Hindawi Publishing Corporation
Advances in Multimedia
Volume 2013, Article ID 175064, 22 pages
http://dx.doi.org/10.1155/2013/175064

Research Article

Multiple Feature Fusion Based on Co-Training Approach and Time Regularization for Place Classification in Wearable Video

Vladislavs Dovgalecs, Rémi Mégret, and Yannick Berthoumieu

IMS Laboratory, University of Bordeaux, UMR5218 CNRS, Bâtiment A4, 351 cours de la Libération, 33405 Talence, France

Correspondence should be addressed to Vladislavs Dovgalecs; vladislavs.dovgalecs@gmail.com

Received 27 April 2012; Revised 29 September 2012; Accepted 5 December 2012

Academic Editor: Anastasios Doulamis

Copyright © 2013 Vladislavs Dovgalecs et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The analysis of video acquired with a wearable camera is a challenge that multimedia community is facing with the proliferation of such sensors in various applications. In this paper, we focus on the problem of automatic visual place recognition in a weakly constrained environment, targeting the indexing of video streams by topological place recognition. We propose to combine several machine learning approaches in a time regularized framework for image-based place recognition indoors. The framework combines the power of multiple visual cues and integrates the temporal continuity information of video. We extend it with computationally efficient semisupervised method leveraging unlabeled video sequences for an improved indexing performance. The proposed approach was applied on challenging video corpora. Experiments on a public and a real-world video sequence databases show the gain brought by the different stages of the method.

1. Introduction

Due to the recent achievements in the miniaturization of cameras and their embedding in smart devices, a number of video sequences captured using such wearable cameras increased substantially. This opens new application fields and renews the problematics posed to the Multimedia research community earlier. For instance, visual lifelogs can record daily activities of a person and constitute a rich source of information for the task of monitoring persons in their daily life [1–4]. Recordings captured using wearable camera depict a view that is inside-out, close to the subjective view of the camera wearer. It is a unique source of information, with applications such as a memory refresh aid or as an additional source of information for the analysis of various activities and behavior related events in healthcare context. This often comes at the price of contents with very high variability, rapid camera displacement, and poorly constrained environments in which the person moves. Search for specific events in such multimedia streams is therefore particularly challenging. As was shown in [5, 6], multiple aspects of the video content and its context can be taken into account to provide a complete view of activity related events: location, presence of objects or persons, hand movements, and external information such as Global Positioning System (GPS), Radio Frequency Identification (RFID), or motion sensor data. Amongst these, location is an important contextual information, that restricts the possible number of ongoing activities. Obtaining this information directly from the video stream is an interesting application in multimedia processing since no additional equipment such as GPS or RFID is needed. In some applications, this may be even be a constraint, since the access to such modalities is limited in practice by the available devices and the installation of any invasive equipment in the environment (such as home) may not be welcome.

Considering the high cost of labeling data for training when dealing with lifelogs, and therefore the low amount of such labeling, inferring place recognition information from such content is a particularly great challenge. For instance, in the framework presented in [7], video lifelog recordings are made in an unknown environment and ground truth location information is limited to small parts of the recording. In such setup, the information sources are short manual annotations and large unlabeled recording parts. The use of unlabeled data to improve recognition performance was up to now reserved to more generic problems and was not evaluated in within the context of wearable video indexing. Efficient usage of this information for place recognition in wearable video indexing therefore defines the problem of the present work.

In this paper, we propose a novel strategy to incorporate and take advantage of the unlabeled data for place recognition. It takes into account both unlabeled data, multiple features and time information. We present a complete processing pipeline from low-level visual data extraction up to the visual recognition. The principal contribution of this work constitutes a novel system for robust place recognition in weakly annotated videos. We propose a combination of the Co-Training algorithm with classifier fusion to obtain a single classification estimate that exploits both multiple features and unlabeled data. In this context, we also study a range of confidence computation techniques found in the literature and introduce our own confidence measure that is designed to reduce the impact of uncertain classification results. The proposed system is designed as such that each component is evaluated separately and its presence is justified. It will be shown that each component yields an increase in classification performance, both separately as well as in a combined configuration, as demonstrated on public and our challenging in-house datasets.

The system we propose in this paper is motivated by the need to develop a robust image-based place recognition system as a part of high-level activity analysis system developed within the IMMED project [7]. As a part of this project, a wearable video recording prototype (see Figure 1) video annotation software and activity recognition algorithms were developed as well but will be left out of the scope. More detail on the latter can be found in [7, 8].

The paper is organized as follows. In Section 2, we review related work from the literature with respect to visual recognition, multiple feature fusion, and semisupervised learning. In Section 3, we present the proposed approach and algorithms. In Section 4, we report the experimental evaluations done on two databases in real life conditions and show the respective gains from the use of (a) multiple features, (b) unsupervised data, and (c) temporal information within our combined framework.

2. Literature Review

2.1. Activity Monitoring Context

2.1.1. Motivation. With this subsection we aim to put our work in the context of activity detection and recognition in video. Several setups have been used for that matter: ambient and wearable sensors.

2.1.2. Monitoring Using Ambient Sensors. Activity recognition systems have emerged quickly due to recent advances in large video recording and in the deployment of high computation power systems. For this application field, most of proposed methods originated from scene classification where static image information is captured and categorized.

Authors in [9] use the SVM classifier to classify local events such as "walking" and "running" in a database consisting of very clean and unoccluded video sequences. Perhaps in a more challenging setup, human behavior recognition is performed in [10] by proposing specially crafted sparse spatiotemporal features adapted for temporal visual data description. Conceptually a similar approach is proposed in [11] where each event in a soccer game is modeled as a temporal sequence of Bag of Visual Words (BOVWs) features used in a SVM classifier, termed strings, which are then compared using the string kernel.

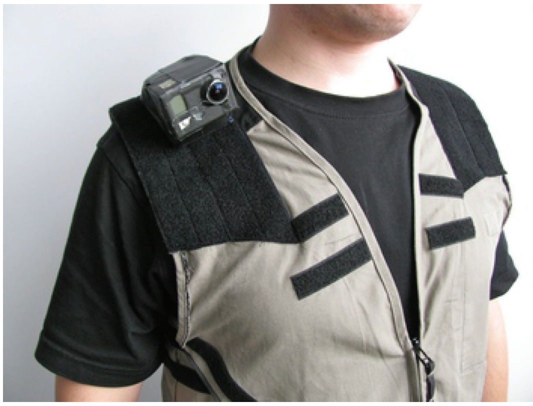


Figure 1: Wearable camera recording prototype used in the IMMED project.

Detection and recognition of events in real-world industrial workflows is a challenging problem because of great intraclass variability (complex classifiers required), unknown event start/end moments, and requirement to remember the whole event history which violates Markovian assumptions of conditional independence (e.g., HMM-based algorithms). The problem is alleviated in [12], where authors propose an online worker behavior classification system that integrates particle filter and HMM.

2.1.3. Monitoring Using Wearable Sensors. Alternatively, activity information can be also obtained from simple on-body sensors (e.g., acceleration) and wearable video.

Authors in [13] investigated two methods for activity context awareness in weakly annotated videos using 3D accelerometer data. The first one is based on multi-instance learning by grouping sensor data into bags of activities (instead of labeling every frame). The second one uses a graph structure for feature and time similarity representation and label information transfer in those structures. Results favor label propagation based methods in multiple feature graphs.

Visual lifelog indexing by human actions [2, 3, 14, 15] is proposed recently in healthcare with expansion of the Alzheimer disease. Early attempts to answer the challenge was done in [1, 4] as a part of the SenseCam and IMMED projects proposing lightweight devices and event segmentation algorithms. A motion-based temporal video segmentation algorithm with HMM at the core [8] identified strong correlation between activities and localization. This study reveals the complexity of the issue which consists in learning a generative model from few training data, extension to larger scale, and in difficulty to recognize short and infrequent activities. These and related issues were addressed in [16] with Hierarchical HMM which simultaneously fusing complementary low-level and midlevel (visual, motion, location, sound, and speech) features and the contribution of an automatic audio-visual stream segmentation algorithm. Results validate the choice of two-level modeling of activities using Hierarchical HMM and reveal improvement in recognition performance when working with temporal segments. Optimal feature fusion strategies using the Hierarchical HMM are studied in [17]. The contributed intermediate level fusion at the observation level, where all features are treated separately, compares positively to more classic early and late fusion approaches.

This work is a part of an effort to detect and recognize person's activities from wearable videos in the context of healthcare within the IMMED (http://immed.labri.fr/) project [7] and continued within the Dem@Care (http://www.demcare.eu/) project. Localization information is one of multiple possible cues to detect and recognize activities solely from the egocentric point of view of the recording camera. Amongst these location estimation is an important cue, which we now discuss in more detail.

2.2. Visual Location Recognition

2.2.1. Motivation. Classifying the current location using visual content only is a challenging problem. It relates to several problematics that have been already addressed in various contexts such as image retrieval from large databases, semantic video retrieval, image-based place recognition in robotics, and scene categorization. A survey [18] on image and video retrieval methods covers the paradigms such as semantic video retrieval, interactive retrieval, relevance feedback strategies, and intelligent summary creation. Another comprehensive and systematic study in [19] that evaluates multimodal models using visual and audio information for video classification reveals the importance of global and local features and the role of various fusion methods such as ensembles, context fusion, and joint boosting. We will hereafter focus on location recognition and classification.

To deal with location recognition from image content only, we can consider two families of approaches: (a) from retrieval point of view where a place is recognized as similar to an existing labeled reference from which we can infer the estimated location; (b) from a classification point of view where the place corresponds to a class that can be discriminated from other classes.

2.2.2. Image Retrieval for Place Recognition. Image retrieval systems work by a principle that visual content presented in a query image is visually similar or related to a portion of images to be retrieved from database.

Pairwise image matching is a relatively simple and attractive approach. It implies query image comparison to all annotated images in the database. Top ranked images are selected as candidates and after some optional validation procedures the retrieved images are presented as a result to the query. In [20] SIFT feature matching followed by voting and further improvement with spatial information is performed to localize indoors with 18 locations with each of them presented with 4 views. The voting scheme determines locations whose keypoints were most frequently classified as the nearest neighbors. Additionally, spatial information is modeled using HMM bringing in neighbor location relationships. A study [21] for place recognition in lifelogs images found that the best matching technique is to use bi-directional matching which nevertheless adds computational complexity. This problem is resolved by using robust and rapid to extract SURF features which are then hierarchically clustered using the k-means algorithm in a vocabulary tree. The vocabulary tree allows the rapid descriptor comparison of query image descriptors to those of the database and where the tree leaf note descriptor votes for the database image.

The success of matching for place recognition depends greatly on the database which should contain a large amount of annotated images. In many applications, this is a rather strong assumption about the environment. In the absence of prior knowledge brought by completely annotated image database covering the environment, topological place recognition discretizes otherwise continuous space. A typical approach following this idea is presented in [22]. Authors propose the gradient orientation histograms of the edge map as image feature with a property that visually similar scenes are described by a similar histogram. The Learning Vector Quantization method is then used to retain only the most characteristic descriptors for each topological location. An unsupervised approach for robot place recognition indoors is adapted in [23]. The method partitions the space into convex subspaces representing room concepts by using approximated graph-cut algorithm with the possibility for user to inject can group and cannot group constraints. The adapted similarity measure relies on the 8-point algorithm constrained to planar camera motion and followed by robust RANSAC to remove false matches. Besides high computation cost, the results show good clustering capabilities if graph nodes representing individual locations are well selected and the graph is properly built. Authors recognize that at larger scale and more similarly looking locations, more false matching images may appear.

2.2.3. Image Classification for Place Recognition. Training visual appearance model and using it to classify unseen images constitutes another family of approaches. Image information is usually encoded using global or local patch features.

In [24] an image is modeled as a collection of patches, each of which is assigned a codeword using a prebuilt codebook, yielding a bag of codewords. The generic Bag of Word [25] approach has been quite successful as global features. One of its main advantages is the ability to represent possibly very complex visual contents and address scene clutter problem. It is flexible enough to accommodate both discrete features [25] and dense features [26], while letting the possibility to include also weak spatial information by spatial binning as in [27]. Authors in [6] argue that indoor scenes recognition require location-specific global features and propose a system recognizing locations by objects that are present in them. An interesting result suggests that the final recognition performance can be boosted even further as more object information is used in each image. A context-based system for place and object recognition is presented in [5]. The main idea is to use context (scene gist) as a prior and then use it as a prior infer what objects can be present in a scene. The HMM-based place recognition system requires a considerable amount of training data, possible transition probabilities, and so forth, but integrates naturally temporal information and confidence measure to detect the fact of navigating in unknown locations. Probabilistic Latent Semantic Analysis (pLSA) was used in [28] to discover higher level topics (e.g., grass, forest, water) from low-level visual features and building novel low dimensional representation used afterwards in k-Nearest Neighbor classifier. The study shows superior classification performance by passing from low-level visual features to high-level topics that could be loosely attributed to the context of the scene.

2.2.4. Place Recognition in Video. Place recognition from recorded videos brings both novel opportunities and information but also poses additional challenges and constraints. Much more image data can be extracted from video while in practice some small portion of it can be labeled manually. An additional information that is often leveraged in the literature is the temporal continuity of the video stream.

Matching-based approach has been used in [29] to retrieve objects in video. Results show that simple matching produces a large number of false positive matches but the usage of stop list to remove most frequent and most specific visual words followed by spatial consistency check significantly improves retrieval result quality. In [30] belief functions in the Bayesian filtering context are used to determine the confidence of a particular location at any time moment. The modeling involves sensor and motion models, which have to be trained offline with sufficiently large annotated database. Indeed, the model has to learn the model of allowed transitions between places, which require the annotated data to represent all possible transitions to be found in the test data.

An important group of methods performing simultaneous place recognition and mapping (SLAM) is widely used in robotics [31, 32]. The main idea in these methods is to simultaneously build and update a map in an unknown environment and track in real time the current position of the camera. In our work, the construction of such map is not necessary and may prove to be very challenging since the environment can be very complex and constantly changing.

2.3. Multiple Feature Learning

2.3.1. Motivation. Different visual features capture different aspects of a scene and correct choice depends on the task to solve [33]. To this end, even humans perform poorly when using only one information source of perception [34]. Therefore, instead of designing a specific and adapted descriptor for each specific case, several visual descriptors can be combined in a more complex system while yielding increased discrimination power in a wider range of applications. Following the survey [35], two main approaches can be identified for the fusion of multiple features, depending on whether the fusion is done in the feature space (early fusion), or in the decision space (late fusion).

2.3.2. Early Fusion. Early fusion strategies focus on the combination of input features before using them in a classifier. In the case of kernel classifiers, the features can be seen as defining a new kernel that takes into account several features at once. This can be done by concatenating the features into a new larger feature vector. A more general approach, Multiple Kernel Learning (MKL), also tries to estimate the optimal parameters for kernel combination in addition to the classifier model. In our work we evaluated the SimpleMKL [36] algorithm as a representative algorithm of the MKL family. The algorithm is based on gradient descent and learns a weighted linear combination of kernels. This approach has notably been applied in the context of object detection and classification [37–39] and image classification [40, 41].

2.3.3. Late Fusion. In the late fusion, strategy several base classifiers are trained independently and their outputs are fed to a special decision layer. This fusion strategy is commonly referred to as a stacking method and is discussed in depth in the multiple classifiers systems literature [42–46]. This type of fusion allows to use multiple visual features, leaving their exploitation to an algorithm which performs automatic feature selection or weighting respective to the utility of each feature.

It is clear that nothing prevents using an SVM as base classifier. Following the work of [47], it was shown that SVM outputs, in the form of decision values, can be combined the linearly using Discriminative Accumulation Scheme (DAS) [48] for confidence-based place recognition indoors. The following work evolved by relaxing the constraint of linearity of combination using a kernel function on the outputs of individual single feature outputs giving rise to Generalized DAS [49]. Results show a clear gain of performance increase when using different visual features or completely different modalities. Other works follow a similar reasoning but use different combination rules (max, product, etc.), as discussed in [50]. A comprehensive comparison of different fusion methods in the context of object classification is given in [51].

2.4. Learning with Unlabeled Data

2.4.1. Motivation. Standard supervised learning, with single or multiple features, is successful if enough labeled training samples are presented to the learning algorithm. In many practical applications, the amount of training data is limited while a wealth of unlabeled data is often available and is largely unused. It is well known that the classifiers learned using only training data may suffer from overfitting or incapability to generalize on the unlabeled data. In contrast, unsupervised methods do not use label information. They may detect a structure of the data; however, a prior knowledge and correct assumptions about the data is necessary to be able to characterize a structure that is relevant for the task.

Semisupervised learning addresses this issue by leveraging labeled as well as unlabeled data [52, 53].

2.4.2. Graph-Based Learning. Given a labeled set $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and an unlabeled set $U = \{\mathbf{x}_j\}_{j=l+1}^{l+u}$, where $\mathbf{x} \in \mathcal{X}$ and $y \in \{-1, +1\}$, the goal is to estimate class labels for the latter. The usual hypothesis is that the two sets are sampled i.i.d. according to the same joint distribution $p(\mathbf{x}, y)$. There is no intention to provide estimations on the data outside the sets L and U. Deeper discussion on this issue can be found in [54] and in references therein.

In graph-based learning, a graph composed of labeled or unlabeled nodes (in our case representing the images) and interconnected by edges encoding the similarities is built. Application specific knowledge is used to construct such graph in such a way that the labels of nodes connected with a high weight link are similar and that no or a few weak links are present between nodes of different classes. This graph therefore encodes information on the smoothness of a learned function f on the graph, which corresponds to a measure of compatibility with the graph connectivity. The use of the graph Laplacian [55, 56] can then be used directly as a connectivity information to propagate information from labeled nodes to unlabeled nodes [57], or as a regularization term that penalizes nonsmooth labelings within a classifier such as the Lap-SVM [58, 59].

From a practical point of view, the algorithm requires the construction of the full affinity matrix W where all image pairs in the sequences are compared and the computation of the associated Laplacian matrix L, which requires large amounts of memory in $O(n^2)$. While theoretically attractive, the direct method scales poorly with the size of the graph nodes which seriously restricts its usage on a wide range of practical applications working.

2.4.3. Co-Training from Multiple Features. The Co-Training [60] is a wrapper algorithm that learns two discriminant classifiers in a joint manner. The method trains iteratively two classifiers such that in each iteration the highest confidence estimates on unlabeled data are fed into the training set of another classifier. Classically two views on the data or two single feature splits of a dataset are used. The main idea is that the solution or hypothesis space is significantly reduced if both trained classifiers agree on the data and reduce the risk of overfitting since each classifier also fits the initial labeled training set. More theoretical background and analysis of the method is given in Section 3.3.2.

The algorithm of Co-Training was proposed in [60] as a solution to classify Web pages using both link and word information. The same method was applied to the problem of Web image annotation in [61, 62] and automatic video annotation in [63]. Generalization capacity of Co-Training on different initial labeled training sets was studied in [64]. More analysis on theoretical properties of Co-Training method can be found in [65] such as rough estimates of maximal number of iterations. A review on different variants of the Co-Training algorithm is given [66] together with their comparative analysis.

2.4.4. Link between Graph and Co-Training Approaches. It is interesting to note the link [67, 68] between Co-Training method and label propagation in a graph since adding the most confident estimations in each Co-Training iteration can be seen as label propagation from labeled nodes to unlabeled nodes in a graph. This view of the method is further discussed and practically evaluated in [69] as a label propagation method on a combined graph built from two individual views.

Graph-based methods are limited by the fact that graph edges encode low-level similarities that are computed directly from the input features. The Co-Training algorithm uses a discriminative model that can be adaptive to the data with each iteration and therefore achieve better generalization on unseen unlabeled data. In the next section, we will build a framework based on the Co-Training algorithm to propose our solution for image-based place recognition.

In this work we attempt to leverage all available information from image data that could help to provide cues on camera place recognition. Manual annotation of recorded video sequences requires a lot of human labor. The aim of this work is to evaluate the utility of unlabeled data within the Co-Training framework for image-based place recognition.

3. Proposed Approach

In this section we present the architecture of the proposed method which is based on the Co-Training algorithm and then discuss each component of the system. The standard Co-Training algorithm (see Figure 2) allows to benefit from the information in the unlabeled part of the corpus by using a feedback loop to augment the training set, thus producing augmented performance classifiers. In the standard algorithm formulation, the two classifiers are still separate, which does not leverage their complementarity to its maximum. The proposed method addresses this issue by providing a single output using late classifier fusion and time filtering for temporal constrain enforcement.

We will present the different elements of the system in the order of increasing abstraction. Single feature extraction, preparation, and classification using SVM will be presented in Section 3.1. Multiple feature late fusion and a proposed extension to take into account the time information will be introduced in Section 3.2. The complete algorithm combining those elements with the Co-Training algorithm will be developed in Section 3.3.

3.1. Single Feature Recognition Module. Each image is represented by a global signature vector. In the following sections, the visual features $\mathbf{x}_i^{(j)} \in \mathcal{X}^{(j)}$ correspond to numerical representations of the visual content of the images where the superscript (j) denotes the type of visual features.

3.1.1. SVM Classifier. In our work we rely on Support Vector Machine (SVM) classifiers to carry out decision operations. It aims at finding the best class separation instead of modeling potentially complex within class probability densities as in generative models such as Naive Bayes [70]. The maximal margin separating hyperplane is motivated from the

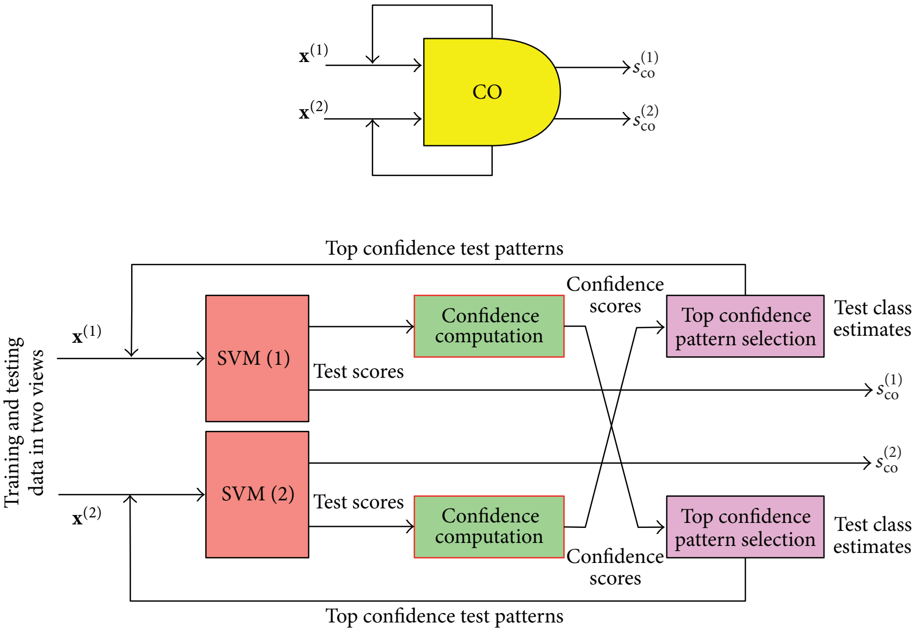


Figure 2: Workflow of the Co-Training algorithm.

statistical learning theory viewpoint by linking the margin width to classifier's generalization capability.

Given a labeled set $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, where $\mathbf{x} \in \mathbb{R}^d$, $y \in \{-1,+1\}$, a linear maximal margin classifier $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ can be found by solving

$$\min_{\mathbf{w},b,\xi} \sum_{i=1}^{l} \xi_i + \lambda \|\mathbf{w}\|^2 \qquad (1)$$
$$\text{s.t. } y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0,\ \forall i,\ i = 1,\ldots,l$$

for hyperplane $\mathbf{w} \in \mathbb{R}^d$ and its offset $b \in \mathbb{R}$. In regularization framework, the loss function called Hinge loss is

$$\ell(\mathbf{x}, y, f(\mathbf{x})) = \max\left(1 - y_i f(\mathbf{x}_i), 0\right), \quad \forall i, i = 1,\ldots,l \quad (2)$$

and the regularizer

$$\Omega_{\text{SVM}}(f) = \|\mathbf{w}\|^2. \quad (3)$$

As it will be seen from discussion, the regularizer plays an important role in the design of learning methods. In the case of an SVM classifiers, the regularizer in (3) reflects the objective to be maximized—maximum margin separation on the training data.

3.1.2. Processing of Nonlinear Kernels. The power of the SVM classifier owes its easy extension to the nonlinear case [71]. Highly nonlinear nature of data can be taken into account seamlessly by using kernel trick such that the hyperplane is found in a feature space induced by an adapted kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)\rangle$ in Reproducing Kernel Hilbert Space (RKHS). The implicit mapping $\mathbf{x} \mapsto \Phi(\mathbf{x})$ means that we can no longer find an explicit hyperplane $\{\mathbf{w}, b\}$ since the mapping function is not known and may be of very large dimensionality. Fortunately, the decision function can be formulated in so-called dual representation [71] and then the solution minimizing regularized risk according to the Representer theorem is

$$f_k(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b, \quad k = 1,\ldots,c, \quad (4)$$

where l is the number of labeled samples.

Bag of Words descriptors have been used intensively for efficient and discriminant image description. The linear kernel does not provide the best results with such representations, which has been more successful with kernels such as the Hellinger kernel, χ^2-kernel, or the intersection kernel [6, 27, 33]. Unfortunately, training with such kernels using the standard SVM tools is much less computationally efficient than using the linear inner product kernel, for which efficient SVM implementations exist [72]. In this work, we have therefore chosen to adapt the input features to the linear context, using two different techniques. For the BOVW (Bag of Visual Words) [25] and SPH (Spatial Pyramid Histogram) [27] features, a Hellinger kernel was used. This kernel admits an explicit mapping function, using a square root transformation $\phi([x_1 \cdots x_d]^T) = [\sqrt{x_1} \cdots \sqrt{x_d}]^T$. In this particular case, a linear embedding $\mathbf{x}' = \phi(\mathbf{x})$ can be computed explicity and will have the same dimensionality as input feature. For the CRFH (Composed Receptive Field Histogram) features [26], the feature vector has very large number of dimensions, but is also extremely sparse, with between 500 and 4000 nonzeros coefficients from many millions of features in total. These features could be transformed into a linear embedding using Kernel Principal Component Analysis [73], in order to reduce it to a 500-dimension linear embedding vector.

In the following, we will therefore consider that features are all processed into a linear embedding $\mathbf{x}_i$ that is suitable for efficient linear SVM. Utility of this processing will be evident in the context of the Co-Training algorithm, which requires multiple retraining and prediction operations of two visual feature classifiers. Other forms of efficient embedding, proposed in [38], could be also used to reduce learning time. This preprocessing is done only once, right after feature extraction from image data. In order to simplify the explanations, we will slightly abuse notation by denoting directly by $\mathbf{x}_i$ the linearized descriptors without further indication in the rest of this document.

3.1.3. Multiclass Classification. Visual place recognition is a truly multiclass classification problem. The extension of the binary SVM classifier to $c > 2$ classes is considered in a one-versus-all setup. Therefore c independent classifiers are trained on the labeled data, each of which learns the separation between one class and the other classes. We will denote by f_k the decision function associated to class $k \in [|\ 1, \ldots c\ |]$. The outcome of the classifier bank for a sample $\mathbf{x}$ can be represented as a scores vector $\mathbf{s}(\mathbf{x})$ by concatenating individual decision scores:

$$\mathbf{s}(\mathbf{x}) = \left(f_1(\mathbf{x}), \ldots, f_c(\mathbf{x})\right). \tag{5}$$

In that case, the estimated class of a testing sample $\mathbf{x}_i$ is estimated from the largest positive score:

$$\hat{y}_i = \arg\max_{k=1,\ldots,c} f_k(\mathbf{x}_i). \tag{6}$$

3.2. Multiple Feature Fusion Module and Its Extension to Time Information. In this work, we follow a late classifier fusion paradigm with several classifiers being trained independently on different visual cues and fusing the outputs for a single final decision. We motivate this choice compared to early fusion paradigm as it will allow easier integration at the decision level of augmented classifiers obtained by the Co-Training algorithm, as well as providing a natural extension to inject temporal continuity information of video.

3.2.1. Objective Statement. We denote the training set by $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and the unlabeled set of patterns by $U = \{\mathbf{x}_j\}_{j=l+1}^{l+u}$ where $\mathbf{x} \in \mathcal{X}$ and the outcome of classification is a binary output: $y \in \{-1, +1\}$.

The visual data may have p multiple cues describing the same image I_i. Suppose that p cues has been extracted from an image I_i:

$$\mathbf{x}_i \longrightarrow \left(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(p)}\right), \tag{7}$$

where each cue $\mathbf{x}_i^{(j)}$ belongs to an associated descriptor space $\mathcal{X}^{(j)}$.

Denote also by p the decision functions $f^{(1)}, f^{(2)}, \ldots, f^{(p)}$, where $f^{(j)} \in \mathcal{F}^{(j)}$ are trained on the respective visual cues and are providing estimation $\hat{y}_k^{(j)}$ on the pattern $\mathbf{x}_k^{(j)}$. Then for a visual cue t and c class classification in one-versus-all setup, a score vector can be constructed:

$$\mathbf{s}^t = \left(f_1^t(\mathbf{x}), \ldots, f_c^t(\mathbf{x})\right). \tag{8}$$

In our work we adopt two late fusion techniques: Discriminant Accumulation Scheme (DAS) [47, 48] and SVM-DAS [49, 74].

3.2.2. Discriminant Accumulation Scheme (DAS). The idea of DAS is to combine linearly the scores returned by the same class decision function across multiple visual cues $t = 1, \ldots, p$. The novel combined decision function for a class j is then a linear combination:

$$f_j^{\mathrm{DAS}}(\mathbf{x}) = \sum_{t=1}^{p} \beta_t f_j^t(\mathbf{x}), \tag{9}$$

where the weight β is attributed to each cue according to its importance in the learning phase. The novel scores can then be used in decision process, for example, using max score criterion.

The DAS scheme is an example of parallel classifier combination architectures [44] and implies a competition between the individual classifiers. The weights β_t can be found using a cross-validation procedure with the normalization constraint:

$$\sum_{t=1}^{p} \beta_t = 1. \tag{10}$$

3.2.3. SVM Discriminant Accumulation Scheme. The SVM-DAS can be seen as a generalization of the DAS by building a stacked architecture of multiple classifiers [44] where individual classifier outputs are fed into a final classifier that provides a single decision. In this approach every classifier is trained on its own visual cue t and produces a score vector as in (8). Then the single feature score vectors $\mathbf{s}_i^t$ corresponding to one particular pattern $\mathbf{x}_i$ are concatenated into a novel multifeatures scores vector $\mathbf{z}_i = [\mathbf{s}_i^1, \ldots, \mathbf{s}_i^p]$. A final top-level classifier can be trained on those novel features:

$$f_j^{\mathrm{SVMDAS}}(\mathbf{z}) = \sum_{i=1}^{l} \alpha_{ij} y_i k(\mathbf{z}, \mathbf{z}_i) + b_j. \tag{11}$$

Notice that the use of kernel function enables a richer class of classifiers modeling possibly nonlinear relations between base classifier outputs. If a linear kernel function is used,

$$k_{\mathrm{SVMDAS}}\left(\mathbf{z}_i, \mathbf{z}_j\right) = \left\langle \mathbf{z}_i, \mathbf{z}_j \right\rangle = \sum_{t=1}^{p} \left\langle \mathbf{s}_i^t, \mathbf{s}_j^t \right\rangle, \tag{12}$$

then the decision function in (11) can be rewritten by exchanging sums:

$$\begin{aligned} f_j^{\mathrm{SVMDAS}}(\mathbf{z}) &= \sum_{i=1}^{l} \alpha_{ij} y_i k(\mathbf{z}, \mathbf{z}_i) + b_j \\ &= \sum_{i=1}^{l} \left\langle \mathbf{s}_i^t, \mathbf{s}_j^t \right\rangle \sum_{t=1}^{p} \alpha_{ij} y_i + b_j. \end{aligned} \tag{13}$$

Denoting $\mathbf{w}_j^t = \sum_{i=1}^{l} \alpha_{ij} y_i \mathbf{s}^t$, we can rewrite the decision function using input patterns and the learned weights:

$$f_j^{\text{SVMDAS}}(\mathbf{z}) = \sum_{t=1}^{p} \sum_{k=1}^{l} \mathbf{w}_{jk}^t f_j^t(\mathbf{x}). \tag{14}$$

The novel representation reveals that using a linear kernel in the SVMDAS framework renders a classifier with weights being learned for every possible linear combination of base classifiers. The DAS can be seen as a special case in this context but with significantly less parameters. Usage of a kernel such as RBF or polynomial kernels can result in even richer class of classifiers.

The disadvantage of such configuration is that a final stage classifier needs to be trained as well and its parameters tuned.

3.2.4. Extension to Temporal Accumulation (TA). Video content has a temporal nature such that the visual content does not usually change much in a short period of time. In the case of topological place recognition indoors, this constraint may be useful as place recognition changes are encountered relatively rarely with respect to the frame rate of the video.

We propose to modify the classifier output such that rapid class changes are discouraged in a relatively short period of time. This leads to lower the proliferation of occasional temporally localized misclassifications.

Let $s_i^t = f^{(t)}(\mathbf{x}_i)$ be the scores of a binary classifier for visual cue t and h a temporal window of size $2\tau + 1$. Then temporal accumulation can be written as

$$s_{i,\text{TA}}^t = \sum_{k=-\tau}^{\tau} h(k)\, s_{i+k}^t \tag{15}$$

and can be easily generalized to multiple feature classification by applying it separately to the output of the classifiers associated to each feature $\mathbf{s}^t$, where $t = 1, \ldots, p$ is the visual feature type. We use an averaging filter of size τ, defined as

$$h(k) = \frac{1}{2\tau + 1}, \quad k = -\tau, \ldots, \tau. \tag{16}$$

Therefore, input of the TA are the SVM scores obtained after classification and output are again the processed SVM scores with temporal constraint enforced.

3.3. Co-Training with Time Information and Late Fusion. We have already presented how to perform multiple feature fusion within the late fusion paradigm, and how it can be extended to take into account the temporal continuity information of video. In this section, we will explain how to additionally learn from labeled training data and unlabeled data.

3.3.1. The Co-Training Algorithm. The standard Co-Training [60] is an algorithm that iteratively trains two classifiers on two view data $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})$ by feeding the highest confidence score z_i estimates from the testing set in another view classifier. In this semisupervised approach, the discriminatory power of each classifier is improved by another classifier's complementary knowledge. The testing set is gradually labeled round by round using only the highest confidence estimates. The pseudocode is presented in Algorithm 1 which could be also extended to multiple views as in [53].

The power of the method lies in its capability of learning from small training sets and grows eventually its discriminative properties on the large unlabeled data set as more confident estimations are added into the training set. The following assumptions are made:

(1) the two distinct visual cues bring complementary information;

(2) the initially labeled set for each individual classifier is sufficient to bootstrap the iterative learning process;

(3) the confident estimations on unlabeled data are helpful to predict the labels of the remaining unlabeled data.

Originally the Co-Training algorithm performs until some stopping criterion is met unless N iterations are exceeded. For instance, a stopping criteria could be a rule that stops the learning process when there are no confident estimations to add or there have been relatively small difference from iteration $t - 1$ to t. The parameter-less version of Co-Training works till the complete exhaustion of the pool of unlabeled samples but requires a threshold on confidence measure, which is used to separate high and low confidence estimates. In our work we use this variant of the Co-Training algorithm.

3.3.2. The Co-Training Algorithm in the Regularization Framework

Motivation. Intuitively, it is clear that after a sufficient number of rounds both classifiers will agree on most of the unlabeled patterns. It remains unclear why and what mechanisms make such learning useful. It can be justified from the learning theory point of view. There are less possible solutions or classifiers from the hypothesis space that agree on unlabeled data in two views. Recall that every classifier individually should fit its training data. In the context of the Co-Training algorithm each classifier should be somehow restricted by another classifier. The two trained classifiers, that are coupled in this system, effectively reduce possible solution space. Each of those two classifier is less likely to be overfitting since each of them has been initially trained on its training while taking into account the training process of another classifier that is carried out in parallel. We follow the discussion from [53] to give more insights about this phenomena.

Regularized Risk Minimization (RRM) Framework. Better understanding of the Co-Training algorithm can be gained from the RRM framework. Let's introduce the Hinge loss

```
INPUT:
  Training set $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$;
  Testing set $U = \{\mathbf{x}_i\}_{i=1}^{u}$;
OUTPUT:
  $\widehat{y}_i$—class estimations for the testing set $U$;
  $f^{(1)}, f^{(2)}$—trained classifiers;
PROCEDURE:
  (1) Compute visual features $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})$ for every image $I_i$ in the dataset
  (2) Initialize $L_1 = \{(\mathbf{x}_i^{(1)}, y_i)\}_{i=1}^{l}$ and $L_2 = \{(\mathbf{x}_i^{(2)}, y_i)\}_{i=1}^{l}$
  (3) Initialize $U_1 = \{\mathbf{x}_i^{(1)}\}_{i=1}^{u}$ and $U_2 = \{\mathbf{x}_i^{(2)}\}_{i=1}^{u}$
  (4) Create two work sets $\widetilde{U}_1 := U_1$ and $\widetilde{U}_2 := U_2$
  (5) Repeat until the sets $\widetilde{U}_1$ and $\widetilde{U}_2$ are empty (CO)
      (a) Train classifiers $f^{(1)}, f^{(2)}$ using the sets $L_1, L_2$ respectively;
      (b) Classify the patterns in the sets $\widetilde{U}_1$ and $\widetilde{U}_2$ using the classifiers $f^{(1)}$ and $f^{(2)}$ respectively;
          (i) Compute scores $s_{\text{test}}^{(1)}$ and confidences $z^{(1)}$ on the set $\widetilde{U}_1$
          (ii) Compute scores $s_{\text{test}}^{(2)}$ and confidences $z^{(2)}$ on the set $\widetilde{U}_2$
      (c) Add the $k$ top confidence estimations $\overline{L}_1 \subset \widetilde{U}_1, \overline{L}_2 \subset \widetilde{U}_2$
          (i) $L_1 := L_1 \cup \overline{L}_1$
          (ii) $L_1 := L_1 \cup \overline{L}_1$
      (d) Remove the $k$ top confidence patterns from the working sets
          (i) $\widetilde{U}_1 := \widetilde{U}_1 \setminus \overline{L}_1$
          (ii) $\widetilde{U}_2 := \widetilde{U}_2 \setminus \overline{L}_2$
      (e) Go to step (5).
  (6) Optionally: perform Temporal Accumulation (TA) according to (15)
  (7) Perform classifier output fusion (DAS)
      (a) Compute fused scores $\mathbf{s}_{\text{test}}^{\text{DAS}} = (1-\beta)\,\mathbf{s}_{\text{test}}^{(1)} + \beta\mathbf{s}_{\text{test}}^{(2)}$;
      (b) Output class estimations $\widehat{y}_i$ from the fused scores $\mathbf{s}_{\text{test}}^{\text{DAS}}$
```

Algorithm 1: The CO-DAS and CO-TA-DAS algorithms.

function $\ell(\mathbf{x}, y, f(\mathbf{x}))$ commonly used in classification. Let's also introduce empirical risk of a candidate function $f \in \mathcal{F}$:

$$\widehat{R}(f) = \frac{1}{l}\sum_{i=1}^{l} \ell(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) \tag{17}$$

which measures how well the classifier fits the training data. It is well known that minimizing only training error, the resulting classifier is very likely to overfit. In practice regularized risk (RRM), minimization is performed instead:

$$f^{\text{RRM}} = \arg\min_{f\in\mathcal{F}} \widehat{R}(f) + \lambda\Omega(f), \tag{18}$$

where $\Omega(f)$ is a nonnegative functional or regularizer that returns a large value or penalty for very complicated functions (typically the functions that fit perfectly to the data). The parameter $\lambda > 0$ controls the balance between a fit to the training data and the complexity of the classifier. By selecting a proper regularization parameter, overfitting can be avoided and better generalization capability on the novel data can be achieved. A good example is the SVM classifier. The corresponding regularizer $\Omega_{\text{SVM}}(f) = (1/2)\|\mathbf{w}\|^2$ selects the function that maximizes the margin.

The Co-Training in the RRM. In semisupervised learning we can select a regularizer such that it is sufficiently smooth on unlabeled data as well. Keeping all previous discussion in mind, indeed a function that fits the training data and is respecting unlabeled data will probably perform better on future data. In the case of the Co-Training algorithm, we are looking for two functions $f^{(1)}, f^{(2)} \in \mathcal{F}$ that minimize the regularized risk and agree on the unlabeled data at the same time. The first restriction on the hypothesis space is that the first function should not only reduce its own regularized risk but also agree with the second function. We can then write a two-view regularized risk minimization problem as

$$\begin{aligned}
&\left(\widetilde{f}^{(1)}, \widetilde{f}^{(2)}\right) \\
&\quad = \arg\min_{f^{(1)},f^{(2)}} \sum_{t=1}^{2}\left(\frac{1}{l}\sum_{i=1}^{l}\ell\left(\mathbf{x}_i, y_i, f^{(t)}(\mathbf{x}_i)\right)\right. \\
&\qquad\qquad \left. + \lambda_1\Omega_{\text{SVM}}\left(f^{(t)}\right)\right) \\
&\qquad + \lambda_2\sum_{i=1}^{l+u}\ell\left(\mathbf{x}_i, f^{(1)}(\mathbf{x}_i), f^{(2)}(\mathbf{x}_i)\right),
\end{aligned} \tag{19}$$

where $\lambda_2 > 0$ controls the balance between an agreed fit on the training data and agreement on the test data. The first part of (19) states that each individual classifier should fit the

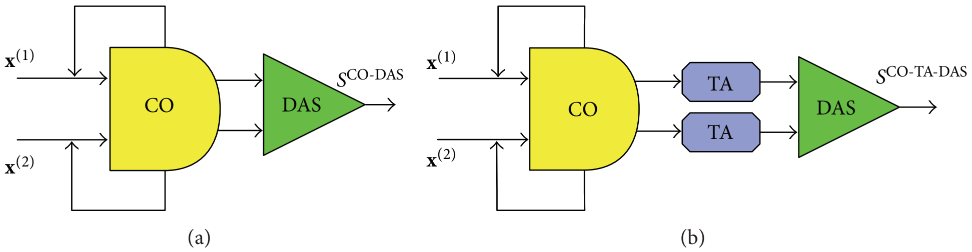


FIGURE 3: Co-Training with late fusion (a); Co-Training with temporal accumulation (b).

given training data but should not overfit, which is prevented with the SVM regularizer $\Omega_{\text{SVM}}(f)$. The second part is a regularizer $\Omega_{\text{CO}}(f^{(1)}, f^{(2)})$ for the Co-Training algorithm, which incurs penalty if the two classifiers do not agree on the unlabeled data. This means that each classifier is constrained both by its standard regularization and is required to agree with another classifier. It is clear that an algorithm implemented in this framework elegantly bootstraps from each classifiers training data, exploits unlabeled data, and works with two visual cues.

It should be noted that the framework could be easily extended to more than two classifiers. In the literature the algorithms following this spirit are implementing multiple view learning. Refer to [53] for the extension of the framework to multiple views.

3.3.3. Proposition: CO-DAS and CO-TA-DAS Methods. The Co-Training algorithm has two drawbacks in the context of our application. The first drawback is that it is not known in advance which of the two classifiers performs the best and if complementarity properties had been leveraged to their maximum. The second drawback is that no time information is used unless the visual features are constructed to capture this information.

In this work we will use the DAS method for late fusion while it is possible to use the more general SVMDAS method as well. Experimental evaluation will show that very competitive performances can be obtained using the former, much more simpler method. We propose the CO-DAS method (see Figure 3(a)), which addresses the first drawback by delivering a single output. In the same framework, we propose the CO-TA-DAS method (see Figure 3(b)), which additionally enforces temporal continuity information. Experimental evaluation will reveal relative performances of each method with respect to baseline and with respect to each other.

The full algorithm of the CO-DAS (or CO-TA-DAS if temporal accumulation is enabled) method is presented in Algorithm 1.

Besides the base classifier parameters, one needs to set the threshold k for the top confidence sample selection, temporal accumulation window width τ, and the late fusion parameter β. We express the threshold k as a percentage of the testing samples. The impact of this parameter is extensively studied in Sections 4.1.4 and 4.1.5. The selection of the temporal accumulation parameter is discussed in Section 4.1.3. Finally, discussion on the selection of the parameter β is given in Section 4.1.2.

3.3.4. Confidence Measure. The Co-Training algorithm relies on confidence measure, which is not provided by an SVM classifier out of the box. In the literature, several methods exist for computing confidence measure from the SVM outputs. We review several methods of confidence computation and contribute a novel confidence measure that attempts to resolve an issue which is common to some of the existing measures.

Logistic Model (Logistic). Following [75], class probabilities can be computed using the logistic model that generalizes naturally to multiclass classification problem. Suppose that in one-versus-all setup with c classes, the scores $\{f^k(\mathbf{x})\}_{k=1}^{c}$ are given. Then probability or classification confidence is computed as

$$P(y = k \mid \mathbf{x}) = \frac{\exp\left(f^k(\mathbf{x})\right)}{\sum_{i=1}^{c} \exp\left(f^i(\mathbf{x})\right)} \tag{20}$$

which ensures that probability is larger for larger positive score values and sum to 1 over all scores. This property allows to interpret the classifier output as a probability. There are at least two drawbacks with this measure. This measure does not take into account the cases when all classifiers in one-versus-all setup reject the pattern (all negative score values) or accept (all positive scores). Finally, forced score normalization to sum up to one may not transfer all dynamics (e.g., very small or very large score values).

Modeling Posterior Class Probabilities (Ruping). In [76] a parameter-less method was proposed which assigns score value:

$$z = \begin{cases} p_+, f(\mathbf{x}) > 1, \\ \dfrac{1 + f(\mathbf{x})}{2}, -1 \le f(\mathbf{x}) \le 1, \\ p_-, f(\mathbf{x}) < 1, \end{cases} \tag{21}$$

where p_+ and p_- are the fractions of positive and negative score values, respectively. Authors argue that interesting dynamics relevant to confidence estimation happen in the region of margin and the patterns classified outside the margin have a constant impact. This measure has sound theoretical background in a two-class classification problem but it does not cover multiclass case as required by our application.

Score Difference (Tommasi). A method that does not require additional preprocessing for confidence estimation was proposed in [77] and thresholded to obtain a decision corresponding to "no action," "reject," or "do not know" situation for medical image annotation. The idea is to use the contrast between the two top uncalibrated score values. The maximum score estimation should be more confident if other score values are relatively smaller. This leads to a confidence measure using the contrast between the two maximum scores:

$$z = f^{k^*}(\mathbf{x}) - \max_{k=1,\dots,c,k \neq k^*} f^k(\mathbf{x}). \qquad (22)$$

This measure has a clear interpretation in a two-class classification problem where larger difference between the two maximal scores hints for better class separability. As it is seen from equation, there is an issue with the measure if all scores are negative.

Class Overlap Aware Confidence Measure. We noticed that class overlap and reject situations are not explicitly taken into account in neither of confidence measure computation procedures. The one-versus-all setup for multiple class classification may yield ambiguous decisions. For instance, it is possible to obtain several positive scores or all positive or all negative scores.

We propose a confidence measure that penalizes class overlap (ambiguous decisions) at several degrees and also treats two degenerate cases. By convention, confidence should be higher if a sample is classified with less class overlap (fewer positive score values) and further from the margin (larger positive value of a score). Cases with all positive or negative scores may be considered as degenerate $z_i \leftarrow 0$.

The computation is divided in two steps. First we compute the standard Tommasi confidence measure:

$$z_i^0 = f^{j^*}(\mathbf{x}_i) - \max_{i=1,\dots,c,i \neq j^*} f^i(\mathbf{x}) \qquad (23)$$

then the measure z_i^0 is modified to account for class overlap

$$z_i = z_i^0 \max\left(0, 1 - \frac{p_i - 1}{C}\right), \qquad (24)$$

where $p_i = \mathrm{Card}(\{k = 1, \dots, c \mid f^k(\mathbf{x}_i) > 0\})$ represents the number of classes for which $\mathbf{x}_i$ has positive scores (class overlap). In case of $\forall k$, $f^k(\mathbf{x}_i) > 0$ or $f^k(\mathbf{x}_i) < 0$, we set $z_i \leftarrow 0$.

Compared to the Tommasi measure, the proposed measure additionally penalizes class overlap which is more severe if the test pattern receives several positive scores. Compared to logistic measure, samples with no positive scores yield zero confidence, which allows to exclude them and not assign doubtful probability values.

Constructing our measure, we assume that a confident estimate is obtained if only one of binary classifiers return a positive score. Following the same logic, confidence is lowered if more than one binary classifiers return a positive score.

4. Experimental Evaluation

In this section we evaluate the performance of the methods presented in the previous section on two datasets. Experimental evaluation is organized in two parts: (a) on the public database IDOL2 in Section 4.1 and (2) on our in-house database IMMED in Section 4.2, respectively. The former database is relatively simple and is expected to be annotated automatically with small error rate, whereas the latter database is recorded in a challenging environment and is a subject of study in the IMMED project.

For each database, two experiment setups are created: (a) randomly sampled training images across all corpus and (b) more realistic video-versus-video setup. First experiment allows for the gradual increase of supervision which gives insights of place recognition performance for algorithms under study. The second setup is more realistic and is aimed to validate every place recognition algorithm.

On the IDOL2 database, we extensively assess the place recognition performance for each independent part of the proposed system. For instance, we validate the utility of multiple features, effect of temporal smoothing, unlabeled data, and different confidence measures.

The IMMED database is used for validation purposes on which we evaluate all methods and summarize their performances.

Datasets. The IDOL2 database is a publicly available corpus of video sequences designed to assess place recognition systems of mobile robots in indoor environment.

The IMMED database represents a collection of video sequences recorded using a camera positioned on the shoulder of volunteers and capturing their activities during observation sessions in their home environment. These sequences represent visual lifelogs, for which indexing by activities is required. This database presents a real challenge for image-based place recognition algorithms, due to the high variability of the visual content and the unconstrained environment.

The results and discussion related to these two datasets are presented in Sections 4.1 and 4.2, respectively.

Visual Features. In this experimental section, we will use three types of visual features that have been used successfully in image recognition tasks: Bag of Visual Words (BOVWs) [25], Composed Receptive Field Histograms (CRFHs) [26], and Spatial Pyramid Histograms (SPHs) [27].

In this work we used 1111 dimensional BOVW histograms which was shown to be sufficient for our application and feasible from the computation point of view. The visual vocabulary was built in a hierarchical manner [25] with 3 levels and 10 sibling nodes to speed up the search of the tree. This allows to introduce visual words ranging from more general (higher level nodes) to more specific (leaf nodes). The effect of overly frequent visual words is addressed with the use of common normalization procedure tf-idf [25] from text classification.

The SPH [27, 78] descriptor harnesses the power of the BOVW descriptor but addresses its weakness when it comes to spatial structure of the image. This is done by

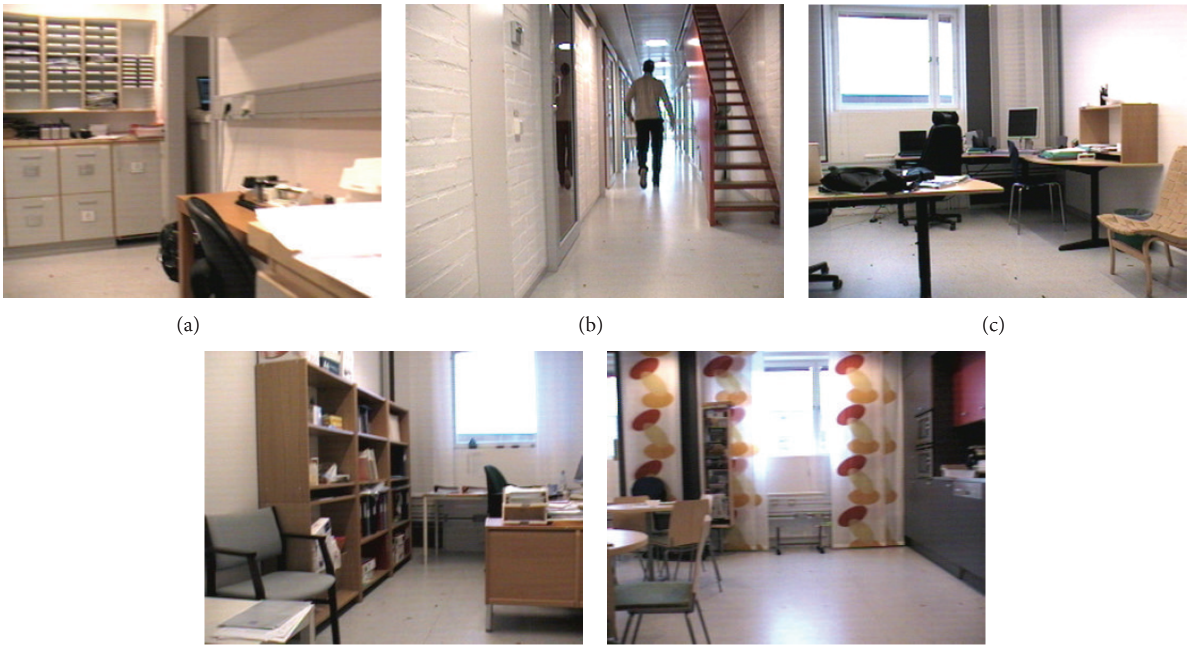


Figure 4: IDOL2 dataset sample images: (a) Printer Area, (b) Corridor, (c) Two-Person Office, (d) One-Person Office, and (e) Kitchen.

constructing a pyramid where each level defines coarse to fine sampling grid for histogram extraction. Each grid histogram is obtained by constructing standard BOVW histogram with local features SIFT sampled in a dense manner. The final global descriptor is composed of concatenated individual region and level histograms. We empirically set the number of pyramid levels to 3 with the dictionary size of 200 visual words, which yielded in 4200 dimensional vectors per image. Again, the number of dimensions was fixed such that maximum of visual information is captured while reducing computational burden.

The CRFH [26] descriptor describes a scene globally, by measuring responses returned after some filtering operation on the image. Every dimension of this descriptor effectively counts the number of pixels sharing similar responses returned from each specific filter. Due to multidimensional nature and the size of an image, such descriptor often results in a very high dimensionality vector. In our experimental evaluations, we used second order derivatives filter in three directions, at two scales with 28 bins per histogram. The total size of global descriptor resulted in very sparse up to 400 million dimension vectors. It was reduced to a 500-dimensional linear descriptor vector using KPCA with an χ^2 kernel [73].

4.1. Results on IDOL2. The public database KTH-IDOL2 [79] consists of video sequences captured by two different robot platforms. The database is suitable to evaluate the robustness of image-based place recognition algorithms in controlled real-world conditions.

4.1.1. Description of the Experimental Setup. The considered database consists of 12 video sequences recorded with the "minnie" robot (98 cm above ground) using a Canon VC-C4 camera at a frame rate of 5 fps. The effective resolution of the extracted images is 309 × 240 pixels.

All video sequences were recorded in the same premises and depict 5 distinct rooms—"One-Person Office," "Two-Person Office," "Corridor," "Kitchen," and "Printer Area". Sample images depicting these 5 topological locations are shown in Figure 4.

The annotation was performed using two annotation setups: random and video versus video. In both setup three image sets were considered: labeled training, validation set, and an unlabeled set. The unlabeled set is used as the test set for performance evaluation. The performance is evaluated using the accuracy metric, which is defined as the number of correctly classified test images divided by the total number of test images.

Random Sampling Setup. In the first setup, the database is divided into three sets by random sampling: training, validation, and testing. The percentage of training data with respect to the full corpus defines the *supervision level.* We consider 8 *supervision levels* ranging from 1% to 50%. The remaining images are split randomly in two halves and used, respectively, for validation and testing purposes. In order to account for the effects of random sampling, 10-fold sampling is made at each supervision level and the final result returned as the average accuracy measure.

It is expected that global place recognition performance raises from mediocre performance at low supervision to its maximum at high supervision level.

Video-versus-Video Setup. In the second setup, videos sequences are processed in pairs. The first video is completely annotated while the second is used for evaluation purposes.

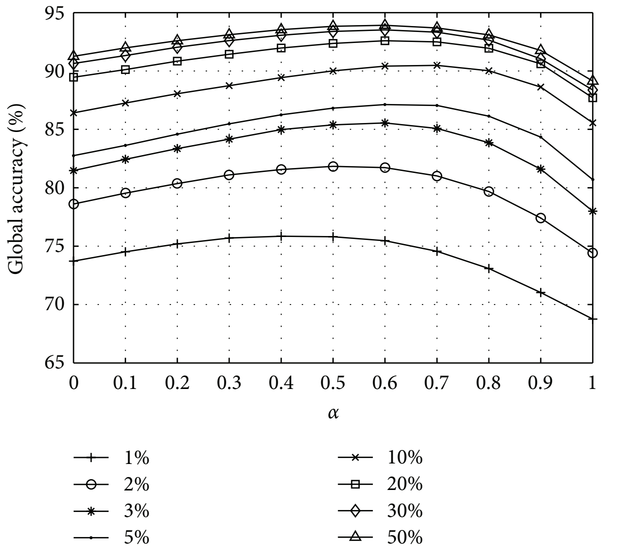


Figure 5: Effect of the DAS late fusion approach on the final performance, for various supervision levels. Plot of the accuracy as a function of the parameter α that balances the fusion between SPH features (3 levels) if $\alpha = 0$ and CRFH if $\alpha = 1$ (IDOL2 dataset, random setup).

The annotated video is split randomly into training and validation sets. With 12 video sequences under consideration, evaluating on all possible pairs amount to $132 = 12 \times 11$ pairs of video sequences. We differentiate three sets of pairs: "EASY," "HARD," and "ALL" result cases. The "EASY" set contains only the video sequence pairs where the light conditions are similar and the recordings were made in a very short span of time. The "HARD" set contains pairs of video sequences with different lighting conditions or video sequences recorded with a large time span. The "ALL" set contains all the 132 video pairs to provide an overall averaged performance.

Compared to random sampling setup, the video-versus-video setup is considered more challenging and thus lower place recognition performances are expected.

4.1.2. Utility of Multiple Features. We study the contribution of multiple features for the task of image-based place recognition on the IDOL2 database. We will present a complete summary of performances for baseline single feature methods compared to early and late fusion methods. These experiments were carried out using the random labeling setup only.

The DAS Method. The DAS method leverages two visual feature classifier outputs and provides a weighted score sum in the output on which class decision can be made. In Figure 5, the performance of DAS using SPH Level 3 and CRFH feature embeddings is shown as a function of fusion parameter α at different supervision levels. Interesting dynamics can be noticed for intermediary fusion values that suggest for feature complementarity. The fusion parameter α can be safely set to an intermediary value such as 0.5 and the final performance would exceed that of every single feature classifier alone at all supervision levels.

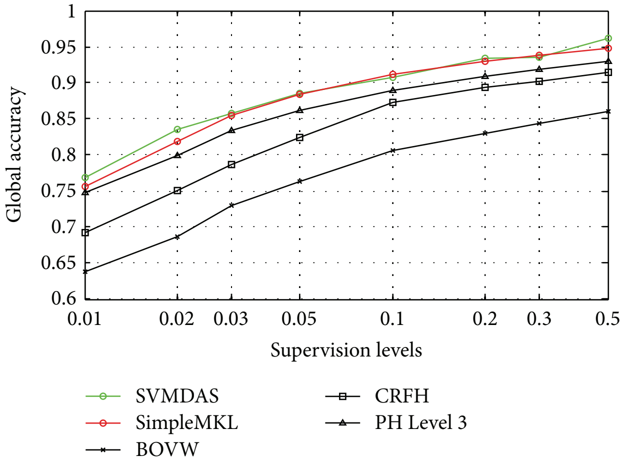


Figure 6: Comparison of single (BOVW, CRFH, and SPH) and multiple feature (SVMDAS, SimpleMKL) approaches for different supervision levels. Plot of the accuracy as a function of the supervision level (IDOL2 dataset, random setup).

The SVMDAS Method. In Figure 6, the effect of the supervision level on the performances of classification is shown for single feature and multiple features approaches. It is clear that all methods perform better if more labeled data is supplied, which is an expected behavior. We can notice differences in the performances on the 3 single feature approaches, with SPH providing the best performances. Both SVMDAS (late fusion approach) and SimpleMKL (early fusion approach) operate fusion over the 3 single features considered. They outperform the single feature baseline methods. There is practically no difference between the two fusion methods on this dataset.

Selection of the Late Fusion Method. Although not compared directly, the two late fusion methods DAS and SVMDAS deliver very comparable performances. Maximum performance comparison (at best α for each supervision) of the DAS (Figure 5) to those of the SVMDAS (Figure 6) confirms this claim on this particular database. Therefore, the choice of the DAS method for the following usage in the final system is motivated by this result and by simplified fusion parameter selection.

4.1.3. Effect of Temporal Smoothing

Motivation. Temporal information is an implicit attribute of video content which has not been leveraged up to now in this work. The main idea is that temporally close images should carry the same label.

Discussion on the Results. To show the importance of the time information, we present the effect of the temporal accumulation (TA) module on the performance of single feature SVM classification. In Figure 7, the TA window size is varied from no temporal accumulation up to 300 frames. The results

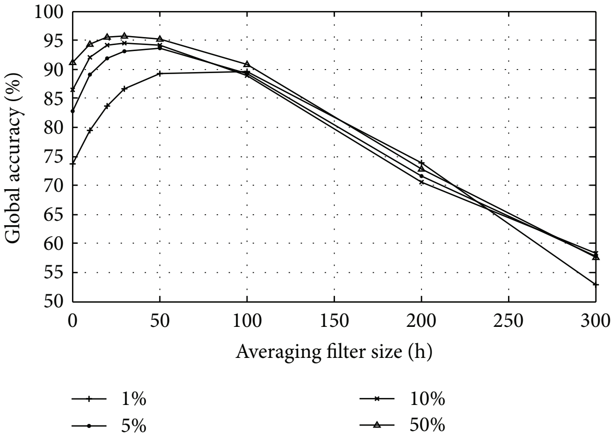


Figure 7: Effect of the filter size in temporal accumulation. Plot of the accuracy as a function of the TA filter size. (IDOL2 dataset, SPH Level 3 features).

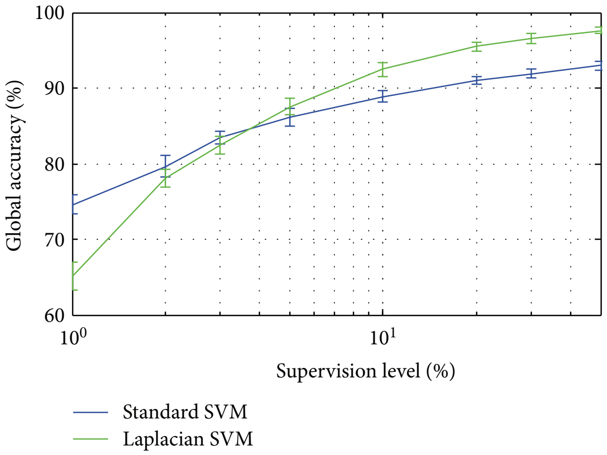


Figure 8: Comparison of standard single feature SVM with semi-supervised Laplacian SVM with RBF kernel on SPH Level 3 visual features (IDOL2 dataset, random setup).

show that temporal accumulation with a window size up to 100 frames (corresponding to 20 seconds of video) increases the final classification performance. This result shows that some minority of temporally close images, which are very likely to carry the same class label, obtain an erroneous label and temporal accumulation could be a possible solution. Assuming that only a minority of temporal neighbors are classified incorrectly makes the temporal continuity a strong cue for our application and should be integrated in the learning process as will be shown next.

Practical Considerations. In practice, the best averaging window size cannot be known in advance. Knowing the frame rate of the camera and relatively slow room change, the filter size can be set empirically to a number of frames that are captured in one second, for example.

4.1.4. Utility of Unlabeled Data

Motivation. The Co-Training algorithm belongs to a semisupervised learning algorithm grouped. Our goal is to assess its capacity to leverage unlabeled data in practice. First, we compare standard single feature SVM to a semisupervised SVM using graph smoothness assumption. Second, we study the proposed CO-DAS method. Third, we are interested to observe the evolution of performance if multiple Co-Training iterations are performed. Finally, we present a complete set of experiments on the IDOL2 database comparing single feature and multifeature baselines compared to the proposed semisupervised CO-DAS and CO-TA-DAS methods.

Our primary interest is to show how a standard supervised SVM classifier compares to a state-of-the-art semisupervised Laplacian SVM classifier. Performance of both classifiers is shown in Figure 8. The results show that semisupervised counterpart performs better if a sufficiently large initial labeled set of training patterns is given. The low performance at low supervision compared to standard supervised classifier can be explained by an improper parameter setting. Practical application of this method is limited since the total kernel matrix should be computed and stored in the memory of a computer, which scales as $O(n^2)$ with number of patterns. Computational time scales as $O(n^3)$, which is clearly prohibitive for medium and large sized datasets.

Co-Training with One Iteration. The CO-DAS method proposed in this work avoids these issues and scales to much larger datasets due to the use of a linear kernel SVM. In Figure 9, performance of the CO-DAS method is shown where we used only one Co-Training iteration. Left and right panels illustrate the best choice of the amount of selected high confidence patterns for classifier retraining and the DAS fusion parameter selection by a cross-validation procedure, respectively. The results show that performance increase using only one iteration of Co-Training followed by DAS fusion is meaningful if the relatively large amount of top confidence patterns are fed for classifier retraining at low supervision rates. Notice that the cross-validation procedure selected CRFH visual feature at low supervision rate. This may hint for overfitting since the SPH descriptor is a richer visual descriptor.

Co-Training with More Iterations. Interesting additional insights on the Co-Training algorithm can be gained if we perform more than one iteration (see Figure 10). The figures show the evolution of the performance of a single feature classifier after it was iteratively retrained from the standard baseline up to 10 iterations where a constant portion of high confidence estimates were added after each iteration. The plots show an interesting increase of performance with every iteration for both classifiers with the same trend. First, this hints that both initial classifiers are possibly enough bootstrapped with initial training data and the two visual cues are possibly conditionally independent as required for the Co-Training algorithm to function properly. Secondly, we notice a certain saturation after more than 6-7 iterations in most cases which may hint that both classifiers achieved complete agreement levels.

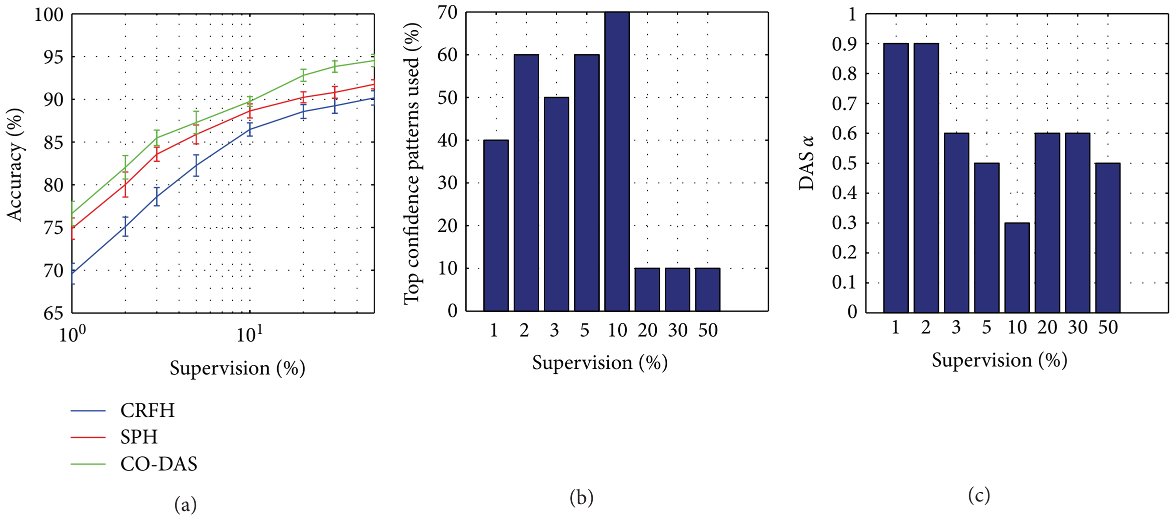


Figure 9: Effect of supervision level on the CO-DAS performance and optimal parameters. (a) Accuracy for CO-DAS and single feature approaches. (b) Optimal amount of selected samples for the Co-Training feedback loop. (c) Selected DAS α parameter for late fusion. (IDOL2 dataset, random setup).

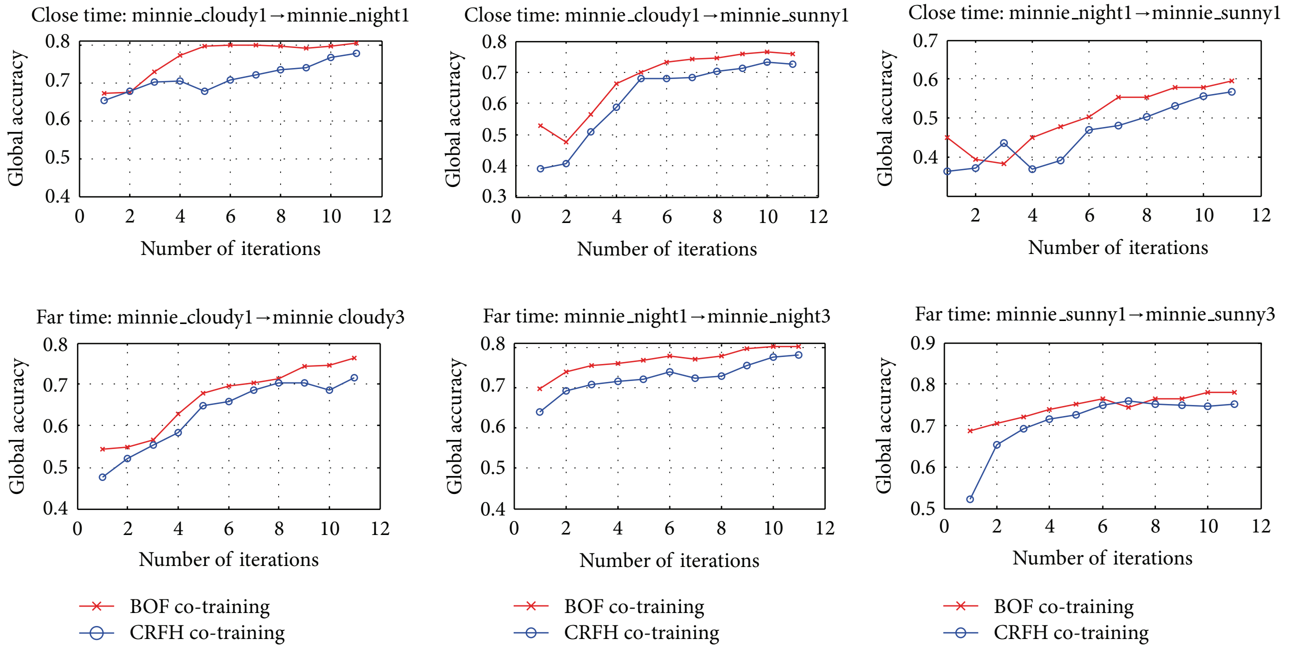


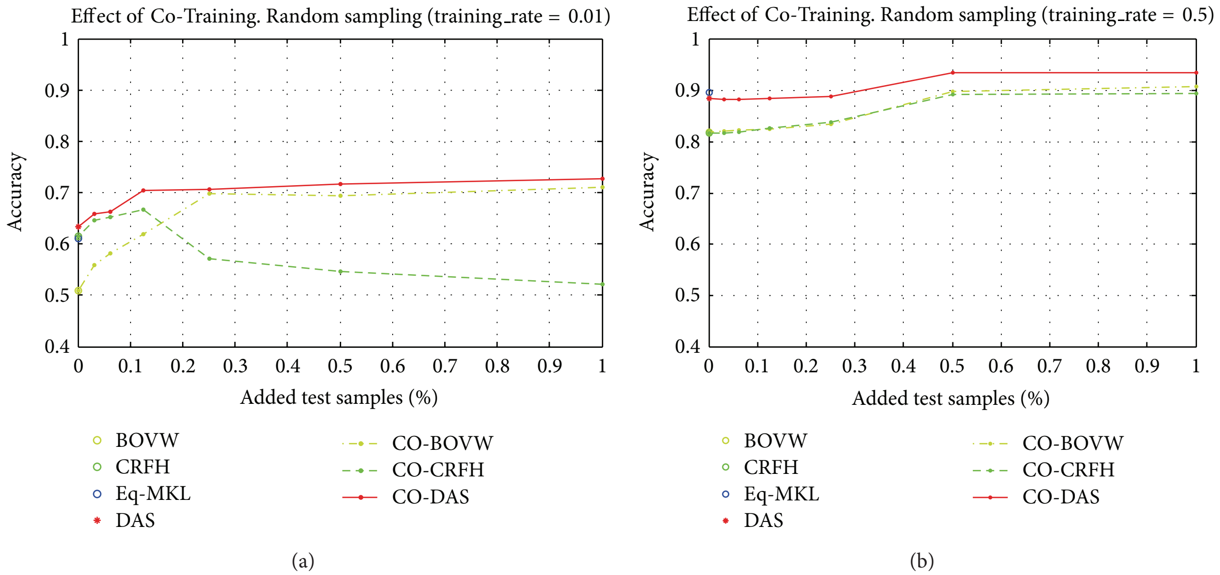
Figure 10: Evolution of the accuracy of individual inner classifiers of the Co-Training module as a function of the number of feedback loop iterations. (IDOL2 dataset, video-versus-video setup). The plots are shown for six sequence pairs: (top) same lighting conditions, (bottom) different lighting conditions.

Conclusion. Experiments carried out this far show that unlabeled data is indeed useful for image-based place recognition. We demonstrated that a better manifold leveraging unlabeled data can be learned using semi-supervised Laplacian SVM with the assumption of low density class separation. This performance comes at high computational cost, large amounts of required memory, and demands careful parameter tuning. This issue is solved by using more efficient Co-Training algorithm, which will be used in the proposed place recognition system.

4.1.5. Random Setup: Comparison of Global Performance

Motivation. Random labeling setup represents the conditions with training patterns being scattered across the database. Randomly labeled images may simulate situation when some small portions of video are annotated in a frame by frame manner. In its extreme, few labeled images from every class may be labeled manually.

In this context, we are interested in the performance of the single feature methods, early and late fusion methods,



(a)

(b)

Figure 11: Comparison of the performance of single features (BOF, CRFH), early (Eq-MKL) and late fusion (DAS) approaches, with details of Co-Training: performances of the individual inner classifiers (CO-BOF, CO-CRFH) and the final fusion (CO-DAS). Plot of the average accuracy as a function of the amount of Co-Training feedback. The performances are plotted for 1% (a) and 50% (b) of labeled data (IDOL2 dataset, random labeling). See text for more detailed explanation.

and the proposed semi-supervised CO-DAS and CO-TA-DAS methods. In order to simulate various supervision levels, the amount of labeled samples varies from a low (1%) to relatively high (50%) proportion of the database. The results depicting these two setups are presented in Figures 11(a) and 11(b), respectively. The early fusion is performed using the MKL by attributing equal weights for both visual features.

Low Supervision Case. The low supervision configuration (Figure 11(a)) is clearly disadvantageous for the single feature methods achieving approximately 50% and 60% of the correct classification for BOVW and CRFH based SVM classifiers. An interesting performance increase can be observed for the Co-Training algorithm leveraging 10% of the top confidence estimates in one re-training iteration, achieving, respectively, 10% and 8% increase for the BOVW and CRFH classifiers. This indicates that the top confidence estimates are not only correct but are also useful for each classifier by improving its discriminatory power on less confident test patterns. Curiously, the performance of the CRFH classifier degrades if more than 10% of high confidence estimates are provided by the BOVW classifier, which may be a sign of increasing the amount of misclassifications being injected. The CO-DAS method successfully performs the fusion of both classifiers and addresses the performance drop in the BOVW classifier, which is achieved by a weighting in favor of the more powerful CRFH classifier.

High Supervision Case. At higher supervision levels (Figure 11(b)), the performance of single feature supervised classifiers is already relatively high reaching around 80% of accuracy for both classifiers, which indicates that a significant amount of visual variability present in the scenes has been captured. This comes as no surprise since at 50% of video annotation in random setup. Nevertheless, the Co-Training algorithm improves the classification by additional 8-9%. An interesting observation for the CO-DAS method shows clearly the complementarity of the visual features even when no Co-Training learning iterations are performed. The high supervision setup permits as much as 50% of the remaining test data annotation for the next re-training rounds before reaching saturation at approximately 94% of accuracy.

Conclusion. These experiments show an interest of using the Co-Training algorithm in low supervision conditions. The initial supervised single feature classifiers need to be provided with sufficient number of training data to bootstrap the iterative re-training procedure. Curiously the initial diversity of initial classifiers determines what performance gain can be obtained using the Co-Training algorithm. This explains why at higher supervision levels the performance increase of a re-trained classifier pair may not be significant. Finally, both early and late fusion methods succeed to leverage the visual feature complementarity but failed to go beyond the Co-Training based methods, which confirms the utility of the unlabeled data in this context.

4.1.6. Video versus Video: Comparison of Global Performance

Motivation. Global performance of the methods may be overly optimistic if annotation is performed only in a random labeling setup. In practical applications a small bootstrap video or a short portion of a video can be annotated instead. We study in a more realistic setup the case with one video being used as training and the place recognition method evaluated on a different video.

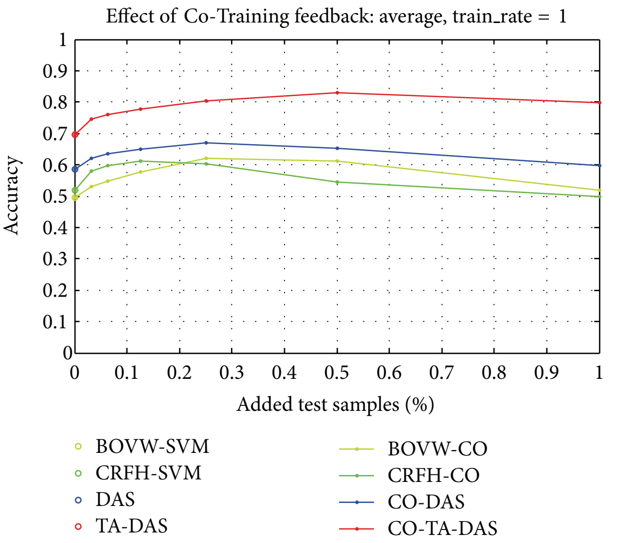


FIGURE 12: Comparison of the global performances for single feature (BOVW-SVM, CRFH-SVM), multiple feature late fusion (DAS), and the proposed extensions using temporal accumulation (TA-DAS) and Co-Training (CO-DAS, CO-TA-DAS). The evolution of the performances of the individual inner classifiers of the Co-Training module (BOVW-CO, CRFH-CO) is also shown. Plot of the average accuracy as a function of the amount of Co-Training feedback. The approaches without Co-Training appear as the limiting case with 0% of feedback (IDOL2 dataset, video-versus-video setup, ALL pairs).

Discussion on the Results. The comparison of the methods in video-versus-video setup is showed in Figure 12. The performances are compared showing the influence of the amount of samples used for the Co-Training algorithm feedback loop. The baseline single feature methods perform around equally by delivering approximately 50% of correct classification. The standard DAS fusion boosts the performance by additional 10%. This confirms the complementarity of the selected visual features in this test setup.

The individual classifiers trained in one Co-Training iteration exceed the baseline and are comparable to performance delivered by standard DAS fusion method. The improvement is due to the feedback of unlabeled patterns in the iterative learning procedure. The CO-DAS method successfully leverages both improvements while the CO-TA-DAS additionally takes advantage of the temporal continuity of the video (a temporal window of size $\tau = 50$ was used).

Confidence Measure. On this dataset, a good illustration concerning the amount of high confidence is showed in Figure 12. It is clear that only a portion of the test set data can be useful for classifier re-training. This is governed by two major factors—quality of the data and robustness of the confidence measure. For this dataset, the best portion of high confidence estimates is around 20–50% depending on the method. The best performing TA-CO-TA-DAS method can afford to annotate up to 50% of testing data for the next learning iteration.

Conclusion. The results show as well that all single feature baselines are outperformed by standard fusion and simple Co-Training methods. The proposed methods CO-DAS and CO-TA-DAS perform the best by successfully leveraging two visual features, temporal continuity of the video, and working in semi-supervised framework.

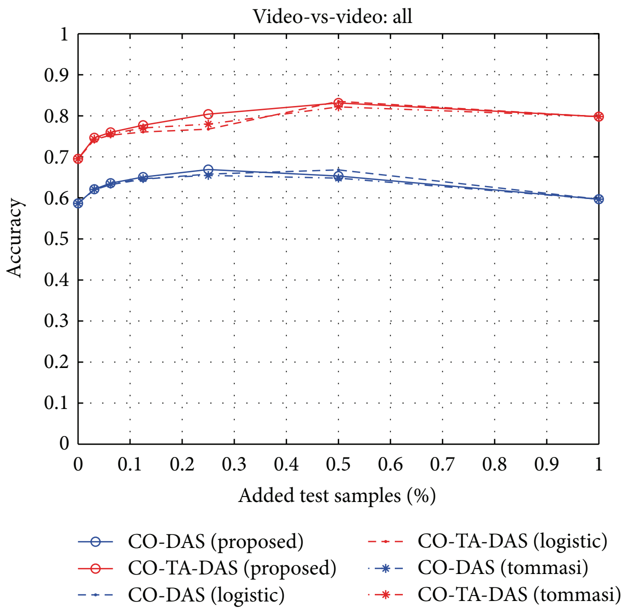


FIGURE 13: Comparison of the performances of the types of confidence measures for the Co-Training feedback loop. Plot of the average accuracy as a function of the amount of Co-Training feedback (video-versus-video setup, ALL pairs).

4.1.7. Effect of the Type of Confidence Measures. Figure 13 represents the effect of the type of confidence measure used in Co-Training on the performances, for different amounts of feedback in the Co-Training phase. The performances for the Ruping approach is not reported, as it was much lower than the other approaches. A video-versus-video setup was used, with the results averaged over all sequence pairs. The three approaches produce a similar behavior with respect to the amount of feedback: first an increase of the performances, when mostly correct estimates are added to the training set, then a decrease when more incorrect estimates are also considered. When coupled with temporal accumulation, the proposed confidence measure has a slightly better accuracy for moderate feedback. It was therefore used for the rest of experiments.

4.2. Results on IMMED

4.2. Results on IMMED. Compared to the IDOL2 database, the IMMED database poses novel challenges. The difficulties arise from increased visual variability changing from location to location, class imbalance due to room visit irregularities, poor lighting conditions, missing or low quality training data, and the large amount of data to be processed.

4.2.1. Description of Dataset. The IMMED database consists of 27 video sequences recorded in 14 different locations in real-world conditions. The total amount of recordings

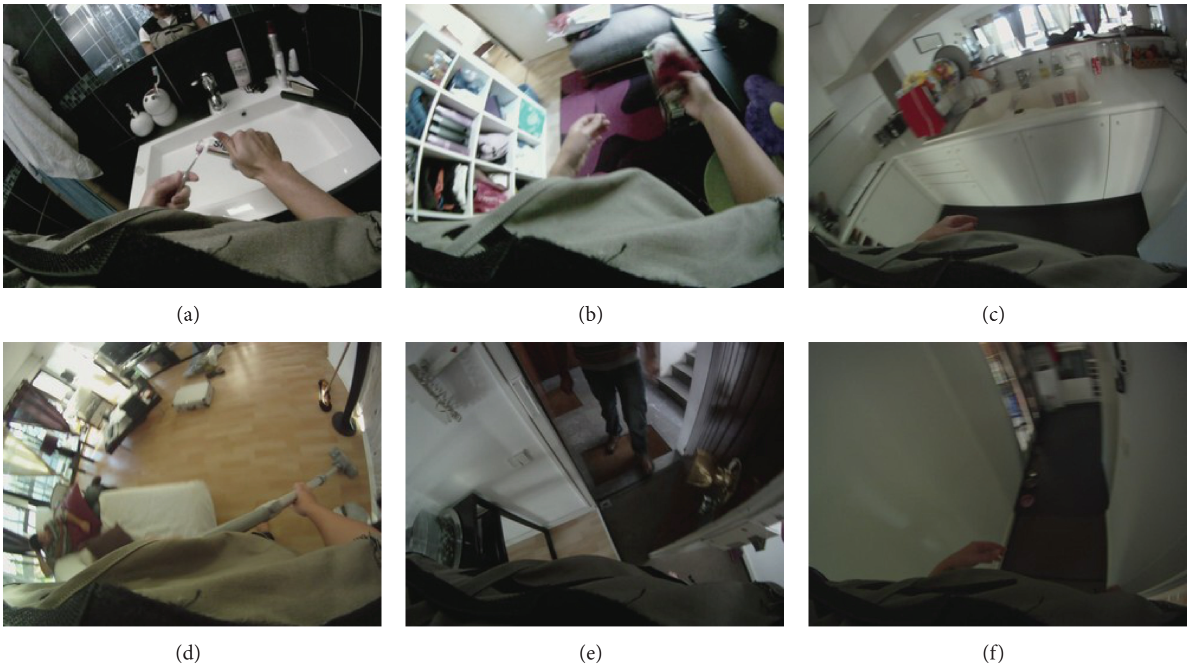


Figure 14: IMMED sample images: (a) bathroom, (b) bedroom, (c) kitchen, (d) living room, (e) outside, and (f) other.

exceeds 10 hours. All recordings were performed using a portable GoPro video camera at a frame rate of 30 frames per second with the frame of the resolution of 1280 × 960 pixels. For practical reasons, we downsampled the frame rate to 5 frames per second. Sample images depicting the 6 topological locations are depicted in Figure 14.

Most locations are represented with one short bootstrap sequence depicting briefly the available topological locations for which manual annotation is provided. One or two longer videos for the same location depict displacements and activities of a person in its ecological and unconstrained environment.

Across the whole corpus, the bootstrap video is typically 3.5 minutes long (6400 images) while the unlabeled evaluation videos are 20 minutes long (36000 images) in average. A few locations are not given a labeled bootstrap video; therefore, a small randomly annotated portion of the evaluation videos covering every topological location is provided instead.

The topological location names in all the video have been equalized such that every frame could carry one of the following labels: “bathroom,” “bedroom,” “kitchen,” “living room,” “outside,” and “other”.

4.2.2. Comparison of Global Performances

Setup. We performed automatic image-based place recognition in realistic video-versus-video setup for each of the 14 locations. To learn optimal parameter values for employed methods, we used the standard cross-validation procedure in all experiments.

Due to a large number of locations, we report here the global performances averaged for all locations. The summary of the results for single and multiple feature methods is provided in Tables 1 and 2, respectively.

Table 1: IMMED dataset: average accuracy of the single feature approaches.

Feature/approach	SVM	SVM-TA
BOVW	0.49	0.52
CRFH	0.48	**0.53**
SPH	0.47	0.49

Baseline-Single Feature Classifier Performance. As show in Table 1, single feature methods provide relatively low place recognition performance. Surprisingly potentially the more discriminant descriptor SPH is less performant than its more simple BOVW variant. A possible explanation to this phenomena may be that due to the low amount of supervision and a classifier trained on high dimensional SPH features simply overfits.

Temporal Constraints. An interesting gain in performance is obtained if temporal information is enforced. On the whole corpus, this performance increase ranges from 2 to 4% in global classification accuracy for all single feature methods. We observe the same order of improvement in multiple feature methods as MKL-TA, for early feature fusion, and DAS-TA, for late classifier fusion. This performance increase over the single feature baselines is constant for the whole corpus and all methods.

Multiple Feature Exploitation. Comparing MKL and DAS methods for multiple feature fusion shows interest in favor

Table 2: IMMED dataset: average accuracy of the multiple feature approaches.

Feature/approach	MKL	MKL-TA	DAS	DAS-TA	CO-DAS	CO-TA-DAS
BOVW-SPH	0.48	0.50	0.51	**0.56**	0.50	0.53
BOVW-CRFH	0.50	0.54	0.51	0.56	0.54	**0.58**
SPH-CRFH	0.48	0.51	0.50	0.54	0.54	**0.57**
BOVW-SPH-CRFH	0.48	0.51	0.51	**0.56**	—	—

of the late fusion method when compared to single feature methods. We observe little performance improvement when using MKL, which can be explained by increased dimensionality space and thus more risk of overfitting. Late fusion strategy is more advantageous compared to respective single feature methods in this low supervision setup by bringing up to 4% with no temporal accumulation and up to 5% with temporal accumulation. Therefore multiple feature information is best leveraged in this context by selecting late classifier fusion.

Leveraging the Unlabeled Data. Exploitation of unlabeled data in the learning process is important when it comes to low amounts of supervision and great visual variability encountered in challenging video sequences. The first proposed method termed CO-DAS aims to leverage two visual features while operating in semi-supervised setup. It clearly outperforms all single feature methods and improves on all but BOVW-SPH feature pair compared to DAS by up to 4%. We explain this performance increase by successfully leveraged visual feature complementarity and improved single feature classifiers via Co-Training procedure. The second method CO-TA-DAS incorporates temporal continuity a priori and boosts performances by another 3-4% in global accuracy. This method effectively combines all benefits brought by individual features, temporal video continuity, and taking advantage of unlabeled data.

5. Conclusion

In this work we have addressed the challenging problem of indoor place recognition from wearable video recordings. Our proposition was designed by combining several approaches, in order to deal with issues such as low supervision and large visual variability encountered in videos from a mobile camera. Their usefulness and complementarity were verified initially on a public video sequence database IDOL2 then applied to the more complex and larger scale corpus of videos collected for the IMMED project which contains real-world video lifelogs depicting actual activities of patients at home.

The study revealed several elements that were useful for successful recognition in such video corpuses. First, the usage of multiple visual features was shown to improve the discrimination power in this context. Second, the temporal continuity of a video is a strong additional cue, which improved the overall quality of indexing process in most cases. Third, real-world video recordings are rarely annotated manually to an extent where most visual variability present within a location is captured. Usage of semi-supervised learning algorithms exploiting labeled as well as unlabeled data helped to address this problem. The proposed system integrates all acquired knowledge in a framework which is computationally tractable, yet takes into account the various sources of information.

We have addressed the fusion of multiple heterogeneous sources of information for place recognition from complex videos and demonstrated its utility on the challenging IMMED dataset recorded in real-world conditions. The main focus of this work was to leverage the unlabeled data thanks to a semi-supervised strategy. Additional work could be done in selecting more discriminant visual features for specific applications and more tight integration of the temporal information in the learning process. Nevertheless, the obtained results confirm the applicability of the proposed place classification system on challenging visual data from wearable videos.

Acknowledgments

This research has received funding from Agence Nationale de la Recherche under Reference ANR-09-BLAN-0165-02 (IMMED project) and the European Community's Seventh Framework Programme (FP7/2007–2013) under Grant Agreement 288199 (Dem@Care project).

References

[1] A. Doherty and A. F. Smeaton, "Automatically segmenting lifelog data into events," in *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '08)*, pp. 20–23, May 2008.

[2] E. Berry, N. Kapur, L. Williams et al., "The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: a preliminary report," *Neuropsychological Rehabilitation*, vol. 17, no. 4-5, pp. 582–601, 2007.

[3] S. Hodges, L. Williams, E. Berry et al., "SenseCam: a retrospective memory aid," in *Proceedings of the 8th International Conference on Ubiquitous Computing (Ubicomp '06)*, pp. 177–193, 2006.

[4] R. Mégret, D. Szolgay, J. Benois-Pineau et al., "Indexing of wearable video: IMMED and SenseCAM projects," in *Workshop on Semantic Multimodal Analysis of Digital Media*, November 2008.

[5] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, vol. 1, pp. 273–280, October 2003.

[6] A. Quattoni and A. Torralba, “Recognizing indoor scenes,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’09)*, pp. 413–420, June 2009.

[7] R. Mégret, V. Dovgalecs, H. Wannous et al., “The IMMED project: wearable video monitoring of people with age dementia,” in *Proceedings of the International Conference on Multimedia (MM ’10)*, pp. 1299–1302, ACM Request Permissionss, October 2010.

[8] S. Karaman, J. Benois-Pineau, R. Mégret, V. Dovgalecs, J.-F. Dartigues, and Y. Gaëstel, “Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases,” in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR ’10)*, pp. 4113–4116, August 2010.

[9] C. Schüldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local SVM approach,” in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR ’04)*, pp. 32–36, August 2004.

[10] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, October 2005.

[11] L. Ballan, M. Bertini, A. del Bimbo, and G. Serra, “Video event classification using bag of words and string kernels,” in *Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP ’09)*, pp. 170–178, 2009.

[12] D. I. Kosmopoulos, N. D. Doulamis, and A. S. Voulodimos, “Bayesian filter based behavior recognition in workflows allowing for user feedback,” *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 422–434, 2012.

[13] M. Stikic, D. Larlus, S. Ebert, and B. Schiele, “Weakly supervised recognition of daily life activities with wearable sensors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2521–2537, 2011.

[14] D. H. Nguyen, G. Marcu, G. R. Hayes et al., “Encountering SenseCam: personal recording technologies in everyday life,” in *Proceedings of the 11th International Conference on Ubiquitous Computing (Ubicomp ’09)*, pp. 165–174, ACM Request Permissions, September 2009.

[15] M. A. Perez-QuiNones, S. Yang, B. Congleton, G. Luc, and E. A. Fox, “Demonstrating the use of a SenseCam in two domains,” in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL ’06)*, p. 376, June 2006.

[16] S. Karaman, J. Benois-Pineau, V. Dovgalecs et al., Hierarchical Hidden Markov Model in Detecting Activities of Daily Living in Wearable Videos for Studies of Dementia, 2011.

[17] J. Pinquier, S. Karaman, L. Letoupin et al., “Strategies for multiple feature fusion with Hierarchical HMM: application to activity recognition from wearable audiovisual sensors,” in *Proceedings of the 21 International Conference on Pattern Recognition*, pp. 1–4, July 2012.

[18] N. Sebe, M. S. Lew, X. Zhou, T. S. Huang, and E. M. Bakker, “The state of the art in image and video retrieval,” in *Proceedings of the 2nd International Conference on Image and Video Retrieval*, pp. 1–7, May 2003.

[19] S.-F. Chang, D. Ellis, W. Jiang et al., “Large-scale multimodal semantic concept detection for consumer video,” in *Proceedings of the International Workshop on Multimedia Information Retrieva (MIR ’07)*, pp. 255–264, ACM Request Permissions, September 2007.

[20] J. Košecká, F. Li, and X. Yang, “Global localization and relative positioning based on scale-invariant keypoints,” *Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 27–38, 2005.

[21] C. O. Conaire, M. Blighe, and N. O’Connor, “Sensecam image localisation using hierarchical surf trees,” in *Proceedings of the 15th International Multimedia Modeling Conference (MMM ’09)*, p. 15, Sophia-Antipolis, France, January 2009.

[22] J. Košecká, L. Zhou, P. Barber, and Z. Duric, “Qualitative image based localization in indoors environments,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. II-3–II-8, June 2003.

[23] Z. Zovkovic, O. Booij, and B. Krose, “From images to rooms,” *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 411–418, 2007.

[24] L. Fei-Fei and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’05)*, pp. 524–531, June 2005.

[25] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’06)*, vol. 2, pp. 2161–2168, 2006.

[26] O. Linde and T. Lindeberg, “Object recognition using composed receptive field histograms of higher dimensionality,” in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR ’04)*, vol. 2, pp. 1–6, August 2004.

[27] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’06)*, vol. 2, pp. 2169–2178, 2006.

[28] A. Bosch and A. Zisserman, “Scene classification via pLSA,” in *Proceedings of the 9th European Conference on Computer Vision (ECCV ’06)*, May 2006.

[29] J. Sivic and A. Zisserman, “Video google: a text retrieval approach to object matching in videos,” in *Proceedings of the 9th IEEE International Conference On Computer Vision*, pp. 1470–1477, October 2003.

[30] J. Knopp, *Image Based Localization [Ph.D. thesis]*, Chech Technical University in Prague, Faculty of Electrical Engineering, Prague, Czech Republic, 2009.

[31] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, “A solution to the simultaneous localization and map building (SLAM) problem,” *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229–241, 2001.

[32] L. M. Paz, P. Jensfelt, J. D. Tardós, and J. Neira, “EKF SLAM updates in O(n) with divide and conquer SLAM,” in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA ’07)*, pp. 1657–1663, April 2007.

[33] J. Wu and J. M. Rehg, “CENTRIST: a visual descriptor for scene categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[34] H. Bulthoff and A. Yuille, “Bayesian models for seeing shapes and depth,” Tech. Rep. 90-11, Harvard Robotics Laboratory, 1990.

[35] P. K. Atrey, M. Anwar Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: a survey,” *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[36] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, “SimpleMKL,” *The Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[37] S. Nakajima, A. Binder, C. Müller et al., "Multiple kernel learning for object classification," in *Workshop on Information-based Induction Sciences*, 2009.

[38] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 606–613, October 2009.

[39] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao, "Group-sensitive multiple kernel learning for object categorization," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 436–443, October 2009.

[40] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 902–909, Laboratoire Jean Kuntzmann, LEAR, INRIA Grenoble, June 2010.

[41] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao, "Multiple kernel active learning for image classification," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '09)*, pp. 550–553, July 2009.

[42] A. Abdullah, R. C. Veltkamp, and M. A. Wiering, "Spatial pyramids and two-layer stacking SVM classifiers for image categorization: a comparative study," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '09)*, pp. 5–12, June 2009.

[43] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[44] L. Ilieva Kuncheva, *Combining Pattern Classifiers. Methods and Algorithms*, Wiley-Interscience, 2004.

[45] A. Uhl and P. Wild, "Parallel versus serial classifier combination for multibiometric hand-based identification," in *Proceedings of the 3rd International Conference on Advances in Biometrics (ICB '09)*, vol. 5558, pp. 950–959, 2009.

[46] W. Nayer, *Feature based architecture for decision fusion [Ph.D. thesis]*, 2003.

[47] M.-E. Nilsback and B. Caputo, "Cue integration through discriminative accumulation," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. II578–II585, July 2004.

[48] A. Pronobis and B. Caputo, "Confidence-based cue integration for visual place recognition," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '07)*, pp. 2394–2401, October-November 2007.

[49] A. Pronobis, O. Martinez Mozos, and B. Caputo, "SVM-based discriminative accumulation scheme for place recognition," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA '08)*, pp. 522–529, May 2008.

[50] F. Lu, X. Yang, W. Lin, R. Zhang, R. Zhang, and S. Yu, "Image classification with multiple feature channels," *Optical Engineering*, vol. 50, no. 5, Article ID 057210, 2011.

[51] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proceedings of the 12th International Conference on Computer Vision*, pp. 221–228, October 2009.

[52] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Department of Computer Sciences, University of Winsconsin, Madison, Wis, USA, 2008.

[53] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*, Morgan and Claypool Publishers, 2009.

[54] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, Mass, USA, 2006.

[55] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[56] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[57] D. Zhou, O. Bousquet, T. Navin Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems*, vol. 16, pp. 321–328, 2004.

[58] S. Melacci and M. Belkin, "Laplacian support vector machines trained in the primal," *The Journal of Machine Learning Research*, vol. 12, pp. 1149–1184, 2011.

[59] B. Nadler and N. Srebro, "Semi-supervised learning with the graph laplacian: the limit of infinite unlabelled data," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS '09)*, 2009.

[60] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT' 98)*, pp. 92–100, October 1998.

[61] D. Zhang and W. Sun Lee, "Validating co-training models for web image classification," in *Proceedings of SMA Annual Symposium*, National University of Singapore, 2005.

[62] W. Tong, T. Yang, and R. Jin, "Co-training For Large Scale Image Classification: An Online Approach," *Analysis and Evaluation of Large-Scale Multimedia Collections*, pp. 1–4, 2010.

[63] M. Wang, X.-S. Hua, L.-R. Dai, and Y. Song, "Enhanced semi-supervised learning for automatic video annotation," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '06)*, pp. 1485–1488, July 2006.

[64] V. E. van Beusekom, I. G. Sprinkuizen-Kuyper, and L. G. Vuurpul, "Empirically evaluating co-training," Student Report, 2009.

[65] W. Wang and Z.-H. Zhou, "Analyzing co-training style algorithms," in *Proceedings of the 18th European Conference on Machine Learning (ECML '07)*, pp. 454–465, 2007.

[66] C. Dong, Y. Yin, X. Guo, G. Yang, and G. Zhou, "On co-training style algorithms," in *Proceedings of the 4th International Conference on Natural Computation (ICNC '08)*, vol. 7, pp. 196–201, October 2008.

[67] S. Abney, *Semisupervised Learning for Computational Linguistics*, Computer Science and Data Analysis Series, Chapman & Hall, University of Michigan, Ann Arbor, Mich, USA, 2008.

[68] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (ACL '95)*, pp. 189–196, University of Pennsylvania, 1995.

[69] W. Wang and Z.-H. Zhou, "A new analysis of co-training," in *Proceedings of the 27th International Conference on Machine Learning*, pp. 1135–1142, May 2010.

[70] C. M. Bishop, *Pattern Recognition and Machine Learning. Information Science and Statistics*, Springer, Secaucus, NJ, USA, 2006.

[71] B. Scholkopf and A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, Mass, USA, 2002.

[72] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: a library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[73] A. J. Smola, B. Schölkopf, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[74] A. Pronobis, O. Martínez Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 298–320, 2010.

[75] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and predictionvolume," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[76] S. Rüping, A Simple Method For Estimating Conditional Probabilities For SVMs. *American Society of Agricultural Engineers*, 2004.

[77] T. Tommasi, F. Orabona, and B. Caputo, "An SVM confidence-based approach to medical image annotation," in *Proceedings of the 9th Cross-Language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access (CLEF '08)*, pp. 696–703, 2009.

[78] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1458–1465, October 2005.

[79] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "The KTH-IDOL2 database," Tech. Rep., Kungliga Tekniska Hoegskolan, CVAP/CAS, 2006.