Pattern Recognition Letters xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Learning explicit and implicit visual manifolds by information projection

Song-Chun Zhu [a,b,*], Kent Shi [a], Zhangzhang Si [a]

[a] University of California at Los Angeles, Los Angeles, CA 90095, USA
[b] Lotus Hill Research Institute, Ezhou, Hubei 436000, China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Texture
Texton
Image Manifold
Primal sketch
Visual learning
Information projection

ABSTRACT

Natural images have a vast amount of visual patterns distributed in a wide spectrum of subspaces of varying complexities and dimensions. Understanding the characteristics of these subspaces and their compositional structures is of fundamental importance for pattern modeling, learning and recognition. In this paper, we start with small image patches and define two types of atomic subspaces: explicit manifolds of low dimensions for structural primitives and implicit manifolds of high dimensions for stochastic textures. Then we present an information theoretical learning framework that derives common models for these manifolds through information projection, and study a manifold pursuit algorithm that clusters image patches into those atomic subspaces and ranks them according to their information gains. We further show how those atomic subspaces change over an image scaling process and how they are composed to form larger and more complex image patterns. Finally, we integrate the implicit and explicit manifolds to form a primal sketch model as a generic representation in early vision and to generate a hybrid image template representation for object category recognition in high level vision. The study of the mathematical structures in the image space sheds lights on some basic questions in human vision, such as atomic elements in visual perception, the perceptual metrics in various manifolds, and the perceptual transitions over image scales.

This paper is based on the J.K. Aggarwal Prize lecture by the first author at the International Conference on Pattern Recognition, Tempa, FL. 2008.

1. Introduction

1.1. Quest for structures of the image space

In pattern recognition, people often extract many features, as many as one could come up with, from input data, treat them as independent points in a vector space, lump them together for classification, and justify them by error rates in the end. It is taken for granted that we are not obligated to understand analytically the ingredients of these features. To the contrary, people feel rather proud that they can solve problems without analytically studying the features or the space structures. Yes, why should we "solve a problem more than necessary!" But, if we can solve the problems by just trying good features, then why are we still here designing new features on a daily bases since the birth of pattern recognition? For example, in object recognition, new features are invented in every computer vision and pattern recognition conference. Given the very high dimensions of images, even the seemingly humble request of finding distinct features for weak classifiers turns out to be very hard to meet for many object categories. For example, for detecting vehicles in streets using Adaboost (Freund and Schapire, 1997), we may run out of weak classifiers rather quickly.

This approach sounds very similar to the Chinese herb clinics, which have been practiced for more than a thousand years. A herb clinic typically has hundreds of remedies including almost anything one can try on: barks, roots, stems, leaves, bugs, worms, and shells which are like our features. They are selected, mixed in calculated proportions and boiled to a soup – darker and more bitter than the strongest coffee. With the enormous number of possible combinations, one is always hopeful to try some of them for any given new or unknown illness. But good recipes are tough to find!

The herb clinics are nowadays adopting terms in modern medicine, such as virus, genes, and molecules. Similarly, why could not we spend some time understanding the structures of the image space, the ingredients of the features, and the mechanism of image composition?

In this paper, we do not intend to engage the long standing debates on generative versus discriminative methods. Instead we take a short journey to explore the image space and to report some characteristics of its atomic subspaces, and then we show how we can pursue models for them under a common information theoretical principle, and integrate them into more complex

* Corresponding author. Address: Lotus Hill Research Institute, Ezhou, Hubei 436000, China. Fax: +1 310 206 5658.
E-mail addresses: sczhu@stat.ucla.edu (S.-C. Zhu), kentshi@stat.ucla.edu (K. Shi), zzsi@stat.ucla.edu (Z. Si).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.07.020

representations for generic images and object categories. Although the image space is very complex, the tools and principles for understanding them could be simple.

1.2. Manifolds in image space: implicit, explicit, and hybrid

Considering an image $\boldsymbol{I}$, for simplicity, we start with a small patch of $N = 11 \times 11$ pixels. Depending on where we look, the patch could be a simple primitive, e.g. patch A at the nose of a hedgehog in Fig. 1, or a texture, e.g. patch B in the hedgehog body.

If we map patches A and B to the N-dimensional space of all image patches, it is not hard to realize that they are from two very different subspaces.

Patch A lies in a 4-dimensional subspace where all the image patches correspond to the same geometric pattern of an edge segment. Each image patch can be represented by the following variables: central location of the edge segment, (x, y), orientation θ, and intensity contrast a. We denote these variables by $w = (x, y, \theta, a)$. In general, we have the following definition:

Definition 1. An *explicit manifold* is a subspace of image patches defined by an explicit function $g(w)$ with small distortions ϵ,

$$\Omega^{\mathrm{ex}} = \{\boldsymbol{I} : \boldsymbol{I} = g(w) + \epsilon; w \in W\}. \quad (1)$$

Each image patch $\boldsymbol{I}$ in the explicit manifold Ω^{ex} is represented or identified by a low-dimensional variable w, that can take values within a range W. ϵ corresponds to the precision of representation or perception. As w varies in W, $g(w)$ spans a low-dimensional manifold in the image space. The left panel of Fig. 1 shows a number of geometric primitives, where each column shows a primitive at the top, followed by some instances below. Each geometric primitive corresponds to an explicit manifold, with a different functional form of $g()$ and an associated range W. The instances in each column belong to the same manifold, and each instance is indexed by a particular value of w. Sometimes, an explicit manifold is also called an equivalent class invariant to a set of transformations associated with $g()$.

Patch B belongs to a subspace of a much higher dimension, where the patches are perceptually equivalent and share some common statistical properties, e.g., the histograms of Gabor filtered responses. Let $\mathrm{H}()$ extract the histograms of filtered responses from image $\boldsymbol{I}$ and $\boldsymbol{h}$ be a specific value of the histograms that is shared by all the image patches in Ω^{im}.

Definition 2. An *implicit manifold* is defined by statistical constraints with a small statistical fluctuation ϵ,

$$\Omega^{\mathrm{im}} = \{\boldsymbol{I} : \mathrm{H}(\boldsymbol{I}) = \mathrm{h} + \epsilon\}. \quad (2)$$

The fluctuation decreases with the patch size N.

The image patches in the subspace Ω^{im} cannot be represented or identified by a small number of variables. That is, these image patches lose their individual identities, and they are collectively described by statistics $\boldsymbol{h}$, in the sense that all the image patches in Ω^{im} share the same $\boldsymbol{h}$. The subspace Ω^{im} is very different from the explicit manifold Ω^{ex}. With a rather liberal use of the term "manifold," we call Ω^{ex} the implicit manifold, in the sense that image patches $\boldsymbol{I}$ in Ω^{ex} cannot be explicitly identified or differentiated by a small number of variables, and they are defined by an implicit function $\mathrm{H}(\boldsymbol{I}) = \mathrm{h}$, instead of an explicit function $\boldsymbol{I} = g(w)$ as in the explicit manifold.

Ω^{im} is also called the Julesz *ensemble* in (Zhu et al., 2000), It is similar to the micro-canonical ensemble in statistical physics which defines a huge set of microscopic states using a small number of macroscopic properties as constraints. Ω^{im} can induce a general family of Markov random fields called the FRAME model for texture (Zhu et al., 1997). In Ω^{im}, $\boldsymbol{I}$ is the microscopic state and $\boldsymbol{h}$ is the macroscopic (or statistically invariant) property which are considered sufficient statistics in human perception. That is, according to the well-known psychophysicist Bella Julesz, texture images are perceptually equivalent if they share certain statistical properties.

The explicit manifolds and implicit manifolds are two extremes of the image patterns. The explicit manifolds contain pure geometric structures, and the implicit manifolds contain pure stochastic textures. For that reason, we also refer to them as pure or atomic manifolds. When we look at the area around of eye of the hedgehog or a larger patch, the image patch may contain both geometric structures and stochastic textures. So such image patches belong to what we call a hybrid or composite manifold.

It is worth pointing out that many terms in various disciplines refer to the same thing from different perspectives. For example, manifolds in mathematics, ensembles in statistical physics, equivalent (invariant) classes in geometry or control theory, clusters in pattern recognition, subspaces in machine learning. In statistics a probability model p is also said to focus on a set or ensemble Ω_p. Thus we have the first set of terminologies:

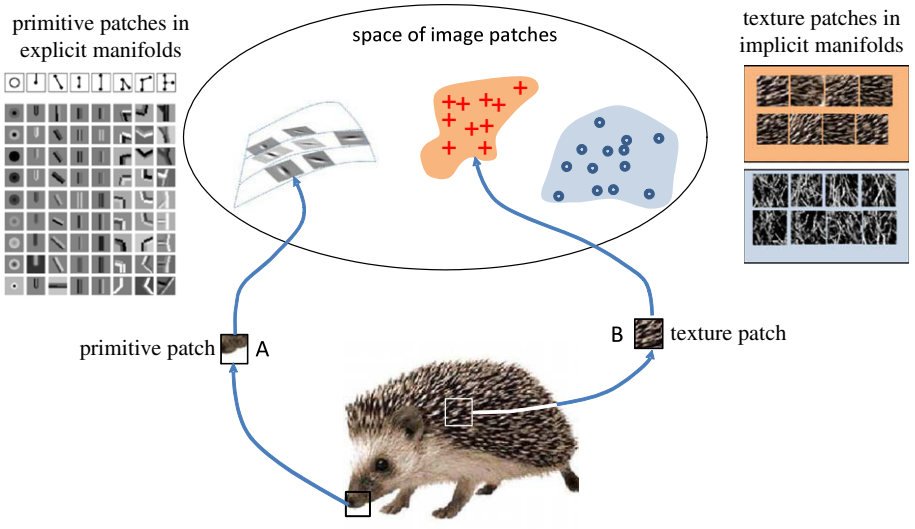


Fig. 1. Pure manifolds in the space of image patches. Patches A and B belong to two distinct types of subspace. See text for interpretation. Picture adopted from Si (2009).

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

$$\text{manifold} \leftrightarrow \text{subspace} \leftrightarrow \text{cluster} \leftrightarrow \text{equivalence class} \leftrightarrow \text{ensemble} \leftrightarrow \text{model } \Omega_p.$$

The symbol $\leftrightarrow$ means "the two concepts can be used interchangeably". These terms may sound confusing sometimes, but we should not be too rigorous or sensitive about these names. In fact, if we tolerate different perspectives, we can benefit from the diversity brought to our field over the years with useful tools associated with them. In pattern recognition, the term "cluster" has never been defined precisely. We believe that the study of manifolds and their compositions will provide a better description for the structures of the image space.

At this point, people may raise many questions. Below are some urgent ones.

1. How many explicit and implicit manifolds can we find in the space of daily photos? These manifolds are supposed to be the basic components for image coding, recognition, and perception.
2. How are they related to each other in the space? Can we sort these manifolds along some axis?
3. How do we model these manifolds in the image space? How do we measure their volume and weight their mass?
4. How do we compose atomic manifolds to form larger composed manifolds? The latter host images from object categories.

We shall discuss these questions along our short journey in exploring the fascinating image space which we still do not know too much about.

1.3. The spectrum of manifolds and manifold transition

It has long been accepted that daily pictures, such as face images under different expressions and lighting conditions or images of a vehicle taken by motion camera, lie in low dimensional appearance manifolds. This is the pillow of many well-known dimension reduction techniques, such as Isomap, local linear embedding (LLE) (Roweis and Saul, 2000). People who applied LLE to image patches cropped from daily photos will be disappointed. The reason is intuitively discussed in the previous subsection, the image space is not a low dimensional manifold but contains a wide spectrum of manifolds with compositions. These manifolds have varying dimensions. Fig. 2 shows image patches of 11 categories from image primitives in "low entropy" classes: edges, bars, parallel lines; to entropy objects: cat, dog, lion, tiger; and to "high entropy" textures: fur, carpet and grass. When all the images are normalized to have zero mean intensity, the two extremes of the spectrum are (1) the set of images with constant pixel intensities over the image lattice, which has zero/minimal dimension, and (2) the set of images with pixel intensities i.i.d. uniformly distributed, which has full N dimensions.

Here we need to clarify another set of terms which are highly related to each other: volume of a manifold Ω, its entropy $\mathscr{H}$ – a term used in statistical physics and information theory, and its intrinsic dimension d – a term adopted in mathematics, coding and learning. In later sections, we will show that the volumes of the explicit and implicit image manifolds are measured in two distinct ways. In both cases, $\mathscr{H}, d$ and the log-volume of Ω are measures for the massiveness of the manifolds,

$$\text{entropy } \mathscr{H} \leftrightarrow \text{dimension } d \leftrightarrow \log \text{volume} \log|\Omega|.$$

Intuitively, one may also call it the degree of freedom in pattern recognition.

In this paper, we use entropy as an axis to map all these manifolds. Sometimes, people confuse entropy with information. Information has to be defined for a task. For vision tasks, both primitives (in the low entropy classes) and textures (in the high entropy classes) are considered boring and less informative than objects (in the middle entropy classes). As we will discuss in later section (see Fig. 18), we have a smaller number of classes at the two ends of the entropy spectrum and a much larger number of classes in the middle entropy regime. The latter are hybrid manifolds and have complexer structures. We dub it the "middle-entropy crisis" of computer vision and pattern recognition. Understanding the structures of such hybrid manifolds shall shed lights on the search of good features and algorithms for pattern recognition.

The entropy is inherently related to image scaling (zooming). Fig. 3 shows sequences of snapshots of maple and ivy leaves in the process of zooming out. At closer distance, each image contains a single image leaf and thus represents image primitives from the low entropy regime of geometric primitives. As the camera zooms out, each image contains a few leaves (i.e. objects) with pedals. At further distances, each image captures hundreds or thousands of leaves and becomes texture where the individual leaves can no longer be identified. At the limit, if we have had large enough maple forest or ivy wall, the intensity of each pixel is the sum of photons from hundreds of leaves, and the image should converge to Gaussian noise because of the central limit theorem.

As the camera zooms out, more leaves come into the images, which become more complex, and the image entropy increases. As images are discrete signals with finite resolution, details of the leaves get lost and become imperceptible. Our perception has to drop explicit variables for geometric structures and change the representation as well as the perceptual metric. In this process, if we crop image patches from images at different scales, the manifolds that contain these image patches will change from explicit

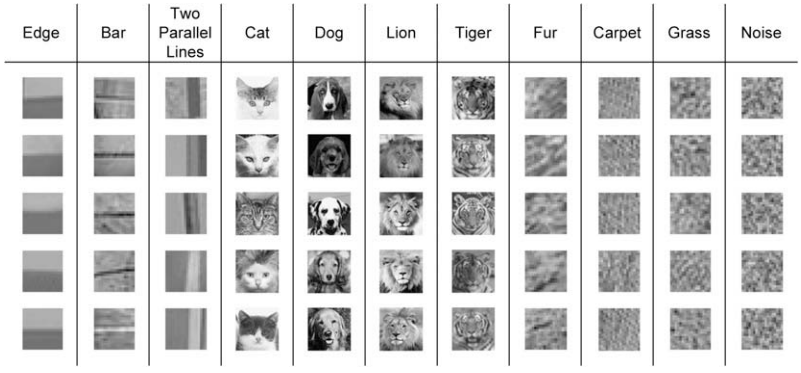


Fig. 2. Example patches of 11 categories ranging from the classes of low-entropy patches, such as edges and bars, to the classes of high-entropy patches, such as textures. Object categories, such as animal faces, often lie in the middle entropy classes.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

Fig. 3. Image scaling causes perceptual transitions between manifolds. From left to right, as the we zoom out from the leaves, our perception of the image patches transits from a primitive for single leaf instance in a low entropy class (or explicit image manifold) to textures in high entropy classes (or implicit image manifolds).

to hybrid and finally to implicit manifolds. Readers interested in this aspect are referred to an early paper (Wu et al., 2008) for a discussion about the information scaling, imperceptibility, and perceptual transitions.

To summarize, we have the third set of terminologies for the axis,

$$\text{entropy regime transition} \leftrightarrow \text{camera zooming} \leftrightarrow \text{image scaling}.$$

In Fig. 3, the leaves have similar sizes in a narrow depth range, thus images at each scale reside in classes/manifolds of similar entropy and the entropy transition is obvious. Intuitively, we may think of the spectrum of manifolds as distributed in different entropy regimes in the image space. By analogy, the structures of the image space may be similar to the cosmology picture in Fig. 4. In our universe, mass and energy are distributed in various forms. In some subspaces, like the stars, the distributions are of high densities and low volumes; while in other subspaces, like the nebulas, the distributions have low densities and high volumes. By analogy, the stars correspond to the explicit manifolds for image primitives and the nebulas correspond to the implicit manifolds for textures.

1.4. Pursuing manifolds in the image space

So far, we have shown that the image manifolds have vastly different dimensions and characteristic structures, and some low dimensional manifolds may be submerged in high dimensional manifolds. In the literature, the conventional *K*-means clustering methods and other recent methods for subspace learning (Ma et al., 2007) all assume that clusters have similar linear structures, and thus cannot be applied to such image space. We need a new way to find these manifolds or clusters. This is illustrated in Fig. 5.

Fig. 4. By analogy, a picture of the universe with mass distributed on stars (high density, low volume) and nebulous (low density, high volume).

Suppose that in the image space, represented by the big ellipse in Fig. 5, there is an unknown target manifold Ω_f to be clustered for a pattern, and it is governed by an underlying "true" probability $f(\boldsymbol{I})$. Ω_f is a subspace represented by the red[1] closed curve. We learn a sequence of models to approach f in a stepwise manner, i.e. pursuit, starting from an initial probability $q(\boldsymbol{I})$:

$$q = p_0 \to p_1 \to p_2 \to \cdots \to p_k \to f. \tag{3}$$

These models represent a series of manifolds, shown by the dashed blue curves approaching Ω_f,

$$\Omega_{p_0} \to \Omega_{p_1} \to \Omega_{p_2} \to \cdots \to \Omega_{p_k} \to \Omega_f. \tag{4}$$

When Ω_p coincides with Ω_f, we said the manifold is captured.

There are two ways for pursuing the manifolds as Fig.5 illustrates in (a) and (b) respectively. For an implicit manifold, we may start with Ω_{p_0} the whole image space, and at each time, we add a new constraint to shrink the manifold. With more constraints added, Ω_p will capture Ω_f from outside. For Ω^{im} defined in Eq. (2), these constraints augment to $\mathtt{h} = (\mathtt{h}_1, \ldots, \mathtt{h}_K)$. For an explicit manifold, we start locally with a single point or small ball inside Ω_f, and at each step we expand some dimensions w in Eq. (1) and fill Ω_f from inside. We will present the model pursuit framework by information projection and then show two case studies for the two types of pursuit in Section 3.

The reason for choosing the two pursuit strategies is quite intuitive. As implicit manifolds are of very high dimensions, thus it is fast to capture them through constraints (reducing entropy or volume), while the explicit manifolds are of much lower dimensions, thus it is more effective to capture them by expansion (increasing the volume).

By analogy, when a teacher is grading a final exam with the full mark being 100 (just like our full dimension N), for a very strong student, the teacher will start with 100, and substract a few points here and there for errors, and then count the final grade like

$$\text{strategy a}: \quad 100 - 3 - 0 - 2 - 1 - 0 - 0 - 2 - 0 - 0 - 0 = 92.$$

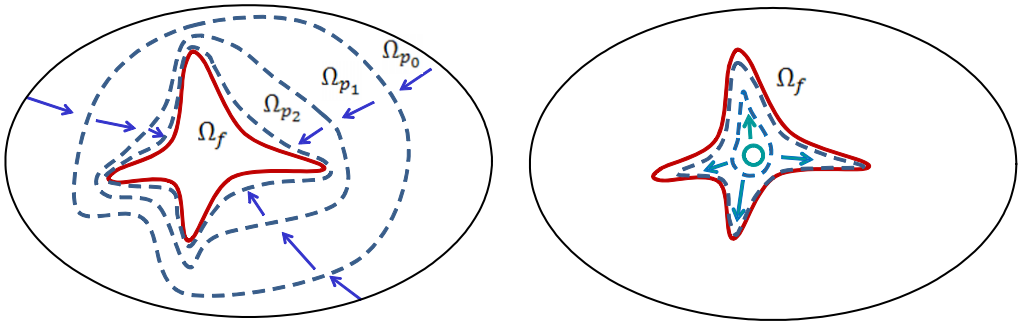
For a very weak student, the teacher starts will count from 0 and add a few points for credits,

$$\text{strategy b}: \quad 0 + 5 + 0 + 0 + 2 + 0 + 0 + 2 + 0 + 0 + 1 = 10.$$

In practice, students at both ends are easy to grade. Students in the middle range around the pass/no-pass line need the most work. This is exactly like the manifold pursuit where we encounter the hybrid manifolds in the middle entropy regime for objects.

[1] For interpretation of colour in Figs. 5, 9, 11, 14, 21, and 22, the reader is referred to the web version of this article.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020



(a) pursuit of implicit manifold (b) pursuit of explicit manifold

Fig. 5. The red curve represents a manifold Ω_f to be pursued. (a) An implicit manifold is pursued through a sequence of models by shrinking from the whole image space. (b) An explicit manifold is pursued through a sequence of models by expanding from a single point or small ball.

1.5. Plan of the paper

So far, we have introduced our motivation for studying the structures of the images space, the characteristics of manifolds or subspaces, an entropy axis for mapping the manifolds into various regimes and its relation to scaling, and the intuitive ideas about modeling and manifold pursuit.

The plan for the rest of the paper is the following.

Firstly, we discuss some related work in the literature in Section 2 to set up the background and context. We overview a number of streams in psychology, coding, image modeling, and applied math which have investigated similar topics.

Secondly, we present the theoretical framework for manifold pursuit by information projection in Section 3. We show that this is a general modeling and learning scheme which has been practiced in several fields under different names. We show two case studies: one on Markov random fields, and the other on learning active basis models for object templates.

Thirdly, in Section 4, we apply the manifold pursuit algorithm in Section 3 to the space of image patches and present experiments for clustering the implicit and explicit manifolds from the space. Also we show the manifolds in a sequence of scaled images to illustrate the transition of these manifolds.

Fourthly, we present two case studies in Section 5 that integrate the implicit and explicit manifolds for image representation. One is the primal sketch model for generic images at the middle level (Guo et al., 2007), and the other is the mixed templates for object categories (Si, 2009). The two cases demonstrate that the two atomic manifold can be combined to represent general images.

Finally we conclude the paper with a discussion and connection to a more general framework at the higher level: stochastic image grammar embedding in a hierarchical And-Or graph structure. The study of the mathematical structures in the image space sheds lights on some basic questions in human vision, such as atomic elements in visual perception, the perceptual metrics in various manifolds, and the perceptual transitions over image scales.

2. Related research streams in the literature

We overview some interesting work in psychology, natural image modeling, coding, and applied mathematics, which investigated related topics.

2.1. Studies in early vision: texture, texton, and primal sketch

In the 1960s, a well-known psychophysicist Julesz (1928–2003) asked a fundamental question about texture perception: what are the essential feature statistics so that two texture images sharing the same statistics are perceptually equivalent. In today's terminology, a texture is a set of images that share the same feature statistics $\mathrm{H}(\boldsymbol{I}) = \mathrm{h}$. This is the implicit manifold that we defined in Eq. (2) and named the Julesz ensemble (Zhu et al., 2000). Julesz's texture quest was not very fruitful, since there was very limited knowledge about the neural functions (such as Gabor filters) in selecting the features and statistics $\boldsymbol{h}$. Given some statistical constraints $\boldsymbol{h}$, one needs to generate arbitrary (unbiased) images that share the same $\boldsymbol{h}$. In statistics, this is to draw fair sample from the manifold Ω^{im}, so one needs to establish Markov random fields for various $\boldsymbol{h}$ and use Markov chain Monte Carlo methods for sampling from the models. Such mathematical tools are necessary for studying the implicit manifolds, but they were simply not available at that time.

Julesz noticed that early vision (about 100–200 ms) seems to be very sensitive to certain elements while indifferent to others. Fig. 6 shows two examples designed by Julesz. In (a), One can detect the arrows from the triangles instantly (i.e. constant time) regardless of the number of triangles (distracters) in the background, while in (b) one has to search for the 'S' in a number of '10's. The search time increases linearly with the number of distractors. Julesz concluded that there must be a set of atomic elements for human perception, which he called "textons." In our terminology, each texton is an explicit manifold. Later psychophysical experiments showed that textons are adaptive (Karni and Sagi, 1991) and can be learned through repeated exposure to such elements. Such phenomenon is also quite common in recognizing symbols in language. For example, when western travelers in China look at a Chinese newspaper or magazine, the Chinese characters appear to be textures, while the Chinese people see the characters as textons.

As a pioneer, Julesz had touched the essence of texture and textons. In his 1995 book (Julesz, 1995), he wrote a dialogue with himself and was apparently puzzled by the textures and the textons, which in our opinion, are two different types of manifolds studied by distinct branches of mathematics with different tools. The textures are modeled by Markov random fields with analysis tools from statistical physics, while textons are studied by coding theories and tools from harmonic analysis. In computer vision, textons are represented by vectors of filter responses (Leung and Malik, 1999) or transformed component analysis (Frey and Jojic, 1999). A comprehensive review and comparison is given in (Zhu et al., 2005). Since Julesz, there has been no real successful work investigating and comparing the atomic textures and textons in generic images. Our result presented in Section 4 is the first direct experiment in the literature that compares and competes between textures and textons and ranks these manifolds numerically.

In his monumental book (Marr, 1982), Marr inherited Julesz's texton notion and proposed the concept of image primitives as basic perceptual tokens, such as edges, bars, junctions, and

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

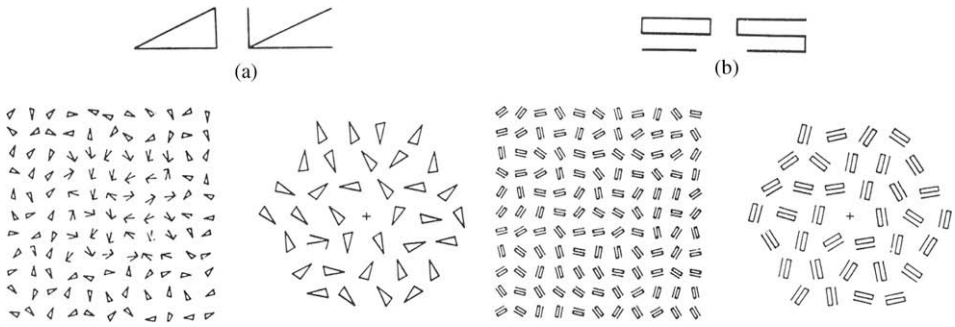


Fig. 6. Two examples from Julesz's experiments on textons.

terminators. Inspired by the Nyquest sampling theorem in signal processing, Marr went a step further and asked for a token representation which he named "primal sketch" as a perceptually lossless conversion from the raw image. He tried to reconstruct the image with zero-crossings unsuccessfully and his effort was mostly limited by the lack of proper models of texture. In Section 5.1, we will present a mathematical model for primal sketch based on our early work (Guo et al., 2007), which integrates the implicit and explicit manifolds seamlessly. We refer to two early papers on texton (Zhu et al., 2005) and primal sketch (Guo et al., 2007) for detailed discussions.

In summary, texture, texton, and primal sketch are important concepts in the early stage of visual perception. In this paper, they correspond to the implicit, explicit, and hybrid manifolds in the image space.

2.2. Studies in the statistics of natural images – tips of the iceberg

It has long been noticed since the 1960s that image signals do not observe the prominent Gaussian distributions. For example, the distribution of the gradients of image intensity $\nabla \boldsymbol{I}$ has higher kurtosis and heavier tails than Gaussian, and often remains invariant when the images are down-scaled. This has inspired much research in the 1990s and early 2000s studying the statistics of natural images (Ruderman, 1994; Zhu and Mumford, 1997; Huang and Mumford, 1999; Geman and Koloydenko, 1999; Lee et al., 2003; Mumford and Gidas, 2001). Here, by natural images, people usually mean photos taken in natural scenes which have a rich set of objects of various sizes in a long range of distance from the camera, e.g. trees in a forest.

Fig. 7 shows two typical results that are non-Gaussian distributions. The heavy tails in (a) and spikes in (b) indicate the existence of structures in images. Filters (such as gradients, Gabor) receive very high responses at such structural locations (such as image primitives) and thus contribute to the tails or spikes of the histogram.

These histograms are marginal (projected) statistics of the image manifolds that we are discussing, and are the tips of the iceberg – the underlying structures of the image space. This simple evidence argues against the Gaussian assumptions and quadratic metrics that are common in computer vision and pattern recognition, including Gaussian MRF, *K*-mean clustering, etc., and call for non-Gaussian models.

It is worth noting that these studies focused on low dimensional statistics over local features. The observed scale invariance is pooled over all locations in an image which consists of objects in a large range of scales. When we scale the image by downsampling, the larger objects become smaller, the local statistics remain relatively stable or invariant while the perception of individual object changes over scales. Such invariance does not conflict with our observations in the scale transitions on objects (say the maple leaves) over scales in Fig. 3.

2.3. The puzzle of feature learning: sparse coding vs. MRF

The natural image statistics motivated a new round of efforts in image modeling in the past 15 years. Two types of models are adopted to account for the statistics.

The first model is the sparse coding by Field (1987) and Olshausen and Field (1996), who argued that the high kurtosis in natural images indicates a sparsity principle which directly contributes to the receptive fields of simple cells found in the prime visual cortex area V1. Fig. 8a shows some examples of the image base functions learned in an unsupervised manner from natural images using a simple sparse coding model. These patches resemble the Gabor functions and in our opinion belong to the explicit manifolds. We notice that their algorithm involved some preprocessing stage that suppressed the high frequency texture signals.

The second model is the non-parametric Gibbs (MRF) model proposed by Zhu and Mumford (1997). This model is constructed through the minimax entropy principle (Zhu et al., 1997) with statistical constraints so that the model reproduces exactly the

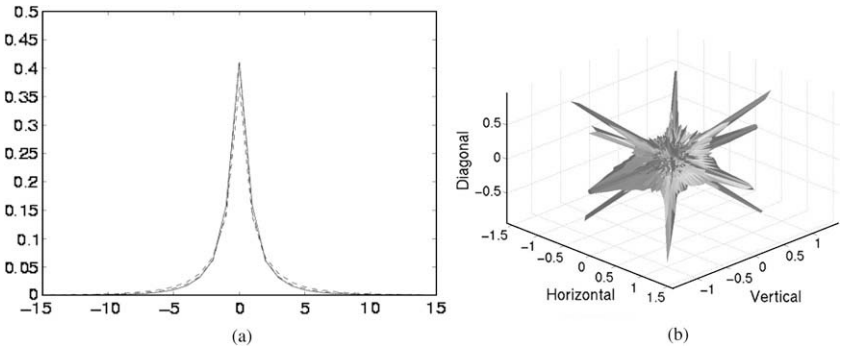
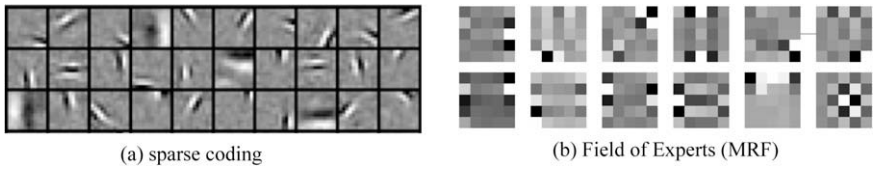


Fig. 7. (a) The histogram of image gradient $\nabla \mathrm{I}$ of images at four resolutions, from Zhu and Mumford (1997). (b) The iso-probability surface plot on a 3D histograms of range depth image patches (2×2 minus mean) of natural scenes, from Huang and Mumford (1999).

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020



(a) sparse coding (b) Field of Experts (MRF)

Fig. 8. (a) Learned base functions in sparse coding by Olshausen and Field (1996). (b) Learned features in a Field of Experts (Markov random fields) by Roth and Black (2005).

observed statistics in Fig. 7a. Interestingly this model automatically selects features with the most information gain from a pre-defined set. It selects the Laplacian of Gaussian (LoG) and gradients, i.e. the second and first order image derivatives, whose histograms provide the more informative statistics against the uniform distribution (noise images). In 1999–2005, Roth and Black further enlarged the pre-defined set of features and let the model learn arbitrary features freely as in Olshausen and Field's experiments. They used a factorized version of the Zhu-Mumford model and called it the Field of Experts (FoE). Fig. 8b shows some examples of the top features learned by the FoE model which are close to checkerboard pattern. In our opinion, these features are different versions of the LoG and Gradient filters in a predefined patch size. For example, LoG and Gradients selected in (Zhu and Mumford, 1997) are also checkboard like feature in 3×3 or 2×1 patches respectively. But they are nothing like the Gabor patches in Fig. 8a.

How could two learning models, motivated by the same statistical observations, end up choosing two completely different sets of features as the most informative representation? Which set has the true or better features?

To comprehend this puzzle, we again have to remind readers of the implicit and explicit manifolds in the image space and the two manifold pursuit strategies discussed in Fig. 5. The sparse coding model looks for explicit base functions to reconstruct the observed images, starting from a constant image, and they capture the structures or primitives which are good at representing the explicit manifolds for image structures. In contrast, in learning the MRF models, one seeks for more informative features to distinguish natural images against noise images in pursuing the implicit manifolds, and therefore selects the LoG or gradients features. So both are correct, and they pursue the manifolds from two different ways, as we discussed in Fig. 5.

Fig. 9 illustrates this idea intuitively. Suppose we have some Gaussian clusters and we plot their eigenvalues in a decreasing order. If a cluster is of a very low dimension, like images of human faces, then we choose a few largest eigenvalues (see the blue curve) whose eigenvectors most effectively reconstruct the face.

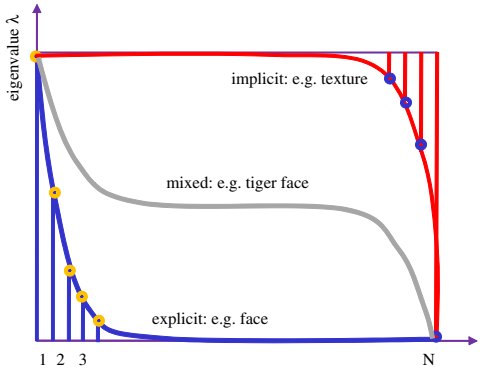


Fig. 9. Plot of eigenvalues in decreasing order for some Gaussian clusters of low dimension (blue curve) or high dimension (red curve). One should choose the largest eigenvectors for the low dimensional clusters and the smallest eigenvectors for the high dimensional clusters. PCA are special cases for the manifold pursuit and related feature selection.

This PCA example corresponds to the sparse coding model except that the base functions in the sparse coding model are over-complete and not orthogonal to each other. If a cluster is of a very high dimension, like stochastic textures whose eigenvalue plot will be like the red curve, then it won't be effective to choose the top eigenvectors, instead one ought to use the smallest eigenvalues whose eigenvectors are often the checkerboard patterns. Both could be considered as principal component analysis (PCA). The first is the usual case for constructing lower dimension clusters, and the second is the opposite case where we constrain the model from uniform. This is again like the two grading strategies used by teachers as we discussed in Section 1.4. We refer to a recent paper (Weiss and Freeman, 2007) for more formal account and comparison between the two learning schemes: sparse coding and Markov random fields.

2.4. Image scaling and perceptual transitions

Objects appear at arbitrary scales or sizes in images and evoke very distinctive perceptions and representations at different scales, Fig. 3 demonstrates the perception of maple leaves changes from primitives to texture. Although there is a long thread of research in image scale space, the first work that linked the perceptual transition to the entropy was done by Wu et al. in 2007–2008 (Wu et al., 2008).

Let W denotes the variables describing the whole scene, say the locations, shapes and appearance for tens of thousands of maple leaves. W generates the image $\boldsymbol{I} = g(W)$ deterministically by a rendering function $g()$. Many details are lost due to occlusion and image discretization in the rendering process. Visual perception is to estimate W from $\boldsymbol{I}$ following a posterior probability in the Bayesian framework,

$$W \sim p(W|\boldsymbol{I}). \tag{5}$$

The symbol $x \sim p(x)$ means "x follows a probability p".

Suppose at a certain scale, we have $\boldsymbol{I}$ and W following probability $p(\boldsymbol{I})$ and $p(W)$ respectively. The entropy of the posterior probability $p(W|\boldsymbol{I})$, averaged over the images $\boldsymbol{I}$ in a certain scale, reflects our uncertainty or inability to compute W precisely.

Definition 3. The *imperceptibility* of description W from an image I in an image ensemble is,

$$\text{imperceptibility}: \ \mathscr{H}(W|\boldsymbol{I}) = \sum_{W} \sum_{\boldsymbol{I}} p(W, \boldsymbol{I}) \log p(W|\boldsymbol{I}).$$

It was shown in (Wu et al., 2008) that the imperceptibility increases when the image is downsampled to a lower resolution where we denote the images by $\boldsymbol{I}_{-}$. For example, the shapes of maple leaves may not be visible. Thus our representation need to reduce its complexity from W to W_{-} by dropping or combining some variables, so that $\mathscr{H}(W_{-}|\boldsymbol{I}_{-})$ returns to below a certain threshold. The underlying assumption in this Bayesian inference is that visual perception, human or machine, do not handle variables of high ambiguities. For example, we don't attempt to recognize a person's face if the person is very far away, and even for close faces we do not define the exact boundary between the upper part of nose and the rest of the face.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

For a scene where the elements have a narrow range of sizes, such as the maple scenes, at a critical scale, a catastrophic transition (Wang and Zhu, 2008) occurs when we discard all the shape variables in W (describing the explicit manifolds for leaves) and switch to a statistical description $W_- = \mathtt{h}$ for implicit manifolds. In the latter case, $\boldsymbol{h}$ is an representation for the overall texture impression without noticing the individual primitives.

$$\text{explicit } (W = \{w\}, \boldsymbol{I}) \overset{\text{zoom-out}}{\rightarrow} \text{implicit } (W_- = \{\mathtt{h}\}, \boldsymbol{I}_-). \tag{6}$$

This intuitively explains the transition between explicit manifolds and implicit manifolds over the scaling process. For natural scenes which contain objects in a continuous scale following certain distributions, for example, the object radius $r \sim 1/r^3$ in the scene follows a density $p(r) \propto 1/r^3$ (Mumford and Gidas, 2001), the individual object must undergo the above perceptual transitions during the zooming process, but the overall local statistics averaged over the entire image remain invariant. We refer to Wang and Zhu (2008) and Wu et al. (2008) for more discussion.

To summarize this section, we have discussed a few puzzling topics of significant importance in the literature: (1) texture, texton and primal sketch; (2) high kurtosis and sparsity in natural image statistics; (3) seemingly contrast features learned by sparse coding vs Markov random fields; and (4) image scaling and perceptual transitions. All these issues are related to or explained by the explicit and implicit manifolds.

3. Information projection and manifold pursuit

In this section, we present the manifold learning framework and pursuit algorithm, following the introduction in Section 1.4.

3.1. Learning by information projection

Suppose we have a target manifold Ω_f governed by a probability $f(\boldsymbol{I})$, and it is represented by a number of observed examples. In the context of discriminative pattern recognition, they are called positive examples.

$$\Omega_f \supset \{\boldsymbol{I}_m^{\text{obs}}; m = 1, 2, \ldots, M\} \sim f(\boldsymbol{I}). \tag{7}$$

The objective of manifold pursuit is to find a sequence of models, starting from an initial reference model q, that would gradually approach $f(\boldsymbol{I})$,

$$q = p_0 \rightarrow p_1 \rightarrow \cdots \rightarrow p_k \text{ to } f \tag{8}$$

in terms of minimizing the Kullback–Lebler divergence $KL(f\|p)$.

We have introduced two pursuit strategies in Section 1.4, more specifically in Fig. 5. Both pursuit strategies follow the same learning procedure and principle, and only differ in their initial models q and the selected feature statistics.

Fig. 10 illustrates the learning procedure by information projection in the space of probability distribution. Note that we have been talking about image space and subspaces where each point is an image. Now we are dealing with a new space for probabilities where each point is a probability distribution $q(\boldsymbol{I}), p(\boldsymbol{I})$, or $f(\boldsymbol{I})$. In this probability space, the Kullback–Leibler divergence between two probabilities plays the same role as squared distance in the Euclidean space.

The learning proceeds iteratively. At each step k, we augment the current model p_{k-1} to p_k by adding some statistical constraint that p_{k-1} does not observe, i.e.

$$E_{p_k}[r_k(\boldsymbol{I})] = E_f[r_k(\boldsymbol{I})] \neq E_{p_{k-1}}[r_k(\boldsymbol{I})]. \tag{9}$$

$r_k(\boldsymbol{I})$ is some function of image $\boldsymbol{I}$, for example, the response of a Gabor filter at certain location in the image, or $r_k(\boldsymbol{I})$ can be a vector for the histogram of Gabor filter responses. In practice we can always approximate the expectation by the sample mean,

$$E_f[r_k(\boldsymbol{I})] \approx \bar{r}_k = \frac{1}{M}\sum_{m=1}^{M} r_k(\boldsymbol{I}_m^{\text{obs}}). \tag{10}$$

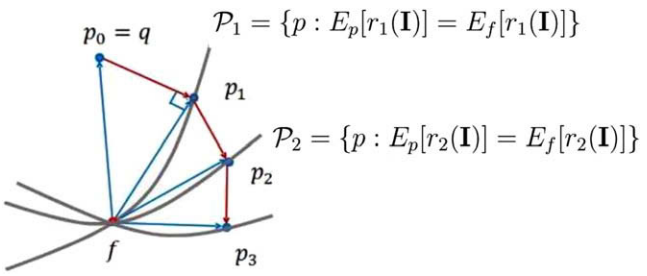


Fig. 10. Learning by information projection in the space of probabilities. Each point is a probability model. With more constraints added, the KL-divergence reduces monotonically.

Definition 4. We denote the set of all probabilities p that satisfy the condition as *candidate models* in the probability space,

$$\mathscr{P}_k = \{p : E_p[r_k(\boldsymbol{I})] = E_f[r_k(\boldsymbol{I})]\}. \tag{11}$$

$\mathscr{P}_k$ is represented by the curve in Fig. 10. For example, both f and p_k lie on $\mathscr{P}_k$ in Fig. 10 as they satisfy the constraint.

Now, we hope to project p_{k-1} to $\mathscr{P}_k$ perpendicularly, and thus find the p^* on $\mathscr{P}_k$ that is closest to p_{k-1},

$$p^* = \arg\min_{p \in \mathscr{P}_k} KL(p\|p_{k-1}). \tag{12}$$

Solving this constrained optimization problem by Lagrange multiplier, we have

$$p_k(\boldsymbol{I}; \Theta_k) = \frac{1}{z_k} p_{k-1}(\boldsymbol{I}; \theta_{k-1}) e^{-\lambda_k r_k(\boldsymbol{I})}, \tag{13}$$

λ_k is the parameter and z_k normalizes the probability to 1. This new model p_k may no longer observe the previous constraints, for example, p_k is not on $\mathscr{P}_{k-1}$.

The three points p_{k-1}, p_k and f form a triangle with right angle, as Fig. 10 shows. This is the famous Pythagorean theorem (see Della Pietra et al. (1997) and Csiszár and Shields (2004)).

Theorem 1. *For the exponential probability family $\{p_k\}$ constructed above, we have*

$$KL(f\|p_{k-1}) = KL(f\|p_k) + KL(p_k\|p_{k-1}), \quad \forall k. \tag{14}$$

As long as one can find informative features $r_k(\boldsymbol{I})$ so that $p_{k-1} \neq p_k$, then $KL(p_k\|p_{k-1}) > 0$ and the pursuit process converges to f monotonically.

After K iterations, we obtain a model,

$$p(\boldsymbol{I}; \Theta) = q(\boldsymbol{I}) \prod_{k=1}^{K} \frac{1}{z_k} e^{-\lambda_k r_k(\boldsymbol{I})}. \tag{15}$$

Or we can rewrite it as

$$\frac{p(\boldsymbol{I}; \Theta)}{q(\boldsymbol{I})} = \prod_{k=1}^{K} \frac{1}{z_k} e^{-\lambda_k r_k(\boldsymbol{I})} = \frac{1}{Z} \exp\left\{-\sum_{k=1}^{K} \lambda_k r_k(\boldsymbol{I})\right\}. \tag{16}$$

$Z = z_1 z_2 \cdots z_K$ and $\Theta = (\lambda_1, \ldots, \lambda_K)$. The above learning process sequentially projects the current model to a number of constrained spaces, and thus is called "information projection".

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

Each iteration of the learning process includes two steps.

1. Min-step: given the feature constraint r_k, we compute the parameter λ_k by finding the p_k on Ω_k that is closest to p_{k-1},

$$\lambda_k^* = \arg\min_{p_k \in \Omega_k} KL(p_k \| p_{k-1}). \quad (17)$$

2. Max-step: choosing an informative feature and statistics r_k, which reveals the biggest difference between p_k and p_{k-1}:

$$r_k^* = \arg\max KL(p_k \| p_{k-1}). \quad (18)$$

As Eq. (14) shows that $KL(p_k \| p_{k-1}) = KL(f \| p_{k-1}) - KL(f \| p_k)$, this step is a greedy way of minimizing the KL-divergence between f and p.

There are variations of the above information projection precess, which differ in two major ways.

1. The choice of the initial or reference probability q. For implicit manifold or texture modeling, one often starts with q being the uniform probability over the entire image space. In contrast, for explicit manifolds q is chosen to be focused on a point with an ϵ radius. See illustrations in Fig. 5. For the former case, the constraint is a "push" operator that shrinks the volume of Ω_p at each step, while for the latter case, the constraint is a "pull" operator that expands the volume of Ω_p at each step.
2. One may choose to accumulate the statistical constraints and thus let Ω_k observe all the existing statistical constraints,

$$\Omega_k = \{p : E_p[r_n(\boldsymbol{I})] = E_f[r_n(\boldsymbol{I})],\ n = 1, 2, \ldots, k\}. \quad (19)$$

The Pythagorean theorem holds true for the above construction. This leads to the minimax entropy learning scheme in (Zhu et al., 1997). It was also used in language modeling in (Della Pietra et al., 1997).

In the following, we show two case studies to illustrate the above learning process: one for implicit manifolds and one for explicit manifolds.

3.2. Case study I: the FRAME model for texture modeling

In the first case study, we illustrate the pursuit of implicit manifolds for texture modeling following the work of FRAME model (Zhu et al., 1997) and Julesz ensemble (Zhu et al., 2000).

The feature dictionary $\Delta^{\mathrm{im}} = \{F_k\}$ consists of Gabor sine and cosine filters, Laplacian of Gaussian (LoG) and gradient filters of varies sizes. The features extracted are filter responses $\langle F_k(x,y), \boldsymbol{I}\rangle$, where k indexes the scale and orientation of the filter, and (x,y) is the location. We assume that the texture is homogeneous, so we pool the filter responses over the image domain to form a histogram $r_k(\boldsymbol{I}) = \mathrm{h}_k(\boldsymbol{I})$ for each filter k. So we have the implicit manifolds for texture in a sequence,

$$\Omega_k = \{p : E_p(\mathrm{h}_i(\boldsymbol{I})) = E_f(\mathrm{h}_i(\boldsymbol{I})), i = 1, 2, \ldots, k,\ \forall (x,y)\}, \quad (20)$$

where f is the true distribution that generates the observed image, and $E_f(\mathrm{h}_i(\boldsymbol{I}))$ can be approximated by the corresponding histogram of the observed image, because f is stationary.

We initialize the learning process with $k = 0$, and we take p_0 to be the uniform distribution of the entire image space. Each step, set $k \leftarrow k+1$, we choose from Δ^{im} a new filter F_k and its histogram h_k which reveals the biggest difference between the current model p_{k-1} and the true distribution f, then we keep adding h_k to augment the model. As k increases, it gets closer to f as Fig. 5 illustrates.

To verify the learning process, we draw typical samples, by Markov chain Monte Carlo simulation, from the sequence of image manifolds $\Omega_k^{\mathrm{im}},\ k = 0, 1, 2, \ldots, 6$. These typical images are shown in Fig. 11. As k increases, the sampled images become perceptually more similar to the input image in (a).

As another way to visualize the learning process, we randomly choose 10,000 image patches of 10×10 pixels from the sample images at each learned stage $\Omega_k^{\mathrm{im}}, k = 0, 1, \ldots, 6$, as well as the original image. We applied PCA analysis and plot the eigenvalues in decreasing order for each manifold in the right panels of Fig. 11. The eigenvalues are scaled in the figure so the first eigenvalue would equal 1. The red-dotted curve is the eigenvalue plot of noise patches which is almost flat as expected. As k increases, the curves converges to the green curve for the input texture.

For a large image lattice, each learned p_k is equivalent to the uniform distribution over an implicit image manifold $\Omega_k^{\mathrm{im}} = \{\boldsymbol{I} : \mathrm{h}_i(\boldsymbol{I}) = \mathrm{h}_i, i = 1, \ldots, k\}$, where $\mathrm{h}_i(\boldsymbol{I})$ is the histogram of filter responses pooled over the image $\boldsymbol{I}$. The entropy of p_k is the log-volume of the ensemble Ω_k^{im}. This volume decreases as k increases. This example for pursuing implicit manifold confirms our intuitive ideas discussed in Fig. 9 and the grading strategy a in Section 1.4.

In the above example, the first two steps have the most effective compression along some dimensions, that is, the volume of the manifold shrinks at each step. In fact, the effectiveness of a filter F_k and thus its statistics h_k is measured by reduction of volume in logarithm,

$$\text{Info. gain} : \mathrm{Ig}^{\mathrm{im}}(\mathrm{h}_k) = \log \frac{|\Omega_{k-1}^{\mathrm{im}}|}{|\Omega_k^{\mathrm{im}}|}. \quad (21)$$

Because of the equivalence between entropy of p_k and the log-volume of Ω_k^{im}, the above information gain is

$$\mathrm{Ig}^{\mathrm{im}}(\mathrm{h}_k) = \mathrm{entropy}(p_{k-1}) - \mathrm{entropy}(p_k). \quad (22)$$

The pursuit of implicit manifold is a greedy process of entropy reduction process. This information gain is computed numerically by the following formula in (Zhu et al., 1997), following a Taylor expansion of entropy(p_k) at entropy(p_{k-1}),

$$\mathrm{Ig}^{\mathrm{im}}(\mathrm{h}_k) = N/2 \cdot (\mathrm{h}_k - \mathrm{h}_o)' \Sigma_o^{-1} (\mathrm{h}_k - \mathrm{h}_o), \quad (23)$$

where N is the number of images in $\boldsymbol{I}$, h_o is the statistics (histogram of filter responses) according to the current model p_{k-1}, and Σ_o is a covariance matrix of h_k. Thus the larger the distance between the observed statistics h_k and the current statistics h_o, the bigger the information gain. We refer to Zhu et al. (1997) for details of the modeling and learning process.

3.3. Case study II: the active basis model for object template

In the second case study, we illustrate the pursuit of explicit manifolds for learning deformable templates using the Active Basis model in our recent work (Wu et al., 2007, 2009). Suppose we observe images $\{\boldsymbol{I}_m^{\mathrm{obs}}, m = 1, \ldots, M\}$ from an object category. For simplicity, let us assume that these images are of the same size, and the objects in these images appear at the same position and scale and in the same pose. Our goal is to learn a common template from these training images.

Similar to the FRAME model, the feature dictionary $\Delta^{\mathrm{ex}} = \{F_k\}$ consists of Gabor sine and cosine filters. However, in the Active Basis model, the Gabor filters play the role of basis functions for spanning the explicit manifolds. More specifically, let $B_{x,y,s,\alpha}$ be the Gabor wavelet centered at location (x,y), at scale s and orientation α. Then the active basis representation is as follows:

$$\mathrm{B} = (B_k = B_{x_k, y_k, s, \alpha_k},\ k = 1, \ldots, K). \quad (24)$$

$\boldsymbol{B}$ is viewed as a deformable template, because we allow each basis function B_k to deform to $B_{x_k + \Delta x_{m,k}, y_k + \Delta y_{m,k}, s, \alpha_k + \Delta \alpha_{m,k}}$, where $(\Delta x_{m,k}, \Delta y_{m,k})$ is the shift of B_k in location, and $\Delta \alpha_{m,k}$ is the shift of B_k in orientation. We restrict the shifts $(\Delta x_{m,k}, \Delta y_{m,k}, \Delta \alpha_{m,k})$ to be

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

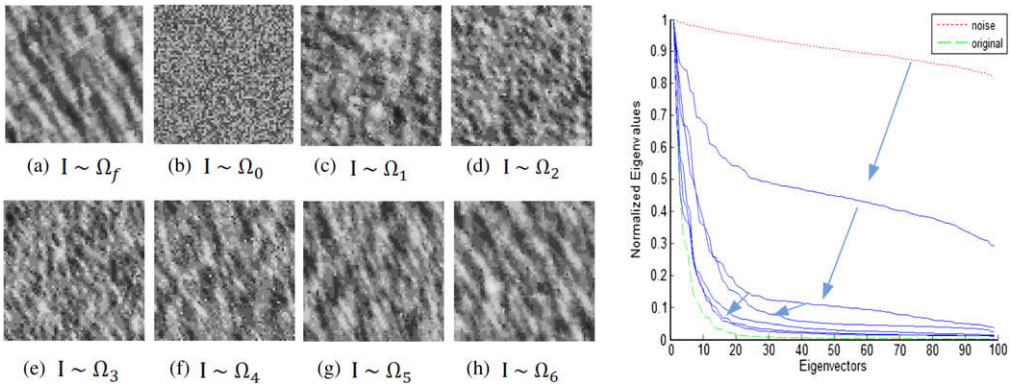


Fig. 11. (a) is the original input image, (b) is the initial noise image, and (c)–(h) are the synthesized image after adding 1–6 features into the model. The right panel plots of eigenvalues for the images patches from the synthesized image sequence. The red dotted line is for the noise image, and green broken line for the original input image. Eigenvalues are scaled so that the first eigenvalue would equal to 1.

within a limited range. In detecting the template, these local deformations are computed through a local maximization process that chooses the maximum response r_k for B_k over the deformation range.

The Active Basis $\boldsymbol{B}$ defines an explicit manifold of images,

$$\Omega^{\text{ex}} = \left\{ \boldsymbol{I} : \boldsymbol{I} = \sum_{k=1}^{K} c_k B_{x_k+\Delta x_k, y_k+\Delta y_k, s, \alpha_k+\Delta\alpha_k} \right\}, \tag{25}$$

where $w = (c_k, \Delta x_k, \Delta y_k, \Delta\alpha_k, k = 1, \ldots, K)$ are the variables. This explicit manifold is highly non-linear, because of the shifts in locations and orientations of the basis elements.

Though Ω^{ex} is spanned by K independent basis functions, these basis functions are like the spikes pointing to various dimensions and thus the volume of Ω^{ex} is quite small even if K is large. In the following, we define a probability distribution for Ω^{ex} following the information projection procedure.

We construct a sequence of constraints on the individual response of basis function B_k.

$$\mathscr{P}_K = \{p : E_p[s(r_k(\boldsymbol{I}))] = E_f[s(r_k(\boldsymbol{I}))], k = 1, 2, \ldots, K, \}, \tag{26}$$

where $s(r)$ is a sigmoid transformation that increases monotonically from 0 to a saturation level, and $E_f[s(r_k(\boldsymbol{I}))]$ can be estimated by the sample mean,

$$E_f[s(r_k(\boldsymbol{I}))] \approx \frac{1}{M} \sum_{m=1}^{M} s(r_k(\boldsymbol{I}_m^{\text{obs}})). \tag{27}$$

The learning process leads to the following model,

$$p(\boldsymbol{I}; \Theta) = q(\boldsymbol{I}) \prod_{k=1}^{K} \frac{1}{z_k} e^{-\lambda_k s(r_k(\boldsymbol{I}))}. \tag{28}$$

$\Theta = (\lambda_1, \ldots, \lambda_K)$ is the parameters. We choose $q(\boldsymbol{I})$ to be a uniform distribution in the image space centered at a flat image with small perturbations. For example, we may take patches from natural images and these patches are dominated by flat regions as we shall show in the next section.

Fig. 12 illustrates the pursuit process for a deer template. It consists of 50 basis functions represented by strokes.

Intuitively, if more image instances have a feature (i.e. high response r_k) at a common location and orientation, then the corresponding basis function B_k has a higher information gain.

In the following, we briefly derive the information gain for the Active Basis model in Eq. (28), and refer to Wu et al. (2009) for more details. In (Wu et al., 2009), some simplification steps are taken by assuming the basis function are non-overlapping and conditionally independent given the overall alignments. Thus we have both p and q in factorized forms,

$$\frac{p(\boldsymbol{I})}{q(\boldsymbol{I})} = \prod_{k=1}^{K} \frac{p(r_k(\boldsymbol{I}))}{q(r_k(\boldsymbol{I}))}, \tag{29}$$

with $p(r_k(\boldsymbol{I}))$ and $q(r_k(\boldsymbol{I}))$ being 1D probabilities. Thus the information gain for selecting a basis function B_k is simply,

$$\text{Ig}^{\text{ex}}(B_k) = KL(p(r_k) \| q(r_k)). \tag{30}$$

Fig. 13 illustrates the information gain for an car example. In this example, $q(r_k)$ is the same for all r_k and is focused around zero,

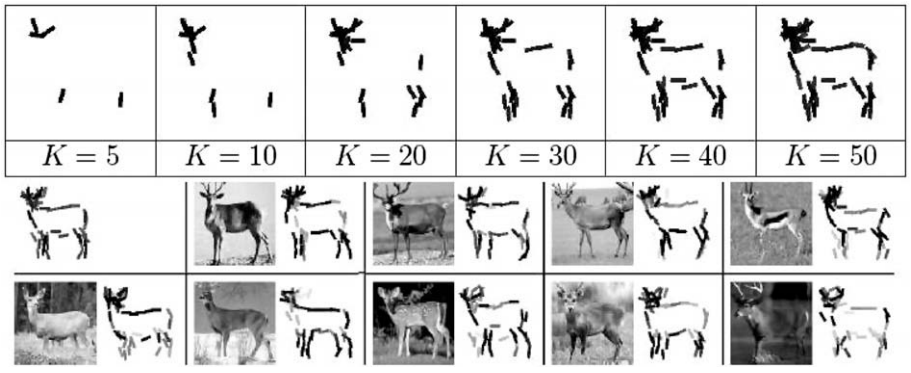


Fig. 12. Top: the process of learning the Active Basis templates for a deer image. Basis function vectors are selected in the order of their information gains. Bottom: nine examples of the deer images with their deformed templates on the right.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

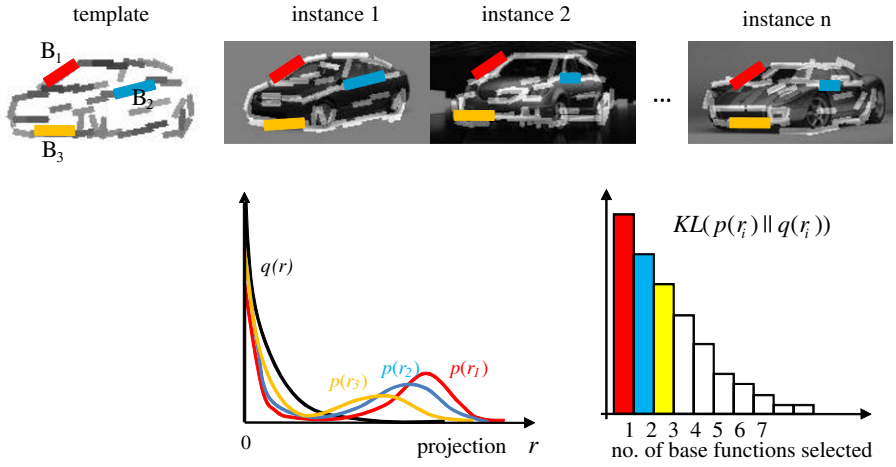


Fig. 13. Active Basis pursuit and measuring the information gain for each basis function. The three basis functions B_1, B_2, B_3 represent some common car structures and their responses follow distributions $p(r_1), p(r_2)$ and $p(r_3)$ respectively in contrast to the null model $q(r)$. The KL-divergence between $p(r_k)$ and $q(r_k)$ measures the information gain of choosing B_k.

while $p(r_k)$ may have a bump due to high responses at the observed image instances.

Following the parametric model in Eq. (28), we have

$$\frac{p(r_k)}{q(r_k)} = \frac{1}{z_k} e^{-\lambda_k s(r_k(\boldsymbol{I}))}. \tag{31}$$

The information gain for choosing B_i is then,

$$\mathrm{Ig}^{\mathrm{ex}}(B_k) = \lambda_i E_f[s(r_k(\boldsymbol{I}))] - \log z_k. \tag{32}$$

$E_f[s(r_k(\boldsymbol{I}))]$ is estimated in Eq. (27) and λ_k and z_k are scalars which can be estimated by Monte Carlo methods on positive training images.

For the source code, data and further details of the above results, please refer to http://www.stat.ucla.edu/~ywu/ActiveBasis.html.

3.4. Manifold pursuit: a push and pull process

So far, we have introduced manifold learning by information projection, and shown examples for pursuing both implicit and explicit manifolds. The model pursuit in the probability space in Fig. 10 is an abstract view which may be less intuitive. Now we further discuss the pursuit in the image space and interpret the manifold pursuit as a push and pull process.

Let Ω_f be the underlying image manifold that we are pursuing and it is governed by a probability $f(\boldsymbol{I})$, and the current model $p(\boldsymbol{I})$ corresponds to an image manifold Ω_p. When $p = f, \Omega_p$ coincides with Ω_f, we say the manifold is "caught" successfully. As the image space is of very high dimensions, we cannot visualize the shape of Ω_f or Ω_p, instead we project them to lower dimensional space and observe their marginal distributions. For example, Fig. 7 shows the 1D and 3D marginal statistics for the set of natural images. Let r be the response of an image to a filter (an axis in the image space), and for simplicity we denote the marginal statistics of r by $E_p[r]$ and $E_f[r]$ respectively (note they are not the expectations of the response), and they are shown by the dashed blue curves and solid red curves respectively in Fig. 14.

If p is matched to f, then a necessary but not sufficient condition is to match their marginal probabilities

$$E_p[r] = E_f[r]. \tag{33}$$

This is exactly what we do in each iteration of the information projection process.

1. *The push process*. In pursuing implicit manifolds like texture in case study I, one starts from a uniform probability p_0 and thus $E_p[r]$ is "fat" while $E_f[r]$ is "slim" and peaked at a single point. We need to push $E_p[r]$ to fit $E_f[r]$ by compressing the volume of Ω_p. In an extreme case, if $E_f[r]$ is so slim and becomes an impulse function (or direct delta function), then its means Ω_f is perpendicular to the filter. So the pushing process compresses a whole dimension.
2. *The pull process*. In pursuing explicit manifolds like the active basis in case study II. $E_f[r]$ has a much longer tail than $E_p[r]$ because the structures aligned with the filter generate large responses. The matching process pulls $E_p[r]$ to fit $E_f[r]$, and thus produces a spike along the axis of the filter, just like the spikes shown in Fig. 7b.

The above push and pull processes are refined interpretations to the two manifold pursuit strategies discussed in Fig. 5 and show that we can use the statistical constraints to both compress and

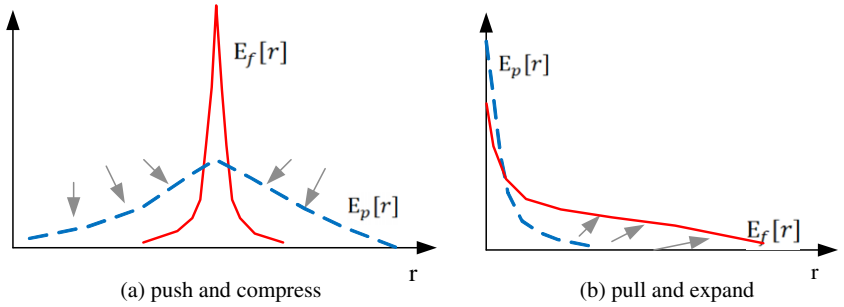


Fig. 14. Two 1D marginal statistics. The blur dashed curves are the marginal probability $E_p[r]$ of model p, and the red curves are the marginal probability $E_f[r]$ of model f. (a) the push process, and (b) the pull process. See text for interpretation.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

expand the manifolds. In the hybrid template learning case which we shall introduce in Section 5.2, the push and pull processes will alternate to reshape Ω_p so that it matches to Ω_f. The convergence is guaranteed by the Pythagorean theorem.

4. Pursuing atomic manifolds in the space of image patches

In this section, we pursue the explicit and implicit manifolds in the space of image patches. We choose small patches as they mostly belong to either a pure explicit or a pure implicit manifold. These manifolds are the atomic structures in the image space, contain the prevailing textures and textons in our visual environments, and they are composed to form large subspaces for complex image patterns. We are particularly interested in knowing the most popular atomic manifolds in the space of natural (or daily) images and sorting them according to their information gains.

4.1. The space of image patches

Let Ω denote the space of all image patches of N pixels, say, $N = 11^2$–19^2. We are interested in studying a subspace $\Omega_f \subset \Omega$ governed by a probability $f(\boldsymbol{I})$. For example, Ω_f contains patches cropped from generic images that we see in daily life or photos that we download from the Internet. We observe a large number of patches $\{\boldsymbol{I}_m^{\text{obs}} : m = 1, 2, \ldots, M\}$ and assume that Ω_f is made up of both explicit and implicit manifolds, plus some leftover image patches that are rare and complex patterns

$$\Omega_f = \cup_{s=1}^{S} \Omega_s \cup \Omega_{\text{leftover}}, \tag{34}$$

Ω_s denotes an implicit or explicit manifold, and the "leftover" image patches in Ω_f are explained by the background model $q(\boldsymbol{I}) = \text{Unif}[\Omega]$, i.e. uniform probability over the entire space Ω.

We assume that Ω_s are non-overlapping, and estimate the frequency of each manifold by

$$f_s = E_{f(\boldsymbol{I})}[\mathbb{1}(\boldsymbol{I} \in \Omega_s)] \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(\boldsymbol{I}_m \in \Omega_s), \tag{35}$$

$\mathbb{1}()$ is an indicator function. This is a reasonable assumption in high dimensional spaces. Sometimes a low dimensional subspace (cluster) is submerged in a cluster of higher dimensions, the volume of the formal is negligible in comparison to the later.

Our objective is to pursue a probability model p to approximate$f$, with an initial uniform distribution q over Ω. The Kullback–Leibler divergence is

$$KL(f\|q) = -\sum_{s=0}^{S} f_s \log \frac{f_s}{|\Omega_s|} + \mathbb{E}_f[\log f(\boldsymbol{I})]. \tag{36}$$

So we can measure the information gain of Ω_k by

$$l_s = f_s \log \frac{f_s}{|\Omega_s|}. \tag{37}$$

Note that $|\Omega| = L^N$ is a constant with L being the number of gray levels. Therefore the pursuit process seeks the manifold Ω_s with large frequency (i.e. heavy) and small volume (i.e. tight).

In the following, we shall calculate l_s for the explicit and implicit manifolds. To do so, we need to estimate their volumes $|\Omega_s|, s = 1, 2, \ldots, S$.

(1) *Volumes of the explicit manifolds.* For an explicit manifold, the images are represented by an Active Basis model with $K = 1, 2, 3, 4$ strokes, such as edges, bars, junctions, and cross etc. $\Omega_s = \{\boldsymbol{I} : \boldsymbol{I} = g_s(w_s) + \epsilon\}$, with $w_s = (w_{s,1}, \ldots, w_{s,K})$, we define its volume as

$$\log |\Omega_s| = \sum_{i=1}^{K} L_i, \tag{38}$$

where L_i is the log-volume of the space that $w_{s,i}$ spans, or the coding length of $w_{s,i}$.

$$L_i = \log |\Omega_{x_i}| + \log |\Omega_{y_i}| + \log |\Omega_{\theta_i}| - \sum_{a_i} p(a_i) \log p(a_i),$$

where $\Omega_{x_i} \times \Omega_{y_i} \times \Omega_{\theta_i}$ is the deformation space of the i-th stroke, and $p(a_i)$ is the probability for the contrast a_i. We can also measure the information gain based on the likelihood ratio of the fitted active basis model as discussed in Section 3.3.

(2). Volumes of the implicit manifolds. For an implicit manifold $\Omega_s = \{\boldsymbol{I} : \mathtt{H}_s(\boldsymbol{I}) = \mathtt{h}_s + \epsilon\}$. $\mathtt{h}_s$ denotes the normalized histograms. Its volume can be estimated according to the information gain in Eq. (23):

$$\log |\Omega_s| = \log |\Omega| - N/2 \cdot (\mathtt{h}_s - \mathtt{h}_o)' \Sigma_o^{-1} (\mathtt{h}_s - \mathtt{h}_o), \tag{39}$$

where $\mathtt{h}_o$ is the histograms of filtered responses from noise images, and $\log |\Omega| = N \log L$. We can also compute the information gain using the method discussed in Section 5.2.

4.2. The pursuit algorithm

For the observed image patches in Ω_f, we first apply an EM-type clustering algorithms using the implicit and explicit models separately. Thus we decompose Ω_f into a set of candidate explicit manifolds (clusters) $\boldsymbol{\Omega}^{\text{ex}} = \{\Omega_s^{\text{ex}}\}$, and also decompose Ω_f into a set of candidate implicit manifolds (clusters) $\boldsymbol{\Omega}^{\text{im}} = \{\Omega_s^{\text{im}}\}$. The two sets of manifolds overlap with each other.

These candidate manifolds are ranked by their information gains l_s discussed in the previous subsection. Currently in this experiment, we calculate the information gain of the explicit manifold based on the log-likelihood ratio of the fitted active basis model. We iteratively select the manifold that has a maximum information gain, and mark its cluster members (image patches) as "explained". Then patches in this cluster are then eliminated from all other clusters whose information gains are re-calculated to only count the un-explained ones. This procedure is carried on until the gain of the newly selected manifold is small than a threshold, or when all example image patches are "explained".

Input: $\boldsymbol{\Omega}^{\text{ex}} = \{\Omega_1^{\text{ex}}, \ldots, \Omega_M^{\text{ex}}\}$ and $\boldsymbol{\Omega}^{\text{im}} = \{\Omega_1^{\text{im}}, \ldots, \Omega_N^{\text{im}}\}$
Output: $\boldsymbol{\Omega} = \{\Omega_1, \ldots, \Omega_S\}$
1. Initialize $\boldsymbol{\Omega} = \emptyset, S \leftarrow 0$
2. Repeat
3. Calculate information gain l_k for all Ω_k^{ex}
4. Calculate information gain l_k for all Ω_k^{im}
5. Select Ω_K with highest gain $l_{\max}$, remove it from $\boldsymbol{\Omega}^{\text{ex}}$ or $\boldsymbol{\Omega}^{\text{im}}$
6. For each $\Omega_k \in \boldsymbol{\Omega}^{\text{ex}} \cup \boldsymbol{\Omega}^{\text{im}}$
7. $\Omega_k \leftarrow \Omega_k - \Omega_K$.
8. $K \leftarrow K + 1$
9. Until $l_{\max} < \tau$, or $\boldsymbol{\Omega}^{\text{ex}} = \emptyset$ and $\boldsymbol{\Omega}^{\text{im}} = \emptyset$

4.3. Pursuit experiment I: manifolds in generic images

In this section, we report the manifold pursuit experiments on three sets of images. For more detailed description refer to a Ph.D. dissertation (Shi, 2008).

The images in this experiment are from Flickr.com, Corel image database and our own collection. The images are approximately 400×600 pixels, and we crop image patches of 19×19 pixels.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

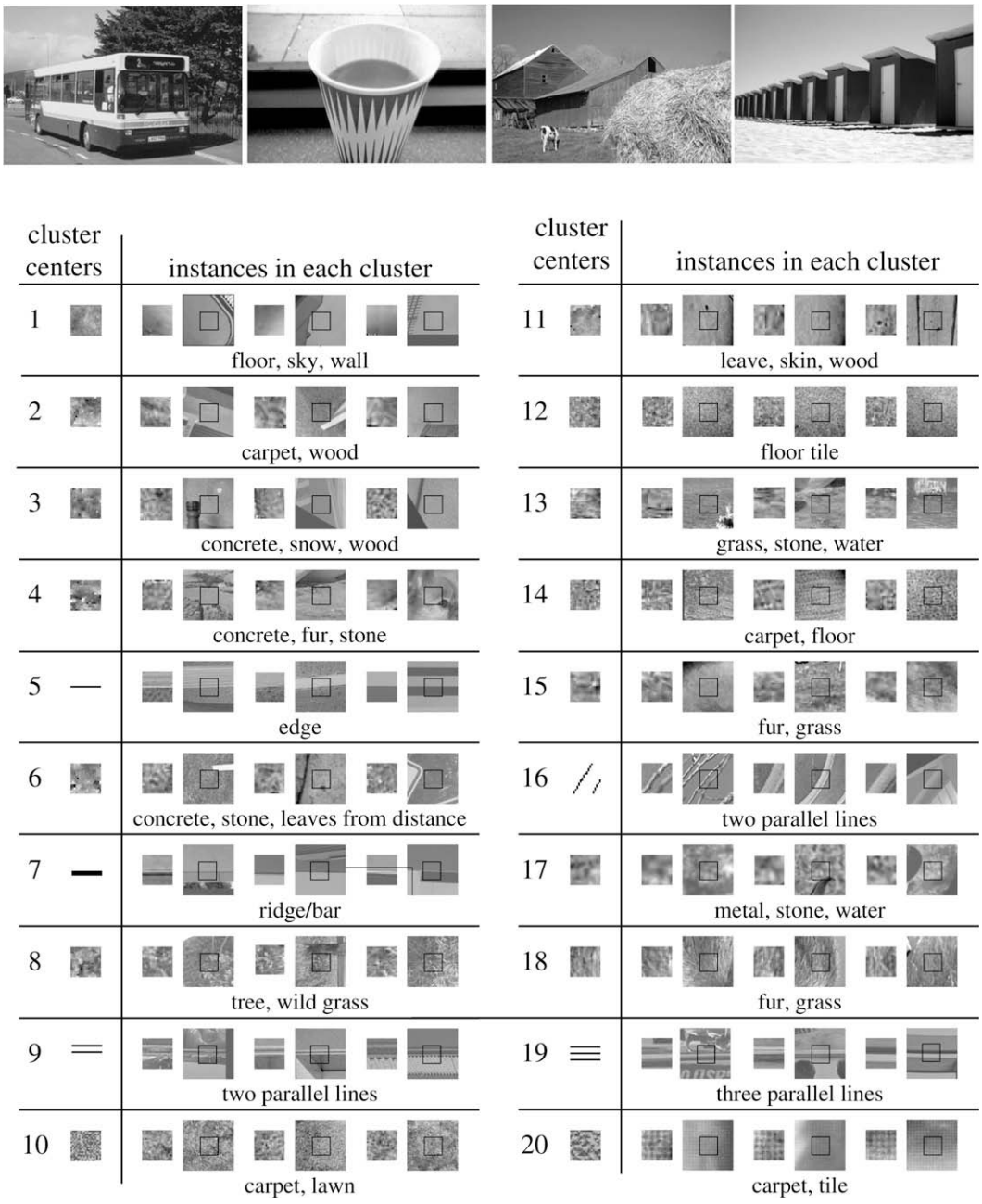


Fig. 15. Top: examples of generic images. Bottom: the top 20 clusters with prototypes of the manifolds sequentially selected, and the instances of image patches on these manifolds. The two types of manifolds are selected in a mixed order. For illustration purpose, we have beautified the explicit shape templates by minimally adjusting the positions of edge elements.

The first image set consists of 200 generic natural images. Fig. 15 shows examples of collected natural images, together with top 20 image manifolds learned using the method described in Section 4.2. The left most column displays a template or prototype image for each manifold, and to the right of it we show three of its instances. Each instance is shown with its context.

The two types of manifolds are selected in a mixed order. Fig. 16 shows the relative frequencies f_k and information gains l_k of the sequentially selected manifolds. The frequencies are highly uneven. The top selected manifolds are dominated by implicit manifolds. Only 5 of the top 20 manifolds are explicit, and they combine to contribute to less than 15% of the total image patches. Clean boundaries of objects are often much more informative perceptually (Marr, 1982) and useful in object recognition. However, the explicit manifolds we found using generic image set only contain very simple structures such as edges, bars and parallel lines. We did not obtain complex structures such as T-junctions, which is in part due to their extremely low frequencies.

To see what types of manifolds are important for describing images with man-made objects, we selected 30 such images as our second image set, which includes both indoor and outdoor scenes with buildings, furniture etc. The top 15 manifolds found are shown in Fig. 17. Comparing with the manifolds shown in Fig. 15, the most glaring difference is that there are far more explicit manifolds. In addition to edges and bars, "L"-junctions and "T"-junctions are also found. This demonstrates that explicit manifolds are more prominent in images containing man-made objects than they are in more generic images. We would like to point out that even though "L"-junctions and "T"-junctions have much higher frequencies in this image set, their frequencies are still small.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

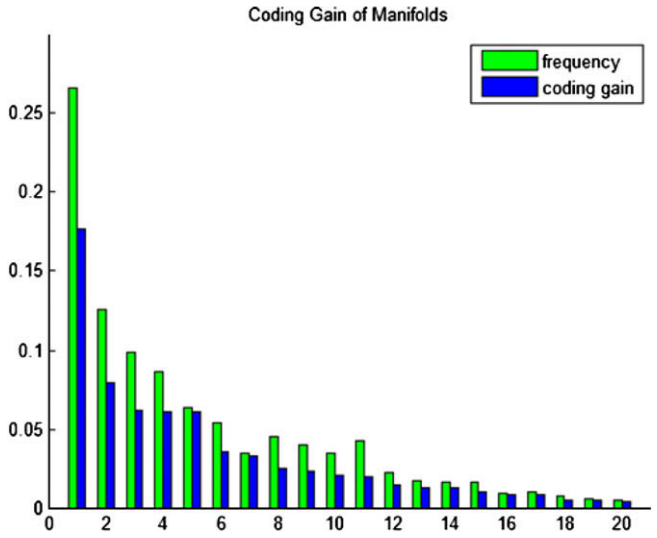


Fig. 16. Plot of frequencies and information gains of the 20 sequentially selected manifolds.

4.4. Pursuit experiment II: manifold transition in scales

For the third image set, we study the images with visual patterns of different scales. Wu et al. (2008) studied this problem using a dead leaves model (Matheron, 1975) by generating sets of 512×512 images containing multiple occluding squares of varies sizes, with one of the sets shown in the bottom of Fig. 18. This simulates the maple leaf example shown in Fig. 3. The image at the first scale is generated by randomly placing squares of various sizes onto the image. The side length of the squares is s and it takes values from [64,256] with the frequency proportional to $1/s^3$. That is, the large squared are occluded by much more squares of smaller size. The squares are placed at random until all pixels are covered at least once. The pixel intensity$t$ is constant within each square, with t randomly sampled from a uniform distribution [0,255]. Image in each subsequent scale is a downsampled version of the image in the previous scale. The resolution is lowered by 1/2, and the intensity of each pixel (x, y) is generated by taking the average of the four pixels $(2x-1, 2y-1), (2x-1, 2y), (2x, 2y-1), (2x, 2y)$ from the previous scale. In their study, they found that the per pixel entropy of the images increases as the scale increases. As

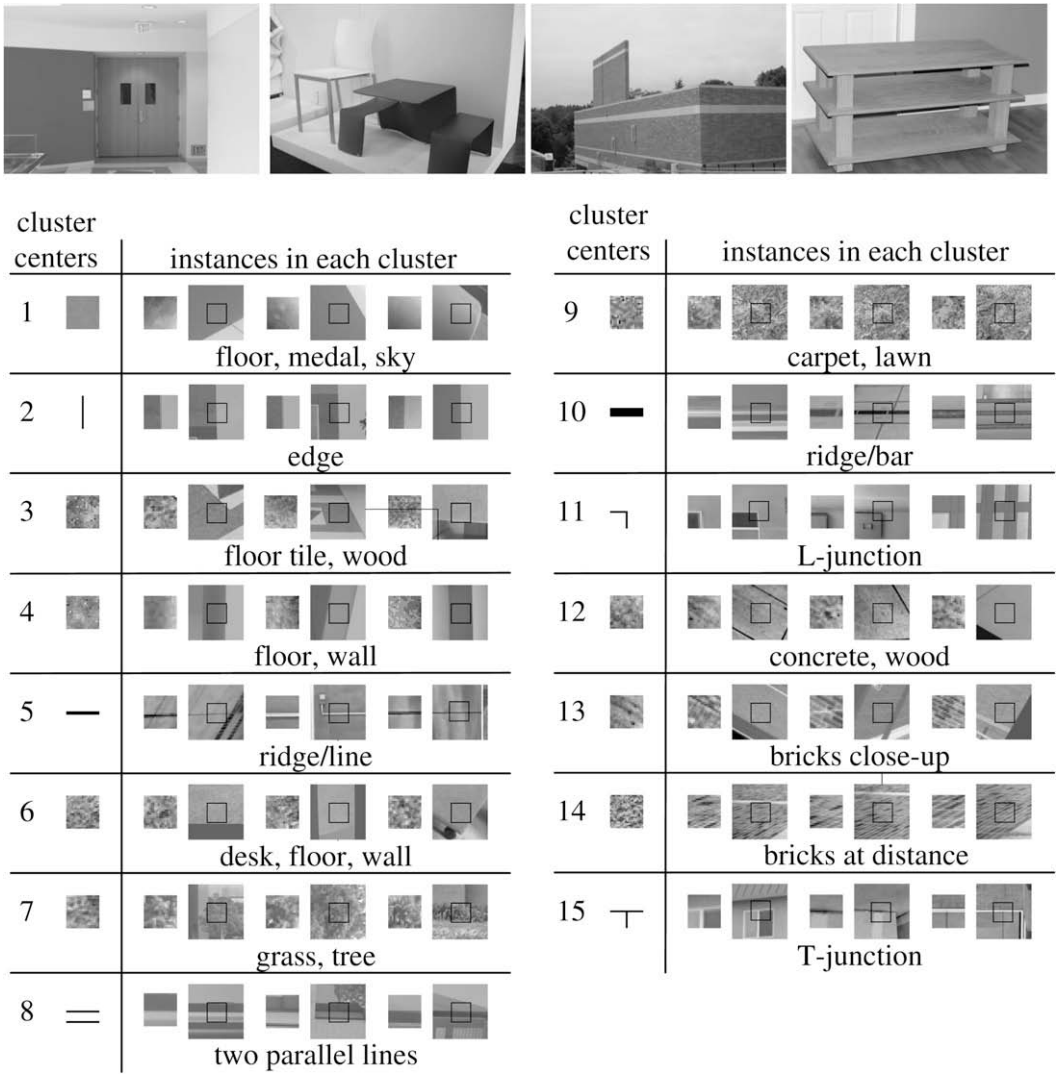


Fig. 17. The prototypes of the manifolds sequentially selected from images primarily compose of man-made objects, The two types of manifolds are selected in mixed order, but explicit manifolds appear much more often, and the implicit manifolds are mostly flat textures.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

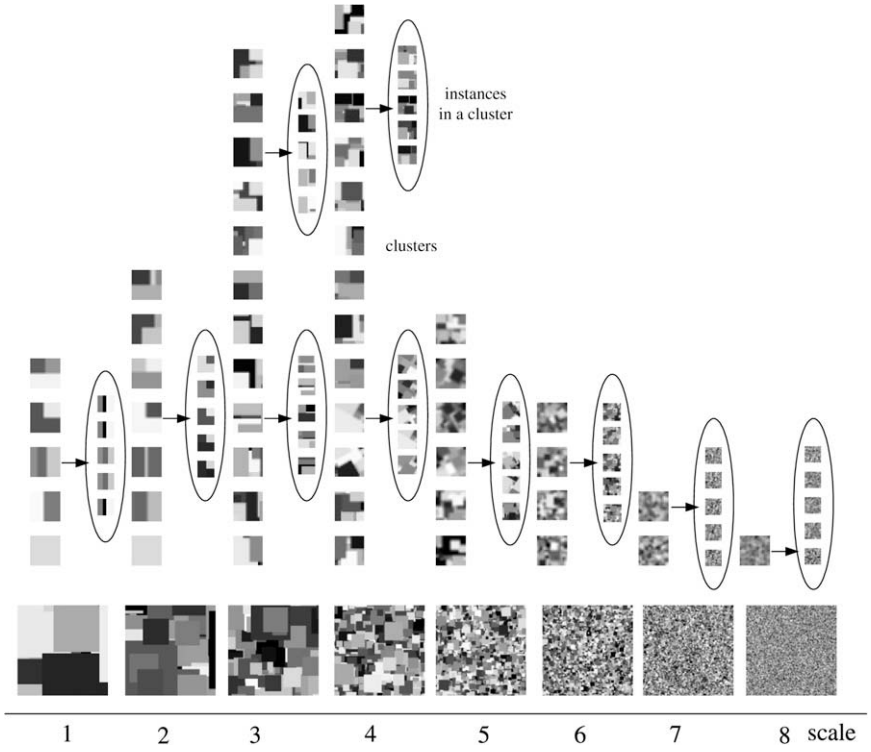


Fig. 18. Bottom: a sequence of images of occluding squares. The resolution of each image is 1/2 of the previous image. Top: examples of manifolds found at each scale, additional patch instances are displayed for selected manifolds to show the within-manifold variance.

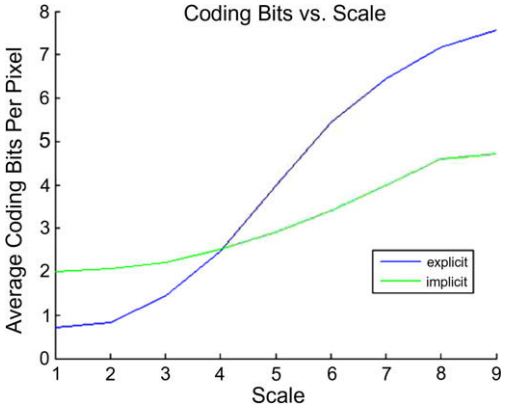


Fig. 19. Coding length changes over scales for the implicit and explicit manifolds.

we discussed before, the images go from cartoon like pictures in a low entropy regime, to an object like middle entropy regime, and to a texture like high entropy regime, and end at Gaussian noise at scale 8. At scale 8, each pixel is the normalized sum of $2^7 \times 2^7$ pixels at scale 1. Even the largest squares are longer destroyed in the downsampling and discretization process.

Studying this dataset reveals some interesting results for transitions of manifold as well as our models for representation.

Firstly, we estimate the number of manifolds needed to encode at each scale. Again we perform the manifold pursuit procedure, and we allow the explicit and implicit manifolds compete against each other in order to form the optimal set of manifolds. Manifolds identified in each scale are shown in Fig. 19. We only display the top manifolds with frequencies greater than 0.5%. These manifolds give a good picture of how many manifolds is needed because these top manifolds already cover the vast majority of the images (greater than 95%).

As Fig. 18 shows, scales 1–2 contain only explicit manifolds, and scales 6–8 are exclusively implicit manifolds. The scales 3–5 have both types. In addition, the number of manifolds peaks around scale 4. This means that we only need a few manifolds in our dictionary to efficiently code very high or very low resolution images, but more manifolds are needed to code images of middle resolution. This suggests that the "middle resolution", which is also where typical patterns of visual objects appear, contains most interesting information.

Secondly, we compare the coding efficiencies of the two types of manifolds at different scales by computing the coding efficiency of the two manifolds. We tabulate the total number of pixels T_k covered by member image patches of each explicit manifold Ω_i^{ex} or implicit manifold Ω_i^{im}. Overlapping pixels contained by multiple image patches are counted as $1/N$ of a pixel toward each of the N image patches that cover it. Pixels that are not covered by any image patches belonging to an explicit manifold or implicit manifold are placed into the default background manifold Ω_0, where pixels are coded with the maximum coding length of 8. Given this information, the average description length per pixel for the whole image by using only explicit manifold and only implicit manifolds can be compute by $L = \sum_{k=1}^{K} \frac{T_k}{N^2} l_k$, where l_k is the description length of manifold Ω_i, and N^2 represents the total number of pixels in the image.

The coding lengths for both explicit and implicit methods increase as the scale increases, because the entropy of the images increases as we increase scale, thus it is inevitable that manifolds obtained from high-resolution images will have larger volumes, regardless how we model it. But it is clearly more efficient to use explicit manifolds to represent the high resolution images, and use the implicit manifolds to represent the low resolution images. The two curves intersect between scale 4 and 5, indicating that the

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

coding efficiencies of the two manifolds are comparable for images at the medium resolutions.

5. Integrating the explicit and implicit manifolds

In this section, we show two cases for integrating image patches from the explicit and implicit manifolds to form larger representations: (1) the primal sketches for representing generic images, and (2) the hybrid image templates for representing object recognition.

5.1. Case study III: the primal sketch model for generic image representation

As reviewed in Section 2.1, the primal sketch was conjectured by Marr (1982) as a symbolic and perceptually lossless representation for generic images and it was considered the perceptual model for early vision. A mathematical model was proposed by Guo et al. (2007). We briefly show how this model integrates the texture and textons (or image primitives), or in our terms, patches from explicit and implicit manifolds.

Fig. 20 illustrates the primal sketch model from Guo et al. (2007), the image domain Λ is divided into two disjoint parts: the sketchable part $\Lambda_{\rm sk}$ for structures in (e) and non-sketchable part for textures in (d):

$$\Lambda = \Lambda_{\rm sk} \cup \Lambda_{\rm nsk}, \quad \boldsymbol{I} = (\boldsymbol{I}_{\rm sk}, \boldsymbol{I}_{\rm nsk}). \tag{40}$$

The sketchable part is further divided into a number of domains (usually 5×11 pixels) for image primitives, such as blobs, edges, bars, and junctions in (b):

$$\Lambda_{\rm sk} = \cup_i \Lambda_{{\rm sk},i}, \quad \text{with } \boldsymbol{I}_{{\rm sk},i} = B_i^*(w_i) + \epsilon, \quad \boldsymbol{I}_{{\rm sk},i} \in \Omega_{i*}^{\rm im}. \tag{41}$$

In the above notation, each patch $\boldsymbol{I}_{{\rm sk},i}$ is mapped to (or coded by) a closest explicit manifold $\Omega_{i*}^{\rm ex}$ with a primitive B_i^* indexed by its explicit variables $w_i = (x_i, y_i, \theta_i, a_i)$ for translation, rotation, and contrast.

The non-sketchable part is also divided into a few texture regions shown by different gray levels in (c). The shape of these regions may be irregular unlike the primitives. The image in each region belongs to an implicit manifold since $\boldsymbol{I}_{{\rm nsk},j}$ under its boundary condition $\boldsymbol{I}_{\rm sk}$ has certain statistics

$$\Lambda_{\rm nsk} = \cup_j \Lambda_{{\rm nsk},j}, \quad \text{with } \mathrm{h}(\boldsymbol{I}_{{\rm nsk},j} | \boldsymbol{I}_{\rm sk}) = \mathrm{h}_j^* + \epsilon, \quad \boldsymbol{I}_{{\rm nsk},j} \in \Omega_{j*}^{\rm im}. \tag{42}$$

In the above notation, each texture region $\boldsymbol{I}_{{\rm nsk},j}$ is mapped to (or coded by) a closest implicit manifold $\Omega_{j*}^{\rm im}$ with statistics h_j^*. The texture areas can be synthesized by sampling the images from the implicit manifold $\Omega_{j*}^{\rm im}$ using a Markov Chain Monte Carlo method. The sampling is conditional on the sketchable part $\boldsymbol{I}_{\rm sk}$.

In Fig. 20, both structures in (e) and textures in (d) are represented by explicit and implicit manifold respectively. The two parts are combined to yield an image in (f), which is perceptually almost lossless to the input image in (a), although the texture parts are very different in terms of pixel intensities. We refer to Guo et al. (2007) for details of the primal sketch model. This model is also related to the well-known Mumford-Shah model that integrates a MRF term for smoothness and an edge term for boundary (Mumford and Shah, 1989).

In summary, the primal sketch model decomposes the image into patches and each patch is indexed to either an explicit manifold or an implicit manifold. This is a very parsimonious representation and it needs much less bytes than jpeg coding (see the bit count in (Guo et al., 2007)). It has the following properties in comparison to vector quantization and image coding in the literature.

Firstly, one may view the primal sketch representation as a vector quantization process. Unlike conventional vector quantization where the reconstruction errors are measured in a single space, say the squared distance in Euclidean space for image coding, the reconstruction errors in primal sketch are measured in differ metrics. A primitive patch is reconstructed to an ϵ precision in the explicit manifold for errors in location, orientation and intensity difference. A texture patch is reconstructed to an ϵ precision in an implicit manifold measured by histogram difference.

Secondly, it is drastically different from wavelet coding or sparse coding and is more effective in image reconstruction. For structure patches, the image primitives are much sparser (i.e. more over-complete) than the wavelet dictionary. Each patch is represented by only one primitive that can account for sharp object boundaries. For texture patches, the image is reconstructed up to a statistical histogram that produces textures perceptually equivalent to the input texture areas.

5.2. Case study IV: learning hybrid image templates for object recognition

In the fourth case study, we show another application of combining explicit/implicit manifolds for representing hybrid image

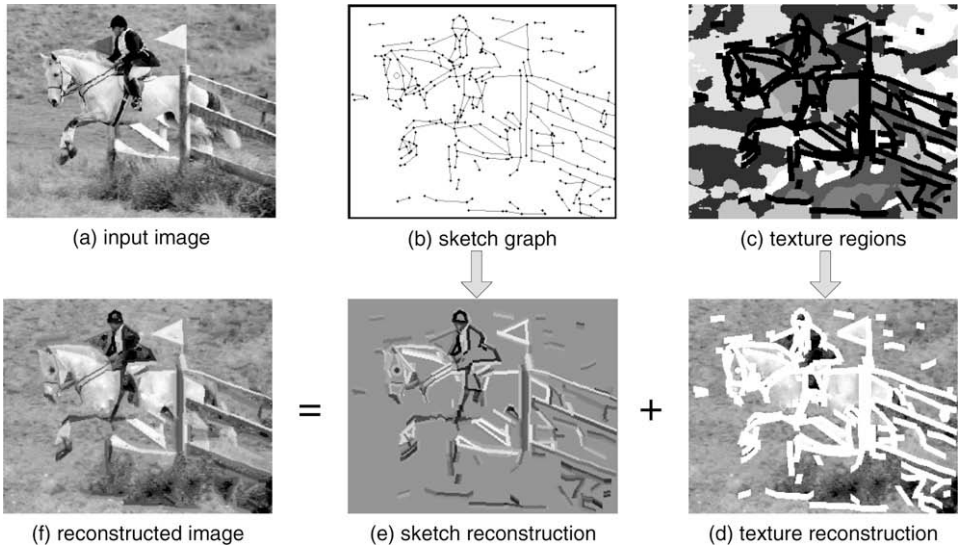


Fig. 20. Example of the primal sketch representation, from Guo et al. (2007).

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

templates for object recognition in high level vision. More discussion about the hybrid template is referred to a recent paper (Si, 2009).

The hybrid template is an extension of the Active Basis model presented in Section 3.3. The latter only uses Gabor basis functions from the explicit manifolds for representing structural elements in the objects. Now we use both primitives from explicit manifolds and texture patches from implicit manifolds. For each object we may have multiple deformable templates to account for different views or configurations. Each template is learned from a set of training images $\{\boldsymbol{I}_m^{\text{obs}} : i = 1, 2, \ldots, M\}$. These images are instances of the object and are well aligned in position, orientation and scale with arbitrary background. Different templates of the object can be learned through an EM-like clustering procedure.

Like primal sketch, the image domain of a hybrid template is divided into a number of K non-overlapping patches $\Lambda = \cup_{i=1}^{K} \Lambda_i$. The image in a patch Λ_i is denoted by $\boldsymbol{I}_{\Lambda_i}$. A patch can be either a primitive from the explicit manifold or a texture from the implicit manifold, just like the hedgehog example in Fig. 1.

Fig. 21 shows our results of the learned hybrid templates for eight object categories. Explicit manifolds and implicit manifolds largely complement each other in explaining the object boundary and interior clutters.

In the following, we briefly introduce the modeling and learning process.

If a patch $\boldsymbol{I}_{\Lambda_i}$ is from an explicit manifold, we define its feature response by,

$$r_i^{\text{ex}}(\boldsymbol{I}) = \rho^{\text{ex}}(\boldsymbol{I}_{\Lambda_i}, B_i), \tag{43}$$

where B_i is a basis function normalized to unit norm, $\rho^{\text{ex}}(\boldsymbol{I}_{\Lambda_i}, B_i) = \|\boldsymbol{I}_{\Lambda_i} - c_i B_i\|^2$, is an Euclidean distance with $c_i = \langle \boldsymbol{I}, B_i \rangle$. Because $\|\boldsymbol{I}_{\Lambda_i} - c_i B_i\|^2 = \|\boldsymbol{I}_{\Lambda_i}\|^2 - |c_i|^2$, we can model $|c_i|^2$ directly, and let $r_i^{\text{ex}}(\boldsymbol{I}) = |c_i|^2$. We allow B_i to slightly perturb its locations and orientations in order to better fit $\boldsymbol{I}$.

If a patch $\boldsymbol{I}_{\Lambda_i}$ is from an implicit manifold, we define the feature response,

$$r_i^{\text{im}}(\boldsymbol{I}) = \rho^{\text{im}}(\mathtt{h}(\boldsymbol{I}_{\Lambda_i}), \mathtt{h}_i), \tag{44}$$

where $\mathtt{h}(\boldsymbol{I}_{\Lambda_i})$ is the histogram of the responses from Gabor filters at different orientations pooled within $\boldsymbol{I}_{\Lambda_i}$, and $\mathtt{h}_i$ is the typical histogram. We can use ℓ_2 distance between $\mathtt{h}(\boldsymbol{I}_{\Lambda_i})$ and $\mathtt{h}_i$ as the feature response $r_i^{im}(\boldsymbol{I})$.

Our objective is to learn a model $p(\boldsymbol{I})$ against a background model $q(\boldsymbol{I})$. We transform the image $\boldsymbol{I}$ to a new set of variables,

$$\boldsymbol{I} \mapsto (R, \overline{R}), \quad \text{with } R = (r_1, \ldots, r_K). \tag{45}$$

R is a vector for the K responses from the explicit or implicit patches, and $\overline{R}$ is the remaining dimensions. This is a non-linear transform that transfers the space into a subspace for R and the remaining subspace for $\overline{R}$. The pursuit process is to seek for the most informative patches in the image so that $p(R)$ is very different from the background model $q(R)$, while in the remaining subspace, $p(\overline{R})$ is of no difference to $q(\overline{R})$, so $p(\overline{R}) = q(\overline{R})$. By canceling the Jacobian term, we have

$$\frac{p(\boldsymbol{I})d\boldsymbol{I}}{q(\boldsymbol{I})d\boldsymbol{I}} = \frac{p(R)dR}{q(R)dR} \cdot \frac{p(\overline{R})d\overline{R}}{q(\overline{R})d\overline{R}} \cdot \frac{\left|\frac{\partial \boldsymbol{I}}{\partial(R,\overline{R})}\right|}{\left|\frac{\partial \boldsymbol{I}}{\partial(R,\overline{R})}\right|} \tag{46}$$

$$= \frac{p(R)dR}{q(R)dR}. \tag{47}$$

As the images are well-aligned, the feature responses $\{r_i\}$ do not overlap and thus are independent of with each other conditional on the overall location and alignment. So we obtain a probability

$$p(\boldsymbol{I}) = q(\boldsymbol{I}) \prod_{i=1}^{K} \frac{p_i(r_i)}{q_i(r_i)}, \tag{48}$$

where $p_i(r_i)$ is a 1-dimensional distribution of r_i pooled over the M training images of the object, and $q_i(r_i)$ is a 1-dimensional distribution pooled over generic images (i.e. daily photos that do not have the object). Thus the task is decomposed as learning a series of 1D models through the information projection process,

$$p_i(r_i) = \frac{1}{z_i} q_i(r_i) e^{-\lambda_i s(r_i)}.$$

The learning process is to select B_i or $\mathtt{h}_i$ sequentially according to their information gains. Intuitively we seek for patches whose responses have large $KL(p_i \| q_i)$.

In experiments, we use the same set of Gabor filters $\{F_k\}$ as in the Active Basis and texture modeling. We perform feature selection for both sketch and texture and rank them by information gains, until a maximum number of features (60) is reached or the information gain is smaller than a threshold, currently set to a

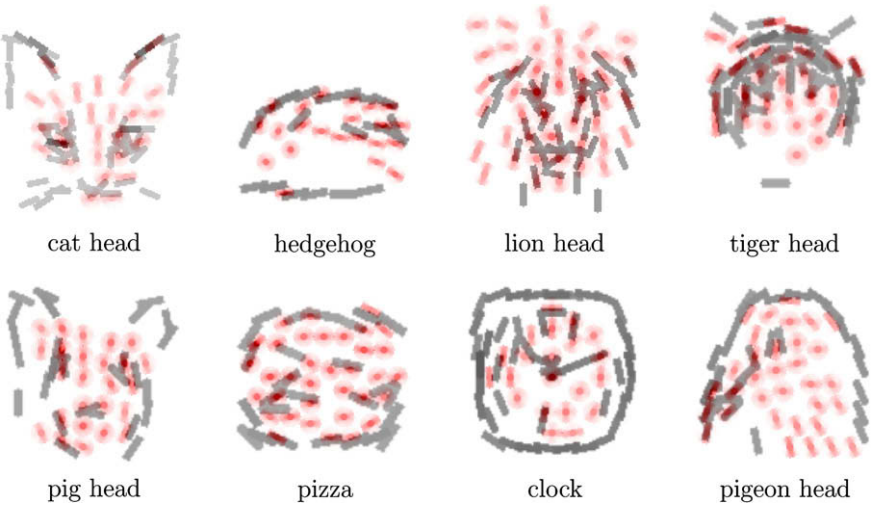


Fig. 21. Learned hybrid templates of eight object categories. Bold block bars denote sketches (explicit manifolds), while the blurred red blobs describe local textures (implicit manifolds). The sketch features capture the global shape, while the texture features capture additional information in the image appearance. Results are from Si (2009).

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

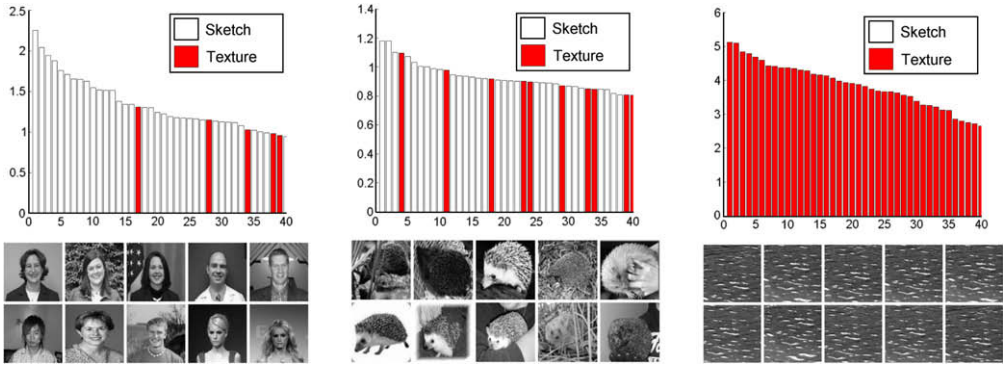


Fig. 22. Competition of sketch (explicit) and texture (implicit) features in learning hybrid templates. Each figure plots the information gains of selected features, ranked in descending order: hollow black/white bars for primitive patches and solid red bars for texture patches. For image categories with clear and regular shape, e.g. human head/shoulder, primitives dominate the information gain. For texture categories with cluttered structures, the texture patches dominate. The hedgehog category is a typical case where the two types of patches alternate. Results are from Si (2009).

heuristic number 0.2, which is universal across different image categories. To ensure the approximate orthogonality of features, the selected primitives patch are enforced to correlate no larger than a threshold 0.1. The texture patches are allowed to overlap 25% so as to pool the feature statistics.

Fig. 22 illustrates the information gain for three categories and each category has $M = 15$ training examples. For image category that has regular shape, like the head and shoulder, the explicit patches dominates. In contrast, the implicit patches dominate the water category, and the hedgehog template is mixed. Eight learned object templates are shown in Fig. 21. More experiments and recognition results are reported in (Si, 2009). For source code and data please refer to the project web page at http://www.stat.ucla.edu/~zzsi/mixed_template.html.

In the following, we briefly compare the hybrid templates to other related templates or object representations in the literature.

1. As a generative representation, deformable shape models and pictorial templates were widely used in the 1970–1980s (Yuille et al., 1992). Appearance is added to shape in the well-known active appearance model in (Cootes et al., 2001). But in such models, the shape (or keypoints) are defined manually and the appearance is modeled by global linear representation, such as PCA. In contrast, the hybrid templates are learned though training images and the shape and texture patches are selected by calculating an information gain in an unsupervised way. Obviously the selection of primitives and textures will change over image scales, as it is discussed in previous sections.

2. As a discriminative representation, many image features are extracted for objects, the most popular one in recognition is the HoG template (Histogram of oriented Gradients) (Dalal and Triggs, 2005), and recently part based HoG models are also studied (Felzenszwalb et al., 2009). The HoG representation divides the image domain into regular $m \times n$ grid with each cell being a small image patch, for instance, 8 pixels. At each pixel, a gradient is calculated, and a histogram is pooled over each cell for different orientations. The histograms from the mn cells are concatenated into a long vector to feed a SVM classifier. In fact, this HoG template bears similarities to the hybrid template here. The differences are (i) the cells or patches are not regularly divided and are allowed to deform in the hybrid templates; (ii) at image primitives, such as edges and bars, the histogram of gradients is dominated by one orientation in the HoG, and thus it is a more expensive representation than the primitives themselves.

In general, the hybrid template is a generative model for object that integrates shape and texture. The implicit and explicit manifolds quantize the space of image patches and provide a very sparse representation.

6. Discussion

Two types of manifolds. The key idea in this paper is to propose a theoretical framework for studying two different types of manifolds in a unified framework. Explicit manifolds are better suited for geometric structures, whereas implicit manifolds are better suited for stochastic textures. They are the atomic structures in the space of image patches and they are composed to form complex representation for larger images.

An entropy spectrum. We map different manifolds in the image space according to their entropy. The explicit manifolds and active basis models are in the low entropy regime, while the implicit manifolds and texture models are in the high entropy regime. The hybrid templates belong to the middle entropy regime where a combinatorial number of objects reside. Image scaling could cause transitions between these manifolds.

Visual vocabulary and AND–OR graph composition. The atomic explicit and implicit manifolds form a leaf level dictionary for the visual vocabulary. They can then be recursively composed into more complicated image categories or visual words, which in turn serve as non-terminal nodes in an hierarchical AND–OR graph representation (Zhu and Mumford, 2006). In this representation, the AND node denotes the composition of its components, while the OR node denotes multiple ways of compositions. We refer to Zhu and Mumford (2006) for a lengthy discussion of the compositional mechanisms.

Acknowledgements

This work at UCLA is supported by NSF Grants IIS-0713652 and DMS-0707055. The work at Lotus Hill Research Institute is supported by 863 Project 2008AA01Z126 and NSFC Grant 60832004. We thank Prof. Yingnian Wu for extensive discussions and insightful suggestions over the entire period of the projects. We thank Liang Lin and Youdong Zhao for their assistance.

References

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Machine Intell. 23 (6), 81–85.
Csiszár, I., Shields, P.C., 2004. Information theory and statistics: A tutorial. Commun. Inf. Theory 1 (4), 417–528.
Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. IEEE Conf. on Computer Vision and Pattern Recognition.
Della Pietra, S., Della Pietra, V., Lafferty, J., 1997. Inducing features of random fields. IEEE Trans. Pattern Anal. Machine Intell. 19 (4), 380–393.
Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D., 2009. Object detection with discriminatively trained part based models. Technical Report, Department of Computer Science, University of Chicago.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020

Field, D.J., 1987. Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Amer. A 4, 2379–2394.
Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Systems Sci. 55 (1), 119–139.
Frey, B., Jojic, N., 1999. Transformed component analysis: Joint estimation of spatial transforms and image components. IEEE Internat. Conf. on Computer Vision.
Geman, D., Koloydenko, A., 1999. Invariant statistics and coding of natural microimages. First Internat. Workshop on Statistical and Computational Theories of Vision, Fort Collins, Co. June, 1999.
Guo, C.E., Zhu, S.C., Wu, Y.N., 2007. Primal sketch: Integrating structure and texture. Computer Vision and Image Understanding, pp. 5–19.
Huang, J., Mumford, D., 1999. Statistics of natural images and models. IEEE Conf. on Computer Vision and Pattern Recognition, 541–547.
Julesz, B., 1995. Dialogues on Perception. MIT Press, Cambridge, MA.
Karni, A., Sagi, D., 1991. Where practice makes perfect in texture discrimination – evidence for primary visual cortex plasticity. Proc. Nat. Acad. Sci. 88, 4966–4970.
Lee, A.B., Pedersen, K.S., Mumford, D., 2003. The nonlinear statistics of high-contrast patches in natural images. Int. J. Comput. Vision 54 (1–3), 83–103.
Leung, T., Malik, J., 1999. Recognizing surface using three-dimensional textons. IEEE Internat. Conf. on Comput. Vision, Corfu, Greece, 1999.
Ma, Y., Derksen, H., Hong, W., Wright, J., 2007. Segmentation of multivariate mixed data via lossy data coding and compression. IEEE Trans. Pattern Anal. Machine Intell. 29 (9), 1546–1562.
Marr, D., 1982. Vision. W.H. Freeman and Company.
Matheron, S.G., 1975. Random Sets and Integral Geometry. John Wiley and Sons.
Mumford, D., Gidas, B., 2001. Stochastic models for generic images. Quart. Appl. Math. 59, 85–111.
Mumford, D.B., Shah, J., 1989. Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. 42 (5), 577–685.
Olshausen, B.A., Field, D.J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (6583), 607–609.
Roth, S., Black, M.J., 2005. Fields of experts: A framework for learning image priors. IEEE Conf. on Computer Vision and Pattern Recognition.
Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.
Ruderman, D.L., 1994. The statistics of natural images. Network: Comput. Neural Systems 5, 517–548.
Shi, K., 2008. Mapping Natural Image Patches by Explicit and Implicit Manifolds. Ph.D. Thesis, Department of Statistics, UCLA.
Si, Z., Gong, H., Wu, Y.N., Zhu, S.C., 2009. Learning mixed image templates for object categories. IEEE Conf. on Computer Vision and Pattern Recognition, Miami, Florida, 2009.
Wang, Y., Zhu, S.C., 2008. Perceptual scale-space and its applications. Internat. J. Comput. Vision 80 (1), 143–165.
Weiss, Y., Freeman, W.T., 2007. What makes a good model of natural images? IEEE Conf. on Computer Vision and Pattern Recognition.
Wu, Y.N., Si, Z., Fleming, C., Zhu, S.C., 2007. Deformable template as active basis. IEEE Internat. on Conf. Computer Vision.
Wu, Y.N., Zhu, S.C., Guo, C.E., 2008. From information scaling of natural images to regimes of statistical models. Quart. Appl. Math. 66, 81–122.
Wu, Y.N., Si, Z., Gong, H., Zhu, S.C., 2009. Learning active basis model for object detection and recognition. Internat. J. Comput. Vision.
Yuille, A.L., Hallinan, P.H., Cohen D., 1992. Feature extraction from faces using deformable templates. Internat. J. Comput. Vision, 8, 99–111.
Zhu, S.C., Mumford, D., 1997. Prior learning and gibbs reaction–diffusion. IEEE Trans. Pattern Anal. Machine Intell. 19 (11), 1236–1250.
Zhu, S.C., Mumford, D., 2006. A stochastic grammar of images. Found. Trends. Comput. Graph. Vision 2 (4), 259–362.
Zhu, S.C., Wu, Y.N., Mumford, D., 1997. Minimax entropy principle and its application to texture modeling. Neural Comput. 9, 1627–1660.
Zhu, S.C., Liu, X.W., Wu, Y.N., 2000. Exploring texture ensembles by efficient markov chain Monte Carlo-toward a 'trichromacy' theory of texture. IEEE Trans. Pattern Anal. Machine Intell. 22 (6), 554–569.
Zhu, S.C., Guo, C.E., Wang, Y., Xu, Z., 2005. What are textons? Internat. J. Comput. Vision 62 (1–2), 121–143.

Please cite this article in press as: Zhu, S.-C., et al. Learning explicit and implicit visual manifolds by information projection. Pattern Recognition Lett. (2009), doi:10.1016/j.patrec.2009.07.020