Epitome Representation of Images
and Notes on Epitome-based Silhouette Generation

Shinko Y. Cheng
Electrical and Computer Engineering
University of California, San Diego
La Jolla CA 92037

Abstract

Describing the appearance of images is a task common to many computer vision applications. A novel appearance and shape model called epitomes was recently reported and seems to draw some attention in academia. Epitomes are condensed miniatures of images composed of an arrangement of patches from the original image. The result is a compact representation of the textural and shape essence of an image. This arrangement of patches is found using a variant of the EM algorithm called “variational EM”. The epitome was introduced in the context of generative image models and structuring the image as various constitutive elements layers.

The purpose of this paper is to describe in some detail epitome construction, the use of variational EM and Bayesian networks in epitome estimation and a layered generative models of images using epitomes. The intended audience of this paper is the student who is interested in epitomic image representation and has some knowledge of the EM algorithm and graphical models.

1. Introduction

Much research was done on finding an appearance model that would be discerning enough to distinguish one image from the next, and flexible enough to categorize the image as having a particular kind of content. Jojic, et al. presents Epitomes, a novel summarization of an image generated from multiple patches from an image, combined into a single miniature image [1]. When used within a hierarchical generative model of an image, that model can generate promising image segmentation by virtue of separating regions of the image that contain particular constituent patches, as opposed to other constituent patches.

This paper will describe epitomes, estimation of epitomes, and its significance as an image summarization technique. This paper will attempt to elucidate the derivations epitome estimation using variational EM algorithm. This paper does not attempt to provide a comprehensive experimental evaluation; it will however, show a few results of the epitome and the likely uses of it based on the epitome construction. The intended audience of this paper is the student reader. Specifically, this paper will derive the update equations and describe the key assumptions for estimating epitomes and epitome-based image segmentation. This paper will describe the variational inference technique, which is a variant of the EM algorithm, called variational EM [4] used to simplify the estimation task computationally. The paper will conclude by trying to expose possible issues involved in utilizing epitome representation of images for the purpose of silhouette generation, and show a few results of epitome estimation of images.

2. Epitome Model Explained

A condensed version of an image \( \mathbf{x} \) of size \( N \times M \) is defined as the epitome \( \mathbf{e} \) of size \( n \times m \). The epitome contains the textural and spatial essence of the original and is found by finding an arrangement of several patches of different shapes and sizes from the original image that best fits into the space of the epitome. The patches in the image may overlap, and the arrangement of these patches in the epitome may overlap.

Given the mapping of each of these patches from the epitome to the original and the epitome itself, it is possible to perform image reconstruction. It is desirable to use patches large enough to capture spatial properties of the image, but small enough to generalize across the image and across a collection of images.

The explanation will begin with a description of the epitome model as a miniature image constructed from a pre-specified selection of patches. Then, patch color estimation is incorporated into the inference algorithm itself, although never changing the original patch’s image coordinates. This allows the formulation to be inserted into larger graphical models for performing other tasks using the essence-extracting characteristic of epitomes. An example of a larger graphical model is the generative layered model of an image for separating the image into two layers.
2.1. Epitome as a Generative Model of Image Patches \( \{Z_k\}_{k=1}^P \)

Assume for the moment that an image \( x \) is given in the form of only a set of patches \( \{Z_k\}_{k=1}^P \). The patch shape can be arbitrary, but for the current discussion, we assume the patch shape to be a square of size \( K \times K \). For each patch \( Z_k \), the generative model uses hidden mappings \( T_k \) that map coordinates \( j \in E_k \) in the epitome to the coordinates \( i \in S_k \) in the patch. An epitomic element contains two parameters, the mean and variance of the epitomic pixel. These variables are shown visually in fig 1.

Given the epitome \( e = (\mu, \phi) \) and the mapping \( T_k \), the patch \( Z_k \) is generated by copying the appropriate pixels from the epitome mean and adding Gaussian noise of the level given in the variance map. This can be formulated as the conditional probability

\[
p(Z_k|T_k, e) = \prod_{i \in S_k} \mathcal{N}(z_{i,k}; \mu_{T_k(i)}, \phi_{T_k(i)})
\]

where \( \mathcal{N}(z_{i,k}; \mu_{T_k(i)}, \phi_{T_k(i)}) \) is a Gaussian distribution over \( z_{i,k} \) with mean \( \mu_{T_k(i)} \) and variance \( \phi_{T_k(i)} \).

The graph representation of the epitome definition as a generative model of image patches is shown in fig 2. The diagram depicts the patches \( \{Z_k\}_{k=1}^P \) as independently observed evidence and the epitome and mappings as hidden variables to be inferred.

The number of possible mappings \( T_k \) is assumed to be finite. Because the patches are square, the patch \( Z_k \) can be mapped to the epitome in \( nm \) possible ways. Recall that the location of the patch at the image end is fixed, because the patches are assumed given. Define the image to epitome coordinate mapping as the function \( T_k(i) = i - T_k \). The \( i \)-th pixel in the patch maps to \( T_k(i) \)-th pixel in the epitome, and \( T_k \) is the two-dimensional shift. A square block copy implies that \( i \) range starts at the top-left and ends at the bottom-right of the image patch \( Z_k \), and \( T_k \) range starts at the top-left and ends at the bottom-right of the epitome \( e \). The mappings can otherwise include rotations, scaling and other deformations of patches or be completely arbitrary.

The given patches are assumed to come in multiples of the largest patch to capture large and small patterns in the image. This allows coarse and fine tessellations of the image patches to compete in the inference process. The competition prevents excessive blurring of the epitome, while at the same time allowing grouping of the image features into larger patterns.

The patches are assumed to be generated independently, so the joint distribution is:

\[
p(\{Z_k,T_k\}_{k=1}^P, e) = \prod_{k=1}^P p(Z_k, T_k, e) = \prod_{k=1}^P p(Z_k|T_k, e)p(T_k, e)
\]

\[
= \prod_{k=1}^P p(Z_k|T_k, e)p(T_k)\frac{p(T_k|e)p(e)}{p(e)}
\]

\[
= p(e) \prod_{k=1}^P p(T_k)p(Z_k|T_k, e)
\]

We arrive at equ. (3) and (4) using Bayes rule on conditional probabilities. We assume that the prior on all epitomes is flat and therefore does not appear in the parameter estimation. The prior on the mappings \( p(T_k) \) is used to favor certain mappings over others such as the desire for larger complete patches rather than several smaller components of that patch. The mappings \( p(T_k) \) can be estimated at the same time as the other parameters. It was found by Jojic, et al, that the simple mappings resulted in a tidy arrangement of patches in the epitome nonetheless.
Maximizing negative Helmholtz Free-Energy as a Variant of the EM Algorithm

The parameters $\{T_k\}_{k=1}^P$ and $e$ are estimated using a variant of the Expectation-Maximization Algorithm. Learning the configuration of the parameters involves finding the parameters that maximize the log-likelihood. A direct way involves finding the log-likelihood of the data $L(\theta|\mathbf{v}) = \log \int_h p(v, h|\theta)$. Starting with a non-trivial guess at the configuration of the parameters, the EM step involves taking the expectation of the complete log-likelihood $\log P(\{Z_k, T_k\}_{k=1}^P|\mathbf{e})$ over the hidden variables $\{T_k\}_{k=1}^P$ given the observed data and parameters $\{Z_k\}_{k=1}^P \mathbf{e}$. Then, the M-step follows, maximizing the resulting (non-complete) likelihood function of the data $L(\mathbf{e}|\{Z_k, T_k\}_{k=1}^P) = \log P(\{Z_k\}_{k=1}^P|\{T_k\}_{k=1}^P, \mathbf{e})$. Replace the new configuration of parameters that maximized the likelihood at this iteration with the non-trivial guess of the parameters for the next E-step iteration. The estimated parameters $\{T_k\}_{k=1}^P$ result from the E-step and the estimated epitome $\hat{\mathbf{e}}$ will result in the M-step while.

The EM algorithm is an approximation because while the direct maximum likelihood technique finds parameters that result in the global maximum, the EM algorithm will potentially find the parameters that result in a local maximum of the likelihood. At each iteration, the likelihood with the newer set of parameters was proven to be greater than the likelihood of any previous iteration [6]. That is to say,

$$L(\theta^{(0)}|z) > L(\theta^{(1)}|z) > \ldots > L(\theta^{(t-1)}|z) = L(\theta^{(t)}|z)$$

where $\theta = e$, $z = \{Z_k, T_k\}_{k=1}^P$ in the case of epitome estimation and the superscript is the iteration number. The likelihood is proven to converge to a maximum value.

A variant of the EM algorithm, dubbed variation EM, involves forming the negative Helmholtz free-energy $F(q, \{T_k\}_{k=1}^P, \mathbf{e})$ which is the lower-bound of the log-likelihood. Take the log-likelihood of the data $L(\theta|\mathbf{v}) = \log \int_h p(v, h|\theta)$ and write it in the form of a marginalization of the joint distribution of the observed and hidden random variables given the parameters, we get

$$\log p(v|\theta) = \log \int_h p(v, h|\theta)$$

$$= \log \int_h q(h) \frac{p(v, h|\theta)}{q(h)}$$

$$\geq \int_h q(h) \log \frac{p(v, h|\theta)}{q(h)}$$

$$= -F(q, p)$$

where $h$ are the hidden and $\mathbf{v}$ are the observed random variables, and $\theta$ the set of deterministic parameters. Noting that $\log(\cdot)$ is a concave function in eqn. (6), we can apply Jensen’s inequality to acquire the lower bound.

The negative of the Helmholtz free-energy can be rearranged to become

$$-F(q, p) = \int_h q(h) \log \frac{p(v, h|\theta)}{q(h)}$$

$$= \int_h q(h) \log \frac{p(h|v, \theta)p(v|\theta)}{q(h)}$$

$$= \int_h q(h) \log \frac{p(h|v, \theta)}{q(h)} + \log p(v|\theta)$$

$$= -D(p||q) + L(\theta|\mathbf{v})$$

Worthy of note here is that the Kullback-Liebler divergence of the two densities $D(p||q) = 0$ if $q(h) = p(h|v, \theta) \forall h$, and therefore the negative Helmholtz free-energy equals the log-likelihood of the data.

The standard E-Step of the standard EM algorithm involves computing the posterior distribution $p(h|v, \theta)$ for calculating the expectation of the complete log-likelihood. The corresponding E-Step in the EM algorithm variant is maximizing the negative Helmholtz free-energy $-F(q, p)$ with respect to the $q(h)$ which has its maximum at $q(h) = p(h|v, \theta)$, which means maximizing the negative Helmholtz free-energy results in calculating the posterior distribution. The standard M-Step involves maximizing the log-likelihood of the observed data with respect to $\theta$. The M-Step in the variant of the EM algorithm involves maximizing $F(q, p)$ in terms of the parameters $\theta$ with the $q(h)$ that maximizes $-F(q, p)$ from the last E-Step, which is the same as maximizing the log-likelihood of the data $L(\theta|\mathbf{v})$.

Furthermore, It has been found by Neal and Hinton that if $-F(q, p)$ has a local (or global) maximum at $q^*(h)$ and $\theta^*$, then a local (or global) maximum of $L(\theta|\mathbf{v})$ occurs at $\theta^*$ as well [4]. This implies that we can maximize $-F(q, p)$ over $q(h)$ and $\theta$ instead of instead of maximizing $L(\theta|\mathbf{v})$ over the parameters $\theta$. 3
Maximizing Negative Helmholtz Free-Energy of the Generative Model of the Epitome

Returning to the estimation of epitomes, the first step in using this variant of the EM algorithm is to find the lower bound of the log-likelihood of the data \( \{ Z_k \}_{k=1}^P \):

\[
\log p( \{ Z_k \}_{k=1}^P ) = \log \sum_{\{ T_k \}_{k=1}^P} \int_e p( \{ Z_k, T_k \}_{k=1}^P, e ) \]

\[
= \log \sum_{\{ T_k \}_{k=1}^P} \int_e q( \{ T_k \}_{k=1}^P, e ) \frac{p( \{ Z_k, T_k \}_{k=1}^P, e )}{q( \{ T_k \}_{k=1}^P, e )} \]

\[
\geq \sum_{\{ T_k \}_{k=1}^P} \int_e q( \{ T_k \}_{k=1}^P, e ) \log \left( \frac{p( \{ Z_k, T_k \}_{k=1}^P, e )}{q( \{ T_k \}_{k=1}^P, e )} \right) \tag{10} \]

\[
= B
\]

where \( e \) is the parameter to be estimated. As above, since \( \log(\cdot) \) is a concave function, we can use Jensen’s inequality in eqn. (10) to arrive at eqn. (11) the Helmholtz free-energy expression [4].

We are interested in the estimate of the deterministic parameter \( e \). The deterministic nature of \( e \) is reflected in the formulation with a delta function \( \delta(e - \hat{e}) \). The mappings are described by their distributions \( q( T_k ) \) where any two mappings are independent of each other. The lower bound is subsequently simplified by using the posterior distribution

\[
q( \{ T_k \}_{k=1}^P, e ) = \delta(e - \hat{e}) \prod_i q(T_k) \tag{12}
\]

which results in

\[
\log p( \{ Z_k \}_{k=1}^P ) \geq \sum_{k=1}^P \log p(T_k) - \log q(T_k) \]

\[
+ \sum_{k=1}^P q(T_k) \sum_{i \in S_k} \log N(z_{i,k}; \hat{\mu}_{T_k(i)}, \hat{\phi}_{T_k(i)}). \tag{13}
\]

This lower bound is now easily maximized over \( q(T_k) \) and \( e \). As shown above, maximizing in this way the lower bound or negative Helmholtz free-energy results in the standard expectation-maximization algorithm. The distribution over the mappings \( T_k \) is set the actual posterior distribution

\[
q(T_k) \sim p(T_k) \prod_{i \in S_k} N(z_{i,k}; \hat{\mu}_{T_k(i)}, \hat{\phi}_{T_k(i)}) \tag{14}
\]

normalized \( q(T_k) \) by summing over all allowed mappings. Evaluating \( q(T_k) \) also comprises of the E-step.

In the M-step, we optimize the lower bound \( B \) of the log-likelihood over the epitome parameters \( e = (\mu, \phi) \). To find the value of the epitome mean \( \hat{\mu}_j \) at epitome coordinate \( j \) that locally maximizes the log-likelihood, take the derivative with respect to \( \hat{\mu}_j \) of the log-likelihood and set to zero. The first term on the right hand side of (13) does not depend on \( \hat{\mu}_j \). We introduce a summation to sum the gaussian distributions that are dependent on \( \hat{\mu}_j \), thus allowing the derivative with respect to \( \hat{\mu}_j \) to move inside the summation. The derivative of \( B \) with respect to \( \hat{\mu}_j \) is then

\[
\frac{\partial B}{\partial \hat{\mu}_j} = \frac{\partial}{\partial \hat{\mu}_j} \sum_{k=1}^P \sum_{i \in S_k} \log N(z_{i,k}; \hat{\mu}_{T_k(i)}, \hat{\phi}_{T_k(i)}) \tag{15}
\]

\[
= \sum_{k=1}^P \sum_{i \in S_k} \sum_{T_k(i) = j} q(T_k) \frac{\partial}{\partial \hat{\mu}_j} \log N(z_{i,k}; \hat{\mu}_j, \hat{\phi}_j) \]

\[
= \sum_{k=1}^P \sum_{i \in S_k} \sum_{T_k(i) = j} q(T_k) \frac{\partial}{\partial \hat{\mu}_j} \left\{ -\frac{1}{2} \left( \log 2\pi \hat{\phi}_j - \frac{(z_{i,k} - \hat{\mu}_j)^2}{\hat{\phi}_j} \right) \right\} \]

\[
= \sum_{k=1}^P \sum_{i \in S_k} q(T_k) \frac{(z_{i,k} - \hat{\mu}_j)}{\hat{\phi}_j} \tag{16}
\]

The variance in the denominator drops out after setting the right hand side to zero, since \( \hat{\phi}_j \) is in every term of the triple summation. The \( \hat{\mu}_j \) that maximizes the lower bound of the log-likelihood is therefore

\[
\hat{\mu}_j = \frac{\sum_{k=1}^P \sum_{i \in S_k} \sum_{T_k(i) = j} q(T_k) z_{i,k}}{\sum_{k=1}^P \sum_{i \in S_k} \sum_{T_k(i) = j} q(T_k)} \tag{16}
\]

Because the block-copy mapping is assumed \( (T_k(i) = i - T_k) \) in this discussion, the summation over \( T_k \) such that \( T_k(i) = j \) simply results in only one term \( T_k = i - j \). But this summation is necessary to express the terms in closed form because of the presence of the approximate posterior term \( q(T_k) \). This additional sum accounts for the general case when multiple \( T_k \) may produce \( T_k(i) = j \).

To find the variance of the epitome \( \hat{\phi}_j \) that locally maximizes the log-likelihood, take the derivative with respect to
\( \hat{\phi}_j \) the lower bound of the log-likelihood and set it equal to zero. Like the case for finding the mean, we introduce a sum to account for terms that depend on \( \hat{\phi}_j \) only.

\[
\frac{\partial B}{\partial \hat{\phi}_j} = \sum_{k=1}^{P} \sum_{i \in S_k} \sum_{T_k : T_k(i) = j} q(T_k) \left\{ - \frac{1}{2} \log 2 \pi \hat{\phi}_j - (z_{i,k} - \hat{\mu}_j)^2 \right\}
\]

\[
= \sum_{k=1}^{P} \sum_{i \in S_k} \sum_{T_k : T_k(i) = j} q(T_k) \left\{ - \frac{1}{2 \hat{\phi}_j} + (z_{i,k} - \hat{\mu}_j)^2 \right\}
\]

The \( \hat{\phi}_j \) that maximizes the lower bound of the log-likelihood is therefore

\[
\hat{\phi}_j = \frac{\sum_{k=1}^{P} \sum_{i \in S_k} \sum_{T_k : T_k(i) = j} q(T_k) (z_{i,k} - \hat{\mu}_j)^2}{\sum_{k=1}^{P} \sum_{i \in S_k} \sum_{T_k : T_k(i) = j} q(T_k)}
\] (17)

In the M-step the epitome mean and variance are computed by (16) and (17).

### 2.2 Epitome as a generative model of an entire image

In the previous section, we assumed that the patches of the image \( \{Z_k\}_{k=1}^{P} \) were independently chosen. Because epitome estimation will ultimately be used in the context of a larger generative model, it is necessary to augment the previous model and introduce the process of combining the independently selected patches by averaging the overlapping patch pixels to yield the image pixel. In the inference step, the patch pixels are made to agree.

The conditional probability of the image pixel \( x_i \) given the patches \( \{Z_k\}_{k=1}^{P} \) is found by

\[
p(x_i | \{Z_k\}_{k=1}^{P}) = N(x_i; \frac{1}{N_i} \sum_{k \in S_k} z_{i,k}, \psi_i)
\] (18)

where \( N_i \) is the number of patches that overlap at coordinate \( i \). The entire model now has the joint distribution

\[
p(x, \{Z_k, T_k\}_{k=1}^{P}, e) = \prod_{k=1}^{P} p(x_i | \{Z_k\}_{k=1}^{P})
\]

\[
= p(e) \prod_{k=1}^{P} p(T_k) \prod_{i \in S_k} N(z_{i,k}; \mu_{T_k(i)}, \phi_{T_k(i)})
\]

\[
\cdot \prod_{i} N(x_i; \frac{1}{N_i} \sum_{k \in S_k} z_{i,k}, \psi_i)
\] (19)

where the first term on the right hand side is the joint distribution of the patches, mappings and epitome defined in the previous section. The content of the patches \( Z_k \) are now hidden and have to be inferred from the observed image \( x \). Since we are interested in only the patches in which the overlapping pixels agree, we introduce the following posterior in the inference process.

\[
q(\{z_{i,k}\}, \{T_k\}, e) = \delta(e - \hat{e}) \prod_{k} q(T_k) \prod_{i \in S_k} \delta(z_{i,k} - \zeta_i)
\] (20)

This posterior assumes that the patch pixels that overlap in the image share the same color \( \zeta_i \). This posterior also reflects the treatment of \( \zeta_i \) as a deterministic parameter to be estimated.

Again, employing the variant of the EM algorithm for parameter estimation, we find the lower bound of the log-likelihood of the data, in this case \( x \), to form

\[
\log p(x) = B \geq \int \sum_{\{z_{i,k}\}, \{T_k\}_{k=1}^{P}, e} \int q(\{z_{i,k}\}, \{T_k\}_{k=1}^{P}, e) p(x, \{Z_k, T_k\}_{k=1}^{P}, e) q(\{z_{i,k}\}, \{T_k\}_{k=1}^{P}, e)
\]

\[
\log \frac{p(x, \{Z_k, T_k\}_{k=1}^{P}, e)}{q(\{z_{i,k}\}, \{T_k\}_{k=1}^{P}, e)}
\] (21)

Figure 3: Generative Model of the Epitome of an Image.
which with the chosen posterior reduces to

\[
B = \sum_{k=1}^{p} q(T_k) \left[ \log p(T_k) - \log q(T_k) \right] \\
+ \sum_{k=1}^{p} q(T_k) \sum_{i \in S_k} \log N(\zeta_i; \hat{\mu}_{T_k(i)}, \hat{\sigma}_{T_k(i)}) \\
+ \sum_{i} \log N(x_i; \zeta_i, \psi_i) \tag{23}
\]

Since the image parameter \( x \) is treated as a deterministic parameter, the E-Step consists the same calculation as in the case of a generative model of only patches, calculating the posterior distribution

\[
q(T_k) \sim p(T_k) \prod_{i \in S_k} N(\zeta_i; \hat{\mu}_{T_k(i)}, \hat{\sigma}_{T_k(i)}) \tag{24}
\]

Taking the derivative of the lower bound of the log-likelihood with respect to \( \hat{\phi}_j, \hat{\mu}_j, \) and \( \zeta_i \) and setting it to zero will result in the following three update equations for the M-Step. The update equation for the color of the patch pixel \( \zeta_i \) is given by

\[
\hat{\zeta}_i = \frac{x_i}{\psi_i} + \frac{\sum_{k=1}^{p} \sum_{u \in S_k \cup u \neq i} q(T_k) \frac{\nu_{T_k(u)}}{\psi_{T_k(u)}}}{\frac{1}{\psi_i} + \sum_{k=1}^{p} \sum_{u \in S_k \cup u \neq i} q(T_k) \frac{1}{\psi_{T_k(u)}}} \tag{25}
\]

where \( u \) is in epitome coordinates and \( i \) is in image coordinates and the expression \( u : u \rightarrow i \) means the patch coordinate \( u \) that maps to the image coordinate \( i \).

The update equations for the mean and variance of the epitome are given by

\[
\hat{\phi}_j = \frac{\sum \sum q(T_k)(\zeta_i - \mu_j)^2}{\sum \sum q(T_k)} \tag{26}
\]

\[
\hat{\mu}_j = \frac{\sum q(T_k)(\zeta_i)}{\sum q(T_k)} \tag{27}
\]

where \( i \) is in patch coordinates and \( j \) is in epitome coordinates.

### 2.3. Shape Epitome and Image Segmentation

In this section, the example of epitomes used as a module in a generative model of an image in the paper by Jojic, et al. is explained. There, epitomes are used to model the texture and color of the appearance and shape of an image.

Fig 4 shows a Bayesian network of this generative model, illustrating the dependency relationships between random variables.

The generative model of overlapping objects consists of one epitome \( e_s \) to model the layer appearances \( s_1, s_2 \) and another epitome \( e_m \), to model the mask image \( m \). The complete image \( x \) is composed of the layer equation

\[
x = s_1 \circ m + (1 - m) \circ s_2 + n
\]

where \( n \) is gaussian distributed noise and \( \circ \) is element-wise multiplication. This produces the following conditional probability

\[
p(x|s_1, s_2, m) = N(x; s_1 \circ m + (1 - m) \circ s_2, \Psi)
\]

### Bayesian Networks and d-connectivity

Recall, the conditions for d-connectivity between nodes in a Bayesian network: A random variable \( a \) is dependent on random variable \( b \) given evidence \( E = \{ e_1, e_2, ..., e_n \} \) if there is a d-connecting path from \( a \) to \( b \) given \( E \). A d-connecting path is defined as a path from nodes \( a \) to \( b \) with respect to the evidence nodes \( E \) if every interior node \( n \) in the path has the property that either

1. it is linear or diverging and not a member of \( E \).
2. or it is converging, and either \( n \) or one of its descendants is in \( E \).

There are three basic kinds of paths between two nodes. They are linear, converging and diverging paths. An example of a linear path between \( a \) to \( b \) is

\[
a \rightarrow n \rightarrow b.
\]

An example of a a diverging path between the same two nodes is

\[
a \rightarrow n \leftarrow b.
\]
For a converging one we would reverse the arrows for
\[ a \leftarrow n \rightarrow b. \]

So, if random variables \( a \) and \( b \) given evidence \( E \) were d-separated (not d-connected) by the conditions named above, then \( a \) and \( b \) given \( E \) are independent and we may write
\[ p(a, b|E) = p(a|E)p(b|E) \]

For more details on the dependence relationships between nodes in a Bayesian network are in [5].

**Estimating Image Layers, Appearance and Shape Epitomes**

With the notion of d-connectivity in mind, the joint distribution can be written in factored form by observing the dependence structure implied in the Bayesian network.

\[
p(x, s_1, s_2, m, e_s, e_m) \\
= p(x, e_s, e_m|s_1, s_2, m)p(s_1, s_2, m) \\
= p(x|s_1, s_2, m)p(e_s, e_m|s_1, s_2, m)p(s_1, s_2, m) \\
= p(x|s_1, s_2, m)p(s_1, s_2, e_s)p(s_2, e_m)p(m|e_m)p(e_m) \\
= p(x|s_1, s_2, m)p(s_1|e_s)p(s_2|e_s)p(m|e_m)p(e_s)p(e_m) \\
= p(x|s_1, s_2, m, e_s)p(s_2|e_s)p(m|e_m)p(e_s)p(e_m) \tag{34}
\]

Bayes rule allows us to separate the joint probability to get eqn (28). For the first term in eqn (29), \( s_1, s_1, s_2, \) and \( m \) “block” the paths existing as evidence between \( x \) and \( e_s, e_m \), violating the first condition for d-connectivity which says interior nodes for linear paths should not be in \( E \) for d-connectivity. Therefore, we can separate the conditional probability in terms of the product of the conditional probability of the image \( x \) and the epitomes \( e_s, e_m \) given the layers \( s_1, s_2, \) and \( m \). We can combine the probabilities of the epitome and layers again using Bayes rule in eqn (30). Because the five nodes (random variables) are disconnected without the image node, \( e_s, s_1, \) and \( s_2 \) as a group are independent from \( e_m \) and \( m \). In eqn (31), \( s_1 \) and \( s_2 \) given \( e_s \) is d-separated by both conditions above therefore independent. Two paths exist between \( s_1 \) and \( s_2 \) in this situation, a diverging path through \( e_s \) and a converging path through \( x \). The diverging path violates the first d-connectivity condition since the interior node, \( e_s \) is evidence. The converging path violates the second condition since \( x \) is not evidence and \( x \) does not have any descendants.

This joint distribution is then marginalized to attain the Helmholtz free-energy lower bound of the log-likelihood of the data \( x \) and optimized. Again employing the variant of the EM algorithm and fixing the layers and epitomes as deterministic parameters of a family of distributions by assuming the posterior
\[
q = \delta(s_1 - \hat{s}_1)\delta(s_2 - \hat{s}_2)\delta(m - \hat{m})\delta(e_s - \hat{e}_s)\delta(e_m - \hat{e}_m) \tag{35}
\]

Therefore,
\[
\log p(x) \geq B = \\
\int_{s_1, s_2, m} q \log \frac{p(x, s_1, s_2, m, e_s, e_m)}{q(s_1)q(s_2)q(m)} \\
= \log p(x|s_1, s_2, \hat{m}) + \log p(s_1|e_s) + \log p(s_2|e_s) + \log p(m|e_m) + \text{const}
\]

The first term in the sum is quadratic in the three images \( s_1, s_2, \) and \( \hat{m} \).
\[
\log p(x|s_1, s_2, \hat{m}) = B_x \\
= \sum_i \log N(x_i; \zeta_k^{s_1}, \zeta_k^{s_2}, \psi_i)
\]

The remaining terms resemble the log-likelihood functions for the generative epitome model given entire images.
\[
\log p(s_1|e_s) = B_i \\
= \sum_{k=1}^{P} q(T_k^{s_1}) \left[ \log p(T_k^{s_1}) - \log q(T_k^{s_1}) \right] \\
+ \sum_{k=1}^{P} q(T_k^{s_1}) \sum_{i \in S_k} \log N(\zeta_i^{s_1}; \hat{\mu}_{T_k^{s_1}}^{s_1}, \hat{\phi}_{T_k^{s_1}}^{s_1})
\]

\[
\log p(s_2|e_s) = B_i \\
= \sum_{k=1}^{P} q(T_k^{s_1}) \left[ \log p(T_k^{s_1}) - \log q(T_k^{s_1}) \right] \\
+ \sum_{k=1}^{P} q(T_k^{s_1}) \sum_{i \in S_k} \log N(\zeta_i^{s_1}; \hat{\mu}_{T_k^{s_1}}^{s_1}, \hat{\phi}_{T_k^{s_1}}^{s_1})
\]

\[
\log p(m|e_m) = B_m \\
= \sum_{k=1}^{P} q(T_k^{s_1}) \left[ \log p(T_k^{s_1}) - \log q(T_k^{s_1}) \right] \\
+ \sum_{k=1}^{P} q(T_k^{s_1}) \sum_{i \in S_k} \log N(\zeta_i^{s_1}; \hat{\mu}_{T_k^{s_1}}^{s_1}, \hat{\phi}_{T_k^{s_1}}^{s_1})
\]

\[
B = B_x + B_1 + B_2 + B_m
\]

For the E-Step, as in the case in sec. 2.2 but for every layer and its epitome, calculate the posterior distribution
\[
q(T_k^{s_1}) \sim p(T_k^{s_1}) \prod_{i \in S_k} N(z_i; k; \mu_{T_k(i)}, \phi_{T_k(i)}) \tag{36}
\]

normalized by the possible mappings \( T_k \)
The M-Step consists of a set of bilinear update equations are iterated over until convergence. These update equations are found by taking the derivative of Helmholtz free-energy of the generative layered image model $B$ with respect to $\hat{\mu}_k^s$, $\hat{\phi}_j^s$, $\hat{\phi}_j^m$, $\hat{\zeta}_e^s$, $\hat{\zeta}_m^s$ and $\hat{\zeta}_m^m$. The update equations are given in table 1 and 2.

3. Examples and Discussion

The epitomes of some sample images are estimated and the results are shown in fig. 5. This code that created these examples is obtainable at http://research.microsoft.com/ jojic/software.htm

The algorithm is initialized by randomly selecting $T$ $K \times K$ patches over the entire image. The epitome mean is initialized to the average pixel value over the entire image plus some gaussian noise with variance equal to the amount of variance over the pixels in the image. The epitome variance is set to all ones. The patch mapping $T_k(i) = i - T_k$ is defined on a torus. In other words, the patch destination coordinates in the epitome are mod $K$. Eqn. 16 and 17 are iterated through a series of convolutions of the approximate posterior probability map and the patch or epitome.

Because the algorithm starts by randomly choosing frames and randomly choosing positions in epitome to place the patches, each epitome estimate will be different even for the same image the position of the patches. The constituting parts in the epitome will be similar. This illustrates the multiple local minima that is produced by the EM algorithm.

Epitomes of the image may be used as a starting point to estimate the epitomes of other images or the same image. An example of epitomes of a sequence of images in a video, where the previous epitome is used as a starting point for estimation of the new epitome is shown in fig. 6. Here we can see that the sequence of epitomes are similar to one another, even though patches were randomly chosen from the image in estimation.

Most note-worthy is the use of variation EM in epitome estimation. The variant of the EM algorithm has permitted the estimation of $e$ without the need to find the exact posterior probability $\rho(\{T_k\}_{k=1}^P, e|\{Z_k\}_{k=1}^P)$ for the standard E-Step which requires finding the marginal $P(\{Z_k\}_{k=1}^P)$ by summing over all possible patch mappings for each patch – a non-trivial task. Instead, variational EM has allowed us to use an approximate posterior, the only requirement being that it is proportional to the complete log-likelihood and sum over all $T_k$ equals 1 [4].

Parameter estimation is performed by forming the joint distribution of the system of the patches and mappings for the epitome or layered generative model of the image, and marginalizing it to attain the lower bound of the log-likelihood of the data. This lower bound is then maximized over the parameters $e$ and the approximate posterior constructed using the complete log-likelihood and normalized by the sum of the function over the hidden variable. The prior on the mappings can have a density that favors larger mappings, but can otherwise be flat. Assuming the densities of the epitome and the color of the patch pixels reflects the manner in which point estimates are found in general parameter estimation problems, assuming the parameters to be deterministic and be of only one configuration.

Of critical concern is the size of the patches. Epitomes of images will be different given the maximum and minimum patch sizes specified. When the patch size is limited to only 1 pixel in size, the epitome and the mappings basically describe the average color and the frequency of each of these colors used in the image. This epitome is then equivalent to the color histogram and spatial arrangement within the epitome means little. When the patch size is as large as the epitome, then the epitome is like an averaging of several regions of the image aligned such that the variance map has small values.

Ideally, the epitome would be generated with the patch size as large as the one of the features in a repeating pattern of features in the image. The ability to have several patch sizes compete with each other is built into the epitome estimation algorithm by allowing some patches to survive while others fade away.

4. Conclusion

Epitomes are a novel appearance model composed of patches from an image. The epitome of an image are found using a variant of the EM algorithm. We have shown most of the steps in deriving the update equations for the epitome as a generative model of image patches, of entire image, and for the image as a generative model of image layers and epitomes.

Incorporating epitomes into a larger generative models, the appearance extracting ability can be used for many other purposes, and deriving update equations using the variational EM inference framework can be relatively straightforward.

References


Table 1: Update equations for epitome-based layered generative model of an Image.

\[
\hat{\mu}_j^s = \frac{\sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{s1}(i) = j) \sum_{i \in S_k} q(T_{k}^{s2}(i) = j) \hat{\zeta}_i^s + \sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{s2}(i) = j) \hat{\zeta}_i^s}{\sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{s1}(i) = j) + \sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{s2}(i) = j)} \tag{37}
\]

\[
\hat{\phi}_j^s = \frac{\sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{s1}(i) = j) \sum_{i \in S_k} q(T_{k}^{s2}(i) = j) (\hat{\zeta}_i^s - \hat{\mu}_j^s)^2 + \sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{s2}(i) = j) (\hat{\zeta}_i^s - \hat{\mu}_j^s)^2}{\sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{s1}(i) = j) + \sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{s2}(i) = j)} \tag{38}
\]

\[
\hat{\mu}_j^m = \frac{\sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{m}(i) = j) \hat{\zeta}_i^m}{\sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{m}(i) = j)} \tag{39}
\]

\[
\hat{\phi}_j^m = \frac{\sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{m}(i) = j) \sum_{i \in S_k} q(T_{k}^{m}(i) = j) (\hat{\zeta}_i^m - \hat{\mu}_j^m)^2}{\sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{m}(i) = j) + \sum_{k=1}^{P} \sum_{i \in S_k} q(T_{k}^{m}(i) = j)} \tag{40}
\]
Table 2: Update equations for epitome-based layered generative model of an Image, continued...

\[ \hat{\zeta}_i^{s1} = \frac{x_i \hat{\zeta}_i^m}{2\psi_i} + \sum_{k=1}^{P} \sum_{u \in S_k} \frac{\hat{\mu}_{T_k^{s1}}(u)}{\phi_{T_k^{s1}}(u)} \]

\[ \frac{(2\hat{\zeta}_i^m - 1)\hat{\zeta}_i^m}{2\psi_i} + \sum_{k=1}^{P} \sum_{u \in S_k} \frac{1}{\phi_{T_k^{s1}}(u)} \]  

\[ \hat{\zeta}_i^{s2} = \frac{-x_i (1 - \hat{\zeta}_i^m)}{2\psi_i} + \sum_{k=1}^{P} \sum_{u \in S_k} \frac{\hat{\mu}_{T_k^{s2}}(u)}{\phi_{T_k^{s2}}(u)} \]

\[ \frac{-\hat{\zeta}_i^m (1 - \hat{\zeta}_i^m) + (1 - \hat{\zeta}_i^m)^2}{2\psi_i} + \sum_{k=1}^{P} \sum_{u \in S_k} \frac{1}{\phi_{T_k^{s2}}(i)} \]  

\[ \hat{\zeta}_i = \frac{(\hat{\zeta}_i^{s2} - x_i)(\hat{\zeta}_i^{s1} + \hat{\zeta}_i^{s2})}{2\psi_i} + \sum_{k=1}^{P} \sum_{u \in S_k} \frac{\hat{\mu}_{T_k^{m}}(i)}{\phi_{T_k^{m}}(i)} \]

\[ \frac{- (\hat{\zeta}_i^{s1} + \hat{\zeta}_i^{s2})^2}{2\psi_i} + \sum_{k=1}^{P} \sum_{u \in S_k} \frac{1}{\phi_{T_k^{m}}(i)} \]  

(41)

(42)

(43)
Figure 5: Sample epitomes of images. These $50 \times 50$ epitomes are generated from 100 $20 \times 20$ patches over 10 iterations of EM. Convergence was achieved with $\sim 4$ iterations.

Figure 6: Sample epitomes of a sequence of images in a video. These $30 \times 30$ epitomes are generated from 100 $20 \times 20$ patches over 4 iterations of EM. Convergence was achieved with $\sim 4$ iterations. Note the similarity of the epitomes from one frame to the next.
A. Original Project Objectives and Milestones

The primary task is to explicate the definition of the appearance epitome of an image and the derived algorithm for epitome estimation. Namely, this paper’s intent is to explain the process of learning the collection of image patches and patch-to-image mappings in the epitome estimation. The secondary task is to utilize the epitome representation in a layered generative image model for silhouette generation, a form of image segmentation. And the intent is to expose the issues involved in this utility. It would then be qualitatively compared with a background subtraction segmentation technique.

A milestones chart for the project is shown in table 3. The explication of the proof of the epitome estimation algorithm is complete. The derivations of the segmentation algorithm are also complete. Only the estimation of the epitome as a generative model of image patches are experimentally verified with the provided code. The results of them are given in the main text.

A.1. Target Questions

The questions the resulting paper will try to address are general questions about inference algorithms and statistical generative models, including what is the appearance and shape epitome? What are generative models and unsupervised variational inference algorithms (EM algorithm)? How do a generative model and an unsupervised variational inference algorithm allow extraction of the epitome, segmentation of the foreground and background, and filling of the occluded parts with the appearance predicted from the epitome?

The paper will also attempt to address specific issues of the algorithms pertaining to the epitome construction, such as how does the extreme and intermediate values of K (patch size), N(epitome size), I (EM iterations) affect the epitome result and its usefulness?

With regards to epitomes and segmentation, the paper will try to understand what is required to perform segmentation an image of a person into a person silhouette? What are the constraints and parameters? Is it effective? What is the rate of computation and how may it be improved? What advantages does this hold over conventional background subtraction segmentation techniques?

A.2. Existing Code and Test Images

Experimental data for image segmentation will be gathered from a digital camera available to me to capture the various situations, such as orange-colored-clothed person in a green room for high contrast example, a person in orange in a
### Milestone Table

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Date</th>
<th>Description of work to be completed</th>
</tr>
</thead>
</table>
| 1         | Oct 28  | - Finalize hypotheses and tasks for the project  
- Read up on Generative Models, Unsupervised Variation Inference Algorithm (M.I. Jordan)  
- *Complete final project proposal* |
| 2         | Nov 11  | - Explicate the proof of the estimation of parameters of the epitome as a generative model of image patches and entire images algorithm. Understand d-connectivity of Bayesian networks |
| 3         | Nov 18  | - Understand the notion of Helmholtz Free Energy and its relationship to EM algorithm.  
- *Epitome Presentation* |
| 4         | Nov 25  | - Devise the steps needed for image segmentation and implement it in Matlab if time permits  
- *Complete polished draft of project report* |
| 5         | Dec 4   | - Complete as much as possible programming of segmentation algorithm. Capture and apply algorithm to a series of test sequences for qualitative analysis.  
- *Submit Final Draft of project report* |

Table 3: Milestones Table

varied background for something more realistic, and finally a plain clothed person in a varied background.

Matlab code for epitome generation and epitome coloring for segmentation is provided at http://www.research.microsoft.com/jojic/software.htm. This software includes the epitome generating algorithm as a generative model from patches, not images, described in Section 2.1 of the paper by Jojic.

### A.3. Previous Work on Epitomes

The parameter estimation technique used in the paper by Jojic [1] for epitomes is described in a general context in [3]. [1] was referred to specifically in Section 2.1 of the epitome paper describing “parameters estimation by marginalizing the joint distribution and optimizing the log likelihood of the data using the approximate posterior to compute the lower bound on the log likelihood.”

The paper that explains an unsupervised variational inference algorithm used to jointly extract the epitome, segments the foreground from the background and fills the occluded parts with the appearance predicted from the epitome is described in [4].