Cross-Modal Localization: Using automotive radar for absolute geolocation within a map produced with visible-light imagery

Peter A. Iannucci, Lakshay Narula, and Todd E. Humphreys
Radionavigation Laboratory
The University of Texas at Austin
Austin, TX, USA
peter.iannucci@austin.utexas.edu, {lakshay.narula, todd.humphreys}@utexas.edu

***Abstract*—This paper explores the possibility of localizing an automotive-radar-equipped vehicle within an urban environment relative to an existing map of the environment created using data from visible light cameras. Such cross-modal localization would enable robust, low-cost absolute localization in poor weather conditions based only on radar even when the vehicle has never previously visited the area. This is because a pre-existing absolutely-referenced visible-light-based map (e.g., constructed from Google Street View images) could be exploited for localization provided that a correspondence between features in this map and the vehicle's radar returns can be established. The greatest challenge presented by cross-modal localization with automotive radar is the extreme sparseness of automotive-radar-produced features, which prevents application of standard computer vision techniques for the cross-modal registration. To the best of the authors' knowledge, cross-modal localization using automotive-grade radar within a visible-light-based map is unprecedented. The current paper demonstrates that it can be used for vehicle localization with horizontal errors below 61 cm (95%).**

***Index Terms*—cross modal; localization; radar-based localization; vision-based localization**

I. Introduction

Abundant visible light imagery—from spaceborne cameras [1], airborne platforms [2], and ground-based cameras [2]—has given rise to large (often public) collections of visible-light images and 3D maps based on these. Localizing a visible light camera by comparing its images against collections of geotagged images is by now a mature technique for urban navigation [3], [4]. Indeed, the recently-released Google Maps augmented reality (AR) pedestrian navigation solution is based exactly on this technique: features extracted from images taken by the user's camera are correlated with similar features extracted from Google's vast trove of Street View imagery, with the result that pedestrians can confidently and precisely navigate within urban environments even when global navigation satellite system (GNSS) coverage is poor.

Given the abundance of geotagged visible-light imagery and the need to perform localization in adverse lighting or weather conditions, there has been longstanding interest in cross-modal image registration [5] and cross-spectral visual odometry [6] and simultaneous localization and mapping (SLAM) [7]. For spaceborne and airborne sensing, vision-to-synthetic-aperture-radar (SAR) image registration has been the focus of a sustained research effort over the past two decades [8]–[10]. For ground vehicles, cross-modal data association has focused exclusively on correspondences between thermal (infrared) imagery and visible light imagery [6], [7], [11].

The authors of the present paper were unable to find prior work in radar-to-visible-light cross-modal odometry or SLAM for ground vehicles. Indeed, as regards the type of low-cost automotive radar common on modern cars, there is but a scant literature in radar-based odometry and SLAM. Apart from recent promising work by the current authors [12], only one other research group has demonstrated useful performance, and this within a highly favorable and compact landscape [13], [14]. Better SLAM and odometry peformance has been demonstrated using high-cost scanning radars on shipborne [15] and automotive [16]–[18] platforms. These rotating radar units produce image-like scans that are amenable to feature extraction. The image-like radar scans shown in [16] and [17] suggest that cross-modal registration between such rotating-radar scans and geotagged visible-light images may be possible, but this has not been explored.

The present work explores the more-challenging task of cross-modal registration between sparse scans produced by automotive-grade radar and geotagged visible-light imagery. Fig. 1 illustrates why this problem is hard: there is only weak correspondence between features extracted from a stereo pair of visible-light cameras and the sparse radar features produced by a trio of automotive radar sensors.

II. Data Collection

A. Sensing Platform

This paper's experimental exploration of cross-modal localization was based on the data set described in [19], collected by the University of Texas Sensorium platform, shown in Fig. 2. The platform is equipped with stereo grayscale visible light cameras and three automotive-grade radar units. The stereo camera pair has a ± 37 deg. field of view, shown in Fig. 1. The coverage patterns of the three radar units are shown in Fig. 3.

Two Antcom G8 triple-frequency GNSS patch antennas are flush-mounted in the cross-track direction on the Sensorium's upper plate, separated by 1.05 meters. The port (driver's

Preprint of the 2020 IEEE/ION Position Location and Navigation Symposium St. Louis, MO, September 23–25, 2020

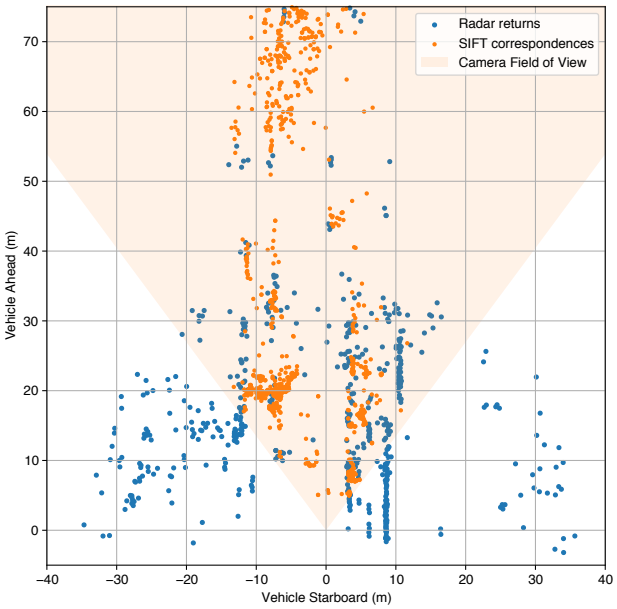

Fig. 1: Left: An image from the left camera in the Sensorium stereo camera pair. Right: Radar point features from the Sensorium's trio of radar units over four subsequent 20-Hz scans (blue) overlaid on scale invariant feature transform (SIFT) features from a single stereo image pair (including the image to the left) projected onto the horizontal plane (orange). The orange shaded area represents the common stereo camera field of view.

side) antenna feeds analog signals to an iXblue ATLANS-C GNSS-disciplined tactical-grade inertial navigation system, about which more details are provided in Sec. II-C.

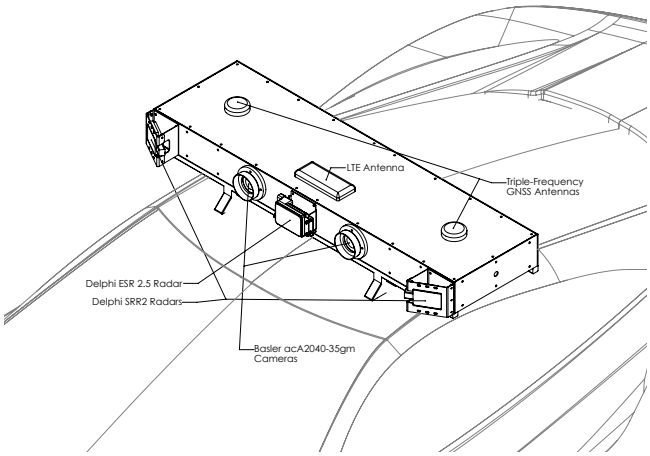

Fig. 2: The University of Texas Sensorium is a platform for automated and connected vehicle perception research. It includes stereo visible light cameras, an industrial-grade inertial measurement unit (IMU), an automotive radar unit, a dual-antenna, dual-frequency software-defined GNSS receiver, 4G cellular connectivity, and a powerful internal computer. The long-range electrically scanning radar (ESR) is shown mounted at the center of the sensorium's front plate. Two short-range radar (SRR2) units are mounted at a 30-deg. outward-facing angle at the ends of the front plate.

B. Collection Interval and Location

Data were collected on May 12, 2019 during approximately 45 minutes of driving in and around the dense urban center of Austin, TX.

The test route, depicted in Fig. 4, runs the gamut of light-to-dense urban conditions, from open-sky to narrow streets with overhanging trees to the high-rise urban city center.

C. Ground Truth Trajectory

A trustworthy ground truth trajectory against which to compare cross-modal localization peformance is an indispensable feature of the experimental setup. The present work adopts the traditional approach of taking the forward-backward smoothed trajectory generated in after-the-fact processing by a coupled real-time-kinematic (RTK) -inertial system with a tactical-grade IMU as the ground truth [20]–[22]. In particular, an iXblue ATLANS-C mobile mapping INS/GNSS system, which incorporates a professional-grade Septentrio AsteRx3 RTK receiver, was used to generate the ground truth [23]. The ATLANS-C was rigidly mounted to the Sensorium and attached to the port GNSS antenna. A cm-accurate lever arm estimate from the inertial sensor to the GNSS antenna was determined. Self-reported 3D accuracy of the ATLANS-C's smoothed estimate varied between 2 and 20 cm (1-sigma) along the test route.

III. Data Processing

This section details a procedure that achieves 61 cm RMS position error at the 95$^{\text{th}}$ percentile cross-modal re-localization using data collected with the University of Texas Sensorium.

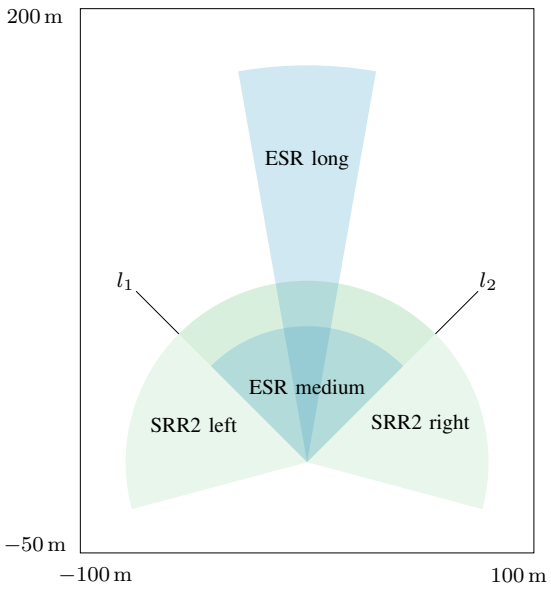

Fig. 3: Coverage patterns for the three Sensorium radar units. The ESR provides simultaneous sensing in a narrow (±10 deg.) long-range (120 m) coverage area and a wider (± 45 deg.) medium-range (60 m) area. The SRR2 units each have a coverage area of ±75 deg. and 80 m. The line l_1 marks the left-most extent of the right SSR2's field of view. Similarly, l_2 marks the right-most extent of the left SSR2's field of view. Each SRR2 is installed facing outwards at an angle of 30 deg.

Fig. 4: Overview of the test route through the urban core of Austin, TX.

What is remarkable here is not the complexity of the procedure, but its simplicity. There is little reason to expect, on physical grounds, that a radar-sensing vehicle equipped with a map constructed from visual markers alone will be capable of accurate localization. The physical mechanisms of generation of radar returns are very different from those for visible-light features, giving very different sensitivity to geometry and material properties, and the interface provided by the sensors is totally dissimilar: pixels on the one hand, and time-of-flight/angle-of-arrival tracking on the other. Nevertheless, the performance of this "simple" cross-modal localization system shows that features in a map constructed from visual markers provide an excellent starting point for inertial-aided radar localization.

This paper does not address the so-called "lost robot" problem, wherein a vehicle with largely or totally unknown coordinates obtains an initial fix. The assumption is rather that the vehicle has an initial fix, whose fidelity is decaying steadily during locomotion due to errors in inertial sensing and odometry. The vehicle uses information from its radar sensors to compensate for this decay in positioning accuracy, leading to an equilibrium. The more information it can extract from the radar sensors, the greater the positioning accuracy at equilibrium. This paper also does not address orientation uncertainty. The assumption is rather that via a combination of gyroscopic and magnetic sensors, the inertial system is largely capable of maintaining its orientation.

The problem can thus be divided into three parts: map-making, map-matching, and inertial tracking. This section details a viable approach to the first two; the third is well-understood.

A. *Map-Making*

In order to build a map from visual markers, one must first characterize the cameras. The Sensorium vision system collects ten frames per second from each camera, with shutter times traceable to GNSS timing and with software exposure control. Each frame has a resolution of 2048×732 pixels and a depth of 8 bits per pixel. The cameras are mounted forward-facing with wide-angle lenses, ultraviolet-blocking filters, and lens hoods, 25 cm to either side of the center of the Sensorium.

1) Calibration: Each camera is modeled as a pinhole with two radial and two tangential distortion coefficients, and four intrinsic parameters: focal length in two axes, and focal center in two axes. The extrinsic transformation matrix between the coordinate systems of the two cameras represents an additional six degrees of freedom. These 22 total degrees-of-freedom within the stereo vision system must be calibrated experimentally by photographing an appropriate test pattern and solving a system of equations. The procedure for this is well-known [24].

An additional six degrees of freedom connect the coordinate system of the cameras to that of the ATLANS-C INS/GNSS device. To an accuracy of roughly one centimeter of distance and a few degrees of angle, these degrees of freedom may be assumed to match their specifications in the Sensorium CAD documents. For the localization system described in this paper, it proved necessary to calibrate out an overall system bias of roughly 40 cm; this may be largely attributed to angular errors in the mounting of various components, which have since been refined.

2) Rectification: Once the intrinsic, inter-camera extrinsic, and distortion parameters of the stereo camera model are determined by calibration, the inverse distortion is applied to each camera frame in the dataset. Ideally, the calibration is "baked in" to the vehicle's onboard processor, and images are un-distorted before being saved to non-volatile storage: this is not a costly step, and can be performed on a low-end graphics processor.

3) Depth Recovery:

a) Feature Extraction: To make a map from visual markers, one must supplement the two-dimensional information present in a camera frame with a depth estimate based on the geometry of the stereo vision system. This is a well-known problem, and many well-known solutions are available. The present work proceeds in three phases, beginning with the Scale Invariant Feature Transform (SIFT) of Lowe [25]. This first phase detects local extrema in an augmented, three-dimensional "x-y-scale" representation of the camera frame, formed by repeated Gaussian filtering and downsampling. These extrema are refined and tested for contrast and selectivity versus small offsets in any direction. The extrema, now known as *keypoints*, become the loci of a "fingerprinting" procedure that summarizes nearby gradient orientations and magnitudes to form a 128-dimensional *descriptor*. Up to this point, all features (i.e. keypoints and descriptors) are independent between the left and right camera frames.

b) Feature Matching and Filtering: In the second phase of depth recovery, the present work computes the two nearest neighbors in 128-dimensional descriptor space of each left-camera descriptor among the set of right-camera descriptors, using the Euclidean distance. (This may produce a many-to-one set of matches). It then discards any best-matches which are not at least 30% better than the second-best match for the same left-camera descriptor, according to Lowe's ratio test, and it discards all but the best match for each individual right-camera descriptor. The matches are now one-to-one between a subset of the original left-camera features and an equal-sized subset of the original right-camera features, and are now known as *stereo correspondences*.

c) Stereo Disparity and Depth: The procedure next computes the delta, in pixels within their respective focal planes, between the positions of the left and right halves of each stereo correspondence. Correspondences with a vertical delta larger than a threshold of four pixels are discarded, since a well-calibrated system should have purely horizontal displacements ("disparities") due to parallax. The horizontal delta finally becomes the denominator in an expression for the distance from the left camera to the location in space where a physical object would have to be in order to gave rise to the observed stereo correspondence.

SIFT may become "cross-eyed". That is, if it detects two independent but similar patterns in the far distance, it may mistake them for a single, nearby object. In this case, there is of course no physical object at the reported distance. The present results do not require further steps to reject cross-eyed stereo correspondences.

The geometric degrees of freedom gathered from the Sensorium CAD documents in Sec.III-A1 allow the stereo correspondences to be transformed from the 3-D coordinate system aligned with the left camera to the 3-D coordinate system aligned with the ATLANS-C INS/GNSS device.

4) Building the Map: One must finally transform these points into a global (i.e. geodetic) coordinate system. The present work did this by interpolating transformation matrices constructed from the 1 Hz latitude, longitude, ellipsoidal height, heading, pitch, and roll estimates tabulated in the post-processed INS/GNSS data. Finally, to enable fast look-ups for later steps, the East and North coordinates of the transformed stereo correspondences are inserted into a two-dimensional K-D tree data-structure.

This K-D tree data-structure is the in-memory representation of the map constructed from visual markers. It may be visualized by forming a two-dimensional histogram of the feature density in an area of interest (left sides of Figs. 5, 7).

An attempt has not yet been made to separate positioning errors caused by the visual map-making process from positioning errors caused by radar processing. Only total errors will be reported.

B. Radar

Automotive radar units like the ESR and SRR2 units are not camera-like. Where an imager typically contains a two-dimensional array of millions of independent incident power detectors, each with angular selectivity better than a milliradian in each of the horizontal and vertial directions, an automotive radar contains a one-dimensional array of inter-dependent radio front-ends, with perhaps 9° of angular selectivity. One factor that sets these two problems apart is *sparsity*: the world as seen through active radar "eyes" is fairly well described as a small number of points of light in darkness. For this reason, an automotive radar unit with 9° of selectivity can nevertheless track a small number of reflectors with single-degree angular resolution–in the horizontal direction. These units are not selective in the vertical direction.

In order to make good use of limited bandwidth on the vehicle's controller area network (CAN) bus, an automotive radar unit will typically run an internal digital signal processor (DSP) pre-coded with proprietary algorithms for the acquisition and tracking of individual reflectors in the environment. Thus, the present work relies on periodic (20 Hz) reports from the radar units, each of which tabulates up to 64 individual trackers, each of which in turn incorporates state machine logic, amplitude estimation, and Doppler velocimetry. The present work ignores all of this information except for range and bearing. It is most likely possible to improve upon the results presented here by taking advantage of more information.

Range and bearing relative to the known locations of the radar units, plus the heading of the vehicle, are sufficient to compute the locations of the tracked objects in two-dimensional East and North vehicle-relative coordinates.

For visualization purposes only, the fully-known vehicle pose was combined with these vehicle-relative coordinates to build the radar "maps" shown on the right side of Figs. 5, 7.

1) Filtering Predicates: A phenomenological examination of the visual and radar maps suggested the need for a number of additional filters to restrict the features used to build each map. In particular, one might expect that stereo correspondences due to foliage and road markings are not good candidates for radar returns. In addition, distant correspondences are poor candidates for map-making, since small angular errors in calibration, amplified by long distances, form large displacements between mapped features and their real-world locations.

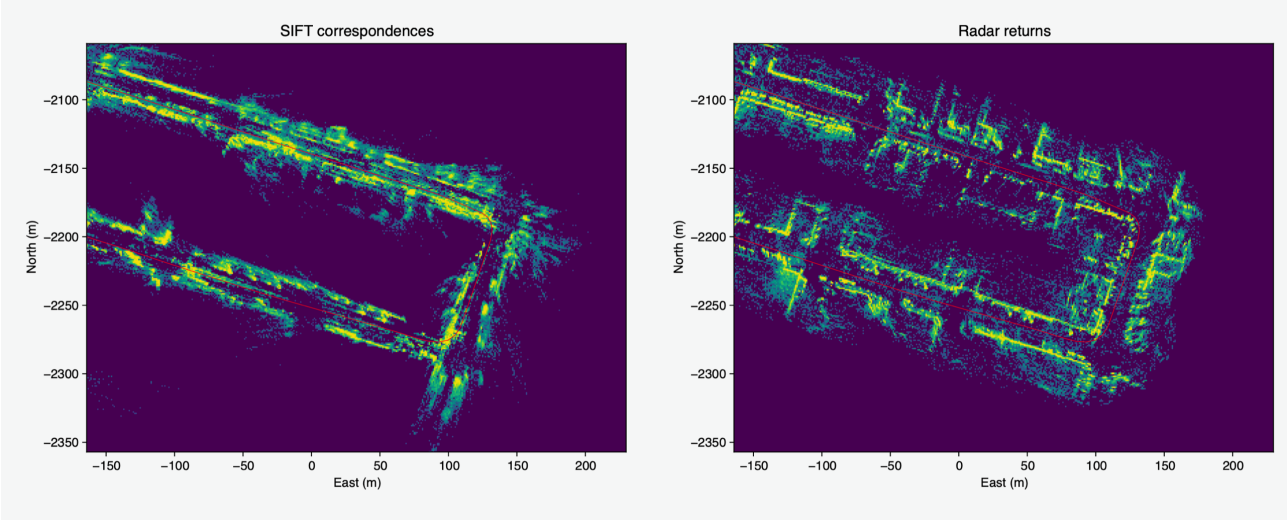

Fig. 5: Maps of portions of 10^{th} and 11^{th} streets and Red River street, Austin, Texas. As seen in the density of stereo vision features (left) and radar returns (right). Remarkably, even though the similarity between these views might, at best, be called messy, the visual map provides a highly accurate basis for radar localization. This is most likely attributable to the accumulation of both radar and visual features along walls and distinctive architectural markers, as well as parked vehicles. Because the binning approach of the present work side-steps the issue of finding a one-to-one association between visual and radar features, the computational cost of maintaining and aligning these maps is modest.

Fig. 6: View of the same region as Fig 5 in perspective, as seen in Apple Maps. Note how the radar map has resolved ground-level architectural details, such as (one of) the L-shaped planters on the west side of Red River street.

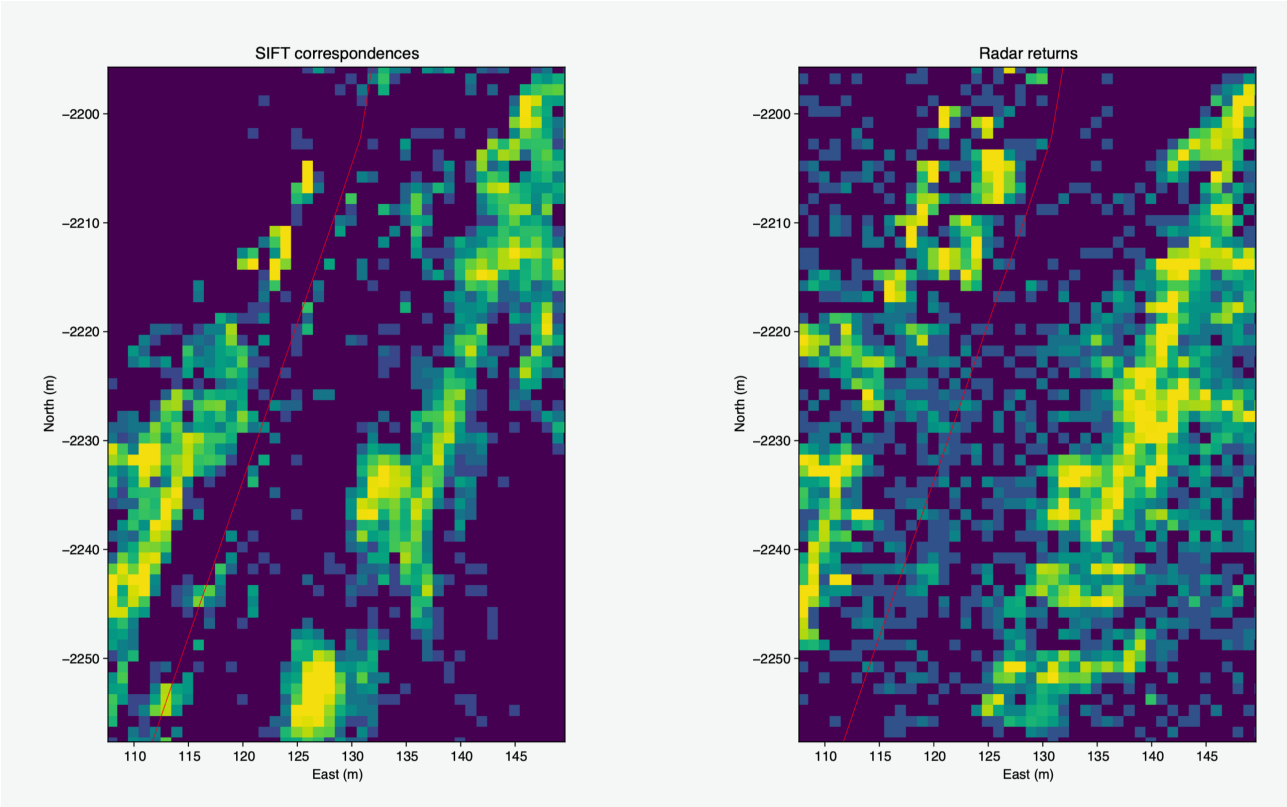

Fig. 7: Close-up of the south-southwesterly leg of the trip down Red River St., once again gridded at $1\,\mathrm{m}$. Both maps resolve the two parked cars on each side of the street.

Fig. 8: Camera view facing south-southwest on Red River St., corresponding to the leg shown in Fig. 7. The metal street sign with its lettering forms a strong cluster of features in both vision and radar, just to the west of the second car. The smooth concrete wall along the sidewalk forms a poor visual feature, but a strong radar feature, where it is visible to the Sensorium's line-of-sight between the parked cars.

These reasons motivate three heuristics: discarding stereo correspondences higher than $3.5\,\mathrm{m}$ from the road surface; discarding correspondences more than $120\,\mathrm{m}$ from the vehicle; and discarding correspondences "too close" to the road. This last is the most subtle, because the road is not flat, and system complexity would increase substantially if one were to require an accurate digital elevation model. As a stopgap, one may define a plane tangent to the road surface directly below the vehicle, and discard all stereo correspondences whose distance h from this plane is less than some function of the radial distance r from the vehicle to the stereo correspondence. In the present work, this lower bound was taken to be $h \geq 10\,\mathrm{cm} + r^2/1000\,\mathrm{m}$, representing the mask needed to avoid including road markings on a half-pipe-shaped road with a radius of curvature of $500\,\mathrm{m}$—that is, a curvature of 4° in $35\,\mathrm{m}$, a crude upper bound for this dataset.

Both the visual and radar maps suffer non-idealities around locations where the test vehicle came to a complete stop. The visual map contained a simple over-density of points collected from these vantages. The radar map, however, showed a number of concentric circles with large numbers of returns from locations where no object was present, or even locations beyond the vehicle's line-of-sight. These erroneous returns may be attributable to standing waves formed between the stationary radar tranceiver and its environment, analogous to the speckle pattern formed by a laser beam on a diffuse reflector. When the vehicle was in motion, these erroneous returns were rapidly eliminated by the radar units' internal DSPs. The solution for both maps was to discard map features collected at a velocity below $0.5\,\mathrm{m/s}$. For the radar, an additional filter excluding returns beyond a range of $50\,\mathrm{m}$ limited the sensitivity of the system to angular calibration errors.

C. Scan Stacking Time

One cannot form a full radar map without prior knowledge of the vehicle's pose trajectory to perform local-to-global coordinate transformations, and so the vehicle cannot simply compare the visual and radar maps. However, with the aid of inertial sensors and odometry, the vehicle can form an accurate estimate of a short segment—perhaps five seconds—of its pose trajectory, measured relative to its current position. Using this trajectory segment, the vehicle can perform a *speculative* local-to-global transformation on radar sensor data from the same $5\,\mathrm{s}$ interval of time to form a "map fragment". The $5\,\mathrm{s}$ interval is the *scan stacking time* of the system, since it accumulates, or stacks, many individual radar scans to form a radar map fragment.

(The stacking time is perhaps a less critical parameter than it may at first appear. Shorter stacking times may be compensated by providing more frequent, less confident position estimates to the downstream tracking filter.)

The radar map fragment is correct in its scale, orientation, distances, and angles, but suffers an overall (two-dimensional) position ambiguity. The vehicle must compare each radar map fragment to the visual map to constrain its position.

D. Map-Matching

To compare a visual map to a radar map fragment, it is possible to conceive of many very convoluted approaches: recognizing materials based on hyperspectral imagery and predicting radar returns; matching geometric primitives such as walls or fences in the visual data and searching for specular reflection points; skeletonizing the visual map via Delaunay triangulation and searching for overlap with a similar skeleton of the radar map fragment; and on, and on. The reason for considering convoluted approaches is the lack of a reason to expect a direct correspondence between points which are visually distinctive, triggering SIFT processing, and points which are highly retro-reflective at $77\,\mathrm{GHz}$, triggering radar tracking. Many visually distinctive features in an urban setting are painted or printed onto walls, signs, and roads. Visually contrasting shades of ink or paint are unlikely to generate substantial radar contrast.

Nevertheless, before trying convoluted solutions, one ought to verify that simple solutions are not adequate. In this case, however, a naïve heuristic works better than it has any right to do. This heuristic is point-cloud density overlap. One discretizes northing and easting, forming a pixel grid. One assigns each three-dimensional feature to a pixel, and counts features in each pixel to compute a two-dimensional number density map. Visual features are counted in one density map, and radar features in another. The visual features have known global coordinates, and so their density map may be said to be parameterized by absolute easting and northing. The radar features only have known relative coordinates, and so their density map is parameterized by easting and northing relative to the location of the vehicle. Given these two pixelized density maps, one computes the cross-correlation as the alignment metric.

E. Justification

Why might a visual map be useful for radar localization? One possibility is that cities simply contain a great deal of radar reflective material, and that a high density of visual features is a good predictor of the presence of a large amount of material, some portion of which is likely to be reflective at radar frequencies. Another possibility is that dense visual features indicate geometric complexity, increasing the number of potential diffractors to yield non-specular radar returns.

One may also offer some justification for the idea of histogram cross-correlation as a generic tool for localization. If the visual map consists of a set of points $\{\vec{g}_i\}$, and the radar map fragment consists of a set of points $\{\vec{h}_i\}$, then a likely hypothesis regarding the vehicle's inertial odometry error $\vec{x}$ ought to approximately satisfy the relation $\vec{g}_i = \vec{h}_j + \vec{x}$ for many pairs (i, j). In fact, this is exactly what the cross-correlation of histograms evaluates: when evaluated at an offset of $\vec{x}$ from the center of the correlogram, it gives the number of pairs (g_i, h_j) that, when rounded to the nearest bin, differ by exactly $\vec{x}$.

If one assumes that each of the events $e_{ij}(\vec{x})$—representing the possibility that a given pair of features (i, j) is consistent with hypothesis $\vec{x}$—is independent and equally likely, then

the log likelihood of the hypothesis $\vec{x}$ is proportional to the number of consistent pairs.

The use of binning has two key advantages: it side-steps both the association and cardinality problems of traditional point-cloud alignment. That is, there is no need to determine a particular one-to-one matching between visual and radar map features, and there is no need for the two maps to have equal numbers of features. Finding good matchings is a hard problem in its own right, even when both feature sets have a common nature. The trade-off is that the binning approach weights each bin in the cross-correlation according to the *product* of the bin occupancies in the visual and radar histograms, rather than according to the number of points in one map which can be matched to a point in the other map. This may be expected to bias the system towards regions of higher feature density.

In the end, the effectiveness of this naïve point-cloud density overlap heuristic is an empirical observation. Qualitatively, the left (visual) and right (radar) halves of Fig. 5 have much the same structure: large concentrations of features on walls, fences, and parked cars.

1) Gridding: The cross-correlation is computed on histograms binned with some finite resolution. What is an appropriate choice for the bin width?

Setting a very large bin width limits the fidelity with which the correlation peak is resolved, making discrimination more challenging. On the other hand, setting a very small bin width requires correlating larger arrays. If the bin width is denoted by δ, then the cost of computing the 2-D cross-correlation using fast Fourier transforms is $\mathcal{O}((1/\delta)^2 \log(1/\delta))$.

Each feature in either map may be viewed as a single noisy observation of the location of some feature on an unknowable "true map". The size of the standard deviation of this noisy observation defines a spatial scale, the noise scale, below which one is unable to make easy discriminations between location hypotheses. In other words, the likelihood function for the vehicle's position (for which the cross-correlation is an heuristic approximation) is smooth over distances comparable to the observation noise scale.

Consider what happens to a histogram of continuously-distributed data as the bin width goes to zero, holding the data constant: the number of features falling into a typical bin goes to zero, and the histogram becomes "sparse". If the cross-correlation is computed with histograms that are zero in most bins, then the cross-correlation will also be zero in most bins, and the few bins which are non-zero will typically have a value of one. This leads, in the limit, to zero discriminating power between good and poor hypotheses, which will all have a test statistic (cross correlation) of one.

Thus, cross-correlating ordinary histograms cannot be a good proxy for the likelihood function if the bin width is set too low. In that case, one should prefer a "splatting" approach, in which the contribution to the histogram of each feature is spread over multiple bins according to the likelihood function of the observation model (for instance, a Gaussian). One should certainly be concerned with the need for such a splatting approach if "shot noise" in the cross-correlation becomes great enough to obscure the correlation peak. This type of shot noise is just visible in Fig. 10 as a slight non-smoothness. In the present work, the number of features acquired in the map-making phase was great enough that the correlogram maintained adequate smoothness with $10\,\mathrm{cm}$ gridding.

2) Peak Fitting: Rather than reporting the location of the maximum value of the correlogram evaluated on a discrete grid, it is possible to obtain a continuous estimate of the vehicle's location from the discrete correlogram. One way to do this is by fitting a polynomial function to the neighborhood of the correlogram peak, and using this polynomial to interpolate the correlogram to non-integer indices.

The first step is to find the location of the correlogram's peak value. Since this work assumes that inertial odometry provides a robust prior distribution on the location of the vehicle, the search for the peak may be narrowed to a small window centered around the inertial estimate of the vehicle position (for instance, a $2.4\,\mathrm{m}$ square). Given this location, the vehicle then extracts a neighborhood of 3×3 bins around the peak, and carries out a least-squares regression for an interpolating quadratic of the form $z = a + bx + cy + dx^2 + ey^2 + fxy$. The peak is then declared to lie at the solution of $\partial z/\partial x = \partial z/\partial y = 0$.

The quadratic regression provides the system with access to a number of additional statistics—including the correlation value at the peak and the eigenvalues of the Hessian matrix at the peak—that could be exploited to improve robustness and more accurately weight position estimates in a downstream tracking filter.

3) Bias Estimation: After evaluating the cross-correlation and peak fitting procedure on each of 10,000 five-second epochs sampled (with overlap) from the Sensorium dataset, it became clear that a systematic bias with a vector magnitude of $40\,\mathrm{cm}$ was present when residuals were plotted in vehicle coordinates. This systematic bias is most likely attributable to angular dimensional inaccuracy in a radar mounting bracket that was incorrectly specified and required manual rework. This source of error could be mitigated by repeating the analysis using each of the three radars separately, one by one, and computing the sensitivity of the correlogram to small angular deflections in the sensors.

IV. Results

The procedure of Sec.III yields a bias-corrected location residual for each sampled epoch. If one fuses cross-modal radar localization using a visual map plus inertial odometry, then these residuals may be interpreted as the difference between the true location of the vehicle on the one hand, and on the other the location estimates supplied to the tracking filter.

A. Root-mean-square Error

The complementary cumulative distribution of these residuals (i.e. the fraction of epochs exceeding any given level of error) is plotted on a log scale in Fig. 9. In 95% of epochs, the error magnitude was no greater than $61\,\mathrm{cm}$.

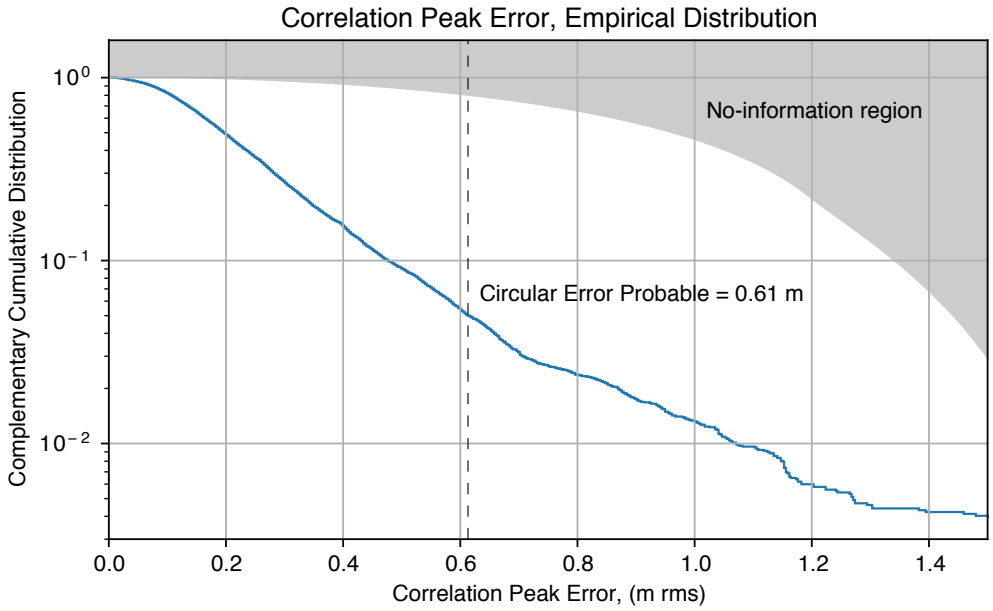

Fig. 9: The complementary cumulative distribution (also known as a survival function) indicates how often (that is, in what fraction of sampled $5\,\mathrm{s}$ epochs) the localization procedure in the text was found to exceed a given level of RMS positioning error. The logarithmic vertical scale makes the tails of the distribution, corresponding to outliers that may cause tracking errors, more visible. Above and to the right, the gray region of no information shows what the complementary cumulative distribution would look like if the cross-correlation localization procedure was non-informative, i.e. uniformly random.

The "null hypothesis", that naïve point-cloud density overlap is no better than random guesswork, corresponds to the gray region in the figure: in this case, the location of the correlogram peak within the $2.4\,\mathrm{m}$-square search window is uniformly random. The complementary cumulative distribution for the given procedure lies far to the left of this "no-information" region, indicating that the correlation procedure is giving strong cross-modal localization performance.

B. Map Feature Distributions

Figs. 5 and 7 spotlight a region of three blocks by one block in a low-rise part of downtown Austin, Texas, visited along the test route. Each frame shows a bird's-eye view of the density of map feature points, either from visual (left) or from radar (right), gridded into $1\,\mathrm{m}$ bins. Dark cells contain no features; bright cells contain many features.

Fig. 7 is a blow-up of the south-southwesterly leg of the turn from the right of each frame in 5. Coordinates are East-North meters relative to a nearby GNSS reference station, and the red curve is the ground-truth vehicle trajectory. For comparison, a perspective view of the same region is shown in Fig. 6.

Clearly visible on both the visual and radar maps are buildings, landscaping features like ledges, fences, and parking lots, and parked cars. Bright streaks along the roads are likely attributable to other traffic.

C. Correlograms

The localization procedure calls for stacking $5\,\mathrm{s}$ of radar scans (roughly 100 scans and 13,000 individual returns) and binning at the $10\,\mathrm{cm}$ level to form a radar map fragment. Fig. 10 shows the result of cross-correlating such a map fragment against a visual map centered on the same point. Each point in the correlogram corresponds to one particular shift that might be applied in an attempt to bring the two maps into alignment. Brighter points indicate better alignment. There is a distinct bright correlation peak, in this case roughly $50\,\mathrm{cm}$ south of ground truth. Part of this correlation peak error is due to the systematic bias, which has not been corrected at this stage in the processing.

To demonstrate a broader range of scenarios, cross-correlations are shown for 16 randomly-sampled epochs from the dataset in Fig. 11. Within each $10\,\mathrm{m}$ frame, a distinct bright correlation peak is visible. A number of frames, notably the last four, show strong linear features corresponding to a one-dimensional translational ambiguity (or rather, near-ambiguity) between the visual map and the radar map fragment. Broad, bright areas like that in the 11^{th} frame indicate translational ambiguity in more than one dimension. In practice, the vehicle must be prepared to survive a number of such potentially non-informative epochs before its inertial odometry uncertainty grows large enough to disrupt tracking.

V. Conclusion

In this paper, accurate cross-modal localization of a radar-equipped vehicle using a map built from visual markers has been demonstrated to be feasible and reliable at the sub-meter level using a combination of careful heuristics and known algorithmic tools. Cross-modal localization represents an unprecedented new low-cost form of alternative navigation,

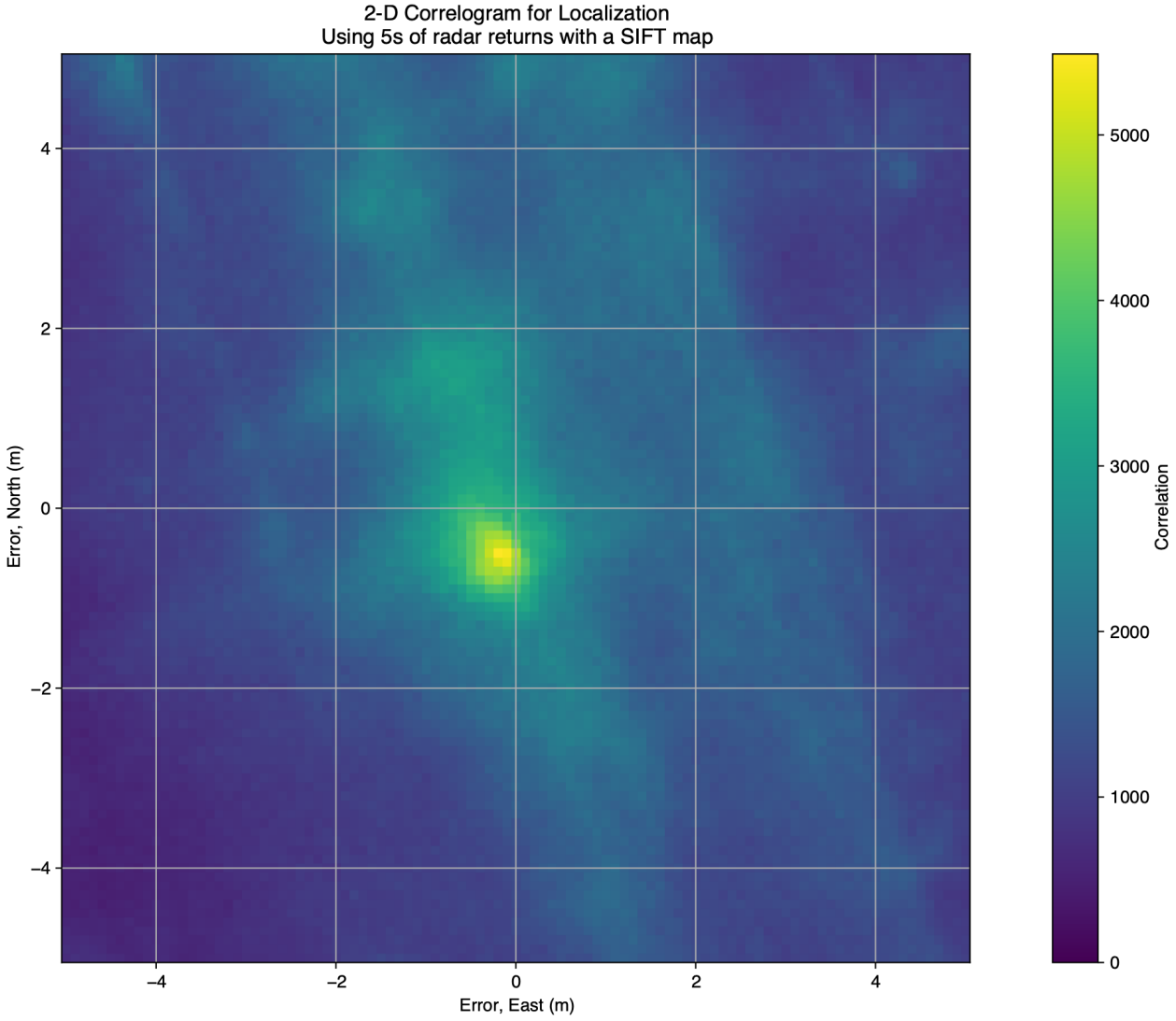

Fig. 10: Result of cross-correlating a visual map, containing at each point the density of visual features at that geographic location, with a radar map fragment, containing at each point the density of radar returns over a $5\,\mathrm{s}$ stacking interval at that vehicle-relative location. Bright spots indicate that many visual features align with many radar returns, conditioned on the hypothesis that the location of the vehicle is given by the coordinates of the bright spot. Subtle gradations indicate a partial ambiguity along the NNW-SSE and WSW-ENE axes. A strong, isolated correlation peak is visible just SSW of the center.

and one which will increase in relevance as sensors continue to evolve and autonomous vehicles face dynamic and difficult-to-map environments, poor weather conditions, and coordination challenges with dissimilar, heterogeneous vehicle platforms.

References

[1] A. Aldeghi, S. Carn, R. Escobar-Wolf, and G. Groppelli, "Volcano monitoring from space using high-cadence Planet cubesat images applied to Fuego volcano, Guatemala," *Remote Sensing*, vol. 11, no. 18, p. 2151, 2019.

[2] J. Liang, S. Shen, J. Gong, J. Liu, and J. Zhang, "Embedding user-generated content into oblique airborne photogrammetry-based 3d city model," *International Journal of Geographical Information Science*, vol. 31, no. 1, pp. 1–16, 2017.

[3] H. Hile, R. Vedantham, G. Cuellar, A. Liu, N. Gelfand, R. Grzeszczuk, and G. Borriello, "Landmark-based pedestrian navigation from collections of geotagged photos," in *Proceedings of the 7th international conference on mobile and ubiquitous multimedia*. ACM, 2008, pp. 145–152.

[4] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, 2018.

[5] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[6] T. Mouats, N. Aouf, A. D. Sappa, C. Aguilera, and R. Toledo, "Multispectral stereo odometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1210–1224, 2014.

[7] M. Magnabosco and T. P. Breckon, "Cross-spectral visual simultaneous localization and mapping (SLAM) with sensor handover," *Robotics and Autonomous Systems*, vol. 61, no. 2, pp. 195–208, 2013.

[8] M. A. Ali and D. A. Clausi, "Automatic registration of SAR and visible band remote sensing images," in *IEEE International Geoscience and Remote Sensing Symposium*, vol. 3. IEEE, 2002, pp. 1331–1333.

[9] T. D. Hong and R. A. Schowengerdt, "A robust technique for precise registration of radar and optical satellite images," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 5, pp. 585–593, 2005.

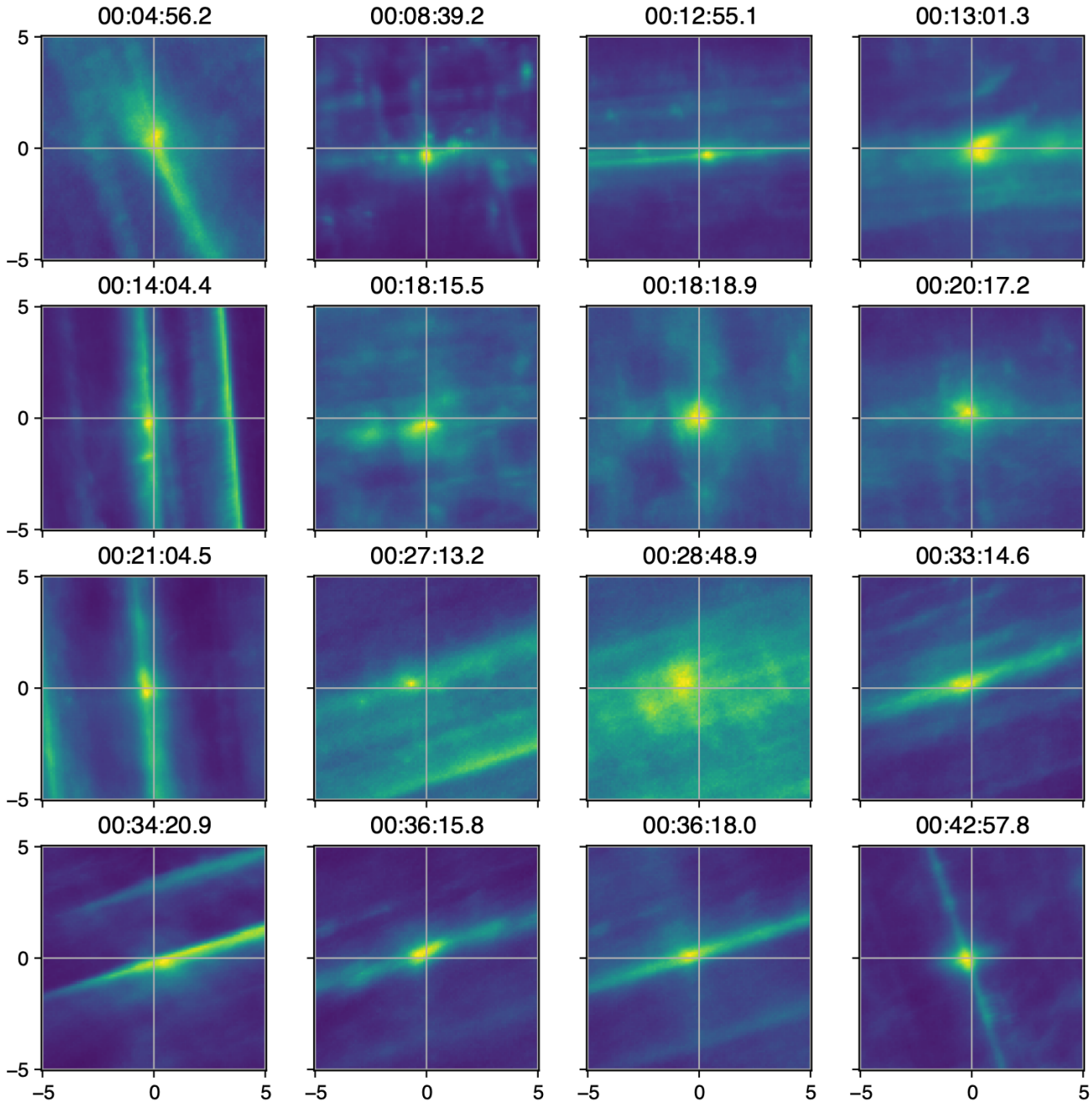

Fig. 11: 5 s correlograms computed for a random sampling of test data epochs when the vehicle was in motion. Partial ambiguities along the axes of the (mirror-imaged) Austin, Texas city grid are particularly visible.

[10] Y. Byun, J. Choi, and Y. Han, "An area-based image fusion scheme for the integration of SAR and optical satellite imagery," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 6, no. 5, pp. 2212–2220, 2013.

[11] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Learning cross-spectral similarity measures with deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[12] L. Narula, P. A. Iannucci, and T. E. Humphreys, "Automotive-radar-based 50-cm urban positioning," in *Proceedings of the IEEE/ION PLANSx Meeting*, St. Louis, MO, 2020.

[13] F. Schuster, M. Wörner, C. G. Keller, M. Haueis, and C. Curio, "Robust localization based on radar signal clustering," in *Intelligent Vehicles Symposium (IV), 2016 IEEE*. IEEE, 2016, pp. 839–844.

[14] F. Schuster, C. G. Keller, M. Rapp, M. Haueis, and C. Curio, "Landmark based radar SLAM using graph optimization," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016, pp. 2559–2564.

[15] J. Callmer, D. Törnqvist, F. Gustafsson, H. Svensson, and P. Carlbom, "Radar SLAM using visual features," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 71, 2011.

[16] S. H. Cen and P. Newman, "Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[17] ——, "Radar-only ego-motion estimation in difficult settings via graph

matching," *arXiv preprint arXiv:1904.11476*, 2019.
[18] R. Aldera, D. De Martini, M. Gadd, and P. Newman, "What could go wrong? introspective radar odometry in challenging environments," 2019.
[19] L. Narula, D. M. LaChapelle, M. J. Murrian, J. M. Wooten, T. E. Humphreys, J.-B. Lacambre, E. de Toldi, and G. Morvant, "TEX-CUP: The University of Texas Challenge for Urban Positioning," in *Proceedings of the IEEE/ION PLANSx Meeting*, St. Louis, MO, 2020.
[20] R. B. Ong, M. G. Petovello, and G. Lachapelle, "Assessment of GPS/GLONASS RTK under various operational conditions," in *Proc. ION GNSS*, 2009, pp. 3297–3308.
[21] T. Li, H. Zhang, Z. Gao, Q. Chen, and X. Niu, "High-accuracy positioning in urban environments using single-frequency multi-GNSS RTK/MEMS-IMU integration," *Remote Sensing*, vol. 10, no. 2, p. 205, 2018.
[22] T. E. Humphreys, L. Narula, and M. J. Murrian, "Deep urban unaided precise Global Navigation Satellite System vehicle positioning," *IEEE Intelligent Transportation Systems Magazine*, 2020.
[23] *Datasheet: ATLANS-C mobile mapping position and orientation solution*, iXblue, 2014.
[24] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1280–1286.
[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.