Knowing Where I Am: Exploiting Multi-Task Learning for Multi-View Indoor Image-based Localization

Guoyu Lu[1]
luguoyu@udel.edu
Yan Yan[2]
yan@disi.unitn.it
Nicu Sebe[2]
sebe@disi.unitn.it
Chandra Kambhamettu[1]
chandrak@udel.edu

[1] Video/Image Modeling and Synthesis Lab
University of Delaware

[2] Department of Information Engineering and Computer Science
University of Trento

Abstract

Indoor localization has attracted a large amount of applications in mobile and robotics area, especially in vast and sophisticated environments. Most indoor localization methods are based on cellular base stations and WiFi signals. Such methods require users to carry additional equipment. Localization accuracy is largely based on the beacon distribution. Image-based localization is mainly applied for outdoor environments to overcome the problem caused by weak GPS signals in large building areas. In this paper, we propose to localize images in indoor environments from multi-view settings. We use Structure-from-Motion to reconstruct the 3D environment of our indoor buildings to provide users a clear view of the whole building's indoor structure. Since the orientation information is also quite essential for indoor navigation, images are localized based on a multi-task learning method, which treats each view direction classification as a task. We perform image retrieval based on the trained multi-task classifiers. Thus the orientation of the image together with the location information is achieved. We assign the pose of the retrieved image to the query image calculated from SfM reconstruction with the use of bundle adjustment to refine the pose estimation.

1 Introduction

Indoor localization systems are applied to navigate people in large and complex indoor environment, such as shopping malls and museums where auxiliary information is necessary to help visitors localize themselves. In some urgent situation, like boarding an airplane and finding the emergency room in a hospital, providing accurate and timely location information is essential for travelers to catch planes and wounded people to get prompt medical assistance. The majority of the current indoor localization methods is based on WiFi and pre-deployed beacons. These methods usually require additional equipment to perform the localization task and the accuracy depends on the distribution of beacons and cellular stations in a large extent. Meanwhile, the WiFi and beacon based methods are lack of the orientation information, which is essential for navigation. GPS is quite successful in outdoor navigation. However, in indoor buildings with roofs and walls, weak GPS signals result in unreliable navigation information. Even in an outdoor large building area, GPS signals from satellite are attenuated by walls.

Image based localization has been mainly applied in outdoor environments in the past to overcome the weak GPS signal problem among large buildings. This method has been introduced to indoor environments in recent time. The main idea is to linearly search the image database consisting of indoor building images and find the best matched image. With the development of Structure-from-Motion (SfM) reconstruction techniques, 3D models are used for localization. Users can easily capture a 2D image with their mobile phone and register the 2D image with the 3D model to get the location information. In this process, features extracted from the 2D images are utilized to match against the features in the SfM 3D model; camera pose can be calculated based on the matching descriptors, providing users the location and orientation information. As the SfM technique does not require the cameras to be calibrated, the related images are easier to obtain, which makes the large scale reconstruction and 3D model based localization possible. Obtaining the location information is only half of the job. A map with the location information can help better perform the navigation task. With this purpose, a 3D model is suitable for localization purposes that facilitate users to understand the 3D building structure and schedule a visiting plan. However, a SfM model for localization usually contains millions of descriptors. Searching the correspondences within this scope is extremely time-consuming. Although k-d trees and visual word methods are applied to accelerate the corresponding search process, the reduced search scope may potentially add incorrect correspondences between 2D features and 3D points.

In this paper, we propose multi-view image based localization, which is a framework based on multi-task learning (MTL). MTL attempts to improve the performance of several specific tasks based on the shared common properties. Current research [1, 26] shows that it is beneficial to learn the tasks simultaneously instead of learning a single task separately when the tasks exhibit commonalities. During the learning process, the shared information across different tasks is extracted to simultaneously learn the multi-related tasks. With the purpose of guiding users with the location and orientation information, we divide the physical view direction into several regions. It is expected that images of the same object captured from different view directions contain similarities with regards to appearance, as well as differences due to the viewing perspectives.

Multi-view image based localization aims to learn the relationship of interior architecture appearance across different viewing directions. Ideally, the tasks within the same group should share the similar features while features extracted from tasks in different groups are expected to be different. Following this idea, images captured from the same direction are classified into one task, including same and different location images. The images captured from the same location across different camera angles are treated as the same group. We learn a multi-view regression model based on the correlated tasks scattering in different groups. During the testing phase, the query image retrieves the most relevant group for achieving the location information. Meanwhile, our MTL regression model assigns a direction to the query image based on multiple tasks for the orientation purpose. As we perform SfM reconstruction prior to the multiple view localization phases, every image used for SfM reconstruction is associated with a camera pose. The camera pose of the most corrected image within the same task, and the same group is assigned to the query image. We further

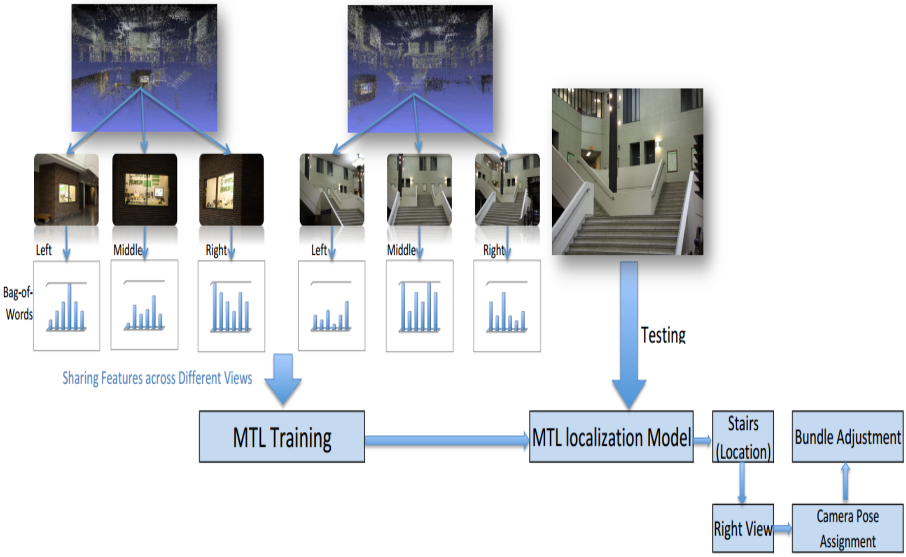


Figure 1: Multi-view image-based localization system

apply bundle adjustment to the query image to refine the assigned camera pose. In this way, we can take benefits from localization methods both based on 2D image and 3D model. The whole multi-view image-based localization framework is illustrated in Figure 1.

To summarize, the contributions of this paper are the following: (i) To our knowledge, this work is the first to address the problem of indoor image-based localization from multi-view settings. (ii) We are the first to propose the multi-task learning approach for multi-view indoor image-based localization. (iii) Both the orientation of the image and the location information can be obtained by exploiting multi-task learning.

2 Related work

Image-based localization is widely applied in localization problems. The biggest difference between GPS and image-based localization solution is that the latter can still be employed in weak GPS signal areas. Robertson et al. [16] utilize a database of the building facade views to calculate the query image pose provided by users. The images in the database are associated with a 3D coordinate system. Shao et al. [19] perform similar research on urban scene image retrieval problems. Zhang et al. [28] extract localization information in an urban area by directly searching for the closest image in an image database. Steinhoff et al. [22] make use of a vocabulary tree [15] to build their localization model to achieve the real-time pose estimation. Schindler et al. [18] select the vocabulary using the most distinctive features to enhance the image retrieval performance on a large street side image database. Xiao et al. [25] improve the object localization accuracy by combining geometric verification with bag-of-words methods. Kluckner et al. [9] propose an image-driven method for automatic extracting buildings and 3D modeling from large-scale aerial imagery based on a fast unsupervised segmentation technique on the use of super-pixels.

With the improvement of Structure-from-Motion (SfM) reconstruction techniques [4, 7, 21], 3D SfM model is utilized on the image-based localization to improve localization accuracy. Irschara et al. [8] retrieve images containing the most descriptors matching the 3D points and Li et al. [11] realize the 3D-to-2D matching through exploiting the mutual visibility information. Wendel et al. [24] introduce an algorithm of monocular visual localization for micro aerial vehicles by generating virtual views in 3D space. Methods [2, 5] based on the Simultaneous Localization and Mapping (SLAM) algorithm [10, 20] estimate camera pose based on the online-built 3D model. However, SLAM is limited to small scenes. Lu et al. [13, 14] propose to map the local feature to a low dimensional Hamming space to simplify the feature matching process and reduce the memory consumption in large scale reconstruction and localization. Sattler et al. [17] propose to directly match the descriptors extracted from 2D images to the descriptors of the 3D SfM model to improve the localization accuracy.

Multi-task learning is an approach which learns a problem jointly with other related problems simultaneously using a shared representation. Multi-task learning often leads to a better model for the main task, because it allows the learner to use the commonality among the tasks. Evgeniou et al. [6] propose a natural extension of single-task SVM through a regularization framework by decomposing the learned SVM parameter w_t into a common part w_0 and an individual part v_t among the tasks. However, these two parts (decision space and correcting space) are in the same feature space. Liang et al. [12] improve the model by assuming the decision space and correcting space are different which is more generic in multi-task settings. Argyrious et al. [1] propose a convex multi-task feature learner by imposing a trace norm regularization. Yan et al. [27] propose a Multi-Task Learning framework for head pose classification, whose Multi-Task classifier is learned based on a graph-guided approach. Based on a new proposed multi-class LDA, Yan et al. [26] also perform action recognition from multiple views.

3 Multi-view image-based localization

The purpose of image-based localization is to provide navigation information based on the query images provided by the users. Simply taking a landmark image in the surroundings and transmitting it to the localization system, users can get the location information, preferably with the orientation information.

Image-based localization is originally about finding the images in the database which contain maximally matching correspondences as the location reference. To achieve a higher accuracy, 3D modeling is applied in the localization framework, which allows camera pose estimation. Sattler et al. [17] propose to match a 2D image against a 3D reconstructed model. All the features extracted from the 2D image are required to find the correspondences with 3D points, which is achieved by searching the nearest neighbor through all the descriptors in the 3D space. Localization systems are usually deployed in large and complex buildings, whose SfM model contains millions of descriptors. The dominant local descriptor for searching correspondences in the localization task is SIFT feature due to its property of invariance to rotation, translation and scaling. Searching correspondences among millions of high dimensional descriptors may consume a large amount of time. Lu et al. [14] propose to simplify the feature matching process by projecting the high dimensional local feature to a low dimensional Hamming space. This process is also accelerated through separating all the 3D descriptors into visual words and search the query descriptor's correspondence within the visual words in [17]. In other words, the acceleration is achieved though reducing the searching scope. Though this can make the correspondence search faster, the potential risk arises that the best correspondence is missed, as the matching descriptor may not be assigned to the same visual word. Meanwhile, even though the search scope is reduced, the descriptor number in some visual words is still large since the descriptors are not evenly partitioned. As shown in [17], rejecting an image requires much longer time than successfully registering an image, which is due to the large RANSAC loop iterations for finding inliers.

In our localization pipeline, we also take the advantage of 3D SfM model for its clear observation and accurate camera pose estimation. Instead of searching through the descriptors of the whole 3D model, we retrieve a best matching image within the training set which are also used for SfM reconstruction and then assign the camera pose of the retrieved image to users, shown in Figure 2. The images are retrieved under a multi-task learning framework. Ideally, after the image retrieval, the camera orientation has already been roughly estimated. This idea is suited with the spirit of multi-task learning, which learns a number of supervised tasks simultaneously. In order to establish correspondences in the SfM reconstruction, multiple images from different angles are captured as the input to the SfM pipeline, especially for landmark areas. The images captured from various angles can be treated as different tasks. In our localization pipeline, we divide the view space into 3 regions: left, middle, and right direction. In each view direction, multiple images with different distance are captured to learn the multi-task regression model.

As mentioned above, the whole view is partitioned into 3 regions. We want to learn the landmark appearance relationship in each region. The whole algorithm is developed from a training set $\tau_i = \{(x_i^t, y_i^t), i = 1,2,3,...N_t\}$, where x_i^t is the a D dimensional feature vector for the i-th sample. In total, there are t tasks ($t = 3$ in our problem). Each task contains N_t samples. $y_i^t \in \{0,1\}$ denotes the sample labels. For our retrieval purpose, the positive samples are denoted as 1 and the negative samples are assigned value 0. Through our MTL framework (details in section 4), each task is assigned a weight vector w_t, which yields a correlation score for each sample by a dot product with the sample feature vector. We learn a MTL regression model for each landmark image group. During the testing phase, our MTL regression model assigns a correlation score to the testing sample for each landmark group. The regression model achieving the highest score for all three tasks is considered the retrieval result for the localization information obtained. Meanwhile, our different tasks represent the view directions. For a certain location, among all the three tasks, one task has an obviously higher score than the other two tasks. Thus, we can achieve the orientation information based on the retrieved task assignment.

In order to obtain a higher location and orientation accuracy, we search the nearest neighbor of the query image within the same location area group and the same task. The camera pose acquired from SfM for the nearest neighbor image is assigned to the query image, which is further refined through bundle adjustment.

4 Multi-task Learning for Multi-view Indoor Image-based Localization

Multi-task learning is an approach which learns a problem jointly with other related problems simultaneously, using a shared representation. Multi-task learning often leads to a better model for the main task, because it allows the learner to use commonality among the tasks.

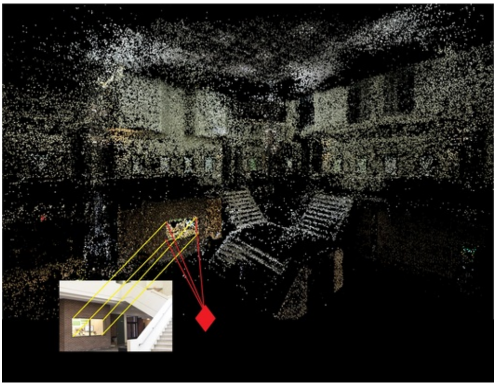


Figure 2: Camera pose estimation based on the retrieved image from the images used for SfM reconstruction

4.1 Regularized Multi-task Learning

Supposing the training data are the union of $t \geq 1$ groups and the training data among the groups have some relationship. Evgeniou et al. [6] propose a framework for multi-task learning based on the minimization of regularization functions similar to the existing ones, such as the one for Support Vector Machines (SVM). They decompose the learned parameter w in SVM into $w + w_t$, $t \in (1, 2, ..., T)$ where w is the common part and w_t is the task-specific part. The decision function is $f_t(x) = (w + w_t)^T \phi(x_i)$. Then the optimization problem is formulated as:

$$\begin{aligned} &\min_{w,w_1,...,w_T} \tfrac{1}{2} w^T w + \tfrac{\beta}{2} \sum_{t=1}^{T} w_t^T w_t + C \sum_{t=1}^{T} \sum_{i=1}^{l_t} \xi_{it} \\ &s.t. \quad y_{it}(w^T \phi(x_i) + w_t^T \phi(x_i)) \geq 1 - \xi_{it}, \\ &\qquad \xi_{it} \geq 0, (i = 1, ..., l_t; t = 1, ..., T) \end{aligned} \tag{1}$$

Here, the subscript *it* means the i the sample of the t th task. We learn all w_t as well as common w simultaneously. β regularizes the difference for w among the tasks and the ξ_{it} is the slack variable measuring the error that each of the final models w_t makes on the data.

4.2 Multi-task SVM

Regularized Multi-task Learning framework could benefit from learning the relationship among similar tasks. However, their decision space and correcting space are the same and without considering the bias term in decision function. Therefore, we map the input feature vector x_i into two different Hilbert spaces. The decision space $\phi(x_i)$ and the correcting space $\phi_t(x_i)$ for every given group t.

The goal is to find the t decision functions $f_t(x) = w^T \phi(x) + b + w_t^T \phi_t(x) + d_t$, where the t different tasks share a common decision function.

Then we extend the traditional SVM to a multi-task setting, we can formalize the optimization problems as:

$$\begin{aligned} &\min_{w,w_1,...,w_T,b,d_1,...,d_T,\xi} \tfrac{1}{2} w^T w + \tfrac{\beta}{2} \sum_{t=1}^{T} w_t^T w_t + C \sum_{t=1}^{T} \sum_{i=1}^{l_t} \xi_{it} \\ &s.t. \quad y_{it}(w^T \phi(x_i) + b + w_t^T \phi_t(x_i) + d_t) \geq 1 - \xi_{it}, \\ &\qquad \xi_{it} \geq 0, \ (i = 1, ..., l_t; \ t = 1, ..., T) \end{aligned} \tag{2}$$

The parameter w, w_t control the common decision function and correcting function capacity. Parameter β is the relative weight which balances these two capacities. C balances the complexity and proportion of non-separable samples. The slack variable ξ_{it} measures the error for the training data of each group models (including the common decision function and the correcting function).

By introducing the Lagrangian multipliers α, μ, the dual form of the objective function is:

$$\begin{aligned} L(\alpha,\mu) = \tfrac{1}{2}w^T w + \beta \sum_{t=1}^{T} w_t^T w_t + C \sum_{t=1}^{T}\sum_{i=1}^{l_t} \xi_{it} \\ + \sum_{t=1}^{T}\sum_{i=1}^{l_t} \alpha_{it}[1 - \xi_{it} - y_{it}(w^T \phi(x_i) + b + w_t^T \phi_t(x_i) + d_t)] - \sum_{t=1}^{T}\sum_{i=1}^{l_t} \mu_{it}\xi_{it} \end{aligned} \tag{3}$$

Based on Karush-Kuhn-Tucker (KKT) conditions, the dual form of the optimization is:

$$\begin{aligned} \max_{\alpha,\mu} L(\alpha,\mu) = \sum_{i=1}^{l} \alpha_i - \tfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j y_i y_j \phi(x_i)\phi(x_j) - \tfrac{1}{2\beta}\sum_{t=1}^{T}\sum_{i,j=1}^{l} \alpha_i\alpha_j y_i y_j \phi_t(x_i)\phi_t(x_j) \\ s.t. \quad \sum_{i=1}^{l_t} \alpha_i y_i = 0, t = 1,\ldots,T \\ \alpha_i + \mu_i = C, i = 1,\ldots,l_t \\ \alpha_i \geq 0, \mu_i \geq 0, i = 1,\ldots,l_t \end{aligned} \tag{4}$$

where w, w_t can be expressed in terms of training samples:

$$w = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)\ , \qquad w_t = \frac{1}{\beta}\sum_{i=1}^{l_t} \alpha_i y_i \phi_t(x_i) \tag{5}$$

Then the decision function is:

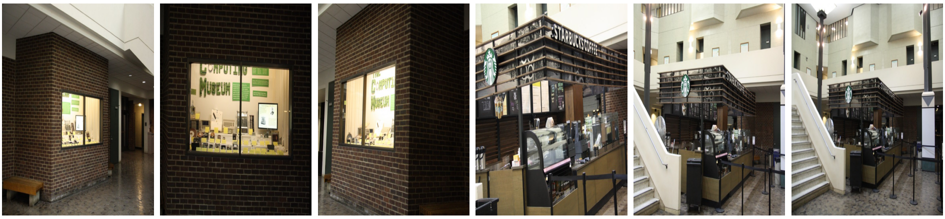
$$f_t(x) = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)\phi(x) + b + \frac{1}{\beta}\sum_{i=1}^{l_t} \alpha_i y_i \phi_t(x_i)\phi_t(x) + d_t, \quad t = 1,\ldots,T \tag{6}$$

The multi-task SVM is a quadratic programming (QP) problem. Vapnik et al. [23] propose a general approach to formalize problems with the group information, known as Learning With Structured Data (LWSD) and its SVM-based optimization formulation. Here we adopt a generalized sequential minimal optimization (SMO) algorithm as [12] to solve this optimization problem.

5 Experiments

We captured 1374 images of the indoor buildings for our system. Among all these images, 600 are dedicated for our multi-task regression model learning and testing which represent landmark areas of our building. 90 of the 600 images are used for testing and another 510 images are utilized for training our multi-view localization model. Except for the 90 testing images, all the rest 1284 images are casted into the Structure-from-Motion reconstruction pipeline to reconstruct the indoor environment. The reconstructed indoor model is shown in Figure 2,which provides us a 3D map for navigation. Every image used for the SfM model is associated with an estimated camera pose.

The 600 images are captured from 10 different locations. Each location contains 60 images with 20 images from each view direction (left, middle and right). The 20 images of

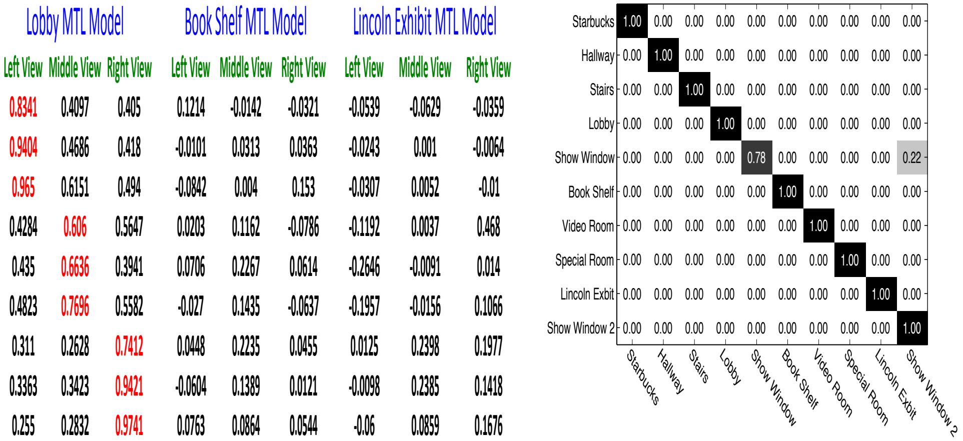


(a) Left view (b) Middle view (c) Right view (d) Close view (e) Middle view (f) Far view

Figure 3: Examples of images in different views (a - c) and different distances (d - f).

one view orientation are captured in different distances. The examples of images in different views and distances are shown in Figure 3.

During the learning process, SURF feature is extracted from each image to describe the image appearance. The bag-of-words model with the code of 500 units is applied to calculate the histogram representing each image through casting the SURF features into each of the 500 unit bins. For each location, we randomly select 3 images from each view direction, totaling 9 images for each location to test our retrieval accuracy. We first present the correlation score of Lobby testing images calculated by 3 different location regression models learned for Lobby, Book shelf and Lincoln Exhibit, as shown in Figure 4(a).



Lobby MTL Model			Book Shelf MTL Model			Lincoln Exhibit MTL Model		
Left View	Middle View	Right View	Left View	Middle View	Right View	Left View	Middle View	Right View
0.8341	0.4097	0.405	0.1214	-0.0142	-0.0321	-0.0539	-0.0629	-0.0359
0.9404	0.4686	0.418	-0.0101	0.0313	0.0363	-0.0243	0.001	-0.0064
0.965	0.6151	0.494	-0.0842	0.004	0.153	-0.0307	0.0052	-0.01
0.4284	0.606	0.5647	0.0203	0.1162	-0.0786	-0.1192	0.0037	0.468
0.435	0.6636	0.3941	0.0706	0.2267	0.0614	-0.2646	-0.0091	0.014
0.4823	0.7696	0.5582	-0.027	0.1435	-0.0637	-0.1957	-0.0156	0.1066
0.311	0.2628	0.7412	0.0448	0.2235	0.0455	0.0125	0.2398	0.1977
0.3363	0.3423	0.9421	-0.0604	0.1389	0.0121	-0.0098	0.2385	0.1418
0.255	0.2832	0.9741	0.0763	0.0864	0.0544	-0.06	0.0859	0.1676

(a)

	Starbucks	Hallway	Stairs	Lobby	Show Window	Book Shelf	Video Room	Special Room	Lincoln Exbit	Show Window 2
Starbucks	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hallway	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Stairs	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lobby	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Show Window	0.00	0.00	0.00	0.00	0.78	0.00	0.00	0.00	0.00	0.22
Book Shelf	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Video Room	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Special Room	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Lincoln Exbit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Show Window 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b)

Figure 4: (a) Correlation score of Lobby testing images calculated from 3 different location regression models. Red color emphasizes the high score of the correct image orientation at the Lobby location. (b) Location prediction confusion matrix.

From Figure 4(a), we can observe that the correlation score of the Lobby's testing images calculated from the Lobby MTL regression model is much higher than the other two locations' MTL regression models. For the correlation score learned from the Lobby MTL model, the left view task achieves the highest score for the left images, the same for middle and right view tasks. From this result, we can see that both location and orientation are correctly estimated. The location prediction confusion matrix of the whole dataset is listed in Figure 4(b).

The confusion matrix shows that our multi-view localization framework can accurately localize the query images. The only mistake happens on the show windows, whose side views are quite similar. The view orientation prediction accuracy is shown in Table 1.

Locations	Starbucks	Hallway	Stairs	Lobby	Show Window	Book Shelf	Video Room	Special Room	Lincoln Exhibit	Show Window2
Prediction	9/9	7/9	9/9	9/9	9/9	9/9	9/9	7/9	9/9	9/9

Table 1: The orientation prediction accuracy. Each location contains 9 test images.

From Table 1, the overall orientation accuracy can achieve 95.6%, which shows that our localization system can accurately predict the image orientation. Learning the tasks of different view directions simultaneously can help to improve the performance of each task by learning the similarities and differences among tasks. Most classification/regression methods cannot perform localization and orientation predictions at the same time. This is a significant advantage for our multi-view localization framework. Meanwhile, through learning the localization model from multiple views, the localization accuracy is greatly improved. We compare our localization accuracy with SVM, as SVM usually generate the state-of-the-art classification performance, shown as Table 2.

Method	Multi-view localization	Single-view SVM (left view)	Single-view SVM (middle view)	Single-view SVM (right view)
Accuracy	97.8%	66.7%	68.9%	61.1%

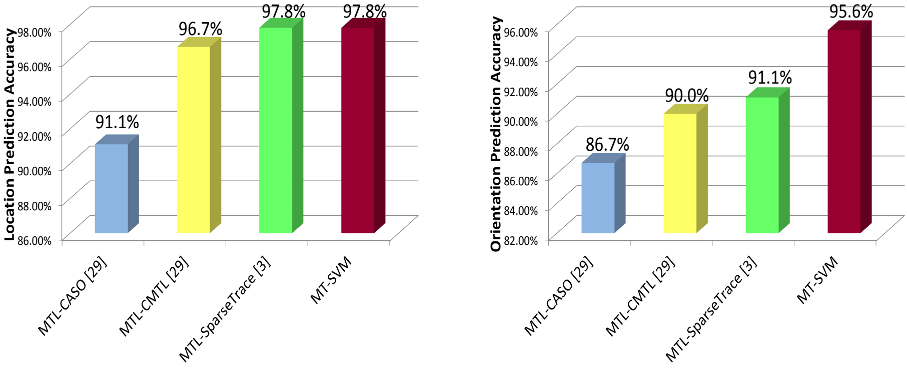
Table 2: The comparison of the localization accuracy with the single-view SVM.

From Table 2, we can observe that the use of multi-view images in learning the localization model can dramatically improve the prediction accuracy. Meanwhile, after learning the prediction model based on the multi-task learning framework, the correlation score can be achieved by simply a dot product, which is much faster than SVM and searching based method, like nearest neighbor search. This property is extremely useful for emergency situation. For prediction our 90 testing samples, the average elapsed time is shown in Table 3.

Method	Multi-view localization	SVM	Nearest Neighbor search
Elapsed time	0.35 (ms)	12.6 (ms)	62 (ms)

Table 3: Time performance comparison with SVM and Nearest Neighbor search.

From Table 2 and Table 3, we can observe that our multi-view localization system significantly outperforms the single-task based systems in both the localization accuracy and the time performance. Our localization model benefits from the simultaneous learning process based on different view tasks, leading to the high localization accuracy and orientation estimation performance. Through learning the different view images simultaneously, the learned weights for each task can accurately predict the orientation and generate relatively high correlation scores for the same location compared with other locations' prediction model. From the high performance of location and orientation prediction accuracy, as well as the time performance, multi-task learning framework is ideal for the localization purpose. We also compare the multi-task SVM method in our usage with other state-of-the-art multi-task learning algorithms (MTL-SparseTrace [3], MTL-CMTL [29], MTL-CASO [29]) for the localization task, shown as Figure 5.



(a) (b)

Figure 5: (a) Comparison of the location prediction accuracy of the Multi-task SVM method with other state-of-the-art Multi-task Learning algorithms. (b) Comparison of the Orientation prediction accuracy of the Multi-task SVM method with the state-of-the-art Multi-task Learning algorithms.

From Figure 5, we can see that the Multi-task SVM algorithm outperforms the other state-of-the-art Multi-task learning methods. Although MTL-SparseTrace algorithm achieves the same location prediction accuracy as our method, the Multi-task SVM algorithm significantly outperforms other methods on orientation prediction accuracy. Regularized Multi-task learning framework with the consideration of bias term in decision function can better facilitate the learning process of the relationship among relevant tasks. Experiments show that the embedding of SVM into multi-task setting can achieve the state-of-the-art performance on the localization task.

6 Conclusion

Making use of the multi-task learning method, we develop a multi-view image based localization system. By separating the view directions into 3 different partitions as tasks, we simultaneously learn the relationship among the tasks, which can improve the prediction accuracy of each view orientation. The learned multi-view regression model can accurately retrieve the location information. After learning the model, our multi-view system can retrieve the location and view orientation information by computing a dot product to assign a correlation score, avoiding large scale correspondences search. Leveraging the 3D localization system, we assign the camera pose of the nearest neighbor image of the same orientation and location used for SfM reconstruction to the query image, with further refinement using bundle adjustment. Embedding our multi-view method into the 3D localization system helps us better achieve the localization information in a 3D map.

References

[1] A. Argyrious and T. Evegenious. Multi-task feature learning. In *Neural Information Processing Systems (NIPS)*, 2007.

[2] R. Castle, G. Klein, and D.W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *Proceedings of the 2008 12th IEEE International Symposium on Wearable Computers(ISWC)*, pages 15–22, 2008.

[3] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4): 22, 2012.

[4] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3001 –3008, june 2011.

[5] M. Cummins and P. Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research(IJRR)*, 27(6): 647–665, 2008.

[6] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *Proc. 10th ACM SIGKDD International Conf. Knowledge Discovery and Data Mining*, 2004.

[7] J.M. Frahm, P. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *Proceedings of the 11th European conference on Computer vision(ECCV)*, pages 368–381, 2010.

[8] A. Irschara, C. Zach, J.M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 2599–2606, 2009.

[9] S. Kluckner and H. Bischof. Image-based building classification and 3d modeling with super-pixels. In *ISPRS Technical Commission III Symposium on Photogrammetry Computer Vision and Image Analysis*, 2010.

[10] J.J. Leonard and H.F. Durrant-Whyte. Simultaneous map building and localization for an autonomous mobile robot. In *Proceedings of the 1991 IEEE/RSJ International Workshop on Intelligent Robots and Systems '91. 'Intelligence for Mechanical Systems*, volume 3, pages 1442–1447, 1991.

[11] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *Proceedings of the 11th European conference on Computer vision(ECCV)*, pages 791–804, 2010.

[12] L. Liang and V. Cherkassky. Connection between svm+ and multi-task learning. In *International Joint Conference on Neural Networks*, 2008.

[13] G. Lu, V. Ly, and C. Kambhamettu. Structure-from-motion reconstruction based on weighted hamming descriptors. In *The International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014.

[14] G. Lu, N. Sebe, C. Xu, and C. Kambhamettu. Memory efficient large-scale image-based localization. *Journal of Multimedia Tools and Applications*, 2014.

[15] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 2, pages 2161–2168, 2006.

[16] D. Robertson and R. Cipolla. An image-based system for urban navigation. In *Proceedings of the 2004 British Machine Vision Conference(BMVC)*, pages 819–828, 2004.

[17] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *Proceedings of the 2011 IEEE International Conference on Computer Vision(ICCV)*, pages 667–674, 2011.

[18] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 1–7, june 2007.

[19] H. Shao, T. Svoboda, T. Tuytelaars, and L. Van Gool. Hpat indexing for fast object/scene recognition based on local appearance. In *Proceedings of the 2003 International Conference on Image and Video Retrieval(CIVR)*, pages 71–80, 2003.

[20] R.C. Smith and P. Cheeseman. On the representation and estimation of spatial uncertainty. *The International Journal of Robotics Research(IJRR)*, 5(6):56–68, 1986.

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

[22] U. Steinhoff, O. Dusan, R. Perko, B. Schiele, and A. Leonardis. How computer vision can help in outdoor positioning. In *Proceedings of the 2007 European conference on Ambient intelligence(AmI)*, pages 124–141, 2007.

[23] V.N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[24] A. Wendel, A. Irschara, and H. Bischof. Natural landmark-based monocular localization for mavs. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5792–5799, may 2011.

[25] J. Xiao, J. Chen, D. Yeung, and L. Quan. Structuring visual words in 3d for arbitrary-view object localization. In *Proceedings of the 10th European Conference on Computer Vision(ECCV)*, pages 725–737, 2008.

[26] Y. Yan, G. Liu, E. Ricci, and N. Sebe. Multi-task linear discriminant analysis for multi-view action recognition. In *International Conference on Image Processing (ICIP)*, 2013.

[27] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[28] W. Zhang and J. Kosecka. Image based localization in urban environments. In *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT)*, pages 33–40, 2006.

[29] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *Neural Information Processing Systems (NIPS)*, pages 702–710, 2011.