EPITOME AND ITS APPLICATIONS

BY

XINQI CHU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:

Professor Thomas S. Huang

ABSTRACT

Due to the lack of explicit spatial consideration, the existing epitome model may fail for image recognition and target detection, which directly motivates us to propose the so-called spatialized epitome in this thesis. Extended from the original simple 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, a new EM learning procedure is derived under the framework of variational approximation. The learning procedure can generate an optimized summary of the image appearance based on patches and automatically cluster the spatial distribution of the similar patches. From the spatialized epitome, we present a principled (parameter-free) way of inferring the probability of a new input image under the learned 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 tough vision tasks, e.g., misalignment/cross-pose face recognition, and vehicle detection with a few training samples. We also apply this model to image colorization which not only increases the visual appeal of grayscale images, but also enriches the information contained in scientific images that lack color information. Most existing methods of colorization require laborious user interaction for scribbles or image segmentation. To eliminate the need for human labor, we develop an automatic image colorization method using epitome. Built upon a generative graphical model, epitome is a condensed image appearance and shape model which also proves to be an effective summary of color information for the colorization task. We train the epitome from the reference images and perform inference in the epitome to colorize grayscale images, rendering better colorization results than previous methods.

To my family and Nan Shen, for their love and support.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation and sincere gratitude to my supervisors, Professor Thomas S. Huang, Research Scientist Tian-Tsong Ng and Assistant Professor Shuicheng Yan, for their invaluable advice and great support. These individuals encouraged me to get through the tough time during the research. I not only learned specialized knowledge from them, but also the spirit of exploring scientific problems. This work would not have been possible without their excellent guidance and persistent encouragement.

I also would like to thank my collaborators Yingzhen Yang, Zhaowen Wang and Jianchao Yang for their intriguing insights and valuable suggestions.

I am grateful to the Department of Electrical and Computer Engineering (ECE), University of Illinois at Urbana-Champaign, for kindly providing excellent courses, resources and other facilities for my research.

Finally, I would like to thank my parents, my ex-girlfriend Nan Shen, and my friends Eik Shern Lee, Honghai Yu, Xin Zhang, Jing Shen, Qing Duan, Benjamin Chidester and Patrick Johnstone for their kind support.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 LEARNING A SPATIALIZED EPITOME 5

CHAPTER 3 INFERENCE BASED ON SPATIALIZED EPITOME . 13

CHAPTER 4 EXPERIMENTS 16

CHAPTER 5 COLORIZATION BY EPITOME 23

CHAPTER 6 CONCLUSIONS 32

REFERENCES 33

CHAPTER 1

INTRODUCTION

Recently, *epitome* has been successfully applied in computer vision as a patch-based generative model of image(s) or video [1, 2]. As a maximum likelihood representation for image data, it can be considered as a tradeoff 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 [2]. These models are learned by compiling patches drawn from input images into a condensed image model. It was shown in [3] that the image epitome is an image summary of high "completeness." The epitome idea has also found its use in representing audio information [4] and human activities [5]. Jigsaw proposed in [6] took the epitome beyond square patches and modeled local spatial coherence. The epitome model was also extended to location recognition [7], 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 [3]. Most recently, epitome priors are investigated for image parsing in which non-overlapping patches are associated with labels of object classes [8].

Under the generative model framework, the learned 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 reconstruction and recognition during the inference process due to the lack of spatial constraints. For example, epitome was used to recover the

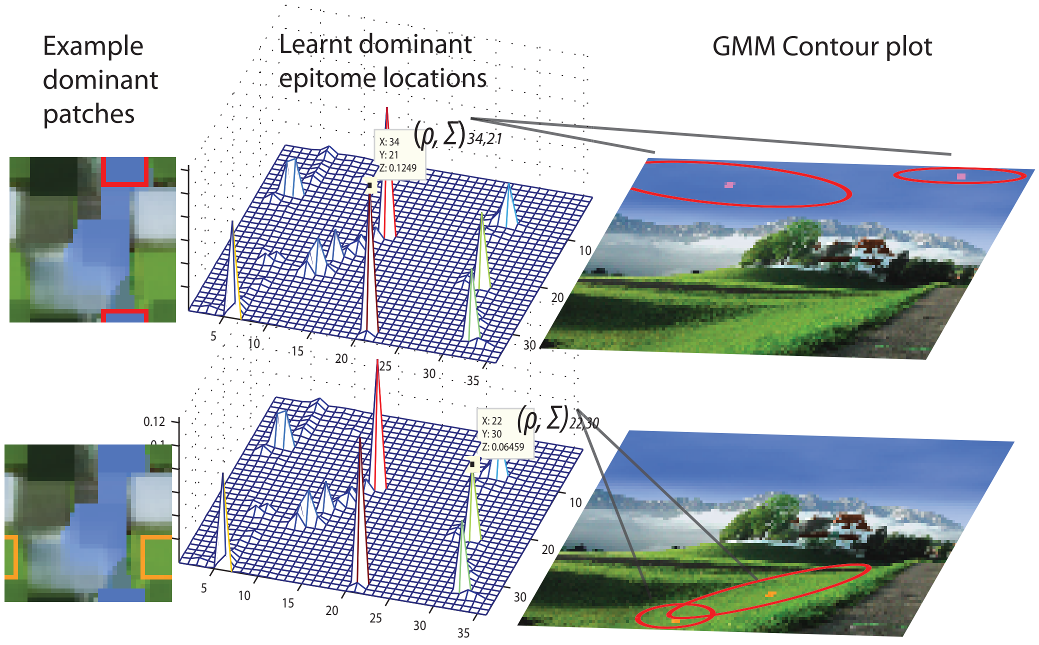

Figure 1.1: A 36 × 36 spatialized epitome (in the first column) is learned from the image in the third column. The distribution in the middle column shows the positions of the significant patches. Note that most locations are of zero value due to regularization. The leftmost image in each row highlights 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.

occluded part of the object in a video by replacing the occlusion with the patches learned 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. 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 [9], 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 rebuilding from compressed image, Wang et al. [10] 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 aforementioned 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 1.1 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 learned 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 by applying 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 thesis can be summarized as follows:

1. A new graphical model of epitome which combines the information about patch appearance and its associated spatial distributions.

2. An EM procedure to learn the optimized appearance summary and cluster the spatial distributions of image patches.

3. A likelihood probability by image inference from the spatialized epitome.

4. Investigation on applying the spatialized epitome for a few tough vision tasks including colorization.

The rest of this thesis is structured as follows: In Chapter 2, we present the spatialized epitome model and the derivation of the learning procedure. We derive the inference process in Chapter 3. Experiments, including the comparisons with the original epitome, on face recognition with misalignments, cross-pose face recognition, and car detection with and without occlusions, are presented in Chapter 4. The application of epitome for colorization is presented in Chapter 5. Conclusions are presented in Chapter 6.

CHAPTER 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 [2, 9], 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 Chapter 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 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 2.1. The generation of each patch (intensity) is formulated as:

$$P(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) = \prod_{i \in S_k} \mathcal{N}(z_{i,k}; \mu_{\mathcal{T}_k(i)}, \phi_{\mathcal{T}_k(i)}), \tag{2.1}$$

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:

$$P(\mathbf{y}_k|\mathcal{T}_k, C_{kr} = 1) = \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \boldsymbol{\Sigma}^r_{\mathcal{T}_k=e}), \tag{2.2}$$

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:

$$p(\mathbf{y}_k|\mathcal{T}_k, C_k) = \prod_{r=1}^{R} \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \boldsymbol{\Sigma}^r_{\mathcal{T}_k=e})^{C_{kr}}. \tag{2.3}$$

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

$$p(C_k|\mathcal{T}_k) = \prod_{r=1}^{R} \tilde{\pi}_{\mathcal{T}_k=e,r}^{C_{\mathcal{T}_k=e,r}}. \tag{2.4}$$

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 learned, 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:

$$\begin{aligned} &p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^{P}, \mathbf{e}, \boldsymbol{\pi}) = \\ &\qquad p(\mathbf{e}, \boldsymbol{\pi}) \prod_{k=1}^{P} p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) p(\mathbf{y}_k|\mathcal{T}_k, C_k) p(C_k, \mathcal{T}_k), \end{aligned} \tag{2.5}$$

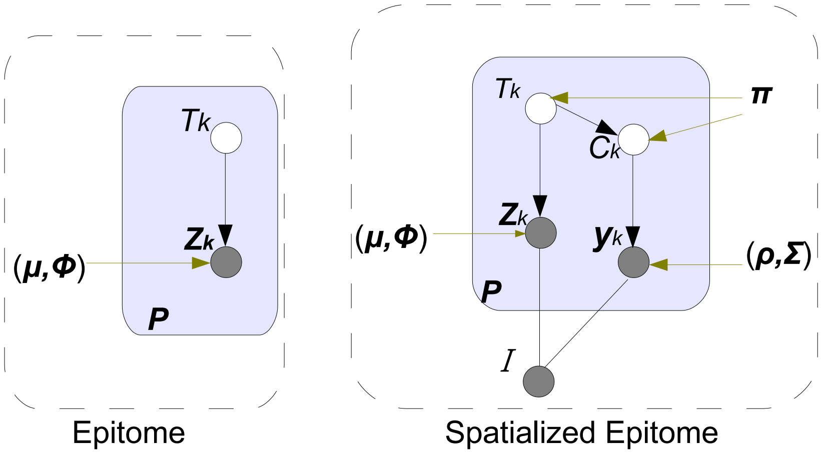

Figure 2.1: The graphical model representations of the epitome and the spatialized epitome. The boxes are "plates" representing replicates.

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

$$\begin{aligned}\log P(\{\mathbf{Z}_k, \mathbf{y}_k\}_{k=1}^{P}) &= \log \sum_{\{C_k,\mathcal{T}_k\}} \int_{\mathbf{e},\boldsymbol{\pi}} p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^{P}, \mathbf{e}, \boldsymbol{\pi}) d(\mathbf{e}, \boldsymbol{\pi}) \\ &= \log \sum_{\{C_k,\mathcal{T}_k\}} \prod_{k=1}^{P} p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) p(\mathbf{y}_k|\mathcal{T}_k, C_k) p(C_k, \mathcal{T}_k). \quad (2.6)\end{aligned}$$

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 [11]:

$$\begin{aligned}&\log P(\{\mathbf{Z}_k, \mathbf{y}_k\}_{k=1}^{P}) = \log \sum_{\{C_k,\mathcal{T}_k\}} \frac{q(\{\mathcal{T}_k, C_k\}_{k=1}^{P}) p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^{P})}{q(\{\mathcal{T}_k, C_k\}_{k=1}^{P})} \\ &\geq \sum_{\{C_k,\mathcal{T}_k\}} q(\{\mathcal{T}_k, C_k\}_{k=1}^{P}) \log \frac{p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^{P})}{q(\{\mathcal{T}_k, C_k\}_{k=1}^{P})} \\ &= \sum_{\{C_k,\mathcal{T}_k\}} q(\{\mathcal{T}_k, C_k\}_{k=1}^{P}) \log p(\{\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k\}_{k=1}^{P}) \\ &\qquad - \sum_{\{C_k,\mathcal{T}_k\}} q(\{\mathcal{T}_k, C_k\}_{k=1}^{P}) \log q(\{\mathcal{T}_k, C_k\}_{k=1}^{P}) = B. \quad (2.7)\end{aligned}$$

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 [11], we have

$$\log P(\{\mathbf{Z}_k, \mathbf{y}_k\}_{k=1}^P) \geq B =$$

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

$$- \sum_{\{C_k,\mathcal{T}_k\}} q(\{\mathcal{T}_k, C_k\}_{k=1}^P) \log q(\{\mathcal{T}_k, C_k\}_{k=1}^P)$$

$$= \sum_{k=1}^P \sum_{C_k,\mathcal{T}_k} q(\mathcal{T}_k, C_k) [\log p(\mathcal{T}_k, C_k) +$$

$$\log p(\mathbf{y}_k|\mathcal{T}_k, C_k) + \log p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}})] - E. \quad (2.8)$$

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 $p(C_k, \mathcal{T}_k)$, $p(\mathbf{y}_k|\mathcal{T}_k, C_k)$ and $p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}})$ 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

$$q(\mathcal{T}_k, C_k) = p(\mathcal{T}_k, C_k|\mathbf{Z}_k, \mathbf{y}_k, \hat{\mathbf{e}}) = \frac{p(\mathbf{Z}_k, \mathcal{T}_k, C_k, \mathbf{y}_k, \hat{\mathbf{e}})}{p(\mathbf{Z}_k, \mathbf{y}_k, \hat{\mathbf{e}})}$$

$$= \frac{p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}}) p(\mathbf{y}_k|\mathcal{T}_k, C_k) p(C_k, \mathcal{T}_k)}{p(\mathbf{Z}_k, \mathbf{y}_k, \hat{\mathbf{e}})}$$

$$\sim p(\mathbf{Z}_k|\mathcal{T}_k, \hat{\mathbf{e}}) p(\mathbf{y}_k|\mathcal{T}_k, C_k) p(C_k, \mathcal{T}_k)$$

$$= \prod_{i \in S_k} \mathcal{N}(z_{i,k}; \mu_{\mathcal{T}_k(i)}, \phi_{\mathcal{T}_k(i)}) \prod_{r=1}^R \mathcal{N}(\mathbf{y}_k; \rho^r_{\mathcal{T}_k=e}, \mathbf{\Sigma}^r_{\mathcal{T}_k=e})^{C_{kr}} p(C_k, \mathcal{T}_k). \quad (2.9)$$

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

$$\begin{aligned} B_1 = &\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} q_k \log p(\mathbf{Z}_k | \mathcal{T}_k, \hat{\mathbf{e}}) = \\ &= \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i \in S_k} q_k \left[-\frac{1}{2} \log 2\pi\phi_j - \frac{(z_{i,k} - \mu_j)^2}{2\phi_j} \right]. \end{aligned} \quad (2.10)$$

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:

$$\mu_j = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i \in S_k} q(\mathcal{T}_k, C_k) z_{i,k}}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i \in S_k} q(\mathcal{T}_k, C_k)}, \quad (2.11)$$

and the corresponding updating rule for ϕ_j is:

$$\phi_j = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i \in S_k} q(\mathcal{T}_k, C_k)(z_{i,k} - \mu_j)^2}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k(i)=j} \sum_{i \in S_k} q(\mathcal{T}_k, C_k)}. \quad (2.12)$$

This is similar to the original epitome updating rules.

b) Update GMM Means and Covariances

From Eq. (2.8), the bound for the GMM term is simplified as:

$$\begin{aligned} B_2 = &\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) \log p(\mathbf{y}_k | \mathcal{T}_k, C_k) = \\ &= \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) \sum_{r=1}^{R} C_{kr} \log \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \boldsymbol{\Sigma}^r_{\mathcal{T}_k=e}). \end{aligned} \quad (2.13)$$

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

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\rho}^r_e} \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) \sum_{r=1}^{R} C_{kr} \log \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \boldsymbol{\Sigma}^r_{\mathcal{T}_k=e}) \\ &= \frac{\partial}{\partial \boldsymbol{\rho}^r_e} \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} \log \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k=e}, \boldsymbol{\Sigma}^r_{\mathcal{T}_k=e}) \\ &= \sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} (\mathbf{y}_k - \boldsymbol{\rho}^r_e)^T (\boldsymbol{\Sigma}^r_e)^{-1} = 0. \end{aligned} \quad (2.14)$$

From the equation 2.14, we can obtain the updating rule for $\boldsymbol{\rho}_e^r$ as:

$$(\boldsymbol{\rho}_e^r)^T = \frac{\sum_{k=1}^P \sum_{C_k, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr} \mathbf{y}_k^T}{\sum_{k=1}^P \sum_{C_k, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr}}. \tag{2.15}$$

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

$$\begin{aligned}
&\frac{\partial}{\partial(\boldsymbol{\Sigma}_e^r)^{-1}} \sum_{k=1}^P \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} \log \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}_{\mathcal{T}_k=e}^r, \boldsymbol{\Sigma}_{\mathcal{T}_k=e}^r) \\
&\qquad = \frac{\partial}{\partial(\boldsymbol{\Sigma}_e^r)^{-1}} \sum_{k=1}^P \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} [- \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_e^r| - \\
&\qquad\qquad \frac{1}{2} (\mathbf{y}_k - \boldsymbol{\rho}_e^r)^T (\boldsymbol{\Sigma}_e^r)^{-1} (\mathbf{y}_k - \boldsymbol{\rho}_e^r)] \\
&\qquad = \sum_{k=1}^P \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) C_{kr} [+ \frac{1}{2} \boldsymbol{\Sigma}_e^r - \frac{1}{2} (\mathbf{y}_k - \boldsymbol{\rho}_e^r)^T (\mathbf{y}_k - \boldsymbol{\rho}_e^r)] = 0.
\end{aligned} \tag{2.16}$$

Therefore we obtain the updating rule for $\boldsymbol{\Sigma}_e^r$ as,

$$\boldsymbol{\Sigma}_e^r = \frac{\sum_{k=1}^P \sum_{C_k, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr} (\mathbf{y}_k - \rho_e^r)(\mathbf{y}_k - \boldsymbol{\rho}_e^r)^T}{\sum_{k=1}^P \sum_{C_k, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr}}. \tag{2.17}$$

c) Update mixing coefficients

From Eq. (2.8), the term related to mixing coefficients can be expressed:

$$B_3 = \sum_{k=1}^P \sum_{C_k, \mathcal{T}_k} q(\mathcal{T}_k, C_k) \log p(\mathcal{T}_k, C_k). \tag{2.18}$$

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:

$$\begin{aligned}
&\frac{\partial}{\partial \pi_{er}} (B_3 + \lambda (\sum_{e,r} \pi_{er} - 1)) \\
&= \frac{\partial}{\partial \pi_{er}} \sum_{k=1}^P \sum_{C_k=r, \mathcal{T}_k=e} q(\mathcal{T}_k, C_k) \log p(\mathcal{T}_k = e, C_k = r) + \lambda \\
&= \sum_{k=1}^P q(\mathcal{T}_k = e, C_k = r) \frac{1}{\pi_{er}} + \lambda = 0.
\end{aligned} \tag{2.19}$$

Table 2.1: 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,

$$\pi_{er} = \frac{\sum_{k=1}^{P} q(\mathcal{T}_k = e, C_k = r)}{P}. \tag{2.20}$$

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 2.1. 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.

$$p(\{\boldsymbol{\rho}_e^r, \boldsymbol{\Sigma}_e^r\}_{r=1}^R, \boldsymbol{\pi}_e) = b(\boldsymbol{\gamma}_e) \prod_{r=1}^{R} (\pi_e^r)^{\gamma_e^r - 1} \\ \prod_{r=1}^{R} \mathcal{N}\left(\boldsymbol{\rho}_e^r | \boldsymbol{\nu}_e^r, \frac{\boldsymbol{\Sigma}_e^r}{\eta_e^r}\right) Wi((\boldsymbol{\Sigma}_e^r)^{-1} | \boldsymbol{\beta}_e^r, \tau_e^r), \tag{2.21}$$

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 [12]. By incorporating the prior, the updating rules are derived to be:

$$(\boldsymbol{\rho}_e^r)^T = \frac{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k = e} q(\mathcal{T}_k, C_k) C_{kr} \mathbf{y}_k^T + \eta_e^r \boldsymbol{\nu}_e^r}{\sum_{k=1}^{P} \sum_{C_k, \mathcal{T}_k = e} q(\mathcal{T}_k, C_k) C_{kr} + \eta_e^r}, \tag{2.22}$$

$$\boldsymbol{\Sigma}_e^r = \frac{\sum_{k=1}^{P}\sum_{C_k,\mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr}(\mathbf{y}_k - \rho_e^r)(\mathbf{y}_k - \rho_e^r)^T}{\sum_{k=1}^{P}\sum_{C_k,\mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr} + 2\tau_e^r - 2} + \frac{\eta_e^r(\boldsymbol{\mu}_e^r - \boldsymbol{\nu}_e^r)(\boldsymbol{\mu}_e^r - \boldsymbol{\nu}_e^r)^T + 2\boldsymbol{\beta}_e^r}{\sum_{k=1}^{P}\sum_{C_k,\mathcal{T}_k=e} q(\mathcal{T}_k, C_k) C_{kr} + 2\tau_e^r - 2}, \quad (2.23)$$

$$\pi_{er} = \frac{\sum_{k=1}^{P} q(\mathcal{T}_k = e, C_k = r) + \gamma_e^r - 1}{P + \sum_{r=1}^{R}\gamma_e^r - R}. \quad (2.24)$$

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 1.1.

CHAPTER 3

INFERENCE BASED ON SPATIALIZED EPITOME

3.1 Inference

We denote the set of learned 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:

$$\begin{aligned}
&\log P(I|\mathcal{D}) \simeq \log P(I|\hat{\boldsymbol{\Theta}}) = \log P(I|\boldsymbol{\rho}, \boldsymbol{\Sigma}, \hat{\mathbf{e}}, \boldsymbol{\pi}) \\
&= \log P(\{\mathbf{Z}_k, \mathbf{y}_k\}_{k=1}^{P} | \boldsymbol{\rho}, \boldsymbol{\Sigma}, \hat{\mathbf{e}}, \boldsymbol{\pi}) \\
&= \log \prod_{k=1}^{P} P(\mathbf{Z}_k, \mathbf{y}_k | \boldsymbol{\rho}, \boldsymbol{\Sigma}, \hat{\mathbf{e}}, \boldsymbol{\pi}) \\
&= \sum_{k=1}^{P} \log \sum_{C_k, \mathcal{T}_k} P(\mathbf{Z}_k, \mathbf{y}_k, C_k, \mathcal{T}_k | \boldsymbol{\rho}, \boldsymbol{\Sigma}, \hat{\mathbf{e}}, \boldsymbol{\pi}) \\
&= \sum_{k=1}^{P} \log \sum_{C_k, \mathcal{T}_k} p(\mathbf{Z}_k | \mathcal{T}_k, \mathbf{e}) p(\mathbf{y}_k | \mathcal{T}_k, C_k) P(C_k, \mathcal{T}_k) \\
&= \sum_{k=1}^{P} \log \sum_{C_k, \mathcal{T}_k} \prod_{i \in S_k} \mathcal{N}(z_{i,k}; \mu_{\mathcal{T}_k(i)}, \phi_{\mathcal{T}_k(i)}) \\
&\qquad\qquad\qquad \prod_{r=1}^{R} \mathcal{N}(\mathbf{y}_k; \boldsymbol{\rho}^r_{\mathcal{T}_k = e}, \boldsymbol{\Sigma}^r_{\mathcal{T}_k = e})^{C_{kr}} P(\mathcal{T}_k, C_k). \qquad (3.1)
\end{aligned}$$

This inference formulation is similar to the way of evaluating the probability value of seeing a new data under a learned GMM. The first step of this derivation follows [13]. The third step uses the assumption that all the patches are independently sampled. The calculated probability value in Eq. 3.1 indicates how likely the probe image is generated by the learned model, and can be directly used for image recognition and object detection purposes.

3.2 Recognition and Detection

Suppose there are N epitomes with parameters $\{\mathbf{\Theta}_i\}_{i=1}^N$ learned 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:

$$p(\mathcal{C}|I) = \frac{p(I|\mathcal{C})p(\mathcal{C})}{p(I)} \sim p(I|\mathcal{C}), \tag{3.2}$$

and select the one with the maximum posterior value:

$$\hat{\mathcal{C}} = \arg\max_i P(I|\mathcal{C} = i) = \arg\max_i P(I|\mathbf{\Theta}_i), \tag{3.3}$$

where $P(I|\mathbf{\Theta}_i)$ can be calculated from Eq. (3.2) which is in turn calculated by Eq. (3.1).

Detection If we scan the input image with multi-scale windows (W), we can perform object detection. In this way, Eq.(3.2) becomes

$$p(\mathcal{C}|W) = \frac{p(W|\mathcal{C})p(\mathcal{C})}{p(W)} \sim p(W|\mathcal{C}). \tag{3.4}$$

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

3.3 Epitomic reestimation

Using existing epitome for image reestimation, 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 [2, 9]. This may give an incorrect estimation when the occluded part has different appearances from nearby patches.

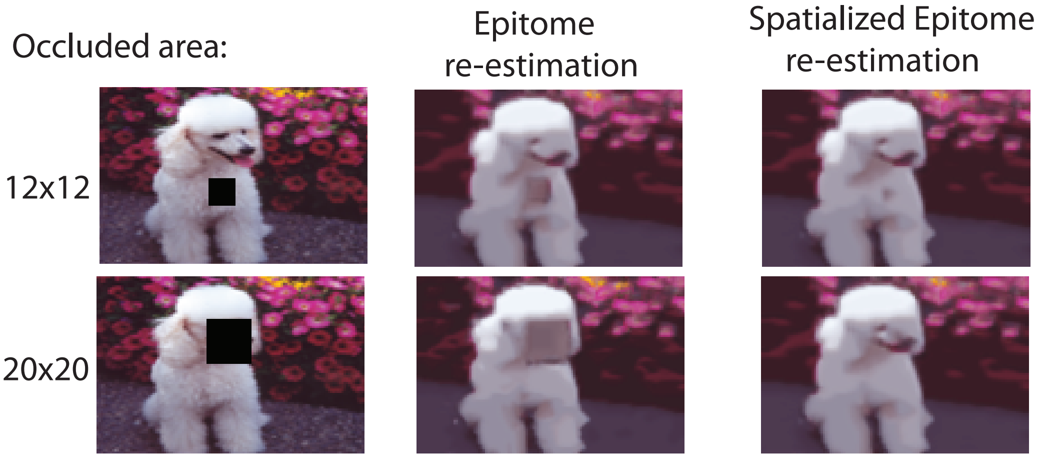

Figure 3.1: The comparison of image reestimation results between epitome and spatialized epitome. Both 40×40 epitomes are learned with patch sizes of 8×8 and 4×4 which is also the patch size used in the reestimation process. During the reestimation process, $40,000$ patches are uniformly sampled from the input image to ensure that all the coordinates are covered for the reestimated image. Since the original epitome just uses a color/intensity similarity to estimate the position posterior, the patches probabilistically chosen from the epitome generate artifacts in the occluded region. In contrast, the spatialized epitome estimates the position posterior based on both intensity similarity and location information; thus, many fewer artifacts are generated due to the spatial constraint. For non-uniform image regions with occlusion, e.g. the second row, spatialized epitome can also restore the occluded region with proper patches.

The reestimation process of spatialized epitome will automatically solve this problem as the position posterior $q(\mathcal{T}_k, C_k)$ takes also the spatial arrangement into account as in Eq. (2.9) in image reestimation. The comparison of existing epitome and spatialized epitome on image reestimation from partially occluded image is given in Figure 3.1.

CHAPTER 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 reestimation 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 tough 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 in the following sections. We will provide functional codes such as spatialized epitome learning, inference and synthesis to reproduce the results in this thesis. The codes for the current state-of-the -art results on misalignment face recognition are also provided to facilitate future works.

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 learned 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

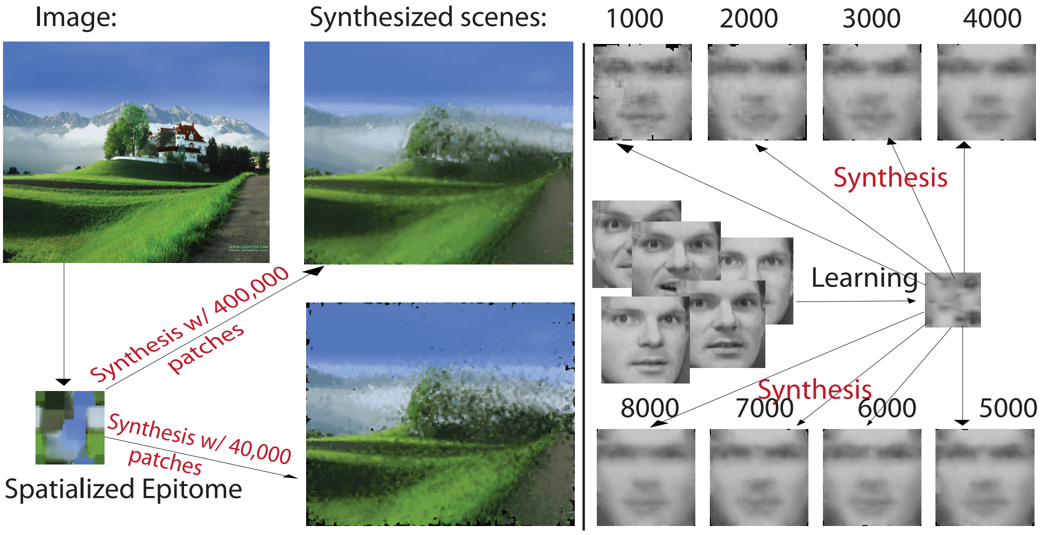

Figure 4.1: The left half of the figure shows the synthesis results for a spatialized epitome learned from a scene image. At the right half of the figure, we show synthesis results for a spatialized epitome model learned from multiple images from the same person.

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 [9, 2]. Following the same principle in Chapter 3, the inferred probability of seeing a new image with original epitome is:

$$\log P(I|\mathcal{D}) \simeq \log P(I|\hat{\mathbf{e}}) = \log P(\{\mathbf{Z}_k\}_{k=1}^{P}|\hat{\mathbf{e}}) \\ = \sum_{k=1}^{P} \log \sum_{\mathcal{T}_k} \prod_{i \in S_k} \mathcal{N}(z_{i,k}; \mu_{\mathcal{T}_k(i)}, \phi_{\mathcal{T}_k(i)}) P(\mathcal{T}_k). \quad (4.1)$$

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 and manually normalized to the size of 32 × 32 pixels. Both original and spatialized epitomes are evaluated with two different patch sizes.

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

Table 4.1: Recognition accuracy rates (%) on two face databases.

Database:	ORL		PIE	
Patch Size:	4×4	6×6	4×4	6×6
Epitome	12.0	15.5	8.2	11.2
Spatialized	67.5	88.5	74.1	78.8

We can observe from Table 4.1 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. (3.1). 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 CMU PIE and ORL databases. We randomly pick five images of each subject for training, the remaining five images of each person serve as probe images. Then an 18×18 artificial occlusion is generated at a random position in each probe image. Seven images are randomly selected from the probe set and the occlusion detection results are shown in Figure 4.2, where the first row shows the original face images, the second row shows the images with occlusions, the third row shows the detected occlusion regions, and the fourth row shows the reconstructed images by the spatialized epitome.

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,

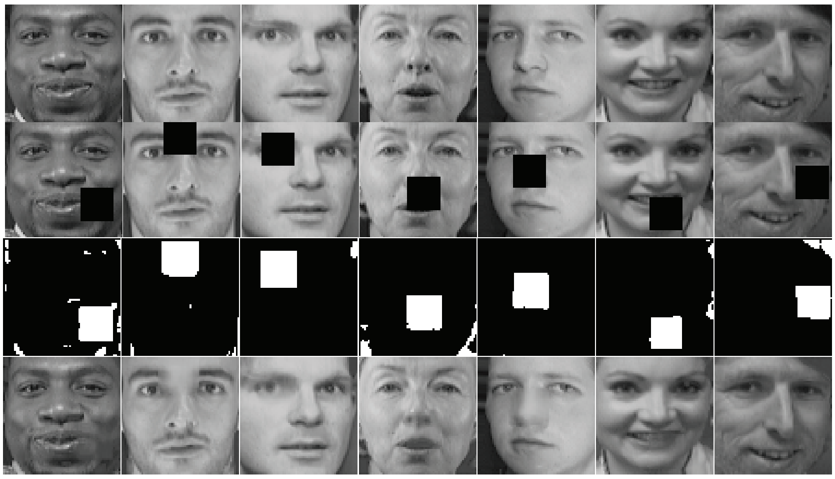

Figure 4.2: Examples of occlusion detection.

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 [14]. 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 learned from patches of images of different expressions (ORL) or different poses (PIE), so the deformation is learned to account for misalignments on the patch level; and (3) the misalignment effect is reduced from the image level to patch level. We evaluate the performance of our algorithm with respect to each of the misalignment factor, e.g., translation, scaling, and rotation as well as the mixed spatial misalignments to simulate the misalignments brought by the automatic face alignment process. 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 four images from each subject is used for training while the remaining six images of each person are artificially misaligned with a rotation $\alpha \in [-5°, 5°]$, 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

Table 4.2: 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

Table 4.3: 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

a uniform distribution. The performance of our algorithm for each misalignment factor is evaluated in Table 4.2 and compared with baselines algorithms such as PCA and LDA (the results come from [14] with four training samples). In the mixed spatial misalignment configuration, the aforementioned effects are added in a random order to the original test image, and the results are shown in Table 4.2.

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, three images with illumination index 8, 11, 21 from each of the two near frontal poses, namely c05 and c29 are chosen as training set. three images from each of the five 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 neighbor classifier) baselines are listed in Table 4.3.

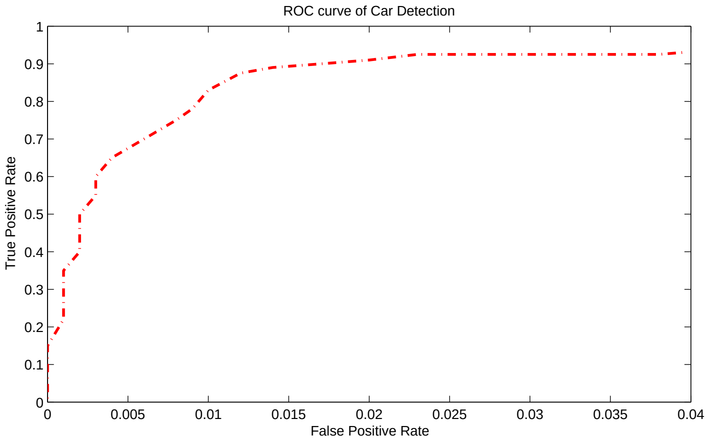

Figure 4.3: The ROC curve of car detection.

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. (3.1). 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 the ROC curve as shown in Figure 4.3. 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.

In these experiments, we have shown the strong abilities of spatialized epitome for image representation, pattern recognition, and object detection. Especially, the tests on some tough vision tasks like misaligned and cross-pose face recognition demonstrate the advantages of the spatialized epitome

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

in adapting to variations in real-world conditions.

CHAPTER 5

COLORIZATION BY EPITOME

Colorization adds color to grayscale images by assigning color values to images which only contain a grayscale channel. It not only increases the visual appeal, but also enhances the information conveyed by scientific images. For example, the grayscale images acquired by scanning electron microscopy (SEM) can be made more illustrative by adding different colors to different parts of the images. However, the manual colorization is tedious and time consuming, so it is not suitable for batch process. To overcome this problem, we propose an automatic colorization method by epitome. We train the epitome from one manually colorized nano mushroom-like image, and use that epitome to automatically colorize the other nano mushroom-like image, which eliminates the need for human labor and makes the batch colorization process possible.

Based on the source of the color information used to colorize the grayscale images, existing colorization techniques fall into two main categories: user scribble based methods and color transfer methods. The user scribble based method in [15] asked users to draw color scribbles in the grayscale image, and the algorithm propagated the user-provided color to the whole image requiring that similar neighboring pixels should receive similar color. Later, L. Qing et al. [16] proposed a method which required less human intervention. The user scribbles were employed for texture segmentation and user-provided color was propagated within each segment. Using a similar color image as a reference, the color transfer methods such as [17] performed colorization by transferring the color from the reference image to the grayscale image, either automatically or with user intervention. However, the pixel-level matching based on luminance value and neighborhood statistics adopted by [17] suffered from spatial inconsistency and the user-provided swatches were required to guide the matching process in many cases. Using [18] improved the spatial consistency by an image space voting scheme. Their method first transferred color to a few pixels in the target image with high confidence, then applied the method in [15] to colorize the whole image, treating the colorized pixels in the first step as the scribbles. However, their method required a robust segmentation of the reference image, which was difficult in many cases without user intervention.

Similar to [17], our automatic colorization method transfers the color information from the reference image to the target grayscale image. Since most of existing colorization methods need user interactions for color selection or segmentation, a robust and automatic colorization algorithm is preferable. In order to approach this problem, it is worthwhile to exploit the biological characteristics of human visual system. The average human retina contains many more rods than cones [19] (92 million rods versus 4.6 million cones). Rods are more sensitive to cones but they are not sensitive to color, so that most of visually significant variation arises only from luminance differences. This fact suggests that we do not need to search the whole reference image for the color patches to colorize the target image, instead we can reduce the search space for color patches, or equivalently find an effective color summary of the reference image, to improve the efficiency and alleviate color assignment ambiguity. In [17], such a color summary is a set of source color pixels randomly sampled, which is, however, subject to noise in the raw pixels.

In order to find an effective and compact summary of the color information in the reference image, we adopt the condensed image appearance and shape representation, i.e. epitome [20]. Epitome consolidates self-similar patches in the spatial domain, and the size of the epitome is much smaller than that of the image it models. By virtual of the generative graphical model, epitome can be interpreted as a tradeoff between template and histogram for image representation and it has been applied to many computer vision tasks such as object detection, location recognition, and synthesis [21, 22]. Epitome summarizes a large number of raw patches in the reference image by only representing the most constitutive elements. In our epitomic colorization scheme, the color patches used to colorize the target grayscale image are retrieved from the epitome trained with the reference image, rather than from the raw image patches. Epitome proves to be an effective summary of the color information in the reference image, which produces more satisfactory colorization results than [17] in the experiments.

5.1 Description of Automatic Colorization by Epitome

Given a reference color image cI and the target grayscale image gI, we aim to automatically colorize gI with the color information from cI. We achieve this goal by first training an epitome e from the reference image, then performing inference in e so as to transfer the color information of the color patches of $\hat{\mathbf{e}}$ to the corresponding grayscale patches of gI. Note that the grayscale channel of gI is retained as the luminance channel after the color transfer process. We will illustrate the training and inference process in detail in the following subsections.

5.2 Training the Epitome

Epitome is a latent representation of an image, which comprises hidden variables and parameters required to generate the image patches according to the epitome graphical model. Epitome summarizes a large set of raw image patches into a condensed representation of a size much smaller than the original image, and it approaches this goal in a manner similar to Gaussian mixture model with overlapping means and variances.

The epitome e of an image I of size $M \times N$ is a condensed representation of size $M_e \times N_e$ where $M_e < M$ and $N_e < N$. The epitome contains two parameters: $\mathbf{e} = (\boldsymbol{\mu}, \boldsymbol{\phi})$. $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$ represent the Gaussian mean and variance, respectively, and both are of size $M_e \times N_e$. Suppose Q patches are sampled from the reference image, i.e. $\{\mathbf{Z}_k\}_{k=1}^{Q}$, and each patch $\mathbf{Z}_k$ contains pixels with image coordinates $\mathbf{S}_k$. Similar to [20], the patches are square and we use fixed patch size throughout this chapter. These patches are densely sampled and they can be overlapping with each other to cover the entire image. We associate each patch $\mathbf{Z}_k$ with a hidden mapping $\mathcal{T}_k$ which maps the image coordinates $\mathbf{S}_k$ to the epitome coordinates, and all the Q patches are generated independently from the epitome parameters and the corresponding hidden mappings as follows:

$$p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) = \prod_{i \in \mathbf{S}_k} \mathcal{N}(z_{i,k}; \boldsymbol{\mu}_{\mathcal{T}_k(i)}, \boldsymbol{\phi}_{\mathcal{T}_k(i)}), k = 1..Q \qquad (5.1)$$

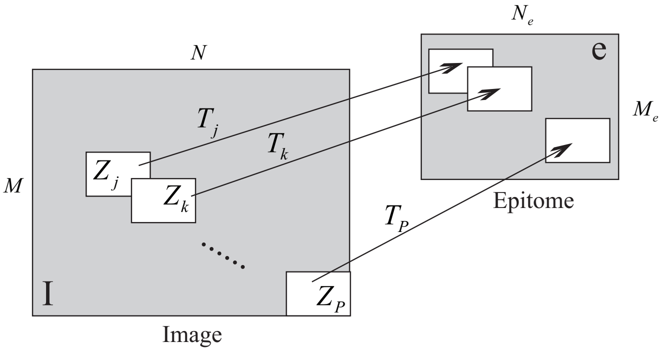

Figure 5.1: The mapping $\mathcal{T}_k$ maps the image patch $\mathbf{Z}_k$ to its corresponding epitome patch with the same size, and $\mathbf{Z}_k$ can be mapped to any possible epitome patch according to $\mathcal{T}_k$.

and

$$\prod_{k=1}^{Q} p(\{\mathbf{Z}_k\}_{k=1}^{Q} | \{\mathcal{T}_k\}_{k=1}^{Q}, \mathbf{e}) = \prod_{k=1}^{Q} p(\mathbf{Z}_k | \mathcal{T}_k, \mathbf{e}). \tag{5.2}$$

where $z_{i,k}$ is the pixel with image coordinates i from the k-th patch. Since $z_{i,k}$ is independent of the patch number k, we simply denote it as z_i in the following text. $\mathcal{N}(\cdot; \mu, \phi)$ represents a Gaussian distribution with mean $\hat{\mu}$ and variance $\hat{\phi}$

$$\mathcal{N}(\cdot; \hat{\mu}, \hat{\phi}) = \frac{1}{\sqrt{2\pi\hat{\phi}}} \exp^{-\frac{(\cdot - \hat{\mu})^2}{2\hat{\phi}}}.$$

Based on Eq.(5.1), the hidden mapping $\mathcal{T}_k$ can be interpreted as a hidden variable that indicates the location of the epitome patch from which the observed image patch $\mathbf{Z}_k$ is generated, and it behaves similar to the hidden variable in the traditional Gaussian mixture models that specifies the Gaussian component from which a specific data point is generated. Also, $\mathcal{T}_k$ maps the image patch to its corresponding epitome patch, and the number of possible mappings that each $\mathcal{T}_k$ can take, denoted as L, is determined by all the discrete locations in the epitome ($L = M_e \times N_e$ in our setting). Figure 5.1 illustrates the role that the hidden mapping variables play in the generative model, and Figure 5.2 shows the epitome graphical model, which again demonstrate its similarity to Gaussian mixture models. $\boldsymbol{\pi} \triangleq \{\pi_l\}_{l=1}^{L}$ indicates the prior distribution of the hidden mapping. Suppose $\mathcal{T}_{k,l}$ is the l-th mapping that $\mathcal{T}_k$ can take, then

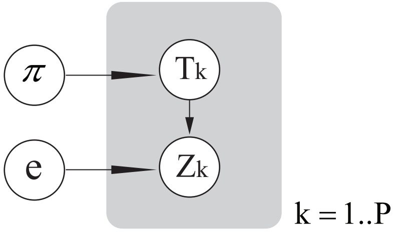

Figure 5.2: The epitome graphical model.

$$p(\mathcal{T}_k) = \prod_{l=1}^{L} {\pi_l}^{\delta\left(\mathcal{T}_k = \mathcal{T}_{k,l}\right)},$$

which holds for any $k \in \{1..Q\}$. δ is an indicator function and δ equals to 1 when its argument is true, and 0 otherwise.

Our goal is to find the epitome $\hat{\mathbf{e}}$ that maximizes the log likelihood function:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \log p\left(\{\mathbf{Z}_k\}_{k=1}^{Q} | \mathbf{e}\right). \tag{5.3}$$

Given the epitome $\mathbf{e}$, the likelihood function for the complete data, i.e. the image patches $\{\mathbf{Z}_k\}_{k=1}^{Q}$ and the hidden mappings $\{\mathbf{Z}_k\}_{k=1}^{Q}$, is derived in the following according to the epitome graphical model:

$$\begin{aligned} p(\{\mathbf{Z}_k, \mathcal{T}_k\}_{k=1}^{Q} | \mathbf{e}, \boldsymbol{\pi}) &= \prod_{k=1}^{Q} p(\mathbf{Z}_k, \mathcal{T}_k | \mathbf{e}, \boldsymbol{\pi}) \\ &= \prod_{k=1}^{Q} p(\mathcal{T}_k) p(\mathbf{Z}_k | \mathcal{T}_k, \mathbf{e}) \\ &= \prod_{k=1}^{Q} \prod_{l=1}^{L} \left[\pi_l \prod_{j \in \mathbf{S}_k} \mathcal{N}(z_j; \boldsymbol{\mu}_{\mathcal{T}_{k,l}(j)}, \boldsymbol{\phi}_{\mathcal{T}_{k,l}(j)}) \right]^{\delta\left(\mathcal{T}_k = \mathcal{T}_{k,l}\right)} \end{aligned} \tag{5.4}$$

We use the expectation-maximization algorithm [23] to maximize the likelihood function Eq.(5.3) and learn the epitome $\hat{\mathbf{e}}$, following the procedure introduced in [24].

The E-step: The posterior distribution of the hidden variables, i.e. the hidden mapping is

$$
\begin{aligned}
q(\mathcal{T}_k) &\triangleq p(\mathcal{T}_k|\mathbf{Z}_k, \mathbf{e}, \boldsymbol{\pi}) \\
&= \frac{p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e})p(\mathcal{T}_k)}{\sum_{\mathcal{T}_k} p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e})p(\mathcal{T}_k)} \\
&= \frac{\prod_{l=1}^{L} \left[\pi_l \prod_{j\in\mathbf{S}_k} \mathcal{N}(z_j; \boldsymbol{\mu}_{\mathcal{T}_{k,l}(j)}, \boldsymbol{\phi}_{\mathcal{T}_{k,l}(j)})\right]^{\delta\left(\mathcal{T}_k=\mathcal{T}_{k,l}\right)}}{\sum_{\mathcal{T}_k} \prod_{l=1}^{L} \left[\pi_l \prod_{j\in\mathbf{S}_k} \mathcal{N}(z_j; \boldsymbol{\mu}_{\mathcal{T}_{k,l}(j)}, \boldsymbol{\phi}_{\mathcal{T}_{k,l}(j)})\right]^{\delta\left(\mathcal{T}_k=\mathcal{T}_{k,l}\right)}}.
\end{aligned}
\tag{5.5}
$$

We observe that $q(\mathcal{T}_k)$ corresponds to the responsibility in Gaussian mixture models.

The M-step: We obtain the expectation of the log-likelihood function for the complete data with respect to the posterior distribution of the hidden mapping from the E-step as follows:

$$
\begin{aligned}
&E\left[\log p\left(\{\mathbf{Z}_k, \mathcal{T}_k\}_{k=1}^{Q}|\mathbf{e}, \boldsymbol{\pi}\right)\right] \\
&= \sum_{k=1}^{Q}\sum_{l=1}^{L} q(\mathcal{T}_k = \mathcal{T}_{k,l}) \cdot [\log \pi_l + \log p\,(\mathbf{Z}_k|\mathcal{T}_k = \mathcal{T}_{k,l}, \mathbf{e})].
\end{aligned}
\tag{5.6}
$$

Maximizing Eq.(5.6) with respect to $(\mathbf{e}, \boldsymbol{\pi})$, we get the following update of the parameters of the epitome and $\boldsymbol{\pi}$:

$$
\boldsymbol{\mu}_j = \frac{\sum\limits_{k=1}^{Q} \sum_{i\in\mathbf{S}_k} \sum_{\mathcal{T}_k} \delta(\mathcal{T}_k(i) = j)q(\mathcal{T}_k)z_i}{\sum\limits_{k=1}^{Q} \sum_{i\in\mathbf{S}_k} \sum_{\mathcal{T}_k} \delta(\mathcal{T}_k(i) = j)q(\mathcal{T}_k)} \tag{5.7}
$$

$$
\boldsymbol{\phi}_j = \frac{\sum\limits_{k=1}^{Q} \sum_{i\in\mathbf{S}_k} \sum_{\mathcal{T}_k} \delta(\mathcal{T}_k(i) = j)q(\mathcal{T}_k)(z_i - \boldsymbol{\mu}_j)^2}{\sum\limits_{k=1}^{Q} \sum_{i\in\mathbf{S}_k} \sum_{\mathcal{T}_k} \delta(\mathcal{T}_k(i) = j)q(\mathcal{T}_k)} \tag{5.8}
$$

$$
\pi_l = \frac{\sum\limits_{k=1}^{Q} p\left(\mathcal{T}_k = \mathcal{T}_{k,l}\right)}{Q}, l = 1..L. \tag{5.9}
$$

The index j indicates the epitome coordinates in Eq.(5.7) and Eq.(5.8).

We alternate between E-step and M-step until convergence or the maximum number of iterations (20 in our experiments) is achieved, and then obtain the resultant epitome $\hat{\mathbf{e}}$ from the reference image cI.

Note that the preceding training process is applicable for a single type of feature of cI. We use two types of feature to train the epitome, i.e. the YIQ channels and the dense sift feature [25]. We convert cI from the RGB color space to the YIQ color space where Y channel represents the luminance and IQ channels represent chrominance information. Moreover, dense sift feature is computed for each sampled patch. A $K \times K$ patch is evenly divided into $R \times R$ grids, and the orientation histogram of the gradients with eight bins is calculate for each grid, which results in an $8R^2$-dimensional dense sift feature vector for each patch. R is typically set to be 3 or 4. We then train the epitome $\mathbf{e} = \left(\mathbf{e}^{YIQ}, \mathbf{e}^{dsift}\right)$ for the YIQ channels and the dense sift feature, and the epitome for YIQ channels ($\mathbf{e}^{YIQ}$) share the same hidden mapping with the epitome for the dense sift feature ($\mathbf{e}^{dsift}$) in the inference process [21]:

$$p(\mathbf{Z}_k|\mathcal{T}_k, \mathbf{e}) = p(\mathbf{Z}_k^{YIQ}|\mathcal{T}_k, \mathbf{e}^{YIQ})^{\lambda} p(\mathbf{Z}_k^{dsift}|\mathcal{T}_k, \mathbf{e}^{dsift})^{1-\lambda}, \tag{5.10}$$

where $\mathbf{Z}_k^{YIQ}$ and $\mathbf{Z}_k^{sift}$ represent the YIQ channel and the dense sift feature of patch $\mathbf{Z}_k$ respectively, $\mathbf{e}^{YIQ}$ and $\mathbf{e}^{dsift}$ represent the epitome trained from the YIQ channels and dense sift feature of cI respectively. $0 \le \lambda \le 1$ is a parameter balancing the preference between color and dense sift feature.

5.3 Colorization by Epitome

With the epitome $\hat{\mathbf{e}}$ learned from the reference image, we colorize the target grayscale image gI by inference in the epitome graphical model. Similar to the epitome training process, we densely sample $\hat{Q}$ patches $\{\hat{\mathbf{Z}}_k\}_{k=1}^{\hat{Q}}$ from gI (these patches cover the entire gI). With the hidden mapping associated with patch $\hat{\mathbf{Z}}_k$ denoted as $\hat{\mathcal{T}}_k$, the most probable mapping of the patch $\hat{\mathbf{Z}}_k$, i.e. $\hat{\mathcal{T}}_k^*$, is formulated as follows:

$$\hat{\mathcal{T}}_k^* = \arg\max_{\hat{\mathcal{T}}_k} p\left(\hat{\mathcal{T}}_k|\hat{\mathbf{Z}}_k, \hat{\mathbf{e}}, \boldsymbol{\pi}\right) \tag{5.11}$$

which is essentially the same as the E-step Eq.(5.5). We take the grayscale channel of gI as the luminance channel (Y channel) of itself. Since the color information (IQ channels) is absent in gI, we only use the epitomes corresponding to the Y channel and the dense sift feature to evaluate the right-hand side of Eq.(5.12). The color information is then transferred from the epitome patch, whose location is specified by $\hat{\mathcal{T}}_k^*$, to the grayscale patch $\hat{\mathbf{Z}}_k$. We denote the target image after colorization as gI_c. Since $\{\hat{\mathbf{Z}}_k\}_{k=1}^{\hat{Q}}$ can be overlapping with each other, the final color (the value of IQ channels) of a pixel i in image gI_c is averaged according to:

$$gI_c\left(i\right) = \frac{\sum\limits_{k=1}^{\hat{Q}} \sum\limits_{j \in \hat{S}_k} \delta\left(j = i\right) \hat{\mathbf{e}}_{\hat{\mathcal{T}}_k^*(j)}^{IQ}}{\sum\limits_{k=1}^{\hat{Q}} \sum\limits_{j \in \hat{S}_k} \delta\left(j = i\right)}, \tag{5.12}$$

where $\hat{S}_k$ is the image coordinates of patch $\hat{\mathbf{Z}}_k$, and $\mathbf{e}_{\hat{\mathcal{T}}_k^*(j)}^{IQ}$ represents the value of the IQ channels in the epitome $\mathbf{e}$ at location $\hat{\mathcal{T}}_k^*\left(j\right)$.

5.4 Experimental Results

We show colorization results in this section. As mentioned in section 5.2, we use square patches of size $K \times K$, and the size of epitome is half of the size of the reference image. We densely sample patches with horizontal and vertical gap of ωK pixels, where ω is a parameter between $[0, 1]$ and it controls the number of sampled patches.

Figure 5.3 shows the result of colorization for the dog image. We convert the original image to grayscale as the target image. The patch size is 12×12 and the parameter λ balancing between the color and the dense sift feature is 0.5. We compare our method to [17] which transfers color from the reference image to the target image by pixel-level matching. The result produced by [17] lacks spatial continuity and we observe small artifacts throughout the whole image. On the contrary, our method renders a colorized image very similar to the ground truth. This example also demonstrates that the learned epitome, which is a summary of a large number of sampled patches, contains sufficient color information for colorization.

Figures 5.4 and 5.5 show the colorization result for the nano mushroom-like images and the cheetah. The patch size is chosen as 12 × 12 and 15 × 15, respectively, and λ is set to be 0.8 for both cases. While [17] still generates artifacts around the top and bottom of the mushroom-like structure, while our method produces a much more spatially coherent result. Moreover, we transfer the correct color for the cheetah to the target image, which results in a more natural colorization result than that of [17].

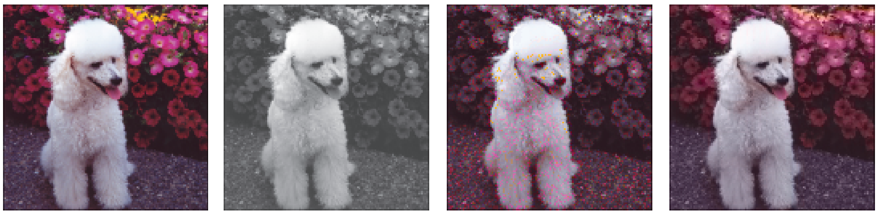

Figure 5.3: The result of colorizing the dog. From left to right: the reference image, the target image (obtained by converting the reference image to the grayscale), the result by [17], and our result.

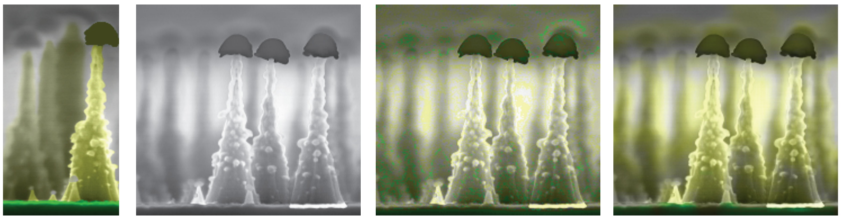

Figure 5.4: The result of colorizing the nano mushroom-like images.

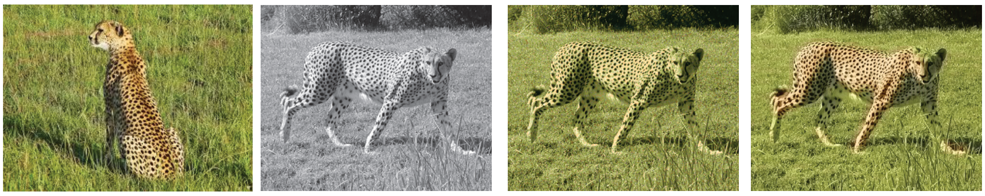

Figure 5.5: The result of colorizing the cheetah.

CHAPTER 6

CONCLUSIONS

In this thesis, we proposed a new graphical model for epitome, i.e. the spatialized epitome. The new epitome model integrates both the local appearance and spatial arrangement for image representation. Employing the powerful generative model framework in both learning and inference, the spatialized epitome is flexible for image representation, discriminative for pattern recognition, adaptive to variation, and robust for object detection. Experiments on several tough vision tasks have shown its superiority over the original epitome model in image modeling. In addition, we present an automatic colorization method using epitome in this thesis. While most existing colorization methods require tedious and time-consuming user intervention for scribbles or segmentation, our epitomic colorization method is automatic. Epitomic colorization exploits the color redundancy by summarizing the color information in the reference image into a condensed image shape and appearance representation. Experimental results shows the effectiveness of our method.

REFERENCES

[1] V. Cheung, B. Frey, and N. Jojic, “Video epitomes,” in *CVPR*, 2005.

[2] N. Jojic, B. Frey, and A. Kannan, “Epitomic analysis of appearance and shape,” in *ICCV*, 2003.

[3] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, “Summarizing visual data using bidirectional similarity,” in *CVPR*, 2008.

[4] A. Kapoor and S. Basu, “The audio epitome: A new representation for modeling and classifying auditory phenomena,” in *ICASSP*, 2005.

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

[6] C. R. Anitha Kannan and J. Winn, “Clustering appearance and shape by learning jigsaws,” in *NIPS 19*. Cambridge, MA: MIT Press., 2006.

[7] K. Ni, A. Kannan, A. Criminisi, and J. Winn, “Epitomic location recognition,” in *CVPR*, 2008.

[8] J. Warrell, S. Prince, and A. Moore, “Epitomized priors for multilabeling problems,” in *CVPR*, 2009.

[9] V. Cheung, N. Jojic, and D. Samaras, “Capturing long-range correlations with patch models,” in *CVPR*, 2007.

[10] H. Wang, Y. Wexler, E. Ofek, and H. Hoppe, “Factoring repeating content within and among images,” in *ACM SIGGRAPH*, 2008.

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

[12] D. Ormoneit and V. Tresp, “Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates,” in *IEEE Trans. on Neuro Networks*, 1998.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. Wiley Interscience, 2001.

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

[15] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, 2004.

[16] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Rendering Techniques*, 2007, pp. 309–320.

[17] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 277–280, 2002.

[18] R. Irony, D. Cohen-Or, and D. Lischinski, "Colorization by example," in *Rendering Techniques*, 2005, pp. 201–210.

[19] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, "Human photoreceptor topography," *Journal of Comparative Neurology*, vol. 292, pp. 497–523, 1990.

[20] N. Jojic, B. J. Frey, and A. Kannan, "Epitomic analysis of appearance and shape," in *ICCV*, 2003, pp. 34–43.

[21] K. Ni, A. Kannan, A. Criminisi, and J. Winn, "Epitomic location recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 12, pp. 2158 –2167, Dec. 2009.

[22] X. Chu, S. Yan, L. Li, K. L. Chan, and T. S. Huang, "Spatialized epitome and its applications," in *CVPR*, 2010, pp. 311–318.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE CVPR*, 2006.