Object Recognition with Hierarchical Stel Models

Alessandro Perina[1,2], Nebojsa Jojic[2], Umberto Castellani[1], Marco Cristani[1,3], and Vittorio Murino[1,3]

[1] University of Verona
[2] Microsoft Research
[3] Italian Institute of Technology

Abstract. We propose a new generative model, and a new image similarity kernel based on a linked hierarchy of probabilistic segmentations. The model is used to efficiently segment multiple images into a consistent set of image regions. The segmentations are provided at several levels of granularity and links among them are automatically provided. Model training and inference in it is faster than most local feature extraction algorithms, and yet the provided image segmentation, and the segment matching among images provide a rich backdrop for image recognition, segmentation and registration tasks.

1 Introduction

It is well understood that image registration, segmentation and recognition are related tasks [17,23,18,3], and yet, the engineering paradigm suggests the decomposition of the general vision problem into components, first to be considered (and even applied) in isolation, and then, sometimes, combined as modules.

In some cases, the modular approach is highly successful. For example, algorithms for registration of multiple images of a static scene have recently matured to the point where they can be directly used in a variety of applications (e.g., photosynth.net). The registration algorithms typically do not attempt to solve the recognition or the segmentation problems, and are not readily applicable to registering images of different scenes or objects so that they can be used as modules in recognition algorithms. Still, the feature extraction stage, e.g. SIFT, in these technologies has found its way to object recognition research, but not as a tool for image registration. Under the assumption that registration of images of similar (but not identical) objects would be hard, the image features are compared as if they do not have a spatial configuration, i.e., as bags of visual words (BOW) [1] randomly scattered across the image.

The initial success of BOW models was extended when the researchers attempted to encode at least some spatial information in the models, even if the required spatial reasoning would be short of full image registration. Such models are often computationally expensive. For example, [2] forms vocabularies from pairs of nearby features called "doublets" or "bigamy". Besides taking co-occurrences into account this approach benefits from some geometric invariance,

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part VI, LNCS 6316, pp. 15–28, 2010.

but it is expensive even when feature pairs are considered, and the cost grows exponentially for higher order statistics. In [4] a codebook of local appearances is learned in way that allows reasoning about which local structures may appear on objects of a particular class. However, this process has to be supervised by human-specified object positions and segmentations. Generative part-based models like [6,23] are in principle learnable from unsegmented images, but are computationally expensive as they solve combinatorial search problems. Among the more computationally efficient approaches, the spatial pyramid method [7] stands out. The images are recursively subdivided into rectangular blocks, in a fixed, image-independent way, and the bag-of-words models are applied separately in these blocks. Image similarity is then defined based on the feature histogram intersections. This representation is combined with a kernel-based pyramid matching scheme [8], which efficiently computes approximate global geometric correspondence between sets of features in two images. Having defined an image kernel, or a similarity measure for two images, a variety of off-the-shelf learning algorithms can be used for classification (e.g., the nearest neighbor method, which simply labels the unlabeled test image with the label of the most similar labeled image). While the spatial pyramid indirectly registers images for computation of such a kernel, this registration is limited by the use of a fixed block-partition scheme for all images.

In this paper, we propose a related approach to defining image similarities, which can guide object recognition, but also segmentation and registration tasks. The similarities between two different images are broken down to different regions, but these regions are not rigidly defined by a pyramid kernel, nor do they require combinatorial matching between images as in [11]. Instead, they are computed using a novel hierarchical model based on the probabilistic index map/stel models [10,9,5,18], which consider the segmentation task as a joint segmentation of an image collection, rather than individual images, thus avoiding a costly combinatorial matching of segments across images. Our new hierarchical stel model (HSM) also contains multiple levels of segmentation granularity, linked across the hierarchy, and provides a rich backdrop for image segmentation, registration and recognition tasks, as any new image can be segmented in various class-specific ways under under this set of generative models. In particular, we propose a similarity kernel based on the entire stel hierarchy across all classes and granularity levels, and we demonstrate that the computation of this kernel for two test images implicitly matches not only image segments, but even the object parts at a much finer granularity than that evident in a segmentation under any class model. Not only that such use of HSM leads to high recognition rates, but it also provides surprisingly accurate unsupervised image segmentation, and unusually informative registration of entirely different images.

2 The Basic Probabilistic Index Map/Stel Model

The basic probabilistic index map, PIM [10], or as it is also called, structure element (*stel*) model, assumes that each pixel measurement x_i, with its 2-D coordinate i, has an associated discrete variable s_i, which takes a label from the interval

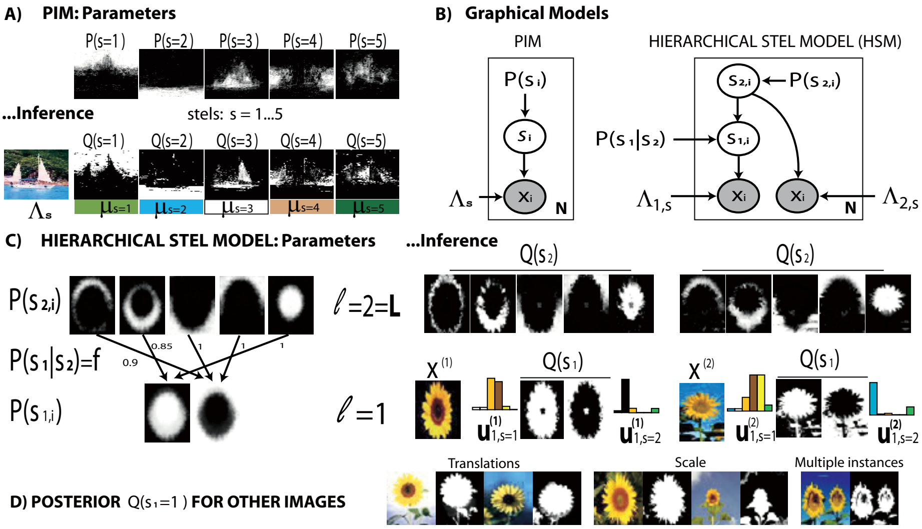

Fig. 1. PIM and Hierarchical stel model (HSM) illustration

$[1, S]$. Such a labeling splits the image into S stels so that s-th stel is a collection of pixel coordinates i, which may be scattered across the image, or grouped together into coherent blobs, and for which the index s_i is set to the desired stel label s, i.e., $\Omega(s) = \{i|s_i = s\}$. Fig. 1A shows some examples of stels: $\Omega(s = 2)$ represents the sea, $\Omega(s = 3)$ the schooner. The stel assignments are almost exclusively considered in a probabilistic fashion. In the simplest case, the distribution over possible assignments of image coordinates to stels is modeled by a set of location-specific distributions $P_i(s_i)$ that describe which image coordinates are more likely to belong to particular stels *a priori*. Such a probabilistic index maps ties the stel partitions in different images of the same type. The posterior stel distribution $Q(s_i)$ describes how this prior belief about class-specific image partition gets altered given the pixel measurements in a particular image (see Fig. 1A). The image evidence that the model detects is the image self-similarity within a stel: the pixels with the same stel label s are expected to follow a tight distribution over image measurements, defined by parameters Λ_s. Each distribution Λ_s can be modeled, for example, as a Gaussian $\Lambda_s = (\mu_s, \sigma_s)$ (in Fig.1 we only show the means μ_s) or in other more complex ways [18,9]. The collection $\{\Lambda_s\}$ of all stel parameters, organized by the stel index, is referred to as a palette. The palette for two different images of the same class can be completely different. Instead of local appearance similarity, the model insists on consistent segmentation through the stel prior. For example stel $\Omega(3)$ in all images of pedestrians may capture the lower part of the background and $\Omega(1)$ the torso of the pedestrian in the foreground (Fig. 3). Differences in local appearance of these parts are explained away as differences in the palettes associated with the images. Moreover, the stel prior is easily learned from a collection of images starting from a noninformative initialization, which allows for efficient segmentation of new images in a fashion consistent with the joint segmentation of the training images. Another view of this model is that captures correlated changes of pixels, as in [24], but in a much more computationally efficient way.

This basic model is easily enriched with transformation variables [18,9] which alleviate the requirement for rough pre-alignment of images. However, even the basic model has a remarkable ability to deal with somewhat misaligned images without the help of extra variables. For example, Fig. 1C-bottom illustrates the basic PIM model of the sunflower category, in which the images undergo significant transformations (scale, translations, multiple instances). Without help with accounting for these transformations explicitly, the prior $P(\{s_i\})$ is soft after learning, but strong enough to tie the segmentations together into consistent stels. Of course, this robustness to image transformation is limited. In case of very fine image segmentations with large number of stels, and/or very large image transformations, and/or a sparse training set, the part correspondence may be highly unreliable. Adding transformation variables could help in such cases, but in this paper we advocate an even more efficient approach that follows a traditional computer vision concept: coarse-to-fine hierarchies.

3 Hierarchical Stel Model (HSM)

Modeling transformation variables is inherently expensive in any model. The cost of dealing with image translation is of the order $N \log N$, where N is the number of pixels, but if we also need to take care of scale, rotation, or even affine transformations, the expense may accumulate quickly. In this paper, our goal is to extend the natural ability of stel models to capture all but the largest transformations. If for instance, the model is not sensitive to the transformations present in the fairly well-aligned Caltech database, then the extra transformation variables only need to model coarse translation in large images (relative to the object size), and capture scale at several coarse levels.

To achieve such an increased invariance to image transformation, we consider stel models at multiple levels of granularity so that the more refined models are linked to the coarser models. This modification confers two advantages to the stel models:

- If the alignment at some level of granularity is failing, the coarser levels may still be useful.
- The higher quality of the alignment of stels at a coarse granularity will guide the alignment at a finer granularity, making these more useful.

Hierarchical stel model captures a hierarchy of stel partitions at L different granularity levels indexed by ℓ: $\Omega^{\ell}(s) = \{i | s_{\ell,i} = s\}$. The index label s can be chosen from sets of different cardinality for stels at different levels of hierarchy. For example, in Fig. 1C we show two levels of hierarchical stel model with two stels in level $\ell = 1$ and five in level $\ell = 2$. The stel partitions are linked hierarchically by distributions $P(s_{\ell,i} = a | s_{\ell+1,i} = b) = f^{\ell}_{a,b}$ which are *not* spatially varying. In Fig. 1C this linking conditional distributions are defined by a 5×2 table of conditional probabilities $f^{1}_{a,b}$, but only a few strongest weights are illustrated by arrows. The image $\{x_i\}$ is linked to each of these stel assignments directly, as if it was generated L times[1] (Fig. 1B).

Given the prior $P^{\ell+1}(\{s_i\})$ for level $\ell+1$ in the same form as in the basic site-specific PIM/stel model of the previous section, the prior for the level below satisfies:

$$P_i^{\ell}(s_{\ell,i}=a)=\sum_b P_i^{\ell+1}(s_{\ell+1,i}=b)\cdot f_{a,b}^{\ell}. \tag{1}$$

In this way, each successive level provides a coarser set of stels, created by (probabilistic) grouping of stels from the previous level according to $f_{a,b}^{\ell}$; only at the finest granularity the stel prior is location-specific, as in the previous section,

$$P(\{s_{L,i}\}_{i=1}^{N})=\prod_i P_i(s_{L,i}). \tag{2}$$

As before, the conditional links between the image observation and the stel assignment at $P(x_i|s_{\ell,i}=s)$ depend only on the s-th palette entry at the hierarchy level ℓ, and not on the pixel coordinate, thus allowing the palette to affect the appearance of all the stel's pixels in concert. For added flexibility, the palette entries capture a mixture of colors. Image colors in the dataset are clustered around 32 color centers, and the real-valued pixel intensities are replaced by discrete indices to these centers in all our experiments. Each palette entry $\Lambda_{\ell,s}$ is thus a histogram consisting of 32 probabilities $\{u_{\ell,s}(k)\}$, and

$$P(x_i=k|s_{\ell,i}=s)=u_{\ell,s}(k). \tag{3}$$

The joint probability over all variables in the model is

$$P=\prod_i P(s_{L,i})\prod_{\ell=0}^{L-1} f_{s_{\ell,i},s_{\ell+1,i}}^{\ell}\prod_{\ell=0}^{L} p(x_i|s_{\ell,i}) \tag{4}$$

where level $\ell=0$ trivially reduces to a bag of words representation as the stel variables across the image are constant $s_{0,i}=1$. Following the same strategy as [10] we can easily write the free energy $F=\sum Q\log\frac{Q}{P}$ for this graphical model assuming a factorized posterior $Q=\prod_{\ell,i}Q(s_{\ell,i})$, take appropriate derivatives, and derive the following inference rules for minimizing the free energy for a single image given the prior over stel hierarchy:

$$Q(s_{\ell,i}=s)\propto P(s_{\ell,i}=s)\cdot u_{\ell,s}(x_i) \quad u_{\ell,s}(k)\propto\sum_i Q(s_{\ell,i}=s)\cdot[x_i=k], \tag{5}$$

where [] is an indicator function. The above updates are image-specific; each image has in fact its own palette of histograms which allows images with very different colors to be segmented following the same stel prior (Fig. 1C).

[1] The motivation for multiple generation of x_i from multiple levels of hierarchy comes from the observation that modeling multiple paths from hidden variables to the data, or, for that matter, among hidden variables in the higher levels, alleviates local minima problems in learning [19].

Given a collection of images indexed by t, and the posterior distributions $Q(s_\ell^t)$ computed as above, the hierarchical stel distribution is updated as

$$f_{a,b}^\ell \propto \sum_{t,i} Q(s_{\ell+1,i}^t = b) \cdot Q(s_{\ell,i}^t = a) \qquad P(s_{L,i} = s) \propto \sum_t Q(s_{L,i}^t = s). \tag{6}$$

These updates are iterated and the model is learned in an unsupervised way form a collection of images. As the result, all images are consistently segmented into stels at multiple levels of hierarchy. As the palettes are image-specific in the model, the images can have completely different colors and still be consistently segmented. The hierarchical representation of stels reduces the errors in segmentation, and provides a rich information about part correspondence for image comparison, and, therefore, recognition.

4 Hierarchical Stel Kernel (HSK)

The HSM can be trained for many different image classes indexed by c. A pair of images (whether they are in one of the training sets for the stel models or not) can be segmented into stels under any of the resulting models $P_c(\{s_{\ell,i}\})$ by iterating the two equations (5). The pair of resulting posterior distributions $Q_c(s_{\ell,i}^A), Q_c(s_{\ell,i}^B)$ for each combination of class c and granularity level ℓ provides a coarse correspondence for regions in the two images (Fig. 2).

This rich information can be used in numerous ways, but we limit our analysis and experiments here to one of the simplest approaches, inspired by the spatial pyramid match kernel [7], which propose course-to-fine spatial feature matching schema based on comparing histograms of image features in different parts of the image and weighting and accumulating evidence of feature sharing. As in [7], we compute image features in images and represent them using the same codebook of 300 visual words. But, instead of partitioning each image image using the same set of rectangular blocks of different sizes, we use the image-specific segmentations induced by HSM models. Then similarity in image features in two different images is considered important if these features tend to be within the same posterior stel under many models.

Specifically, the feature indices $k \in [1, 300]$ are assigned to locations on a grid that covers every fifth pixel along both image dimensions. In a given image, within the s-th stel under the model of class c, at a hierarchy level ℓ an unnormalized histogram of image features $h_{c,\ell,s}(k)$ is computed as

$$h_{c,\ell,s}(k) = \sum_i Q_c(s_{\ell,i}) \cdot n_{i,k} \tag{7}$$

where $n_{i,k}$ is equal to 1 if a feature of index k is present at location i, 0 otherwise. Given two images A and B, their histogram similarities within the corresponding stels are defined by the histogram intersection kernel [8] defined as

$$K(A, B) = \min_k(h_{c,\ell,s}^A(k), h_{c,\ell,s}^B(k)), \tag{8}$$

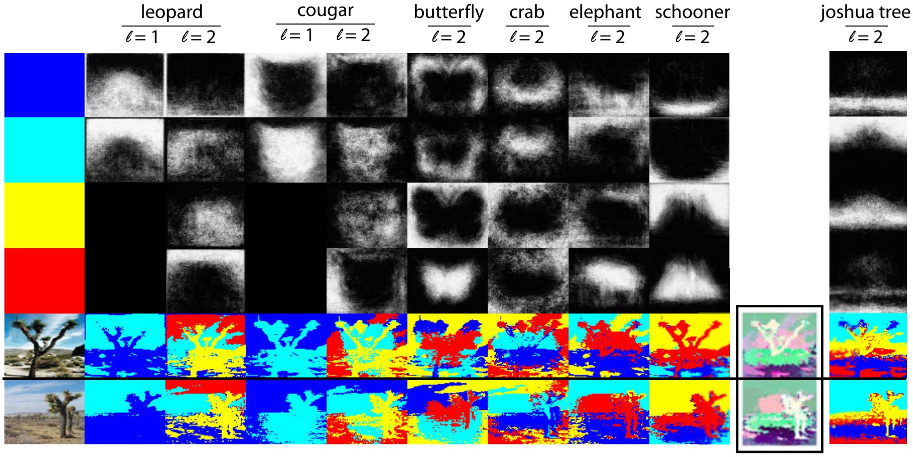

Fig. 2. Segmentations of two images from the Joshua tree category under various stel models trained on Caltech 101 images. The prior stel distributions are illustrated on top. The stels are assigned different colors (blue, light blue, yellow and red), to illustrate the mode of each posterior stel assignment, which is based both on the prior and on the image evidence. Although none of the individual segmentations under the leopard, cougar, butterfly, crab, elephant, and schooner models fits these models very well, the two images are for the most part consistently segmented under these models: If the different stel assignments a pixel gets under these different models are considered a discrete multi-dimensional label, and if these multi-dimensional labels of all pixels are projected through a random matrix onto 3D colors, so that the similar consistent labels across models and levels of hierarchy result in a similar color, then the two joshua tree images end up colored as shown in the rectangular box. This illustrates that the tree bark has consistent stel assignment in two images more often than not, and similar correspondence among other parts of the two scenes are visible. In contrast, a single segmentation, even under the model trained on Joshua tree images (the last column), does not provide a refined part correspondence.

because this provides computational advantages. To compute a single measure of similarity for two images under all stels of level ℓ, we sum all the similarities, weighting more the matches obtained in finer segments:

$$K_c^{HSK}(A,B) = \sum_{l=0}^{L} \frac{1}{2^{L-\ell}} \cdot \sum_s \min_k (h^A_{c,\ell,s}(k), h^B_{c,\ell,s}(k)), \qquad (9)$$

In multi class classification tasks, we define the hierarchical stel kernel (HSK) as the sum of the kernels for individual classes $K^{HSK} = \sum_c K_c^{HSK}$. There are two reasons for this operation. First, when image similarities are computed for classification tasks, one or both images may not be labeled as belonging to a particular class, and so considering all classes simultaneously is needed. Second, even if one of the images belongs to a known class (an exemplar used in classification, for instance) and the other's class is to be predicted, multiple segmentations of the images under different class models provides useful additional alignment information (Fig. 2). When insufficient data is used for training stel models (e.g., 15 training images for Caltech101), the segmentation under any given class may be noisy, and so pulling multiple segmentations may help. Natural images share similar structure: Consider for example portraits of dogs and humans, or structure of different classes of natural scenes, where the background is broken into horizontal stripes in images of schooners and cars alike. Thus, using many stel tessellations under many classes reinforces proper alignment of image parts.

Furthermore, as Fig. 5B illustrates, the alignment becomes finer than under any single model, even than the finest level of stel hierarchy under the model for the *correct* class. To illustrate this, we note that because the posterior $Q(s)$ tends to be peaky, i.e. close to 0 or 1 for most pixels, for any class we have

$$\begin{aligned} K_c^{HSK}(A,B) &\approx \sum_{l=0}^{L} \frac{1}{2^{L-\ell}} \cdot \sum_{i,j} \min_k(n_{k,i}^A, n_{k,j}^B) \times \big(\sum_s \min_{A,B}(Q(s_{\ell,i}^A = s), Q(s_{\ell,j}^B = s)\big) \\ &= \sum_{i,j} F_{i,j} \times M_{i,j} \end{aligned} \tag{10}$$

where $M_{i,j} = \sum_{\ell=0}^{L} \frac{1}{2^{L-\ell}} \big(\sum_s \min_{A,B}(Q(s_{\ell,i}^A = s), Q(s_{\ell,j}^B = s)\big)$ represents the level of expected similarity between the i-th pixel in image A and j-th pixel in image B based simply on how often the stel labels for these two pixels are shared across the hierarchy, and $F_{i,j} = \min_k(n_{k,i}^A, n_{k,j}^B)$ represents feature similarities (i.e., matches) between the coordinate i in one image and coordinate j in the other, independently of any segmentation. Finally we can write

$$K^{HSK} = \sum_{i,j} F_{i,j} \times \sum_c M_{i,j}^c. \tag{11}$$

Here we have that $F_{i,j} > 0$ if in locations i and j the same feature index is present. This feature match is more rewarded through weight $\sum_c M_{i,j}^c$ if i and j share the same stels across different models and granularity levels. Figure 5 illustrates these two components, $F_{i,j}$ and $\sum_c M_{i,j}^c$, of the similarity kernel on the pixel level. First, in Fig. 5A we show how combining three arbitrary classes creates enough context not only to find the corresponding segment for pixel i in the first image, but to actually refine this matching across pixels j in the second. For the selected i, marked by a square, $\sum_c M_{i,j}^c$ is represented as an image over coordinates j in the second image. In the second image, as well as in match maps $\sum_c M_{i,j}^c$, the cross represents the pixel $j = i$ so that the misalignment of the two faces is evident. While the inference under the face class may be sufficient to roughly match large regions among the images, the stel segmentations based on three classes' segmentations narrow down the correspondence of the marked pixel (right eye) to the eye regions of the face in the second image and a spurious match in the background which happened to have a similar color to the facial region. For easier visualization we illustrated only three select stels from the three classes. In Fig. 5B for this example, and several more, we show what happens when all stels and all classes are used as in the equations above. For two facial images, the supplemental video shows correspondence

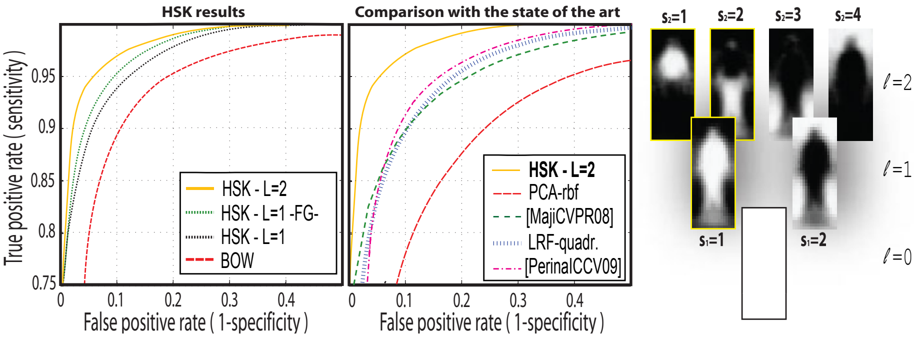

Fig. 3. Pedestrian classification. Left: ROC plots comparing HSM/HSK and other approaches. Right: the learned HSM parameters.

of various pixels in the same manner (The pixel in the first image is marked by a cursor, and the mapping in the second image is shown as a heat map).

Finally in Fig. 5C, we show jointly the mapping of three pixels i_1, i_2, i_3 in the first image by placing the appropriate match maps M in the R, G, and B channels of the image. As the result, when the entire stel hierarchy under all classes is used to evaluate $\sum M$, the regions around the eyes, and especially around the right eye in the second image are colored red, while the regions in the lower part of the face, especially lips, are colored green, and the background elements are colored blue, indicating that the entire stel model hierarchy can localize the face parts beyond the granularity of any single model and any single level of hierarchy. For comparison, M obtained for the face class only and butterfly class only are shown. To illustrate in the same manner the spatial pyramid kernel [7], we compute similar decomposition into the expected matching of pixels based on block image segmentation, and the feature matching of pixels. The complete kernel under both HSM and the spatial pyramid is the sum over all pixels of the product $M_{i,j} \cdot F_{i,j}$ and so these products are also illustrated in the figure.

Inference and learning complexity in stel models is linear in the number of image coordinates, stels and classes. The total computation time is considerably faster than SIFT feature computation. Furthermore, the quality of image matching does not decay much if we use only 30 out of 101 classes.

5 Experiments

We evaluated our approach on Caltech28, Calteh101 and Daimler pedestrian datasets. We compared with the classification results provided by the datasets' creators and with the other feature organization paradigms, namely Bag of words (BW), Stel organization (SO) and Spatial Pyramids (SPK), as well as other state-of-the art methods. We considered both classification and unsupervised segmentation tasks. We used support vector machines as discriminative classifiers, feeding the kernels as input.

5.1 Pedestrian Classification: Daimler Dataset

We evaluated our method on pedestrian classification using the procedure of [12]. We trained a hierarchical stel model with $S_1 = 2$ and $S_2 = 4$ on the training set for each class (See Fig. 3 for an illustration). Having trained HSM on the training data, stel inference can be performed on test images, so that pairwise similarities (the kernel matrix) can be computed for all pairs of images (training and test). For the feature code book, we used the dictionary of Haar wavelets [13]. Given input images of size 18 x 36 and their posterior distributions $Q(s_1^t)$ and $Q(s_2^t)$, we compute w_l^t convolving the image x^t with wavelets of scales 4 x 4 (l=1) and 8 x 8 (l=2). We only encoded the magnitude in the feature vectors. As described above, image features and stel segmentations are used to compute the kernel matrix and this matrix is fed to a standard SVM classification algorithm. The ROC plots are shown in Fig. 3. As expected, results improve as we go from L = 0 (AUC, Area under the curve, 0.954) to a multi-level setup (L > 0). We repeated the classification only keeping into account the foreground wavelet coefficients. When L=1 the accuracy is significantly improved by considering only the foreground, but for L=2 it does not, as the hierarchical stel kernel already reaches impressive performance without emphasizing foreground in classification. Though matching at the highest pyramid level seems to account for most of the improvement (AUC 0.9751), using all the levels together confers a statistically significant benefit (AUC 0.9854). The ROC plot on the right of figure 3 compares HSK with several recent approaches including [12] which reviews standard pedestrian classification algorithm and features, [15] which uses a hybrid generative-discriminative approach based on PIM [10], and [14] which employs spatial pyramids kernel on a multi-level version of the HOG descriptor [16].

5.2 Unsupervised Segmentation and Supervised Recognition of Caltech 28 Images

Caltech 28 [17] is composed of 28 classes of objects among the subset of Caltech 101 categories that contain more than 60 images. The chosen categories contain objects with thin regions (e.g. flamingo, lotus), peripheral structures (e.g. cup), objects that are not centered (e.g. leopards, dalmatians, Joshua trees). None of the chosen classes contains background artifacts that make them easily identifiable. For each class, we randomly selected 30 images for training and 30 images for testing. To serve as discrete features to match, we extracted SIFT features from 15x15 pixel windows computed over a grid with spacing of 5 pixels. These features were mapped to W=300 codewords as discussed in Section 4. We trained a hierarchical model for each class using $S_1 = 3$ and $S_2 = 5$ and then

Table 1. Classification accuracies on Caltech 28

HSK L=1	HSK L=1	HSK L=2	[9]	SPK [7]	BW	[17]
$S_1 = 3$	$S_1 = 5$	$S_1 = 3, \ S_2 = 5$	-	L=2	-	-
73,15%	74,57%	**78,10%**	65,12%	65,43%	56,01%	69%

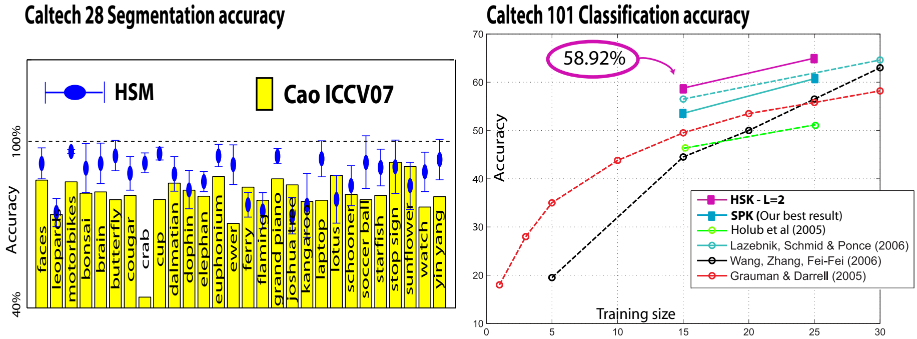

Fig. 4. Classification results for the Caltech experiments. On the left we report the segmentation accuracy for each class of Caltech 28 obtained by [17] (yellow bars) and by HSM (blue dots with confidence level). On the right, we compare recognition rates on Caltech 101 images with related spatial-reasoning methods using similar local features.

performed inference on the test images. We calculated the kernel between all pairs of images as discussed in Section 4 and the used a standard SVM that uses the class labels and kernels to determine the missing class labels of images in the test set. We compared the results of several set ups of HSK and with: *i*) the bag of words classifier BW, *ii*) the spatial pyramid kernel (SPK, [7]), and *iii*) a classifier based on the single level stel partition (SO, S=5, [9]). All the methods are compared using the same core-kernel (histogram intersection) and the same feature dictionary. First, we compared these related methods repeating the classification 10 times with a randomly chosen training-testing partition. Then we performed t-tests and found:

$$BW <<^{1\cdot10^{-3}} SPK <<^{3\cdot10^{-3}} HSK >>^{5\cdot10^{-4}} SO >>^{4\cdot10^{-3}} BW^{2} \qquad (12)$$

Where $>>^p$ stands for greater with statistical significance with p-value equal to p. HSK's advantage here is due to the segmentations provided by HSM, which explain away a lot of object transformations (see Fig.1C, bottom) and capture meaningful object partitions. Mean classification accuracies are summarized in table 1. As a further test on Caltech 28 we tackled image segmentation, simply using the posterior stel segmentation induced by the coarsest level of HSM ($S_1 = 2$). Each class of images is fit independently as described in Section 3. After training, the posterior stel distributions are used as image segmentations. We compared our results with [17], which provides the manual labeling of pixels. In figure 4 we compare the segmentation accuracy over different classes. The overall test accuracy of our unsupervised method is 79,8%, outperforming the supervised method of [17] with test accuracy of 69%.

[2] SO and SPK have been found statistically equal.

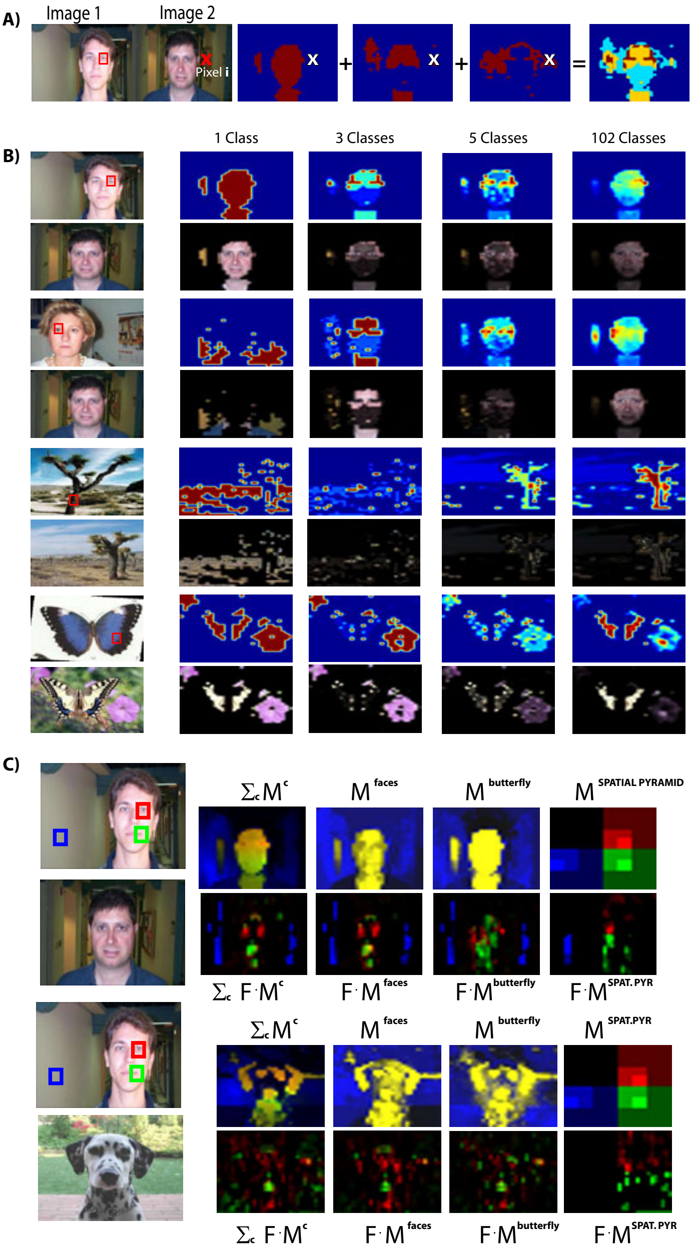

Fig. 5. Image correspondences implicitly captured by the hierarchical stel kernel. In A and B, the pairs of images are shown with the pixel of interest in the first image labeled by a square. In B, for each pair, the stel-based match matrix M, which is only based on color stel models, is shown as averaged under 1,3,5, and 102 classes randomly selected from Caltech 101. Below each M matrix we show it multiplied with the target image. C illustrates the correspondence of multiple points for two image pairs.

5.3 Recognition Rates on Caltech 101

Our final set of experiment is on the Caltech 101 dataset. For the sake of comparison, our experimental setup is similar to [7]. Namely, we randomly select 30 images from each category: 15 of them are used for training and the rest are used for testing. We compare our method to only those recognition approaches that do not combine several other modalities. Results are reported in figure 4 The successfully recognized classes include the ones with rotation artifacts, and the natural scenes (like joshua tree and okapi), where segmentation is difficult. The least successful classes are animals, similarly to [7]. This is likely not due to problems of segmentation, but discretized feature representation [20]. Since our goal is mainly to compare our representation with SPK we report the results we have obtained using the SPK authors's implementation of the feature extraction and quantization. Note that due to a random selection of images, we did not recreate the exact classification result of SPK, but our HSK similarity measure outperforms both our implementation of the SPK and the best published SPK result.

6 Conclusions

We propose a new generative model, and a new image similarity kernel based on a linked hierarchy of stel segmentation. The goal of our experiments was primarily to demonstrate the spatial reasoning that can be achieved with our method, and which goes beyond block comparisons, and even beyond segment matching and closer to registration of very different images. Therefore we compared our method using the same discretized features as in the literature describing efficient spatial reasoning approaches. However, we expect that the better local feature modeling may improve classification performance, as for example, [20] proposes. Still, even with current discretized features, the hierarchical stel models can be used efficiently and with high accuracy in segmentation and classification tasks. We expect that our image representation will find its applications in multikernel approaches but may also find other applications due to its ability to combine image recognition, segmentation, and registration. For example [21,22] are based on SPK and could be easily used with our method.

References

1. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR 2005 (2005)
2. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: ICCV 2005 (2005)
3. Russell, B.C., Efros, A., Sivic, S., Freeman, W.T., Zisserman, A.: Segmenting Scenes by Matching Image Composites. In: NIPS 2009 (2009)
4. Leibe, B., et al.: An implicit shape model for combined object categorization and segmentation. In: Ponce, J., Hebert, M., Schmid, C., Zisserman, A. (eds.) Toward Category-Level Object Recognition. LNCS, vol. 4170, pp. 508–524. Springer, Heidelberg (2006)

5. Ferrari, V., Zissermann, A.: Learning Visual Attributes. In: NIPS 2007 (2007)
6. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1842, pp. 18–32. Springer, Heidelberg (2000)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR 2006 (2006)
8. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: ICCV 2005 (2005)
9. Jojic, N., Perina, A., Cristani, M., Murino, V., Frey, B.J.: Stel component analysis: Modeling spatial correlations in image class structure. In: CVPR 2009 (2009)
10. Jojic, N., Caspi, Y.: Capturing image structure with probabilistic index maps. In: CVPR 2004 (2004)
11. Russell, B., et al.: Using Multiple Segmentations to Discover Objects and their Extent in Image Collections. In: CVPR 2006 (2006)
12. Munder, S., Gavrila, D.: An experimental study on pedestrian classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 1863–1868 (2006)
13. Papageorgiou, C., Poggio, T.: A trainable system for object detection. Int. J. Comput. Vision 38, 15–33 (2000)
14. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR 2008 (2008)
15. Perina, A., et al.: A Hybrid Generative/discriminative Classification Framework Based on Free-energy Terms. In: ICCV 2009 (2009)
16. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: ICCV 2005 (2005)
17. Cao, L., Fei-Fei, L.: Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes. In: ICCV 2007 (2007)
18. Winn, J., Jojic, N.: LOCUS: Learning Object Classes with Unsupervised Segmentation. In: ICCV 2005 (2005)
19. Jojic, N., Winn, J., Zitnick, L.: Escaping local minima through hierarchical model selection: Automatic object discovery, segmentation, and tracking in video. In: CVPR 2006 (2006)
20. Boiman, O., Shechtman, E.: In Defense of Nearest-Neighbor Based Image Classification. In: CVPR 2008 (2008)
21. Yang, J., Yuz, K., Gongz, Y., Huang, T.: Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In: CVPR 2009 (2009)
22. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active Learning with Gaussian Processes for Object Categorization. In: ICCV 2007 (2007)
23. Sun, M., Su, H., Savarese, S., Fei-Fei, L.: A multi-view probabilistic model for 3d object classes. In: CVPR 2009 (2009)
24. Stauffer, C., Miller, E., Tieu, K.: Transform-invariant image decomposition with similarity templates. In: NIPS 2002 (2002)