Capturing long-range correlations with patch models

Vincent Cheung
Elec. & Comp. Eng.
University of Toronto
vincent@psi.toronto.edu

Nebojsa Jojic
Machine Learning & Applied Stat.
Microsoft Research
jojic@microsoft.com

Dimitris Samaras
Computer Science Dept.
Stony Brook University
samaras@cs.sunysb.edu

Abstract

The use of image patches to capture local correlations between pixels has been growing in popularity for use in various low-level vision tasks. There is a trade-off between using larger patches to obtain additional high-order statistics and smaller patches to capture only the elemental features of the image. Previous work has leveraged short-range correlations between patches that share pixel values for use in patch matching. In this paper, long-range correlations between patches are introduced, where relations between patches that do not necessarily share pixels are learnt. Such correlations arise as an inherent property of the data itself. These long-range patch correlations are shown to be particularly important for video sequences where the patches have an additional time dimension, with correlation links in both space and time. We illustrate the power of our model on tasks such as multiple object registration and detection and missing data interpolation, including a difficult task of photograph relighting, where a single photograph is assumed to be the only observed part of a 3D volume whose two coordinates are the image x and y coordinates and the third coordinate is the illumination angle θ. We show that in some cases, the long-range correlations observed among the mappings of different volume patches in a small training set are sufficient to infer the possible complex intensity changes in a new photograph due to illumination angle variation.

1. Introduction

Patches have been used to capture local correlations between pixels in various low-level vision tasks, with perhaps the most notable early example in [6]. To capture correlations that span a longer range, larger patches can be used, though this has adverse side-effects, including for example, increased difficulty in patch matching. When multiple patches from different images are matched there are many correlations between the mapping pairs. If two patches match in two images, then it is likely that shifting both patches one pixel in the same direction also leads to a matched pair, since there is a significant amount of overlap between the pixels in the patches. Such local coherence ideas have been used in [1, 13]. In case of videos and 3D patches, local correlations are even stronger because of the added time dimension.

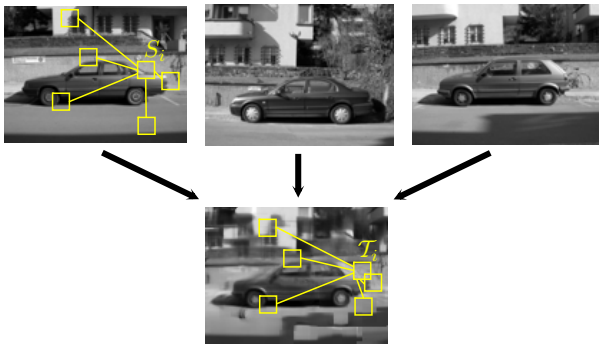

Figure 1. The epitome of three car images learnt using long-range patch correlations. The epitome ends up being a merge of the three cars, with both the front and back of the car reaching a compromise between the different shapes of the cars. Also interesting to note is how the epitome merges the three backgrounds. The patch indicated by $\mathcal{S}_i$ in the top-left image is connected to several other randomly chosen patches in the image, to which, relative patch distances should be generally maintained during patch matching to the epitome. The corresponding patch in the epitome is shown as $\mathcal{T}_i$ and the matching is constrained by the matches for the patches connected to $\mathcal{S}_i$. A video illustrating this patch correspondence during learning is available at http://www.psi.toronto.edu/∼vincent/patchcorr.html.

Visual data also exhibits strong long range correlations in images which can relate patches that do not have any pixels in common. Elastic matching, *e.g.* [2], has been used in the past to register image pairs by leveraging the fact that mappings are generally smooth between images to overcome erroneous correspondences due to noise or lack of identifying features. By reducing the analysis to a subset of patches with relatively high mutual distances, it is possible to produce elastic matching of large structures using a small number of image features. This is achieved by assuming that relative offsets of image features are only slightly perturbed between two images. Relative positions of features are also useful for object recognition. For example, in constellation models [3, 12], the relative locations of a small number of detected features from an image are used to facilitate object recognition. Elastic image matching, which only some of

its many forms have been mentioned here, has been one of the most used tools in vision.

In this paper, we are concerned with the use of similar elastic constraints, but with the goal of modeling correlations among the mappings of all data patches to a common learned representation of a category of images (Fig. 1), *e.g.* an epitome [4, 8, 10]. Thus, the model we propose captures the full probability distribution of the data, making it possible to mine the long-range image correlations in various inference tasks, including data registration, data likelihood computation (for tasks such as classification or detection), and missing data interpolation.

For example, one of the tasks we can perform using inference in our model is the simulation of illumination changes on an object in a single photograph. The illumination training data consists of video sequences of other static objects and varying illumination angles, and the patches mapped to a common epitome are three-dimensional. Our model estimates the appearance and mapping constraints through space and time among the video cubes in the training data and then estimates a video sequence which satisfies these constraints and whose central frame is equal to the given photograph. This creates plausible illumination changes on the object in the photograph.

Previously, such image relighting tasks typically required an expensive, brute force, hardware solution as in [5], where the subject sits in a dome and photos of the subject are taken from many different angles, from which any illumination can be reconstructed by taking combinations of these images. There are several limitations to this approach including the a priori knowledge of the desired illumination change, so deceased individuals cannot be re-lit; the subject must remain still and be tolerant to strobe lights; and the subject must fit in the dome, so entire scenes cannot be re-lit. Less hardware-dependent and more computation-oriented alternatives to re-lighting an image or video sequence have also proposed. For instance, generic face surface geometry and reflectance models have been used for relighting faces [11, 14]. However, once the problem changes to re-lighting something other than a human face, such as an animal, a piece of cloth, or an entire scene, these approaches become more difficult to follow, as they need object-specific surface geometry models. With the exception of a small number of object categories (perhaps only human faces), such models are fairly rare. The richness of the 3-D face modeling literature is the best indication of the difficulty of acquiring such models. The examples of relighting we show in this paper are example-based - given a small number of examples (sometimes even just one) of how the image of an object changes with smooth variation of illumination angles, we can construct plausible similar changes on another similar object. This removes the need for full modeling of the surface geometry of the objects. Instead, the correlations in the patches form the training data provide sufficient constraints to infer plausible image changes due to illumination angle changes.

In addition to face and cloth photograph-relighting, we show results on simulating a walk through a hallway given one photograph of a hallway, and learning epitomes of cars and faces, all using the same trainable model of data patches.

2. Flexible patch configurations

As discussed in the introduction, the issue of varying geometric configurations of object features has repeatedly been encountered in vision research. In this paper we are particularly concerned with how this variability can be accounted for in patch models that describe learnable probability density functions of images. In particular, as described in Fig. 1, we construct an epitome model in which patches from different locations in the image have correlated mappings to the epitome locations. While the discussion in this section is limited to 2-D images for concreteness, it is trivial to extend these ideas to N-D structures.

2.1. Review of epitome models

The original epitome model [8] proposes that a set of pixels from image z with indices in the set $\mathcal{S}$, *i.e.*, the set $z_{\mathcal{S}} = \{z_{\mathbf{u}} | \mathbf{u} \in \mathcal{S}\}$[1], can be described by specific individual probability distributions taken from epitome (e) locations in the set $\mathcal{T}$:

$$p(z_{\mathcal{S}}|e_{\mathcal{T}}) = \prod_k p(z_{\mathcal{S}(k)}|e_{\mathcal{T}(k)}), \tag{1}$$

or simply,

$$p(z_{\mathcal{S}}|e_{\mathcal{T}}) = \prod_k e_{\mathcal{T}(k)}(z_{\mathcal{S}(k)}), \tag{2}$$

where it is assumed that the sets $\mathcal{S}$ and $\mathcal{T}$ are ordered and of equal sizes, and the k-th index in one set corresponds to the k-th index in the other. Given a number of these correspondences between different subsets of pixels in training images $\mathcal{S}_i$ and subsets of epitome locations $\mathcal{T}_i$, learning an optimal epitome reduces to assembling the required sufficient statistics. For example, if the distributions at each epitome location $e_{\mathbf{u}}$ are Gaussians, $p(z_{\mathbf{v}}|e_{\mathbf{u}}) = e_{\mathbf{u}}(z_{\mathbf{v}}) = \mathcal{N}(z_{\mathbf{v}}; \mu_{\mathbf{u}}, \sigma^2_{\mathbf{u}})$, then the mean $\mu_{\mathbf{u}}$ of the Gaussian at epitome location $\mathbf{u}$ is simply equal to the average of all image pixels that map there,

$$\mu_{\mathbf{u}} = \frac{\sum_i \sum_k [\mathbf{u} = \mathcal{T}_i(k)] z_{\mathcal{S}_i(k)}}{\sum_i \sum_k [\mathbf{u} = \mathcal{T}_i(k)]}, \tag{3}$$

where [ ] is Iverson's indicator function, *i.e.*, $[true] = 1$, $[false] = 0$. When the correspondences are not given, but the nature of these correspondences is described so as to limit the possibilities[2], the mapping for each set $\mathcal{S}$ can be

[1] Boldcase $\mathbf{u}$ and $\mathbf{v}$ represent 2D indices describing image coordinates, *i.e.*, $\mathbf{u} = (x, y)$

[2] For example, one way to limit the space of allowed correspondence is to consider subsets $\mathcal{S}_i$ in the data that are rectangular patches of a certain size, *i.e.*, $\mathcal{S}_i = \{\mathbf{u} = (x, y) | X_i \leq x < X_i + \delta, Y_i \leq y < Y_i + \delta\}$ and the corresponding epitome subsets $\mathcal{T}$ are defined to also be rectangular patches starting at some epitome location X_j, Y_j.

inferred using an early estimate of the epitome, which leads to soft posterior mapping of image subsets $\mathcal{S}_i$ to the corresponding epitome subsets $\mathcal{T}_i$,

$$q(\mathcal{T}_i = \mathcal{T}) = p(\mathcal{T}_i = \mathcal{T}|\mathcal{S}_i, z) \propto p(z_{\mathcal{S}_i}|e_{\mathcal{T}})p(\mathcal{T}), \quad (4)$$

where $p(\mathcal{T})$ is the a priori probability that epitome patch $\mathcal{T}$ is used to describe any of the data, and the posterior distribution is established by normalizing the above expression over all possible sets $\mathcal{T}$.

The epitome is then re-estimated using this soft mapping. For example, in case of Gaussian epitome entries, the means are estimated as weighted average of *all* pixels, with the weights defined by mapping probabilities. Since each set of image coordinates $\mathcal{S}_i$ may map to any set of epitome coordinates $\mathcal{T}$, with probability $q(\mathcal{T}_i = \mathcal{T})$, the sufficient statistics reflect this by weighting with these probabilities [8],

$$\mu_{\mathbf{u}} = \frac{\sum_i \sum_{\mathcal{T}} q(\mathcal{T}_i = \mathcal{T}) \sum_k [\mathbf{u} = \mathcal{T}(k)] z_{\mathcal{S}_i(k)}}{\sum_i \sum_{\mathcal{T}} q(\mathcal{T}_i = \mathcal{T}) \sum_k [\mathbf{u} = \mathcal{T}(k)]}. \quad (5)$$

The variance $\sigma^2_{\mathbf{u}}$ at each location is estimated in a similar fashion [8].

Iterating mapping inference and epitome re-estimation leads to joint epitome learning and data registration [8]. The learning procedure quilts and averages patches from various locations from one or more images to create a compact model of all patches, similar to [6]. The model can easily be used for ordered data of different dimensionalities than two, *e.g.* 3D epitomes were used to model videos [4], and 1D epitomes were used to estimate an HIV vaccine [9].

The rules of establishing pixel correspondence (choosing various image locations $\mathcal{S}$ and their corresponding epitome locations $\mathcal{T}$) are left general in these early papers, although particular applications usually considered regular small patches of image pixels to form various sets $\mathcal{S}_i$, and the same size patches in the epitome. This made the search for optimal mapping of each image patch linear in the size of the epitome, as effectively, only the position of the epitome patch is required to fully describe the mapping regardless of the patch size. This choice also limited the spatial extent in which image correlations are nicely captured by the epitome to several patch sizes. Due to the overlap of patches both in the input image(s) and in the epitome, the textures that form in the epitome upon learning capture structures larger than the patch sizes, but often much smaller than the object size.

The basic formulation of the model allows the pixel coordinates in $\mathcal{S}_i$ to come from disconnected parts of the image, and the mapping rules that limit the space of possible sets of epitome coordinates $\mathcal{T}$ to include rotation, shearing, and other transformations. This would allow capturing more complex geometric changes that span a larger spatial extent in the image. To the best of our knowledge, while the inclusion of more sophisticated geometric transformations has been studied before, the use of non-contiguous patches has not been investigated due to the explosion of the numbers of possible image subsets $\mathcal{S}_i$ to be considered. Recently, patches of arbitrary (and inferred) shape have been used in

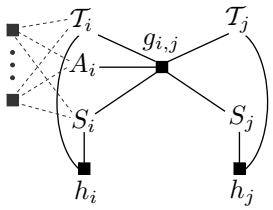

Figure 2. Factor graph of the long-range correlations patch model.

epitome structures dubbed jigsaws [10], but these patches are still contiguous and do not capture global correlations in images. Without directly capturing longer range correlations in the data, be it images, videos, or other ordered datasets, the epitome models will fail to capture global scale phenomena of the objects they were trained on.

To resolve this problem, instead of using non-contiguous patches to capture within each single mapping, $\mathcal{S} \rightarrow \mathcal{T}$, the correlations in distant parts of the image, we propose to model correlations among *different* mappings, $\mathcal{S}_i \rightarrow \mathcal{T}_i$. This allows us to capture long-range correlations in the image while still having relatively simple individual patches and mappings.

2.2. The mapping field

The use of simple rectangular patches to represent data has significant computational advantages, especially for higher dimensional data, as discussed, for example, in [4]. Rectangular patches allow the use of fast Fourier transform tricks and efficient image correlation computations necessary to efficiently perform otherwise very expensive computations. Smaller patches of other shapes can be simulated using the masking variables [8], or, with a higher computational cost, but some other benefits, using jigsaw models [10]. Different patches of data coordinates $\mathcal{S}_i$ have the associated mapped epitome coordinates $\mathcal{T}_i$. The original epitome model assumed independence of variables $\mathcal{T}_i$, as the patch overlap naturally enforced the appropriate agreements in mappings of nearby patches. Similar local agreement is enforced in the jigsaw model in a way that allows patches to be arbitrarily shaped.

In the model we propose in this paper, we capture the constraints on the mappings $\mathcal{T}_i$ and $\mathcal{T}_j$ of *distant* patches $\mathcal{S}_i$ and $\mathcal{S}_j$ through agreement factors $g_{i,j} = g(\mathcal{T}_i, \mathcal{T}_j, \mathcal{S}_i, \mathcal{S}_j)$ (Fig. 2), which have high value if the mappings $\mathcal{T}_i$, $\mathcal{T}_j$ keep a similar geometric configuration as $\mathcal{S}_i$, $\mathcal{S}_j$. The factors h_i correspond to the usual epitome likelihoods *e.g.* $h_i = e_{\mathcal{T}_i}(\mathcal{S}_i)$. Intuitively, this is represented in Fig. 1, where the patches connected to $\mathcal{S}_i$ constrain its matching to $\mathcal{T}_i$ as it is desirable not only for the two patches to be similar, but also to maintain the relative locations of the matching patches. The likelihood of the entire image is proportional to the product of all factors (only some of which are shown in Fig. 2),

$$p(z_{\mathcal{S}_1}, z_{\mathcal{S}_2}, ..., z_{\mathcal{S}_I}) \propto \prod_{i=1}^{I} h_i \prod_{j \in N_i} g_{i,j}, \quad (6)$$

where I is the total number of image patches considered, and N_i represents the set of data patches j connected to patch i in the model. While this set can be arbitrary for each patch i, in our experiments we chose a particular (but randomly chosen) relative configuration and use it for all patches in the image.

There is a number of ways to parameterize the relative geometric configuration of the patches, and some alternatives we have not tested will be discussed later, but first, we go over the choice of factors g and long-range interaction neighborhoods N in our experiments. The basic property that factors g are enforcing is that the relative positions of the coordinates in $\mathcal{S}_i$, $\mathcal{S}_j$ are preserved in the mappings $\mathcal{T}_i$, $\mathcal{T}_j$, *i.e.*, $\mathcal{S}_i(k) - \mathcal{S}_j(k) \approx \mathcal{T}_i(k) - \mathcal{T}_j(k)$[3]. If each patch is kept rectangular, this is equivalent to $\bar{\mathcal{S}}_i - \bar{\mathcal{S}}_j \approx \bar{\mathcal{T}}_i - \bar{\mathcal{T}}_j$, where bar denotes taking the mean of the coordinates in the set, since $\Delta\mathcal{S} = \mathcal{S}_i(k) - \mathcal{S}_j(k)$ is constant for all elements, and the same is true for $\Delta\mathcal{T}$. If the mapping inference enforces a preference to keeping the relative positions of the chosen patches, the epitomes would reflect longer-range correlations in images. However, the images often undergo geometric deformations due to angle of view changes and object deformations, which can violate some of these constraints, and to account for that, we can allow for different variances on the Gaussians that enforce them, $g_{i,j} = \mathcal{N}(\bar{\mathcal{T}}_i - \bar{\mathcal{T}}_j; \bar{\mathcal{S}}_i - \bar{\mathcal{S}}_j, \mathbf{\Phi}_{i,j})$. In this way, the mappings $\mathcal{S}_j \to \mathcal{T}_j$ for the neighbors of $\mathcal{S}_i$, *i.e.*, $j \in N_i$ will effect the mapping $\mathcal{S}_i \to \mathcal{T}_i$.

In our experiments, the neighborhood N_i consists of K patch (rather than pixel) indices (usually 10-20). There are roughly as many different rectangular patches as there are pixels in the image, since the patch can be centered at any pixel except those close to image boundaries. Thus patches can be indexed by their central pixels. To choose a neighborhood for each patch $S_{\mathbf{i}}$, where $\mathbf{i}$ now represents a 2-D coordinate of the central pixel, we first choose K *random* 2-D offsets $\mathbf{\Delta}_k$ up to some maximal distance d (*e.g.* half or quarter of the image size), *i.e.*, $||\mathbf{\Delta}_k|| \leq d$ for all k, and define $N_{\mathbf{i}}$ as an ordered set with $N_{\mathbf{i}}(k) = \mathbf{i} + \mathbf{\Delta}_k$. In other words, to construct the field of mapping constraints, each patch i is connected to interacting neighbors in the same relative configuration, but the mapped epitome patches $\mathcal{T}_{\mathbf{j}}$, $\mathbf{j} \in N_{\mathbf{i}}$ may not follow fixed configurations due to the uncertainty captured in the 2-D covariance matrix $\mathbf{\Phi}_{i,j}$ in the Gaussians $g_{i,j}$.

The K Gaussians $g_{i,j}$ for some i should have linked parameters, since they should all depend on the local deformation at i. Furthermore, the assumption $\bar{\mathcal{S}}_i - \bar{\mathcal{S}}_j \approx \bar{\mathcal{T}}_i - \bar{\mathcal{T}}_j$ is too rigid, both because the possible squishing of the texture in the epitome and because of the local image foreshortening and object deformations due to viewing angle changes and other effects. To account for this, we introduce a hidden transformation A_i which affects each of the patch links, *i.e.*, factors $g_{i,j}$,

$$g_{i,j} = \mathcal{N}(\bar{\mathcal{T}}_i - \bar{\mathcal{T}}_j; A_i(\bar{\mathcal{S}}_i - \bar{\mathcal{S}}_j), \mathbf{\Phi}_{i,j}). \tag{7}$$

In our experiments this transformation is linear and thus A_i is a matrix. The prior on this matrix can be included so as to prefer identity (not shown in Fig. 2). When, as in our experiments, each patch is connected to a large number of interacting neighbors (N_i contains a sufficiently large number of patches), A_i is inferrable. In our experiments we link parameters $\mathbf{\Phi}_{i,j}$ for different patches,

$$\mathbf{\Phi}_{m,N_m(k)} = \mathbf{\Phi}_{n,N_n(k)} = \mathbf{\Phi}_k. \tag{8}$$

In other words the links in the same relative configuration (the same $\mathbf{\Delta}_k$) share the same covariance structure. This allows learning the relative extent of the image correlations – the links that tend to lead to low correlation (*e.g.* because they reach to far in some direction) will simply have high variance captured in $\mathbf{\Phi}_k$.

As in some previous patch models, to account for image intensity changes (darkening or brightening of the patches, for example), we add two scalar hidden variables a, b that control the patch contrast in the factors h_i:

$$h_i = e_{\mathcal{T}_i}(a_i z_{\mathcal{S}_i} + b_i). \tag{9}$$

3. Mapping inference

Next we discuss inference in the epitome model with long-range patch correlations defined by (6), (7), (8) and (9). This model is a Markov random field (but unlike in most vision applications with more frequent and further-reaching links), with the epitome as the representation of the observation likelihoods. A number of techniques for inference in MRFs have been studied in the past, and most of them can be adopted here, including sampling, loopy belief propagation, and variational techniques (for review and some comparisons of probabilistic inference techniques see [7]). We have experimented with a simple variational technique[4], which factorizes the posterior distribution as $Q = \prod_i q(A_i)q(a_i, b_i|\mathcal{T}_i)q(\mathcal{T}_i)$ and further assumes that $q(a_i, b_i|\mathcal{T}_i)$ and $q(A_i)$ are delta functions. The resulting update rules are:

$$\begin{aligned}
q(\mathcal{T}_i) &\propto \tilde{h}_i(\mathcal{T}_i) \sum_{\mathcal{T}_j | j \in N_i} \prod_{j \in N_i} q(\mathcal{T}_j) g_{i,j}(\mathcal{T}_i, \mathcal{T}_j, \tilde{A}_i), \\
\tilde{h}_i(\mathcal{T}_i) &= \arg\max_{a,b} h_i(\mathcal{T}_i, a, b), \\
\tilde{A}_i &= \arg\max_{A_i} \sum_{\mathcal{T}_i} q(\mathcal{T}_i) \\
&\qquad \sum_{\mathcal{T}_j | j \in N_i} \prod_{j \in N_i} q(\mathcal{T}_j) g_{i,j}(\mathcal{T}_i, \mathcal{T}_j, \tilde{A}_i).
\end{aligned} \tag{10}$$

These equations do not update the belief $q(\mathcal{T}_i)$ about where each patch $\mathcal{S}_i$ should map only according to the epitome

[3] While this could be achieved by simply merging the patches $\mathcal{S}_i$ and $\mathcal{S}_j$ into one non-contiguous patch $\mathcal{S}$and imposing constraints on the epitome mapping $\mathcal{T}$, the patches in the epitome may no longer have a fixed shape, thus making it impossible to use cumulative sum and other computational tricks to perform efficient computation of the patch likelihoods h_i for all possible patches $\mathcal{T}_i$.

[4] Due to a large number of links, belief propagation yields essentially equivalent messages.

likelihoods for different possible patches $\mathcal{T}_i$ as in (4). Instead they take into account the probable mappings of the patches in N_i to skew the inference so as to have these patches in the proper geometric configuration with $\mathcal{T}_i$. Using the best matching contrast parameters a,b also allows the inference to be somewhat invariant to illumination changes. Finally, $\tilde{A}_i$ captures shearing of the image as it affects patch $\mathcal{S}_i$. Depending on the strength of the links defined by $\mathbf{\Phi}_k$, whose learning is discussed in the appendix, this shearing may be only local or more global.

Note that the epitome e involved in computation of h_i can either be learned or preset. For instance, in Fig. 5 we simply use an example of a video which we feel sufficiently epitomizes the class of data of interest and define the mean of the epitome e to be equal to that video, and use a small uniform value for all epitome variance. Then, the inference rules above, when iterated can be used to map other videos to it. To also learn the epitome from data, the original update rules, *e.g.* (5), only need to be changed slightly to account for the contrast variables (see the appendix). As in the previous work, the epitome update is iterated with the inference equations above.

4. Interpolating missing data

In (6) we model a selection of data patches. In our experiments, the image patches we considered are all image or video patches of a certain size. In many applications, a model of individual pixels is required, and the fact that each pixel belongs to several patches needs to be resolved. We follow the recipe from [8] and [4] – the patches $z_{\mathcal{S}}$ are in a *hidden* image, while the observed image, at each pixel contains the average of appropriate pixels in all patches $z_{\mathcal{S}}$ that overlap it. The patch agreements are enforced in the inference distribution, rather than in the model. In our case, to the factors h and g described above, we add an extra factor $f_{\mathbf{u}}$ *per pixel* $x_{\mathbf{u}}$ of the observed image x,

$$f_{\mathbf{u}} = \mathcal{N}(x_{\mathbf{u}}; \frac{\sum_i \sum_k [\mathbf{u} = \mathcal{S}_i(k)] z_{\mathcal{S}_i(k)}}{\sum_i \sum_k [\mathbf{u} = \mathcal{S}_i(k)]}, \rho^2_{\mathbf{u}}), \qquad (11)$$

with the total image likelihood proportional to

$$p(x) \propto (\prod_{\mathbf{u}} f_{\mathbf{u}})(\prod_i h_i \prod_j g_{i,j}). \qquad (12)$$

The variational posterior is factorized as $Q = \prod_{\mathbf{u}} q(z_{\mathbf{u}}) \prod_i q(A_i) q(a_i, b_i | \mathcal{T}_i) q(\mathcal{T}_i)$, with a single part of the posterior $q(z_{\mathbf{u}}) = \delta(z_{\mathbf{u}} - \nu_{\mathbf{u}})$ for each particular pixel $z_{\mathbf{u}}$ in the hidden image, regardless of how many patches $z_{\mathcal{S}_i}$ it may be in. This enforces the agreement of overlapping patches in the posterior distribution over all hidden variables. The posterior, as well as model parameters are estimated by minimizing the free energy

$$F = \sum_{hiddens} Q \log \frac{(\prod_{\mathbf{u}} f_{\mathbf{u}})(\prod_i h_i \prod_j g_{i,j})}{Q}. \qquad (13)$$

Not only does the model describe the likelihood[5] of image pixels rather than patches (still capturing a number of pixel correlations), but it also makes possible the inference of hidden pixels $z_{\mathbf{u}}$. Inferring these hidden pixels has various applications such as denoising and superresolution as in [4], which are all achieved by setting some of the variances $\rho^2_{\mathbf{u}}$ to large values. However, the inference procedure in our model will involve enforcing long-range correlations in the image. While this property should be helpful in previous applications of patch models, even more ambitious tasks can be attempted – the ones for which accounting for long range correlations in the data is crucial. Some of these tasks will be illustrated in the experimental section.

The inference of the hidden image pixels $z_{\mathbf{u}}$ reduces to estimation of parameters $\nu_{\mathbf{u}}$:

$$\nu_{\mathbf{u}} = \frac{\frac{x_{\mathbf{u}}}{\rho^2_{\mathbf{u}}} + \sum_{i,k|\mathcal{S}_i(k)=\mathbf{u}} q(\mathcal{T}_i) \frac{\mu_{\mathcal{T}_i(k)}}{\sigma^2_{\mathcal{T}_i(k)}}}{\frac{1}{\rho^2_{\mathbf{u}}} + \sum_{i,k|\mathcal{S}_i(k)=\mathbf{u}} q(\mathcal{T}_i) \frac{1}{\sigma^2_{\mathcal{T}_i(k)}}}, \qquad (14)$$

which balances the votes from different epitome patches with the observed value for the pixel based on the ratio of appropriate noise or uncertainty parameters (variances σ^2 for epitome 'votes' and ρ^2 for the votes from the observed image), as well as the uncertainties about mapping $q(\mathcal{T}_i)$. The other update rules (10) remain the same, except that instead of patches $z_{\mathcal{S}_i}$, patches of variational hidden image means $\nu_{\mathcal{S}_i}$ are used to compute h_i.

We have performed a number of experiments on halucinating plausible guesses for large chunks of missing data, by setting variances $\rho^2_{\mathbf{u}}$ for the missing data to high values. For instance, in one of the experiments, the data x is assumed to be a video of a hallway walk-through (with all the motion due to the cameraman's walking), but only a *single* frame is given. In this case, the coordinates $\mathbf{u} = (x, y, t)$ are 3D, the patches $\mathcal{S}_i$ are all video cubes of a certain size, and the variances $\rho^2_{x,y,t}$ are set to a high value everywhere except when $t = 0$, where it is set to a small value, thus overpowering the epitome predictions. For the epitome e we simply used a sequence of a walk-through of another hallway to be its mean, and set the epitome variances to a same small value everywhere (training the epitome on a larger number of such sequences would probably lead to better results), and then iteratively applied equations (10,14) until convergence. After each application of these equations, the inferred video ν resembles the original video which we used as an epitome more an more, both in terms of the local video texture resulting from quilting patches $e_{\mathcal{T}}$ and in terms of how the quilting of such patches in one part of the video volume influences the choice of the patches in another, distant, part of the volume. Thus, the resulting sequence $\nu_{x,y,t}$ contains the given photograph as its frame 0, since the low variances $\rho^2_{x,y,t=0}$ require it, but from $t = -7$ to $t = 7$ it adds new frames that agree with frame 0 so that

[5] Strictly speaking, the model is not normalized because of the factors g, but the same inference procedures can still be used; the imbalance due to g factors is uniformly distributed over the data, due to the fixed relative configuration of the neighborhoods N_i

the sequence contains the motion of the hall's walls out of the field of view, zooming motion of the texture close to the center of the field of view, as well as the same rocking motion of the human walk, present in the epitomic example.

In another experiment, the data and epitome coordinates are x, y, θ, where θ is a illumination angle, and the same procedure is used to perform single-example photograph relighting. Due to complex long-range correlations in these two types of data, inference of missing data using traditional patch quilting would be impossible.

5. Experiments

5.1. Epitome learning

In the original image epitome, a variety of patch sizes could be used during learning. Using large patches it is possible to capture large features, but because larger image structures also undergo larger deformations, the use of large patches also introduces significant amount of blurring. Small patches, on the other hand can capture repeating details that are easier to register, but the epitomes tend to have smaller structures and more discontinuities. When learning epitomes with long-range patch correlations, it is possible to capture large image structures using smaller patches, and thus achieve sharper epitomes and higher epitome likelihoods. The large epitome structures are the result of a combination of the global correlations provided by the mapping field we introduce in this paper, and the local correlations provided simply by patch overlaps, as in the original epitome.

The epitome shown in Fig. 1 was learnt from just three images of cars. The mean of the three images was used to initialize the epitome and after learning, the resulting epitome is a morph of the three cars. No alignment of the images was done beforehand. The long-range patch correlations caused the patches from these three cars to essentially agree upon an alignment of their features. The car images and the epitome all have a resolution of 120x90 pixels, and patches of size 10x10 with 10 random correlation links were used during learning. An example of these patch correlation links is outlined on top of the image on the left. The patch, $\mathcal{S}_i$ is randomly linked to a couple other patches, such that their corresponding epitome mappings $\mathcal{T}_i$ keep a similar spatial configuration. Fig. 3 illustrates the impact of long-range patch correlations in learning the epitome. The face epitome from [8] is shown on the left. Continuing the learning procedure for just a few more iterations, but including the constraints g on patch mappings by iterating (10, 17), results in the epitome on the right. The contiguous features in the new epitome are significantly larger, with a prototypical sharp face emerging near the center that does not look like any one single face in the database.

We can also examine where patches in an image match in the epitome. With the patch correlation constraints, we expect patches of a human face to match to contiguous areas of the epitome, as opposed to patches scattered all around the epitome. But, when a non-human face is matched with the epitome, we expect constraints to be violated and patches would not match to the epitome in the same manner as would a human face. To show where patches in an image match in the epitome, the denominator in (5) can be used as a transparency mask on the epitome as shown in Fig. 4. The human face on the left causes a large contiguous area of the epitome to be used frequently. The large forehead of the subject also results in the high use of a bright patch above the main face area of the epitome. The middle image shows a digitally created image of a cyclops. The usage of the bottom half of the prototypical face in the epitome is normal, but only one side of the upper half of the face is needed. Without modeling long-range correlations in epitome mappings, we would expect that both eyes would be used with equal probability, but because of these modeling constraints, for the most part, only half of the face in the epitome is used. Finally, an image of a dog is shown on the right. As it does not resemble a human face, the patch usage area in the epitome is quite deviant from that of a human face. Classification can be performed by computing the likelihood under the epitome for each of these images and the images shown have been ordered according to their likelihood with the human face on the left with the highest likelihood of the three.

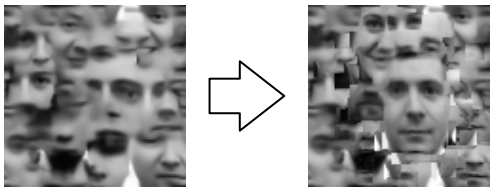

Figure 3. The effect of long-range patch correlations in learning the epitome. Starting with a traditional epitome on the left, conducting a few iterations of learning with patch correlations leads to the epitome on the right, which starts to show larger image structures, including a sharp whole prototypical face to which many of the images are mapped.

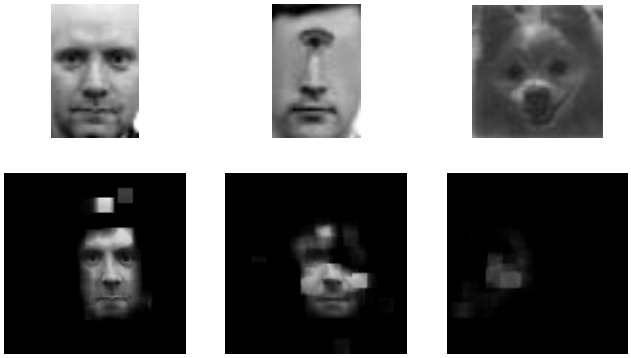

Figure 4. Face mapping. Three images of potential human faces are shown along the top with the corresponding matched areas of the epitome on the right from Fig. 3 below each one. The epitome is 'lit-up' proportional to how well the patches in each of the images matches to areas in the epitome.

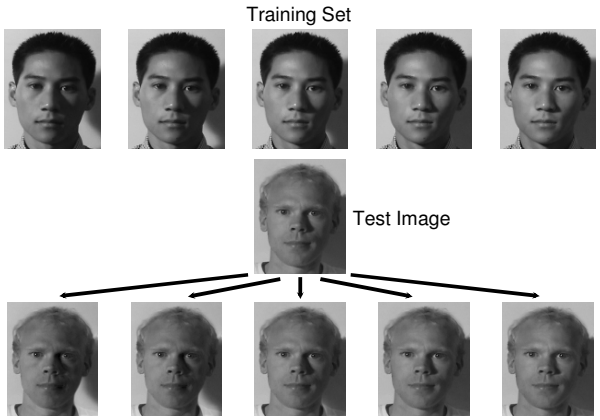

Figure 5. Changing the illumination of a face. Given a single test image and a guiding training set, a synthesized sequence is generated that reflects a changing illumination of the single image by iterating (10, 15, 16, 17, 14). The synthesis is plausible despite the absence of the use of any geometry or domain knowledge. Not only do the sharp shadows on the face move as expected, but the projected shadow behind the head also moves in a plausible manner. The training video of size 105x130 was used to synthesize 28 frames of size 100x125 from the target frame using patches of size 10x10x5 with 30 correlation links. The video sequence is available at http://www.psi.toronto.edu/∼vincent/patchcorr.html.

5.2. Image illumination manipulation

It is often desired to change the illumination of a subject in order for it to appear consistent with other elements of a differently illuminated scene. Information from a training sequence can be leveraged and used to interpolate changes in illumination of an image. Fig. 5 shows an example. The top row shows several frames of a video sequence exhibiting a change in illumination. The single test image shown in the middle is then extrapolated to mimic the illumination change of the training sequence and the frames corresponding to those in the training set are shown.

The illumination change is transferred onto the image through patch matching between the image and the video sequence and subsequent transferral of the illumination change that the patches exhibit in the adjacent frames of the training sequence via the 3D nature of the patches. The result can then grow outwards in an iterative fashion. Because this is an extrapolation from a single image, it is difficult, especially in frames far from the original seed, to maintain the coherence of the patch matching. Using long-range correlations between the patches is essential in maintaining consistency in the results. The shadows in Fig. 5 move in a plausible fashion. Patches of size 10x10x5 were used with 30 correlation links.

The second face illumination example shown in Fig. 6 shows a wider range of illumination change over a different subject. The first row of results serves to demonstrate the need for the long-range correlations as that is the result without the correlation links, while the sequence in the bottom row incorporate such links.

In the final illumination experiment shown in Fig. 7, an analogous operation was performed with a rippled piece of clothing. The geometry of folded cloth is very complex and would be difficult to model. Again, the illumination change is transferred from a sample video sequence in order to extrapolate the change of illumination angle of the source lighting for a single image. The complexity of the subject posed a difficult problem, but even then, the shadows can be seen moving in a plausible manner.

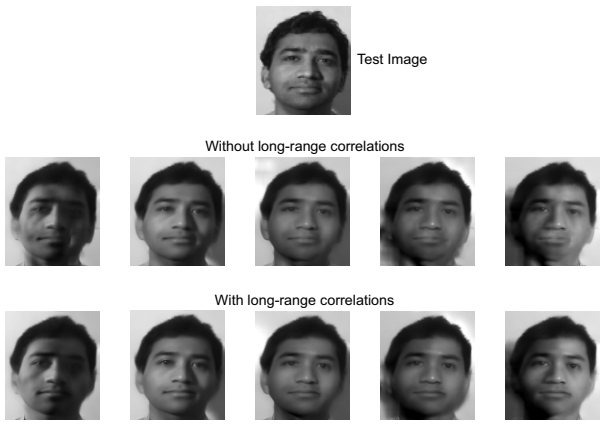

Figure 6. The necessity for long-range correlations in patch matching. The same experiment done in Fig. 5 is performed here with a different test image. Synthesis results with and without long-range patch correlations are shown. See the website for the video.

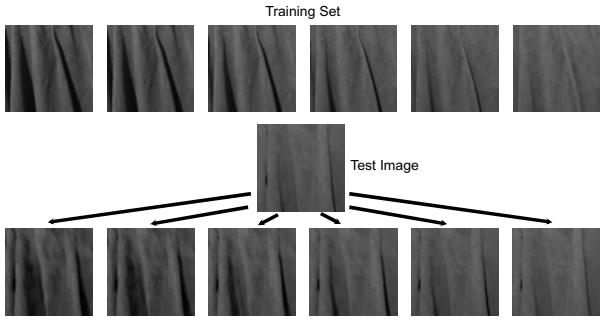

Figure 7. Changing the illumination of cloth. Given a single test image of draped cloth and a guiding sequence, the illumination of the single image is changed. Both the training set and the test image were of size 150x150, from which 74 frames were extrapolated using patches of size 15x15x5 with 50 correlation links. The video result can be found on the website.

5.3. Image walk-through

The same algorithm can be used in a variety of other synthesis applications. In Fig. 8, walking through a given image of a hallway is simulated given a training video. Note that the image is not simply enlarged, as parallax effects are apparent. As with the previous applications, no knowledge of geometry or domain knowledge is given to the algorithm. The patch correlations are sufficient to generate a plausible synthesis.

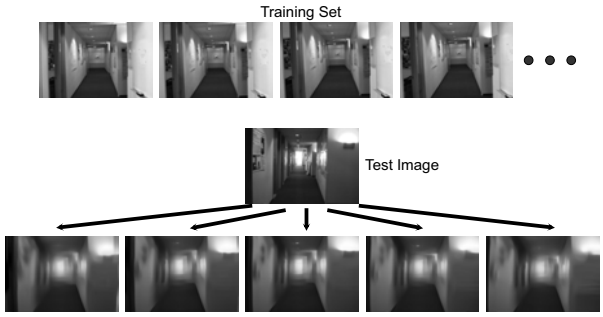

Figure 8. Walking through an image. Given a single image of a hallway, it is desired to mimic walking through the scene. Instead of just enlarging the image, it should appear as if the camera is moving down the hallway. This effect is achieved by quilting patches from a training sequence utilizing correlation links between patches to aid in matching. Patches of size 5x5x3 with 20 random correlation links were used to synthesize a plausible movement of walls, lights, and fixtures given the single 180x120 seed frame. See the video on the website for the full effect.

6. Conclusion

We have introduced a powerful new patch-based model that accounts for the varying geometric configurations of object features to describe learnable probability density functions of visual data. The representation power of our model has been illustrated in a variety of tasks including multiple object registration and detection, as well as extreme missing data problems, such as relighting and walking through an image, where a single image frame is extrapolated to a video sequence, the video results of which can be found on the project webpage[6]. These tasks can be achieved without explicitly incorporating domain knowledge because our simple data-driven model captures sufficiently short- and long-range correlations among the data patches.

Acknowledgements

V. Cheung was financially supported by an NSERC Canada Graduate Scholarship and D. Samaras was supported by grants from NSF (ACI-0313184) and DOJ (2004-DD-BX-1224).

Appendix

For a diagonal transformation A_i, with diagonal elements A_{i_m}, the update equation is given by

$$\tilde{A}_{i_m} = \frac{\sum_{j \in N_i} \sum_{\mathcal{T}_i} \sum_{\mathcal{T}_j} q(\mathcal{T}_i) q(\mathcal{T}_j) (\bar{\mathcal{T}}_{i_m} - \bar{\mathcal{T}}_{j_m})(\bar{\mathcal{S}}_{i_m} - \bar{\mathcal{S}}_{j_m})}{\sum_{j \in N_i} (\bar{\mathcal{S}}_{i_m} - \bar{\mathcal{S}}_{j_m})^2}. \tag{15}$$

The update equation to account for uncertainties in the correlation links is given by

$$\mathbf{\Phi}_k = \frac{\sum_{i,j|j \in N_i} \sum_{\mathcal{T}_i} \sum_{\mathcal{T}_j} q(\mathcal{T}_i) q(\mathcal{T}_j) D_{ij}^2}{\sum_{i,j|j \in N_i} 1}, \tag{16}$$

[6] http://www.psi.toronto.edu/∼vincent/patchcorr.html

where

$$D_{ij} = \bar{\mathcal{T}}_i - \bar{\mathcal{T}}_j - A_i(\bar{\mathcal{S}}_i - \bar{\mathcal{S}}_j).$$

Learning under the contrast model requires a reversal of the scaling and addition used during matching:

$$\mu_{\mathbf{u}} = \frac{\sum_i \sum_{\mathcal{T}} q(\mathcal{T}_i = \mathcal{T}) \sum_k [\mathbf{u} = \mathcal{T}(k)](z_{\mathcal{S}_i(k)} - b_i)/a_i}{\sum_i \sum_{\mathcal{T}} q(\mathcal{T}_i = \mathcal{T}) \sum_k [\mathbf{u} = \mathcal{T}(k)]}. \tag{17}$$

References

[1] M. Ashikhmin. Synthesizing natural textures. In *Proc. Symposium on Interactive 3D Graphics*, pages 217–226, 2001.

[2] C. Broit. *Optimal registration of deformed images*. PhD thesis, University of Pennsylvania, 1981.

[3] M. C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. ECCV*, pages 628–641, 1998.

[4] V. Cheung, B. J. Frey, and N. Jojic. Video epitomes. In *Proc. IEEE CVPR*, 2005.

[5] P. Debevec, A. Wenger, C. Tchou, A. Gardner, J. Waese, and T. Hawkins. A lighting reproduction approach to live-action compositing. In *SIGGRAPH*, pages 547–556, 2002.

[6] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, pages 56–65, 2002.

[7] B. J. Frey and N. Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(9):1392–1416, 2005.

[8] N. Jojic, B. J. Frey, and A. Kannan. Epitomic analysis of appearance and shape. In *Proc. IEEE ICCI*, 2003.

[9] N. Jojic, V. Jojic, B. J. Frey, C. Meek, and D. Heckerman. Using 'epitomes' to model genetic diversity: Rational design of hiv vaccine cocktails. In *Proc. NIPS*, 2005.

[10] A. Kannan, J. Winn, and C. Rother. Clustering appearance and shape by learning jigsaws. In *Proc. NIPS*, 2006.

[11] A. Shashua and T. Riklin-Raviv. The quotient image: class-based re-rendering and recognition withvarying illuminations. *IEEE Trans. PAMI*, 23(2):129–139, 2001.

[12] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. ECCV*, pages 18–32, 2000.

[13] Y. Wexler, E. Shechtman, and M. Irani. Space-time video completion. In *Proc. IEEE Comp. Vision and Pattern Recognition*, pages 120–127, 2004.

[14] L. Zhang, S. Wang, and D. Samaras. Face synthesis and recognition under arbitrary unknown lighting using a spherical harmonic basis morphable model. In *Proc. IEEE CVPR*, pages 209–216, 2005.