2013 IEEE International Conference on Robotics and Automation
Karlsruhe, Germany
May 6-10, 2013

An Approach-Path Independent Framework for Place Recognition and Mobile Robot Localization in Interior Hallways

Khalil M. Ahmad Yousef[1], Johnny Park[2] and Avinash C. Kak[3]

***Abstract*— Our work provides a fast approach-path-independent framework for the problem of place recognition and robot localization in indoor environments. The approach-path independence is achieved by using highly viewpoint-invariant 3D junction features extracted from stereo pairs of images; these are based on stereo reconstructions of the JUDOCA junctions extracted from the individual images of a stereo pair. The speed in place-recognition and robot-localization is achieved by using a novel cylindrical data structure — we refer to it as the Feature Cylinder — for representing either all of the 3D junction features found in a hallway system during the learning phase of the robot or a set of locale signatures derived from the data. For the case when all data is placed on the Feature Cylinder, we can use the 3D-POLY polynomial-time in a hypothesize-and-verify approach to place recognition. On the other hand, in the locale signature based approach, we can use the same data structure for constant-time place recognition.**

***Index Terms*— Viewpoint independent indoor recognition, robot localization, junction detection, hypothesis generation and verification**

I. INTRODUCTION

The problem of place recognition and robot localization has attracted much research attention lately for both outdoor and indoor environments. For example, for outdoors, there now exist several contributions that have demonstrated the use of satellite imagery and the photo collections available from the Internet for solving the place recognition problem. These approaches depend on either the geo-tagging of the images or the GPS information associated with the images. The matching of a sensed image (also referred to as a query image) with the images in a collection for the purpose of place recognition is frequently based on a point-cloud representation of the interest-point descriptors extracted from the images [11], [18], [9].

In contrast with the place recognition work in outdoor environments, place recognition research in indoor environments has received relatively little attention [7]. The indoor problems are more daunting because we do not have access to geo-tagged resource of images such as the satellite or Internet image databases for outdoors. Additional difficulties faced by indoor robots consist of the problems caused by illumination and other environment variations such as those caused by the additions/removals of wall hangings, etc.

The indoor-environment techniques used in the past for place recognition and robot localization run the gamut from beacons to ultrasonic sensors, from single-camera vision to multi-camera vision, from laser-based distance-to-a-point measurements to full-scale ladar imaging. The key idea in the techniques based on beacons, as with WiFi or radar [3], is to first construct a database of measured signal strengths at the different locations in an indoor environment and then to carry out place recognition by comparing the recorded signals with the signals stored in the database.

When using images for place recognition and robot localization, we see two different types of approach in the literature: (1) point cloud based; and (2) salient landmark based. The former typically create a database of SIFT or SURF interest points [2] to represent all of the interior space. Place recognition then consists of matching the point descriptors in a query image (which is the image recorded at the current location of the robot) with the descriptors stored in the database. The latter approach, the one based on landmarks, is similar in spirit to the former [16], in the sense that you still create a database of points and their descriptors, except that the points are now distinct and meaningful to humans visually.

There are two issues that are basic to the effectiveness of any image-based approach to place recognition and robot localization:

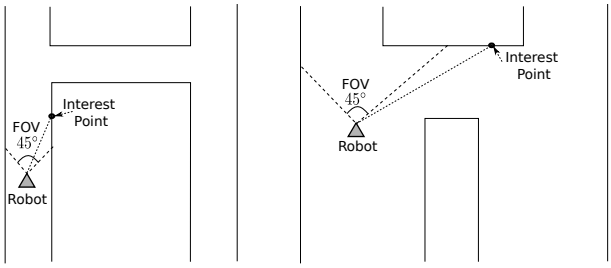
- The extent of approach-path invariance; and
- Whether or not the indexing strategy used in the global database of point clouds (or landmarks) allows for fast retrieval of the correct place in response to a query set of points (or landmarks).

We claim that the approach-path invariance made possible by SIFT or SURF like interest points, while probably sufficient for a system of narrow hallways, is unlikely to yield correct results for more complex interior space. The viewpoint invariance associated with SIFT and SURF like features usually extends to $\pm 30^\circ$ from the direction from which the image was recorded. Given a narrow system of hallways, as in Fig. 1(a), one may assume that the robot's camera will generally subtend a view angle of 45° on a wall and, with that view angle, a $\pm 30^\circ$ invariance of the point descriptors would be sufficient for the needed approach-path independence (although one could argue that if you needed place recognition independent of the direction of traversal in a hallway system of the type depicted, you would need a viewpoint invariance that is much larger than $\pm 30^\circ$ for the place recognition algorithms to work).

But now consider a more complex interior space, such as the one shown in Fig. 1(b). Now we have much wider hallways and the hallways meet in large halls.

As should be obvious from the geometrical construction,

[1,2,3] are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA https://engineering.purdue.edu/RVL/



(a) (b)

Fig. 1: (a) An example of a narrow system of hallways (b) An example of a wider system of hallways

especially if the reader keeps in the mind the fact that in wide hallways and large halls the robot is less likely to view the space with more or less the same orientation (as could happen in narrow hallways) on account of the maneuvering needed for collision avoidance, the viewpoint invariance made possible by the use of interest points like SIFT and SURF will not be sufficient for place recognition and robot localization.

With regard to the second important issue related to the use of image-based strategies for place recognition and robot localization, the approaches we have seen so far simply use a bag-of-words approach for creating the database of descriptors associated with the interest points (or the landmarks). The problem with the bag-of-words approach is that it does not ameliorate the exponential combinatorics of matching the interest points extracted from a query image with the candidate descriptors stored in the database.

In this paper, we address both of these fundamental issues. With regard to approach-path independence, we show that our 3D junction features based on the JUDOCA junctions [8] extracted from the images possess far greater invariance than several other popular interest points such as SIFT and SURF. (We refer to these features as 3D-JUDOCA for obvious reasons.) And with regard to the indexing of the global database for fast matching, we present a new data structure called the Feature Cylinder for representing either all of the 3D-JUDOCA features found in a hallway system during the learning phase of the robot or a set of locale signatures derived from the data. For the case when all data is placed on the Feature Cylinder, we use the polynomial-time 3D-POLY algorithm [4], which is guaranteed to return the correct location of the robot in low-order polynomial time. On the other hand, when the data structure is used to represent just the locale signatures, we can achieve constant-time place recognition and robot localization. The former approach works well for hallway systems that are as large as those found in typical institutional buildings. Whether or not it would scale up to hallways of arbitrary size and complexity is open to question. On the other hand, the signature based framework has greater potential in terms of scalability, but its robustness is yet to be fully explored.

The remaining sections are organized as follows: Section II presents an overview of the related work for the recognition and localization problem. Sec. III introduces the 3D junction features, their extractions from stereo images and discusses their properties and viewpoint-invariance evaluation. Section IV explains what we mean by locales and their associated signatures. Section V focuses on describing our novel Feature Cylinder data structure. Section VI describes our hypothesize-and-verify matching algorithms based on using the Feature Cylinder. Section VII presents our experimental results. Lastly, Section VIII draws conclusions from this study and presents some future work directions.

II. Related Work

Recent research in indoor and outdoor place recognition and robot localization is represented by the work reported in [17], [18], [11], [7] and [6]. Wu *et al*. [17] have proposed a technique based on viewpoint invariant patches (VIP) that are extracted from orthogonal projections of 3D textures obtained from dense 3D reconstruction of a scene using Structure from Motion (SfM). A key aspect of their work is that, using SIFT descriptors, each VIP feature uniquely defines a camera pose hypothesis vis-a-vis the 3D scene. Their matching algorithm is based on a hierarchical matching method called Hierarchical Efficient Hypothesis Testing (HEHT). HEHT is applied sequentially to prune out the matches based on first the scale, then the rotation, and lastly, the translation. All possible hypotheses are exhaustively tested to determine the final set of inlier VIP correspondences. In contrast to this work, the technique we present in this paper does not require *dense* 3D reconstruction to extract the features. Our features are instead obtained from sparse 3D reconstructions. Additionally, compared to HEHT, our matching approaches are based on fast algorithms — 3D-POLY when we place all of the feature data on the Feature Cylinder and signature-based matching when we use locale signatures — that should yield faster localization results even when complex environments are involved.

In [7] and [6], Elias and Elnahas propose a fast localization approach in indoor environments. Their work is based on using 2D JUDOCA features for localizing a user roaming inside a building wearing a camera-phone. An affine based correlation approach is used as their image matching algorithm. The problem with this approach is that their correlation based matching requires exhaustive search of all the features, which is extremely slow compared to the matching algorithms we employ in this paper. Additionally, we are using 3D junction features that are more robust to viewpoint changes compared to the 2D junction features used by these authors.

Another approach proposed by Wu *et al*. [18] employs a visual word based recognition scheme for image localization assuming unknown scales and rotations in satellite imagery. The visual words, each a SIFT descriptor, are indexed for more efficient retrieval in response to a query image. For expediting the retrieval process, they also use an inverted index in which the keys are the descriptors and the entries for each key consist of all the image identifiers that are known to contain that key. The unknown scale and rotation

are handled by comparing the query image with the database at multiple scales and rotations through a hypothesize and verify approach. Along the same lines is the work reported in [11]. A potential shortcoming of these approaches is that they do not provide performance guarantees with regard to the speed with which a match for a query image can be established if it is present in the database. Comparatively speaking, our matching algorithms come with low-order polynomial-time guarantees with regard to this performance measure.

III. 3D Junction Features (3D-JUDOCA)

In this section we introduce 3D-JUDOCA, the 3D junction features based on a stereo reconstruction of the 2D JUDOCA features extracted from stereo pairs of images. As we will show, 3D-JUDOCA features possess much greater viewpoint invariance (and therefore approach-path invariance) in comparison with other features.

A. Extracting 3D Junctions from Stereo Images

3D junction features are extracted from stereo pairs of images obtained from a calibrated stereo camera mounted on the robot.

The 3D-JUDOCA features are derived from the 2D JUDOCA junctions that are extracted from the individual images of a stereo pair. The algorithm for extracting 2D JUDOCA junction features is described in [8] and [5]. Basically, a 2D JUDOCA junction feature is defined by a triangle corresponding to the vertex where two edge fragments meet. The JUDOCA algorithm draws a circular mask of radius λ around the vertex and then finds the points of intersection of the two edge fragments with the circle. The point where the edge fragments emanating from the vertex meet the circle are referred to as the anchors, as shown in Fig. 2(a).

To form a 3D-JUDOCA feature from the 2D JUDOCA junctions, epipolar constraints are first applied to create a candidate list of junctions in the right image for any junction in the left image. Subsequently, the NCC (normalized cross-correlation) metric is used to prune this candidate list. This is followed by the use of the RANSAC algorithm to get rid of the outliers from the candidate list. These processing steps are very much along the lines of what is described in [10]. After finding the matching 2D junction in the right image for any given junction in the left image, stereo triangulation is applied to both the vertices and the anchors of a corresponding pair of junctions in order to create a truly 3D junction that we call a 3D-JUDOCA feature. Fig. 2(b) shows an orthonormal view of 3D-JUDOCA features derived from several 2D JUDOCA stereo correspondences. The normalization is done very similar to the method described in [17] with the help of an orthographic camera using the normal vector to the plane containing all the highlighted 3D-JUDOCA features. In Fig. 2(c), we show an example from an indoor hallways highlighting some of the extracted 3D-JUDOCA features.

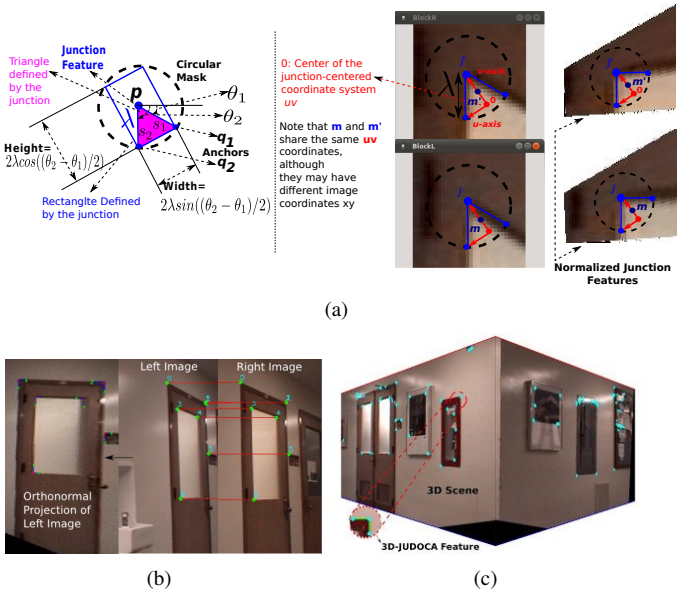


Fig. 2: (a) An example of JUDOCA junction extraction highlighting some of the junction descriptors (b) An orthonormal view of several 3D-JUDOCA features (c) An example of 3D-JUDOCA features in an indoor hallways

B. A Descriptor-based Representation of a 3D-JUDOCA Feature

Being a geometrical entity in 3D, a 3D-JUDOCA feature is viewpoint invariant (to 3D stereo-based measurements, obviously). This makes any place recognition based on 3D-JUDOCA robust to viewpoint changes as long as the robot follows the basic strategy of using its stereo cameras to extract such features for matching with the database collection of similar features.

Each 3D-JUDOCA feature contains the minimum set of points required to compute a 3D transformation that consists of a rotation matrix R and a translation vector T. What that implies is that a single 3D-JUDOCA feature correspondence between the data collected at a given position of the robot and the database uniquely defines the 3D transformation between the current position of the robot and the position of the robot corresponding to the matching features in the database. This property plays a critical role in constructing candidate hypotheses regarding the location of a given query image as will be seen later when presenting our hypothesize-and-verify matching algorithms in Sec. VI.

We associate a descriptor with each 3D-JUDOCA feature. This descriptor consists of the following: (i) the 2D/3D locations of the junction and the associated anchors (q_1, q_2); (ii) the orientations (θ_1, θ_2), the strengths (s_1, s_2) of the edges forming the junction, and the center of the junction in a local coordinate frame in relation to the junction vertex [8]; (iii) the width and the height of the minimum bounding rectangle for the junction, as shown in Fig. 2(a); (iv) the average color of the triangle formed at the junction; (v) pointers to the neighboring 3D-JUDOCA features; and (vi) the normal

vector to the plane confined to the 3D-JUDOCA feature triangle. Some of these attributes that go into a descriptor are highlighted in Fig. 2(a). Part of this descriptor is derived from the 2D junction textures in the original stereo image pair. Thus, for that part of the descriptor to be viewpoint invariant, the image textures associated with the 2D junctions are first normalized by an affine transformation. In fact, the 3D information — the normal vector to the plane in 3D that is formed by the 3D-JUDOCA feature triangle — is used to construct an orthographic camera and the normalization is carried out as discussed in [17]. The right side of Fig. 2(a) shows the normalization to the junction textures that are shown in the middle of the same sub-figure.

C. Viewpoint Invariance of 3D-JUDOCA

In this subsection, we provide an evaluation of the robustness of 3D-JUDOCA features against viewpoint changes. We compare 3D-JUDOCA with other well known features for representing interest points in images, these being BRISK [13], SIFT, SURF, 2D JUDOCA, Harris-Affine, Hessian-Affine, Intensity Based Region (IBR), Edge Based Region (EBR), and Maximally Stable Extremal Region (MSER). The metric we used is the repeatability metric [1]. Our evaluation setup is similar to the method used in [14] and [17]. Our test data is a sequence of stereo image pairs of indoor storage cabinets taken with increasing angles between the optical axis and the cabinets' normal. Each of the stereo pair of images has a known homography to the first stereo image, which was taken with image plane fronto-parallel to cabinets. Using this homography we extract a region of overlap between the first stereo image and each other stereo pairs. We extract features in this area of overlap and measure the repeatability using the following equation:

$$Repeatability = 100 \times \frac{N_i}{N_o} \tag{1}$$

where N_i is the number of inlier correspondences found in the overlapping region and N_o is the number of features in the overlapping region found in the fronto-parallel view. Fig. 3 shows this dataset (top row). Only the left images of the stereo pairs are shown. Also shown are the projection and overlapping regions of the images numbered (2-5) to the reference image 1.

Fig. 4 shows that the 3D-JUDOCA features generate a significantly larger repeatability over a wide range of angles compared to other feature detectors. This establishes the robustness of 3D-JUDOCA to viewpoint changes.

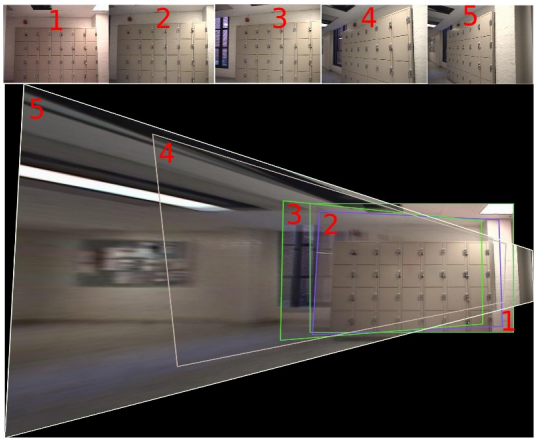


Fig. 3: The dataset used for evaluating the 3D-JUDOCA features against viewpoint changes

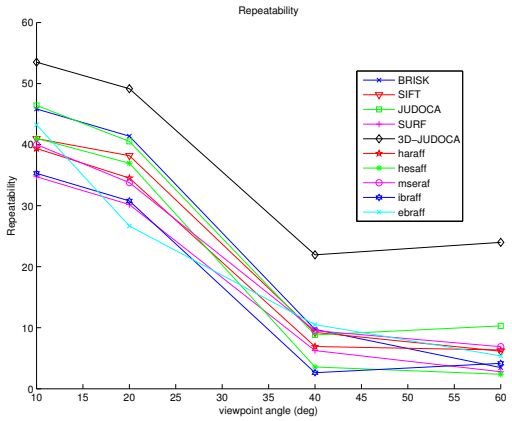


Fig. 4: Repeatability of different feature detectors measured across viewpoint angle

IV. The Notion of an Indoor Locale and its Associated Signatures

A locale in a system of an indoor hallways is defined as an indoor region rich in visual features (3D-JUDOCA) visible to the robot. Typically, in order to decide whether to characterize a point in a hallway as an identifiable locale, the robot situates itself in the middle of the hallway as it orients itself straight down the hallway. The robot subsequently analyzes the stereo images from that vantage point for its visual content in terms of the 3D-JUDOCA features extracted and their height distribution. Based on a user-defined threshold on the number of features found, a locale is identified. Fig. 5 shows an example of an identified locale in a hallway.

The robot associates with each locale its spatial location with respect to the world frame which corresponds to the location and the orientation of the robot at the beginning of the training phase. Each locale is characterized with the following two signatures: (1) a Differential Signature (DS) that is a height-based histogram of the differences in the 3D-JUDOCA feature counts collected from the left side and from the right side up to a certain threshold distance beyond the current position of the robot; and (2) A Radial Signature (RS) that is a radial histogram of the 3D-JUDOCA features at the current location of the robot. As mentioned previously, the robot tries to center itself the best it can between the hallway walls and orients itself so that it is looking straight down a hallway before constructing these signatures. Construction of a DS is illustrated in Fig. 6 for the locale depicted in Fig. 5. Fig. 7, on the other hand, depicts the construction of an RS for the same locale.

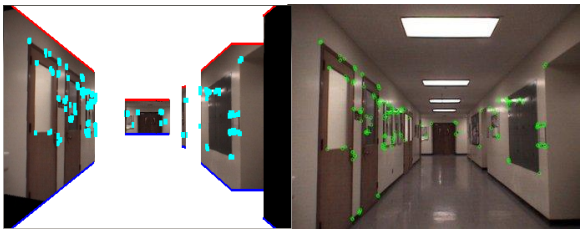


Fig. 5: An example of a locale in a hallway. The left image shows the 3D-JUDOCA points in a 3D reconstruction of the locale. One of the images used for the reconstruction is shown on the right

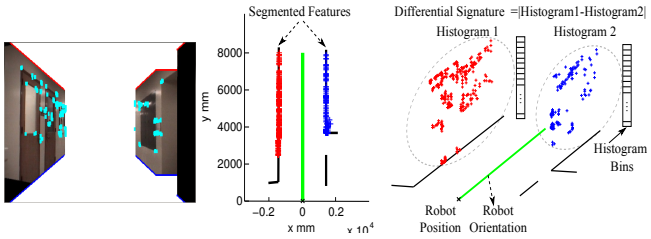


Fig. 6: Illustration of how a differential signature is computed

V. The Feature Cylinder Data Structure

We now describe the Feature Cylinder data structure that is used in two different modes in our work for the purpose of expediting place recognition. A Feature Cylinder (FC), inspired by the notion of a Feature Sphere in 3D-POLY [4], consists of an abstract cylinder that is tessellated both radially and vertically as shown in Fig. 8. The tessellation process depends on the user-defined sampling parameters θ, l that are associated respectively with the cylinder circumference and height sampling rates. Each cell of the cylinder holds a pointer to a unique directional attribute of a feature. We associate a floor-to-ceiling height with the FC. The FC is used for either representing the 3D-JUDOCA features or for representing the differential signatures of the locales during the learning phase of the robot. When 3D-JUDOCA features are mapped directly on to FC, we first associate a *principal direction* and *height* with each feature and then place pointer to the feature in the corresponding cell of FC.

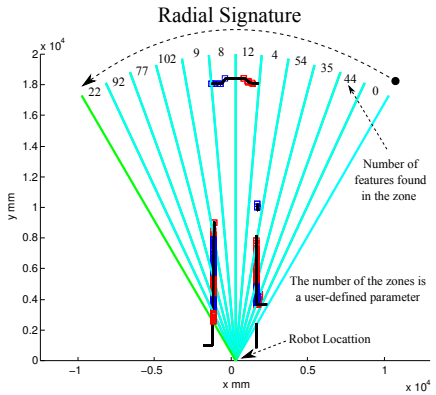


Fig. 7: Illustration of how a radial signature is computed

Note that the notion of a principal direction associated with a feature is exactly the same notion used in [4]. Basically, the principal direction of a point feature is defined as the normalized position vector of that feature with respect to the centered coordinate system of the FC. The principal direction of a feature gives us a fix on its directional orientation with respect to the other features that are likely to be seen in the same general portion of the interior space. The directional reference for this purpose is the orientation of the robot when it first starts to learn about the environment (world coordinate frame). On the other hand, when a locale signature is mapped to FC, each vertical facet of FC — a vertical facet consists of all the cells that are at the same angular orientation — stores the height distribution of a DS. That is, the differential count at a given height in a DS is stored in the corresponding cell of the FC along with a pointer to ID of the DS and its locale. Note that of the two signatures at each locale, only the DS is mapped on to FC. The other signature, RS, is stored in a bag for a direct comparison with a query RS at testing time.

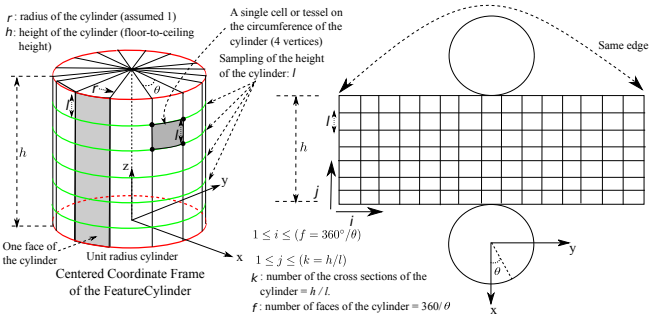


Fig. 8: The construction of the Feature Cylinder and its tessellation (Note: f needs to be at least 3)

A tessel mapping function (TMF) is defined to control the mapping mentioned above for either the 3D-JUDOCA features or the locales' differential signatures. So, for a given 3D-JUDOCA feature represented by its (x,y,z) coordinates (with respect to the centered coordinate frame of the FC), the TMF maps this feature to the nearest tessel on the cylinder using only two parameters i, j as follows (Obviously this mapping should depend on the uncertainty associated with the feature):

$$i = \begin{cases} \lfloor \phi/\theta \rfloor & if\ \phi \geq 0 \\ \lfloor (360^\circ + \phi)/\theta \rfloor & if\ \phi < 0 \end{cases}, j = \lfloor z/l \rfloor \quad (2)$$

where $\phi = tan^{-1}(\Phi_2/\Phi_1)$ denotes the directional angle of the feature that is computed from its principle direction, which is given by $\Phi = \dfrac{\langle x,y,z \rangle}{\sqrt{x^2+y^2+z^2}} = \langle \Phi_1, \Phi_2, \Phi_3 \rangle$. Fig. 9 shows an example of the FC in indoor hallways and how the TMF maps one of the 3D-JUDOCA features to the nearest tessel on the cylinder surface.

To briefly mention the role played by FC in place recognition with the help of the two locale signatures DS and RS, as mentioned earlier, only the DS is mapped directly onto FC; RS is used only for locale hypothesis formation, as

explained in the next section. As before, the TMF maps the robot's location/orientation at (p_x, p_y, p_ϕ) to FC using Eq. 2, but now only the index i is needed since the index j spans all of the cells at the same angular orientation. In general, this mapping should also take into account the uncertainty associated with the robot orientation at a locale. Associated with a differential count in each cell is a pointer to a locale ID.

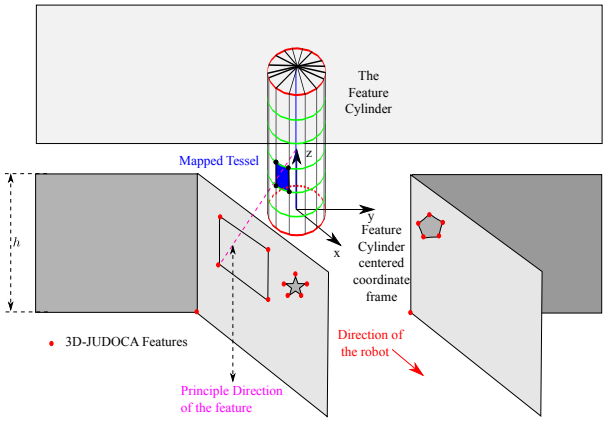


Fig. 9: An example of using the FC to represent the extracted 3D-JUDOCA features in a system of indoor hallways

VI. Matching Algorithms – Hypothesis Generation and Verification

Our framework for place recognition and robot localization may be viewed as a classification problem in which there are two main phases: training (learning) and testing (recognition). During the training phase, 3D-JUDOCA features are extracted from the stereo pairs of images recorded by the robot. We refer to the extracted 3D-JUDOCA features as the model features and to the computed locale signatures as the model signatures. In the testing phase, the goal is to find a match for a query stereo image using one of the two different modes: (1) an FC-based hypothesize-and-verify approach along the lines of the 3D-POLY algorithm; and (2) a hypothesize-and-verify approach in which the RS extracted from the query images is compared with the bag of all RS constructed during training for creating locale hypotheses that are subsequently subject to verification using DS stored on FC.

A. 3D-POLY Hypothesize-and-Verify Algorithm

In this approach, the 3D-JUDOCA features are used directly to generate a 3D transformation hypothesis regarding the current location of the robot with respect to the global frame used in the training phase. Although a single 3D-JUDOCA feature correspondence between the model and the data features is sufficient to generate this hypothesis, as mentioned in Sec. III-B, more robust hypotheses can be generated by matching neighboring groups of data features with groups of model features. Toward that end, we associated with each 3D-JUDOCA feature a set of k-nearest neighbors. The training phase creates a bag of all such $k+1$ feature groupings. The matching between two groups of $k+1$ of features, one from the query data and the other from the training data, is based on scanning the bag and establishing a similarity between the query grouping and a bag grouping on the basis of the 3D-JUDOCA descriptor values presented in Sec. III-B. A successful match leads to the calculation of a 3D transformation hypothesis regarding the current location of the robot. Subsequently, verification of a given transformation hypothesis is carried out by applying the transformation in question to all the other feature groupings in the query data and establishing their presence on FC in low-order polynomial time according to the 3D-POLY algorithm. Basically, the verification can be thought of as placing all of the test data on a test FC, applying the hypothesized transformation to the test FC, and checking for congruences between the test FC and the FC on which all the training data resides.

Regarding the time complexity of the matching process as described above, the overall complexity is obviously directly related to the effort required for hypothesis generation and then for hypothesis verification. The worst-case time complexity for hypothesis generation is obviously $O(n \times m)$, where m is the total number of feature groupings collected during the training time and n the number of feature groupings extracted from the data at a given location of the robot during testing time. In the worst case, we may have to compare every one of the feature groupings from the test data with all of the feature groupings in the model data. Regarding the complexity of verification, it is given by $O(n \times q) \approx O(n)$ where q is the largest number of features placed in a cell of FC at training time and where we have assumed that $q << n$ in general. Combining the hypothesis generation and verification complexities, we get $O(m \times n^2)$ for the overall time complexity of this approach to localization.

B. Signature-Based Hypothesize-and-Verify Algorithm

In the locale signature based approach, a locale hypothesis is formed by using the radial signature (RS) collected from the current position of the robot during test time with all of the radial signatures collected during training time. The test RS is compared with each of the training RSs using the Earth Movers Distance (EMD) metric [15]. The set of hypotheses generated in this manner are evaluated for verification in the order of how strongly they matched the test RS. If n_l is the number of locales, the complexity of the hypothesis generation stage is $O(n_l)$ since each locale is characterized by a single RS. The verification complexity for any given hypothesis is $O(n_h)$ where n_h is the vertical divisions of the FC. Recall, a hypothesis generated by an RS gives us a height-based DS histogram which must be matched with the histograms stored in the vertical facets of the FC. The orientation at which this match takes place must correspond to the transformation hypothesis created by the RS. The computing required for verification is merely the cell-by-cell comparison of the two height-based histograms.

VII. Experimental Results

This section presents experimental support for our matching algorithms for place recognition and robot localization in indoor hallways. After the training phase is over, the matching algorithms we have presented work in real time as the robot asked to localize itself when taken to a random place in the same environment. Since it would be difficult to include such demonstrations in a paper, we will base the experimental results in this section on a database of 1000 pairs of stereo images recorded by our robot with a sampling interval of 2.5m in the hallways of Purdue's MSEE building.[1] Each image in the database has a resolution of 640 x 480 pixels. The average number of 3D-JUDOCA features detected per stereo image pair was around 97. To help the reader visualize the nature of the interior space used for experimental validation, Fig. 10(c) shows a 3D map of the interior space using the framework presented in [12]. This 3D map was built using the same 1000 stereo images that we used for the experimental evaluation we report in this section.

For the hypothesize-and-verify experiments reported here, the tessellation parameters for the FC are: $\theta = 1^\circ, l = 100mm, h = 2700mm$. The training time takes roughly 568 seconds on a 2.67 GHz PC class machine. The experimental evaluation consisted of recording additional 100 pairs of stereo images at known locations and orientations of the robot and then testing whether the robot could figure out those locations and orientations using the matching algorithms presented in this paper. These 100 test stereo images were recorded at different times of the day (in order to allow for different ambient illumination) and at locales of what appeared to be of different visual complexity.

Fig. 10(a) shows an example of one of the query images (only the left image of the stereo pair is shown). Fig. 10(b) shows the position and the orientation of the robot as calculated by both the matching algorithms presented in this paper — the position and the orientation of the robot is illustrated by a reconstruction of the locale using the same framework that is presented in [12]. The recognition processing time for the query image that is shown in Fig. 10(a) was 1.35 secs for the 3D-POLY based matching framework and 0.98 secs for the RS/DS locale signatures based matching framework. Table I shows the average localization error and average processing time for all the 100 stereo images in the test dataset with the 3D-POLY based matching algorithm.

TABLE I: The average localization error and average processing time for the 100 test images using the 3D-POLY based matching algorithm

Average Localization Error		Average Processing Time (*sec*)
Position (*cm*)	Heading (*deg*)	
18	1.5	1.3

The localization and place recognition with the RS/DS locale signature approach requires a slightly different procedure — it requires that the interior be sampled at a higher rate than was the case for the previous evaluation. The reason for that is the fact that a signature match can only yield a position and orientation corresponding to one of the training signatures. Therefore, if the training images are recorded at too coarse a sampling interval, the robot may fail to match a test-time signature with any of the training signatures.[2] Therefore, our results with the locale signature are based on a training dataset of 333 stereo images recorded with a sampling interval of 0.25m in the RVL hallways of our building. For the purpose of visualization, we show in Fig. 11 a 3D map reconstruction obtained from these 333 images using the framework described in [12]. For each pair of stereo images in the training dataset, we recorded the position and the orientation of the robot. For testing, we recorded separately 50 new stereo pairs of images (along with the position and the orientation of the robot for each pair to serve as the ground-truth for evaluation). Table II shows the overall localization results obtained for the 50 test stereo images.

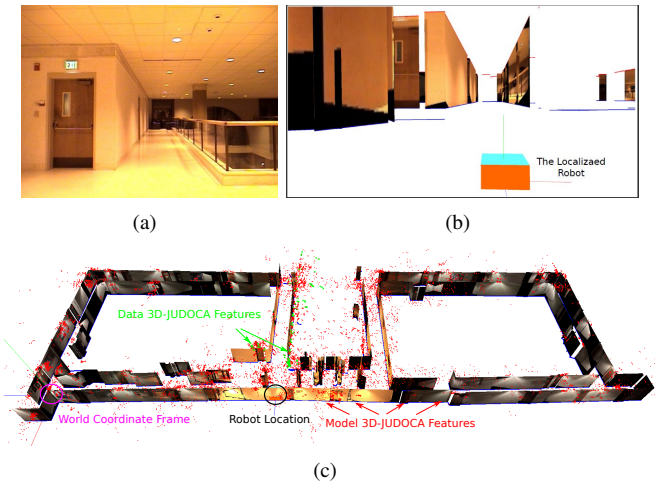


Fig. 10: The lower frame shows a 3D map of the indoor environment that was constructed from the 1000 stereo pairs of images used in the evaluation of the 3D-POLY algorithm for robot self localization. The upper two frames illustrate on the left a sample test image (only one image of the stereo pair is shown) and, on the right, the localization achieved for the test image.

TABLE II: The average localization error and average processing time for the 50 query images using the RS/DS locale signatures based matching framework

Average Localization Error		Average Processing Time (*sec*)
Position (*cm*)	Heading (*deg*)	
32	0.17	1.05

Finally, we want to demonstrate an example of the de-

[1]This database is being publicly available at `https://engineering.purdue.edu/RVL/Research/Research.html`

[2]This also implies that the localization error with the signature-based approach is lower-bounded by the sampling interval used at the training time for signature collection.

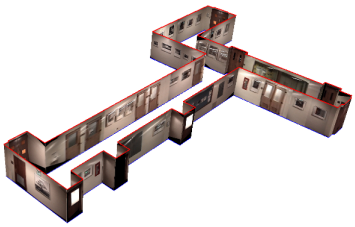
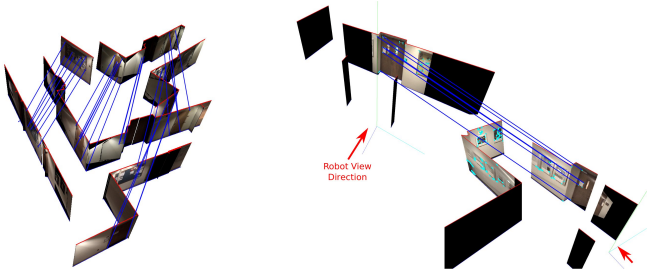


Fig. 11: Shown here is the 3D map reconstructed from the 333 pairs of training images used in the evaluation of the locale signature based approach to robot self localization.

gree of viewpoint invariance of our approaches to robot self localization. Fig. 12 demonstrates two experiments for two different scenes specifically selected to measure the extent to which the 3D-JUDOCA features and our matching approaches are robust against viewpoint changes. In each experiment, we used two image sequences of each scene with different viewpoints. The viewpoint change was 45° in the first experiment and 90° in the second one. As seen in Fig. 12, our approach using 3D-JUDOCA features was able to successfully recognize and match these views. Additional offsets were added to the scenes in order to have clear visualization of the matches. We compared such matching that achieved when replaced the 3D-JUDOCA feature by the 3D-SIFT features. 3D-SIFT descriptors were constructed from the more popular 2D SIFT feature descriptors in a manner similar to how we constructed 3D-JUDOCA features from 2D-JUDOCA features. A least-squares method with RANSAC was used to evaluate the 3D transformation between the point matches obtained with 3D-SIFT. In general, the matches achieved with 3D-SIFT had fewer inliers compared to 3D-JUDOCA features. For the two cases shown in Fig. 12, the results obtained with 3D-SIFT and with 3D-JUDOCA were comparable for the case on the left. However, 3D-SIFT failed for the case at the right because of the large change in the viewpoint.



(a) Scene1: 45° viewpoint change (b) Scene2: 90° viewpoint change

Fig. 12: Two experiments for recognizing the same scenes under different viewpoint changes showing that our approach is robust against viewpoint changes

VIII. Conclusions

This paper presented two different hypothesize-and-verify approaches for fast place recognition and self-localization by indoor mobile robots. The 3D-JUDOCA features we used give us the large viewpoint invariance we need and the feature cylinder data structure gives us fast verification of hypotheses regarding the location/orientation of the robot. Our evaluation demonstrates that the proposed matching approaches work well even under large viewpoint changes.

All of the work presented in this paper is based on the premise that a robot wants to carry out place recognition and self-localization with zero prior history. This is a worst case scenario. In practice, after a robot has recognized a place and localized itself, any subsequent attempts at doing the same would need to examine a smaller portion of the search space compared to the zero-history case. Our goal is to create a complete framework that allows prior history to be taken into account when a robot tries to figure out where it is in a complex indoor environment.

References

[1] Detectors evaluation. http://www.robots.ox.ac.uk/∼vgg/research/affine/.

[2] H. Andreasson, A. Treptow, and T. Duckett. Localization for mobile robots using panoramic vision, local features and particle filter. In *ICRA 2005*, pages 3348–3353. IEEE, 2005.

[3] P. Bahl and V. Padmanabhan. RADAR: An in-building rf-based user location and tracking system. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 775–784. IEEE, 2000.

[4] C. Chen and A. Kak. A robot vision system for recognizing 3d objects in low-order polynomial time. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6):1535–1563, 1989.

[5] R. Elias. Sparse view stereo matching. *Pattern Recognition Letters*, 28(13):1667–1678, 2007.

[6] R. Elias and A. Elnahas. An accurate indoor localization technique using image matching. In *Intelligent Environments, 2007. IE 07. 3rd IET International Conference on*, pages 376–382. IET, 2007.

[7] R. Elias and A. Elnahas. Fast localization in indoor environments. In *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*, pages 1–6. IEEE, 2009.

[8] R. Elias and R. Laganière. Judoca: Junction detection operator based on circumferential anchors. *IEEE Transactions on Image Processing*, 21(4):2109–2118, 2012.

[9] F. Fraundorfer, C. Wu, J. Frahm, and M. Pollefeys. Visual word based location recognition in 3d models using distance augmented weighting. In *Fourth International Symposium on 3D Data Processing, Visualization and Transmission*, volume 2, 2008.

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[11] A. Irschara, C. Zach, J. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR 2009*, pages 2599–2606. IEEE, 2009.

[12] H. Kwon, K. M. Yousef, and A. C. Kak. Building 3d visual maps of interior space with a new hierarchical sensor-fusion architecture. *Robotics and Autonomous Systems*. Paper under review.

[13] S. Leutenegger, M. Chli, and R. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision (ICCV), 2011*, pages 2548–2555. IEEE, 2011.

[14] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1):43–72, 2005.

[15] O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, 2009.

[16] A. Tapus and R. Siegwart. A cognitive modeling of space using fingerprints of places for mobile robot navigation. In *ICRA 2006*, pages 1188–1193. IEEE, 2006.

[17] C. Wu, B. Clipp, X. Li, J. Frahm, and M. Pollefeys. 3D Model Matching with Viewpoint-Invariant Patches (VIP). In *CVPR 2008*, pages 1–8. IEEE, 2008.

[18] C. Wu, F. Fraundorfer, J. Frahm, J. Snoeyink, and M. Pollefeys. Image localization in satellite imagery with feature-based indexing. *ISPRS*, 2008.