Neurocomputing 173 (2016) 83–92

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Where am I in the dark: Exploring active transfer learning on the use of indoor localization based on thermal imaging

Guoyu Lu [a,*], Yan Yan [b], Li Ren [a], Philip Saponaro [a], Nicu Sebe [b], Chandra Kambhamettu [a]

[a] Department of Computer and Information Sciences, University of Delaware, USA
[b] Department of Information Engineering and Computer Science, University of Trento, Italy

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form
9 June 2015
Accepted 5 July 2015
Available online 10 August 2015

Keywords:
Image-based localization
Active transfer learning
Thermal imaging

ABSTRACT

Indoor localization is one of the key problems in robotics research. Most current localization systems use cellular base stations and Wifi signals, whose localization accuracy is largely dependent on the signal strength and is sensitive to environmental changes. With the development of camera-based technologies, image-based localization may be employed in an indoor environment where the GPS signal is weak. Most of the existing image-based localization systems are based on color images captured by cameras, but this is only feasible in environments with adequate lighting conditions. In this paper, we introduce an image-based localization system based on thermal imaging to make the system independent of light sources, which are especially useful during emergencies such as a sudden power outage in a building. As thermal images are not obtained as easily as color images, we apply active transfer learning to enrich the thermal image classification learning, where normal RGB images are treated as the source domain, and thermal images are the target domain. The application of active transfer learning avoids random target training sample selection and chooses the most informative samples in the learning process. Through the proposed active transfer learning, the query thermal images can be accurately used to indicate the location. Experiments show that our system can be efficiently deployed to perform indoor localization in a dark environment.

1. Introduction

The purpose of indoor localization [1] is to help the users navigate in a large and complicated environment. For example, in an airport, travelers want to receive prompt navigation information in order to board their plane. Currently, most indoor localization systems are based on GPS. The GPS signals are scattered by the roofs and walls; hence, the strength of signals is attenuated, which potentially negatively affects the localization accuracy. Another technology is based on pre-deployed beacons and requires a high beacon distribution density. Furthermore, the radio signal based indoor localization systems can usually locate a person while failing to tell the user orientation information.

Although image-based localization is mainly used in outdoor environments to make up the deficiency of weak GPS signals, recent research has studied indoor localization as well. The basic idea of indoor image-based localization is matching a query image with all the images in a database to find the nearest neighbor image in a descriptor space. Structure-from-Motion (SfM) techniques enable the 3D model to be utilized in the localization system. Based on the matching between 2D images and a 3D reconstruction model, the 3D coordinate of the camera can be returned together with the pose estimation information. SfM techniques do not require the camera to be calibrated as in stereo reconstruction; thus images used for SfM reconstruction are easier to obtain. Due to the same reason, updating the SfM reconstruction model is easier than other techniques. A user only needs to capture an image and then the localization system matches the image against the 3D model in the whole localization process. Meanwhile, the 3D model can perform the functionality of a 3D map that helps users better understand the building structure for visit planning purposes. A 3D model, however, usually contains millions of descriptors, and searching through the whole descriptor space for correspondences would potentially consume a lot of time, making the system less practical.

During the Structure-from-Motion process, every image used in the reconstruction is assigned with pose estimation information, including position and orientation. Instead of searching through the whole point cloud of a 3D model, we search the nearest neighbor of the query image among the images used for the 3D SfM reconstruction and assign the pose information of the returned image to the query image. Then we can assign 3D coordinates based on the 2D image retrieved, making use of the advantages of both the 3D model and the 2D images.

* Corresponding author.
E-mail address: luguoyu@udel.edu (G. Lu).

http://dx.doi.org/10.1016/j.neucom.2015.07.106
0925-2312/

Current image-based indoor localization systems make the assumption that light resources are always sufficiently present. The reality, however, is that the lighting conditions may vary dramatically from place to place. In many emergent situations, such as a power outage of a building, there can be little to no lighting. This makes the localization and navigation tasks challenging for both robots and humans. In dealing with this problem, we propose to use thermal-infrared imagery for indoor image-based localization. Imagery from a long wave infrared camera is based on the temperature of the object and is not dependent on the lighting. Since indoor buildings are composed by different objects with different materials, e.g. glasses and wooden tables, the surface temperature of the indoor objects also varies. By recognizing the object shapes, thermal imagery is an ideal choice to perform localization tasks in a dark environment.

Unlike common RGB color images, capturing thermal images require more effort. To overcome some of the limitations of the thermal cameras, we have to invest more time into focusing and framing a scene. This process is labor intensive, which means that far fewer samples were taken. Our thermal camera has to be plugged into a PC with certain ports, so the hardware is a limitation for the ease of data collection. It is very expensive to collect sufficient images to perform accurate localization. To solve this problem, we propose performing transfer learning between color images as the source domain and thermal images as the target domain. Transfer learning based image localization aims to leverage the useful information from visible images to thermal images. During training, there are usually fewer target training samples than the source samples. By leveraging the source task, the learning of the target task is enhanced. In the traditional transfer learning, the target samples are selected randomly. However, not all the target samples are equally informative. Active transfer learning selects the most informative target samples to train the model, which provides a higher localization accuracy by avoiding learning the model through randomly selected samples. We captured color and thermal images in different locations and train the location classification model based on the active transfer learning algorithm. Here, each location is treated as a landmark group for classification. The whole process is illustrated in Fig. 1.

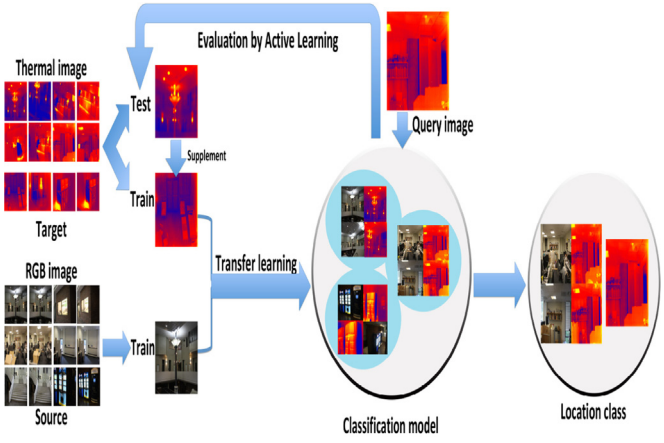


Fig. 1. Our image-based indoor localization system based on active transfer learning: the target domain is the thermal images and the source domain is RGB images. The target domain is separated into a training set and testing set. Through transfer learning, the source domain training set is adapted into a target domain to train a classification model for the thermal images. Using the trained classification model, the most informative target sample in the test set is added into the training set through active learning. Utilizing the enriched training set, the system learns a new classification model. After the target training set reaches a certain number of samples, the classification model is finalized to classify the query images into one of the location classes.

To summarize, the contributions of our paper are as follows: (1) We present a framework for solving the indoor image-based localization problem in a dark environment; (2) We apply active transfer learning on indoor image-based localization problem in order to adapt the training set to be most informative. (3) Finally, we jointly use the common RGB image and thermal images together to learn a better model for indoor localization based on thermal images.

The rest of the paper is organized as the following: Section 2 provides the related work on image-based localization, the usage of thermal image in computer vision problems and the transfer learning; Section 3 introduces our image localization system based on the transfer learning between RGB images and thermal images; Section 4 presents the proposed active transfer learning method; Section 5 provides our image localization experiments result; Section 6 discusses the failure case of our method; Section 7 concludes the paper.

2. Related work

Image-based Localization: Image-based localization [2] is widely applied to localization problems, especially in weak GPS signal areas. This paper calculates the pose of a query image by utilizing a database of building facades and associates a 3D-coordinate system with images in the database. Schindler et al. [3] selected the vocabulary [4] using informative features to improve image-retrieval performance on a large street-side image database. Xiao et al. [5] further improved localization accuracy by using geometric verification with a bag-of-words method.

3D SfM models [6,49] are used for image-based localization problems in enhancing the accuracy. Li et al. [7] used mutual visibility information for 3D-to-2D matching. Sattler et al. [1] directly matched descriptors of 2D images to descriptors of 3D model. Irschara et al. [8] proposed to retrieve images containing the most descriptors matching the 3D points. The proposed localization framework achieves a high image-registration rate by accelerating the matching process. Gronat et al. [9] changed the place recognition as a classification problem based on a classifier trained by geo-tagged images. Lu et al. [42,48] improved the localization accuracy and memory efficiency through local feature processing.

Indoor Localization: For indoor image-based localization, many different techniques have been tried. Ravi et al. [10] matched a query image to a database using color histograms. Kosecka et al. [11] detected edges of room images, generating edge histograms for each image for matching. Liu et al. [12] used a Transfer Regression Model to localize. Kawaji et al. [13] applied online transfer learning in humanoid robots for object recognition. Wannous et al. [14] proposed an automatic indexing method of the content stream of a camera mounted on the shoulder based on the presence in specific 3D places related to instrumental activities to detect the activity related places. The system in [15] involved RANSAC and applied ASIFT features to perform affine invariant image matching. Yu et al. [16] proposed combining the color, person detection, face recognition, and non-background information for localization. Lu et al. [50] separated the entire view into several directions and learned a multi-view localization system to predict location and orientation.

Transfer Learning: Transfer learning deals with applying knowledge learned from some existing tasks to new domains which share some commonalities. Sun et al. [17] used Wifi based indoor localization with transfer learning when variation in signal distributions causes the old localization model to be inaccurate. Jiang et al. [18] presented a cross-domain SVM algorithm which adapts previously learned support vectors from one domain to facilitate

classification in another domain. Shi et al. [19] applied the knowledge transferred from other domains in the largest extent to help learn the current domain, meanwhile queried experts in the smallest extent. Xu et al. [20] developed a system that optimizes the kernel weights in closed-form based on the equivalence between group-lasso and multi-kernel learning. Soek et al. [21] proposed to perform indoor location estimation with environmental changes. Kimura et al. [22] applied online transfer learning in humanoid robots for object recognition. Liu et al. [23] classified scenes using multimodal fusion with transfer learning based on the use of SURF and MRHM features. Duan et al. [24] incorporated a cross-domain kernel learning framework into existing kernel methods, which minimizes both the structural risk functional and the mismatch between data distributions from source and target domains. Similarly, they incorporated adaptive Multiple Kernel into this framework to perform visual event recognition task [25]. Yan et al. [51,52,53] also used multi-task learning to solve head pose, daily activity and event recognition problems. Chattopadhyay et al. [26] developed an integrated model that performs transfer and active learning simultaneously through single convex optimization, which reduces distribution difference between the set containing re-weighted source and the queried target domain data and the set of unlabeled target domain data. To deal with the problem of samples being insufficient, Han et al. [27,30] proposed frameworks of video recognition by semi-supervised feature selection to better identify the relevant video features. Similarly, Yang et al. [28,29] mined label correlations and visual similarities to selected data as diverse as possible to overcome the issue of small labeled data in the seed set. Meanwhile, Han et al. [31] added a joint $l_{2,1}$-norm on multiple feature selection matrices to ensemble different classifiers' loss functions into a joint optimization framework, where $l_{2,1}$ norm minimization is also used in discriminative feature selection [32,33]. Chang et al. [54,55] explored semantic information to perform event detection.

Thermal Imaging: Thermal images are largely used in human tracking. Bertozzi et al. [34] implemented an experimental vehicle equipped with infrared camera to detect pedestrians. Nanda et al. [35] proposed an effective probabilistic template to capture the variations in human shape, especially in the case where the contrast is low and body part is missing. In recent research, thermal imaging was deployed on mobile platforms designed for search and rescue tasks [36,37]. Fehlman et al. [38] introduced the use of thermal sensor to detect and classify the non-heat generating objects used for mobile robot navigation. Cielniak et al. [39] proposed an approach to track multiple persons on a mobile robot based on both color and thermal vision sensors. Vidas et al. [40] generated dense 3D models using the combination of RGB-D camera and thermal cameras. Furthermore, the thermal videos were applied to build real-time human detection on the autonomous mobile platform [41].

In our work, we make use of both visible and thermal images to perform the indoor localization task. The use of transfer learning largely enhances the localization accuracy compared with simply using thermal images. We are able to perform indoor localization tasks even in a totally dark environment.

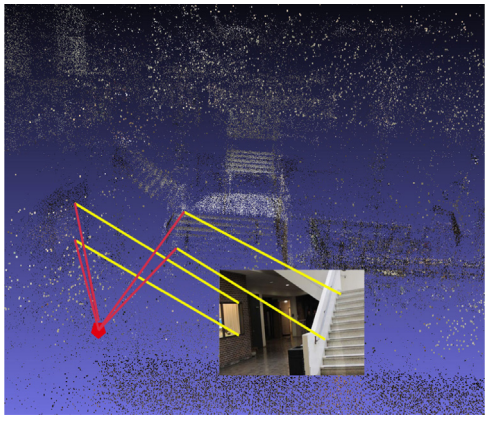


Fig. 2. Correspondences between 2D image and 3D model and the camera pose estimation.

3. Localization based on thermal images

3.1. Basic localization pipeline

An image-based localization system provides robots or human beings the location information based on the captured images. Taking a common color image as the query, the location coordinate is returned along with the orientation information.

With the use of a 3D model generated from SfM, camera pose is estimated based on correspondences, as showed in Fig. 2. Sattler et al. [1] proposed a direct 2D-to-3D matching scheme to perform localization tasks. In this framework, features extracted from 2D images are cast into the corresponding visual-word of the descriptors from the 3D model to compute correspondences. Within each visual-word, a k-d tree is built to search for the approximate nearest neighbor. This system requires all the 3D descriptors stored in memory for searching. A large Structure-from-Motion reconstruction model, however, usually contains millions of descriptors. Storing all the descriptors in the memory would potentially consume large amounts of computation resource. Lu et al. [42] proposed to project the descriptors into Hamming space and correspondingly changed the localization pipeline to reduce the memory cost during the localization process. Furthermore, the matching process is accelerated by simplifying the descriptor distance computation method. Nevertheless, during the searching process, the corresponding descriptor might not get returned due to incorrect visual-word assignment. In our paper, we cluster landmark building areas into groups. For a new query image, we classify the query image into one of the landmark areas. Within the landmark building area, we retrieve the best matching image among all the images used for SfM reconstruction. The camera pose associated with the matching image will be assigned to the query image as the camera pose. In this way, we can utilize the clear view of the whole 3D model, as well as the accurate 3D coordinates for navigation. Furthermore, by performing image classification, we get rid of searching through all the descriptors of the 3D model. Bundle adjustment between query and returning images is conducted to refine the camera pose.

3.2. Localization based on thermal imaging

For vision-based robot navigation systems, the lighting condition is a critical problem. Existing vision-based methods are under the assumption that sufficient lighting is provided. In reality, however, lighting is not always adequately supplied. The most common situation is that in shadows or weak lighting area, the captured images are usually blurred due to the long exposure, resulting in inaccurate localization accuracy. The extreme case would be a power outage. In this situation, the indoor environment would be totally dark. Existing vision based navigation methods will not be applicable anymore.

To address this problem, we utilize the thermal images captured by long wave infrared cameras to perform the localization. Thermographic cameras create the images based on a much larger wavelength than color images. In indoor conditions, since the temperature of objects differs from each other, the infrared radiation varies, resulting

Fig. 3. Example of infrared in the dark environment. The left image is the thermal image captured in the same dark environment as the right visible image.

in an image showing the shapes of various objects. Because thermal images are independent of the lighting condition, the use of thermal camera would be ideal for performing indoor localization in a dark environment. Fig. 3 shows an example of a thermal image captured in the dark environment. Some sample images we captured from thermal images are provided in Fig. 4.

From Fig. 4, we can still observe the objects in thermal images. That provides us the possibility to localize images in an environment without light. The thermal images, however, are not easy to obtain compared to the common RGB color cameras. In order to capture a clear thermal image from a certain faraway distance with high resolution, a scientific infrared camera usually needs to connect to a PC to be controlled and needs to manually tune the focal length to keep the objects in the image distinct from each other. It is also much more expensive to capture high quality thermal images using the infrared cameras. To learn a better classification model, we require more learning samples with a small cost. A solution for solving this problem is to adapt the knowledge from a related source domain to the target domain, where the training samples of the source domain are easily available while only a few training samples from the target domain can be obtained. In these conditions, we apply transfer learning to learn a better landmark classification model. In our experiments, the RGB images captured by common color cameras are used as the source domain and the thermal images are the target domain.

In the training process, we learn the classification model based on both visible RGB images and the thermal images through transfer learning. We label the whole building area into several landmark areas and group the images based on the landmark labels. When a query thermal image is transmitted to the system, we classify the image into one of the landmark areas to obtain the basic location information. Within each group, the visible images are filtered by bilateral filters, which reduce the small edges. To make the thermal image and color image in the same space, we further use Difference of Gaussian (DoG) filter to process original query thermal images and color images after bilateral filter. The processed image samples are shown in Fig. 5. The filtered thermal image is matched to the best filtered visible image using geometric line feature [43] and RANSAC. The camera pose of the returned image will be assigned to the query image, which will be refined through Bundle Adjustment [44].

4. Active transfer learning

The goal of transfer learning is to explore the *target* domain by learning sufficient *source* and few *target* samples. The few *target* samples are randomly selected. All samples, however, do not contain information equally. We prefer to select samples that contain most information (*i.e.*, most difficult to classify). Active transfer learning seeks to use an efficient sampling strategy to select *target* samples whose labels are obtained by domain expert that helps learn the transfer learning model sufficiently. We provide a brief introduction of the following two parts in this section: (1) AMKL-based transfer learning; (2) active learning for multi-class classification, followed by introducing our proposed active transfer learning method.

4.1. Adaptive mkl-based transfer learning

To adapt a learned SVM into a new domain (*target* domain), we formulate the target decision function as

$$f^T(x) = f^S(x) + \Delta f(x) \tag{1}$$

Here $f^T(x)$ and $f^S(x)$ represent the *target* and *source* decision functions separately. $\Delta f(x)$ is the mismatch between the *source* and *target* domains. x denotes a feature vector.

Duan et al. [45] extended the above *target* decision function as

$$f^T(x) = \sum_{p=1}^{P} \gamma_p f_p(x) + \sum_{m=1}^{M} d_m w_m^T \phi_m(x) + b \tag{2}$$

$f_p(x)$ represents the pth pre-learned classifiers of P. These pre-trained classifiers are trained using labeled data from both source and target domains. γ_p's are the combination coefficients. M kernels with coefficients d_m are linearly combined to model $\Delta f(x)$ with a bias term b. $(.)^T$ is the transpose operator and $\phi_m(x)$ denotes a non-linear feature mapping function which computes base kernels of $k_m(x_i, x_j) = \phi_m^T(x_i)\phi_m(x_j)$.

4.2. Multi-class active learning

Allwein et al. [46] proposed a framework to reduce k-class classification to l binary problem based on margin-based learning.

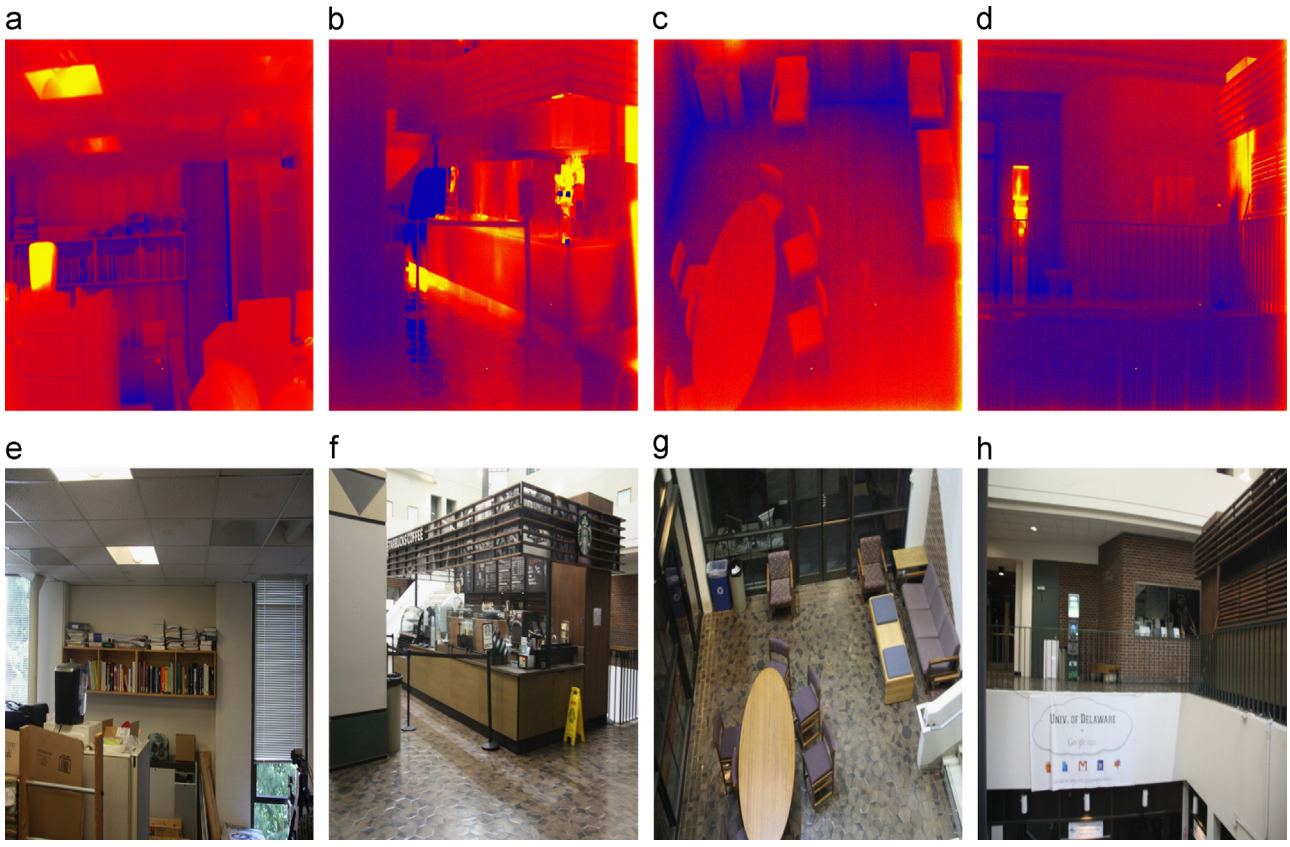


Fig. 4. Example of infrared images along with the corresponding RGB color images.

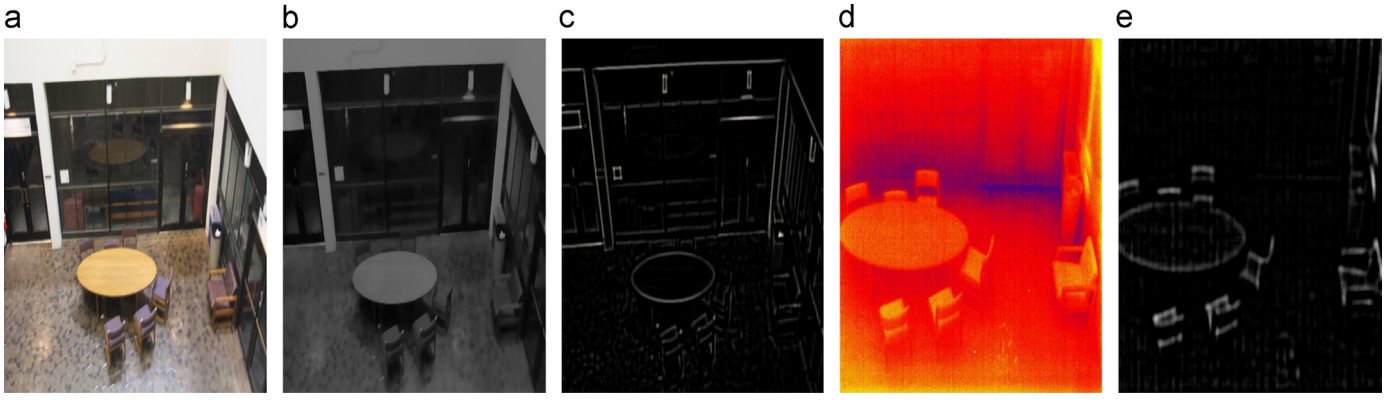


Fig. 5. Examples for processed images for brute force search using geometric line feature. (a) Original color image. (b) Color image after bilateral filter. (c) Color image after bilateral and DoG filter. (d) Original thermal image. (e) Thermal image after DoG filter.

Error correcting output coding (ECOC) selects the most consistent label r regarding the predictions $f(x)$. For example, for the sample x labeled as r, the loss on sample (x, r) is minimized over all label options $r \in Y$, $Y = \{1 \ldots k\}$. The loss function of (x, r) is formulated as

$$d_L(M(r), f(x)) = \sum_{s=1}^{l} L(M(r,s) f(x)) \tag{3}$$

Here L is the loss function, while $M \in \{-1, 0, +1\}^{k \times l}$ denotes the *coding matrix*. $M(r)$ represents the rth row of M. With regards to a k classification problem, the total loss is minimized through the predicted label

$$\hat{y} = \arg\min_{r} d_L(M(r), f(x)) \tag{4}$$

We approximate the expected loss function with the smallest loss across all possible labels. For x whose predicted label is y_x, the least maximum expected loss is

$$\arg\max_{x} \sum_{s=1}^{l} L(M(y_x, s) f(x)) \tag{5}$$

Using active learning, we select samples that generate the least maximum expected loss using the predicted label by expert.

4.3. Active transfer learning framework

We represent our active transfer learning framework in Fig. 6. Labeled *source* data and *target* data (both labeled and unlabeled)

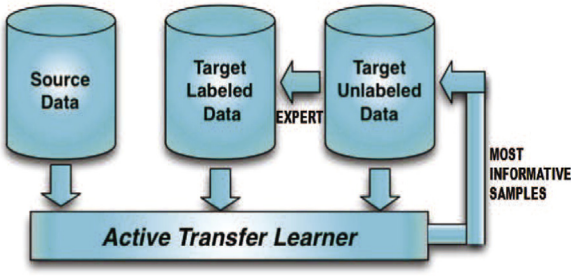


Fig. 6. Active transfer learning framework.

Fig. 7. Infrared camera we use to capture thermal image.

are used to train the transfer classifier. The most informative unlabeled *target* data is chosen and related labels are obtained from expert for transfer learning under *best worst case* model.

Algorithm 1. AMKL-based active transfer learning.

Input: Unlabeled *target* data D^t and labeled *source* data D^s.
Let $D^t = D_l^t \cup D_u^t$. Label one *target* sample per class randomly and
add them to D_l^t.
for $k = 1, \ldots, K$ do
(1) Apply $D^s \cup D_l^t$ to obtain adapted classifier $f_T^m(x)$ for *target*.
The steps involved are enumerated as follows:
- **for** $j = 1, \ldots, T_{max}$ do
  - Calculate the dual SVM variable α_j using kernel matrix $\sum_{m=1}^{M} d_m \tilde{K_m}$.
  - Update the base kernel coefficients d_j as $d_{j+1} = d_j - \eta_j g_j$.
- **end for**

(2) For all the samples $x_i \in D_u^t$, compute the loss function using Eq. (5).
(3) Choose the most informative *target* samples s^* (could also be batch mode)
with least loss.
(4) Obtain sample label y_{s^*} from expert.
(5) Add samples s^* to D_l^t.
(6) Classify the *target* test data with $f_T^m(x)$.
end for

We present details of the active transfer learning procedure in Algorithm 1. We randomly select and label one *target* sample per class in the first stage. Step (1) presents the transfer learning procedure. Labeled *target* samples D_l^t and labeled *source* samples D_l^s are combined together to train an adaptive SVM classifier $f_T^m(x)$ in the *target* domain D^t. $f_T^m(x)$ denotes the Mth of M learners used in MKL framework. At each iteration j, alternative coordinate descent is employed to optimize the dual variable α_j and the kernel coefficients d_j. This procedure iterates T_{max} times. η_j is the learning rate and g_j represents the update direction. Steps (2)–(5) denote the active learning procedure. In Step (2), we estimate loss values for all the unlabeled *target* samples. From Steps (3) to (5), those *target* samples that produce the least loss are labeled by domain expert, followed by adding labeled *target* samples for transfer learning. Depending on the desired accuracy, we repeat this process K times.

5. Experiments

We captured 131 thermal images of an indoor scene using a Xenics Gobi 640 GigE infrared camera, which is an uncooled long wave infrared camera capable of imaging infrared wavelengths between 8 and 14 μm. The resulting image is a thermal map of the environment. It has a resolution of 640×480 and has a 50 mC sensitivity at 30 °C, with a max frame rate of 8 fps. Fig. 7 shows the infrared camera we use to generate thermal images.

All the 131 thermal images are from 13 locations with about 10 images from each location. For the common RGB images, we captured 1284 images of the same building indoor environment, which are used for SfM reconstruction. The 3D model is reconstructed using Bundler [47], consisting of 189,787 reconstructed points and 928,058 descriptors. Each image is associated with a camera pose. The reconstruction result is shown in Fig. 2. Among these 1284 images, 260 images were chosen as the source domain training samples, 20 images from each location.

During the learning process, we randomly choose one thermal image from each location together with all the source domain images to train the classification model. For each location, we train an SVM classifier. The labels are either 1 or -1 depending whether the sample belongs to the class or not. Based on the trained classification model, we evaluate all the test samples based on the loss function in Eq. (5). The most informative target sample will be selected to be added into the target training samples. The whole process was described in Algorithm 1. This process continues until 3 images from each location are added into the training set. We compare our active transfer learning method with other methods (SVM trained with thermal images, SVM trained with visible images, Active SVM trained by thermal images, Active SVM trained by visible images, cross-domain SVM (CD-SVM) [18], active transfer domain knowledge [19] (ATDK), multiple kernel learning (MKL) [20], domain transfer multiple kernel learning (DT-MKL) [24], AMKL random selection [25], joint transfer and batch-mode active learning [26] (JTBAL)) for verifying the effectiveness of our method. The active learning method used for training SVM is the same as our active transfer learning method. The only difference is that we did not apply transfer learning while training active SVM. From Table 1 we can see that our transfer learning classifier works well on the localization task.

From Table 1, our method performs the best compared with other methods. In certain categories, some other methods can also have relatively high performance, but may also be very low in other categories. For example, the SVM trained model with only thermal images can achieve 100% accuracy on the vending machine from the right view (Vending machine right). However, the accuracy can drop to 63.26% in Starbucks from the left view (Starbucks left). From the bold black numbers, notice that our method performs the best in most categories. Except for the category of 'Computer Show Window Left', location classification of our method in each classification category always keeps the highest or the second highest accuracy among all the methods. Since 'Computer Show Window Left' category is mainly made up of the glass, the object behind the glass is not visible in infrared cameras. This might cause reduced location prediction accuracy. The average accuracy and precision score are presented in Figs. 8 and 9. Our system maintains a highest average accuracy score.

Table 1
The location classification accuracy of the 13 locations. The black bold numbers represent the highest location classification accuracy for each location.

Accuracy for locations	Robot show window (%)	Computer show window left (%)	Basement left (%)	Lamp post (%)	Vending machine left (%)	Vending machine right (%)	Computer show window right (%)	VIMS lab (%)	Stairs (%)	Starbucks left (%)	Starbucks right (%)	Basement right (%)	IR lab (%)
Our method	**96.67**	93.84	**99.17**	**98.20**	**94.35**	97.30	**98.20**	**96.67**	**99.17**	**98.20**	**98.20**	**96.58**	98.20
SVM trained by thermal images	91.83	91.83	98.97	86.73	91.83	92.85	92.85	92.85	91.83	95.91	91.83	91.83	92.85
SVM trained by visible images	93.87	91.83	82.65	95.50	91.83	**100**	94.89	92.85	91.83	63.26	91.83	93.87	**98.97**
Active SVM trained by thermal images	92.37	92.37	98.97	91.60	92.97	95.50	94.89	95.04	93.39	95.91	95.50	92.31	94.57
Active SVM trained by visible images	94.56	91.83	86.26	96.40	92.37	**100**	95.37	93.02	92.25	73.28	93.84	93.87	**98.97**
CD-SVM [18]	91.60	96.40	96.95	86.87	90.60	92.37	95.37	92.25	91.60	94.66	91.47	92.31	92.25
ATDK [19]	91.60	96.40	95.67	94.56	92.79	94.56	97.30	94.56	95.50	91.60	93.84	94.87	92.37
MKL [20]	91.47	**96.58**	97.30	88.37	91.47	96.80	96.12	93.02	93.39	94.57	92.25	93.80	93.02
DT-MKL [24]	92.79	**96.58**	99.10	87.60	91.60	95.50	96.18	93.60	92.25	96.58	90.91	94.87	92.37
AMKL [25]	95.67	92.79	**99.17**	96.40	92.79	96.67	97.30	95.04	**99.17**	**98.20**	97.44	95.50	96.58
JTBAL [26]	96.18	93.84	98.97	97.71	**94.35**	96.80	96.12	96.67	97.71	**98.20**	96.40	**96.58**	97.71

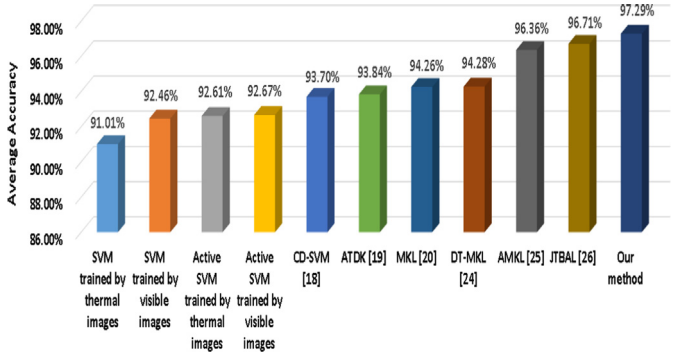


Fig. 8. Average accuracy of our method and other methods.

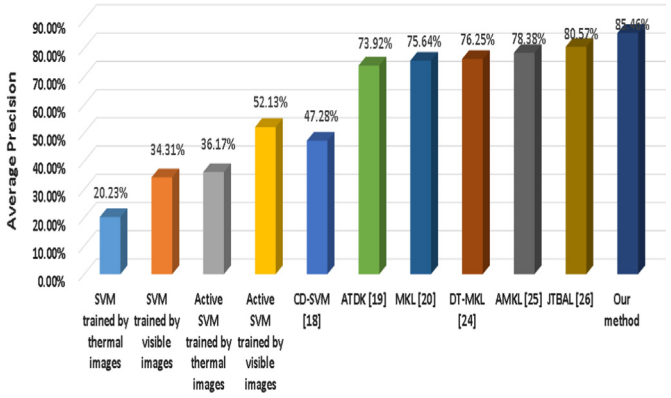


Fig. 9. Average precision of our method and other methods.

From Figs. 8 and 9, we can see that the average accuracy is relatively high for almost all the methods. But the precision for most methods is very low. The reason is that while testing the average accuracy, the negative samples take the majority of all the testing samples. For example, to classify the testing images of stairs, all the other 12 classes are treated as the negative samples. If a classifier judges all the testing samples as negative, it can still keep a high accuracy. But in such a situation, the precision would be extremely low, as all positive samples are mis-classified as negative samples. For this reason, our method is much more useful than other methods for localization purposes.

In our testing cases, images from "Starbucks", "Vending machine" and "Computer show window" are captured from left and right two views, shown as Fig. 10. We treat the images taken from left and right views of the same location as two different classes. For the location recognition purpose, the left and right view of the same location should be treated as the same class. We, however, want to test our system's ability in view recognition, as obtaining the orientation information is also essential in localization. In obtaining the view information together with the location, users can better learn the environment and get to know where to go. For this purpose, we captured some locations' images in left and right views separately and treat them as different classes. Experiments on accuracy show that our method can distinguish the images captured from different directions for the same object. That is exceptionally helpful for the localization tasks, since orientation information is also useful for navigating the robots or human users.

As our active transfer learning method uses SVM as the base classifier, the running time of our model does not obviously differ from other methods which all use SVM as the base classifier, with approximately 9 ms per testing image for classification.

6. Discussion

Infrared cameras are good at generating thermal images based on varying object surface temperatures. For transparent materials such as glass, however, the infrared camera cannot really see the objects behind the material, as exhibited in Fig. 11. Objects behind such transparent materials are not visible under infrared cameras. If the surrounding area is mostly decorated with transparent objects, the thermal images are not an appropriate choice to recognize the location.

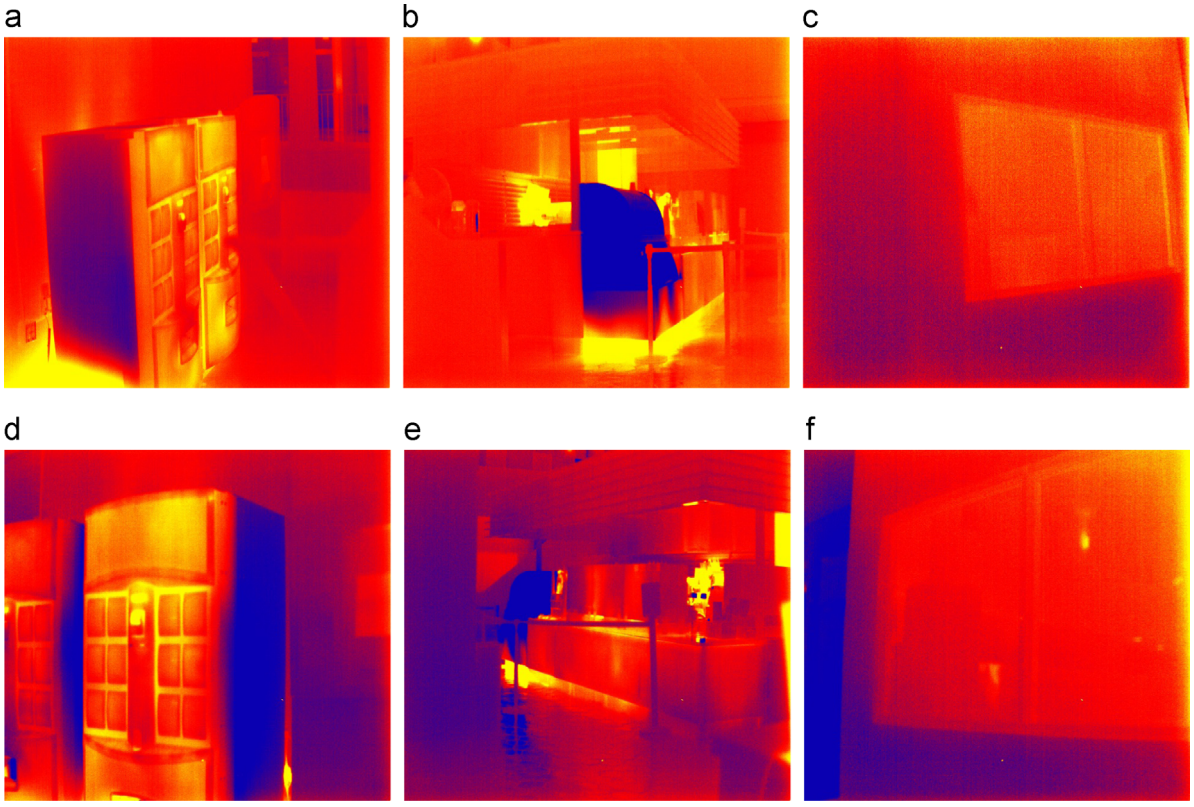


Fig. 10. Examples of infrared images captured from left and right two views. Upper level images are from the left view and lower level images show the right view. Our method can accurately separate the images captured from different directions.

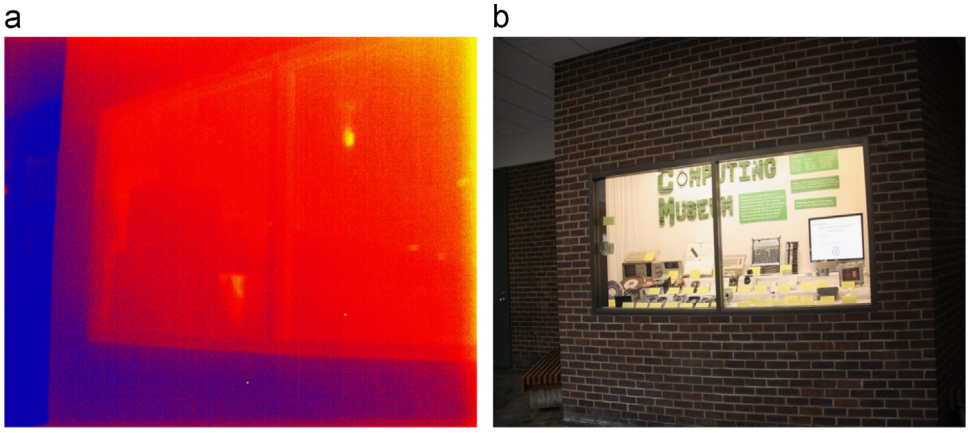


Fig. 11. Images captured by infrared camera and common RGB camera in front of transparent object.

7. Conclusions

In this paper, we developed an indoor image-based localization system based on thermal imaging. This system is aimed at providing localization services in an environment with little light. Existing image-based localization systems assume that the images are always captured under adequate light, which is not always possible. Object shapes can be recognized due to varying object surface temperature. Considering that thermal images are difficult to obtain with our hardware limitations, we apply transfer learning to enrich the training samples by taking RGB color images as source domain, thermal images as target domain, and adapting the knowledge learned from the source to the target. The RGB images are captured at the same places as the thermal images. Color and thermal images representing the same location and direction are classified as the same group. Furthermore, during the learning process, we take advantage of active learning to avoid random selection of target training samples. Through selecting the most informative target samples for training, our location classification accuracy is enhanced. After thermal image classification, the query image is assigned with a camera pose of the best matched image used for SfM reconstruction with further refinement by Bundle Adjustment. Our system is most suitable for dark situations, such as a power outage in emergency, and military use during the night.

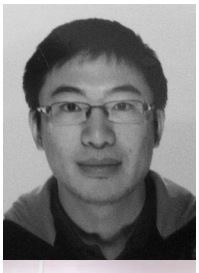
Acknowledgments

This work is funded by NSF CDI Type I Grant 1124664 and by Cooperative Agreement W911NF-11-2-0046 (ARO Proposal no. 59537-EL-PIR).

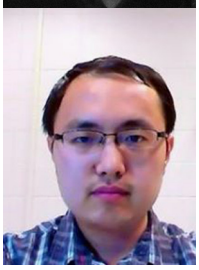
References

[1] T. Sattler, B. Leibe, L. Kobbelt, Fast image-based localization using direct 2d-to-3d matching, in: International Conference on Computer Vision, 2011, pp. 667–674.
[2] D. Robertson, R. Cipolla, An image-based system for urban navigation, in: British Machine Vision Conference, 2004, pp. 819–828.
[3] G. Schindler, M. Brown, R. Szeliski, City-scale location recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–7.
[4] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2161–2168.
[5] J. Xiao, J. Chen, D. Yeung, L. Quan, Structuring visual words in 3d for arbitrary-view object localization, in: European Conference on Computer Vision, 2008, pp. 725–737.
[6] D. Crandall, A. Owens, N. Snavely, D. Huttenlocher, Discrete-continuous optimization for large-scale structure from motion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3001–3008.
[7] Y. Li, N. Snavely, D.P. Huttenlocher, Location recognition using prioritized feature matching, in: European Conference on Computer Vision, 2010, pp. 791–804.
[8] A. Irschara, C. Zach, J. Frahm, H. Bischof, From structure-from-motion point clouds to fast location recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2599–2606.
[9] P. Gronat, G. Obozinski, J. Sivic, T. Pajdla, Learning and calibrating per-location classifiers for visual place recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 907–914.
[10] N. Ravi, P. Shankar, A. Frankel, A. Elgammal, L. Iftode, Indoor localization using camera phones, in: 7th IEEE Workshop on Mobile Computing Systems and Applications, 2006, pp. 49–49.
[11] J. Kosecka, L. Zhou, P. Barber, Z. Duric, Qualitative image based localization in indoors environments, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2003, p. II-3.
[12] J. Liu, Y. Chen, Y. Zhang, Transfer regression model for indoor 3d location estimation, in: Advances in Multimedia Modeling, 2010.
[13] H. Kawaji, K. Hatada, T. Yamasaki, K. Aizawa, Image-based indoor positioning system: fast image matching using omnidirectional panoramic images, in: ACM International Workshop on Multimodal Pervasive Video Analysis, 2010, pp. 1–4.
[14] H. Wannous, V. Dovgalecs, R. Mégret, M. Daoudi, Place recognition via 3d modeling for personal activity lifelog using wearable camera, in: Advances in Multimedia Modeling, 2012.
[15] X. Li, J. Wang, Image matching techniques for vision-based indoor navigation systems: performance analysis for 3d map based approach, in: International Conference on Indoor Positioning and Indoor Navigation, 2012, pp. 1–8.
[16] S. Yu, Y. Yang, A. Hauptmann, Harry potter's marauder's map: localizing and tracking multiple persons-of-interest by nonnegative discretization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[17] Z. Sun, Y. Chen, J. Qi, J. Liu, Adaptive localization through transfer learning in indoor wi-fi environment, in: Seventh International Conference on Machine Learning and Applications, 2008, pp. 331–336.
[18] W. Jiang, E. Zavesky, S.-F. Chang, A. Loui, Cross-domain learning methods for high-level visual concept classification, in: IEEE International Conference on Image Processing, 2008, pp. 161–164.
[19] X. Shi, W. Fan, J. Ren, Actively transfer domain knowledge, in: Machine Learning and Knowledge Discovery in Databases, 2008, pp. 342–357.
[20] Z. Xu, R. Jin, H. Yang, I. King, M. Lyu, Simple and efficient multiple kernel learning by group lasso, in: ICML, 2010, pp. 1175–1182.
[21] H.-S. Seok, K.-B. Hwang, B.-T. Zhang, Feature relevance network-based transfer learning for indoor location estimation, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 41 (5) (2011) 711–719.
[22] D. Kimura, R. Nishimura, A. Oguro, O. Hasegawa, Ultra-fast multimodal and online transfer learning on humanoid robots, in: ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2013, pp. 165–166.
[23] J. Liu, J. Du, X. Wang, Internet tourism scene classification with multi-feature fusion and transfer learning, in: IET International Conference on Communication Technology and Application, 2011, pp. 747–751.
[24] L. Duan, I.W. Tsang, D. Xu, Domain transfer multiple kernel learning, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 465–479.
[25] L. Duan, D. Xu, I.-H. Tsang, J. Luo, Visual event recognition in videos by learning from web data, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1667–1680.
[26] R. Chattopadhyay, W. Fan, I. Davidson, S. Panchanathan, J. Ye, Joint transfer and batch-mode active learning, in: International Conference on Machine Learning (ICML), 2013, pp. 253–261.
[27] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, X. Zhou, Semisupervised feature selection via spline regression for video semantic recognition, IEEE Trans. Neural Netw. Learn. Syst. 26 (2) (2015) 252–264.
[28] Y. Yang, Z. Ma, F. Nie, X. Chang, A.G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, Int. J. Comput. Vis. (2014) 1–15.
[29] Y. Yang, F. Wu, F. Nie, H.T. Shen, Y. Zhuang, A.G. Hauptmann, Web and personal image annotation by mining label correlation with relaxed visual graph embedding, IEEE Trans. Image Process. 21 (3) (2012) 1339–1351.
[30] Y. Han, J. Zhang, Z. Xu, S.-I. Yu, Discriminative multi-task feature selection, in: Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.
[31] Y. Han, Y. Yang, X. Zhou, Co-regularized ensemble for feature selection, in: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, 2013, pp. 1380–1386.
[32] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, X. Zhou, l2, 1-norm regularized discriminative feature selection for unsupervised learning, in: International Joint Conference on Artificial Intelligence (IJCAI), 2011.
[33] Y. Yang, Z. Ma, A.G. Hauptmann, N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks, IEEE Trans. Multimed. 15 (3) (2013) 661–669.
[34] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, M. Meinecke, Pedestrian detection in infrared images, in: IEEE Intelligent Vehicles Symposium, 2003, pp. 662–667.
[35] H. Nanda, L. Davis, Probabilistic template based pedestrian detection in infrared videos, in: IEEE Intelligent Vehicle Symposium, 2002, pp. 15–20.
[36] A. Garcia-Cerezo, A. Mandow, J. Martinez, J. Gómez-de Gabriel, J. Morales, A. Cruz, A. Reina, J. Seron, Development of alacrane: a mobile robotic assistance for exploration and rescue missions, in: IEEE International Workshop on Safety, Security and Rescue Robotics, 2007, pp. 1–6.
[37] M. Andriluka, M. Friedmann, S. Kohlbrecher, J. Meyer, K. Petersen, C. Reinl, P. Schauß, P. Schnitzspan, D. Thomas, A. Vatcheva, et al., Robocuprescue 2009—robot league team darmstadt rescue robot team (germany).
[38] W.L. Fehlman, M.K. Hinders, Mobile robot navigation with intelligent infrared image interpretation, 2009.
[39] G. Cielniak, T. Duckett, A.J. Lilienthal, Data association and occlusion handling for vision-based people tracking by mobile robots, Robot. Auton. Syst. 58 (5) (2010) 435–443.
[40] S. Vidas, P. Moghadam, M. Bosse, 3d thermal mapping of building interiors using an rgb-d and thermal camera, in: IEEE International Conference on Robotics and Automation (ICRA), 2013, pp. 2311–2318.
[41] A. Fernández-Caballero, J.C. Castillo, J. Martínez-Cantos, R. Martínez-Tomás, Optical flow or image subtraction in human detection from infrared camera on mobile robot, Robot. Auton. Syst. 58 (12) (2010) 1273–1281.
[42] G. Lu, N. Sebe, C. Xu, C. Kambhamettu, Memory efficient large-scale image-based localization, Multimed. Tools Appl. 74 (2) (2015) 479–503.
[43] J. Han, E. Pauwels, P. de Zeeuw, Visible and infrared image registration in man-made environments employing hybrid visual features, Pattern Recognit. Lett. 34 (1) (2013) 42–51.
[44] B. Triggs, P. McLauchlan, R. Hartley, A. Fitzgibbon, Bundle adjustment—a modern synthesis, in: Vision Algorithms: Theory and Practice, 2000, pp. 298–372.
[45] L. Duan, D. Xu, I.W. Tsang, Visual event recognition in videos by learning from web data, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[46] E.L. Allwein, R.E. Schapire, Y. Singer, Reducing multiclass to binary: a unifying approach for margin classifiers, J. Mach. Learn. Res. 1 (1) (2000) 113–141.
[47] N. Snavely, S. Seitz, R. Szeliski, Photo tourism: exploring photo collections in 3d, ACM Transit. Graph. 25 (3) (2006) 835–846.
[48] G. Lu, V. Ly, H. Shen, A. Kolagunda, C. Kambhamettu, Improving image-based localization through increasing correct feature correspondences, in: Advances in Visual Computing, 2013, pp. 312–321.
[49] G. Lu, V. Ly, C. Kambhamettu, Structure-from-motion reconstruction based on weighted hamming descriptors, in: International Joint Conference on Neural Networks (IJCNN), 2014, pp. 2367–2374.
[50] G. Lu, Y. Yan, N. Sebe, C. Kambhamettu, Knowing where I am: exploiting multi-task learning for multi-view indoor image-based localization, in: British Machine Vision Conference (BMVC), 2014.
[51] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, N. Sebe, No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classi cation under target motion, in: IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1177–1184.
[52] Y. Yan, E. Ricci, L. Gao, N. Sebe, Egocentric daily activity recognition via multitask clustering, IEEE Transactions on Image Processing (TIP) 24 (10) (2015) 2984–2995.
[53] Y. Yan, E. Ricci, R. Subramanian, L. Gao, N. Sebe, Multi-task linear discriminant analysis for multi-view action recognition, IEEE Transactions on Image Processing (TIP) 23 (12) (2014) 5599–5611.
[54] X. Chang, Y. Yang, Y. Yu, Semantic concept discovery for large-scale zero-shot event detection, in: International Joint Conference on Arti cial Intelligence (IJCAI), 2015.
[55] X. Chang, Y. Yang, Y. Yu, Complex event detection using semantic saliency and nearly-isotonic svm, in: International Conference on Machine Learning (ICML), 2015.

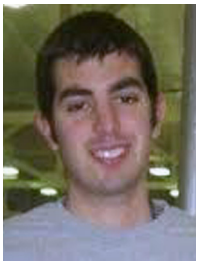
Guoyu Lu is currently pursuing Ph.D. degree in Video/Image Modeling and Synthesis Lab, University of Delaware. He was a student of European Master in Informatics (EuMI) double degree program. He obtained Master degree in Computer Science at University of Trento and Master degree in Media Informatics at RWTH Aachen University. He was a visiting scholar in Auckland University of Technology KEDRI group. He was an intern in Siemens Corporate Research and Bosch Research. His research interest includes Computer Vision, Multimedia Retrieval, and Machine Learning.



Yan Yan received the Ph.D. degree from the University of Trento, Trento, Italy, in 2014, where he is currently a Post-Doctoral Researcher with the Multimedia and Human Understanding Group. His research interests include machine learning and its application to computer vision and multimedia analysis.



Li Ren is currently a Ph.D. degree in University of Delaware, where he also obtained his Master degree in Computer Science. He received his Bachelor degree in Hong Kong Polytechnic University. His research interest includes Computer Vision and Information Retrieval.



Philip Saponaro is currently Ph.D. student in Computer Science in University of Delaware, where he also obtained his Master and Bachelor degree. His research interest includes Computer Vision and Machine Learning.

Nicu Sebe received the Ph.D. degree from Leiden University, Leiden, The Netherlands, in 2001. He is currently with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human behavior understanding. He was the General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference in 2008 and the ACM Multimedia Conference in 2013, and the Program Chair of the International Conference on Image and Video Retrieval in 2007 and 2010, and the ACM Multimedia Conference in 2007 and 2011. He will be the Program Chair of the European Conference on Computer Vision in 2016 and the International Conference on Computer Vision in 2017. He is a Senior Member of the Association for Computing Machinery and a fellow of the International Association for Pattern Recognition.

Chandra Kambhamettu is currently a Professor in the Department of Computer Science, University of Delaware, Newark, where he leads the Video/Image Modeling and Synthesis (VIMS) group. From 1994 to 1996, he was a Research Scientist at the NASA Goddard Space Flight Center (GSFC). His research interests include video modeling and image analysis for biomedical, remote sensing, and multimedia applications. He is best known for his work in motion analysis of deformable bodies, for which he received the NSF CAREER award in 2000. He has published over 200 peer-reviewed papers, supervised ten Ph.D. students and several Masters students in his areas of interest. Dr. Kambhamettu received the Excellence in Research Award from NASA in 1995 while at GSFC. He has served as Area Chair, and has been technical committee member for leading computer vision and medical conferences. He has also served as Associate Editor for the journals Pattern Recognition and Pattern Recognition Letters and the IEEE Transactions on Pattern Analysis and Machine Intelligence.