Localize Me Anywhere, Anytime: A Multi-task Point-Retrieval Approach

Guoyu Lu[1], Yan Yan[2], Li Ren[1], Jingkuan Song[2], Nicu Sebe[2], and Chandra Kambhamettu[1]

[1]VIMS Lab, Computer and Information Science Department, University of Delaware, USA
{luguoyu,renli,chandrak}@udel.edu
[2]MHUG Lab, Department of Computer Science, University of Trento, Italy
{yan.yan,jingkuan.song,niculae.sebe}@unitn.it

Abstract

Image-based localization is an essential complement to GPS localization. Current image-based localization methods are based on either 2D-to-3D or 3D-to-2D to find the correspondences, which ignore the real scene geometric attributes. The main contribution of our paper is that we use a 3D model reconstructed by a short video as the query to realize 3D-to-3D localization under a multi-task point retrieval framework. Firstly, the use of a 3D model as the query enables us to efficiently select location candidates. Furthermore, the reconstruction of 3D model exploits the correlation among different images, based on the fact that images captured from different views for SfM share information through matching features. By exploring shared information (matching features) across multiple related tasks (images of the same scene captured from different views), the visual feature's view-invariance property can be improved in order to get to a higher point retrieval accuracy. More specifically, we use multi-task point retrieval framework to explore the relationship between descriptors and the 3D points, which extracts the discriminant points for more accurate 3D-to-3D correspondences retrieval. We further apply multi-task learning (MTL) retrieval approach on thermal images to prove that our MTL retrieval framework also provides superior performance for the thermal domain. This application is exceptionally helpful to cope with the localization problem in an environment with limited or even no light sources.

1. Introduction

Image-based localization attempts to overcome the deficiencies of Global Positioning Satellite systems (GPS) and some other radio-signal based methods to provide accurate location information. Given 2D images as input, image-based localization techniques will either search through images in a database [19, 25] or directly use a 3D reference model [13, 22] to find a user's location.

Current research has proven that video frames captured by a mobile phone can build dense SfM reconstruction [33] with slightly more than 100 video frames. In this paper, rather than using an image as a query, we capture a short video and reconstruct a simple SfM model as the query. We then establish ***3D-to-3D*** correspondences. Existing localization frameworks build correspondences between images and 3D reconstruction models (***2D-3D-2D*** [18, 23]), (***2D-3D*** [22]), (***3D-2D*** [13, 11]), which ignore the geometry attributes of the query scene. It is expected that a 3D reconstructed scene will contain numerous geometrical properties to enable localization. We integrate outdoor environments, large indoor structures, and room environments into an image-based localization framework to provide easy localization without the limitations of place. Although all three environments are conceptually the same, each nevertheless has typical objects that can largely reduce the localization scope such as trees in outdoors, stairs in a large indoor structure, and tables in a room. Based on the spherical maps of segmented components in a query 3D model, we can quickly select the matching candidates in our dataset. We further perform an accurate localization based on local features (SIFT) of the SfM model among location candidates. State-of-the-art methods [13, 22] search correspondences through descriptors associated with the 3D SfM model by approximate nearest neighbor, which ignores the relationships between descriptors and points. For a given scene, points across different views match one another to reconstruct the 3D model. Those matching features are the shared information and can be explored to improve the query descriptor view-invariance property, which builds more accurate feature correspondences. Based on this idea, we propose the point-retrieval system through a novel multi-task learning (MTL) framework by learning the retrieval scheme for each discriminant point. Our MTL framework learning

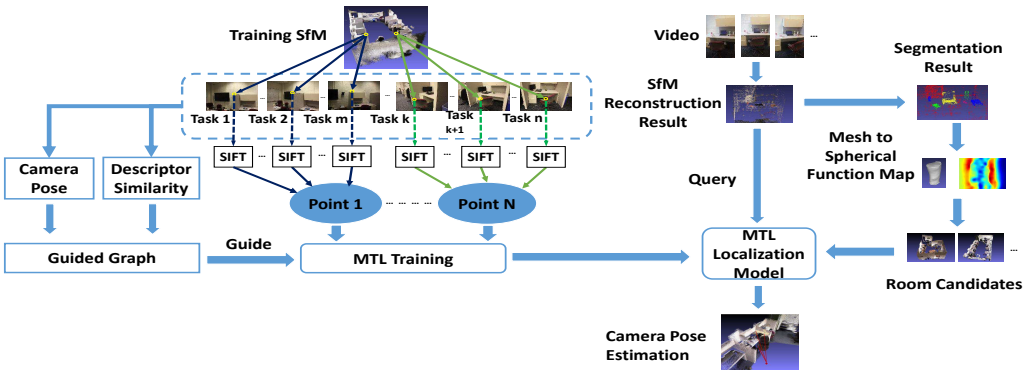

Figure 1. Our localization framework based on the query SfM model reconstructed by video frames. All points with more than 7 descriptors are utilized to learn the MTL point-retrieval system. Segmented objects of the query scene are processed with spherical map to find the location candidates. The final camera pose of the query scene is estimated.

scheme is guided by the shared information of matching features of different view images (multi-task), as well as each image's camera pose, which produces a more accurate point retrieval system. Additionally, the query SfM model provides us with the possibility of quickly identifying discriminant points to find correspondences. During the query process, distinctive points with seven or more descriptors are used as query points that, in turn, perform a majority vote among all the descriptors' correspondences for selecting a corresponding point in the training data (Fig. 1).

In the testing part, we also prove that our MTL approach is applicable to thermal imaging, which solves the localization problem when little light is available, e.g., at night or during power outage. Thermal images are learned together with the visible images using an MPEG-7 edge orientation histogram feature.

To summarize, the contributions of our paper are as follows: 1) We use the geometrical feature from a reconstructed SfM model to select location candidates, largely reducing the searching scope; 2) We use a 3D reconstructed scene as input and perform 3D-to-3D localization; this is in contrast to the state-of-the-art methods, which use 2D images as the query; 3) We learn the relationship between different points and propose a multi-task point-retrieval framework; 4) We show that our MTL model also has superior performance for thermal image on localization.

2. Related work

Image-based localization. Image-based localization, introduced by Robertson *et al.* [19], is widely applied to localization problems. 2D image localization is an image retrieval problem based on feature relevance [32] or semantic relevance [5]. Zamir *et al.* [34] proposed a distance computation method to constrain the local feature matching. Lu *et al.* [15] applied transfer learning from source color image domain to target thermal image domain in order to conduct scene classification. With the help of SfM reconstruction technique, Irschara *et al.* [11] approached localization by retrieving images containing the most descriptors matching the points in 3D space. Li *et al.* [13] proposed 3D-to-2D matching through mutual visibility information obtained between query images and database images. Sattler *et al.* [22] provided a framework that achieved a high image registration rate by directly matching descriptors extracted from 2D images to descriptors of a 3D model. Ventura *et al.* [28] proposed a keyframe-based monocular SLAM on mobile phone, where an external server estimates the keyframes pose. Lu *et al.* [14] increased memory efficiency and speed in localization through local feature processing. Middelberg *et al.* [18] improved this system by keeping a small relevant part to the scene in the mobile device and registering the keyframe to the global map based on a 2D-3D-2D method. Bergamo *et al.* [2] proposed use of random forest to train codebooks for local descriptors corresponding to the 3D points in a SfM reconstruction model. Following a similar idea, Donoser *et al.* [6] proposed an embedded random ferns method to classify query image descriptors into a corresponding point to build the 2D-to-3D correspondences used for localization. RGB-D data based on Kinect is also used in indoor localization [26], though such localization is limited to a small range (within 4 meters for Kinect), unsuitable for many environments. Our method makes use of 3D retrieval and 3D registration for the localization purpose. Our direct 3D-to-3D matching is based on the system learned from SfM reconstruction, which differs from most previous works [20, 36] that are based on exhaustive search for correspondences. Each 3D point in our system is associated with several descriptors. Through a majority vote among these several descriptors, we can build more accurate correspondences.

Multi-task learning. Multi-task learning jointly learns a problem with other related problems simultaneously, often leading to a better model for the main task as learners are able to use the commonality among the tasks. Evge-

niou *et al.* [8] proposed a natural extension of single-task SVM through a regularization framework.To capture the tasks dependencies a common approach is to constrain all the learned models to share a common set of features. This motivates the introduction of a group sparsity term, *i.e.* the ℓ_1/ℓ_2-norm regularizer as in [7]. Since not every task is related to all the others, the MTL algorithm based on a dirty model is proposed in [12] with the aim to identify irrelevant (outlier) tasks. To model complex task dependencies several clustered multi-task learning methods have been introduced [38]. In computer vision, MTL has been previously proposed for tracking [35], daily action recognition [30]. Yan *et al.* [31] used multi-task learning to classify head poses. However, to the best of our knowledge, MTL has never been applied to build point correspondences used in image-based localization.

Compared with previous work, our 3D-to-3D retrieval has two major benefits: 1) reduces the retrieval candidates (section 3); 2) 3D model exploits the relationships among images and points, and thus can improve the discriminative power of descriptors by using multi-task learning (section 4).

3. Location Candidates Selection

3.1. Scene Reconstruction and Segmentation

In this section, we introduce steps for selecting location candidates that largely reduce the search scope. State-of-the-art localization methods [18, 22, 13] capture images on a mobile device and send images to an external server to search feature correspondences due to the heavy computation burden. Because we are performing three-level localizations (outdoor, indoor, and room), searching through all descriptors would be expensive. By employing a 3D SfM reconstruction query model, we can better explore the 3D geometry properties of the scene in order to reduce the search scope. We segment the 3D objects through plane fitting and Euclidean clustering constrained by the point distance and intensity changes. We perform incremental reconstruction of a short video using CUDA, similar to the SfM reconstruction step in [18], followed by segmenting the reconstruction model. The whole process is conducted within 2 seconds (top level in Fig. 2).

3.2. Spherical Map

After object segmentation, we explore object geometry characteristics based on spherical function [24], using Poisson reconstruction applied to the point cloud to build the mesh. The spherical map calculates the distance from the object center to the surface, which can represent 3D shape property in a 2D view. When there is no intersection between the object center and the surface at a certain angle, the distance value r is set to 0 (for details, please refer to [24]). As in Fig. 2, the bottom images indicate the spherical maps of different kind objects (bottles and tree) which have large differences, preserving their shape properties.

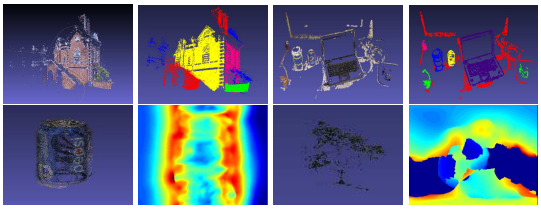

Figure 2. Segmentation samples for an outdoor training scene and a room query scene. Different colors represent various clusters (top). Spherical maps for corresponding point cloud (bottom)

We divide the x and y axes of the spherical map into $16 * 16$ grids, extracting the mean value and the mode value (a total of 512 dimensions per image) for each grid. Each query scene object will match the spherical maps of objects in the training scenes. Every object in the query scene searches the best matching in each location through spherical map. By summing up all spherical maps of objects in the query scene, every location gets a matching distance. The top five scenes with smallest distances to the query scene are then returned for exact camera pose estimation. In state-of-the-art image-based localization algorithms [11, 13, 22], all operations are on a point level whose search space is enormous in large scenes. By focusing the point level search onto object level, we can largely reduce the searching scope, an advantage for large scene localization tasks.

At times, the point cloud is not distributed uniformly on the object surface, resulting in small holes. Poisson reconstruction can fill the holes by interpolating points and building mesh throughout the surface. Spherical maps based on the reconstructed surfaces mainly represent object geometry properties. We want to match the spherical map in the query data to the objects' spherical map in the training dataset to find the matching candidates (Fig.3). For each building, we separately capture images and reconstruct the 3D model. We store all the buildings together. A building is correctly retrieved when the query scene is among the best matched 5 buildings. Existing localization methods search matching candidate through local features. Our method explores the typical objects in various scenes and uses the global feature for each segmented object in the scene. The approach is easy to implement and is robust to different scenes (indoor and outdoor).

From Fig.3 we observe that the video reconstructed objects are segmented and holes are filled by Poisson reconstruction. We demonstrate that all returned location candidates based on the spherical map from the object Poisson reconstruction result are labs or meeting rooms which share commonalities of interior settings (writing desks and swivel chairs in this case).

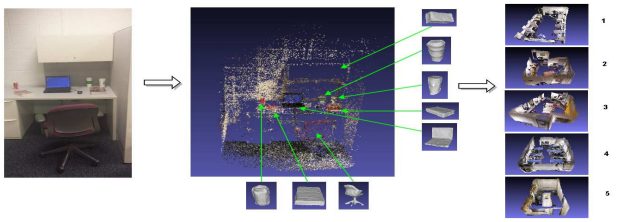

Figure 3. The query scene and retrieved location candidates. A video frame (left). SfM model is the query scene reconstructed from the short video, with object components from Poisson reconstruction of segmented objects from the SfM point cloud (middle). Best matching results based on the spherical maps ranked from high (1) to low (5) (right).

4. Multi-task Point Retrieval

Current research [7] shows that when the tasks exhibit commonalities it is beneficial to learn related tasks simultaneously instead of learning a single task separately. Thus, by taking the matching images that reconstruct 3D points as relevant tasks, the learning of one point's descriptors across images would potentially improve the retrieval accuracy. Based on this motivation, we propose to improve the 3D-to-3D point retrieval framework by applying a novel MTL regression model guided by prior knowledge. Our point retrieval framework primarily divides descriptors into a set of tasks according to the particular image that the descriptors are extracted from. Our MTL approach relies on a graph structure as prior knowledge, which models the similarity degree of descriptors and that of the camera poses between the related images. Note that in our framework, the descriptors of a 3D point are extracted from related images, which share similar structure since they describe the similar appearance of the same scene. On the other hand, the difference between images is caused by their various camera poses. Thus with the pre-defined guided graph, our MTL learning process carefully captures two facts in our case: 1) The patches described by the matching descriptors should be more similar if they construct the same 3D point; 2) The difference among these patches should be caused by changed camera poses while capturing the same scene. Based on the correspondences built by our multi-task point retrieval system, we can estimate the camera pose through 6-point Direct-linear-transform (DLT) under RANSAC (Details in section 5).

4.1. Camera Pose Constraint

Camera pose is the main cause of different view images of the same scene and is an essential constraint for searching correspondences, which is usually ignored. Even capturing the same scene, SIFT descriptors extracted from different viewpoints' images usually do not match. As current image-based localization systems query single images, no camera pose is associated with the query descriptor; however, as we are using an SfM model as the query, each video frame is associated with a camera pose that describes the relative distance and orientation to the scene. Thus, the descriptors extracted from the same image share the same camera pose. From the transformation matrix t and rotation matrix r returned by SfM, we calculate camera position (calculated by $-r' * t$) and orientation (calculated by $r' * [0,\ 0,\ -1]'$) [27], forming a 6-dimensional vector (6-DoF). We use camera pose in the MTL learning framework for guiding graph (section 4.2).

4.2. Point Correspondence Search via MTL

In this paper, we learn the multi-task retrieval framework based on images used for SfM reconstruction. We consider a set of R images for the reconstruction as R related tasks. For each image (task), a regression problem and G regression groups are considered. In our framework, we set all descriptors of a single 3D point as a group. Thus G is equal to the total number of 3D points.

We are given a training set $\mathcal{T}_t = \{(x_{t_n}, l_{t_n})\}_{n=1}^{N_t}$. t_n is the sample index in task t. N_t is the total samples number of task t. Considering each task $t = 1, 2, \ldots, R$, $x_{t_n} \in I\!R^d$ is d-dimensional feature vector, and $l_{t_n} \in \{1, 2, \ldots, G\}$ is the label indicating the group membership. Here we use SIFT as the feature vector for which $d = 128$ in our framework. We also introduce the camera pose vector $p \in I\!R^6$ described above to guide the graph of our model. Let $(\cdot)'$ denote the transpose operator.

Dealing with each task t, we have $\mathbf{x}_t = [x_{t_1}, ..., x_{t_{N_t}}]' \in I\!R^{N_t \times d}$, $\mathbf{y}_t = [y_{t_1}, ..., y_{t_{N_t}}]' \in I\!R^{N_t \times (GR)}$. Here $\mathbf{y}_t$ is the group indicator matrix. We define this matrix as below:

$$(\mathbf{y}_t)_{pq} = \begin{cases} \sqrt{N_t} - \sqrt{\frac{N_{t_q}}{N_t}} & if\ \ l_{t_p} = q - (t-1)G \\ -\sqrt{\frac{N_{t_q}}{N_t}} & otherwise \end{cases} \quad (1)$$

In the above equation, $(\cdot)_{pq}$ represents the p-th row and q-column of matrix $\mathbf{y}_t$. p is the sample index in task t. N_{t_q} is the sample size of q-th group in t-th task, either 1 or 0 in out setting. $N_t = \sum_{q=1}^{G} N_{t_q}$ is total training samples of all groups in t-th task. $(q-(t-1)G)$ represents the group index. $\mathbf{x}_t$ of R tasks are concatenated to be $\mathbf{X} = [\mathbf{x}'_1, \ldots, \mathbf{x}'_R]'$, where $\mathbf{X} \in I\!R^{N \times d}$. The same for $\mathbf{y}_t$ of R tasks, concatenated to be $\mathbf{Y} = [\mathbf{y}'_1, \ldots, \mathbf{y}'_R]', \mathbf{Y} \in I\!R^{N \times (GR)}$, $N = \sum_{t=1}^{R} N_t$. The goal of learning system is to learn a global weight matrix $\mathbf{W} = [\mathbf{w}'_1, \ldots, \mathbf{w}'_R]', \mathbf{W} \in I\!R^{d \times (GR)}$ that achieve the optimization of function $\min_W \frac{1}{2} \|XW - Y\|_F^2$. To make sure that the projection matrix W separates the data from different groups but preserves the similarity among the tasks, we propose to solve the following optimization problem:

$$\min_{\mathbf{W}} \frac{1}{2} \left\| (\mathbf{Y}\mathbf{Y}')^{-1/2} (\mathbf{Y} - \mathbf{X}\mathbf{W}) \right\|_F^2 + \lambda_1 \left\| \mathbf{M}\mathbf{W}' \right\|_F^2 + \lambda_2 \left\| \mathbf{W} \right\|_1 \quad (2)$$

where $\|\cdot\|_F$ represents the Frobenius norm and $\|\cdot\|_1$ denotes the L_1 norm. $(\mathbf{YY}')^{-1/2}$ compensates for each task's various amount of samples and normalizes the first component of the equation. $\mathbf{M}$ is our guided graph representing the prior knowledge, which tells us how one task can be utilized by the other tasks. $\mathbf{M}$ is an edge-vertex incident matrix, and $\mathbf{M} \in I\!R^{|\frac{R(R-1)}{2}| \times GR}$, where:

$$M_{e=(i,j),h} = \begin{cases} \gamma_{ij} & if \quad h = i \\ -\gamma_{ij} & if \quad h = j \\ 0 & otherwise. \end{cases} \tag{3}$$

In Eq. 3, h refers to the task index in all groups of the matrix, and $\gamma_{ij} = (\frac{\|p_i - p_j\|_2}{G'} \sum_{g=1}^{G'} \left\|x_{t_i,g} - x_{t_j,g}\right\|_2)^{-1}$, where G' is the total number of groups that contain samples from both task i and j, *i.e.* γ_{ij} is established by calculating the sum of the normalized Euclidean distance of descriptors between task i and j for all groups that have shared descriptors, then multiplying the distance of camera pose p_i and p_j. After the inversion of this sum, a larger γ_{ij} means the images share more similarity on specific descriptors and the camera pose. Practically, γ_{ij} is normalized into 0-1 in the regularization term.

Through solving the optimization problem, our framework can benefit from the following aspects: first, we correlate all tasks by means of the graph regularization terms, from which one task's information can also be applied to the related tasks; second, feature selection benefits from the sparsity utilized in the learning process, which reduces the effect of less discriminant features and emphasizes the influence of the most discriminant features; third, γ_{ij} enables us to embed the prior information of the shared features into our learning scheme.

Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [1] is adopted in our method to accelerate gradient computation while solving Eq. 2. FISTA is mainly resolving the proximal operator associated to L_1 norm, which is a non-smooth term. The smooth part of the objective function $f(\mathbf{W})$ and the non-smooth part $g(\mathbf{W})$ is denoted as below:

$$f(\mathbf{W}) = \frac{1}{2} \left\| (\mathbf{YY}')^{-1/2} (\mathbf{Y} - \mathbf{XW}) \right\|_F^2$$
$$g(\mathbf{W}) = \lambda_1 \left\| \mathbf{MW}' \right\|_F^2 + \lambda_2 \left\| \mathbf{W} \right\|_1$$

We summarize our algorithm for solving Eq. 2 as Algorithm 1. Based on the learned multi-task point retrieval framework, we use reconstructed SfM model as query. For each query point with seven or more descriptors, we search its correspondence among location candidates. According to the correspondence of each descriptor associated with the query point, we perform a majority vote to select the final corresponding point in the training set. As long as 12 inliers are found by RANSAC between the query model and the training model, we can reliably estimate the camera pose.

Algorithm 1: Accelerated Gradient Algorithm for Solving Eq. 2

INPUT: $\mathcal{T}_t = \{(x_{t_n}, y_{t_n})\}_{n=1}^{N_t}, \forall t = 1, \ldots, R$, λ_1, λ_2, $\mathbf{M}$
Initialize $\mathbf{W}_0$, $\alpha_0 = 1$.
LOOP:
$\alpha_k = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_{k-1}^2})$
$\hat{\mathbf{W}} = \mathbf{W}_k - \frac{2}{L_k} \mathbf{X}'(\mathbf{YY}')^{-1}(\mathbf{XW}_k - \mathbf{Y})$
Solving
$\mathbf{W}_{k+\frac{1}{2}} \leftarrow \min_{\mathbf{W}} \left\| \mathbf{W} - \hat{\mathbf{W}} \right\|_F^2 + \hat{\lambda}_1 \left\| \mathbf{MW}' \right\|_F^2 + \hat{\lambda}_2 \left\| \mathbf{W} \right\|_1$
based on Soft Thresholding [4].
$\mathbf{W}_{k+1} = (1 + \frac{\alpha_{k-1}-1}{\alpha_k}) \mathbf{W}_{k+\frac{1}{2}} - \frac{\alpha_{k-1}-1}{\alpha_k} \mathbf{W}_k$
Until Convergence
Output: $\mathbf{W}$

Compared with Yan et al.[31], our objective functions are different. They decompose learning weights in two. However, we learn a projection matrix W which optimally separates data from different classes, which is sparse (thus filters out noisy features) and has a structure reflecting image view similarity for our specific localization task. Moreover, optimization strategies are different.

5. Experiments

5.1. Dataset

Public datasets integrating outdoor, large indoor and room level 3D SfM buildings are unavailable. Although certain popular public localization datasets (Dubrovnik [13], Rome [13], Vienna [11] and Aachen [18]) use 2D images as a query, these do not fit our need for capturing a short video as a query. Mattausch et al. [17] scanned rooms using microCT to build 3D room models. As there are no features associated with each point, this dataset is not suitable for testing our method. Furthermore, there is no public dataset of thermal images for localization purposes. Therefore to evaluate our methods, we captured our own datasets–more specifically, we captured outdoor buildings, indoor building structures (such as lobbies and corridor), and rooms in the buildings. We captured 16 buildings (including their indoor environments) covering about 60000 m^2 (Aachen dataset [18] covers about 40000 m^2). Additionally, we captured 50 rooms of different purposes (e.g., lecture and meeting rooms, laboratories and lounges) for localization. We captured data in a wide range of domains to ensure the diversity in environments e.g. leisure place, working place, stores, study rooms, sports centers, to name a few (Table 1).

The testing videos are captured by iPhone 6 video camera by 60 fps with a resolution of 1920*1080. We randomly pick places covered by the training set and captured 50 videos each for outdoor, large indoor and room environments with camera translated, totaling 150 videos. The videos range from 120 to 180 frames, each $\sim$ 2- to 3-second

	Building #	Image #	Point #	SIFT descriptor #
Outdoor	16	18,116	5,760,383	29,921,389
Large indoor	16	12,957	2,331,864	11,608,121
Room	50	12,873	2,462,309	12,832,443

Table 1. Details of our dataset

in length, similar to the video frame reconstruction settings in Yu *et al.* [33]. The large public SfM dataset, Rome, Dubrovnik and Vienna dataset [13], are also used in our experiment to make comparison with other methods.

We also tested our MTL regression model on thermal image retrieval problem to prove that our MTL approach can be applied also to the thermal domain and that it provides superior performance. The use of thermal imaging based localization can help solve the problem of localization in the dark. Since current SfM techniques still cannot reconstruct a dense 3D model based on thermal images, we capture images instead of videos for experiments. For the thermal image test, we captured images of 15 landmark locations using a Xenics Gobi 640 GigE camera, an uncooled long wave infrared camera capable of imaging infrared wavelengths between 8-14 μm. Compared to normal RGBD sensor with infrared wavelength less than 1 μm (e.g., Kinect IR wavelength 0.83 μm), LWIR camera can detect objects' temperature. The resulting image is a thermal map of the environment with a resolution of 640x480. For each of landmark locations, we captured 15 visible and 15 thermal images separately. We make use of all 15 visible images and 5 thermal images of each location to learn the retrieval framework, and 10 thermal images of each location for testing.

5.2. Result for 3D MTL point retrieval localization

Our training sets cover the whole scene of each building. As the query scene is always part of the whole scene, the query segmented objects are always covered in the training sets. The location candidates' selection number reflects changes in the true positive building retrieval number. As location candidates' selection number increases, so does the number of the true positive retrieval building (Fig.4).

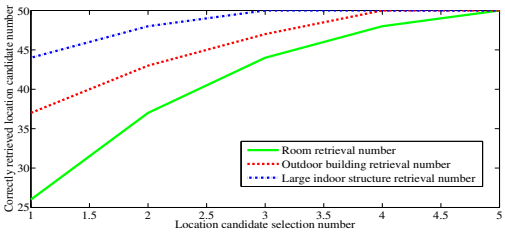

Figure 4. The true positive building retrieval number reflecting change in location candidate selection number

Thus when retrieving rank-1 location candidate through a spherical map, we achieve only 44 correct buildings out of 50 for the indoor large structure, 37 out of 50 for the outdoor buildings, and 26 out of 50 for the rooms. With an increase in the candidate number, the true positive candidate number also increases. When we retrieve the best 5 matching candidates, we can achieve 100% recall rate. Since large indoor environments commonly have unique structures and decorations, it is easier to retrieve the best location candidates. Rooms, however, are more difficult to retrieve, as most are similar in appearance and contain similar objects. To guarantee retrieval of correct candidates, we select the best 5 candidates for further extracting camera pose estimation. We test our processing on a GPU server with a GPU computing accelerator that has 4992 CUDA cores. On average, segmenting the query scene and computing the spherical map approximately takes 0.72 seconds and 0.26 seconds on average. Distance computing of the query spherical map with all spherical maps in our dataset is just less than 0.05 seconds, due to the simplicity of our feature. This step could help significantly reduce the point correspondence searching time, which we will provide later.

In our SfM reconstruction model each point on average has 5 descriptors, and about 30% of the points have at least 7 descriptors. We train our multi-task point retrieval system based on those 30% of the points because they are more discriminant as well as being essential for localization. We test our multi-task point retrieval system based on the query videos. For all points having at least seven descriptors in the SfM reconstruction model by query videos, we test the point retrieval accuracy (Fig. 5).

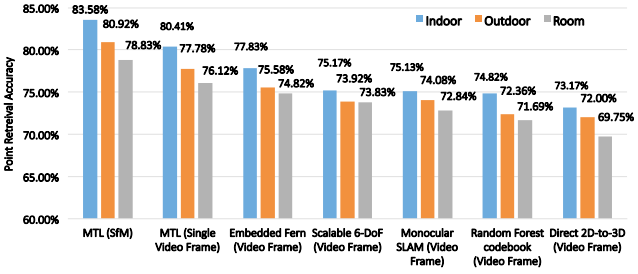

Figure 5. Comparison of average point retrieval accuracy of our method, which is the correct retrieved points among all the points with at least 7 descriptors. (MTL using SfM and single video frame with camera pose constraint) and other methods without camera pose constraint (Embedded Fern [6], Scalable 6-DoF [18], Monocular SLAM [28], Random Forest codebook[2], Direct 2D-to-3D [22]).

By using the SfM scene, we demonstrate that we can improve on point retrieval accuracy and scene registration number through multi-task point retrieval when compared with state-of-the-art methods (Fig.5 and Fig.6). Since large indoor environments have richer decoration and more distinct objects (e.g., display windows, statues), the large indoor structures have better performance in terms of point retrieval accuracy and image registration number. Li and Sattler et al. [13, 22] make the point that with 12 inliers

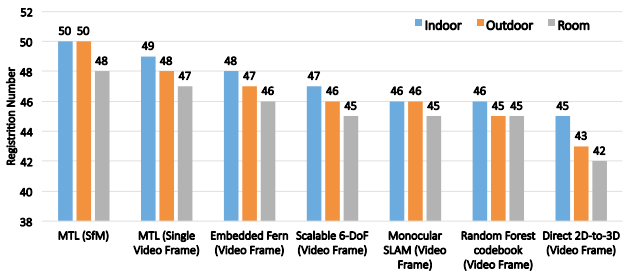

Figure 6. Comparison of total image registration number of our method (MTL using SfM and single video frame with camera pose constraint) and other methods without camera pose constraint (Embedded Fern [6], Scalable 6-DoF [18], Monocular SLAM [28], Random Forest codebook[2], Direct 2D-to-3D [22]).

between query image and the 3D reference SfM model satisfying the RANSAC transformation, camera pose is accurately estimated and the 2D image is registered. We test point-retrieval accuracy based on the top 24 points retrieving the best correlation score among the query SfM scene, as these points provide the most trusted correspondences. As long as 12 inliers are found among these 24 correspondences through RANSAC, we consider the query SfM scene to be registered. The inlier number fitting RANSAC is the correctly retrieved point number. We also compare retrieval speed based on each method, tested in Matlab. (Table 3).

Method	MTL	Embedded Fern [6]	Scalable 6-DoF [18]	Random Forest codebook [2]	Direct 2D-to-3D [22]
Elapsed Time	0.33(s)	1.32(s)	3.48(s)	0.93(s)	6.26(s)

Table 3. Time performance in localizing a scene (3D model for MTL, 2D image for others) by MTL, Embedded Fern [6], Scalable 6-DoF [18], Random Forest codebook[2], Direct 2D-to-3D [22].

Method	Multi-task retrieval	K-D tree search	SVM	Nearest Neighbor search
Elapsed time	0.54 (ms)	3.1 (ms)	5.9 (ms)	35 (ms)

Table 4. Average elapsed time to retrieve the 150 thermal image locations, in comparison with K-D tree search, SVM and Nearest Neighbor search.

As can be seen from Table 3, our method is several times faster than all other methods because retrieval is just a simple dot product between the descriptor and the weight, after learning the weight through MTL. This is a superior time performance, thus making the whole system practical for emergent situations.

We also evaluate the point retrieval accuracy of our approach on Dubrovnik [13], Rome [13], Aachen [18] and Vienna [11] datasets. Similar to the experiment setting of [6], we take 5-fold cross validation and each time we use 4 folds of the data for training, leaving 1 fold for testing. We test all 5 folds and take the average. Every SIFT descriptor is labeled with the corresponding 3D point and has a camera pose associated, which we can use to train our graph-guided multi-task point retrieval system. We show the point

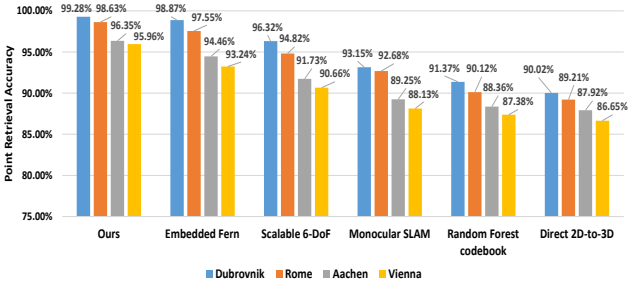

Figure 7. Comparison of average point retrieval accuracy of our MTL point retrieval system with other methods (Embedded Fern [6], Scalable 6-DoF [18], Monocular SLAM [28], Random Forest codebook[2], Direct 2D-to-3D [22]) on Dubrovnik [13], Rome [13], Aachen [18] and Vienna [11] public datasets.

retrieval accuracy comparison result on Fig. 7.

As seen from Fig. 7, our retrieval system achieves the highest point retrieval accuracy than other state-of-the-art methods. The public datasets have a sparser SfM reconstruction model our dataset, which makes the points less confusing and retrieval accuracy higher than our collected data.

5.3. Result for MTL thermal images localization

To further test MTL in the use of image localization, we propose a MTL thermal image localization framework, which addresses the localization problem in the dark. Thermal imaging has been used in tracking and face recognition [3], as well as on robotics [9, 29] for rescue tasks. As the scenes are composed of objects made of various materials, the temperature of objects also differs from each other. Thus, infrared radiation is also different, leading to images showing various objects' shape. Unlike visible images, thermal images are not dependent on lighting condition, thus providing a useful feature in performing the localization in a dark environment.

Using thermal images, however, makes extracting local features difficult, as it contains few color gradient changes. To achieve higher thermal image retrieval accuracy, we propose the thermal image retrieval system together with the visible images through multi-task retrieval. In performing retrieval tasks for both thermal and visible images, we need to find the common feature space. Because the most observable features in the thermal images are the object edges, we perform Difference of Gaussians (DoG) to thermal images in order to extract the edge information. Visible images, however, contain much rich texture information compared to thermal images. To make the edges extracted from visible images and thermal images as similar as possible, we apply the bilateral filter on the visible images to reduce the noise and preserve the strong edges. We have observed that applying DoG directly on visible images results in noisy edges, an effect absent in thermal imaging. After the bilateral filtration, edges of the visible image and edges extracted from

	Robot display window	Computer display window left	Basement left	Lamp post	Vending machine left	Vending machine right	Computer display window right	Vision lab	Stairs	Starbuck left	Starbuck right	Basement right	IR lab	Library outside	CS bulding outside
Ours	**9/10**	**7/10**	**10/10**	**10/10**	**10/10**	**10/10**	**8/10**	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**	**10/10**
SVM - thermal	6/10	6/10	7/10	8/10	8/10	7/10	5/10	**10/10**	9/10	7/10	8/10	9/10	9/10	**10/10**	8/10
SVM - visible	5/10	3/10	6/10	6/10	6/10	5/10	5/10	7/10	6/10	5/10	5/10	6/10	7/10	7/10	7/10
MTL-CASO [38]	7/10	6/10	8/10	9/10	9/10	9/10	6/10	**10/10**	9/10	9/10	9/10	9/10	**10/10**	**10/10**	8/10
MTL-Dirty [12]	6/10	6/10	7/10	9/10	8/10	7/10	7/10	9/10	9/10	7/10	8/10	7/10	9/10	**10/10**	9/10
MTL-FTC [37]	8/10	8/10	8/10	**10/10**	9/10	8/10	7/10	**10/10**	9/10	9/10	8/10	9/10	9/10	**10/10**	8/10
MTL-Robust [10]	8/10	**7/10**	8/10	9/10	9/10	9/10	7/10	**10/10**	9/10	9/10	9/10	9/10	**10/10**	9/10	8/10

Table 2. Location classification accuracy of the 15 locations. Black bold numbers represent the highest location classification accuracy for each location.

thermal images following DoG are quite similar.

As our infrared camera's field-of-view is much smaller than that of our visible image camera, we perform the thermal-visible camera calibration [21] for finding the corresponding part of the thermal image on the visible images.We divide the whole thermal image in $8 * 8$ grids (Fig.8). From the individual grids, we extract a MPEG-7 edge-orientation histogram [16], each feature of which is a 5 dimensional histogram. Thus the whole image is a 320 dimensional histogram feature. We extract the MPEG-7 edge-orientation histogram feature from the matching part of visible image and cast the thermal and the visible MPEG-7 edge-orientation histograms into our multi-task image retrieval system.

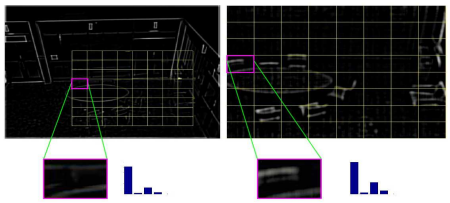

Figure 8. MPEG-7 edge orientation histogram comparison. The matching area to the thermal image in the color image (left). MPEG-7 edge-orientation histogram in the thermal image (right).

During the learning process, we treat both thermal and visible image retrieval as two tasks. We apply the graph guided MTL learning process similar to the description in section 4, but without consideration of camera pose information. During testing, the same MPEG-7 edge-orientation histogram is extracted from the thermal image. The location achieving the highest combination correlation score of thermal and RGB imaging tasks are considered location result. We use 10 thermal images from each location to test retrieval accuracy. The feature extraction process for visible images is similar to that of thermal images, with one main difference: before applying the DoG filter, images are first processed by bilateral filters to filter out the weak edges. In Table 2 we compare our MTL regression model with the single SVM and other state-of-the-art multi-task learning methods (MTL-CASO[38], MTL-Dirty[12], MTL-FTC[37], MTL-Robust[10]) for comparing the thermal image retrieval performance.

We have observed that our method outperforms other methods, achieving 100% accuracy in most locations. For "Robot display window", "Computer display window left" and "Computer display window right" locations, retrieval accuracy is lower, however, the infrared camera cannot properly observe the scene because the glass is reflective. We also test retrieval speed through a variety of methods. As discussed in the 3D SfM localization section, our retrieval system can achieve the correlation score by a dot product, much simpler than most methods. Based on the multi-task learning framework, we achieve more accurate and much faster localization result (Table 4). This shows that even when the thermal images are more difficult to obtain, the use of a visible image can help improve accuracy through learning together with the thermal images. Our method can also be applied to 2D-to-3D matching. However, we want to make use of camera pose to guide the MTL learning, which is obtained through 3D reconstruction. In testing, there are 3 main advantages of using a 3D model for query: 1) we can explore the 3D object geometry to choose location candidates through videos; 2) we can use multiple descriptors from the same point to find more trusted point-to-point correspondences; 3) camera pose estimation from 3D-to-3D is generally more accurate than 2D-to-3D.

6. Conclusion

We present a localization system integrating outdoor, large indoor, and room environments. Our system is based on the SfM model reconstructed from a short video employed as query to perform localization. Using SfM model, we extract spherical maps of segmented objects and select the best location candidates based on spherical maps matching. The use of 3D geometric attributes largely reduces the point level search scope. We further extract discriminant points from SfM model as queries. Concurrently we propose a novel camera pose graph guided multi-task learning method to explore the relationship among points and descriptors, which largely increases the point retrieval accuracy. Moreover, to overcome the problem of localization without sufficient light, we propose using thermal imaging for localization. Thermal image retrieval is learned together with visible images. Experiments show that our methods for both SfM query and thermal query outperform state-of-the-art methods by achieving higher accuracy while being more efficient.

Acknowledgement

This work is partly funded by Cooperative Agreement W911NF-11-2-0046 (ARO Proposal No. 59537-EL-PIR).

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[2] A. Bergamo, S. N. Sinha, and L. Torresani. Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification. In *CVPR*, 2013.

[3] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke. Pedestrian detection in infrared images. In *IVS*, pages 662–667, 2003.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[5] X. Chang, Y. Yang, E. Xing, and Y. Yu. Complex event detection using semantic saliency and nearly-isotonic svm. In *ICML*, pages 1348–1357, 2015.

[6] M. Donoser and D. Schmalstieg. Discriminative feature-to-point matching in image-based localization. In *CVPR*, pages 516–523, 2014.

[7] A. Evgeniou and M. Pontil. Multi-task feature learning. volume 19, page 41, 2007.

[8] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. In *Journal of Machine Learning Research*, pages 615–637, 2005.

[9] A. Garcia-Cerezo, A. Mandow, J. Martinez, J. Gómez-de Gabriel, J. Morales, A. Cruz, A. Reina, and J. Seron. Development of alacrane: A mobile robotic assistance for exploration and rescue missions. In *SSRR*, pages 1–6, 2007.

[10] P. Gong, J. Ye, and C. Zhang. Robust multi-task feature learning. In *SIGKDD*, pages 895–903, 2012.

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

[12] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar. A dirty model for multi-task learning. In *NIPS*, 2010.

[13] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *ECCV*, pages 791–804, 2010.

[14] G. Lu, N. Sebe, C. Xu, and C. Kambhamettu. Memory efficient large-scale image-based localization. *Multimedia Tools and Applications*, 74(2):479–503, 2015.

[15] G. Lu, Y. Yan, L. Ren, P. Saponaro, N. Sebe, and C. Kambhamettu. Where am i in the dark: Exploring active transfer learning on the use of indoor localization based on thermal imaging. *Neurocomputing*, 2015.

[16] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE CSVT*, 11(6):703–715, 2001.

[17] O. Mattausch, D. Panozzo, C. Mura, O. Sorkine-Hornung, and R. Pajarola. Object detection and classification from large-scale cluttered indoor scans. In *Computer Graphics Forum*, volume 33, pages 11–21, 2014.

[18] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt. Scalable 6-dof localization on mobile devices. In *ECCV*, pages 268–283, 2014.

[19] D. Robertson and R. Cipolla. An image-based system for urban navigation. In *BMVC*, pages 819–828, 2004.

[20] S. Salti, A. Petrelli, F. Tombari, and L. D. Stefano. On the affinity between 3D detectors and descriptors. In *3DIMPVT*, pages 424–431, 2012.

[21] P. Saponaro, S. Sorensen, S. Rhein, and C. Kambhamettu. Improving calibration of thermal stereo cameras using heated calibration board. 2015.

[22] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *ICCV*, pages 667–674, 2011.

[23] T. Sattler, B. Leibe, and L. Kobbelt. Improving image-based localization by active correspondence search. In *ECCV*, pages 752–765. 2012.

[24] D. Saupe and D. V. Vranić. *3D model retrieval with spherical harmonics and moments*. Springer, 2001.

[25] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, pages 1–7, june 2007.

[26] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, pages 2930–2937, 2013.

[27] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transition on Graphics*, 25(3):835–846, 2006.

[28] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg. Global localization from monocular slam on a mobile phone. *IEEE TVCG*, 20(4):531–539, 2014.

[29] S. Vidas, P. Moghadam, and M. Bosse. 3D thermal mapping of building interiors using an rgb-d and thermal camera. In *ICRA*, pages 2311–2318, 2013.

[30] Y. Yan, E. Ricci, G. Liu, and N. Sebe. Egocentric daily activity recognition via multitask clustering. *TIP*, 24(10):2984–2995, 2015.

[31] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*, pages 1177–1184, 2013.

[32] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *TPAMI*, 34(4):723–742, 2012.

[33] F. Yu and D. Gallup. 3D reconstruction from accidental motion3D reconstruction from accidental motion. In *CVPR*, pages 3986–3993, 2014.

[34] A. R. Zamir and M. Shah. Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs. *TPAMI*, 36(8):1546–1558, 2014.

[35] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via multi-task sparse learning. In *CVPR*, pages 2042–2049. IEEE, 2012.

[36] W. Zhao, D. Nister, and S. Hsu. Alignment of continuous video onto 3D point clouds. *TPAMI*, 27(8):1305–1318, 2005.

[37] W. Zhong and J. Kwok. Convex multitask learning with flexible task clusters. *ICML*, 2012.

[38] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *NIPS*, pages 702–710, 2011.