Histogram of Oriented Uniform Patterns for robust place recognition and categorization
Ehsan Fazl-Ersi and John K Tsotsos
The International Journal of Robotics Research published online 25 January 2012
DOI: 10.1177/0278364911434936

The online version of this article can be found at:
http://ijr.sagepub.com/content/early/2012/01/24/0278364911434936

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Multimedia Archives

Additional services and information for The International Journal of Robotics Research can be found at:

Email Alerts: http://ijr.sagepub.com/cgi/alerts
Subscriptions: http://ijr.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> OnlineFirst Version of Record - Jan 25, 2012

What is This?
Histogram of Oriented Uniform Patterns for robust place recognition and categorization

Ehsan Fazl-Ersi and John K Tsotsos

Abstract
This paper presents a novel context-based scene recognition method that enables mobile robots to recognize previously observed topological places in known environments or categorize previously unseen places in new environments. We achieve this by introducing the Histogram of Oriented Uniform Patterns (HOUP), which provides strong discriminative power for place recognition, while offering a significant level of generalization for place categorization. HOUP descriptors are used for image representation within a subdivision framework, where the size and location of sub-regions are determined using an informative feature selection method based on kernel alignment. Further improvement is achieved by developing a similarity measure that accounts for perceptual aliasing to eliminate the effect of indistinctive but visually similar regions that are frequently present in outdoor and indoor scenes. An extensive set of experiments reveals the excellent performance of our method on challenging categorization and recognition tasks. Specifically, our proposed method outperforms the current state of the art on two place categorization datasets with 15 and 5 place categories, and two topological place recognition datasets, with 5 and 27 places.

Keywords
visual place categorization, topological place recognition, Histogram of Oriented Uniform Patterns, perceptual aliasing, kernel alignment

1. Introduction
One of the most fundamental requirements for an autonomous mobile robot is localization, that is, the capability of knowing where it is located within its environment. Various research themes have been defined to study different aspects of the localization problem. Topological place recognition and place categorization (also referred to as semantic place labeling) are among the most important ones. While topological place recognition gives the robot the capability of recognizing previously observed places in known environments, place categorization aims at enabling the robot to classify different locations of a new environment into a set of pre-specified categories (e.g. ‘Living room’, ‘Kitchen’ and ‘Bedroom’ for indoor environments) relating them to human-understandable concepts.

The difficulties with traditional robot sensors in dealing with such high-level problems prompted the research towards using more information-rich sensors, such as vision. Within computer vision, robot localization can be seen as the problem of scene recognition, where the key issue is to find an appropriate image representation that is invariant to common changes in dynamic environments (e.g. lighting condition, viewpoint, partial occlusion, etc.) and robust against intra-class variations.

The focus of this paper is to introduce the Histogram of Oriented Uniform Patterns (HOUP), a new scene descriptor for visual place recognition and categorization. For a given image region, its HOUP descriptor is built by performing detailed analysis on the output amplitude of a set of oriented band-pass filters at different locations within the image region. Such detailed analysis is performed using uniform patterns (Ojala et al., 2002) – a specific type of local binary pattern (LBP) – which encodes primitive textual features, such as different types of curved edges, spots, flat regions, etc.

HOUP descriptors are used for image representation within the subdivision framework (i.e. partitioning the image into sub-blocks and computing descriptors for these sub-blocks), which has been used by many authors for...
global non-invariant representations. However, unlike previous methods that use fixed sub-regions and perform trial-and-error heuristics to determine the right subdivision scheme, in our method a feature selection technique based on kernel alignment (Cristianini et al., 2002) is suggested that allows only the most informative image regions (i.e. those that best separate different classes) to contribute to the image representation. The developed feature selection has the advantages of being based only on training data (i.e. no cross-validation is required), and explicitly dealing with multiple classes – unlike the vast majority of available solutions that are designed for two classes (to separate a set of positive samples from negative ones) and require auxiliary techniques (such as one-versus-one or one-versus-all) to be extended to multi-class feature selection.

We further develop a similarity measure for comparing HOUP descriptors, which also takes into account the individual distinctiveness of the descriptors. This is to achieve robustness against perceptual aliasing (Whitehead and Ballard, 1991), so that the descriptors that are similar to each other but are relatively common in the database (e.g. those extracted from repetitive features in the environment, such as bushes, walls, ceiling or floor surfaces, etc.) are assigned a low similarity score.

Our experiments show that while our method outperforms the state of the art on four publicly available place recognition and categorization datasets, it produces image representations that are significantly more compact than those produced by other methods.

The remainder of this paper is organized as follows. In the next section, we review related methods. Section 3 introduces the HOUP descriptors and describes the similarity measure developed to compare them. Our scene representation method based on HOUP descriptors is presented in Section 4, followed by experiments and results in Section 5. Finally, Section 6 concludes this paper by discussing the method and outlining some of the potential directions for future work.

2. Related work

The majority of early works on place categorization (and recognition) use laser range finders to perceive the environment. These methods usually use shape features (describing the geometric layout of the surrounding) to classify different locations into a set of pre-specified place categories. Examples of such methods are the works of Martinez-Mozos et al. (2005) Martinez-Mozos and Burgard (2006) and Friedman et al. (2007), where a set of binary classifiers is trained to recognize specific places, such as ‘Room’, ‘Corridor’ and ‘Doorway’, in the environment. Binary classifiers are often built by boosting simple geometric features using the AdaBoost algorithm (Freund and Schapire, 1995), where each simple geometric feature is a numerical value, computed from the observed beams of a laser range scan, or from a polygon representation of the area covered by these observed beams. During the mapping, the robot moves around, classifies its sensor reading data into one of the learned place categories, and labels its position, according to the label of the activated class. Given that laser beams can be simulated off-line from metric maps, these methods compute place categories for almost every point in the free space, resulting in a labeled topological representation for the environment.

Using laser range scans as observations allows these methods to recognize only a certain type of place (e.g. they are not able to distinguish between places with similar geometric structure). Rottmann et al. (2005) proposed a method that combines laser range features with visual features to enable the robot to support a greater variety of place categories. Motivated by the fact that typical objects appear at different places with different probabilities, they defined visual features as the number of instances of certain categories of objects (including ‘Monitor’, ‘Coffee machine’, ‘Office cupboard’, ‘Face’ and ‘Pedestrian’) observed in the environment. For this purpose, a fast object detector was built for each of the considered object categories, using the object detection method of Viola and Jones (2001). The visual features, along with laser features, are used to classify each location in the environment visited by the robot.

Similar to the work of Rottman et al. (2005), several other recent approaches to place categorization (e.g. Galindo et al., 2005; Vasudevan et al., 2007; Zender et al., 2008) are based on the occurrence statistics of different objects in different places. However, these methods often fail to generalize to new environments that are unseen during the training phase. This is mainly due to the fact that object detection, for the most part, is still an unsolved problem and cannot reasonably deal with the intra-class variations in the appearance of objects. Furthermore, in visual place categorization (and visual place recognition), image data is gathered by autonomous robots without the guidance of the human attention mechanism. Therefore, the informative objects that reveal the place categories are often not sufficiently visible in many image frames, causing the object detection to fail. Finally, the use of occurrence statistics of objects in place representation often leads to ambiguities arising from common objects (e.g. chairs, lamps) being observed in different places (e.g. bedrooms, kitchens).

Given the difficulties with object-based methods, another stream of work suggests that place categories can be estimated from the global configurations in the observed scenes without explicitly detecting and recognizing objects. Such methods can be classified into two general categories: context based and landmark based. Amongst context-based methods is that of Oliva and Torralba (2001), which uses the Discrete Fourier Transform to encode spectral information of the image. The spectral signals from non-overlapping sub-blocks are then compressed to produce the image representation. Torralba et al. (2003) extended this work by using wavelet image decomposition instead of Discrete...
Fourier Transform to produce more compact and precise image representations. They evaluated the performance of their scene representation method for recognizing place categories from data collected by a mobile system and reported reasonable accuracy in recognizing place categories, such as ‘Conference room’, ‘Corridor’ and ‘Office’, which have a lower range of intra-class variations. Recently, Wu and Rehg (2011) proposed the CENTRIST descriptors, which use the Census Transform to capture spatial relations between neighboring pixels. They showed that a spatial hierarchy of such descriptors used with Support Vector Machine (SVM) classifiers performs very well on recognizing a wide variety of place categories.

In landmark-based approaches (e.g. Lazebnik et al., 2006; Bosch et al., 2008; Pronobis et al., 2008; Fazl-Ersi et al., 2009), local image features play the main role in scene recognition. Local features characterize a limited area of the image. However, they usually provide more robustness against common image variations (e.g. viewpoint). Among local feature extraction techniques, the Scale Invariant Feature Transform (SIFT) of Lowe (2004) has dominated the field. Local image features are usually used for scene recognition within the bag-of-features framework, where only the appearances of features are used and their spatial coordinates are discarded. In this framework, the extracted features from the image are matched to a vocabulary of visual words (each representing a category of local image features that are visually similar to each other), resulting in a response vector indicating the frequency of each visual word in the image. Several extensions have been proposed to this basic approach. Lazebnik et al. (2006) proposed the spatial pyramid matching method, which is based on global geometric correspondence. The method works by partitioning the image into increasingly fine sub-blocks and building a bag-of-features representation from each sub-block. The local representations are combined in a principled way to produce the image representation. Bosch et al. (2008) further improved the performance of bag-of-features by proposing a model that builds an intermediate representation based on the probabilistic Latent Semantic Analysis (pLSA) algorithm to produce more distinctive image representation.

While in general, landmark-based methods perform more accurately than context-based methods in scene category recognition, their major drawback is their high dimensionality. Usually high performance is achieved when large number of visual words and sub-blocks are used to produce the image representations. For example, the methods of Lazebnik et al. (2006) and Bosch et al. (2008) work best when images are represented with vectors of dimensionally ~8K and ~25K, respectively.

In this paper we propose a novel context-based scene representation method by introducing the HOUP as a new visual descriptor. HOUP descriptors, similar to the majority of the context-based representation methods (e.g. Oliva and Torralba 2001; Torralba, 2003; Torralba et al., 2003), use oriented band-pass filters to encode spatial information in the scenes, resulting in a huge amount of information about different scene attributes. However, unlike those methods, which use simple statistics, HOUP descriptors perform detailed analysis on the output of the oriented band-pass filters using uniform patterns (discussed in detail in Section 3), to aggregate the encoded information at different locations into a low-dimensional image representation. This paper argues that while the suggested aggregation method based on the uniform patterns boosts the discriminative power and the generalizability of the representations, it produces scene representations with lower dimensionality than the state-of-the-art landmark-based methods.

3. Histogram of Oriented Uniform Patterns

Empirical evidence has shown that the human visual system uses oriented band-pass filters in the early stages of the visual pathway to analyze scenes (Hubel and Wiesel, 1968). Several computational studies have also found this choice of features useful for a wide variety of computer vision applications. Examples are the works of Oliva and Torralba (2001), Torralba (2003) and Torralba et al. (2003), proposing a computational model for computing an abstract representation for scenes (referred to as ‘Gist’), using features such as steerable pyramids and Gabor filters. In this work, similar to the method of Torralba et al. (2003), images are initially encoded using Gabor filters:

$$v_k(x) = \sum_{x'} |i(x')g_k(x - x')|$$ (1)

where $i(x)$ is the input image, $g_k(x)$ are oriented band-pass filters tuned to different orientations at a certain spatial frequency and $v_k(x)$ are the output amplitudes of the filters at the location $x$. While computing a set of Gabor coefficients for all image pixels results in a huge amount of information about different scene attributes, using this information to produce a distinctive yet low-dimensional visual descriptor for the whole image (or its individual sub-blocks) is a major challenge. In this work, rather than taking simple statistics (as employed by Torralba et al., 2003), detailed analysis is performed on the computed coefficients and their joint distributions using uniform patterns.

Uniform patterns are a specific type of LBPs, proposed by Ojala et al. (2002) for grayscale texture classification. The original LBP operator (Ojala et al., 1996) labels each image pixel by subtracting the intensity at that pixel from the intensity at each of its eight neighboring pixels and converting the thresholded results ($\text{threshold} = 0$) to a base-10 number (see Figure 1 for an example). A texture descriptor for the image can then be formed by aggregating the pixel labels into a histogram with $2^b$ bins (one bin for each possible binary pattern that can be formed by eight pixels).

Among all the possible binary patterns for a given neighborhood, a limited number of them occur more frequently in
texture images than others; these binary patterns are called uniform patterns, which often encode primitive texture patterns, such as different types of curved edges, spots, flat regions, etc. Uniform patterns can be identified as LBPs with at most two bitwise transitions (or discontinuities) in the circular presentation of the pattern (see Figure 1 for an example). When using a $3 \times 3$ neighborhood, only 58 of the 256 total patterns are uniform, which yields in a 59-dimensional image representation (i.e. histogram), one dimension for each uniform pattern and one dimension for all the non-uniform patterns.

Computing the histograms of the uniform patterns for the output of each oriented band-pass filter and concatenating them together produces a global representation of the image (or image sub-region), called the HOUP. In our experiments, we compute the Gabor coefficients at six orientations,\(^1\) which yield $6 \times 59 = 354$ dimensional representations (see Figure 2 for an example). To obtain more compact representations, the dimensionality of HOUP descriptors is reduced by projecting them on to the first $M$ principal components, computed from a dataset of training images, where $M$ is chosen such that about 95% of the sum of all eigenvalues in the training data is accounted for by the eigenvalues of the chosen principal components (in our experiments we observed that around 70 principal components are often sufficient to satisfy this condition).

### 3.1. Comparing HOUP descriptors

There is a wide variety of similarity measures that can be used to compare the HOUP descriptors, some of them are general (not descriptor specific), while some are learned to fit available training data. One problem with the majority of the available similarity measures is that they do not explicitly deal with the perceptual aliasing problem: the fact that visually similar objects may appear in the same location in images from different categories or places. An example of perceptual aliasing is illustrated in Figure 3, where several images from different categories have visually similar ‘sky’ regions at a certain sub-block. Comparing each pair of these images using conventional measures, a high similarity score is obtained between descriptors extracted from this sub-block, while in fact the similarities are due to perceptual aliasing.

This work tries to achieve robustness to perceptual aliasing by using a variant of the One-Shot Similarity (OSS) measure of Wolf et al. (2009). Given a pair of HOUP descriptors, the Linear Discriminant Analysis (LDA) algorithm is used to learn a model for each of the descriptors (as single positive samples) against a set of examples $A$. Each of the two learned models is applied on the other descriptor to obtain a likelihood score. The two estimated scores are then combined to compute the overall similarity score between the two descriptors:

$$
S_n(x^n_i, x^n_j) = (x^n_i - \mu_A)^T S_A^{-1} (x^n_i - \mu_A) + (x^n_j - \mu_A)^T S_A^{-1} (x^n_j - \mu_A)
$$

where $\mu_A$ and $S_A$ are the mean and covariance of $A$, respectively.

In the OSS method, $A$ is a fixed set of background examples (i.e. samples from classes other than those to be recognized or classified), to measure the extent to which the two descriptors being compared are closer to one another and different from the typical examples in the background set. In contrast, this paper argues that by replacing $A$ with the complete training set, it is possible to take into account the distinctiveness of the descriptors, which is the key to achieve robustness against perceptual aliasing.\(^2\) If two descriptors are similar to each other but are indistinctive and relatively common in the dataset (e.g. those extracted from repetitive features in the environment, such as bushes, sky regions, walls and ceiling/floor surfaces), they receive a low similarity score (see Figure 4(a) for an example). This is because the individual models that have been learned for the two descriptors cannot separate them well from the typical examples in $A$ and therefore they return lower similarity scores when applied to one another. On the other hand, when two descriptors are distinctive but have lower similarity than the examples of perceptual aliasing, they are still assigned a high similarity score, since they can be separated better from the examples in $A$ (see Figure 4(b) for an example).

### 4. Scene representation

It is essential for a scene representation method to capture structural properties and rough geometrical constraints in the scenes. Spatial structures are very useful in revealing the scene category. Many structural properties can be reflected in the distribution of the primitive patterns (horizontal edges, vertical edges, etc.) and their spatial layout. Rough geometrical constraints are also very helpful in recognizing scene categories. For example, constraints on objects’ elevation (as a function of ground level) can help reduce ambiguity, even when the images are taken from random viewpoints. HOUP descriptors, when computed from the

---

\(^1\)\(^\text{Fig. 1. An example of the local binary pattern (LBP) operator. By using the value of the center pixel as threshold, the neighboring pixels are converted to binary codes 0 and 1, forming an ordered local binary pattern for the center pixel. Given that there are only two bitwise transitions in this pattern, it is a uniform one.}\)

\(^2\)\(^\text{Table}\): 52 28 31 0 1 0 1 0 00111100\rangle_2 = 60

\[\begin{array}{ccc}
52 & 28 & 31 \\
73 & 47 & 17 \\
96 & 66 & 14 \\
\end{array}\]

\[\begin{array}{ccc}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 1 & 0 \\
\end{array}\]
Fig. 2. The processes involved in computing a Histogram of Oriented Uniform Patterns (HOUP) descriptor for a sample image (a). First the output amplitudes of a set of Gabor filters tuned to six different orientations are computed ((b), top row). Next, detailed analysis on the output of oriented Gabor filters is performed by applying the binary operators corresponding to uniform patterns, which label each pixel at each orientation with a value between 0 and 58 ((b), middle row). The HOUP descriptor is then produced by computing the histograms of the uniform patterns for the output of each oriented Gabor filter ((b), bottom row) and concatenating them together. This results in a 354-dimensional description vector for the image, which is subsequently reduced to $\sim 70$ dimensions through Principal Component Analysis (PCA).

Fig. 3. Example of perceptual aliasing in the UIUC scene category recognition database. Visually similar ‘sky’ regions are present in images of different categories (from left to right: ‘Suburb’, ‘Coast’, ‘Forest’, ‘Mountain’, ‘Open country’ and ‘Street’) at the same sub-block (in a $3 \times 3$ regular grid). Comparing each pair of these images using conventional measures, a high similarity score is obtained between descriptors extracted from this sub-block, while in fact the similarities are due to perceptual aliasing.

Fig. 4. Similarity scores computed between individual Histogram of Oriented Uniform Patterns (HOUP) descriptors in two sample pairs of images. As can be seen, while the highlighted regions in (a) are visually more similar to each other than those in (b), the estimated similarity score is significantly lower for (a). This is because the similarity measure used in our method accounts for perceptual aliasing, preventing two similar descriptors that are indistinctive and relatively common in the dataset from receiving a high similarity score.

entire image, discard all information about the spatial layout of the features, and therefore cannot capture the structural properties and rough geometrical constraints in the scenes. This problem is addressed in two ways: (i) using oriented band-pass filters at multiple frequencies to incorporate spatial scales (resulting in one HOUP descriptor for each frequency); and (ii) partitioning the image into sub-blocks and computing a HOUP descriptor from local features within each sub-block (aka the subdivision operation).

While these operations have been used by many authors, both for global image representation and local description of interest points, often arbitrary constants have been
suggested for the implementation parameters (specifically the size and spacing of the sub-blocks), which are determined using trial-and-error heuristics.

In this paper, rather than using constant configurations, we suggest using feature selection so that only the HOUP descriptors extracted from the most informative image regions and Gabor frequencies contribute to the image representation. The majority of the available feature selection methods are designed for binary classification, and need auxiliary techniques (such as one-versus-one or one-versus-all) to be extended to multiple classes. In this paper, a method for feature selection is developed, based on kernel alignment, which explicitly deals with multiple classes.

The notion of kernel alignment was first introduced by Cristianini et al. (2002) as a measure of similarity between classes.

In this paper, rather than using constant configurations, we suggest using feature selection so that only the HOUP descriptors extracted from the most informative image regions and Gabor frequencies contribute to the image representation. The majority of the available feature selection methods are designed for binary classification, and need auxiliary techniques (such as one-versus-one or one-versus-all) to be extended to multiple classes. In this paper, a method for feature selection is developed, based on kernel alignment, which explicitly deals with multiple classes.

The notion of kernel alignment was first introduced by Cristianini et al. (2002) as a measure of similarity between two kernel functions or between a kernel and a target function:

\[
A(K_1, K_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \tag{3}
\]

where \(\langle K_1, K_2 \rangle_F\) is the Frobenius dot product.

Given a set of training images and their class labels, the target kernel, \(K_T\), is defined as \(K_T(I, J) = 1\), if \(I\) and \(J\) belong to the same class, and 0 otherwise. For each candidate feature (image sub-block or Gabor frequency) \(n\), its corresponding HOUP descriptors extracted from the training images form a kernel \(K_n\), by using (2) within a parameterized sigmoid function:

\[
K_n(I, J) = \frac{1}{1 + \exp\left(-\sigma_n s_n(x^n_I, x^n_J)\right)} \tag{4}
\]

where \(s_n(x^n_I, x^n_J)\) is the similarity between the \(n\)th descriptors extracted from images \(I\) and \(J\), and \(\sigma_n\) is the kernel parameter, chosen to maximize \(A(K_n, K_T)\), using the unconstrained non-linear optimization method of Lagarias et al. (1998).

Our informative feature selection starts by identifying the feature whose kernel has the highest alignment to the target kernel. It then proceeds by iteratively searching for the next most informative feature whose kernel delivers the maximal amount of additional alignment to the target kernel with respect to each of the previously selected feature(s):

\[
Q_I = \arg \max_{K_i \in P_I, K_j \in R_I} (A(K_i \cdot K_j, K_T) - A(K_i, K_T)) \tag{5}
\]

where \(P_I\) is the set of candidate features, \(R_I\) is the set of selected features up to iteration \(I\), \(Q_I\) is the feature to be selected in iteration \(I\) and \(K_i \cdot K_j\) is the joint kernel produced by combining \(s_i\) and \(s_j\) (see Equation (6)). By taking \(\min\) over all previously selected features, redundancy is avoided (when a candidate feature is similar to one of the selected features, this minimum will be small, preventing the feature from being selected). The max stage then finds the candidate feature with the largest additional contribution.

The feature selection process ends when no increment in alignment with the target is gained by selecting a new feature, or until the number of selected features reaches a pre-defined limit (50 in our experiments).

Once the most informative features are selected, each image is represented by a collection of HOUP descriptors extracted from the selected image sub-blocks and Gabor frequencies. The similarity between each pair of images is then measured by the weighted sum of the individual similarities computed between their corresponding HOUP descriptors:

\[
S(I, J) = \frac{1}{1 + \exp\left(-\sigma \sum_{n=1}^{N} w_n s_n(x^n_I, x^n_J)\right)} \tag{6}
\]

where \(N\) is the total number of selected features, \(\sigma\) is the kernel parameter and \(w_n\) are the combination weights. \(\sigma\) and \(w_n\) are individually chosen to maximize \(A(K_n, K_T)\), using the optimization algorithm of Lagarias et al. (1998).3

5. Experiments

To evaluate our method, we use four publicly available datasets: the UIUC (Lazebnik et al., 2006) and the VPC (Wu et al., 2009) categorization datasets, and the KTH IDOL (Pronobis et al., 2006) and the USC (Siagian and Itti, 2007) recognition datasets. All processing in our experiments was performed in grayscale; color information, when available, was discarded.

For each experiment, the training images are encoded using Gabor filters at six orientations and three frequencies. Given the Gabor coefficients, each image is partitioned to 5 sub-blocks, and a HOUP descriptor is computed for each sub-block at each frequency level, from the oriented Gabor coefficients of pixels that fall into that sub-block. The resulting pool of 165 candidate features (55 sub-blocks multiplied by 3 frequencies) is used for multi-classification using the one-versus-all rule: a classifier is trained to separate each class from the rest and a test image is assigned to the class whose classifier returns the highest response. For topological place recognition (i.e. our experiments on the KTH IDOL and the USC datasets), since the task is not a categorization and no generalization is sought, we use Nearest-Neighborhood (1-NN) to recognize the testing images.

5.1. Scene categorization: UIUC dataset

In our first experiment, we use the UIUC dataset, as this is the most commonly used dataset for scene category recognition, which was gradually built by Oliva and Torralba (2001), Fei-Fei and Perona (2005) and Lazebnik et al. (2006). The dataset contains 15 scene categories, with
210–410 images per category (sample images from each category are shown in Figure 5). Following the standard protocol on this dataset, 100 images are randomly selected from each class for training and the rest for testing. The experiment is repeated five times, each time with different randomly selected sets of training and testing images. The training images are used to select informative features, compute the Principal Component Analysis (PCA) basis and estimate the model parameters. The performance in each run is measured by the average of the per-class recognition rates. The final result on this dataset is reported as the mean and standard deviation of the performances obtained in individual runs.

As shown in Table 1, our method outperforms the state of the art on this dataset by a decent margin. The superior performance of our method can be better appreciated when considering that images in our method are represented with vectors that are more compact than those used by the majority of the compared methods. On average (over the five runs), the feature selection chose 43 informative features (from different candidate sub-blocks and frequency levels) to contribute to image representations, resulting in 43 × 70 = 3010 dimensional representations (that is, almost 10 times more compact than the representations proposed by Bosch et al. (2008) and Liu and Shah (2007)).

By analyzing the distribution of the selected features over different subdivision schemes and frequency levels, we observed that while all sub-blocks in 1 × 1 and 2 × 2 grids (at different frequency levels) are selected as informative features, only 48% and 23% of those resulted from 3 × 3 and 4 × 4 subdivisions were identified as informative features, respectively. Sub-blocks in the 5 × 5 grid were rarely selected to contribute to the image representations. Overall, the majority of the informative sub-blocks result from the 2 × 2 and 3 × 3 subdivision schemes (at all frequency levels). Figure 6 shows the significance of all the sub-blocks based on the average weights (over different frequency levels) computed for them. As can be seen, higher weights are assigned to the sub-blocks in 1 × 1 and 2 × 2 grids (since they capture larger image regions), while among the sub-blocks in the 3 × 3 grid, higher weights are assigned to those at the horizontal middle of the grid. Sub-blocks at the horizontal middle have relatively similar weights. This is consistent with the fact that while scene context can place constraints on elevation (a function of ground level), it fails to provide enough constraints on the horizontal location of the salient and distinctive objects in the scene (Torralba, 2003). Sub-blocks in 4×4 and 5×5 grids have much lower weights, perhaps because these sub-blocks are far too specific compared to 2×2 and 3×3 sub-blocks, with individual HOUP descriptors yielding fewer matches. Figure 6 also compares the average weights assigned to each frequency level (over all sub-blocks). The figure shows that the descriptors extracted at higher frequency levels have lower discriminative power.

To better study the effectiveness of HOUP descriptors, we eliminated the effect of the feature selection and subdivision schemes, and compared the descriptors based on their classification performance when each image is represented by a single block (containing the whole image). Table 2 summarizes the results. As can be seen, 70-dimensional HOUP descriptors can perform just as well as 400-dimensional descriptors based on bag-of-features representations, created from dense SIFT features. In contrast to the HOUP descriptors, which are computed from 3×3 neighborhoods, dense SIFT features are extracted from 16 × 16 image patches, encoding more spatial information.
Table 1. Categorization results (%) on the UIUC dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Descriptor</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Lazebnik et al., 2006)</td>
<td>Bag of features – 200 SIFT words</td>
<td>81.1 ± 0.3</td>
</tr>
<tr>
<td>(Lazebnik et al., 2006)</td>
<td>Bag of features – 400 SIFT words</td>
<td>81.4 ± 0.5</td>
</tr>
<tr>
<td>(Liu and Shah, 2007)</td>
<td>Bag of features – 400 concepts</td>
<td>83.3</td>
</tr>
<tr>
<td>(Bosch et al., 2008)</td>
<td>Bag of features – 1200 topics</td>
<td>83.7</td>
</tr>
<tr>
<td>(Wu and Rehg, 2011)</td>
<td>Census Transform histogram</td>
<td>83.88 ± 0.76</td>
</tr>
<tr>
<td>(Zhou et al., 2009)</td>
<td>Hierarchical Gaussianization</td>
<td>85.2</td>
</tr>
<tr>
<td>Our method</td>
<td>Histogram of Oriented Uniform Patterns</td>
<td>86 ± 0.55</td>
</tr>
</tbody>
</table>

SIFT: Scale Invariant Feature Transform.

Fig. 6. Weights computed for different sub-blocks and frequency levels in our experiment on the UIUC scene categorization database. (a) Shows the box plot of the weights computed at each frequency level for all sub-blocks. (b) Shows the average weights computed for all sub-blocks over all frequencies.

at each image location. To compensate for this and to make a more fair comparison, we compute the HOUP descriptors at three frequencies (as discussed in Section 3) and combined them with equal weights. As Table 2 shows, the combination of HOUP descriptors at multiple frequencies, while still having a lower dimensionality than the compared descriptors, performs significantly higher than other descriptors commonly used for scene recognition.

To investigate the extent to which the matching technique, as discussed in Section 3.1, improves the performance of our scene classification method by dealing with perceptual aliasing, we experiment with two other similarity measures that do not account for perceptual aliasing: Histogram Intersection (HI) and Radial Basis Function (RBF). HI has been widely used for scene recognition, especially with bag-of-features representations (e.g. Lazebnik et al., 2006), and RBF is also a common similarity measure in the field, frequently used as kernel basis with SVM. In this experiment, accuracies of 84.4% and 84.8% were obtained for HI and RBF, respectively. The results indicate that our method for comparing HOUP descriptors outperforms common similarity measures, partially (if not fully) due to its capability to incorporate the distinctiveness of the descriptors when comparing them.

Finally, Table 3 shows the confusion table between the 15 categories, produced by one of the runs. As expected, confusion mainly occurs between the indoor place categories (‘Kitchen’, ‘Living room’ and ‘Bedroom’) and also between some of the outdoor categories with similar structural properties (e.g. ‘Open country’ and ‘Coast’).

5.2. Place categorization: VPC dataset

Our second experiment is conducted on the VPC dataset (Wu et al., 2009). The VPC dataset consists of image sequences from six different homes, each with multiple floors. There are between 6000 to 10,000 image frames for each home, belonging to 11 different place categories. Similar to the paper of Wu et al. (2009), in our evaluations we only use images that belong to the five place categories that are present in all homes: ‘Bathroom’, ‘Bedroom’, ‘Dining room’, ‘Kitchen’ and ‘Living room’ (samples from each place category are shown in Figure 7). Evaluation on this dataset is based on a leave-one-out cross-validation strategy, where the experiment is repeated six times; each time the images of one home are used for testing, and the images of the other five homes are combined to form the training set. The performance in each run is measured by the average of the per-category recognition rates. The final result on this dataset is reported as the average of the performances obtained in the six individual runs.

Table 4 shows the detailed categorization accuracy rates for different homes and place categories. As can be seen, the overall performance of our method on the VPC dataset is 45.94%, which is significantly lower than the categorization results obtained for the UIUC dataset (despite the smaller number of categories in the VPC dataset). This is because, unlike the UIUC dataset, images in the VPC dataset are
collected without the guidance of a human attention mechanism. As a result, a large number of captured images are uninformative views of the places (close up of a wall segment) and the key objects that reveal the place categories are often only partially visible in any specific frame.

As shown in Table 4, the highest accuracy rates are achieved for the ‘Bedroom’ and ‘Bathroom’ categories, while the lowest accuracy rates are obtained for the ‘Dining room’ and ‘Living room’ categories. This is because ‘Bedroom’ and ‘Bathroom’ categories in general, and specifically in the VPC dataset, have lower intra-class variations than categories such as ‘Living room’ and ‘Dining room’.

To put our results in context, we compare the performance of our method with the performances of two other methods based on the bag-of-features framework, developed by Wu et al. (2009). Table 5 summarizes the results. As can be seen, our method based on HOUP descriptors performs significantly better than the methods based on SIFT and CENTRIST descriptors, which clearly demonstrates the suitability of the HOUP descriptors for place categorization.

In this experiment, on average 44 informative features were selected to contribute to image representations. We observed that around 80% of the selected features were also selected in our previous experiment on the UIUC database.
Table 3. Confusion table for the UIUC scene categorization dataset. The diagonal entries show the average accuracy rates for individual categories. The entry in the $i$th row and $j$th column is the percentage of images from category $i$ that were mistakenly classified as category $j$.

<table>
<thead>
<tr>
<th></th>
<th>Suburb</th>
<th>Coast</th>
<th>Forest</th>
<th>Highway</th>
<th>City</th>
<th>Mountain</th>
<th>Country</th>
<th>Street</th>
<th>Building</th>
<th>Office</th>
<th>Bedroom</th>
<th>Industrial</th>
<th>Kitchen</th>
<th>Living room</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suburb</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Coast</td>
<td>0</td>
<td>83</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Forest</td>
<td>0</td>
<td>0</td>
<td>94</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Highway</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>City</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>86</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mountain</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Country</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Street</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>92</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Building</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>91</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Office</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Bedroom</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>65</td>
<td>4</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Industrial</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Kitchen</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>78</td>
</tr>
<tr>
<td>Living room</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Store</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Detailed categorization results (%) on the visual place categorization (VPC) dataset.

<table>
<thead>
<tr>
<th></th>
<th>Bathroom</th>
<th>Bedroom</th>
<th>Dining room</th>
<th>Kitchen</th>
<th>Living room</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home 1</td>
<td>79.42</td>
<td>67.70</td>
<td>10.55</td>
<td>21.29</td>
<td>11.31</td>
<td>38.05</td>
</tr>
<tr>
<td>Home 2</td>
<td>43.84</td>
<td>69.45</td>
<td>36.15</td>
<td>53.61</td>
<td>13.75</td>
<td>43.36</td>
</tr>
<tr>
<td>Home 3</td>
<td>94.72</td>
<td>80.31</td>
<td>4.60</td>
<td>42.17</td>
<td>9.76</td>
<td>41.35</td>
</tr>
<tr>
<td>Home 4</td>
<td>57.31</td>
<td>79.03</td>
<td>17.86</td>
<td>71.50</td>
<td>60.98</td>
<td>57.33</td>
</tr>
<tr>
<td>Home 5</td>
<td>73.53</td>
<td>86.42</td>
<td>8.93</td>
<td>28.10</td>
<td>9.76</td>
<td>41.35</td>
</tr>
<tr>
<td>Home 6</td>
<td>63.73</td>
<td>55.15</td>
<td>14.66</td>
<td>62.63</td>
<td>30.61</td>
<td>45.36</td>
</tr>
<tr>
<td>Average</td>
<td>68.76</td>
<td>73.01</td>
<td>15.46</td>
<td>46.55</td>
<td>25.88</td>
<td>45.94</td>
</tr>
</tbody>
</table>

Table 5. Categorization results (%) on the visual place categorization (VPC) dataset in comparison to the other methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Descriptor</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Wu et al., 2009)</td>
<td>Bag of features – SIFT words</td>
<td>35.00</td>
</tr>
<tr>
<td>(Wu et al., 2009)</td>
<td>Bag of features – CENTRIST words</td>
<td>41.87</td>
</tr>
<tr>
<td>Our method</td>
<td>Histogram of Oriented Uniform Patterns</td>
<td>45.94</td>
</tr>
</tbody>
</table>

SIFT: Scale Invariant Feature Transform

The fact that our method performs very well with almost the same set of features on two different datasets, consisting of images with quite different characteristics, is an indication of the generalizability of our method.

5.3. Topological place recognition: KTH IDOL dataset

The purpose of this experiment (and the next one) is different from the previous two in that topological place recognition is not a categorization task, and therefore no robustness to intra-class variations is sought. However, it is still very challenging, especially in dynamic environments, where the visual appearance of scenes varies significantly with changes in lighting conditions and objects being added or removed from the environment (which can change the visual appearance remarkably from the training time).

Our experiments in this section are conducted on the KTH IDOL dataset (Pronobis et al., 2006), which was built in an office environment with five places. Images were captured using two mobile robots, Minnie and Dumbo, under three different lighting conditions: Cloudy, Night and Sunny. For each lighting condition, four image sequences were captured by each robot (containing images from all five rooms), resulting in 24 image sequences. Since the cameras are mounted at different heights on the robots, the images captured from the same location by the two robots look quite different (sample images from this dataset are shown in Figure 8).

Similar to the paper of Wu and Rehg (2011), we use the first two image sequences in each platform-lighting combination and perform three types of experiments: same robots same lighting conditions; same robots different lighting conditions; and different robots same lighting conditions. The experiments are designed to evaluate the ability of our method to generalize over different lighting conditions, viewpoint changes and other variations caused by regular activities in the rooms. As mentioned earlier, since no generalizability (or robustness to intra-class variations) is required for topological place recognition, we use 1-NN instead of SVM in these experiments.

Table 6 summarizes the results obtained by our method in the three experiments, in comparison to the methods of Wu and Rehg (2011) and Pronobis et al. 2006. In each experiment, all possible combinations of the corresponding image sequences were used for training and testing, and the final result is reported as the mean of the performances obtained in each trial.

As Table 6 shows, our method outperforms other compared methods in all experiments. The superior performance of our method is an indication of its robustness to illumination changes (specifically evaluated by the second experiment where the training and testing images are from different lighting conditions), viewpoint variations (specifically evaluated by the third experiment where the training and testing images come from different robots and camera setups) and other common variations caused by moving objects and people (evaluated by all three experiments).

Analyzing the behavior of feature selection for our experiments on the IDOL dataset, we observed that on average, 9, 13 and 23 features were selected for the first, second and third experiments, respectively. The variation in the number of selected features is directly related to the difficulty of the experiment: the lowest number of features being selected for the first experiment (same robots same lighting conditions), which is the easiest one, and the highest number being selected for the third experiment (different robots same lighting conditions), which is the most difficult one. This indicates that our method can automatically calibrate the configuration of the representation to suit the given conditions.
Fig. 8. Example images from the KTH IDOL topological place recognition database, with five places. For each place, the first three sample images (from left to right) were captured by the Dumbo robot at three different lighting conditions (cloudy, night and sunny, respectively) and the forth sample image was captured by the Minnie robot at the night lighting condition. All sample images for each place were captured from relatively the same pose.

classification/recognition task. This is a very interesting characteristic that the majority of the global methods, which use constant subdivision schemes for representation (e.g. Torralba et al., 2003; Renniger and Malik, 2004; Lazebnik et al., 2006; Siagian and Itti, 2007; Wu and Rehg, 2011), do not possess.

5.4. Topological place recognition: USC dataset

In our final experiment, we evaluate our method on another topological place recognition dataset, the USC dataset (Siagian and Itti, 2007), which is more than five times larger than the KTH IDOL dataset, in terms of the number of the topological places. The goal of this experiment is to examine the scalability of our method. The USC dataset, built by Siagian and Itti (2007), contains images from three outdoor sites on the University of Southern California campus, including the ACB site (a rigid and less spacious man-made environment), the AnF site (composed of two adjoining parks) and the FDF site (a largely open area). Each site is manually divided into nine continuous segments/places. Figure 9 shows a sample image from each of the 27 topological places. For each place, 12–15 image sequences are provided, capturing different lighting conditions, small
viewpoint variations and some structural changes (e.g., benches temporarily removed from the parks, service vehicles or cars temporarily parked, storage boxes temporarily placed in different places, etc.). The standard protocol for experimenting on this dataset is to use 9–11 image sequences from each place to train the models and the remaining 4 image sequences (taken on separate days and various lighting conditions) for testing. However, in our experiments, only one image sequence (randomly selected) from each place is used to train our method, while the testing sequences are identical to the standard protocol (and different from the training sequence) to facilitate comparison.

We sought to investigate the extent to which our method is successful in dealing with variations in the test datasets. Table 7 shows the performance of our method in comparison to different state-of-the-art context-based approaches, namely the methods of Oliva and Torralba (2001), Torralba et al. (2003), Renniger and Malik (2004) and Siagian and Itti (2007). The performance is measured by the percentage...
of all test images recognized correctly. As can be seen in Table 7, our method outperforms other context-based methods in recognizing scenes from 27 different places, which is significant given that our method was trained with images from a single appearance condition, while other methods used several image sequences for each place (captured in different conditions) to assure a wide range of testing conditions (see Figure 10 for example of appearance variations caused by changes in the lighting condition). This validates the advantages of our solution in dealing with dynamic changes in the environment, even with the presence of a relatively large number of place classes. Note that the underlying features used by the compared methods to encode spatial information are very similar to those used in the HOUP descriptors. However, they use simple statistics to aggregate the encoded information at different image locations and produce the image representation, while in HOUP descriptors detailed analysis is performed for this purpose.

Furthermore, given the frequency of the repetitive objects (e.g. bushes) in outdoor sites, the superior performance of our method and its robustness to dynamic changes can be partially attributed to our similarity measure, which accounts for perceptual aliasing.

Examining the performance of our method, we observed that the majority of the recognition errors occurred during the transitions between different places/segments, specifically those adjacent places with the possibility of overlapping scenes (e.g. the first and second segments of the ACB site). The remaining errors are mainly due to significant variations in the condition of the environment, especially changes in lighting conditions, which considerably influence the output amplitude of the oriented filters.

### 6. Conclusions

In this paper we presented the HOUP, as a new visual descriptor. We showed that HOUP descriptors, when extracted from informative image sub-blocks and frequency levels, can build image representations that provide strong discriminative power for place recognition, while offering a significant level of generalizability for place categorization. An extensive set of experiments revealed the superior performance of our method, in comparison to the state of the art, on two publicly available classification datasets and two topological place recognition datasets.

The proposed method produces image representations that are global and non-invariant. Therefore, it does not explicitly tackle image geometrical variations (e.g. scale, rotation, etc.). However, as previous research (Oliva and Torralba 2001; Lazebnik et al., 2006) has shown and our own experiments confirm, global representations can be surprisingly effective for identifying the overall scene and discovering its semantics (e.g. the contained objects or activities), even when they are embedded in heavy clutter and are from various poses and appearances.

There are a number of potential directions for future work. The temporal continuity of the image frames in visual place recognition and categorization imposes the constraint that the computed labels should vary smoothly along the robot’s trajectory almost everywhere, while preserving discontinuities at the borders between adjacent visited places in the environment. A number of solutions (e.g. Rottman et al., 2005; Wu et al., 2009; Fazl-Ersi and Tsotsos, 2010; Ranganathan, 2010) have been proposed to use this as a major source of additional information.
for improving the initial categorization results in light of contextual cues. For example, Wu et al. (2009) and Ranganathan (2010) reported improved accuracy rates of 45.62% and 44.88% on the VPC dataset, respectively, with methods that enforce spatial smoothness and take into account the contextual cues. Given these results, we can safely expect that the performance of our method too will be improved by taking advantage of contextual information. While this was outside of the scope of this paper, it certainly suggests a potential direction for future work.

Our method in its current form always assigns a query image to one of the learned classes, even if the query image is heavily ambiguous (e.g. wall close-up examples) or it does not belong to any of the learned classes. One way to address this shortcoming, which will be discussed in more detail in future work, is to output ‘Unknown’ for any query image with a low similarity score to all learned classes, as such images are often from classes other than those learned, or are global examples of perceptual aliasing (e.g. wall close-up images).

Another direction for future work is to evaluate the performance of HOUP descriptors when used within the bag-of-features framework. Given that substantial improvements have been reported in the literature for descriptors such as SIFT and HOG (Histogram of Oriented Gradients) when used within the bag-of-features, we expect to achieve some improvement by exploring this direction.

Finally, we believe that our proposed method can be used beyond place recognition and place categorization. We are especially interested in using the method within a larger scene interpretation system, to take advantage of the useful discriminative information provided by the global scene cues as indirect evidence to capture a more semantically enriched description of the scenes (including the identity of the constituent objects, the type of the activities contained in the scene, etc.), enabling an autonomous robot to automatically build a true semantic representation of the environment.

Notes
1. In an initial experiment, Gabor filters tuned to six orientations (from 0 to 5π/6 with increment of π/6) provided the best performance.
2. The descriptors being compared are not eliminated from A. This is because when A is large, the contribution of the two descriptors being compared to the values of μ_A and S_A is negligible. Therefore, μ_A and S_A need to be computed only once.
3. To estimate σ_A in Equation (4) and w_A in Equation (6), the optimization algorithm of Lagarias et al. (1998) is used, which finds the max/min of a scalar function, starting at an initial estimate. In our implementation, the scalar function returns the alignment between a given kernel and the target kernel for an input parameter σ_A. The initial estimate for σ_A is empirically set to 2.0 for all our experiments. The σ_A that maximizes the alignment is selected as the optimal kernel parameter, and the alignment value corresponding to the optimal σ_A is used as the weight of the kernel, w_A · σ in Equation (6) is similarly chosen.
4. In our experiments, our variation of the OSS kernel is used as a pre-computed kernel with the LIBSVM tool (Chang and Lin, 2001). All SVM parameters are set to the default values suggested by the authors of the LIBSVM, with the exception of c (cost) and w (weight), which are set to 0.6 and 1.3, respectively.
5. Note that while in this experiment the OSS kernel is replaced by RBF and HI kernels, the image representation process (i.e. selecting informative image sub-blocks and wavelet frequencies and computing and combining their corresponding HOUP descriptors) is preserved unchanged.
6. There are in total 75,073 test images, resulted by combining 6. There are in total 75,073 test images, resulted by combining the images from the four testing sequences of each place.

Funding
The authors are grateful for support from the Natural Science and Engineering Council of Canada. The second author also acknowledges support of the Canada Research Chairs Program.

References


