Where am I: Place instance and category recognition using spatial PACT

Jianixn Wu James M. Rehg
School of Interactive Computing, College of Computing, Georgia Institute of Technology
{wujx,rehg}@cc.gatech.edu

Abstract

We introduce spatial PACT *(Principal component Analysis of Census Transform histograms), a new representation for recognizing instances and categories of places or scenes. Both place instance recognition ("I am in Room 113") and category recognition ("I am in an office") have been widely researched. Features that have different discriminative power/invariance tradeoff have been used separately for the two tasks.* PACT *captures local structures of an image through the Census Transform (CT), while large-scale structures are captured by the strong correlation between neighboring CT values and the histogram. The PCA operation ignores noise in the histogram distribution, computes important "primitive shapes", and results in a compact representation. Spatial* PACT*, a spatial pyramid of* PACT*, further incorporates global structures in the image. Our experiments demonstrate that spatial* PACT *outperforms the current state-of-the-art in several place and scene recognition, and shape matching datasets. Besides, spatial* PACT *is easy to implement. It has nearly no parameter to tune, and evaluates extremely fast.*

1. Introduction

Knowing "Where am I" has always being an important research topic in the robotics and computer vision communities. Place recognition (or robot localization/mapping) has been widely studied in robotics [8, 17, 19, 23, 27], which usually requires to find the exact location in a global frame of reference [8, 19], or at least to find a rough location (*e.g.* "I am in Room 113") [17, 22, 23, 27]. Vision researchers, however, work on the other end of the spectrum. Instead of recognizing a precise location or exact instance of a room, usually a category of the place is recognized [2, 4, 9, 16, 18]. In other words, vision methods will output "This is an office" instead of "This is office 113." For this reason, place recognition is usually termed as scene recognition in vision.

The difference in recognition goals also results in different choices of input sensors and data collection procedures. Images are usually used in scene recognition tasks. The images were purposely captured to be in the canonical view (characteristic of the scene category). In robot localization, range sensors are popular. Recently cameras are also frequently used [17, 25, 28]. Since images are acquired by robots, no effort is taken to make sure that they are representative of or distinctive for the place.

Consequently, the image representations are quite different.[1] Robot localization usually employs features with high discriminative power (*e.g.* SIFT features in [19]), while features with higher invariance (*e.g.* "visual codebook" in [9]) are used for scene recognition.

Both research problems have wide applications in real world. It is very convenient if a home service robot could locate itself in specific rooms in a house. Knowing the semantic category of a location will also help recognizing objects in the scene [22], which may further help identify more detailed information of the location. Place instance and category recognition are different but related aspects of a general place recognition problem. In this paper we propose PACT, Principal component Analysis of Census Transform histograms, a representation that unify the needs for recognizing both instances and categories of places. The Census Transform (CT) summarizes local shape information, while the strong constraints among neighboring CT values and the PCA operation compactly encode the global shape in an image patch. We also propose spatial PACT, which encodes rough global spatial arrangement of sub-blocks in an image, and finds the tradeoff between discriminative power and invariance [24] for place recognition tasks. We show that spatial PACT has several important advantages in comparison to state-of-the-art feature representations for scene recognition and categorization:

- Superior recognition performance on multiple standard datasets;
- Significantly fewer parameters to tune;
- Extremely fast evaluation speed (> 50 fps);
- Very easy to implement.

[1] In this paper we will focus on the camera sensor data.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

The rest of the paper is organized as follows. Related methods are discussed in Sec. 2. PACT and spatial PACT are presented in Sec. 3, with experiments shown in Sec. 4. Sec. 5 concludes this paper with discussions of drawbacks of the proposed method and future research.

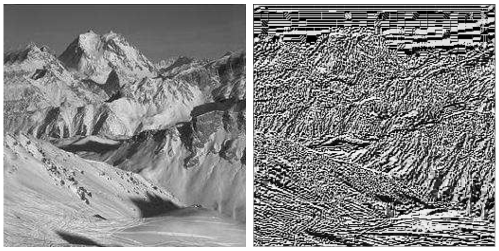
2. Related Work

There is a large literature body in both place instance and category recognition. In this section we will focus on the representation issues in place recognition.

Histograms of various image properties (*e.g.* color [17, 21, 23], or image derivatives [17]) have been widely used in place recognition. However, after SIFT [13] is popularized in the vision community, it nearly dominates the feature choice in place recognition systems [2, 4, 7, 9, 10, 12, 18, 19, 27]. SIFT features are invariant to scale and robust to rotation changes. The 128 dimensional SIFT descriptors have high discriminative power, while at the same time are robust to local variations [15]. It is shown that SIFT significantly outperforms edge points [9], pixel intensities [2, 4], and steerable pyramids [7] in recognizing places and scenes.

It is suggested that recognition of scenes could be accomplished by using "global configurations", without detailed object information [16]. Thus statistical analysis of the distribution of SIFT features are popular in scene recognition. SIFT descriptors are first vector quantized to form the "visual codebook" or "visterms". Different views are held on the quantized SIFT features. Some researchers believe that the codebook represent meaningful semantic aspects of the natural scenes. Liu and Shah [12] used Maximization of Mutual Information co-clustering to cluster SIFT features to form intermediate semantic concepts. Probabilistic Latent Semantic Analysis (pLSA) was also used to detect latent semantic topics [2, 18]. Quelhas *et al.* showed that in a 3 class classification task pLSA generated compact representation and improved recognition. However, Lazebnik, Schmid and Ponce showed that pLSA lowered recognition rates by about 9% in a 15 class scene recognition problem [9]. The k-means algorithm was used to cluster SIFT features, and the cluster centers were used as the codebook in [9]. In place recognition, SIFT features were usually densely sampled, instead of only sampled at interest points [2].

SIFT models represent images as "bag of features", *i.e.* spatial arrangement information among features are completely ignored. However, it was long recognized that spatial arrangements were essential for recognizing scenes. For example, Szummer and Picard divided images into 4×4 blocks, matched blocks separately, and combined the matching results [21]. This strategy significantly improved recognition accuracy. In [9], Spatial Pyramid Matching (SPM) was proposed as a kernel method that systematically integrated the spatial information. Images were repeatedly divided into increasingly finer sub-blocks, and histograms of local features of sub-blocks were integrated using the spatial kernel matching scheme, which took into account rough spatial correspondences.



(a) mountain image (b) Transformed image

Figure 1. An example "Census Transformed image".

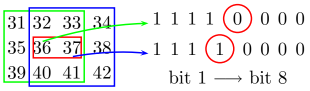


Figure 2. Illustration of constraints between CT values of neighboring pixels. This picture is best viewed in color.

3. Spatial PACT

3.1. PACT: Principal component Analysis of Census Transform histograms

Census Transform (CT) is a non-parametric local transform designed for establishing correspondence between local patches [26]. Census transform compares the intensity values of a pixel with its eight neighboring pixels, as illustrated in Eqn. 1.

$$\begin{array}{c|c|c} 32 & 64 & 96 \\ \hline 32 & \mathbf{64} & 96 \\ \hline 32 & 32 & 96 \end{array} \Rightarrow \begin{array}{ccc} 1 & 1 & 0 \\ 1 & & 0 \\ 1 & 1 & 0 \end{array} \Rightarrow (11010110)_2 \Rightarrow \text{CT} = 214 \quad (1)$$

The eight bits generated from intensity comparisons can be put together in any order (we collect bits from top to bottom, and from left to right), which is consequently converted to a base-10 number in $[0\ 255]$.[2] Just as other non-parametric local transforms which are based on intensity comparisons (*e.g.* ordinal measures [1]), Census Transform is robust to illumination changes, gamma variations, *etc.*

As a visualization method, we create a "Census Transformed image" by replacing a pixel with its CT value. Shown by the example in Fig. 1, the Census Transform retains global structures of the picture (especially discontinuities) besides capturing the local structures as it is designed for. A histogram of the CT values in an image (or image patch) thus encodes both local and global information of the image.

Another important property of the transform is that CT values of neighboring pixels are highly correlated. In the

[2] $x = y$ is treated as if $x > y$. Thus in Eqn. 1 the second bit is set to 1.

example of Fig. 2, we examine the constraint posed by the two center pixels. The Census Transform for pixels valued 36 and 37 are depicted in right, and the two circled bits are both comparing the two center pixels (but in different orders). Thus the two circled bits are constrained to be strictly complement to each other. More generally, bit 5 of CT(x, y) and bit 4 of CT$(x + 1, y)$ must always be complement to each other, since they both compare the pixels at (x, y) and $(x + 1, y)$. There exist many other such constraints. In fact, there are eight such constraints between one pixel and its eight neighboring pixels. Besides these deterministic constraints, there also exist indirect constraints that are more complex. For example, in Fig. 2, the pixel valued 32 compares with both center pixels in computing their CT values (bit 2 of CT(x, y) and bit 1 of CT$(x+1, y)$). Depending on the comparison results between the center pixels, there are probabilistic relationships between these bits.

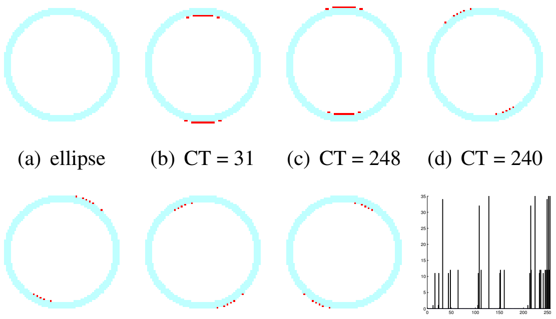
The transitive property of such constraints also make them propagate to pixels that are far apart. For example, in Fig. 2, the pixels valued 31 and 42 can be compared using various paths of comparisons, *e.g.* $31 < 35 < 39 < 40 < 41 < 42$. Similarly, although no deterministic comparisons can be deduced between some pixels (*e.g.* 34 and 39), probabilistic relationships still can be obtained. The propagated constraints make Census Transform histograms implicitly contain information for describing global structures, unlike the histogram of pixel values.

Finally, in the top part of Fig. 1(b), various CT values seemingly quite different are displayed. But in the base-2 format, these CT values all represent homogeneous regions with small variations (*e.g.* $(00001000)_2$). That is, there also exist strong correlations between pairs of CT values. We use Principal component Analysis of Census Transform histograms (PACT) to remove these correlation effects, and to get a more compact representation. We will use PACT as our representation for place recognition, and usually 40 eigenvectors are used in the PCA operation.

In computing histograms and PCA, we remove two bins with CT $= 0, 255$ and normalize the CT histograms and eigenvectors such that they have zero mean and unit norm. Also, we do not subtract mean in PCA for computational efficiency. Our experiments show that this does not cause significant difference in recognition results.[3]

3.2. PACT encodes shape

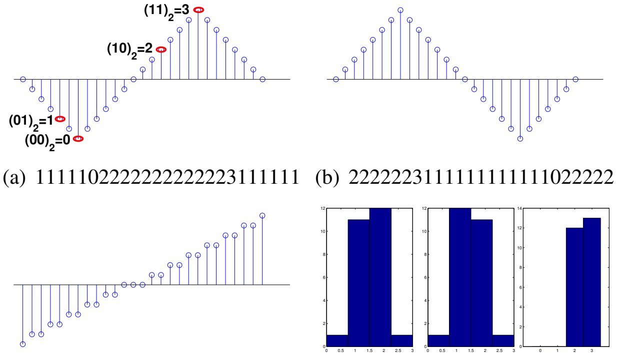
In order to understand why PACT efficiently captures the essence of scene information, it is worthwhile to further examine the distribution of CT values and the PACT. Using 1500 images from the 15 class scene dataset [9], we find that the 6 CT values with highest frequencies are CT $= 31, 248, 240, 232, 15, 23$ (excluding 0 and 255). As shown in Fig. 3(b)-3(g), these CT values captures local 3×3 neighborhoods that have either horizontal or various close-to-diagonal edge structures. It is sort of counter-intuitive that vertical edge structures are not among the top candidates. A possible explanation is that vertical structures are usually appearing to be inclined in pictures because of the perspective nature of cameras.



(a) ellipse (b) CT = 31 (c) CT = 248 (d) CT = 240

(e) CT = 232 (f) CT = 15 (g) CT = 23 (h) histogram

Figure 3. Illustration of Census transforms. 3(a) is an example image of ellipse. 3(b)-3(g) show pixels having the 6 highest frequency CT values (shown in red). 3(h) is the CT histogram of 3(a). This image is best viewed in color.



(a) 1111102222222222222223111111 (b) 2222223111111111111022222

(c) 323232323233232323232323232 (d) Census Transform histograms

Figure 4. Census Transform encodes shape in 1-d. The Census Transform values of (a)-(c) are shown in the caption, and their CT histograms in (d). Both end points are ignored in compute CT. This image is best viewed in color.

Histogram of the example ellipse image (Fig. 3(a)) is shown in Fig. 3(h). It summarizes the distribution of various local structures in the image. Because of the strong correlation of neighboring CT values, the histogram cells are not independent of each other. On the contrary, a histogram implicitly encodes strong constraints of the global structure of the image. For example, if an image has a CT distribution close to that of Fig. 3(h), we would well expect the image to exhibit ellipse shape with a high probability.

A simplification to the one dimensional case better explains the intuition behind our statement. In 1-d there are only 4 possible CT values, and the semantic interpretation of these CT values are obvious. As shown in Fig. 4(a), the four CT values are $CT = 0$ (valley), $CT = 1$ (downhill),

[3] Code is available at `http://www.cc.gatech.edu/~wujx/PACT/PACT.htm`. Please refer to the code for details.

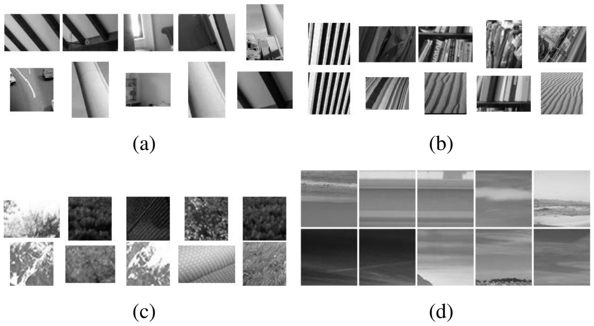


Figure 5. Image patches that have high correlation coefficients with the primitive shapes. Image patches that have high correlation with the same eigenvector are organized in the same subfigure, and they share common shape characteristic.

$CT = 2$ (uphill), and $CT = 3$ (peak). For simple shapes in 1-d, the CT histograms encode lots of shape information and constraints. Downhill shapes and uphill shapes can only be connected by a valley, and uphill shapes require a peak to transit to downhill shapes. Because of these constraints, the only other shapes that has the same CT histogram as that of Fig. 4(a) is those shapes that cut a small portion of the left part of Fig. 4(a) and move it to the right. Images that are different but keep the shapes (*e.g.* Fig. 4(b)) also are similar in their CT histograms (Fig. 4(d)). On the contrary, a large number ($> 1M$) of possible curves have the same intensity histogram as that of Fig. 4(a). Even if we impose smoothness constraints between neighboring pixel intensities, the shape ambiguity is still large (*e.g.* Fig. 4(c) is smooth and has the same intensity histogram as that of Figs. 4(a) and 4(b), but it has different shape and a very different CT histogram).

PCA on the CT histograms (*i.e.* PACT) extracts the most important components among the distribution of CT histograms (in terms of histogram reconstruction). In this sense, eigenvectors with high eigenvalues are important "primitive shapes" which are independent (or, not similar) to each other. In other words , the eigenvectors are the "shape codebook" in the CT histogram space. Since both the histogram and eigenvectors are normalized to have zero mean and unit norm, elements in PACT are the correlation coefficients (*i.e.* similarities) between the input histogram and the primitive shapes. Example image patches whose CT histograms have high correlation coefficients with these primitive shapes are shown in Fig. 5. For example, image patches in Fig. 5(b) all have inclined edges, and Fig. 5(c) are patches of high degree of roughness (consisting of many miniature building blocks, *c.f.* [16]).

3.3. Spatial PACT

A "spatial pyramid" (dividing an image into subregions and integrating correspondence results in these regions) encodes rough global structure of an image and usually improves recognition [9]. It is straightforward to build a spatial pyramid for the proposed PACT representation (spatial PACT, or, sPACT). As shown in Fig. 6, the level 2 split in a spatial pyramid divides the image into $2^2 \times 2^2 = 16$ blocks. We also shift the division (dash line blocks) in order to avoid artifacts created by the non-overlapping division, which makes a total of 25 blocks in level 2. Similarly, level 1 and 0 have 5 and 1 blocks respectively. The image is resized between different levels so that all blocks contain the same number of pixels. PACT in all blocks are then concatenated to form an overall feature vector. For example, if 40 eigenvectors are used in PACT, a level 2 pyramid will result in a feature vector which has $40 \times (25 + 5 + 1) = 1240$ dimensions.

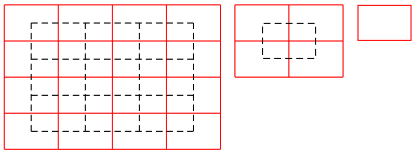


Figure 6. Illustration of the level 2, 1, and 0 split of an image.

After the sPACT feature vectors are extracted from images, we choose different classifiers for recognizing place instances and categories, in order to find the right tradeoff between discriminative power and invariance for both problems. In order to recognize place instances, we use the Nearest Neighbor classifier (1-NN, to be precise). Thus we are looking for places that have not only similar local patches, but also *exact* spatial arrangements of these patches. SVM classifiers are used for category recognition. In category recognition we are only expecting loose spatial information, such as "sky should be above the ground". We expect the generalization ability of SVM to capture such relationships, while avoiding overfitting.

Since the CT values are based on only pixel intensity comparisons, it might be helpful to include a few images statistics, *e.g.* average value and standard deviation of pixels in a block. We append these statistics to spatial PACT in the input to SVM classifiers for scene recognition problem. However, they are not used in the 1-NN classifiers, since the large variation of illumination in place instance recognition tasks will cause these global statistics to be unreliable.

4. Experiments

The spatial PACT representation are tested on 4 datasets: Swedish leaf [20], KTH IDOL [17], 15 class scene category [9], and the 8 class event dataset [10]. In each dataset, the available data are randomly split into a training set and a testing set. The random splitting is repeated 5 times, and the average accuracy is reported. Although color images are available in 3 datasets (leaf, IDOL, and events), we only use the intensity values and ignore color information. No parameter need to be set in the 1-NN classifier. We use

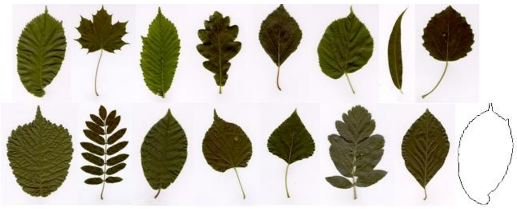


Figure 7. Example images from the Swedish Leaf dataset. The first 15 images are chosen from the 15 leaf species, one per species. The last image is the contour of the first leaf image.

Method	Input	Rates
Shape-Tree [5]	Contour only	**96.28%**
IDSC+DP [11]	Contour only	94.13%
sPACT	Contour only	90.77%
SC+DP [11]	Contour only	88.12%
Söderkvist [20]	Contour only	82.40%
sPACT	Gray-scale image	**97.92%**
SPTC+DP [11]	Gray-scale image	95.33%

Table 1. Results on the Swedish leaf dataset.

LIBSVM [3] for the SVM classifiers. RBF kernels are used in our experiments, and the parameters of SVM are chosen by cross validation.[4]

4.1. Swedish Leaf

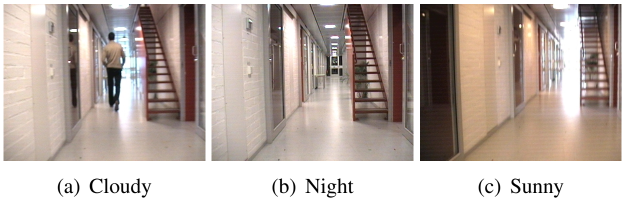
The Swedish leaf dataset [20] collects pictures of 15 species of Swedish leaves (*c.f*. Fig. 7). There are 75 images in each class. Following the protocol of [20], 25 images from each class are used for training and the rest 50 for testing.

This dataset has been used to evaluate shape matching methods [5, 11], in which the contour of leaves (instead of the gray-scale or color leaf picture) are used as input (*e.g.* the last picture in Fig. 7). In the contour image, no other information is available (*e.g.* color, texture) besides shape. We use the contour input to verify our statement that sPACT encodes shape information.

The first 25 images from each class are used to train the PCA eigenvectors. 10 and 40 eigenvectors are used when the inputs are contour and intensity images, respectively. Results on this dataset are shown in Table 1. Although not specifically designed for matching shapes, sPACT can achieve 90.77% accuracy on leaf contours, better than Shape Context+Dynamic Programming (SC+DP). When pictures instead of contours are used as input, sPACT can recognize 97.92% leaves, which outperforms other methods by a large amount.

4.2. KTH IDOL

The KTH IDOL (Image Database for rObot Localization) dataset [14] was captured in a five-room office environment, including a one-person office, a two-person office, a kitchen, a corridor, and a printer area. Images were taken by two Robots: Minnie and Dumbo. The purpose of this dataset is to recognize which room the robot is in based on a single image, *i.e.* a place instance recognition problem.



(a) Cloudy (b) Night (c) Sunny

Figure 8. Example images from the KTH IDOL dataset. Images showed the same location under different conditions. Images were taken by the Minnie robot.

Cameras were mounted at different heights on the robots, which made the pictures taken by the two robots quite different. Image resolution was 320×240. A complete image sequence contained all the images captured by a robot when it was driven through all five rooms. Images were taken under 3 weather conditions: Cloudy, Night, and Sunny. For each robot and each weather condition, 4 runs of robot driving were taken on different days. Thus, there are in total $2 \times 3 \times 4 = 24$ image sequences. Various changes during different robot runs (*e.g.* moving persons, changing weather and illumination conditions, relocated/added/removed furniture make this dataset both realistic and challenging. Fig. 8 shows images taken by the Minnie robot under 3 different weather conditions at approximately the same location, but with substantial visual changes.

In our experiments we use the run 1 and 2 in each robot and weather condition. We perform 3 types of experiments as those in [17]. First we train and test using the same robot, same weather condition. Run 1 is used for training and run 2 for testing, and vice versa. Second we use the same robot for training and testing, but with different weather conditions. These experiments test the ability of sPACT to generalize over variations caused by person, furniture, and illumination. The last type of experiment uses training and testing set under the same weather conditions, but captured by different robots. Note that images taken by the two robots are quite different. The 1-Nearest neighbor classifier is used for this place instance recognition task. Results using level 2 pyramid sPACT and 1-NN are shown in Table 2, compared against results in [17].

In the first type of experiments, both sPACT and the method in [17] attain high accuracy ($> 95\%$), and the two methods are performing roughly equally well. However, in the second type of experiments sPACT has significantly higher accuracies (18% higher in Minnie and 14% higher in Dumbo). The superior performance of sPACT shows that it is robust to illumination changes and other minor variations (*e.g.* moving persons, moved objects in an image, *etc*). The

[4] In detail, we use the `easy.py` tool from LIBSVM and the training set in the first random split to choose parameters.

Train	Test	Condition	sPACT	[17]
Minnie	Minnie	Same	95.35%	**95.51%**
Dumbo	Dumbo	Same	**97.62%**	97.26%
Minnie	Minnie	Different	**90.17%**	71.90%
Dumbo	Dumbo	Different	**94.98%**	80.55%
Minnie	Dumbo	Same	**77.78%**	66.63%
Dumbo	Minnie	Same	**72.44%**	62.20%

Table 2. Average accuracies on recognizing place instances using the KTH-IDOL dataset. Level 2 pyramids are used for sPACT.

	L=0	L=1	L=2	L=3
Minnie	60.51%	85.75%	**90.17%**	90.30%
Dumbo	74.67%	91.75%	**94.98%**	94.67%

Table 3. Average accuracies on the KTH-IDOL dataset using different levels of spatial pyramid. The training and testing set are acquired using the same robot, but different weather conditions. $L = 0$ means not using a spatial pyramid at all.

Dumbo robot achieves a 94.57% accuracy using a single input image without knowing any image histories (a "kidnapped robot" [25]). Thus, after walking a robot in an environment, sPACT enables the robot to robustly answer the question "Whare am I?" based on a single image, a capacity that is very attractive to indoor robot applications. When the training and testing data come from different robots, the performance of both methods drop significantly. This is expected, since the camera heights are quite different. However, sPACT still outperforms the SVM classifier in [17] by about 10%.

We also tested the effects of using different pyramid levels. As shown in Table 3, applying a spatial pyramid matching scheme greatly improves system performances ($L > 0$ vs. $L = 0$). However, the improvement after $L > 2$ is negligible. $L = 3$ performance is even worse than that of $L = 2$ in Dumbo. Our observation corroborates that of Lazebnik, Schmid and Ponce in [9], which used a scene recognition dataset. In the remainder of this paper, we will use $L = 2$ in sPACT.

sPACT can be computed and evaluated quickly. The IDOL dataset has around 1000 images in each image sequence, and sPACT processes at about 50 frames per second on an Intel Pentium 4 2GHz computer for computing the features, and finding the 1-NN match.[5]

4.3. The 15 class scene category dataset

The 15 class scene recognition dataset was built gradually by Oliva and Torralba ([16], 8 classes), Fei-Fei and Perona ([4], 13 classes), and Lazebnik, Schmid and Ponce ([9], 15 classes). This is a scene category dataset (scene classes including office, store, coast *etc*. Please refer to Fig. 9 for category names.) Images are about 300×250 in resolution, with 210 to 410 images in each category. This dataset contains a wide range of scene categories in both indoor and outdoor environments. Unlike the KTH IDOL images which are taken by robots, images in this datasets are taken by people and representative of the scene category. We use SVM and sPACT in this dataset. The first 100 images in each category were used to perform PCA. Same as previous research on this dataset, 100 images in each category are used for training, and the remaining images constitute the testing set. The results are shown in Table 4, where our level 2 pyramid sPACT achieves the highest accuracy.

L	Method	Feature type	Rates
0	SPM [9]	16 channel weak features	45.3 ± 0.5
0	SPM [9]	SIFT, 200 cluster centers	72.2 ± 0.6
0	SPM [9]	SIFT, 400 cluster centers	$\mathbf{74.8 \pm 0.3}$
0	sPACT	CT histogram	73.8 ± 0.8
3	SPM [9]	16 channel weak features	66.8 ± 0.6
2	SPM [9]	SIFT, 200 cluster centers	81.1 ± 0.3
2	SPM [9]	SIFT, 400 cluster centers	81.4 ± 0.5
3	SPM [12]	SIFT, 400 inter. concepts	**83.3**
2	sPACT	PACT, 40 eigenvectors	$\mathbf{83.3 \pm 0.5}$

Table 4. Recognition rates on the 15 class scene dataset.

In [9], low level features were divided into weak features (computed from local 3×3 neighborhoods) and strong features (SIFT features computed from 16×16 image patches). Strong features were shown to have much higher accuracy than weak features (*c.f.* Table 4). The Census Transform is computed from 3×3 local neighborhoods, and falls into the weak feature category. However, when $L = 0$ (not using spatial pyramid), sPACT substantially outperforms the weak features and the strong features with 200 codebook size in [9], and is only inferior to the strong features with 400 codebook size. When a spatial pyramid is used, sPACT has the highest recognition rate (in tie with the strong SIFT features with 400 "intermediate concepts" in [12]). We believe that this is because the strong constraints between neighboring CT values of make PACT able to capture shape information beyond the 3×3 patches.

Confusion matrix from one run on this dataset ($L = 2$ sPACT) is shown in Fig. 9, where row and column names are true and predicted labels respectively. The biggest confusion using sPACT happens between category pairs such as bedroom/living room, industrial/store, and coast/open country, which coincides well with the confusion distribution in [9].

Orientation Histogram Orientation histogram [6] is another representation that uses histogram of quantities computed from 3×3 neighborhoods. We implemented this method with 40 bins. Combined with a level 2 spatial pyramid, Orientation Histogram achieves 76.6% recognition rate, which is signifantly worse than sPACT (83.3%).

Indoor-outdoor classification We also distinguish indoor and outdoor scenes in this dataset. The "industrial" category contains both indoor and outdoor images, and is ignored. The remaining 14 categories are separated as 5 in-

[5] Or 20 fps if include the time for loading the test image from hard drive.

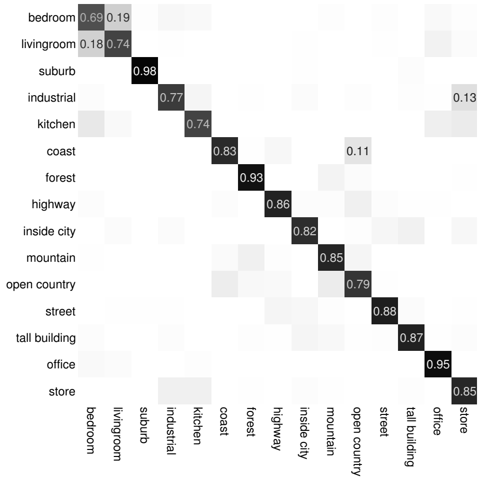


Figure 9. Confusion matrix of the 15 class scene dataset. Only rates higher than 0.1 are shown in the figure.

door categories and 9 outdoor categories. Using $L = 2$ and $L = 0$, sPACT successfully predicts labels for 98.54% and 96.28% of the images, respectively. These recognition rates are much higher than previous results on indoor-outdoor classification datasets (*e.g.* [21, 16]).

Linear classifiers Linear SVM classifiers are also applied to the scene dataset. They achieve accuracy of 82.05% and 74.18%, using sPACT with $L = 2$ and $L = 0$, respectively. The implication of these results are two fold. First, the difference in performance of RBF kernels and linear kernels are quite small.[6] This observation suggests that images from the same category are compact in the sPACT representation space. Second, because of the fast testing speed of linear classifiers and small performance difference, linear SVM classifiers could be used to ensure real-time classification. A further observation is that linear SVM classifiers get 95.32% accuracy on indoor-outdoor classification, without using a spatial pyramid. In other words, the CT histgoram could reliably distinguish the man-made indoor structures and the outdoor natural scenes.[7]

Speed and classifier analysis The time to extract PACT is proportional to the input image size. However, large images can be down-sampled to ensure high speed. Our experiments observed only slight (usually $< 1\%$) performance drop. Also, sPACT is not sensitive to SVM parameters. $(C, \gamma) = (8, 2^{-7})$ is recommended for RBF kernels with probability output, and $C = 2^{-5}$ for linear SVM. Finally, we want to point out that choosing the right classifier for a specific application is very important. If we use SVM for the IDOL dataset or use 1-NN for the scene dataset, recognition rates are about 10% lower than the rates reported above.

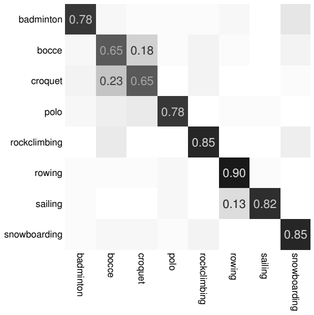


Figure 10. Confusion matrix of the event dataset. Only rates higher than 0.1 are shown in the figure.

4.4. The 8 class event dataset

The event dataset contains images of eight sports: badminton, bocce, croquet, polo, rock climbing, rowing, sailing, and snowboarding [10]. In [10], Li and Fei-Fei used this dataset in their attempt to classify these events by integrating scene and object categorizations (*i.e.* deduce "what" from "where" and "who"). We use this dataset for scene classification purposes only. That is, we classify events by classifying the scenes, and do not attempt to recognize objects or persons.

The images are high resolution ones (from 800x600 to thousands of pixels per dimension). The number of images in each category ranges from 137 to 250. Following [10], we use 70 images per class for training, and 60 for testing. The first 50 images in each category are used to compute the eigenvectors. We use RBF kernel SVM classifiers with level 2 pyramid sPACT features in this dataset.

Overall we achieve 78.50% accuracy on this dataset. In [10], the scene only model achieved approximately 60% accuracy, which is significant lower than the sPACT result. When both scene and object categorization were used, the method in [10] had an accuracy of 73.4%, still inferior to our result. Note that this scene+object categorization used manual segmentation and object labels as additional inputs.

The scene only model of sPACT exhibits different behaviors than the scene+object model in [10], as shown in the confusion matrix in Fig. 10. The most confusing pairs of our method are bocce/croquet, and rowing/sailing. These results are intuitive because these two pairs of events share very similar scene or background. In [10], the most confusing pairs are bocce/croquet, polo/bocce, and snowboarding/badminton. The object categorization helped in distinguishing rowing and sailing. However, it seems that it also confused events that have distinct backgrounds, such as snowboarding and badminton.

5. Conclusions

In this paper we propose PACT, Principal component Analysis of Census Transform histograms, as a representation for recognizing instances and categories of places.

[6] In all the datasets we experimented with, the difference in recognition rates between these two kernel types are smaller than 2%.

[7] However, because of the strong correlation in CT values, it is difficult to translate this linear indoor-outdoor classifier into intuitive semantic interpretations or visualizations.

We show that the Census Transform efficiently captures image structures in the 3×3 local area. We analyze the direct and indirect constraints existing among neighboring CT values. These constraints were shown to propagate to pixels far apart, which enables PACT to implicitly capture the global shape in an image. PACT also handles the strong correlation among pairs of CT values using PCA. We use the one dimensional special case to illustrate why PACT encodes shape and support our statement by experiments on the Sweden leaf dataset. PACT is then combined with the spatial pyramid matching [9] idea. We use spatial PACT to recognize both place instances and categories. On four datasets including both place instance and category recognition tasks, spatial PACT achieves higher accuracies than state-of-the-art methods. Comparing with other representations, sPACT not only exhibits superior performance. It has nearly no parameter to tune and is easy to implement. sPACT also evaluates extremely fast.

There are several limitations of PACT and future research directions to improve it. First, PACT is not invariant to rotations. Although robot acquired images and scene images are usually upright, making PACT rotational invariant will enlarge its application area. Second, we want to recognize place categories in more realistic settings, *i.e.* learning the category concepts using images acquired without human bias. And finally, the PACT representation and learned place category could be applied to facilitate the recognition of objects in the image [22].

Acknowledgments

The authors would like to thank Henrik I Christensen for fruitful discussions, and especially thank the many researchers who kindly published their datasets and software packages online.

References

[1] D. N. Bhat and S. K. Nayar. Ordinal measures for image correspondence. *IEEE TPAMI*, 20(4):415–423, 1998.

[2] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via pLSA. In *ECCV*, volume 4, pages 517–530, 2006.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume II, pages 524–531, 2005.

[5] P. F. Felzenszwalb and J. D. Schwartz. Hierarchical matching of deformable shapes. In *CVPR*, 2007.

[6] W. T. Freeman and M. Roth. Orientation histogram for hand gesture recognition. In *FG workshop*, pages 296–301, 1995.

[7] J. J. Kivinen, E. B. Sudderth, and M. I. Jordan. Learning multiscale representaiton of natural scenes using dirichlet processes. In *ICCV*, 2007.

[8] B. Kuipers and P. Beeson. Bootstrap learning for place recognition. In *AAAI/IAAI*, pages 174–180, 2002.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume II, pages 2169–2178, 2006.

[10] L.-J. Li and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In *ICCV*, 2007.

[11] H. Ling and D. W. Jacobs. Shape classification using the inner-distance. *IEEE TPAMI*, 29(2):286–299, 2007.

[12] J. Liu and M. Shah. Scene modeling using Co-Clustering. In *ICCV*, 2007.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[14] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt. The KTH-IDOL2 database. Technical Report CVAP304, Kungliga Tekniska Hoegskolan, CVAP/CAS, October 2006.

[15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE TPAMI*, 27(10):1615–1630, 2005.

[16] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[17] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen. A discriminative approach to robust visual place recognition. In *IROS*, 2006.

[18] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE TPAMI*, 29(9):1575–1589, 2007.

[19] S. Se, D. G. Lowe, and J. J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *ICRA*, pages 2051–2058, 2001.

[20] O. J. O. Söderkvist. Computer vision classification of leaves from swedish trees. Master's thesis, Linköping University, 2001.

[21] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *CAIVD*, pages 42–51, 1998.

[22] A. B. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *ICCV*, pages 273–280, 2003.

[23] I. Ulrich and I. R. Nourbakhsh. Appearance-based place recognition for topological localization. In *ICRA*, pages 1023–1029, 2006.

[24] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

[25] J. Wolf, W. Burgard, and H. Burkhardt. Robust vision-based localization for mobile robots using an image retrieval system based on invariant features. In *ICRA*, pages 359–365, 2002.

[26] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*, volume 2, pages 151–158, 1994.

[27] Z. Zivkovic, O. Booij, and B. J. A. Kröse. From images to rooms. *Robotics and Autonomous Systems*, 55(5):411–418, 2007.

[28] Z. Zivkovic and B. J. A. Kröse. From sensors to human spatial concepts. *Robotics and Autonomous Systems*, 55(5):357–358, 2007.