2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

A Transfer Learning Approach for Multi-Cue Semantic Place Recognition

Gabriele Costante, Thomas A. Ciarfuglia, Paolo Valigi and Elisa Ricci

***Abstract*— As researchers are striving for developing robotic systems able to move into the 'the wild', the interest towards novel learning paradigms for domain adaptation has increased. In the specific application of semantic place recognition from cameras, supervised learning algorithms are typically adopted. However, once learning has been performed, if the robot is moved to another location, the acquired knowledge may be not useful, as the novel scenario can be very different from the old one. The obvious solution would be to retrain the model updating the robot internal representation of the environment. Unfortunately this procedure involves a very time consuming data-labeling effort at the human side. To avoid these issues, in this paper we propose a novel transfer learning approach for place categorization from visual cues. With our method the robot is able to decide automatically if and how much its internal knowledge is useful in the novel scenario. Differently from previous approaches, we consider the situation where the old and the novel scenario may differ significantly (not only the visual room appearance changes but also different room categories are present). Importantly, our approach does not require labeling from a human operator. We also propose a strategy for improving the performance of the proposed method by fusing two complementary visual cues. Our extensive experimental evaluation demonstrates the advantages of our approach on several sequences from publicly available datasets.**

I. Introduction

One of the main research challenges in mobile robotics is to provide robots the capability to move autonomously in real world unconstrained scenarios. In this context, devising novel methods allowing the robots to adapt their internal knowledge and behavior over time is crucial. Therefore, it is clear how lifelong, online and transfer learning techniques are fundamental components for building mobile robot systems operating in highly dynamic environments. While learning and adaptation are of fundamental importance in many mobile robot tasks, including mapping, navigation, or manipulation, in this paper we focus on the specific application of semantic place recognition from visual cues.

Recent works on visual place recognition [1], [2], [3], [4] have demonstrated that good performance can be obtained in many real world indoor settings, even in challenging scenarios (*e.g.* varying illuminations conditions). The vast majority of place recognition approaches, although based on state-of-the-art supervised learning algorithms, makes a rather simplifying assumption: the robot operates in the same scenario where learning has been performed, *i.e.* training and future data are supposed to be drawn from the same distribution. In practice, in real-world applications, this assumption does not hold. To overcome this issue, knowledge transfer approaches are required with the purpose of improving the performance of the learning algorithm by avoiding expensive labeling of data in the novel domain. In recent years, many transfer learning algorithms have been proposed in the machine learning community [5], [6]. However very few works [7], [8] have considered how these methods can be employed for semantic place classification in robotics. Moreover, these works assume that in the novel domain labeled data are available and that the set of categories (*e.g.* room types) does not change during the learning process.

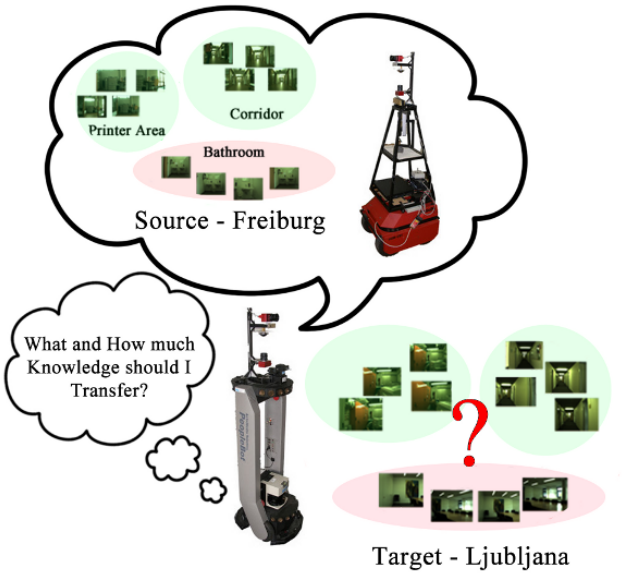


Fig. 1: Illustration of the idea behind the proposed method: the robot performs visual place categorization using images collected in a novel unknown environment (*e.g.* the *target* data) and reasons about transferring knowledge available from a different scenario (*i.e.* the *source* data).

To overcome these limitations, in this paper we present a novel transfer learning approach for semantic place recognition. The main idea behind our method is illustrated in Fig.1. A robot must perform a visual place categorization task (*e.g.* recognizing rooms at university such as offices, corridors, etc) in a novel unknown scenario. As data-labeling is a very time consuming task no information about room categories is available in the new location. Can the robot use data gathered from other sources, *i.e.* images depicting different

G. Costante, T. A. Ciarfuglia, P. Valigi and E. Ricci are with the Department of Electrical and Information Engineering, University of Perugia, Via G. Duranti, 93, 06125, Perugia, Italy {ciarfuglia,valigi,ricci}@diei.unipg.it, gabriele.costante@gmail.com

This work has been partly supported by IIT funds under project HARNESS coordinated by ENEA.

scenarios (*e.g.* rooms of a different university), collected by another robot, to build a robust place recognition system, being able to select only the relevant data and discarding the misleading ones? To answer this question, we propose a new transfer learning approach, allowing a robot to perform place categorization in a novel environment using previous data collected in a different scenario. Importantly the robot is able to automatically determine *what* and *how much* to transfer by quantifying the distance between the distribution of the past and the novel data. The proposed approach is also empowered with a strategy for fusing multiple visual cues, further enhancing recognition accuracy.

There are many concrete scenarios where the proposed risk sensitive transfer approach could make the difference. The RobotEarth platform [9], for instance, aims to collect the robots experiences and share them in the World Wide Web. In this context the robot could easily accesses to a huge amount of data, thus the proposed transfer learning framework gains a great importance. Another possible scenario is the place categorization under dynamic conditions. For instance consider the case of a robot operating under different illumination conditions: the proposed approach permits to adapt the knowledge gathered under daylight conditions for night exploration sessions.

To summarize the main contributions of this paper are: (i) we cast the problem of place categorization in an unknown scenario within a transfer learning framework, (ii) differently from previous works [7], [8], we do not consider a supervised learning approach, thus keeping the human annotation effort as low as possible, (iii) the proposed method permits to quantify the similarity between the data from the original and the novel scenario, thus avoiding the dangerous situations of 'negative transfer' (decrease in recognition accuracy due to the discrepancy between source and target domains), (iv) we show how our transfer learning approach can be extended to integrate informations from multiple modalities, *i.e.* from two different visual cues.

II. Related works

In the last few years several efforts have been made to develop robotic systems with the ability of building robust semantic space representations of environments. In the specific context of place categorization from visual cues, many works have been proposed in the robotics community [1], [2], [3]. Few of them [7], [8], [10] specifically address the domain adaptation challenge. In [7] an algorithm based on support vector machines is proposed for knowledge transfer across two robotic platforms. However, both robots are assumed to perform the same task. Moreover, the robot updates its internal model as new data arrives, discarding progressively the old data and without selecting them according to their importance, *e.g.* the similarity with the novel samples. In [8] a multi-robot transfer learning approach is presented for indoor place categorization which allows the robot to select the type and the amount of information to be transferred. However, both these works [7], [8] rely on supervised learning methods, thus requiring labeled data provided by a human operator. This may be disadvantageous in practical situations where a robot is moved to another scenario, where semantic categories are different from the learned ones and no trained data are available in the new setting. Our approach aims to address this more challenging situation. In [10] a transfer learning approach is proposed for a different application, *i.e.* perceptual classification of objects. Objects are described though an intermediate representation, via a combination of properties derived from different modalities (*e.g.* color, texture, shape). Knowledge transfer between multiple heterogeneous robots is realized by sharing the learned object models.

In the machine learning and the computer vision communities transfer learning techniques have received considerable attention in the last few years [5]. However few works [11], [12] have addressed the problem of transfer learning with different categories in the source and the target data. In this paper, we consider the approach described in [11] and we show how it can be used effectively for semantic place recognition. We also extend the algorithm in [11] to the case where multiple visual cues are used for place categorization. The importance of combining multiple modalities for visual tasks has been demonstrated in several works [2], [13], [14]. For example in the context of semantic place recognition in [2] features extracted from cameras and laser are used for improved performance. However no previous works have addressed this problem into a transfer learning framework.

III. Transfer Learning for Place Recognition

In this section we describe the proposed knowledge transfer approach for place categorization. Suppose that a robot is operating in a completely unknown scenario. In this paper we focus on the task of semantic place recognition in an indoor setting, *i.e.* on the classification of rooms as the robots moves around in a university. While the robot has no *a-priori* information regarding the novel location, we assume that it has access to other data, *e.g.* to videos recorded in other similar scenarios from a different robot, for which labels are available. Can the robot use these data and importantly decide autonomously if these data are useful for the current task, *i.e.* how much the past video sequences are similar to those it observes in the the current scenario?

The main intuition behind out approach articulates in two main steps. First it computes the similarity between the two distributions of images. This step aims to understand if the two locations have similar appearances and sets the risk accordingly to the divergence measure. Then, if the transfer risk is small, in the second step, we can take advantage of the category constraints from the labeled data to perform clustering in the novel scenario and we include them in the optimization function. Otherwise, if the risk is high, we only rely on the images from the new location.

This problem can be formalized as follows. We are given a set $\mathcal{S} = \{(\mathbf{x}_1^s, y_1^s), (\mathbf{x}_2^s, y_2^s), \ldots, (\mathbf{x}_{N_s}^s, y_{N_s}^s)\}$ (the *source data*), where $\mathbf{x}_i^s \in I\!R^D$ are visual features extracted from video frames and $y_i^s \in \{1, 2, \ldots, K_C^S\}$ are the corresponding labels indicating the rooms types (*e.g.* corridor, office, etc),

and a set $\mathcal{T} = \{\mathbf{x}_1^t, \mathbf{x}_2^t, \ldots, \mathbf{x}_{N_t}^t\}$ (the *target* data), where $\mathbf{x}_i^t \in I\!R^D$ are visual features extracted in the novel scenario for which labels are not available. We are interested in learning a model in order to classify the target data. Note that the categories of the target data are not the same of the K_C^S classes in $\mathcal{S}$. As the target and the source data belongs to different distributions, respectively $\mathcal{P}_\mathcal{S}$ and $\mathcal{P}_\mathcal{T}$, we would also like to measure the distance between them in order to quantify the risk of knowledge transfer, *i.e.* of using the source data to build a suitable model for the target data.

A. Clustering-based Transfer Risk

To measure the distance between the source and target distributions $\mathcal{P}_\mathcal{S}$ and $\mathcal{P}_\mathcal{T}$ in this paper we adopt the method proposed in [11]. A popular approach to quantify the distance between two probability distributions is the Kullback-Leibler (KL) divergence, defined as:

$$KL(\mathcal{S},\mathcal{T}) = \sum_x P_\mathcal{T}(x) \log \frac{P_\mathcal{T}(x)}{P_\mathcal{S}(x)} \tag{1}$$

As calculating the KL divergence directly from the data can be quite time consuming, in [11] a more practical solution is proposed, where an approximation of KL distance is computed based on the output of a clustering algorithm operating on the combined data (source and target data together). More specifically the following definition of *Clustering-based KL divergence* is proposed:

$$KL_c(\mathcal{S},\mathcal{T}) = \frac{2}{|\mathcal{T}|} \sum_{c=1}^{|\mathcal{C}|} \left(\frac{|\mathcal{T} \cap C_c|}{|C_c|} \log \frac{|\mathcal{T} \cap C_c|}{|\mathcal{S} \cap C_c|}\right) + \log \frac{|\mathcal{S}|}{|\mathcal{T}|} \tag{2}$$

where $\forall c$ the centroid of the data from $\mathcal{T}$ corresponding to cluster C_c and the centroid of the data from $\mathcal{S}$ corresponding to cluster c are the same. The computation of the clustering-based KL divergence is illustrated in Fig.2. Due to lack of space we refer to the original paper for details on the derivation of (2). Having computed the distance between distributions $\mathcal{P}_\mathcal{S}$ and $\mathcal{P}_\mathcal{T}$, the risk of transferring source data information while learning from target data is defined as:

$$R_{\mathcal{S},\mathcal{T}} = \frac{1}{1 + e^{(\gamma - KL_c(\mathcal{S},\mathcal{T}))}} \tag{3}$$

where γ is a fixed parameter which is set equal to e^2 in our experiments.

B. Transfer Learning with Different Class Labels

The transfer learning approach we adopt in this paper is an extension of the Normalized-Cut algorithm [15]. It amounts into solving the following optimization problem:

$$\min_{\mathbf{U}} \frac{\mathbf{U}^T\mathbf{L}\mathbf{U}}{\mathbf{U}^T\mathbf{D}\mathbf{U}} + \beta((1 - R_{\mathcal{S},\mathcal{T}})\|\mathbf{M}_\mathcal{S}\mathbf{U}\|^2 + R_{\mathcal{S},\mathcal{T}}\|\mathbf{M}_\mathcal{T}\mathbf{U}\|^2) \tag{4}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix, $\mathbf{W}$ is the similarity matrix computed on the entire dataset $\mathcal{S} \cup \mathcal{T}$, $\mathbf{D} = \text{diag}(\mathbf{We})$ and $\mathbf{e}$ is a vector with all the coordinates set to 1. The matrix $\mathbf{M}_\mathcal{S} = [\mathbf{m}_1\mathbf{m}_2 \ldots \mathbf{m}_{N_c^s}]^T$ where $\mathbf{m}_i \in I\!R^{N_s+N_t}$ is a vector with 1 in the i-th position and -1 in the j-th position if the source data points $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same labels. The matrix $\mathbf{M}_\mathcal{T}$ is similarly defined on the target data. However, as for the target data labels are not provided, a preprocessing phase where the target data are clustered with Normalized-Cut [15] is performed. The matrix $\mathbf{M}_\mathcal{T}$ is then defined using as labels the vectors indicating the cluster membership.

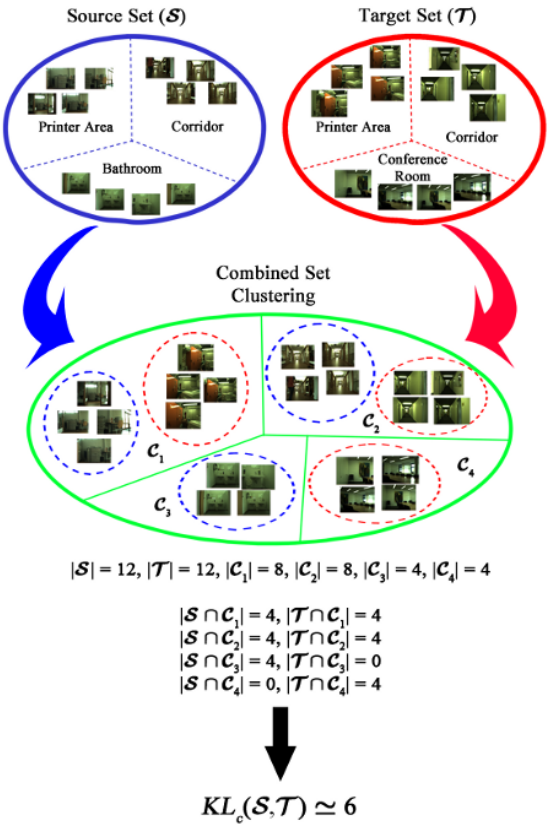


Fig. 2: KL divergence computation. $|\mathcal{S}|$ and $|\mathcal{T}|$ represent respectively the cardinality of the source and the target data set, while $|\mathcal{C}_c|$ indicates the size of cluster c. The term $|\mathcal{S} \cap \mathcal{C}_c|$ ($|\mathcal{T} \cap \mathcal{C}_c|$) represents the cardinality of the intersection between the source (or target) set and the cluster c.

The objective function in (4) is the sum of two terms. The first term simply aims to cluster the entire dataset using Normalized-Cut, while the second term enforces that the learned clustering structure respect some constraints. More specifically two sets of constraints are imposed. One guarantees that the learned projection matrix leads to clusters consistent with the labels of the source data. The second set of constraints imposes some coherence between the novel clustering results and those that are obtained only grouping the target data. The trade-off between transferring source data information and not using it is regulated by the risk $R_{\mathcal{S},\mathcal{T}}$.

Defining the matrix $\mathbf{A} = \mathbf{L} + \beta((1 - R_{\mathcal{S},\mathcal{T}})\mathbf{M}_\mathcal{S}^T\mathbf{M}_\mathcal{S} + R_{\mathcal{S},\mathcal{T}}\mathbf{M}_\mathcal{T}^T\mathbf{M}_\mathcal{T})$ the optimization problem (4) can be reformulated as follows:

$$\min_{\mathbf{U}} \frac{\mathbf{U}^T\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{U}}{\mathbf{U}^T\mathbf{U}} \tag{5}$$

The details of the derivation can be found in the original paper [11]. The resulting transfer learning method is presented

Algorithm 1 Transfer Learning algorithm

Input: source data $\mathcal{S}$, target data $\mathcal{T}$, number of target categories K_C^T, total number of categories K_C, β

procedure COMPUTERISK($\mathbf{W}$, K_C, $\mathcal{S}$, $\mathcal{T}$)
Set $\mathbf{D}$ with $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$
Set $\mathbf{L}$ = $\mathbf{D}$ - $\mathbf{W}$
$\mathbf{U}$ = eig($\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$, K_C)
$\mathbf{U} = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}$
Normalize $\mathbf{U}$ by row where $\mathbf{U}_{ij} = \mathbf{U}_{ij}/\sqrt{\sum_{l=1}^{K_C} \mathbf{U}_{il}^2}$
$\mathcal{C}$ = kmeans($\mathbf{U}$, K_C)
Compute $R_{\mathcal{S},\mathcal{T}}$ using (2) and (3)
end procedure

$\mathbf{W}$ = computeSimilarityMatrix($\mathcal{S}$, $\mathcal{T}$)
$\mathbf{M}_{\mathcal{S}}$ = computeSourceConstraints(y^s)
$R_{\mathcal{S},\mathcal{T}}$ = computeRisk($\mathbf{W}$, K_C, $\mathcal{S}$, $\mathcal{T}$)
Set $\mathbf{D}$ with $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$
Set $\mathbf{L}$ = $\mathbf{D}$ - $\mathbf{W}$
$\mathbf{M}_{\mathcal{T}}$ = computeTargetConstraints($\mathbf{W}$, K_C^T)
$\mathbf{A} = \mathbf{L} + \beta((1 - R_{\mathcal{S},\mathcal{T}})\mathbf{M}_{\mathcal{S}}^T\mathbf{M}_{\mathcal{S}} + R_{\mathcal{S},\mathcal{T}}\mathbf{M}_{\mathcal{T}}^T\mathbf{M}_{\mathcal{T}})$
$\mathbf{U}$ = eig($\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, K_C^T).
$\mathbf{U} = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}$
Normalize $\mathbf{U}$ by row where $\mathbf{U}_{ij} = \mathbf{U}_{ij}/\sqrt{\sum_{l=1}^{K_C^T} \mathbf{U}_{il}^2}$
$\mathcal{C}$ = kmeans($\mathbf{U}$, K_C^T)
Output: Target set label $\mathcal{C}$

in Algorithm (1).

C. Transfer Learning with Complementary Visual Cues

In this paper we extend the transfer learning approach proposed in [11] and described in the previous section in order to operate with two complementary visual cues. In particular in the context of indoor place recognition we adopt two different descriptors: the spatial pyramid matching kernel (SPMK) originally proposed in [16] and the Spatial Principal component Analysis of Census Transform histograms (SPACT) descriptor [17]. A detailed description of the adopted features can be found in the following section.

Given $\mathbf{L}_S = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{W}_S\mathbf{D}_S^{-\frac{1}{2}}$ and $\mathbf{L}_C = \mathbf{D}_C^{-\frac{1}{2}}\mathbf{W}_C\mathbf{D}_C^{-\frac{1}{2}}$, where $\mathbf{W}_S$ and $\mathbf{W}_C$ are respectively the SPMK and the SPACT kernels and $\mathbf{D}_S = \text{diag}(\mathbf{W}_S\mathbf{e})$, $\mathbf{D}_C = \text{diag}(\mathbf{W}_C\mathbf{e})$, the problem of transfer learning can be formulated as follows:

$$\begin{aligned} \max_{\mathbf{U}_S,\mathbf{U}_C} \quad & \textstyle\sum_{i\in\{S,C\}} tr(\mathbf{U}_i^T\mathbf{B}_i\mathbf{U}_i) + \lambda\mathcal{A}(\mathbf{U}_S,\mathbf{U}_C) \\ \text{s.t.} \quad & \mathbf{U}_S^T\mathbf{U}_S = I\ ,\ \mathbf{U}_C^T\mathbf{U}_C = I \end{aligned} \tag{6}$$

with:

$$\mathbf{B}_S = \mathbf{L}_S - \beta_S(1 - R_{\mathcal{S},\mathcal{T}}^S)\mathbf{M}_{\mathcal{S}}^T\mathbf{M}_{\mathcal{S}} + R_{\mathcal{S},\mathcal{T}}^S\mathbf{M}_{\mathcal{T}_S}^T\mathbf{M}_{\mathcal{T}_S} \tag{7}$$

$$\mathbf{B}_C = \mathbf{L}_C - \beta_C(1 - R_{\mathcal{S},\mathcal{T}}^C)\mathbf{M}_{\mathcal{S}}^T\mathbf{M}_{\mathcal{S}} + R_{\mathcal{S},\mathcal{T}}^C\mathbf{M}_{\mathcal{T}_C}^T\mathbf{M}_{\mathcal{T}_C} \tag{8}$$

where λ is an appropriate regularization parameter and $\mathcal{A}(\mathbf{U}_S,\mathbf{U}_C)$ is the agreement term between the two views defined as follows:

$$\mathcal{A}(\mathbf{U}_S,\mathbf{U}_C) = tr(\mathbf{U}_S\mathbf{U}_S^T\mathbf{U}_C\mathbf{U}_C^T) \tag{9}$$

In practice the proposed optimization problem (6) is a sum of two main terms. The first aims to reason about transferring knowledge from source data separately for each modality, the second is meant to impose consistency between the two projected eigenspaces. To solve this problem efficiently an alternating optimization approach is adopted, solving separately for $\mathbf{U}_S$ and $\mathbf{U}_C$. In particular for a given $\mathbf{U}_C$ we get:

$$\begin{aligned} \max_{\mathbf{U}_S} \quad & tr\left\{\mathbf{U}_S^T(\mathbf{B}_S + \lambda\mathbf{U}_C\mathbf{U}_C^T)\mathbf{U}_S\right\} \\ \text{s.t.} \quad & \mathbf{U}_S^T\mathbf{U}_S = I \end{aligned} \tag{10}$$

which can be easily solved using spectral decomposition methods. Similarly when $\mathbf{U}_S$ is fixed an analogous problem must be solved with respect to $\mathbf{U}_C$. The main steps of the proposed multi-cue transfer learning method are shown in Algorithm (2).

Algorithm 2 Multi-Cue Transfer Learning

Input: source data $\mathcal{S}$, target data $\mathcal{T}$, number of target categories K_C^T, total number of categories K_C, β_S, β_C, λ, number of iteration T

$\mathbf{W}_S$ = computeSPMKernel($\mathcal{S}$, $\mathcal{T}$)
$\mathbf{W}_C$ = computeCENTRISTKernel($\mathcal{S}$, $\mathcal{T}$)
$\mathbf{M}_{\mathcal{S}}$ = computeSourceConstraints(y^s)
$R_{\mathcal{S},\mathcal{T}}^S$ = computeRisk($\mathbf{W}_S$, K_C, $\mathcal{S}$, $\mathcal{T}$)
$R_{\mathcal{S},\mathcal{T}}^C$ = computeRisk($\mathbf{W}_C$, K_C, $\mathcal{S}$, $\mathcal{T}$)
$\mathbf{M}_{\mathcal{T}_S}$ = computeTargetConstraints($\mathbf{W}_S$, K_C^T)
$\mathbf{M}_{\mathcal{T}_C}$ = computeTargetConstraints($\mathbf{W}_C$, K_C^T)
Compute $\mathbf{B}_S$ and $\mathbf{B}_C$ using (7) and (8)
$\mathbf{U}_S$ = eig($\mathbf{B}_S$, K_C^T).
for $t = 1, \ldots, T$ **do**
$\mathbf{U}_C$ = eig($\mathbf{B}_C + \lambda\mathbf{U}_S\mathbf{U}_S^T$, K_C^T).
$\mathbf{U}_S$ = eig($\mathbf{B}_S + \lambda\mathbf{U}_C\mathbf{U}_C^T$, K_C^T).
endfor
Normalize $\mathbf{U}_S$ and $\mathbf{U}_C$
$\mathcal{C}$ = kmeans($[\mathbf{U}_S\ \mathbf{U}_C]$, K_C^T)
Output: Target set label $\mathcal{C}$

IV. EXPERIMENTAL RESULTS

A. Datasets

To demonstrate the effectiveness of our approach in different scenarios with several place categories, we select sequences from three datasets: the COLD [18], the KTH-IDOL2 [19] and the VPC [20] datasets.

The first datasets consists of several video sequences gathered in three indoor university environments of different European cities: Freiburg, Ljubljana and Saarbrücken. The video sequences have been collected using three different robotic platforms (an ActivMedia People Bot, an ActiveMedia Pioneer-3 and an iRobot ATRV-Mini) with two Videre Design MDCS2 digital cameras to obtain perspective and omnidirectional views. Each frame is registered with the associated absolute position recovered using laser and odometry data and annotated with a label representing the corresponding place. The acquisition was performed in several rooms of different functionality, under different illumination conditions (cloudy, night and sunny). Each dataset has some place category in common with the other datasets, *e.g.*

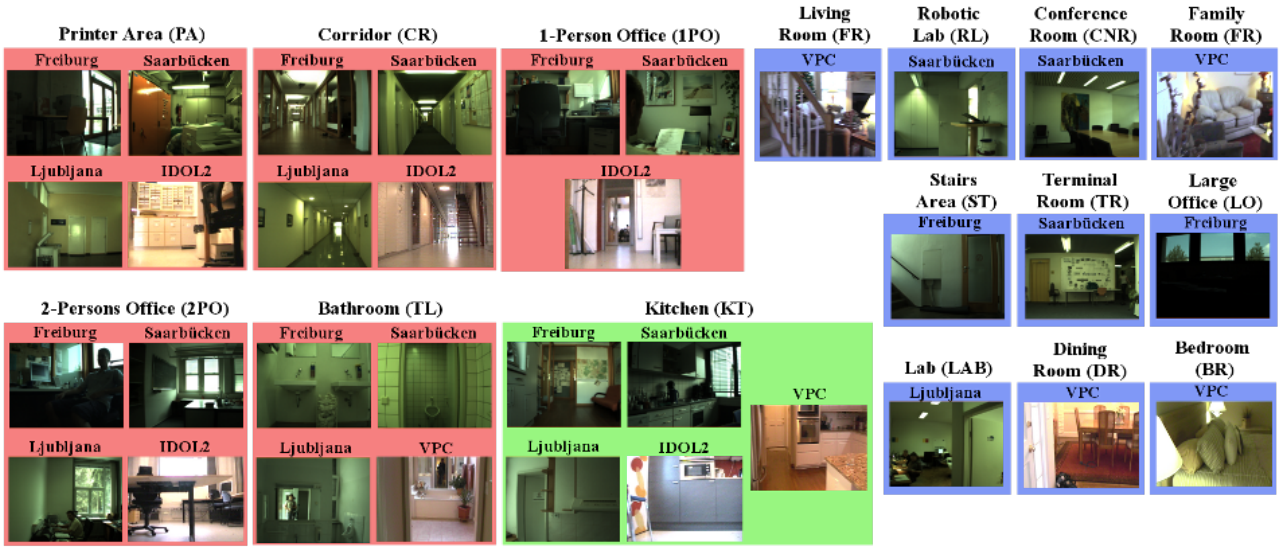


Fig. 3: Sample images for all the place categories extracted in the three datasets used. Red rectangles group together room categories belonging only to a subset of datasets, the green rectangles show the categories which are common to all the datasets and the blue ones highlight specific rooms in each database.

Corridor (CR), Printer Area (PA) or Bathroom (TL), but also contains dataset specific rooms, *e.g.* the Robotics Lab in the Saarbrücken sequences or the Stairs Area in the Freiburg data. Moreover rooms of different datasets associated to the same labels may have very different appearance. An example is the Corridor (CR) class: the separating walls between offices in the Freiburg data are made of glass, while in the Saarbrücken and Ljubljana sequences concrete walls are depicted. Therefore, transfer learning is very challenging.

The IDOL2 dataset is similar to COLD: it contains several image sequences recorded under various weather and illumination conditions. The acquisition was performed in an indoor environment that contains five types of rooms: One-Person Office (OO), Two-person Office (TO), Corridor (CR), Kitchen (KT) and Printer Area (PA). The robotic platforms used were a MobileRobots PeopleBot and a PowerBot equipped with a Canon VC-C4 camera.

Finally the VPC dataset consists of several sequences collected in six houses with different room categories. The dataset was recorded using a camera (JVC GR-HD1) mounted on a mobile tripod.

Figure 3 shows some sample images for all the place categories of the considered datasets.

B. Experimental Setup

To properly evaluate the performance of our method, we choose sequences extracted from all the four datasets and recorded at different illumination conditions. In every experiment we select sequences where the source and the target data have different place categories. We only require that they have at least one specific room type (one class) in common. This is meant to show the validity of our method which operates in the realistic situation where transferring knowledge across different scenarios and determining automatically *how much* to transfer is essential.

We compute two different set of visual features, one based on the state of the art Spatial Pyramid Matching Kernel scheme proposed by [16], the other based on the more recent SPACT descriptor [21]. The SPMK representation has been shown to be very effective and has been widely used for place recognition applications in the context of robotic systems. Specifically the pyramid matching strategy works by dividing the image into a set of increasingly coarser grids and computing a weighted sum of the matches that occurs at each level. Two points are said to match if they are in the same cell, given a certain resolution. According to this scheme the matching kernel is computed calculating the histogram intersection between the vectors formed by concatenating the weighted histograms at all resolutions. More specifically we use the SIFT descriptors [22] to extract interest points from images. Then we create a vocabulary of 400 visual words following the standard Bag-of-Words approach using 800 images as training set. Finally the histograms for each image are constructed projecting the extracted SIFT in the vocabulary at each level of resolution and for each cell. We choose $L = 3$ as the number of pyramid levels. The similarity matrix $\mathbf{W}$ is then obtained computing histogram intersection.

The CENTRIST descriptor [21] was originally proposed for scene classification tasks and has been shown to be very effective as it captures the structural properties of the scenes. The Census Transform is a nonparametric local transform introduced to compare local patches. It compares the intensity of a pixel with its eight neighbors and the binary values obtained replaces the pixel itself. Thus the CENTRIST descriptor has 256 bins where each bin counts the occurrences of a value in the range [0 255] after the application of the Census Transform to the entire image. Following the approach in [21] to obtain our final descriptor we also apply the spatial-pyramid [16] to capture the global structure of the image at a large scale and Principal

Component Analysis (PCA) to reduce the dimensionality of histograms and obtain a more compact representation. Specifically we set the number of pyramid levels to $L = 3$ and the number of principal components equal to 40. The final descriptor is called SPACT (spatial Principal component Analysis of Census Transform histograms). After computing the SPACT histograms, we use the RBF kernel to calculate the similarity matrix $\mathbf{W}$.

In our experiments we first tested the proposed transfer learning approach using a single visual cue. We perform experiments both for the SPMK and the SPACT descriptors. We also compare our approach against two baselines: a *No-Transfer* method which applies a clustering algorithm (specifically Normalized Cut [15]) to the union of the source and the target data and a *Full Transfer* algorithm where the knowledge gathered from source is completely transferred in the target *i.e.* without considering the risk of transferring potentially harmful information. This situation is obtained setting $R(\mathcal{S}, \mathcal{T}) = 0$.

A second series of experiments aim to test the proposed multi-cue approach. As baselines we again consider the *No-Transfer* and *Full Transfer* methods. In this case the two visual cues are simply combined taking the average of the two computed kernels. In both single-cue and multi-cues tests the parameters β in Eqn. 4 and β_S and β_C in Eqn. 6 are set to 1. Since the output of our algorithm consists into a set of clusters representing place categories we measure the performance in terms of clustering accuracy [23]:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N_T} \delta(y_i, map(c_i))}{N_T} \tag{11}$$

where N_T is the total number of images on target data, y_i is the true label for the i-th image, c_i is the cluster label. $\delta(y, c)$ is a function that is 1 if true label and cluster label are the same and 0 otherwise and $map(\cdot)$ is a permutation function that maps cluster labels to true labels. The optimal matching is found using the Hungarian algorithm [24]. Due to the variability introduced by the k-means algorithm, we repeat the clustering step after the spectral decomposition 10 times. The resulting average accuracy is considered.

C. Quantitative Evaluation

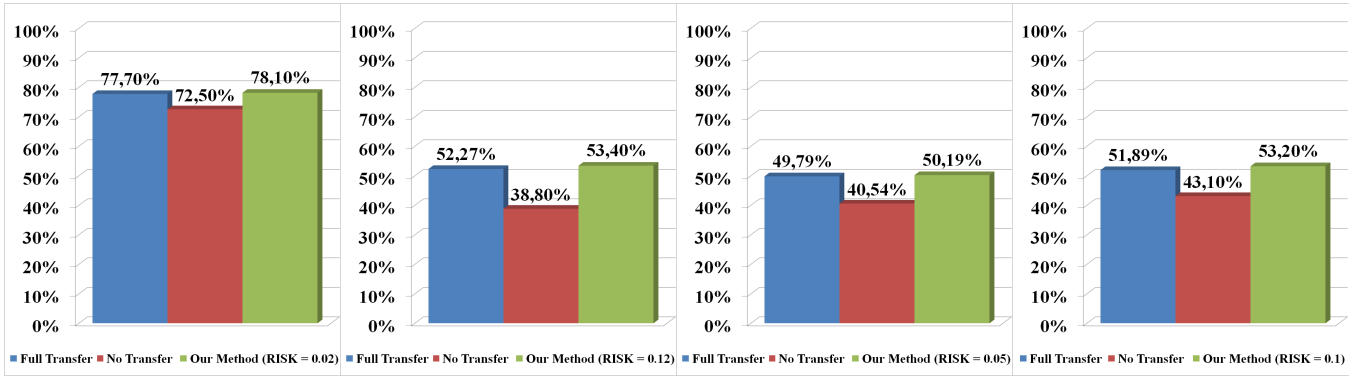
In a first series of experiments we show some place recognition results on target sequences using a single visual cue. We perform several tests on various sequences extracted from all the three datasets. Our aim here is to demonstrate the capability of the proposed method to understand what to transfer, avoiding negative transfer and maximizing the use of information gathered from the source data.

Figures 4 and 5 show the obtained results respectively for the SPMK and SPACT experiments. The labels of the source data provided in the ground truth files in the datasets are used to specify the set of source constraints and define $\mathbf{M}_S$ in (4). We compare our method against the Full-Transfer and the No-Transfer algorithms. It is evident that our transfer learning approach outperforms the two baselines in almost all the experiments. For example in cases where transferring knowledge is helpful the KL divergence is very low and the transfer risk is correctly set to a value close to zero. In the Freiburg-Freiburg experiments (Fig. 4.a) the No-Transfer algorithm achieves an accuracy of 72.50% while the Full-Transfer approach reaches 77.7% with the Spatial Pyramid Matching Kernel features. Similar results are obtained in case of the SPACT features where they get respectively 57.2% and 59.39% (Fig. 5.a). In both configurations, since the source and the target distributions are similar, (*i.e.* rooms have similar visual appearance) the transfer risk is close to zero and our strategy correctly determines that the source knowledge helps clustering the target data: we get 78.1% with the SPMK features and 62.04% with SPACT. In the Freiburg-IDOL2 and Freiburg-Saarbrücken experiments we observe similar results: in the first experiment (Fig.4.b and Fig.5.b) we obtain 53.40% with SPMK features and 68.50% with the SPACT features while the No-Transfer case only reaches 38.8% and 46.70%; in the second test (Fig.4.c and Fig.5.c) our approach correctly classify 50.19% of the frames with the first type of features and 57.65% with the second one, while the No-Transfer algorithm gets 40.54% and 42.60% respectively. We perform a further test on the Freiburg-Saarbrücken sequences (Fig.4.d and Fig.5.d) by changing the illumination conditions of the target sequence in order to prove the robustness of our method: even in this case we improve the clustering accuracy by more than 10% with respect to the No-Transfer algorithm with both set of features.

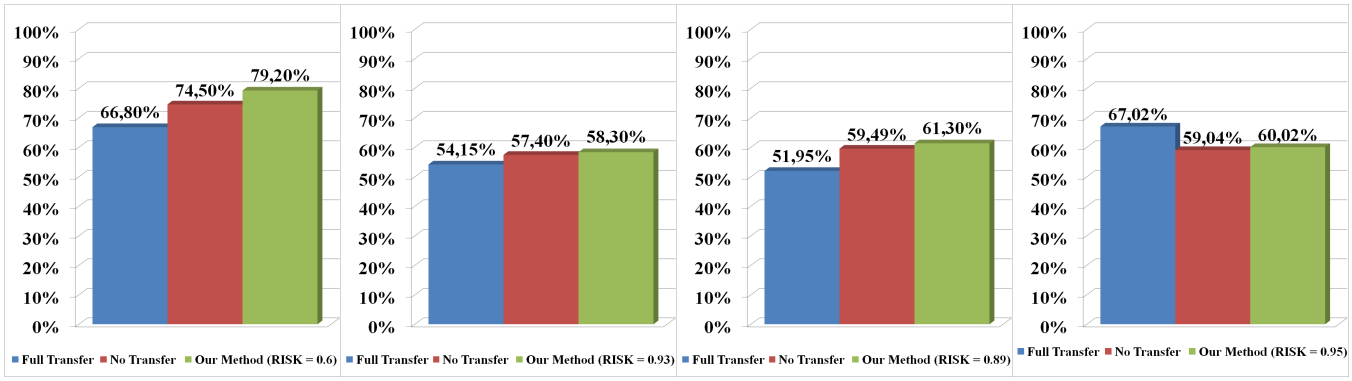
Tests on the Ljubljana-Saarbrücken (Fig.4.f and Fig.5.f) and VPC-Freiburg (Fig.4.g and Fig.5.g) sequences show how our distribution sensitive method avoids negative transfer. In these cases the place categories in the source and the target data are very different so transferring knowledge from the source may worsen clustering performance on target data. The KL divergence detects this situation and set the transfer risk to a value close to 1. Due to this effect, in the first sequence we get 58.30% and 63.40% while the Full-Transfer algorithm reaches 54.15% and 51.20% respectively for the SPMK and SPACT kernels; in the second sequence we obtain 61.30% and 64.50% against 51.95% and 48.93%. The proposed approach also outperforms the No-Transfer baseline, meaning that a small amount of information from the source data could be effectively used for improving the performance on clustering target data. This is reflected by the value of the risk which is slightly lower than 1.

The test on the Ljubljana-Freiburg sequences (Fig.4.e and Fig.5.e) considers another possible scenario. Here the source and the target sequences share some categories, while the other are deeply different. In this case the transfer risk has on an intermediate value between 0 and 1. In this case our approach obtains a higher clustering accuracy with respect to both the No-Transfer and the Full-Transfer methods.

Finally Fig.4.h and Fig.5.h show an example where our method fails. The computation of the clustering-based KL divergence is not accurate and a risk close to 1 is obtained despite the Saarbrücken and the Freiburg sequences share some similar patterns. This is probably due to a large number of categories. In this case the Full-transfer method outperforms both our approach and the No-transfer algorithm.

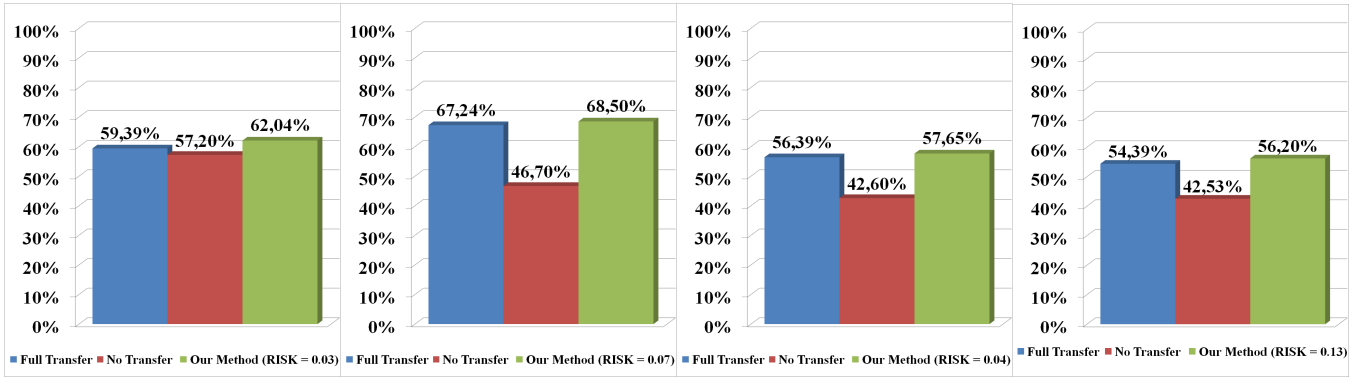


(a) Fr.(Std A) c.1-Fr.(Ext A) c.1 (b) Fr.(Std A) c.1-IDOL2 Minnie c.2 (c) Fr.(Std A) c.1-Saar.(Std A) c.2 (d) Fr.(Std A) c.1-Saar.(Ext A) n.1

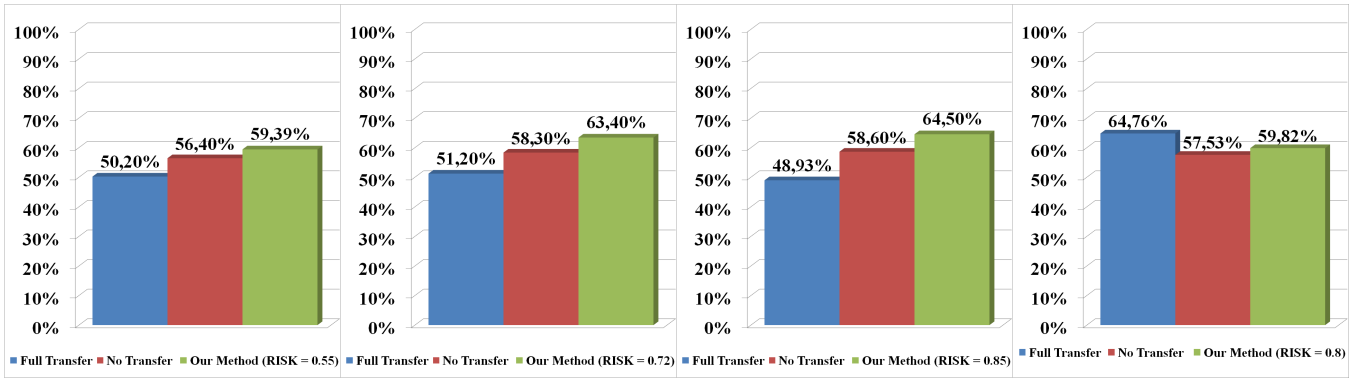


(e) Ljub.(Std A) s.1-Fr.(Std A) c.1 (f) Ljub.(Std A) s.1-Saar.(Ext A) n.1 (g) VPC 1 Floor 2-Fr.(Std A) cloudy1 (h) Saar.(Ext A) s.2-Fr.(Ext A) c.1

Fig. 4: Place recognition accuracy obtained with SPMK features.



(a) Fr.(Std A) c.1-Fr.(Ext A) c.1 (b) Fr.(Std A) c.1-IDOL2 Minnie c.2 (c) Fr.(Std A) c.1-Saar.(Std A) c.2 (d) Fr.(Std A) c.1-Saar.(Ext A) n.1



(e) Ljub.(Std A) s.1-Fr.(Std A) c.1 (f) Ljub.(Std A) s.1-Saar.(Ext A) n.1 (g) VPC 1 Floor 2-Fr.(Std A) c.1 (h) Saar.(Ext A) s.2-Fr.(Ext A) c.1

Fig. 5: Place recognition accuracy obtained with SPACT features.

TABLE I: Place recognition accuracy obtained with a single visual cue and with cues combination.

	SPMK		SPACT		Cues Combination
Freiburg(StdA) cloudy 1 - Freiburg(ExtA) cloudy 1	Risk = 0.02	78.1 % ± 0.1	Risk = 0.03	62.04 % ± 0.11	**82,67 % ± 0.06**
Freiburg(StdA) cloudy 1 - IDOL2 Minnie cloudy 2	Risk = 0.12	53.40 % ± 0.2	Risk = 0.07	68.50 % ± 0.25	**73,13 % ± 0.05**
Freiburg(StdA) cloudy 1 - Saarbrücken(StdA) cloudy 2	Risk = 0.05	50.19 % ± 0.2	Risk = 0.04	57.65 % ± 0.1	**61,27 % ± 0.03**
Freiburg(StdA) cloudy 1 - Saarbrücken(ExtA) night 1	Risk = 0.1	53.20 % ± 0.1	Risk = 0.13	56.20 % ± 0.12	**60,89 % ± 0.04**
Ljubljana(StdA) sunny 1 - Freiburg(StdA) cloudy 1	Risk = 0.6	79.20 % ± 0.19	Risk = 0.55	59.30 % ± 0.14	**85.65 % ± 0.07**
Ljubljana(StdA) sunny 1 - Saarbrücken(ExtA) night 1	Risk = 0.93	58.30 % ± 0.16	Risk = 0.72	63.40 % ± 0.13	**69.61 % ± 0.05**
VPC Home 1 Floor 2 - Freiburg(StdA) cloudy 1	Risk = 0.89	61.30 % ± 0.09	Risk = 0.85	64.50 % ± 0.08	**72.3 % ± 0.07**
Saarbrücken(ExtA) sunny 2 - Freiburg(ExtA) cloudy 1	Risk = 0.95	60,02 % ± 0.09	Risk = 0.8	59,82 % ± 0.11	**62,43 % ± 0.12**

In a second series of experiments we aim to test our distribution sensitive transfer learning approach with complementary visual cues. After the computation of the transfer risk for both SPMK and SPACT features, we aim to combine their contributions to improve performance with respect to the single visual features tests. The value of the parameter λ in Eqn. (6) is set to 0.5 in all our experiments. We compare the multiple cue approach with the single features one. Results are shown in Table I. It is evident how the multi-cue strategy is beneficial for place recognition accuracy. For example in the first experiment (Freiburg(Std A) cloudy 1 - Freiburg(Ext A) cloudy 1) an accuracy of 78.1% and 62.04% is obtained using the SPMK and the SPACT features respectively, while with the proposed multiple cue transfer learning strategy the accuracy increases to a value of 82.67%.

It is worth nothing that the recognition accuracy obtained with the proposed approach is generally lower with respect to that we can get using supervised learning approaches (*e.g.* Support Vector Machines) considered in previous works [2], [7]. However we point out that we operate in a more challenging scenario where the categories in the source and in the target data are different and no labels are provided for the target data. In this way no human intervention for annotating data is required.

V. Conclusions

We presented a novel transfer learning approach for semantic place recognition which operates integrating two complementary visual cues. Our risk sensitive transfer algorithm allows the robot to perform semantic place categorization incorporating in the learning process the previous knowledge collected in different environments. Importantly, with the proposed strategy, we avoid the negative transfer by measuring the similarity among the source and the target data distribution. Our extensive evaluation demonstrates that our approach significantly outperforms the baselines. Future works include developing a more robust approach in alternative to the KL distance to measure the transfer risk and extending the proposed framework to operate in an incremental manner without having all target data available at the beginning and with more than two sets of features, eventually including RGB-D data.

References

[1] A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *ICRA*, 2012, pp. 3515–3522.

[2] A. Pronobis, Ó. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal Semantic Place Classification," *International Journal of Robotics Research (IJRR)*, vol. 29, no. 2-3, pp. 298–320, 2010.

[3] M. M. Ullah, A. Pronobis, B. Caputo, J. Luo, P. Jensfelt, and H. I. Christensen, "Towards robust place recognition for robot localization," in *ICRA*, 2008, pp. 530–537.

[4] T. A. Ciarfuglia, G. Costante, P. Valigi, and E. Ricci, "A discriminative approach for appearance based loop closing," in *IROS*, 2012.

[5] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345 –1359, oct. 2010.

[6] A. K. Rajagopal, S. Ramanathan, R. L. Vieriu, E. Ricci, O. Lanz, K. Ramakrishnan, and N. Sebe, "An Adaptation Framework for Head-Pose Classification in Dynamic Multi-view Scenarios," in *ACCV*, 2012, pp. 652–666.

[7] J. Luo, A. Pronobis, and B. Caputo, "SVM-based transfer of visual knowledge across robotic platforms," in *Proc. ICVS*, 2007.

[8] S. Prasath Elango, T. Tommasi, and B. Caputo, "Transfer Learning of Visual Concepts across Robots: a Discriminative Approach," Idiap, Tech. Rep. Idiap-RR-06-2012, 1 2012.

[9] M. Waibel, M. Beetz, J. Civera, R. D'Andrea, J. Elfring, D. Galvez-Lopez, K. Haussermann, R. Janssen, J. Montiel, A. Perzylo, B. Schiessle, M. Tenorth, O. Zweigle, and R. van de Molengraft, "Roboearth," *Robotics Automation Magazine, IEEE*, vol. 18, no. 2, pp. 69–82, 2011.

[10] Z. Kira, "Inter-robot transfer learning for perceptual classification," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*, 2010, pp. 13–20.

[11] X. Shi, W. Fan, Q. Yang, and J. Ren, "Relaxed Transfer of Different Classes via Spectral Partition," in *Proc. ECML-PKDD*, 2009.

[12] W. Yang, Y. Wang, and G. Mori, "Efficient Human Action Detection Using a Transferable Distance Function," in *ACCV*, 2009.

[13] F. Orabona, J. Luo, and B. Caputo, "Online-batch strongly convex multi kernel learning," in *CVPR*, 2010, pp. 787–794.

[14] S. Duffner, J.-M. Odobez, and E. Ricci, "Dynamic partitioned sampling for tracking with discriminative features," in *BMVC*, 2009.

[15] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *CVPR*, 2006.

[17] J. Wu and J. M. Rehg, "CENTRIST: A Visual Descriptor for Scene Categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, 2011.

[18] A. Pronobis and B. Caputo, "COLD: COsy Localization Database," *International Journal of Robotics Research (IJRR)*, vol. 28, no. 5, pp. 588–594, may 2009.

[19] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "Incremental learning for place recognition in dynamic environments," in *IROS*, 2007.

[20] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual place categorization: Problem, dataset, and algorithm," in *IROS*, 2009, pp. 4763–4770.

[21] J. Wu and J. Rehg, "Centrist: A visual descriptor for scene categorization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1489 –1501, 2011.

[22] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999.

[23] M. Wu and B. Schlkopf, "A local learning approach for clustering," in *NIPS*, 2006.

[24] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1982.