Stel component analysis: Modeling spatial correlations in image class structure

Nebojsa Jojic[1], Alessandro Perina[2], Marco Cristani[2], Vittorio Murino[2,3] and Brendan Frey[4]

Abstract

As a useful concept in the study of the low level image class structure, we introduce the notion of a structure element – 'stel.' The notion is related to the notions of a pixel, superpixel, segment or a part, but instead of referring to an element or a region of a single image, stel is a probabilistic element of an entire image class. Stels often define clear object or scene parts as a consequence of the modeling constraint which forces the regions belonging to a single stel to have a tight distribution over local measurements, such as color or texture. This self-similarity within a region in a single image is typical of most meaningful image parts, even when in different images of similar objects the corresponding parts may not have similar local measurements. The stel itself is expected to be consistent within a class, yet flexible, which we accomplish using a novel approach we dubbed stel component analysis. Experimental results show how stel component analysis can assist in image/video segmentation and object recognition where, in particular, it can be used as an alternative of, or in conjunction with, bag-of-features and related classifiers, where stel inference provides a meaningful spatial partition of features.

1. Introduction

Due to the high sensitivity of pixel intensities to various imaging conditions, images are often represented by indices referring to a set of possible local features. The image features are chosen so that they are more robust to imaging conditions than the straight-forward color intensity measurements, and often, their spatial configuration is discarded, e.g., in bag of- words models [5, 6]. In such models, each image class has a distinctive palette of features typically found in most instances of the class, although the image locations in which these features are found vary substantially. Therefore, the distribution over indices into a single palette describes an entire image class.

[1] jojic@microsoft.com; Microsoft Research, Redmond, WA, USA
[2] {alessandro.perina, marco.cristani, vittorio.murino}@univr.it
Dipartimento di Informatica, Università di Verona, Italy
[3] Istituto Italiano di Tecnologia, Genova, Italy
[4] frey@psi.toronto.edu; University of Toronto, Canada

In contrast, in the probabilistic index map model [1] the feature palette is pertinent only to a single instance of a class, while the indexing configuration is assumed to be relevant to the entire class of images. As illustrated in the cartoon example of the face category in Fig. 1A, this representation allows for "palette invariance" in image modeling and dramatically reduces sensitivity of models to various unimportant, but usually troublesome causes of variability in images, such as illumination, surface color, or texture variability.

This paper provides modeling ideas that go beyond these basic concepts of feature palette and index map modeling, illustrated briefly in Fig. 1. We define index maps as ordered sets of indices $s_{\mathbf{i}} \in 1, \ldots, S$, linked to spatially distinct areas $i \in 1, \ldots, N$ of images or videos, where N is the number of such image areas (e.g. pixels). These indices point to a table of S possible local measurements, referred to as a palette. Probabilistic index maps (PIMs) consider the uncertainty in the indexing operation as well as in the nature of the palette. Each image location i is associated with a prior distribution over indices $p(s_{\mathbf{i}} = s)$. Indices point to palette entries which each describe a distribution over local measurements.

We refer to an area of an image with the same assigned index s as a *structure element*, or *stel*. A stel will be considered probabilistically as a map $q(s_{\mathbf{i}}^t = s)$ over image locations i defining the certainty of pixels of the $t-$th image belonging to the $s-$th stel. Several of these maps, extracted from a face dataset, are shown in Fig. 1A. The preservation of a stel over a coherent image class, such as an object class or a category, a video segment, etc., is defined in flexible terms through the learned prior distribution over stels $p(\{s_{\mathbf{i}}\}^N = s)$. The extent of tolerable variation of the shape of the stel area, which will often be discontiguous, strongly depends on the power of the statistical model over the indices $s_{\mathbf{i}}$. While the PIM model has the advantage over bag-of-features models in terms of capturing spatial structure in images, it does not deal with (1) the possibility of consistent palettes across images of the class (e.g. in a video sequence, subsequent images do have similar palettes), and (2) with possible dependencies among the spatially separated indices $s_{\mathbf{i}}$. As a result of the second drawback, PIM does no capture correlations that exist in structural elements of an image class due to global effects, such as a slight

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

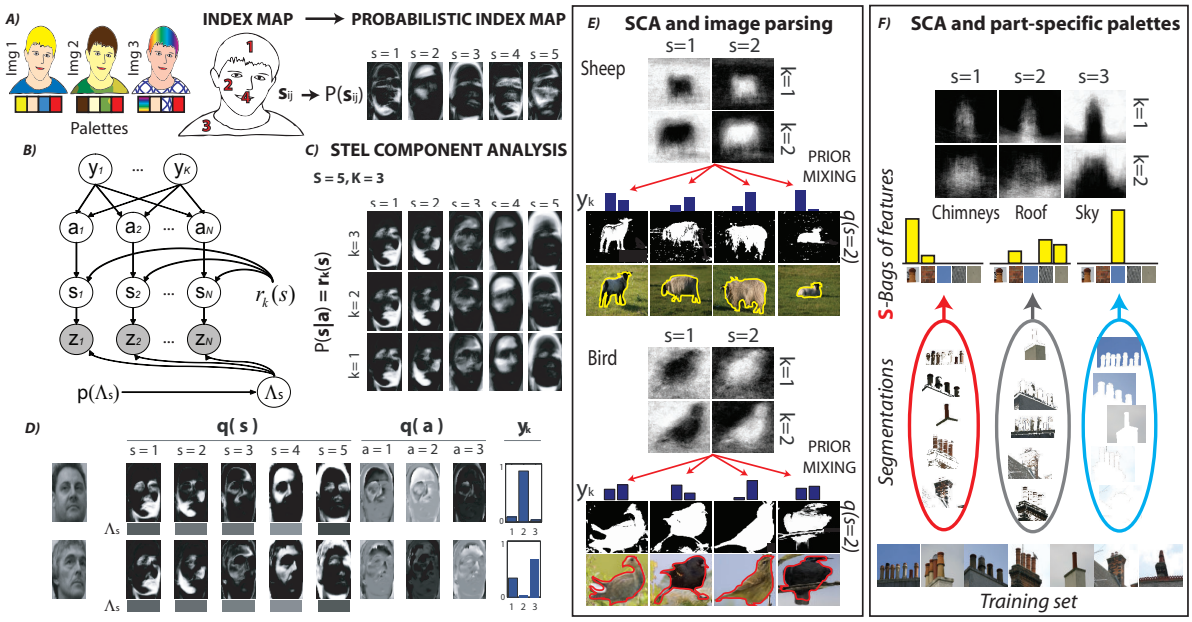


Figure 1. SCA Illustration

change in face proportions (Fig. 1A) which may induce many correlated changes in indices across the image.

We address both of these problems in this paper and propose a new model (Fig. 1B), which we call *stel component analysis*. The model of index variation, in the spirit of principle component analysis and other subspace models used for modeling real-valued pixel intensities, captures correlated variations in discrete indices by blending several component PIMs based on real-valued weights y. This is illustrated in Fig. 1C, where three PIM components are shown, and Fig. 1D, where the blending of these components allows for a better agreement of the observed facial image with the model. The model, described in the next section and in Fig. 1B, was estimated from a set of facial images in an unsupervised manner. As most of the stel structure in this example reflects the grouping of surface normals, the component mixing strengths y capture the varying pose angle for this set of images, as discussed in Experiments. However, for other image categories, different structure may be learned, as illustrated in the rest of the figure. For example, SCA with two stels can be used to segment foreground objects from the background (Fig. 1E). As in case of a PIM relative, LOCUS [2], the segmentations are performed jointly over a set of images from the same category without any supervision, exploiting the self similarity patterns in images to define stels as segments of an image class, rather than individual images. The figure also emphasizes the difference between the prior over stels $p(s)$ and the inferred indices for individual images $q(s_{\mathbf{i}}^t)$, which depend on both the prior and the self similarity properties of an individual image.

In more complex categories, the model benefits from learning a prior over individual image palettes, which is similar to what is achieved in the bag-of-feature models, except that these appearance models can now be part-specific. In Fig. 1F, SCA is applied to roof images, where the prior over individual image palette is represented by different histograms over image features in the three different stels. Here, we illustrate stel segmentation by grouping parts of different images obtained as pointwise products between stel maps $q(s_{\mathbf{i}}^t)$ and the pixel intensities. Each of the stels has a learned prior over image features, allowing a separation of the sky color from other features. This example illustrates the advantages of the model presented here over both bag-of-words models and the PIM model. Where the image class does indeed have consistent features across its instances, our model, unlike PIM, captures this through a prior over palettes. But, unlike the bag-of-words models, our model keeps the features typical of different image parts separated, and the segmentation most appropriate for modeling the image class is inferred jointly with these feature distributions through unsupervised learning.

2. Stel component analysis

To make image models invariant to changes in local measurements, while sensitive to changes in image structure, a measurement $z_{\mathbf{i}}^t$ (e.g. the pixel intensity or a feature) at the

location $\mathbf{i} = (i, j)$ in the t-th image of a certain class (object category or a video clip, for instance), is considered to depend on a hidden index $s_{\mathbf{i}}^t \in \{1, ..., S\}$, Fig. 1B:

$$p(z_{\mathbf{i}}^t | s_{\mathbf{i}}^t = s) = p(z_{\mathbf{i}}^t | \Lambda_s^t) \quad (1)$$

The s-th structure element (stel) indicates pixels $\{\mathbf{i} | s_{\mathbf{i}}^t = s\}$ which follow a shared distribution over local measurements (palette) with parameters Λ_s^t. In the example in Fig. 1D, each palette entry defines a single Guassian model with its mean and variance over intensity levels, $\Lambda_s^t = \{\mu_s^t, \phi_s^t\}$, as was previously done in [1]. The inferred means μ_s^t of such palette entries for several facial images are shown in the lower part of each stel. Palettes Λ_s^t are considered hidden variables in the model, each defining a limited diversity of local measurements within a different image. However, the stels are generated from a single distribution shared among all images of the class $p(\{s_{\mathbf{i}}\})$. Fig. 1A shows the estimated distribution of the form:

$$p(\{s_{\mathbf{i}}\}) = \prod_{\mathbf{i}} p(s_{\mathbf{i}}). \quad (2)$$

To visualize the class stel distribution, for the face pose dataset, in Fig. 1A we show the estimated $p(s_{i,j} = s)$ for $s \in \{1, 2, 3, 4, 5\}$ as five images: In the s-th image, the pixel intensity indicates the probability that the location is mapped to index s. The observed pixel intensity $z_{\mathbf{i}}$ tends to be uniform within a stel in a single image, and can be inconsistent across different facial images, as they may be darker or brighter, for example. However, the stels are relatively consistent over facial images and they represent interesting image structure beyond intensity levels. For example, stel s=2 captures parts of forehead and cheek that have similar surface normals, while the eyebrows and the hair are grouped into stel s=3, despite the variability in hair color across images. This model assumes independence of distributions over indices across different image locations, ignoring the correlations in index variation which often arise even from simple structural variation in the image, such as variation in face proportions, or out of plane head rotation. In models of variation in real-valued, rather than discrete, arrays, such correlations are often captured using a subspace model, e.g. PCA, which achieves this through a linear combination of several components. Since we are concerned here with modeling a distribution over discrete indices, we develop a discrete analogue to eigen images, similar in spirit to multinomial PCA or latent Dirichlet analysis models, but with some important distinctions. Our model is meant to capture spatial structure, and thus it is designed for *ordered* index sets, and it also allows *spatially nonuniform* mixing of the components.

In stel component analysis, the components r_k, $k \in \{1, ..., K\}$ are of the same form as (2), $r_k(\{s_{\mathbf{i}}\}) = \prod_{\mathbf{i}} r_k(s_{\mathbf{i}})$. An example of learned components is shown in Fig. 1B. These components are combined to define the distribution $p(\{s^t\})$ using component strengths $y_k^t \in [0, 1]$, so that $\sum_k y_k^t = 1$. The components strengths are real-valued hidden variables for image t, rather than component priors in a mixture model as in a mixture of probabilistic index maps (MPIM) [1]. Each image is thus defined by a hidden point in the simplex defined by $\sum_k y_k^t = 1$, and this point will rarely fall in a vertex (see Fig. 1), whereas in a mixture model, each image will have a discrete pointer to a single component. To achieve this, as well as to allow spatially nonuniform mixing of components based on real-valued component strengths, we add a layer of discrete hidden variables $a_{\mathbf{i}} \in \{1, .., K\}$ which act as mixture component indicators, *but only locally* for their corresponding image locations $\mathbf{i}$. Hidden component strengths y_k, shared across the pixels of an image, then act as prior probabilities in these local mixture models:

$$\begin{aligned} p(\{s_{\mathbf{i}}^t\} | \{a_{\mathbf{i}}^t\}) &= \prod_{\mathbf{i}} p(s_{\mathbf{i}}^t | a_{\mathbf{i}}^t) \\ p(s_{\mathbf{i}}^t | a_{\mathbf{i}}^t = k) &= r_k(s_{\mathbf{i}}^t); \quad p(a_{\mathbf{i}}^t = k | y_k^t) = y_k^t. \end{aligned} \quad (3)$$

By summing over hidden variables a a desired mixing of components with real-valued weights y_k^t is achieved to form a differently mixed stel distribution for each image. Since different hidden variables $a_{\mathbf{i}}$ can have different values (and thus choose different components r_k in different parts of the image), the mixing is spatially nonuniform, and each variable y_k influences only the total number of image locations $\mathbf{i}$ that choose $r_k(s_{\mathbf{i}})$ as the local prior on the index. This allows dramatically more flexible mixing than in PCA models, making object part alignment across images much easier to achieve without global image transformations.

We should note that if the index maps are considered more broadly than in the sense offered in [1], we can recognize several past approaches to modeling spatial correlations in index maps. Such models have mostly been limited to three basic approaches. In one, a Markov random field is used to define several potentials that govern local spatial correlations among some of the image features e.g., [10]. In the second approach, each feature is given its spatial distribution in an image, which imposes probabilities of seeing a particular index in a particular spot [6, 4]. Finally, in [1, 2], site-specific distribution over indices (2), which assume independence in index variation across image locations, are enriched with transformation/deformation models or are used within a mixture model. The stel component analysis is more flexible than these models, as it captures higher-order statistics than Markov random fields, and can adapt to a variety of image deformations without parameterizing them ahead of time as in [1, 2]. Again, the use of the layer of hidden variables $a_{\mathbf{i}}$ makes the model different from a simple mixture of site-specific models. Index probabilities from different components are blended differently

in different parts of the image, which simple mixture models do not allow. This gives the model more flexibility in parsing images, and, as desired, allows for variable mixing of the components for different images to model smooth geometric changes (Fig. 1).

The joint distribution over all observed variables $\mathbf{z} = \{z_{\mathbf{i}}^t\}$, and hidden variables/parameters $\mathbf{h} = \{\{y_k^t\}, \{a_{\mathbf{i}}^t, s_{\mathbf{i}}^t\}, \{\Lambda_s^t\}, \{r_k\}\}$ is

$$p(\mathbf{z},\mathbf{h}) = \prod_t \left(p(\{y_k^t\}_{k=1}^K) p(\{\Lambda_s^t\}) \prod_{\mathbf{i}} p(z_{\mathbf{i}}^t | s_{\mathbf{i}}^t, \{\Lambda_s^t\}) \prod_k (y_k^t r_k(s_{\mathbf{i}}^t))^{[a_{\mathbf{i}}^t = k])} \right) \quad (4)$$

where $[\cdot]$ is the indicator function. The priors on y_k can be kept flat (as in our experiments), or learned in a Dirichlet form. The prior on PIMs r_k was kept flat, i.e. omitted in equations.

Following the variational inference recipe, we introduce a tunable distribution $q(\mathbf{h})$ over the hidden variables/parameters, define as a bound on the log likelihood $\log p(\mathbf{z})$, the negative free energy $-F = \sum_{\mathbf{h}} q(\mathbf{h}) \log \frac{q(\mathbf{h})}{p(\mathbf{z},\mathbf{h})}$, and pursue the strategy of minimizing this free energy iteratively. We used the simplest of the algorithms from this family, where the approximate posterior distribution $q(\mathbf{h})$ is fully factorized, $q(\mathbf{h}) = \prod_k q(r_k) \prod_{i,t} q(a_{\mathbf{i}}^t) q(s_{\mathbf{i}}^t) \prod_t q(y_k^t) q(\{\Lambda_s^t\})$, with $q(r_k)$, $q(y_k^t)$ and $q(\Lambda_s^t)$ being Dirac functions centered at the optimal values (or vectors) $\hat{r}_k$, $\hat{y}_t^k$, $\{\hat{\Lambda}_s^t\}$. As a result, the (approximate) inference reduces to minimizing the following free energy,

$$F = \sum_t p(\{\hat{\Lambda}_s^t\}) + \sum_{t,\mathbf{i},s} q(s_{\mathbf{i}}^t = s) \log p(z_{\mathbf{i}}^t | s_{\mathbf{i}}^t, \{\hat{\Lambda}_s^t\}) + \\ + \sum_{t,\mathbf{i},a} q(a_{\mathbf{i}}^t = a) \log \hat{y}_a^t + \\ + \sum_{t,\mathbf{i},a,s} q(a_{\mathbf{i}}^t = a) q(s_{\mathbf{i}}^t = s) \log \hat{r}_a(s_{\mathbf{i}}^t = s), \quad (5)$$

which is reduced by each of the following steps:

- The palettes for different stels in a single image t are assigned so as to balance the need to agree with the prior $p(\Lambda)$ with the statistics of the local measurements within a (probabilistic) stel in the image:

$$\hat{\Lambda}_s^t = \arg\max \log p(\{\hat{\Lambda}_s^t\}) + \\ + \sum_i q(s_{\mathbf{i}}^t = s) \log p(z_{\mathbf{i}}^t | s_{\mathbf{i}}^t = s, \{\hat{\Lambda}_s^t\}). \quad (6)$$

  (More details below).

- The stel segmentation of image t is based on the similarity of observed local measurements to what is expected in a particular class stel s according to the estimated palette $\hat{\Lambda}_s^t$ in this particular image, as well as the expected stel assignment based on mixed components $r_k(s)$.

$$q(s_{\mathbf{i}}^t = s) \propto p(z_{\mathbf{i}}^t | s_{\mathbf{i}}^t, \{\hat{\Lambda}_s^t\}) e^{\sum_a q(a_{\mathbf{i}}^t = a) \hat{r}_a(s_{\mathbf{i}}^t = s)}. \quad (7)$$

- The spatially nonuniform component mixing, defined by $q(a)$, is updated so as to balance the agreement with the overall strength y_a^t of the component a in the particular image t, with the agreement of the stel assignment with the stel component r_a:

$$q(a_{\mathbf{i}}^t = a) \propto y_a^t e^{\sum_s q(s_{\mathbf{i}}^t = s) \log r_a(s_{\mathbf{i}}^t = s)}. \quad (8)$$

- The stel component strengths y_a are assigned proportional to their use in the image:

$$y_a^t \propto \sum_{\mathbf{i}} q(a_{\mathbf{i}}^t = a). \quad (9)$$

- The stel components r_a are updated to reflect the assignment statistics over all images:

$$r_a(s) \propto \sum_t q(a_{\mathbf{i}}^t = a) q(s_{\mathbf{i}}^t = s). \quad (10)$$

Local measurements z^t, palette models $p(z^t|\Lambda_s^t)$ and palette priors $p(\Lambda_s)$

The local measurements $z_{\mathbf{i}}$ may vary depending on the application, and can be scalar or multidimensional, discrete or real-valued. To obtain the face model in Fig. 1, as in [1], we assumed that the local measurements are simply the real-valued image intensities, that the palette model $\Lambda_s = (\mu_s, \phi_s)$ is Gaussian, $p(z_{\mathbf{i}}^t | s_{\mathbf{i}}^t = s, \{\hat{\Lambda}_s^t\}) = \mathcal{N}(z_{\mathbf{i}}^t; \mu_s^t, \phi_s^t)$, and that the prior on the palette Λ_s is flat. The palette update is therefore based on sufficient statistics over intensities within stels in individual images (See [1] for details). Alternative local measurements include color, disparity, flow, SIFT [7] or some other local features. As more expressive palette models, we use the histogram representation for discrete local measurements, and the mixture of Gaussians for the real-valued measurements (Fig. 4).

For the case of discrete measurements, we define the palette as a histogram over C possible observations $\{\zeta_j\}$, $j \in \{1, ..., C\}$. The observation distribution is multinomial with parameters $u_j = p(z = \zeta_j)$, and the palettes $\Lambda_s = \{u_{s,j}\}$ are defined by these probabilities. With a flat prior on Λ, the equation (6) reduces to

$$u_{s,j}^t \propto \sum_{\mathbf{i}} q(s_{\mathbf{i}}^t = s) [z_{\mathbf{i}}^t = \zeta_j] \quad (11)$$

When measurements consist of different modalities, which are generally uncorrelated at the local level (except

through higher level variables in the model), they are combined by setting $p(z_{\mathbf{i}}|s) = \prod_m p(z_{m,i}|\Lambda_{m,s}) = \prod_m \prod_j u_{m,s,j}^{[z_{m,i}=\zeta_m,j]}$, where m denotes different modality (e.g., available pixel label and discrete texture features associated with each pixel).

To avoid complete palette invariance, we also add a Dirichlet prior on the histogram palette models:

$$p(\Lambda) = p(\{u_j\}) = \frac{1}{Z(\{\alpha_j\})} \prod_j u_j^{\alpha_j - 1}, \qquad (12)$$

which is estimated from the data iteratively together with other updates. The effect of this prior on the palette updates in (6) for different modalities m is $u^t_{m,s,j} \propto \alpha_{m,s,j} - 1 + \sum_{\mathbf{i}} q(s^t_{\mathbf{i}} = s)[z^t_{m,i} = \zeta_{m,j}]$, and the appropriate update on palette priors α_j can be shown to be:

$$\{\hat{\alpha}_{s,m,j}\} = \arg\max \sum_t (\alpha_{s,m,j} - 1) \log u^t_{m,s,j}, \qquad (13)$$

subject to the appropriate normalization constraint. The addition of the (learnable) prior over palette entry allows the model to discover and exploit consistency of local measurements across instances of a class, if there is any. In case of real-valued measurements of arbitrary dimensionality, the palette entry is defined by a mixture of C Gaussians, and the appropriate palette priors are added similarly as in the case of discrete measurements. The treatment of Gaussian component probabilities in each entry is identical to the treatment of discrete measurement frequencies above, while the mean and covariance matrix have the appropriate conjugate priors (Gaussian and scaled inverse Gamma, respectively). Being a mixture, each palette entry has a hidden variable pointing to one of the C Gaussians.

When the raw local measurement is real-valued, e.g. a filter response, we can still choose to discretize it rather than use a real-valued model. Finally, we often combine discrete and real-valued modalities, in the same way the multiple discrete modalities are combined (Fig. 4).

Relationship to other models. We can express many other models frequently used in vision and elsewhere as special cases by assuming an appropriate number of stels S, the components K, and the palette entry size C. The color histogram model and the bag of words/features model [5, 6] are achieved with $S = 1$. On the other hand, when $S > 1$, but only a single component $r_k(s)$ is used, $K = 1$, and each palette entry represents a single Gaussian, $C = 1$, and the prior over palettes is fixed to flat, our model reduces to a probabilistic index map (PIM) [1]. Finally, the basic ingredient of LOCUS [2] is a model we get when we set $S = 2$ (foreground/background), and use a large C to represent color histograms in each palette entry[1]. If we fix the stel partition $q(s^t_{\mathbf{i}})$ to a division of image into regions by hand, rather than let them be estimated from images themselves, the model becomes similar to [4].

[1]Both LOCUS and PIM contained transformation variables, which capture correlations due to a given set of simple 2D geometric transformations, while stel component analysis learns (approximately) arbitrary correlations in possible index assignments across an image. The palette choices we discuss here apply to all three models.

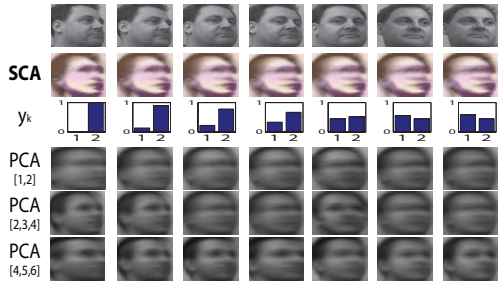
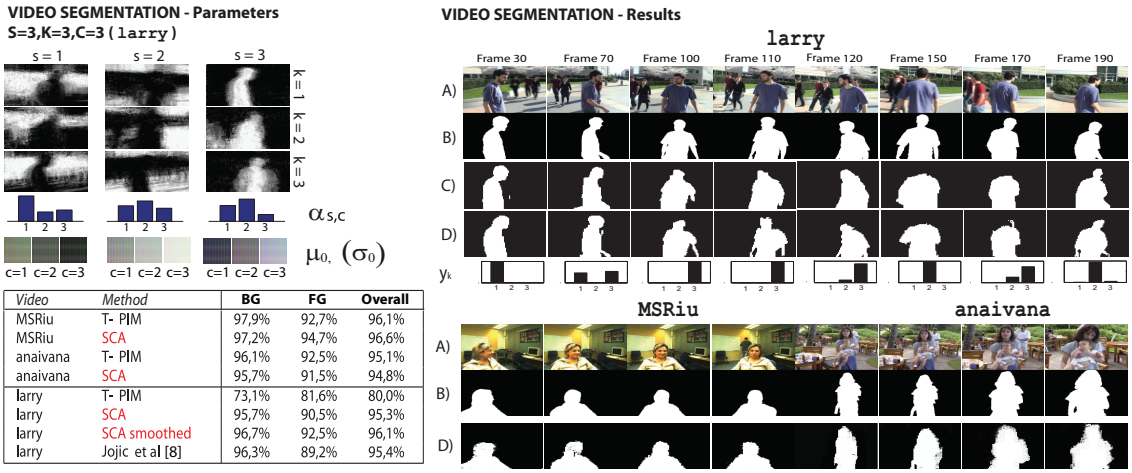


Figure 2. SCA component strengths y_k, K=2, for a set of images of faces with varying pose, have a single degree of freedom ($y_1 + y_2 = 1$) and this degree of freedom captures the pose angle well. Below the y_k strengths, we show images generated from the model using the y_k inferred from the input. The rest of the figure illustrates the PCA reconstruction, which does not manage to separate pose from other causes of variability.

3. Experiments

The main contribution of this paper is a novel representation of images, which allows for unsupervised extraction of image parts, with this process synchronized over many examples of images from a certain class, e.g., a single video clip, or an object category. The representation can be used in a variety of ways in computer vision, often in conjunction with other models, and the purpose of this section is to provide some illustrations.

Head pose angle estimation: Comparison with PCA. In this experiment, we used a database [3] of 250 images of 18 subjects, each captured at 25 head poses (some examples in Fig. 1). The poses in images were manually labeled with the estimated out-of-plane rotation angle (from 0 to $45\,\mathrm{deg}$). In five-fold crossvalidation, we trained both a PCA model and a stel component analyzer (SCA) (K=2, S=7, Gaussian palettes), and chose the optimal predictor of the pose angle based on the component strengths y of PCA and the stel component analyzer. In case of PCA, the predictors we considered used up to the 6 components with highest egienvalues, and, furthermore, to allow for some illumination invariance, we considered sparse variants that also discarded the first, the first two, or the first three components. For both PCA and the stel-based angle prediction, the cross validation included linear regression, robust linear regression, and the nonlinear regression. The SCA outperformed PCA projection as the input to regression in this test, as the average test error for the optimal PCA-based regressor was $9\,\mathrm{deg}$ and the optimal SCA-based regressor had a test error was $8\,\mathrm{deg}$. The standard deviation over the



Video	Method	BG	FG	Overall
MSRiu	T- PIM	97,9%	92,7%	96,1%
MSRiu	SCA	97,2%	94,7%	96,6%
anaivana	T- PIM	96,1%	92,5%	95,1%
anaivana	SCA	95,7%	91,5%	94,8%
larry	T- PIM	73,1%	81,6%	80,0%
larry	SCA	95,7%	90,5%	95,3%
larry	SCA smoothed	96,7%	92,5%	96,1%
larry	Jojic et al [8]	96,3%	89,2%	95,4%

Figure 3. Video segmentation using SCA. The inferred parameters for Larry video are shown on the left. On the right we illustrate the segmentation for three videos `larry`, `MSRiu` and `anaivana`: A) Frames, B) Ground truth mask, C) Segmentation from [8], D) SCA, temporal smoothing of y_k.

folds was twice lower for stel-based method and the difference between methods was statistically significant. As illustrated in Fig. 2, even given a single degree of freedom in the subspace, the stel component analyzer does not use it to model illumination differences, since it is palette-invariant. Rather, the stels capture facial parts of relatively uniform color (an therefore often a uniform surface normal), and the variation in y_1 captures the changes of these parts as they undergo significant geometric changes (Fig. 1C,D). PCA model captures small geometric transformations as well as large illumination changes, but fails to capture significant structural changes, and no single PCA component captures the majority of the angle variation (the most predictive component had the prediction error of $13\,\mathrm{deg}$ vs SCA's $8\,\mathrm{deg}$), and instead the angle has to be inferred from multiple components, and this result ($9\,\mathrm{deg}$) still lags behind SCA's single component inference.

Unsupervised video segmentation and object tracking. The 220 frames of a video sequence used to test hierarchical model selection idea of [8] contain significant illumination changes, background clutter, significant (and confusable) foreground and background motion, as well as dramatic changes in the size and pose of the foreground object (Fig. 3, larry). To analyze the frames of this video using our model, as local pixel measurements we used two modalities: real valued color and optical flow for each pixel, and the model complexity was S=3, K=3,C=3. The comparison was made based on a manual segmentation of every 10th frame into foreground (FG) and background (BG). The parsing of our model agreed with this 'ground truth' in 95% of pixels, comparable with the algorithm of [8], which is based on a much more complex hierarchical model with multiple components specialized for video processing, including also the LOCUS model which on its own significantly underperformed the full mix. If temporal correlations among components y_k are modeled using a simple Brownian motion model SCA achieves the accuracy of 96%, outperforming the state of the art [8], but given that errors in images are highly correlated, and that our ground truth may not be perfect, the difference is not statistically significant.

We have also compared SCA model' ability to deal with misalignments of the object in video frames, and thus track it, with the ability of the transformed PIM model (T-PIM) [1] to do the same. The latter approach is much more computationally intensive, as it requires a search over many possible image transformations. Even when this search is sped up in case of image translations by reducing many operations to efficient convolution computations, the computational burden of T-PIM is significantly higher than that of SCA, whose computational cost grows only linearly with the number of components K, and typically only a handful of components is used to vary a wide variety of geometric changes in stels. In addition to 'Larry' video, the two approaches were compared on two other video sequences shown in the figure. In all three cases, the foreground segmentation using SCA was at least as good as the one achieved by a more expensive search over alignments performed by T-PIM (For T-PIM, we considered three different scales and nine possible shifts, making the algorithm 27 times slower than the basic PIM, and around 9 times slower than SCA).

Object category modeling. We have trained a variety of SCA models ($S \in \{1, 5\}$, $K \in \{1, 2, 3\}$), for sixteen image categories from various datasets[2]. We used three modali-

[2]Eyes, pedestrians, bottles, birds from Labelme; motorbikes, faces, sunflowers, schooner, airplanes from Caltech256; doors, clouds, sheep,

Table 1. The first table shows 16-class recognition results. The same local features were used in all models. SCA results are highlighted in yellow and bag of feature models in green. In the middle, a comparison of inferred stels as shape features with other discriminative features is provided for the entire Caltech101. The bottom table compares SCA with state of the art in pedestrian detection.

Algorithm	Number of stels				
	S=1	**S=2**	**S = 3**	**S = 4**	**S = 5**
SCA - K=1	-	-	58,3%	59,6%	59,0%
SCA - K=2	-	-	66,9 %	68,2%	70,9%
SCA - K=3	-	-	71,3%	74,7%	80,1%
Gen. bag-of-words	19,3%	44,8%	45,1%	44,7%	54,2%
LDA [6]	19,8%	-	59,1%	-	-
SVM	43,2%	-	65,8%	-	-
Spatial Weighting [11]	43,2%	60,1%	61,3%	-	-

Features comparison on complete Caltech-101 dataset							
Shape GB 1	Shape GB 2	Self Sim.	Shape 180	Shape 360	App. Col.	App. Gray	SCA $q(s)$
57%	59%	55%	48%	50%	40%	52%	51%

Features comparison (AUCs) on Pedestrian dataset							
Haar [13]	LRF [13]	MPIM K=2	MPIM K=3	SCA K=2	SCA K=3	PIM	SCA
0.90	0.94	0.91	0.91	0.92	0.93	0.961	0.973

ties for characterizing local measurements $z_{\mathbf{i}}$: real-valued color, and two discrete texture measurements – SIFT measurements quantized into 200 discrete features, and texton features obtained by clustering outputs of a bank of 48 filters into 200 discrete features (see Fig. 4). The images, approximately 100 per class, were split into 70/30 training/test partitions. In addition to comparing the full model's ability to recognize image categories with several state-of-the-art algorithms [6, 11][3], we aim to illustrate the value of spatial parsing of the categories into stels in particular. Therefore, for a variety of models recently used for object modeling without any spatial structure, we show the results obtainable by utilizing the inferred stel segmentation to re-learn separate models in meaningful image parts. For example, a generative bag-of-words model can be applied to the entire image (S=1), or two three parts inferred by SCA with S=3. The inferred posterior distribution $q(s_i^t)$ for each image was used to define the parts in different images, and then three bag-of-words models were estimated, each only from the features found in the appropriate part across different images. Classification was then based on the product of three part likelihoods over features found in a test image tessellated into parts by SCA inference. For S=3, this yields a boost in recognition from 19.3% to 45.1%. Furthermore, the performance increases with the increase in the number of stels, and the classification rate grows to 54.2% for S=5. This kind of classification is equivalent to keeping only the first term of the SCA free energy (5) which deals with matching the distribution over features in different parts to appropriate Dirichlet priors over the entire class. Other popular methods also receive a boost from stel segmentation (Table 1, Top), which could be understood from the example in Fig. 1F - the image segmentation allows the models to expect the sky color at the top of the image, edge features on the chimneys, and roof texture in the roof stel. The best results are obtained for a full SCA model, which optimally matches the features with the stel tessellation, but more importantly, through the last two terms in SCA free energy (5) also punishes large deviations of the inferred tessellation $q(s)$ in a single image from the prior $p(s)$, defined as a mix over components $r_k(s)$ and learned over many images (see Fig. 4). Matching posterior and prior distribution over stels in the free energy (or likelihood) is a probabilistic version of shape matching, which resurrects the question of how valuable the shape features are for object modeling in comparison to texture features. The latter have been recently assumed to be the most informative, with spatial configurations, global shape descriptors, etc, being also helpful but less so.

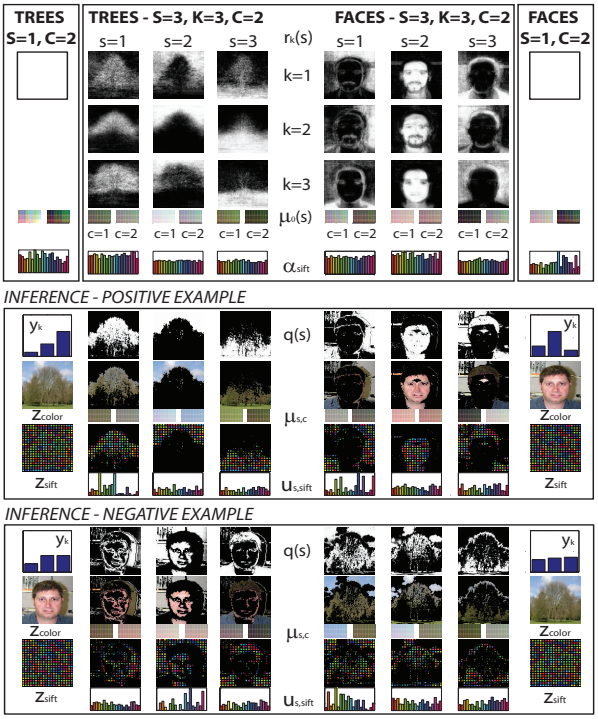


Figure 4. Top: Model parameters for the face and the tree categories with S=4, K=3 are contrasted with the bag-of-words model (S=1). The learned Dirichlet priors for color, and SIFT (α_{sift}) are illustrated with bars whose height is proportional to the strength α_j of different words ζ_j. The rich histogram priors for the standard bag-of-words model (special case of our model with $s = 1$) are broken into tighter priors for appropriately estimated stels $s = 1, ..., 5$. Bottom: Inferred hidden variables under the learned models for two images.

chimneys, car side, car rear and trees from MSR-23 categories.

[3]Background/Foreground segmentation has been obtained using SCA (S=2) in an unsupervised way, while [11] requires the segmentations of the training data.

Classifying Caltech101 categories without local features. To investigate this further, we followed the same training and test recipe, but on all 101 categories from Caltech101, for which the best features, used discriminatively provide classification rates of 40-59% as shown in the Tab.1 (The numbers are due to [9]: GB features correspond to geometric blur [12] which captures some of the spatial configuration in feature distributions, and App. Color and Gray are SIFT features [7] calculated from color and gray images, the rest of the features capture gradient orientations, and thus mostly local shape features). In analyzing Caltech101 images, we only used color as a local measurement, to perform inference within a single image, but we performed classification using only the inferred stel segmentation, $q(s)$, without parts of the likelihood that have to do with matching of image measurements to those expected for the category. This corresponds to dropping out the first two terms from the free energy (5) which deal with evaluating the uniformity of observed features z_i and their agreement with the prior over the entire class defined by Λ. Therefore, the only terms kept are the last two terms concerned with the KL distance between the prior $p(s)$ and the inferred stel tesselation for the image $q(s)$. Such classification yields accuracy of 25%. However, the discriminative use of inferred stels, through SVM classification using only inferred stels as features resulted in classification accuracy of 51%, making the global shape features defined by stel segmentation of comparable quality to the top features used in object classification. This is encouraging, as these features capture rather different aspects of images and could thus be used in multi-feature approaches which previously yielded best results on this dataset [9]. It is important to note here that for an already trained SCA model, inference of stels $q(s)$ for any new given image consists of only a 4-5 iterations of Eq. (7)-(9), as the SCA components r_k, and palette priors are linked to the entire category, not a single image. Thus, inference for a single image is linear in the number of pixels, and is in fact more computationally efficient than the computation involved in methods that require a large number of filter banks or SIFT extraction, which SCA does not require when the considered image measurement is just color.

The SCA model also outperforms state of the art in pedestrian detection (Table 1, bottom).

4. Conclusions

We have introduced a novel model that captures the correlations in spatial structure of an image class. Instead of relying on consistency of image features across images from the same class, the model mines self similarity patterns within individual images. Inference in this model leads to consistent segmentation of images into structural elements (stels), shared across the entire class, even when the images differ dramatically in their local colors and features. Significant variation in stels can be tolerated by a subspace model, stel component analyzer, which captures correlated changes in image structure and thus avoids over-generalization that the PIM model was prone to when faced with significant structural variation. The model can be inferred from the data in an unsupervised manner and this affords this representation of images significant advantages in a variety of computer vision applications, some of which have been illustrated above. In addition, the analysis described here can help scale up the approaches to computer vision which depend on manual segmentation of image parts. Recent sociological innovations have made it possible to recruit a large number of volunteers to manually segment images through collaborative games, or by motivating them by other types of incentives. However, even in these cases, the SCA model may prove to be an invaluable tool for refinement of user-provided segmentations, such as the ones obtainable from LabelMe database.

References

[1] N. Jojic and C. Caspi, "Capturing image structure with probabilistic index maps," CVPR 2004, pp. 212-219.

[2] J. Winn and N. Jojic, "LOCUS: Learning Object Classes with Unsupervised Segmentation," ICCV 2005, pp. 756-763.

[3] D. Graham, N. Allinson. "Characterizing Virtual Eigensignatures for General Purpose Face Recognition," (in) Face Recognition: From Theory to Applications

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

[5] D.M. Blei, A.Y. Ng, M.I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., 2003

[6] L. Fei-Fei, P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," IEEE CVPR 2005.

[7] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," IJCV, 2004

[8] N. Jojic, J. Winn, L. Zitnick, "Escaping local minima through hierarchical model selection: Automatic object discovery, segmentation, and tracking in video," IEEE CVPR 2006

[9] M. Varma, D. Ray "Learning The Discriminative Power-Invariance Trade-Off" ICCV 2007

[10] J. Shotton, J. Winn, C. Rother, A. Criminisi "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation" ECCV 2006

[11] M. Marszaek, C. Schmid "Spatial Weighting for Bag-of-Features" IEEE CVPR 2006

[12] A.C. Berg, J. Malik "Geometric Blur for Template Matching" IEEE CVPR 2001

[13] S. Munder and D. M. Gavrila "An experimental study on pedestrian classification" *IEEE TPAMI*, 2006