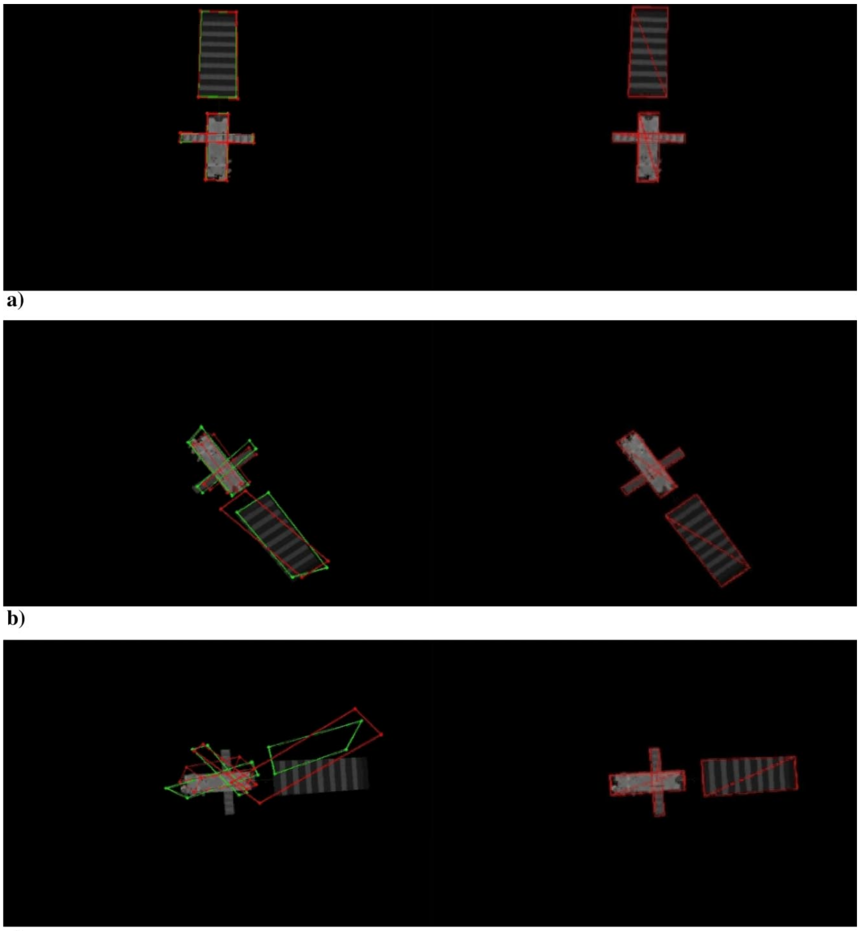
Fig. 15 Time plot of Envisat pose estimation (left) and estimation error (right) between estimated pose computed from monochromatic monocular camera images and ground truth.

15 provide the pose estimation and estimation error of the resulting simulated Envisat tumbling motion. Figures 14a provides the grayscale images captured by a virtual camera in the simulation environment. The simplified 3-D model is projected onto the images as visual overlays in red over one tumbling cycle. Figure 14b provides the equivalent 3-D plot of the Envisat motion computed using only the captured images compared to the ground truth (GT) in the Envisat spacecraft body frame. The position and orientation of the Envisat relative to the camera expressed in the camera frame are shown in Figs. 15a and 15c, respectively, where red lines are the computed pose and blue lines are the ground truth. The position and orientation errors are shown in Figs. 15b and 15d, respectively. The camera pose is defined as the position from the camera frame to the Envisat body frame expressed in the camera frame, as well as orientation of the Envisat body frame rotated from the camera frame using the Pitch-Yaw-Roll (PYR) Euler angle rotation sequence. The offset error in the X and Y directions is less than 0.1 and 0.5 m, respectively. There is a total root-sum-squared (RSS) lateral error of 0.6% out of 80 m distance between the camera and Envisat. There is a Z-axis drift of less than 4 m or 5%. Figure 15d shows the orientation error is bounded within 15 and 6 deg for roll and wobble, respectively. With respect to the absolute mean orientation, the approximate error is 10 and 3.3% for roll and wobble, respectively. It is not surprising that the boresight error is larger than the lateral when using a single camera because without triangulation or time-of-flight measurement, the depth is dependent on the region area and image resolution, where a large distance between the target vehicle and the camera will result in a larger bore sight measurement error. An initial translational pose offset shown in Fig. 14b is caused by an orientation offset when expressed in the camera frame. Due to the RSS distance between the two vehicles being 80 m, any small rotation error has been amplified into larger translation errors when expressed in the Envisat body frame. Figure 15c shows the orientation offset during peak oscillations. The orientation offset is largely due to the shadowing of the Envisat lower solar panel reducing the area of the spacecraft. This error highlights the sensitivity of the region-based method accuracy to the observable area.

2. *Pose Estimation Technique Comparison*

We compare our region-based pose estimation method with a popular point-based method called *efficient perspective-N point* (EPNP) [78]. The EPNP method computes the relative pose of an object by using geometric constraints from four virtual points generated using image features. For the pose estimation comparison, we use a combination of scale invariant feature transform (SIFT) [28] image feature points combined with homography transform, EPNP, and random sample and consensus [79] for the final pose estimation calculation. Details of the point-based estimation pipeline are provided in the work of Shi et al. [80]. The main benefit of the region-based method is adding stability to the pose estimation compared to the point-based method. Some examples of the pose estimation are shown in Fig. 16.[§] The top figure shows the two methods computing

[§] Data available online at https://youtu.be/8--Km–FOmC8E [retrieved 29 January 2021].


a)

b)

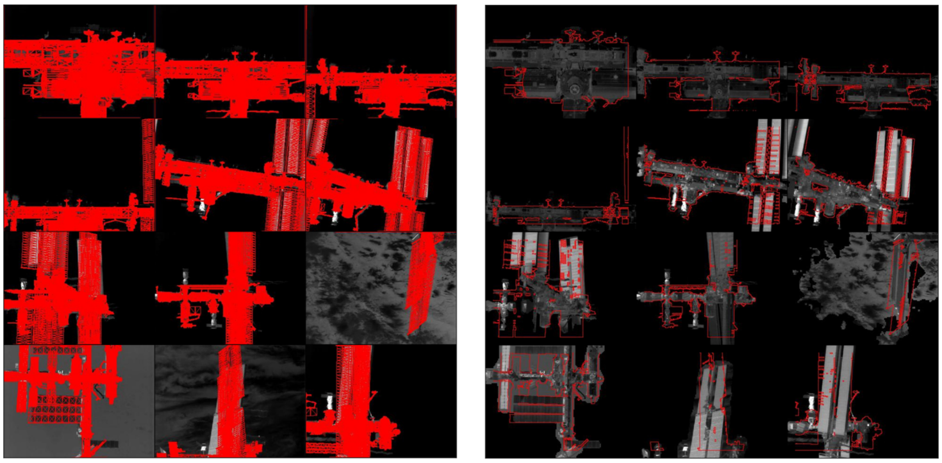
c)

Fig. 16 Pose estimation comparison between SIFT + EPNP (left) and our proposed method (right). Feature key-point efficient prospective-*N*-point method is less stable than the proposed regional method.

the correct pose of the Envisat spacecraft; whereas in the bottom two images, the homography–EPNP method fails to produce the proper pose (where the green lines and points are the initial step homography transform and the red lines and points are the EPNP matching). In the middle image, the homography transformed correctly and the EPNP failed to produce an acceptable pose; whereas in the bottom image, both homography and EPNP failed to produce the right results. The main reasons are because homography will not always be able to find a satisfactory transformation, and EPNP precision has dependencies on the amount of perspective observed from the image points. The proposed region-based method, on the other hand, is computed smoothly throughout the sequence because the blob region is more stable than the point features.

3. *ISS Pose Estimation*

The NASA ISS local orbital coordinate system or the local-vertical/local-horizontal (LVLH) [81] is in the general direction of


a) Pose overlay

b) Projected silhouette

Fig. 17 STS-135 ISS undocking and proximity flyby sequence pose estimation: estimated pose 3-D projected image overlay (left), and level-set function silhouette overlay (right).

the ISS body frame with minor attitude offsets. Figure 17 shows the STS-135 ISS undocking and the proximity flyby sequence IR image pose estimation; it provides the six-degree-of-freedom motion sequence of the SSO undocking from the ISS initially departing along the *forward V* bar (LVLH X), and then it turns to the negative H bar (LVLH Y) and performs a proximity flyby *overhead* of the negative R bar (LVLH $-Z$) from the ISS's *starboard* to *port* [82]. The *fst*+ saliency mask is applied to the input IR image. Figure 18 provides the 3-D reconstruction of the camera trajectory expressed in the ISS body coordinate system (ISSBCS) [81].

The initial condition estimated pose needs to be reinitialized every 10 frames on average to avoid error buildup. Certain phases of the proximity operations are more robust than others; for example the initial V-bar departure and the negative R-bar overhead transition. Other relative motions are not as precise; namely, the V-bar to H-bar turn, the turn to the R-bar, and turning from the R-bar to the H-bar. The V-bar departure and R-bar transition are mostly translational motion, whereas the turning sequences require changing the roll and pitch-axis Euler angles as shown in Fig. 18, where the camera frame with respect to the ISSBCS is shown in red squares, a to-scale simplified ISS model outline is shown in blue lines, and the ISSBCS is located at the ISS mass center and is in the general direction of the ISS LVLH coordinate system with slight attitude offsets. The pose estimation error is mainly caused by the unknown ISS subcomponent configuration: for example, the solar panel pan and tilt angles; the SSRMS joint configuration; the radiation panel orientations; and the Soyuz module attachment location, shape, and the Soyuz module attachment location, shape, and its solar panel deployment position. All these unknowns add to the projection error and reduce the odds of achieving a perfect alignment. Furthermore, in some frames, the ISS is not in full view. The regional method performs best when the entire vehicle is displayed in the image with distinctive shape features. Conversely, when the projected region almost entirely covers the image (such as frames 1 and 366 in Fig. 17) or only a tiny portion of the vehicle is displayed (such as frames 585 and 731), the regional method performs poorly due to ambiguity in the region silhouette. In Fig. 17, sequence frames 1, 74, 147, 220, 293, 366, 439, 512, 585, 658, 731, and 793 are displayed, respectively. Figure 17a is the projection overlay of the 3-D model on the ISS IR image. Figure 17b is the silhouette outline of the projection overlay while using the *fst*+ saliency detection.

Based on the experiment results, it is recommended to include additional correction methods to increase prediction accuracy, such as using feature localization when the target is near; and there is an abundant amount of image features within the boundary region that can be used to clarify silhouette ambiguity. Additionally, the pose estimation precision can improve by combining stochastic filters with the camera pose estimation, taking into account predictions in the dynamic motion. This can be done, namely, by implementing a Kalman filter. Local histogram template matching may possibly also reduce region ambiguity, as demonstrated by Hexner and Hagege [48] and Tjaden et al. [50].

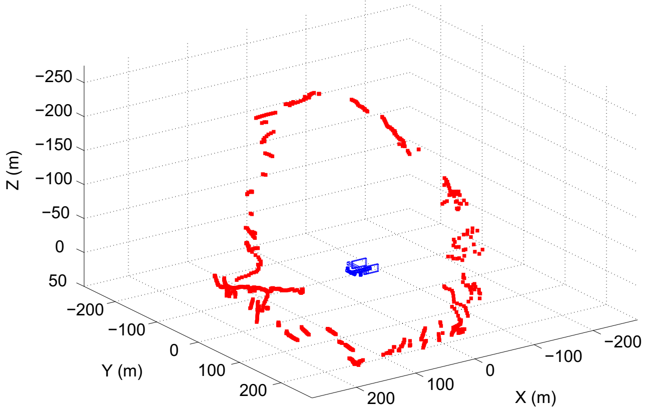


Fig. 18 STS-135 ISS undocking and proximity flyby 3-D reconstruction from monochromatic monocular camera-only pose estimation.

VII. Conclusions

In this paper, a novel image-driven approach to distinguish nadir and nonnadir camera pointing was provided. Specifically, several saliency detection models were developed, including the LAPLACE foreground mask algorithm, as well as the *fst* and *fst*+ enhanced graph manifold ranking models for working with infrared grayscale images. The current evaluation of the *fst*+ model shows performance exceeding traditional and state-of-the-art saliency detection methods, and the method is 11 times faster than the original graph manifold ranking method. Using these new algorithms, an end-to-end real-time spacecraft state generation model was then presented that includes saliency foreground extraction and region-based pose estimation using level-set segmentation and pixel statistics. Innovative initialization and gradient descent formulations for stable and efficient pose estimation convergence were introduced. Finally, using this approach for pose estimation of a complex six-degree-of-freedom uncooperative target spacecraft motion was successfully demonstrated. Future work will include the improvement of the robustness of this pose estimation method by using stochastic filters and dynamic model predictions.

Acknowledgments

This research was jointly funded by the Natural Sciences and Engineering Research Council Scholarship CGSD3-453738-2014, the Canadian Space Agency Space Technology Development Program, and the Ontario Centres of Excellence Voucher for Innovation and Productivity II award 24053.

References

[1] Ruel, S., Luu, T., and Berube, A., “Space Shuttle Testing of the TriDAR 3D Rendezvous and Docking Sensor,” *Journal of Field Robotics*, Vol. 29, No. 4, 2012, pp. 535–553.

[2] Liu, C., and Hu, W., “Relative Pose Estimation for Cylinder-Shaped Spacecrafts Using Single Image,” *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 50, No. 4, 2014, pp. 3036–3056.

[3] Ventura, J., Fleischner, A., and Walter, U., “Pose Tracking of a Noncooperative Spacecraft During Docking Maneuvers Using a Time-of-Flight Sensor,” *AIAA Guidance, Navigation and Control Conference and Exhibit*, AIAA Paper 2016-0875, Jan. 2016.

[4] Zhang, L., Zhu, F., Hao, Y., and Pan, W., “Optimization-Based Non-Cooperative Spacecraft Pose Estimation Using Stereo Cameras During Proximity Operations,” *Applied Optics*, Vol. 56, No. 15, 2017, pp. 4522–4531.

[5] Casonato, G., and Palmerini, G. B., “Visual Techniques Applied to the ATV/ISS Rendezvous Monitoring,” *2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No.04TH8720)*, Vol. 1, IEEE Publ., Piscataway, NJ, March 2004, p. 625.
https://doi.org/10.1109/AERO.2004.1367648

[6] Mokuno, M., and Kawano, I., “In-Orbit Demonstration of an Optical Navigation System for Autonomous Rendezvous Docking,” *Journal of Spacecraft and Rockets*, Vol. 48, No. 6, 2011, pp. 1046–1054.

[7] Wu, Y., Yang, G., Lin, J., Raus, R., Zhang, S., and Watt, M., “Low-Cost, High-Performance Monocular Vision System for Air Bearing Table Attitude Determination,” *Journal of Spacecraft and Rockets*, Vol. 51, No. 1, 2014, pp. 66–75.

[8] Zhang, G., Kontitsis, M., Filipe, N., Tsiotras, P., and Vela, P., “Cooperative Relative Navigation for Space Rendezvous and Proximity Operations Using Controlled Active Vision,” *Journal of Field Robotics*, Vol. 32, No. 2, 2016, pp. 205–228.

[9] Fourie, D., Tweddle, B., Ulrich, S., and Saenz-Otero, A., “Flight Results of Vision-Based Navigation for Autonomous Spacecraft Inspection of Unknown Objects,” *Journal of Spacecraft and Rockets*, Vol. 51, No. 6, 2014, pp. 2016–2026.

[10] Lim, T. W., Ramos, P. F., and O’Dowd, M. C., “Edge Detection Using Point Cloud Data for Noncooperative Pose Estimation,” *Journal of Spacecraft and Rockets*, Vol. 54, No. 2, 2017, pp. 500–505.

[11] Lourakis, M., and Zabulis, X., “Satellite Visual Tracking for Proximity Operations in Space,” *Proceedings of 14th ESA Workshop on Advanced Space Technologies for Robotics and Automation*, Leiden, The Netherlands, June 2017, pp. 1–8.

[12] Sharma, S., Ventura, J., and D’Amico, S., “Robust Model-Based Monocular Pose Initialization for Noncooperative Spacecraft Rendezvous,” *Journal of Spacecraft and Rockets*, Vol. 55, No. 6, 2018, pp. 1414–1429.

[13] Curtis, D., and Cobb, R., "Satellite Articulation Tracking Using Computer Vision," *Journal of Spacecraft and Rockets*, Vol. 56, No. 5, 2019, pp. 1478–1491.

[14] Bodin, P., Larsson, R., Nilsson, F., Chasset, C., Noteborn, R., and Nylund, M., "PRISMA: An In-Orbit Test Bed for Guidance, Navigation, and Control Experiments," *Journal of Spacecraft and Rockets*, Vol. 46, No. 3, 2009, pp. 615–623.

[15] Modenini, D., "Five-Degree-of-Freedom Pose Estimation from an Imaged Ellipsoid of Revolution," *Journal of Spacecraft and Rockets*, Vol. 56, No. 3, 2019, pp. 952–958.

[16] Christian, J., "Accurate Planetary Limb Localization for Image-Based Spacecraft Navigation," *Journal of Spacecraft and Rockets*, Vol. 54, No. 3, 2017, pp. 708–730.

[17] Shi, J., Ulrich, S., and Ruel, S., "Unsupervised Method of Infrared Spacecraft Image Foreground Extraction," *Journal of Spacecraft and Rockets*, Vol. 56, No. 6, 2019, pp. 1847–1856.

[18] Liu, Z., Xiang, Q., Tang, J., Wang, Y., and Zhao, P., "Robust Salient Object Detection for RGB Images," *Visual Computer Journal*, Vol. 36, Dec. 2019, pp. 1823–1835.

[19] Huang, K., and Gao, S., "Image Saliency Detection via Multi-Scale Iterative CNN," *Visual Computer Journal*, Vol. 36, Aug. 2019, pp. 1355–1367.

[20] Kisantal, M., Sharma, S., Park, T., Izzo, D., Märtens, M., and D'Amico, S., "Satellite Pose Estimation Challenge: Dataset, Competition Design and Results," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 56, No. 5, 2020, pp. 4083–4098.

[21] Rosenhahn, B., Perwass, C., and Sommer, G., "Foundations About 2D-3D Pose Estimation," *CVonline: The Evolving Distributed, Non-Proprietary, On-Line Compendium of Computer Vision*, 2004.

[22] Lepetit, V., and Fua, P., "Monocular Model-Based 3D Tracking of Rigid Objects: A Survey," *Foundations and Trends in Computer Graphics and Vision*, Vol. 1, No. 1, 2005, pp. 1–89.

[23] Capuano, V., Ryan Alimo, S., Ho, A., and Chung, S., "Robust Features Extraction for On-Board Monocular-Based Spacecraft Pose Acquisition," *AIAA Guidance, Navigation, and Controls Conference and Exhibit*, AIAA Paper 2019-2005, Jan. 2019.

[24] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Publ., Piscataway, NJ, 2016, pp. 770–778.

[25] Szegedy, C., Ioffe, S., and Vanhoucke, V., "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," Preprint, submitted 23 Feb. 2016, https://arxiv.org/abs/1602.07261.

[26] Harris, C., and Stephens, M., "A Combined Corner and Edge Detector," *Proceedings of the 4th Alvey Vision Conference*, Vol. 15, 1988, pp. 147–151.

[27] Rosten, E., and Drummond, T., "Machine Learning for High-Speed Corner Detection," *European Conference on Computer Vision*, 2006, pp. 430–443.

[28] Lowe, D., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, 2004, pp. 91–110.

[29] Alcantarilla, P., Nuevo, J., and Bartoli, A., "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 7, 2011, pp. 1281–1298.

[30] Choi, C., and Christensen, H., "RGB-D Object Pose Estimation in Unstructured Environments," *Robotics and Autonomous System*, Vol. 75, Jan. 2016, pp. 595–613.

[31] Zivkovic, Z., and van der Heijden, F., "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction," *Pattern Recognition Letters*, Vol. 27, No. 7, 2006, pp. 773–780.

[32] Godbehere, A., Matsukawa, A., and Goldberg, K., "Visual Tracking of Human Visitors Under Variable-Lighting Conditions for a Responsive Audio Art Installation," *American Control Conference*, 2012, pp. 4305–4312.

[33] Long, J., Shelhamer, E., and Darrel, T., "Fully Convolutional Networks for Semantic Segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Publ., Piscataway, NJ, 2015, pp. 3431–3440.

[34] Badrinarayanan, V., Handa, A., and Cipolla, R., "Segnet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling," *Computing Research Repository*, Vol. abs/1505.07293, 2015.

[35] Yang, J., and Yang, M., "Top-Down Visual Saliency via Joint CRF and Dictionary Learning," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 39, No. 3, 2017, pp. 576–588.

[36] Li, L., Zhou, F., Zheng, Y., and Bai, X., "Saliency Detection Based on Foreground Appearance and Background-Prior," *NeuroComputing*, Vol. 301, Aug. 2018, pp. 46–61.

[37] Krizhevsky, A., Sutskever, I., and Hinton, G., "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, Vol. 25, 2012, pp. 1097–1105.

[38] Li, J., Tian, Y., Chen, X., and Huang, T., "Measuring Visual Surprise Jointly from Intrinsic and Extrinsic Contexts for Image Saliency Estimation," *International Journal of Computer Vision*, Vol. 120, No. 1, 2016, pp. 44–60.

[39] Cheng, M., Mitra, N., Huang, X., Torr, P., and Hu, S., "Global Contrast Based Salient Region Detection," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 37, No. 3, 2015, pp. 569–582.

[40] Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., and Měch, R., "Minimum Barrier Salient Object Detection at 80 FPS," *Proceedings of the IEEE International Conference on Computer Vision*, IEEE Publ., Piscataway, NJ, 2015, pp. 1404–1412.

[41] Yang, C., Zhang, L., Lu, H., Ruan, X., and Yang, M., "Saliency Detection via Graph-Based Manifold Ranking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Publ., Piscataway, NJ, 2013, pp. 3166–3173.

[42] Strand, R., Ciesielski, K., Malmberg, F., and Saha, P., "The Minimum Barrier Distance," *Computer Vision and Image Understanding*, Vol. 117, No. 4, 2013, pp. 429–437.

[43] Jung, C., Kim, W., Yoo, S., and Kim, C., "A Novel Monochromatic Cue for Detecting Regions of Visual Interest," *Journal of Image and Vision Computing*, Vol. 32, Nos. 6–7, 2014, pp. 405–413.

[44] Yacoob, Y., and Davis, L., "Segmentation Using Meta-Texture Saliency," *2007 IEEE 11th International Conference on Computer Vision*, IEEE Publ., Piscataway, NJ, 2007, pp. 801–808.

[45] Dambreville, S., Sandhu, R., Yezzi, A., and Tannenbaum, A., "A Geometric Approach to Joint 2D Region-Based Segmentation and 3D Pose Estimation Using a 3D Shape Prior," *SIAM Journal on Imaging Sciences*, Vol. 3, No. 1, 2010, pp. 110–132.

[46] Prisacariu, V., and Reid, I., "PWP3D: Real-Time Segmentation and Tracking of 3D Objects," *International Journal of Computer Vision*, Vol. 98, No. 3, 2012, pp. 335–354.

[47] Perez-Yus, A., Puig, L., Lopez-Nicolas, G., Guerrero, J., and Fox, D., "RGB-D Based Tracking of Complex Objects," *International Workshop on Understanding Human Activities Through 3D Sensors*, Springer, Cham, 2016, pp. 115–127.

[48] Hexner, J., and Hagege, R., "2D-3D Pose Estimation of Heterogeneous Objects Using a Region Based Approach," *International Journal of Computer Vision*, Vol. 118, No. 1, 2016, pp. 95–112.

[49] Tjaden, H., Schwanecke, U., and Schömer, E., "Real-Time Monocular Segmentation and Pose Tracking of Multiple Objects," *European Conference on Computer Vision*, 2016, pp. 423–438.

[50] Tjaden, H., Schwanecke, U., and Schömer, E., "Real-Time Monocular Pose Estimation of 3D Objects Using Temporally Consistent Local Color Histograms," *IEEE International Conference on Computer Vision*, IEEE Publ., Piscataway, NJ, 2017, pp. 124–132.

[51] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," *Automatica*, Vol. 9, No. 1, 1979, pp. 62–66.

[52] Zhou, D., Weston, J., and Gretton, A., "Ranking on Data Manifolds," *Advances in Neural Information Processing Systems*, Vol. 16, 2004, pp. 169–176.

[53] Wan, X., Yang, J., and Xiao, J., "Manifold-Ranking Based Topic-Focused Multi-Document Summarization," *Proceedings of International Joint Conference on Artificial Intelligence*, 2007, pp. 2903–2908.

[54] Xu, B., Bu, J., Chen, C., Cai, D., He, X., Liu, W., and Luo, J., "Efficient Manifold Ranking for Image Retrieval," *ACM Special Interest Group on Information Retrieval*, 2011.

[55] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S., "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 34, No. 11, 2012, pp. 2274–2281.

[56] Margolin, R., Tal, A., and Zelnik-Manor, L., "What Makes a Patch Distinct?" *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Publ., Piscataway, NJ, NJ, 2013, pp. 1139–1146.

[57] Hou, X., and Zhang, L., "Saliency Detection: A Spectral Residual Approach," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Publ., Piscataway, NJ, 2007, pp. 1–8.

[58] Osher, S., and Sethian, J., "Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations," *Journal of Computational Physics*, Vol. 79, No. 1, 1988, pp. 12–49.

[59] Mumford, D., and Shah, J., "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems," *Communications on Pure and Applied Mathematics*, Vol. 42, No. 5, 1989, pp. 577–685.

[60] Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, E., and Willsky, A., "Model-Based Curve Evolution Technique for Image Segmentation," *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, IEEE Publ., Piscataway, NJ, 2013, 2001, pp. 463–468.

[61] Chan, T., and Vese, L., "Active Contours Without Edges," *IEEE Transactions on Image Processing*, Vol. 10, No. 2, 2001, pp. 266–277.

[62] Cremers, D., Rousson, M., and Deriche, R., "A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape," *International Journal of Computer Vision*, Vol. 72, No. 2, 2007, pp. 195–215.

[63] Bibby, C., and Reid, I., "Robust Real-Time Visual Tracking Using Pixel-Wise Posteriors," *European Conference on Computer Vision*, 2008, pp. 831–844.

[64] Nelder, J., and Mead, R., "A Simplex Method for Function Minimization," *Computer Journal*, Vol. 7, No. 4, 1965, pp. 308–313.

[65] Hager, W., and Zhang, H., "A New Conjugate Gradient Method with Guaranteed Descent and an Efficient Line Search," *SIAM Journal on Optimization*, Vol. 16, No. 1, 2005, pp. 170–192.

[66] Zhai, Y., and Shah, M., "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues," *Proceedings of the 14th ACM International Conference on Multi-Media*, 2016, pp. 815–824.

[67] Guo, C., Ma, Q., and Zhang, L., "Spatio-Temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Soc., Piscataway, NJ, 2008.

[68] Achanta, R., Estrada, F., Wils, P., and Süsstrunk, S., "Salient Region Detection and Segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Soc., Piscataway, NJ, 2008.

[69] Borji, A., Cheng, M., Jiang, H., and Li, J., "Salient Object Detection: A Benchmark," *IEEE Transactions on Image Processing*, Vol. 24, No. 12, 2015, pp. 5706–5723.

[70] Achanta, R., Hemami, S., Estrada, F., and Süsstrunk, S., "Frequency-Tuned Salient Region Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Publ., Piscataway, NJ, 2009, pp. 1597–1604.

[71] Achanta, R., and Süsstrunk, S., "Saliency Detection Using Maximum Symmetric Surround," *IEEE International Conference on Image Processing*, IEEE Publ., Piscataway, NJ, Sept. 2010.

[72] Cheng, M., Zhang, G., Mitra, N., Huang, X., and Hu, S., "Global Contrast Based Salient Region Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Publ., Piscataway, NJ, 2011, pp. 409–416.

[73] Cheng, M., Warrell, J., Lin, W., Zheng, S., Vineet, V., and Crook, N., "Efficient Salient Region Detection with Soft Image Abstraction," *IEEE International Conference on Computer Vision*, IEEE Publ., Piscataway, NJ, 2013, pp. 1529–1536.

[74] Shi, J., Yan, Q., Xu, L., and Jia, J., "Hierarchical Image Saliency Detection on Extended CSSD," *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 38, No. 4, 2016, pp. 717–729.

[75] Fawcett, T., "An Introduction to ROC Analysis," *Pattern Recognition Letters*, Vol. 27, No. 8, 2006, pp. 861–874.

[76] Meyer, F., "Color Image Segmentation," *IET 1992 International Conference on Image Processing and its Applications*, IET, England, U.K., 2002, pp. 303–306.

[77] Rother, C., Kolmogorov, V., and Blake, A., "GrabCut-Interactive Foreground Extraction Using Iterated Graph Cuts," *ACM Transactions on Graphics*, Vol. 23, No. 3, 2004, pp. 309–314.

[78] Lepetit, V., "EPnP: An Accurate O(n) Solution to the PnP Problem," *International Journal of Computer Vision*, Vol. 81, No. 2, 2009, pp. 155–166.

[79] Fischler, M., and Bolles, R., "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Graphics and Image Processing*, Vol. 24, No. 6, 1981, pp. 381–395.

[80] Shi, J., Ulrich, S., and Ruel, S., "Spacecraft Pose Estimation Using a Monocular Camera," *Proceedings of the International Astronautical Congress*, Paper IAC-16-C1.3.4, Sept. 2016.

[81] "Space Station Reference Coordinate Systems," NASA International Space Station Program, Rev. H, SSP-30219, June 2005.

[82] Goodman, J., "Rendezvous and Proximity Operations of the Space Shuttle," NASA Technical Reports Server, July 2005.

J. A. Christian
Associate Editor