Spatialized Epitome and Its Applications

Xinqi Chu[1,4], Shuicheng Yan[2], Liyuan Li[1], Kap Luk Chan[3], Thomas S. Huang[4]
[1]Institute for Infocomm Research, [2]Department of ECE, National University of Singapore
[3] Nanyang Technological University, Singapore,
[4]ECE Department, University of Illinois at Urbana-Champaign

Abstract

Due to the lack of explicit spatial consideration, existing epitome model may fail for image recognition and target detection, which directly motivates us to propose the so-called spatialized epitome in this paper. Extended from the original graphical model of epitome, the spatialized epitome provides a general framework to integrate both appearance and spatial arrangement of patches in the image to achieve a more precise likelihood representation for image(s) and eliminate ambiguities in image reconstruction and recognition. From the extended graphical model of epitome, an EM learning procedure is derived under the framework of variational approximation. The learning procedure can generate an optimized summary of the image appearance with spatial distribution of the similar patches. From the spatialized epitome, we present a principled way of inferring the probability of a new input image under the learnt model and thereby enabling image recognition and target detection. We show how the incorporation of spatial information enhances the epitome's ability for discrimination on several vision tasks, e.g.*, misalignment/cross-pose face recognition and vehicle detection with a few training samples.*

1. Introduction

Recently, *epitome* has been successfully applied in computer vision as a patch-based generative model of image(s) or video [3, 7]. As a maximum likelihood representation for image data, it can be considered as a trade-off representation in-between template and histogram. The balance between visual resemblance and generalization of image and video can be adjusted by the sizes of epitome and patch. It has attracted more and more attention in computer vision due to its impressive abilities in many vision tasks.

The "epitomes" were first introduced as simple appearance and shape models in [7]. These models are learned by compiling patches drawn from input images into a condensed image model. It was shown in [11] that the image epitome is an image summary of high "completeness". The

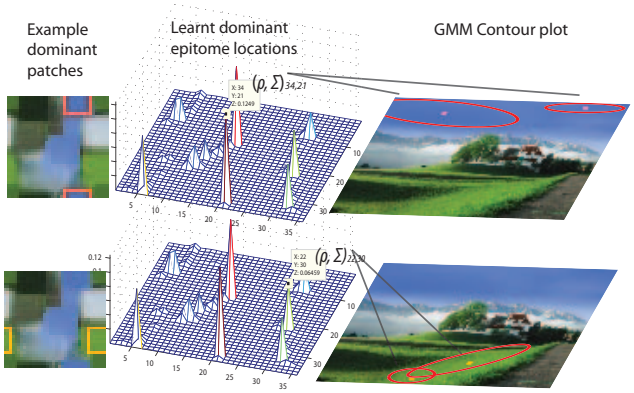

Figure 1. A 36×36 spatialized epitome (in the first column) is learnt from the image in the third column. The distribution in the middle column shows the positions of the significant patches. Note that most locations are of zero value due to regularization. The leftmost image in each row highlights a significant patch in the spatialized epitome. Its associated Gaussian Mixture which represents the spatial arrangement of the significant patch in the input image is shown as ellipse contours in the third column.

epitome idea has also found its use in representing audio information [8] and human activities [5]. Jigsaw proposed in [1] took the epitome beyond square patches and modeled local spatial coherence. The epitome model was also extended to location recognition [9], where it uses each of the entire input image as a patch in which the mappings are fixed during learning and inference. The image frames from a panoramic video are automatically stitched together to form a panorama due to epitome's ability in exploring image similarities [11]. Most recently, epitome priors are investigated for image parsing in which non-overlapping patches are associated with labels of object classes [14].

Under the generative model framework, the learnt epitome is a condensation of image patches, which are however not able to regenerate a meaningful image without guidance by an input image to give a meaningful spatial layout. The input image serves as a location map during the learning and inference process. Since the expected mapping posteriors are only estimated from patch-similarity measurements in inference, it will often cause ambiguities in reconstruc-

tion and recognition during the inference process due to the lack of spatial constraints. For example, epitome was used to recover the occluded part of the object in a video by replacing the occlusion with the patches learnt from the nearby images without occlusions. However, the conventional Epitome model can only assign a patch in the model to a patch in the image according to the patch-wise similarity of intensity. When the occluded area contains patches that are of different appearance from nearby patches in the image, the model would generally fail to assign the correct patch to replace the occlusion, see Figure 3. Therefore, the epitome might not be applicable for recognition/detection tasks because of this ambiguity caused by the lack of information about where the patches come from and how similar-patches are distributed on the input images. In [4], a few pairs of long-range patches are randomly selected for each patch for spatial constraints in image reconstruction. Such pairs represent a few specific spatial correlations. They cannot model the general spatial distributions of similar patches, and, in worse cases, may capture false correlation between two long-range patches, *e.g*. the foreground patch with background patch. As for re-building from compressed image, Wang *et al*. [12] proposed to record the fixed mapping to copy the patches from the epitome to the image locations. The flexibility and optimality of image summarization and inference by generative model are lost in such a hard-coding approach.

Motivated by the above observations, we propose a new graphical model of epitome to integrate information about the appearance summary and spatial arrangement of patches in the image(s). A set of Gaussian Mixtures is introduced into the original graphical model of epitome to relate the appearance and shape with their spatial arrangements on the input images, see Figure 1 for illustration. In this way, the model is self-contained with appearance, shape, as well as patch spatial distribution in input images. So by sampling the learnt model itself, the spatialized epitome is capable of synthesizing the scenes and objects it "saw" during training (See Section 4.1). With spatial constraints included in the epitome model, the misalignment problem with various variations can be solved automatically because the proposed model allows the patches to organize adaptively during inference. To evaluate on a few tough vision tasks, we investigate to apply the proposed spatialized epitome for misaligned face recognition and cross-pose face recognition which means to recognize people with poses unseen in the training set. The main contributions of this paper can be summarized as follows:

1. An improved epitome model which combines the patch appearance information with its associated spatial distribution.
2. An EM procedure to learn an optimized appearance summary and the spatial distributions of image patches.
3. An inference procedure for spatialized epitome.
4. Investigation on applying the spatialized epitome for a few vision tasks.

The rest of this paper is structured as follows: In Section 2, we present the spatialized epitome model and the derivation of the learning procedure. Inference process for recognition and detection is presented in Section 3. Experiments, including the comparisons with the original epitome, on face recognition with misalignments, cross-pose face recognition, and car detection are presented in Section 4. The paper is concluded in Section 5 and limitations are discussed.

2. Learning a Spatialized Epitome

An image does not merely consist of patches, and it is also about how the patches are spatially arranged. In existing epitome [7, 4], for each patch $\mathbf{Z}_k$, the likelihood probability was calculated by an intensity similarity. Therefore, the process of inference and reconstruction on an input image is purely guided by intensity-similarity measure with respect to the training images regardless of how patches are arranged in the training or probe image. We show the problem of this under-constrained process in Figure 3.

Here we present a generative model combining both patch appearances and arrangements in an image or a collection of images. Suppose P patches are sampled from M images, denote each patch as $\mathbf{Z}_k$. The corresponding mapping random variable is denoted as $\mathcal{T}_k$, which is hidden and unknown. The patch is sampled from the position $\mathbf{y}_k$ in the original image, so $\mathbf{y}_k$ is observed. For each patch in the epitome, we use a Gaussian Mixture Models (GMM) to model the image locations from which the patches are originated. If the size of the epitome is a, then we have $a \times R$ such GMMs. C_k is a R-dimensional binary random variable in which a particular element C_{kr} is equal to 1 and all other elements are equal to 0 when the component r is active. For each observed location $\mathbf{y}_k$, there is a corresponding latent variable C_k. We now define the generative process:

1. Choose a position in the epitome, $\mathcal{T}_k \sim Cat(\boldsymbol{\pi})$;
2. For each of the chosen position $\mathcal{T}_k$,
   (a) Choose a patch $\mathbf{Z}_k$ from $p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e})$;
   (b) Choose a component C_k from the GMMs for the given location $\mathcal{T}_k$: $C_k \sim p(C_k|\mathcal{T}_k)$;
   (c) Choose a coordinate $\mathbf{y}_k$ from the component C_k for patch $\mathbf{Z}_k$: $\mathbf{y}_k \sim p(\mathbf{y}_k|\mathcal{T}_k, C_k)$.

This process is illustrated in Figure 2. The generation of each patch (intensity) is formulated as:

$$P(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) = \prod_{i \in S_k} \mathcal{N}(z_{i,k}; \mu_{\mathcal{T}_k(i)}, \phi_{\mathcal{T}_k(i)}) \qquad (1)$$

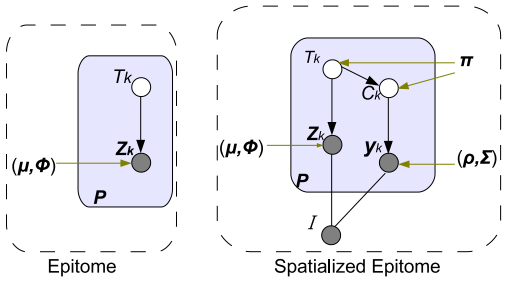

Figure 2. The graphical model representations of the epitome and the spatialized epitome. The boxes are "plates" representing replicates.

where S_k is the set of the coordinates of all pixels in the patch $\mathbf{Z}_k$. The generation of the coordinate of each patch is formulated as:

$$P(\mathbf{y}_k|\mathcal{T}_k, C_{kr}=1) = \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \boldsymbol{\Sigma}^r_{\mathcal{T}_k=e}), \quad (2)$$

where e represents the location in the epitome that the patch maps to, and the superscript r indicates the rth component of the GMM. Write it in a compact distribution form:

$$p(\mathbf{y}_k|\mathcal{T}_k, C_k) = \prod_{r=1}^{R} \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \boldsymbol{\Sigma}^r_{\mathcal{T}_k=e})^{C_{kr}}, \quad (3)$$

Given the mapping $\mathcal{T}_k$ of the patch $\mathbf{Z}_k$, there are several Gaussian components in the location $\mathcal{T}_k = e$ to choose from, where e denotes a particular location in the epitome. The probability distribution of choosing each Gaussian component given the location e is

$$p(C_k|\mathcal{T}_k) = \prod_{r=1}^{R} \tilde{\pi}_{\mathcal{T}_k=e,r}^{C_{\mathcal{T}_k=e,r}}. \quad (4)$$

Since $p(C_k, \mathcal{T}_k) = p(C_k|\mathcal{T}_k)p(\mathcal{T}_k)$ and the prior on both parameters shall be learnt, we use the joint distribution of C_k and $\mathcal{T}_k$ to perform parameter estimation on the mixing coefficients.

2.1. Learning procedure for spatialized epitome

For the P patches generated independently, we have the joint distribution:

$$p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^P, \mathbf{e}, \boldsymbol{\pi}) = p(\mathbf{e}, \boldsymbol{\pi}) \prod_{k=1}^{P} p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) p(\mathbf{y}_k|\mathcal{T}_k, C_k) p(C_k, \mathcal{T}_k), \quad (5)$$

where $\boldsymbol{\pi}$ are the parameters of the mixing proportions on $\mathcal{T}_k$ and C_k. Since we cannot observe C_k and $\mathcal{T}_k$, we sum over all possible values that they might be taking, and

$$\log P(\{\mathbf{Z}_k, \mathbf{y}_k\}_{k=1}^P) = \log \sum_{\{C_k, \mathcal{T}_k\}} \int_{\mathbf{e}, \boldsymbol{\pi}} p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^P, \mathbf{e}, \boldsymbol{\pi}) d(\mathbf{e}, \boldsymbol{\pi}) = \log \sum_{\{C_k, \mathcal{T}_k\}} \prod_{k=1}^{P} p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) p(\mathbf{y}_k|\mathcal{T}_k, C_k) p(C_k, \mathcal{T}_k)$$

Now we first assume that the prior on the parameters are flat. We use variational approximation to put the $\log$ inside the $\sum$ for tractable optimization, the auxiliary distribution $q(\{\mathcal{T}_k, C_k\}_{k=1}^P)$ is put into the likelihood of data and then use the Jensen's Inequality [2]:

$$\begin{aligned}
&\log P(\{\mathbf{Z}_k, \mathbf{y}_k\}_{k=1}^P) = \\
&\log \sum_{\{C_k, \mathcal{T}_k\}} \frac{q(\{\mathcal{T}_k, C_k\}_{k=1}^P) p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^P)}{q(\{\mathcal{T}_k, C_k\}_{k=1}^P)} \\
&\geq \sum_{\{C_k, \mathcal{T}_k\}} q(\{\mathcal{T}_k, C_k\}_{k=1}^P) \log \frac{p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^P)}{q(\{\mathcal{T}_k, C_k\}_{k=1}^P)} \\
&= \sum_{\{C_k, \mathcal{T}_k\}} q(\{\mathcal{T}_k, C_k\}_{k=1}^P) \log p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^P) \\
&- \sum_{\{C_k, \mathcal{T}_k\}} q(\{\mathcal{T}_k, C_k\}_{k=1}^P) \log q(\{\mathcal{T}_k, C_k\}_{k=1}^P) = B. \quad (6)
\end{aligned}$$

Since $q(\{\mathcal{T}_k, C_k\}_{k=1}^P) = \prod_{k=1}^P q(\mathcal{T}_k, C_k)$ due to the independence assumption by variational mean field theory [2], we have

$$\begin{aligned}
&\log P(\{\mathbf{Z}_k, \mathbf{y}_k\}_{k=1}^P) \geq B = \\
&\sum_{\{C_k, \mathcal{T}_k\}} \prod_{k=1}^{P} q(\mathcal{T}_k, C_k) \log \prod_{k=1}^{P} p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) p(\mathbf{y}_k|\mathcal{T}_k, C_k) p(C_k, \mathcal{T}_k) \\
&\quad - \sum_{\{C_k, \mathcal{T}_k\}} q(\{\mathcal{T}_k, C_k\}_{k=1}^P) \log q(\{\mathcal{T}_k, C_k\}_{k=1}^P) \\
&= \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) [\log p(\mathcal{T}_k, C_k) + \\
&\quad \log p(\mathbf{y}_k|\mathcal{T}_k, C_k) + \log p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}})] - E. \quad (7)
\end{aligned}$$

When $q(\mathcal{T}_k, C_k) = p(\mathcal{T}_k, C_k|\mathbf{Z}_k, \mathbf{y}_k, \hat{\mathbf{e}})$, the lower bound is tight and the entropy $E = 0$ which can be proved by substituting the posterior into the bound. Note that here we can update $p(C_k, \mathcal{T}_k)$, $p(\mathbf{y}_k|\mathcal{T}_k, C_k)$, $p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}})$ independently. By iteratively optimizing the bound B, we can derive an EM procedure to learn the spatialized epitome.

The E-Step: By setting the auxiliary distribution to be the

posterior of hidden variables, there is

$$q(\mathcal{T}_k, C_k) = p(\mathcal{T}_k, C_k|\mathbf{Z}_k, \mathbf{y}_k, \hat{\mathbf{e}}) = \frac{p(\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k, \hat{\mathbf{e}})}{p(\mathbf{Z}_k, \mathbf{y}_k, \hat{\mathbf{e}})}$$
$$= \frac{p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}})p(\mathbf{y}_k|\mathcal{T}_k, C_k)p(C_k, \mathcal{T}_k)}{p(\mathbf{Z}_k, \mathbf{y}_k, \hat{\mathbf{e}})}$$
$$\sim p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}})p(\mathbf{y}_k|\mathcal{T}_k, C_k)p(C_k, \mathcal{T}_k)$$
$$= \prod_{i\in S_k} \mathcal{N}(z_{i,k}; \mu_{\mathcal{T}_k(i)}, \phi_{\mathcal{T}_k(i)}) \prod_{r=1}^{R} \mathcal{N}(\mathbf{y}_k; \rho^r_{\mathcal{T}_k=e}, \mathbf{\Sigma}^r_{\mathcal{T}_k=e})^{C_{kr}} p(C_k, \mathcal{T}_k). \quad (8)$$

The M-Step: Note the equal sign indicates that the bound is tight at this moment, the bound B can be separated into three parts: $B = B_1 + B_2 + B_3$, where B_1 is related to the epitome appearance, B_2 is related to spatial distributions, and B_3 is related to mixing weights. Hence, we can derive the update rules for the three sets of parameters separately.

a) Updating the appearance

Only the term B_1 in B relates to the epitome appearance $\hat{\mathbf{e}}$. Let us denote the estimated distribution $q(\mathcal{T}_k, C_k)$ as q_k for simplicity. B_1 can be expressed as

$$B_1 = \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} q_k \log p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}}) =$$
$$= \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i\in S_k} q_k \left[-\frac{1}{2} \log 2\pi\phi_j - \frac{(z_{i,k} - \mu_j)^2}{2\phi_j} \right]. \quad (9)$$

Finding the solution for $\partial B_1/\partial \hat{\mathbf{e}} = 0$ is equivalent to finding the solutions for $\frac{\partial B_1}{\partial \mu_j} = 0$ and $\frac{\partial B_1}{\partial \phi_j} = 0$, respectively. Hence, the updating rule for μ_j can be obtained as:

$$\mu_j = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i\in S_k} q(\mathcal{T}_k, C_k) z_{i,k}}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i\in S_k} q(\mathcal{T}_k, C_k)}, \quad (10)$$

and the corresponding updating rule for ϕ_j is:

$$\phi_j = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i\in S_k} q(\mathcal{T}_k, C_k)(z_{i,k} - \mu_j)^2}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i\in S_k} q(\mathcal{T}_k, C_k)}. \quad (11)$$

This is similar to the original epitome updating rules.

b) Update GMM Means and Covariances

From Eq. (7), the bound for the GMM term is simplified as:

$$B_2 = \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) \log p(\mathbf{y}_k|\mathcal{T}_k, C_k) =$$
$$= \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) \sum_{r=1}^{R} C_{kr} \log \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \mathbf{\Sigma}^r_{\mathcal{T}_k=e}). \quad (12)$$

Set the derivative w.r.t $\boldsymbol{\rho}^r_{\mathcal{T}_k=e}$ to be 0, *i.e.* $\frac{\partial B_2}{\partial \boldsymbol{\rho}^r_e} = 0$, there is

$$\frac{\partial}{\partial \boldsymbol{\rho}^r_e} \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) \sum_{r=1}^{R} C_{kr} \log \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \mathbf{\Sigma}^r_{\mathcal{T}_k=e})$$
$$= \frac{\partial}{\partial \boldsymbol{\rho}^r_e} \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} \log \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \mathbf{\Sigma}^r_{\mathcal{T}_k=e})$$
$$= \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} (\mathbf{y}_k - \boldsymbol{\rho}^r_e)^T (\mathbf{\Sigma}^r_e)^{-1} = 0. \quad (13)$$

From the above equation, we can obtain the updating rule for $\boldsymbol{\rho}^r_e$ as:

$$(\boldsymbol{\rho}^r_e)^T = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr} \mathbf{y}_k^T}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr}}. \quad (14)$$

Applying the same deduction for the GMM mean, we take derivative w.r.t $(\mathbf{\Sigma}^r_e)^{-1}$ and set to be 0:

$$\frac{\partial}{\partial (\mathbf{\Sigma}^r_e)^{-1}} \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} \log \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \mathbf{\Sigma}^r_{\mathcal{T}_k=e})$$
$$= \frac{\partial}{\partial (\mathbf{\Sigma}^r_e)^{-1}} \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} [-\log 2\pi - \frac{1}{2} \log |\mathbf{\Sigma}^r_e| - \frac{1}{2} (\mathbf{y}_k - \boldsymbol{\rho}^r_e)^T (\mathbf{\Sigma}^r_e)^{-1} (\mathbf{y}_k - \boldsymbol{\rho}^r_e)]$$
$$= \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} [+\frac{1}{2} \mathbf{\Sigma}^r_e - \frac{1}{2} (\mathbf{y}_k - \boldsymbol{\rho}^r_e)^T (\mathbf{y}_k - \boldsymbol{\rho}^r_e)] = 0. \quad (15)$$

Therefore we obtain the updating rule for $\mathbf{\Sigma}^r_e$ as,

$$\mathbf{\Sigma}^r_e = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr} (\mathbf{y}_k - \rho^r_e)(\mathbf{y}_k - \boldsymbol{\rho}^r_e)^T}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr}}. \quad (16)$$

c) Update mixing coefficients

From (7), the term related to mixing coefficients can be expressed:

$$B_3 = \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) \log p(\mathcal{T}_k, C_k). \quad (17)$$

Denoting $p(\mathcal{T}_k = e, C_k = r) = \pi_{er}$, we can maximize the bound B_3 subject to $\sum_{e,r} p(\mathcal{T}_k = e, C_k = r) = 1$ as:

$$\frac{\partial}{\partial \pi_{er}} (B_3 + \lambda (\sum_{e,r} \pi_{er} - 1))$$
$$= \frac{\partial}{\partial \pi_{er}} \sum_{k=1}^{P} \sum_{C_k=r, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) \log p(\mathcal{T}_k = e, C_k = r) + \lambda$$
$$= \sum_{k=1}^{P} q(\mathcal{T}_k = e, C_k = r) \frac{1}{\pi_{er}} + \lambda = 0. \quad (18)$$

Table 1. The number of parameters for spatialized epitome model.

Epitome($\hat{\mathbf{e}}$)	Gaussians($\boldsymbol{\rho}, \boldsymbol{\Sigma}$)	Mixing Coefficients ($\boldsymbol{\pi}$)
$N \times N \times 2$	$N \times N \times 2$	$N \times N \times R$

Then, we can obtain $\lambda = -P$ and the updating rule of the mixing coefficient as,

$$\pi_{er} = \frac{\sum_{k=1}^{P} q(\mathcal{T}_k = e, C_k = r)}{P}. \quad (19)$$

2.2. Bayesian regularization and priors

Suppose we have R Gaussian components at one epitome location e. The number of parameters for our epitome with a size of $N \times N$ is $N^2 \times (R+4)$. The details are listed in Table 1. Since we have a finite training set and a relatively large set of parameters, in order to avoid overfitting, on each location in the epitome we put a Dirichlet-Normal-Wishart prior on the three sets of parameters $\{\boldsymbol{\rho}_e^r, \boldsymbol{\Sigma}_e^r\}_{r=1}^R$ and $\boldsymbol{\pi}_e$, *i.e.*

$$p(\{\boldsymbol{\rho}_e^r, \boldsymbol{\Sigma}_e^r\}_{r=1}^R, \boldsymbol{\pi}_e) = b(\boldsymbol{\gamma}_e) \prod_{r=1}^{R} (\pi_e^r)^{\gamma_e^r - 1} \prod_{r=1}^{R} \mathcal{N}\left(\boldsymbol{\rho}_e^r | \boldsymbol{\nu}_e^r, \frac{\boldsymbol{\Sigma}_e^r}{\eta_e^r}\right) Wi((\boldsymbol{\Sigma}_e^r)^{-1} | \boldsymbol{\beta}_e^r, \tau_e^r), \quad (20)$$

where $b(\boldsymbol{\gamma}_e)$ is the normalizing factor of the Dirichlet distribution and $Wi(.|)$ denotes a Wishart distribution. By determining appropriate values for the hyper-parameters $\{\gamma_e^r, \boldsymbol{\nu}_e^r, \boldsymbol{\Sigma}_e^r, \eta_e^r, \boldsymbol{\beta}_e^r, \tau_e^r\}$ we state our beliefs about the data generation process in terms of a prior distribution. The use of such prior is justified in [10]. By incorporating the prior, the updating rules are derived to be:

$$(\boldsymbol{\rho}_e^r)^T = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k = e} q(\mathcal{T}_k, C_k) C_{kr} \mathbf{y}_k^T + \eta_e^r \boldsymbol{\nu}_e^r}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k = e} q(\mathcal{T}_k, C_k) C_{kr} + \eta_e^r}; \quad (21)$$

$$\boldsymbol{\Sigma}_e^r = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k = e} q(\mathcal{T}_k, C_k) C_{kr} (\mathbf{y}_k - \rho_e^r)(\mathbf{y}_k - \rho_e^r)^T}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k = e} q(\mathcal{T}_k, C_k) C_{kr} + 2\tau_e^r - 2} + \frac{\eta_e^r (\boldsymbol{\mu}_e^r - \boldsymbol{\nu}_e^r)(\boldsymbol{\mu}_e^r - \boldsymbol{\nu}_e^r)^T + 2\boldsymbol{\beta}_e^r}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k = e} q(\mathcal{T}_k, C_k) C_{kr} + 2\tau_e^r - 2}; \quad (22)$$

$$\pi_{er} = \frac{\sum_{k=1}^{P} q(\mathcal{T}_k = e, C_k = r) + \gamma_e^r - 1}{P + \sum_{r=1}^{R} \gamma_e^r - R}. \quad (23)$$

The prior penalizes singularities in the log-likelihood function in the case when an epitome patch has only one corresponding patch in the image(s). We also encode our prior belief that the covariance matrices of GMMs are diagonal with diagonal values to be the width of the training image. We adjust the strength of the prior by modifying γ, β and τ which are functions of the equivalent sample size in Bayesian terms.A sparsity inducing prior (Dirichlet) with $\alpha = 0.05$ is used so that most of the mixing coefficients tend to zero and the corresponding Gaussian components will not contribute in modeling the distributions, as shown in Figure 1.

3. Inference Based on Spatialized Epitome

We denote the set of learnt parameters $\{\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\Sigma}}, \hat{\mathbf{e}}, \hat{\boldsymbol{\pi}}\}$ of training set $\mathcal{D}$ as $\hat{\boldsymbol{\Theta}}$. Given the data of a training set $\mathcal{D}$, the probability of seeing a given probe image can be directly calculated as:

$$\begin{aligned}
&\log P(I|\mathcal{D}) \simeq \log P(I|\hat{\boldsymbol{\Theta}}) = \log P(I|\boldsymbol{\rho}, \boldsymbol{\Sigma}, \hat{\mathbf{e}}, \boldsymbol{\pi}) \\
&= \log P(\{\mathbf{Z}_k, \mathbf{y}_k\}_{k=1}^P | \boldsymbol{\rho}, \boldsymbol{\Sigma}, \hat{\mathbf{e}}, \boldsymbol{\pi}) \\
&= \log \prod_{k=1}^{P} P(\mathbf{Z}_k, \mathbf{y}_k | \boldsymbol{\rho}, \boldsymbol{\Sigma}, \hat{\mathbf{e}}, \boldsymbol{\pi}) \\
&= \sum_{k=1}^{P} \log \sum_{C_k, \mathcal{T}_k} P(\mathbf{Z}_k, \mathbf{y}_k, C_k, \mathcal{T}_k | \boldsymbol{\rho}, \boldsymbol{\Sigma}, \hat{\mathbf{e}}, \boldsymbol{\pi}) \\
&= \sum_{k=1}^{P} \log \sum_{C_k, \mathcal{T}_k} p(\mathbf{Z}_k | \mathcal{T}_k, \mathbf{e}) p(\mathbf{y}_k | \mathcal{T}_k, C_k) P(C_k, \mathcal{T}_k) \\
&= \sum_{k=1}^{P} \log \sum_{C_k, \mathcal{T}_k} \prod_{i \in S_k} \mathcal{N}(z_{i,k}; \mu_{\mathcal{T}_k(i)}, \phi_{\mathcal{T}_k(i)}) \\
&\qquad \prod_{r=1}^{R} \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}_{\mathcal{T}_k = e}^r, \boldsymbol{\Sigma}_{\mathcal{T}_k = e}^r)^{C_{kr}} P(\mathcal{T}_k, C_k). \quad (24)
\end{aligned}$$

This inference formulation is similar to the way of evaluating the probability value of seeing a new data under a learnt GMM. The first step of this derivation follows [6]. The third step uses the assumption that all the patches are independently sampled. The above calculated probability value indicates how likely the probe image is generated by the learnt model, and can be directly used for image recognition and object detection purposes.

Recognition Suppose there are N epitomes with parameters $\{\boldsymbol{\Theta}_i\}_{i=1}^N$ learnt from N classes of visual objects. Denote the label of the input image to be $\mathcal{C}$ and we assume no prior knowledge on label $\mathcal{C}$, so the recognition is achieved by computing the label posterior $p(\mathcal{C}|I)$ using:

$$p(\mathcal{C}|I) = \frac{p(I|\mathcal{C}) p(\mathcal{C})}{p(I)} \sim p(I|\mathcal{C}), \quad (25)$$

and select the one with the maximum posterior value:

$$\hat{\mathcal{C}} = \arg\max_i P(I|\mathcal{C} = i) = \arg\max_i P(I|\boldsymbol{\Theta}_i), \quad (26)$$

where $P(I|\boldsymbol{\Theta}_i)$ can be calculated from (25) which is in turn calculated by (24).

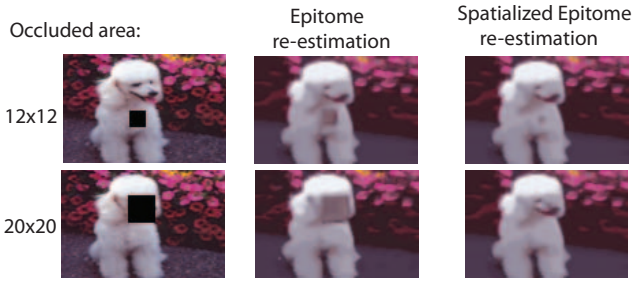

Figure 3. The comparison of image re-estimation results between epitome and spatialized epitome. Both 40×40 epitomes are learnt from the *original image* with patch sizes of 8×8, 4×4, and 2×2 which are also the patch sizes used in the re-estimation process. During the re-estimation process, 40000 patches are uniformly sampled from the input image to ensure that all the coordinates are covered for the re-estimated image. For occlusion in non-uniform image regions *e.g.* the second row, spatialized epitome can also restore the occluded region with proper patches after a number of iterations.

Detection If we scan the input image with multi-scale windows (W), we can perform object detection. In this way, (25) becomes

$$p(\mathcal{C}|W) = \frac{p(W|\mathcal{C})p(\mathcal{C})}{p(W)} \sim p(W|\mathcal{C}), \qquad (27)$$

The mean-shift approach can be used to select local maxima to locate the target objects in the image.

Epitomic re-estimation Using existing epitome for image re-estimation, for each patch $\mathbf{Z}_k$, the inference step evaluates how likely each epitome patch is to generate $\mathbf{Z}_k$. Then the estimation step will replace the initialized values of $\mathbf{Z}_k$ with the average votes from the epitome patches according to $q(\mathcal{T}_k)$. Consequently, the estimated texture will be more consistent with the epitome texture. This is how denoising, video super-resolution and other video repairing applications are achieved. However, the position posterior $q(\mathcal{T}_k)$ is evaluated purely based on the intensity similarity between the epitome patches and the image patches [7, 4]. This may give incorrect estimation when the occluded part has different appearances from nearby patches.

The re-estimation process of spatialized epitome solves this problem as the position posterior $q(\mathcal{T}_k, C_k)$ takes also the spatial arrangement into account as in Eq. (8) in image re-estimation. The comparison of existing epitome and spatialized epitome on image re-estimation from partially occluded image is given in Figure 3.

4. Experiments

In the proposed spatialized epitome, the correlation between the local appearance and spatial arrangement is introduced. This makes it possible to employ epitome for image recognition, object detection, and image re-estimation from partial occlusions. To evaluate the performance of the spatialized epitome, several experiments were conducted, including the comparison with existing epitome on face recognition, and applications to several vision tasks, *e.g.*, face recognition with misalignments, cross-pose face recognition, occlusion detection, and car detection with a few training samples. The details are described below.

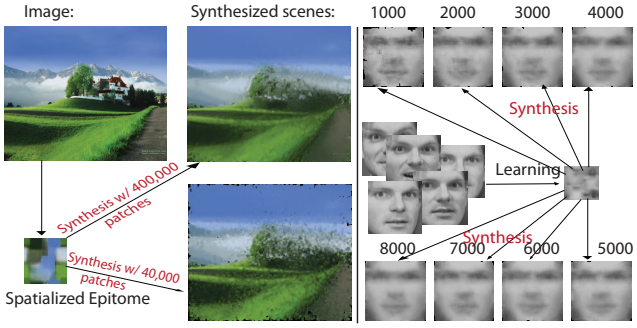

Figure 4. The left half of the figure shows the synthesis results for a spatialized epitome learnt from a scene image. At the right half of the figure, we show synthesis results for a spatialized epitome model learnt from multiple images from the same person.

4.1. Synthesis

Being a self-contained generative model, with both patch intensity and associated spatial distribution, images can be synthesized by ancestral sampling of the proposed model. We show the synthesis results for a scene epitome model (where scene images often consist of large number of redundant patches) as well as for a face epitome model learnt from multiple images of the same person in Figure 4.1.

4.2. Generative face recognition

In this experiment, we evaluate the effectiveness of our spatialized epitome formulation by face recognition. This generative method does not need to go through any feature extraction or dimensionality reduction step but just uses the intensity image as the input and give out the results in probability terms. In order to evaluate the effectiveness of including spatial information, we need to derive a recognition algorithm for the original epitome proposed in [4, 7]. Following the same principle in Section 3, the inferred probability of seeing a new image with original epitome is:

$$\log P(I|\mathcal{D}) \simeq \log P(I|\hat{\mathbf{e}}) = \log P(\{\mathbf{Z}_k\}_{k=1}^{P}|\hat{\mathbf{e}})$$
$$= \sum_{k=1}^{P} \log \sum_{\mathcal{T}_k} \prod_{i \in S_k} \mathcal{N}(z_{i,k}; \mu_{\mathcal{T}_k(i)}, \phi_{\mathcal{T}_k(i)}) P(\mathcal{T}_k). \qquad (28)$$

Table 2. Recognition accuracy rates (%) on two face databases.

Database:	ORL		PIE	
Patch Size:	4×4	6×6	4×4	6×6
Epitome	19.5	27.5	14.7	20.9
Spatialized	76.5	88.5	74.1	78.8

In this experiment, two benchmark face databases, *e.g*. ORL and CMU PIE [1] are used. The ORL database contains 400 images of 40 persons, where each image is manually cropped and normalized to the size of 32×32 pixels. The CMU PIE (Pose, Illumination, and Expression) database contains more than 40,000 facial images of 68 people. In our experiment, a subset of five near frontal poses (C27, C05, C29, C09, and C07) with illumination indexed as 08 and 11 are used. Images with these two indices subject to small illumination variations from one another because our intensity-based model is not illumination invariant. The images are manually normalized to the size of 32×32 with unit norm. Both original and spatialized epitomes are evaluated with two different patch sizes. We can observe from Table 2 that the incorporation of spatial information considerably increases the recognition accuracy. Therefore, the performance of original epitome in later more complex applications are not evaluated.

4.3. Occlusion detection

For a facial image with occlusions, the occluded parts can be revealed by evaluating the likelihood for one patch or a set of few nearby patches by Eq. (24). The set of patch samples with the probabilities lower than a certain threshold are considered to be the patches that are occluded. In this experiment we examine the occlusion detection capability of our spatialized epitome formulation on the ORL database. We randomly pick 5 images of each subject for training, the remaining 5 images of each person serve as probe images. Then an 18×18 artificial occlusion is generated at a random position in each probe image. In this experiment, re-estimation is performed on the detected occlusion area only. Seven images are randomly selected from the probe set and the occlusion detection results are shown in Figure 5, where the 1st row shows the original face images, the 2nd row shows the images with occlusions, the 3rd row shows the detected occlusion regions, and the 4th row shows the reconstructed images by the spatialized epitome of the corresponding person.

4.4. Face recognition with misalignments

In most of the techniques for face recognition, explicit semantics is assumed for each feature. But for computer vision tasks, *e.g*., face recognition, the explicit semantics of the features may be degraded by *spatial misalignments*. face cropping is an inevitable step in an automatic face recognition system, and the success of subspace learning for face recognition relies heavily on the performance of the face detection and face alignment processes. Practical systems or even manual face cropping, may bring considerable image misalignments, including translations, scaling and rotation, which consequently change the semantics of two pixels with the same index but in different images [13]. To a certain extent, the spatialized epitome proposed here can naturally adapt to misaligned inputs because: 1) a moderate amount of coordinate shifts caused by the misalignments can also have a high probability value under a Gaussian mixture distribution as long as the "data point" is still in the vicinity; 2) the spatialized epitome is learnt from patches of images of different expressions (ORL) or different poses (PIE), so the deformation is learnt to account for misalignments on the patch-level; and 3) the misalignment effect is reduced from the image-level to a patch-level. These experiments are also conducted on two benchmark face databases, *e.g*. ORL and PIE with spatial misalignments for the testing data and no misalignments for the training data. A set of 4 images from each subject is used for training while the remaining 6 images of each person are artificially misaligned with a rotation $\alpha \in [-5^\circ, 5^\circ]$, a scaling $s \in [0.95, 1.05]$, a horizontal shift $T_x \in [-1, +1]$ or a vertical shift $T_y \in [-1, +1]$. The value of each of the misalignment factor is drawn from a uniform distribution. In the mixed spatial misalignment configuration, the above mentioned effects are added in a random order to the original test image, and the results are shown in Table 3 with baselines algorithms such as PCA and LDA (the results come from [13] with 4 training samples).

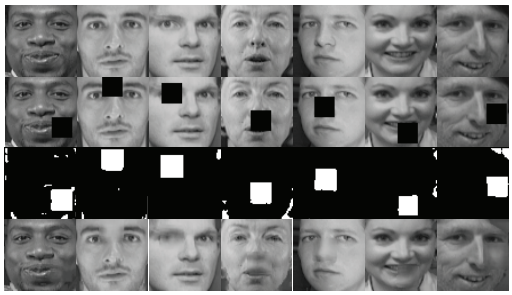

Figure 5. Examples of occlusion detection.

Table 3. Recognition accuracy rates (%) on two databases with mixed misalignments. The patch size of 6×6 is used in both learning and recognition.

Database:	ORL			PIE		
Methods	PCA	LDA	Ours	PCA	LDA	Ours
Results	63.2	51.7	88.0	65.9	54.0	67.9

[1] Available at http://www.face-rec.org/databases/.

Table 4. Cross-pose recognition accuracy rates (%) on PIE database. Each column shows the respective results for each pose. The patch size of 6×6 is used in both learning and recognition.

Methods:	c09	c27	c07	Overall
PCA	34.3	36.1	33.4	34.6
LDA	65.3	66.3	49.1	60.2
Ours	82.4	66.2	72.1	73.6

4.5. Cross-pose face recognition

In the real world scenario, we may often have to recognize a face with a pose that we have not seen before. We show in this experiment that our spatialized epitome can adapt to unknown pose variations to a certain extent. Here we use a different subset of the PIE database. For each subject in the PIE database, 3 images with illumination index 8, 11, 21 from each of the two near frontal poses, namely c05 and c29 are chosen as training set. 3 images from each of the 5 different poses (c09, c27, c07, c37, and c11) for each subject are then selected for testing. In both learning and testing, we use patch size of 6×6. Detailed results and comparison with PCA and LDA (with K-Nearest Neighbour classifier) baselines are listed in Table 4.

4.6. Car detection

In order to show the detection ability of our spatialized epitome, the UIUC side-view car dataset [2] was used for evaluation. Six representative cars are chosen for learning the car model. During learning, we use gradient images which are extracted from the six Gaussian-smoothed positive training images. We slide the window of size 30×90 over the entire query image and calculate the probability value given by Eq. (24). The windows that have probability values above a threshold t are considered to be the locations of the cars. We evaluate performance by comparing the bounding box of detection to the "ground truth" bounding box B_t in manually annotated data. We follow the procedure adopted in the Pascal VOC competition, and compute the area ratio a of $B_p \bigcap B_t$ and $Bp \bigcup Bt$. If $a > 0.5$, then B_p is considered a true positive. By varying the threshold on this confidence, we compute ROC curve as shown in Figure 4.6. Our method achieves reasonable performance under a less restrictive condition which requires a few training samples and no negative training samples are needed. In this case, conventional supervised learning algorithms are not applicable.

5. Conclusions

In this paper, we proposed a new graphical model for epitome, *i.e*. spatialized epitome. It integrates both the local appearance and spatial arrangement for image representation. Experiments on several vision tasks have shown its superiority over the original epitome model. Especially, the tests on misaligned and cross-pose face recognition demonstrates the advantages of the spatialized epitome in adapting to variations in real world conditions. Several limitations on this model can be noticed, as an object model, it is neither scale-invariant nor illumination-invariant. Furthermore, each model instance learnt has considerably more parameters than that of other techniques, especially discriminative ones, *e.g*., a hyperplane. The computational complexity for inference is also quite high as it must go through all possible values of hidden variables for each patch as in Eqn 24.

[2] http://l2r.cs.uiuc.edu/ cogcomp/Data/Car/

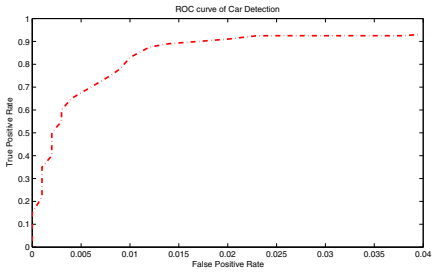

Figure 6. The ROC curve of car detection.

Acknowledgment

Partially supported by NRF/IDM Program, under research Grant NRF2008IDMIDM004-029. Xinqi Chu thanks Junyan Wang for helpful comments on this work.

References

[1] C. R. Anitha Kannan, JohnWinn. Clustering appearance and shape by learning jigsaws. In *NIPS 19*, Cambridge, MA, 2006. MIT Press.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] V. Cheung, B. Frey, and N. Jojic. Video epitomes. In *CVPR*, 2005.

[4] V. Cheung, N. Jojic, and D. Samaras. Capturing long-range correlations with patch models. In *CVPR*, 2007.

[5] N. Cuntoor and R. Chellappa. Epitomic representation of human activities. In *CVPR*, 2007.

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley Interscience, 2 edition, 2001.

[7] N. Jojic, B. Frey, and A. Kannan. Epitomic analysis of appearance and shape. In *ICCV*, 2003.

[8] A. Kapoor and S. Basu. The audio epitome: a new representation for modeling and classifying auditory phenomena. In *ICASSP*, 2005.

[9] K. Ni, A. Kannan, A. Criminisi, and J. Winn. Epitomic location recognition. In *CVPR*, 2008.

[10] D. Ormoneit and V. Tresp. Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates. *T-NN*, 1998.

[11] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *CVPR*, 2008.

[12] H. Wang, Y. Wexler, E. Ofek, and H. Hoppe. Factoring repeating content within and among images. In *ACM SIGGRAPH*, 2008.

[13] H. Wang, S. Yan, T. Huang, J. Liu, and X. Tang. Misalignment-robust face recognition. In *CVPR*, 2008.

[14] J. Warrell, S. Prince, and A. Moore. Epitomized priors for multi-labeling problems. In *CVPR*, 2009.