Joint Object and Pose Recognition using Homeomorphic Manifold Analysis

Haopeng Zhang[1], Tarek El-Gaaly[2], Ahmed Elgammal[2] and Zhiguo Jiang[1]
1 Image Processing Center, School of Astronautics, Beihang University
Beijing Key Laboratory of Digital Media
Beijing, 100191, China
2 Department of Computer Science, Rutgers University
Piscataway, NJ 08854, USA

Abstract

Object recognition is a key precursory challenge in the fields of object manipulation and robotic/AI visual reasoning in general. Recognizing object categories, particular instances of objects and viewpoints/poses of objects are three critical subproblems robots must solve in order to accurately grasp/manipulate objects and reason about their environments. Multi-view images of the same object lie on intrinsic low-dimensional manifolds in descriptor spaces (*e.g.* visual/depth descriptor spaces). These object manifolds share the same topology despite being geometrically different. Each object manifold can be represented as a deformed version of a unified manifold. The object manifolds can thus be parametrized by its homeomorphic mapping/reconstruction from the unified manifold. In this work, we construct a manifold descriptor from this mapping between homeomorphic manifolds and use it to jointly solve the three challenging recognition sub-problems. We extensively experiment on a challenging multi-modal (*i.e.* RGBD) dataset and other object pose datasets and achieve state-of-the-art results.

1 Introduction

Visual object recognition is a challenging problem with many real-life AI applications. The difficulty of the problem is due to variations in appearance between objects within the same category, and between varying poses of the same object. Under this perceptual problem of visual recognition lie three subproblems that are each quite challenging. The first is *category recognition*, which identifies the category of a particular object (*e.g.* is this a mug or bottle?). The second is *instance recognition* (*e.g.* is this John's pen or Jackie's pen?). The third subproblem is *pose recognition* which identifies the viewpoint/pose of an object (*e.g.* where is the handle of the mug?).

Traditional 3D pose estimation algorithms often solve the recognition and pose estimation problems simultaneously using 3D object model-bases, hypothesis and test principles, or invariants, *e.g.* geometric hashing (Lamdan and Wolfson 1988). Such models are incapable of dealing with large within-class variability and have been mainly focused on recognizing instances. This limitation led to the development, over the last decade, of very successful categorization methods mainly based on local features and parts. Such methods loosely encode the geometry, *e.g.* methods like pictorial structure (Felzenszwalb and Huttenlocher 2005); or does not encode the geometry at all, *e.g.* bag of words (Willamowski et al. 2004; Sivic et al. 2005).

Most research on generic object recognition from local features has focused on recognizing objects from a single viewpoint or from limited viewpoints, *e.g.* front, side and rear views of cars, *etc*. Recently, there has been an increasing interest in object categorization from multiple views, as well as recovering object pose in 3D, e.g. (Thomas et al. 2006; Savarese and Fei-Fei 2007; Sun et al. 2009; Ozuysal, Lepetit, and Fua 2009; Liebelt and Schmid 2010; Payet and Todorovic 2011). Almost all the work on pose estimation and multi-view recognition from local features is based on formulating the problem as a classification problem where view-based classifiers and/or viewpoint classifiers are trained. Very few works formulate the problem of pose estimation as a regression problem and the works that do, learn the regression function within each category, e.g. (Torki and Elgammal 2011). In the domain of multi-modal data, recent work by (Lai et al. 2011b) uses synchronized multi-modal photometric and depth information (*i.e.* RGBD) to achieve significant performance in object recognition. They build an object-pose tree model from RGBD images and perform hierarchical inference. Although performance of category and instance recognition is significant, object pose recognition performance is less so. The reason is the same; a classification strategy for pose recognition results in coarse pose estimates and does not fully utilize the information present in the continuous distribution of descriptor spaces.

The contribution of this paper is in the way we formulate the problem of view-invariant recognition and pose estimation. We pose the problem as a *style* and *content* separation problem in an unconventional way. The intuitive way is to model the category as the *content* and the viewpoint as a *style* variable. Instead we model the viewpoint as the content and the category as a style variable. This unintuitive way is justified from the point of view of learning the visual manifold of the data. The manifold of different views is intrinsically low-dimensional with a known topology. We can show that view manifolds of all objects are deformed versions of each other. In contrast, the manifold of all object categories can be of infinite dimensions and hard to model given within-class object variability and the enormous num-

ber of categories. Therefore, we propose to model the category as a style variable over the view manifold of objects.

(Tenenbaum and Freeman 2000) formulated the separation of style and content based on a bilinear model. They presented a computational framework for model fitting using SVD. A more general multilinear model was used by (Vasilescu and Terzopoulos 2002) to decompose multiple orthogonal factors. However, in bilinear and multilinear models, the content and style factors are discrete classes. Separating style where content is a continuous manifold was introduced in (Elgammal and Lee 2004), in the context of human motion analysis, where the separation was done in a manifold parameterization space. In this paper, we adapt a similar approach to the problem of object recognition, where we model the viewpoint as a continuous content manifold and separate object style variables as view-invariant descriptors for recognition. This results in a generative model of object appearance as a function of multiple latent variables, one describing the viewpoint and lies on a low-dimensional manifold, and the other describing the category/instance and lies on a low-dimensional subspace.

In this paper we focused on the case of a camera looking at an object on a turntable setting, which results in a one-dimensional view manifold, *i.e.*, one degree of freedom (1DOF) and generalization to a viewing sphere centered around the object (2DOF). Generalization to recover the full six degrees of freedom (6DOF) of a camera is not obvious. Recovering the full 6DOF camera pose is possible for a given object instance, which can be achieved by traditional model-based method. However, this is a quite challenging task for the case of generic object categories. There are various reasons why to consider only the case of 1DOF and 2DOF and not the 6DOF. First, It quite hard to have training data that covers the space of poses in that case; all the state-of-the-art dataset are limited to few views of at most a turntable with couple of different altitudes. Second, practically, we do not see objects in all possible poses, in many applications the poses are quite limited to a viewing circle or sphere. Even humans will have problems recognizing objects in unfamiliar poses. Third, for most applications, it is not required to know the 6DOF pose, 1DOF pose might be enough. Definitely for categorization 6DOF is not needed. In this paper we show that we can learn on a viewing circle and generalize very well to a large range of views around it.

The organization of the paper is as follows. Section 2 summarizes the homeomorphic framework and its application to object recognition. Section 3 describes learning the model. Section 4 describes using this model to solve for category, instance and pose. Experimental results are shown in Section 5 to validate the novelty of our approach.

2 Factorized Model for Object Recognition

The objective of our framework is to learn a manifold representation for multi-view objects that supports category, instance and viewpoint recognition. In order to achieve this, given a set of images captured from different viewpoints, we aim to learn a generative model that explicitly factorizes the following:

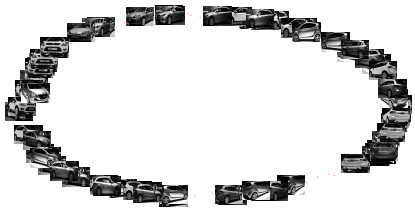

Figure 1: Embedding of view manifold from (Torki and Elgammal 2011). The view manifold converges to a circle in 2D embedding space

1. Viewpoint variable (within-manifold parameterization): smooth parameterization of the viewpoint variations, invariant to the object's category.

2. Object variable (across-manifold parameterization): parameterization at the level of each manifold that characterizes the object's instance/category, invariant to the viewpoint.

The underlying principle, is that multiple views of an object lie on intrinsic low-dimensional manifolds (view manifold) in the descriptor space (input space). The view manifolds of different objects are distributed in that descriptor space. To recover the category, instance and pose of a test image we need to know which manifold this image belongs to and the intrinsic coordinates of that image within the manifold. This basic view of object recognition and pose estimation is not new, and was used in the seminal work of (Murase and Nayar 1995). PCA was used to achieve linear dimensionality reduction of the visual data, and the manifolds of different object were represented as parameterized curves in the embedding space.

What is novel in our framework, is that we use the view manifold deformation as an invariant that can be used for categorization and modeling the within-class variations. Let us consider the case where different views are obtained from a viewing circle, *e.g.* camera viewing an object on a turntable. The view manifold of the object is a 1D closed manifold embedded in the input space (denoted as visual manifold). How that simple closed curve deforms in the input space is a function of the object geometry and appearance. The visual manifold can be degenerate, *e.g.* imaging a textureless sphere from different views results in the same image, *i.e.* the visual manifold in this case is degenerate to a single point. Therefore, capturing and parameterizing the deformation of a given object's view manifold tells us information about the object category and within category variation. If the views are obtained from a full or part of the view-sphere centered around the object, it is clear that the resulting visual manifold should be a deformed sphere as well (assuming the cameras are facing toward the object).

Let us denote the view manifold of object instance s in the input space by $\mathcal{D}^s \subset \mathbb{R}^D$. D is the dimensionality of the input space. Assuming that all manifolds $\mathcal{D}^s$ are not degenerate (we will discuss this issue shortly), then they are all

topologically equivalent, and homeomorphic to each other[1]. Moreover, suppose we can achieve a common view manifold representation across all objects, denoted by $\mathcal{M} \subset \mathbb{R}^e$, in an Euclidean embedding space of dimensionality e. All manifolds $\mathcal{D}^s$ are also homeomorphic to $\mathcal{M}$. In fact all these manifolds are homeomorphic to a unit circle in 2D for the case of a viewing circle, and a unit-sphere in 3D for the case of full view sphere.

We can achieve a parameterization of each manifold deformation by learning object-dependent regularized mapping functions $\gamma_s(\cdot) : \mathbb{R}^e \rightarrow \mathbb{R}^D$ that map from $\mathcal{M}$ to each $\mathcal{D}^s$. Given a Reproducing Kernel Hilbert Space (RKHS) of functions and its corresponding kernel $K(\cdot,\cdot)$, from the representer theorem (Kimeldorf and Wahba 1970; Poggio and Girosi 1990) it follows that such functions admit a representation of the form

$$\gamma_s(\boldsymbol{x}) = \boldsymbol{C}^s \cdot \psi(\boldsymbol{x}) \ , \tag{1}$$

where $\boldsymbol{C}^s$ is a $D \times N_\psi$ mapping coefficient matrix, and $\psi(\cdot) : \mathbb{R}^e \rightarrow \mathbb{R}^{N_\psi}$ is a nonlinear kernel map, as will be described in Sec 3.

In the mapping (Eq. 1), the geometric deformation of manifold $\mathcal{D}^s$, from the common manifold $\mathcal{M}$, is encoded in the coefficient matrix $\boldsymbol{C}^s$. Therefore, the space of matrices $\{\boldsymbol{C}^s\}$ encodes the variability between different object manifolds, and can be used to parameterize such manifolds. We can parameterize the variability across different manifolds in a subspace in the space of coefficient matrices. The general form of our generative model is

$$\gamma(\mathbf{x}, \mathbf{s}) = \mathcal{A} \times_2 \mathbf{s} \times_3 \psi(\mathbf{x}). \tag{2}$$

In this model $\mathbf{s} \in \mathbb{R}^{d_s}$ is a parameterization of manifold $\mathcal{D}^s$ that signifies the variation in category/instance of an object. $\mathbf{x}$ is a representation of the viewpoint that evolves around the common manifold $\mathcal{M}$. $\mathcal{A}$ is a third order tensor of dimensionality $d \times d_s \times N_\psi$, where $\times_i$ is the mode-i tensor product as defined in (Lathauwer, de Moor, and Vandewalle 2000). In this model, both the viewpoint and object latent representations, $\mathbf{x}$ and $\mathbf{s}$, are continuous. Given a test image y recovering the category, instance and pose reduces to an inference problem where the goal is to find $\mathbf{s}^*$ and $\mathbf{x}^*$ that minimizes a reconstruction error, *i.e.*,

$$\arg\min_{\mathbf{s},\mathbf{x}} \|y - \mathcal{A} \times_2 \mathbf{s} \times_3 \psi(\mathbf{x})\|^2. \tag{3}$$

Once $\mathbf{s}$ is recovered, an instance classifier and a category classifier can be used to classify y.

Learning the model is explained in Section 3. Here we discuss and justify our choice of the common manifold embedded representation. Since we are dealing with 1D closed view manifolds, an intuitive common representation for these manifolds is a unit circle in $\mathbb{R}^2$. A unit circle has the same topology as all object view manifolds (assuming no degenerate manifolds), and hence, we can establish a homeomorphism between it and each manifold.

Dimensionality reductions (DR) approaches, whether linear (such as PCA (Jolliffe 1986) and PPCA (Tipping and Bishop 1999)) or nonlinear (such as isometric feature mapping (Isomap) (Tenenbaum 1998), Locally linear embedding (LLE) (Seung and Lee 2000), Gaussian Process Latent Variable Models GPLVM (Lawrence 2004)) have been widely used for embedding manifolds in low-dimensional Euclidean spaces. DR approaches find an optimal embedding (latent space representation) of a manifold by minimizing an objective fn. that preserves local (or global) manifold geometry. Such low-dimensional latent space is typically used for inferring object pose or body configuration. However, since each object has its own view manifold, it is expected that the embedding will be different for each object. On the other hand, using DR to embed data from multiple manifolds together will result in an embedding dominated by the inter-manifold distance and the resulting representation cannot be used as a common representation.

Embedding multiple manifolds using DR can be achieved using manifold alignment, *e.g.* (Ham, Lee, and Saul 2005). If we embed aligned view manifolds for multiple objects where the views are captured from a viewing circle, we observe that the resulting embedding will converge to a circle. Similar results were shown in ((Torki and Elgammal 2011)), where a view manifold is learned from local features from multiple instances with no prior alignment (shown in Fig 1). This is expected since each object view manifold is a 1D closed curve in the input space, *i.e.* a deformed circle. Such deformation depends on object geometry and appearance. Hence it is expected that the latent representation of multiple aligned manifolds will converge to a circle. This observation empirically justifies the use of a unit circle as a general model of object view manifold in our case. Unlike DR where the goal is to find an optimal embedding that preserves the manifold geometry, in our case we only need to preserve the topology while the geometry is represented in the mapping space. This facilitates parameterizing the space of manifolds. Therefore, the unit circle represents an *ideal* conceptual manifold representation, where each object manifold is a deformation of that ideal case. In some sense we can think of a unit circle as a prior model for all 1D view manifolds. If another degree of freedom is introduced which, for example, varies the pitch angle of the object on the turntable then a sphere manifold would capture the conceptual geometry of the pose and be topologically-equivalent.

There are several reasons why we learn the mapping in a generative manner from the unit circle to each object manifold (not the other way). First, this direction guarantees that the mapping is a function even in the case of degenerate manifolds (or self intersections) in the input space. Second, mapping from the unit circle results in a common RKHS of functions. All the mappings will be linear combinations of the same finite set of basis functions in $\mathbb{R}^e$. This facilitates factorizing the manifold geometry variations in the space of coefficients in Eq 2.

[1] A function $f : X \rightarrow Y$ between 2 topological spaces is called a homeomorphism if it is a bijection, continuous, and its inverse is continuous. In our case the existence of the inverse is assumed but not required for computation, *i.e.*, we do not need the inverse for recovering pose. We mainly care about the mapping in a generative manner from $\mathcal{M}$ to $\mathcal{D}^s$.

3 Learning the Model

Conceptual Manifold Embedding

Let the sets of input image sequences be $Y^k = \{\mathbf{y}_i^k \in \mathbb{R}^d, i = 1, \cdots, N_k\}$ and their corresponding points on the conceptual unified embedding space be $X^k = \{\mathbf{x}_i^k \in \mathbb{R}^e, i = 1, \cdots, N_k\}$. d is the dimensionality of the input space (*i.e.* descriptor space) and e is the dimensionality of the conceptual embedding. The image sequences do not necessarily have to be same length. For clarity and without loss of generality, we assume the input is captured from viewpoints: $\Theta = \{\theta_i^k \in [0, 2\pi), i = 1, \cdots, N_k\}$ on a viewing circle. The k-th image sequence is embedded on a unit circle such that $\mathbf{x}_i^k = [\cos\theta_i^k, \sin\theta_i^k] \in \mathbb{R}^2, i = 1, \cdots, N_k$. Notice that by embedding on a unit circle the metric input space is not preserved, but the topology of the manifold is.

Homeomorphic Manifold Mapping

Given an input sequence Y^k and its embedding coordinates X^k on a unit circle, we learn a regularized nonlinear mapping function from the embedding to the input space, *i.e.* a function $\gamma_k(\cdot) : \mathbb{R}^e \rightarrow \mathbb{R}^d$ that maps from embedding space, with dimensionality e, into the input space with dimensionality d and satisfies $\mathbf{y}_i^k = \gamma_k(\mathbf{x}_i^k), i = 1, \cdots, N_k$. From the representer theorem (Kimeldorf and Wahba 1970) we know that a nonlinear mapping function that minimizes a regularized risk criteria admits a representation in the form of linear combination of basis functions around arbitrary points $\mathbf{z}_j \in \mathbb{R}^e, j = 1, \cdots, M$ on the manifold (unit circle). In particular we use a semi-parametric form for the function $\gamma(\cdot)$. Therefore, for the l-th dimension of the input, the function γ_k^l is an RBF interpolant from $\mathbb{R}^e$ to $\mathbb{R}$. This takes the form

$$\gamma_k^l(\mathbf{x}) = p^l(\mathbf{x}) + \sum_{j=1}^{M} \omega_j^l \cdot \phi(|\mathbf{x} - \mathbf{z}_j|), \tag{4}$$

where $\phi(\cdot)$ is a real-valued basis function, ω_j are real coefficients and $|\cdot|$ is the 2^{nd} norm in the embedding space. p^l is a linear polynomial with coefficients c^l, *i.e.* $p^l(\mathbf{x}) = [1 \quad \mathbf{x}] \cdot c^l$. The polynomial part is needed for positive semi-definite kernels to span the null space in the corresponding RKHS. The polynomial part is essential for regularization with the choice of specific basis functions such as Thin-plate spline kernel (Kimeldorf and Wahba 1971). The choice of the centers is arbitrary (not necessarily data points). Therefore, this is a form of Generalized Radial Basis Function (GRBF) (Poggio and Girosi 1990). Typical choices for the basis function include thin-plate spline, multiquadric, Gaussian[2], biharmonic and tri-harmonic splines. The whole mapping can be written in a matrix form

$$\gamma^k(\mathbf{x}) = \mathbf{C}^k \cdot \psi(\mathbf{x}), \tag{5}$$

where $\mathbf{C}^k$ is a $d \times (M + e + 1)$ dimensional matrix with the l-th row $[\omega_1^l, \cdots, \omega_M^l, {c^l}^T]$. The vector $\psi(x) = [\phi(|x - z_1|) \cdots \phi(|x - z_M|), 1, x^T]^T$ represents a nonlinear kernel map from the embedded conceptual representation to a kernel induced space. To ensure orthogonality and to make the problem well posed, the following condition constraints are imposed: $\Sigma_{i=1}^M \omega_i p_j(x_i) = 0, j = 1, \cdots, m$, where p_j are the linear basis of p. Therefore, the solution for $\mathbf{C}^k$ can be obtained by directly solving the linear system:

$$\begin{pmatrix} \mathbf{A} + \lambda\mathbf{I} & \mathbf{P}_x \\ \mathbf{P}_t^T & \mathbf{0}_{(e+1)\times(e+1)} \end{pmatrix}_k {\mathbf{C}^k}^T = \begin{pmatrix} \mathbf{Y}_k \\ \mathbf{0}_{(e+1)\times d} \end{pmatrix}, \tag{6}$$

$\mathbf{A}$, $\mathbf{P}_x$ and $\mathbf{P}_t$ are defined for the $k-th$ set of object images as: $\mathbf{A}$ is a $N_k \times M$ matrix with $\mathbf{A}_{ij} = \phi(|x_i^k - z_j|), i = 1, \cdots, N_k, j = 1, \cdots, M$, $\mathbf{P}_x$ is a $N_k \times (e+1)$ matrix with i-th row $[1, {\mathbf{x}_i^k}^T]$, $\mathbf{P}_t$ is $M \times (e+1)$ matrix with i-th row $[1, \mathbf{z}_i^T]$. Y_k is a $N_k \times d$ matrix containing the input images for set of images k, *i.e.* $\mathbf{Y}_k = [\mathbf{y}_1^k, \cdots, \mathbf{y}_{N_k}^k]$. Solution for $\mathbf{C}^k$ is guaranteed under certain conditions on the basic functions used.

Decomposition

Each coefficient matrix $\mathbf{C}^k$ captures the deformation of the view manifold for object instance k. Given learned coefficients matrices $\mathbf{C}^1, \cdots, \mathbf{C}^K$ for each object instance, the category parameters can be factorized by finding a low-dimensional subspace that approximates the space of coefficient matrices. We call the category parameters/factors *style* factors as they represent the parametric description of each object view manifold.

Let the coefficients be arranged as a $d \times (M + e + 1) \times K$ tensor $\mathcal{C}$. The form of the decomposition we are looking for is:

$$\mathcal{C} = \mathcal{A} \times_3 \mathbf{S} \tag{7}$$

A is a $d \times (M + e + 1) \times J$ tensor containing category bases for the RBF coefficient space and $\mathbf{S} = [\mathbf{s}^1, \cdots, \mathbf{s}^K]$ is $J \times K$. Its columns contains the instance/category parameterization. This decomposition can be achieved by arranging the mapping coefficients as a $(d(M + e + 1)) \times K$ matrix:

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^1 & \cdots & \mathbf{c}_1^K \\ \vdots & \ddots & \vdots \\ \mathbf{c}_{M+e+1}^1 & \cdots & \mathbf{c}_{M+e+1}^K, \end{pmatrix} \tag{8}$$

$[\mathbf{c}_1^k, \cdots, \mathbf{c}_{M+e+1}^k]$ are the columns of $\mathbf{C}^k$. Given $\mathbf{C}$, category vectors and content bases can be obtained by SVD as $\mathbf{C} = \mathbf{USV}^T$. The viewpoint bases are the columns of $\mathbf{US}$ and the object instance/category vectors are the rows of $\mathbf{V}$.

4 Inference of Category, Instance and Pose

Given a test image $\mathbf{y} \in \mathbb{R}^d$ represented in a descriptor space, we need to solve for both the viewpoint parameterization $\mathbf{x}^*$ and the object instance parameterization $\mathbf{s}^*$ that minimize Eq. 3. This is an inference problem and various inference algorithms can be used. Notice that, if the instance parameters $\mathbf{s}$ is known, Eq. 3 reduces to a nonlinear 1D search for viewpoint $\mathbf{x}$ on the unit circle that minimizes the error. This can be regarded as a solution for viewpoint estimation, if the object is known. On the other hand, if $\mathbf{x}$ is known, we can obtain a least-square closed-form approximate solution for $\mathbf{s}^*$. An EM-like iterative procedure was proposed in (Elgammal and Lee 2004) for alternating between the two factors. If dense multiple views along a view circle of an object

[2] A Gaussian kernel does not need a polynomial part

are available, we can solve for $\mathbf{C}$ in Eq. 8 and then obtain a closed-form least-square solution for the instance parameter $\mathbf{s}^*$ as

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} ||\mathbf{C} - \mathcal{A} \times s||$$

In the case where we need to solve for both $\mathbf{x}$ and $\mathbf{s}$, given a test image, we use a particle filter (Arulampalam et al. 2002) to solve the inference problem (with K category samples $\mathbf{s}^1, \mathbf{s}^2, \cdots, \mathbf{s}^K$ in the category factor space and L viewpoint samples $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^L$ on the unit circle). To evaluate the performance of each particle we define the likelihood of a particle $(\mathbf{s}^k, \mathbf{x}^l)$ as follows:

$$w_{kl} = exp\frac{-||\mathbf{y} - \mathcal{A} \times \mathbf{s}^k \times \psi(\mathbf{x}^l)||^2}{2\sigma^2} \quad (9)$$

We marginalize the likelihood to obtain the weights of $\mathbf{s}^k$ and $\mathbf{x}^l$ as

$$W_{s_k} = \frac{\sum_{l=1}^{L} w_{kl}}{\sum_{k=1}^{K}\sum_{l=1}^{L} w_{kl}}, W_{x_l} = \frac{\sum_{k=1}^{K} w_{kl}}{\sum_{k=1}^{K}\sum_{l=1}^{L} w_{kl}}. \quad (10)$$

We resample style and viewpoint particles according to W_s and W_x from Normal distributions in order to reduce the reconstruction error. In the case of classification and instance recognition, once the parameters $\mathbf{s}$ are known, k-nearest neighbor classifier is used to find the closest matching category or instance.

Multimodal Fusion

For each individual channel (e.g. RGB and depth), a homeomorphic manifold generative model is built. Our model can be extended to include multiple modalities of information as long as there is smooth variation along the manifold as the viewpoint/pose changes. We combine visual information (*i.e. RGB*) and depth information by using a combined objective function that encompasses the reconstruction error in each mapping. This is done by running the training separately on each channel and combining the objective functions. The combined reconstruction error becomes:

$$\begin{aligned} E_{rgbd}(x, s) = \lambda_{rgb}||y_{rgb} - \mathcal{A}_{rgb} \times s_{rgb} \times \psi(\mathbf{x})||^2 \\ +\lambda_d||y_d - \mathcal{A}_d \times s_d \times \psi(\mathbf{x})||^2 \end{aligned} \quad (11)$$

Notice that the two terms share the same viewpoint variable $\mathbf{x}$. λ_{rgb} and λ_d were selected empirically. Since visual data has less noise than depth (which commonly exhibits missing depth values, *i.e.* holes), we bias the visual reconstruction error term of Eq. 11.

5 Experiments and Results

To validate our approach we experimented on 3 datasets: 3DObjects (Savarese and Fei-Fei 2007), Multi-View Car Dataset (Ozuysal, Lepetit, and Fua 2009), and RGB-D dataset (Lai et al. 2011a) .

Dense Viewpoint Estimation: Multi-View Car Dataset is a challenging dataset that captures 20 rotating cars in an auto show. It provides finely discretized viewpoint ground truth. For quantitive evaluation of our framework for pose estimation, we use the Mean Absolute Error (MAE) between estimated and ground truth viewpoints. To compare with classification-based viewpoint estimation approaches (which use discrete bins) we also compute the percentage of test samples that satisfy $AE < 22.5^\circ$ ($AE = |EstimatedAngle - GroundTruth|$) to achieve an equivalent of a 16 bin viewpoint classifier. We also compute the percentage of test samples that satisfy $AE < 45^\circ$ to achieve an equivalent of an 8 bin viewpoint classifier. We used 35 RBF centers along a 2D unit circle to define the kernel map $\psi(\cdot)$ in Eq 1. We represented the input using HOG (Dalal and Triggs 2005) features. Table 2 shows the view estimation results in comparison to the state of the art and hence clearly shows the significant improvement we achieve.

Table 1: Category recognition performance % based on mean absolute error in pose angles on 3DObject dataset and comparison with state-of-the-art

Class	**Ours**	(Savarese and Fei-Fei 2007)
Bicycle	**99.79**	81.00
Car	**99.03**	70.00
Cellphone	66.74	76.00
Iron	75.78	77.00
Mouse	48.60	87.00
Shoe	**81.70**	62.00
Stapler	**82.66**	77.00
Toaster	**86.24**	75.00

Table 2: Results on Multi-View Cars in Mean Abs. Err. (MAE) and comparison with state-of-the-art

Method	MAE	% of AE $< 22.5^\circ$	% of AE $< 45^\circ$
(Ozuysal, Lepetit, and Fua 2009)	46.48	41.69	71.20
(Torki and Elgammal 2011) - leave-one-out	35.87	63.73	76.84
(Torki and Elgammal 2011) - 50% split	33.98	70.31	80.75
Ours - leave-one-out	**19.34**	**90.34**	**90.69**
Ours - 50% split	**24.00**	**87.77**	**88.48**

Sparse pose estimation: We used the car subset of the 3D-Objects dataset (typically used for pose estimation) to test our approach for viewpoint estimation using sparse training samples on the view circle. This dataset contains only 8 sparse views. We used HOG features as input. We follow the same setup as (Savarese and Fei-Fei 2007; Sun et al. 2009): 5 training sequences and 5 sequences for testing (160 training and 160 testing images). For fair comparison, we report our results in terms of $AE < 45^\circ$, equivalent to a 8 bin classifier. The reported accuracy is 52.5% in (Savarese and Fei-Fei 2007), 66.625% in (Sun et al. 2009) and 85.38% in (Payet and Todorovic 2011). The best accuracy is reported by (Torki and Elgammal 2011) as 77.5% for $AE < 45^\circ$. Using our homeomorphic manifold analysis framework, we achieve 93.13% for $AE < 45^\circ$. This shows the ability of our framework to model the visual manifold, even with sparse views.

Categorization and pose estimation: We used the entire 3DObjects dataset to evaluate the performance of our framework on both object categorization and viewpoint estimation. 3DObjects contains 10 very different everyday objects (shown in Table 1). Similar to (Savarese and Fei-Fei 2007; Sun et al. 2009), we test our model on a 8-category classi-

Table 3: Summary of Results on RGBD dataset using RGB/D and RGB+D

RGB[+D] Methods	Category	Instance	Avg Pose	Med Pose	Avg Pose (C)	Med Pose (C)	Avg Pose (I)	Med Pose (I)
PF (RGB)	92.00	74.36	61.59	89.46	80.36	93.50	82.83	93.90
PF (Depth - DHOG)	74.49	36.18	26.06	0.00	66.36	86.60	72.04	90.03
PF (Depth - VFH)	27.88	13.36	7.99	0.00	57.79	62.75	59.82	67.46
PF (RGB+D)	93.10	74.79	**61.57**	**89.29**	**80.01**	**93.42**	**82.32**	**93.80**
Baseline (RGB+D) (Lai et al. 2011b)	94.30	78.40	53.30	65.20	56.80	71.40	68.30	83.20
Baseline (RGB+D) (El-Gaaly et al. 2012)	-	-	-	-	74.76	86.70	-	-

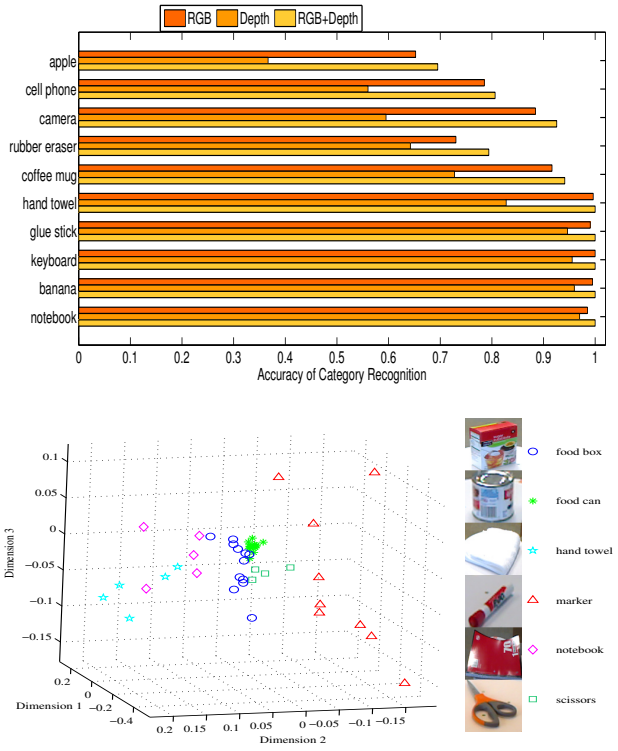

Figure 2: Top: Category recognition using different modes for a subset of categories. Bottom: Sampled instances from 6 different categories in RGB-D dataset. Notice: flatter objects lie to the left and more rounded shapes to the right

fication task (excluding heads & monitors), and the farthest scale is not considered. To learn each category, we randomly select 7/10 object instances for learning and the remaining 3 instances for testing. Average recognition results for cross-validation performed 45 times are shown in Table 1. We achieve an avg. recognition accuracy of 80.07% on 8 classes and an avg. viewpoint estimation performance of 73.13% on the entire test set which satisfies $AE < 45°$.

RGB-D Object Dataset

The largest and most challenging multi-modal multiview dataset available is the RGB-D dataset (Lai et al. 2011a). It consists of 300 instances of 51 tabletop object categories. Each object is rotated on a turn-table and captured using a Kinect sensor (Kinect 2010), providing synchronized visual and depth images. For each object the camera is positioned at 3 height angles (*i.e.* elevation angle): 30°,45°,60°. Training is done using 30°and 60°sequences and testing is done using 45°sequences.

We use HOG features for both RGB channels and depth channel. We also experimented with an additional more recent depth descriptor called Viewpoint Feature Histogram (VFH) (Rusu et al. 2010), computed on point cloud data.

Table 3 summarizes the results of our approach, and compares to 2 state-of-the-art baselines. For training/ testing, we follow the exact same procedures as (Lai et al. 2011b). We used 30 uniformly sampled viewpoints to learn our model. In the case of category and instance recognition (column 2 & 3), we achieve similar results to state-of-the-art (Lai et al. 2011b). We find that $\approx 57\%$ of the categories exhibit better category recognition performance when using RGB+D, as opposed to using RGB only (set of these categories shown in Fig. 2-top). Fig. 2-bottom shows a very nice illustration of sample instances in the object style latent space.

Following from (Lai et al. 2011b); the entire test set was used. The results are shown in Table 3. Incorrectly classified objects were assigned pose accuracies of 0. Avg. and Med. Pose (C) are computed only on test images whose categories were correctly classified. Avg. and Med. Pose (I) were computed only using test images that had their instance correctly recognized. All the object pose estimations significantly out-performs the state-of-the-art (Lai et al. 2011b; El-Gaaly et al. 2012). This verifies that the modeling of the underlying continuous pose distribution is very important in pose recognition.

Lime and bowl categories were found to have better category recognition accuracy using depth alone than using either visual-only or visual and depth together. This can be explained by the complete lack of visual features on their surfaces. Some object instances were classified with higher accuracy using depth only also. There were 19 (out of 300) of these instances, including: lime, bowl, potato, apple and orange. These instances have textureless surfaces with no distinguishing visual features and so the depth-only approach was able to utilize shape information to achieve higher accuracy.

In Table 3 we see that depth HOG (DHOG) performs quite well in all the pose estimation experiments except for where misclassified categories or instances were assigned 0 (column 3 & 4). DHOG appears to be a simple and effective descriptor to describe noisy depth images captured by the Kinect in the dataset. It achieves better accuracy than (Lai et al. 2011b) in the pose estimation. Similar to (Lai et al. 2011b), recursive median filters were applied to depth images to fill depth holes. This validates the modeling of the

underlying continuous distribution which our homeomorphic manifold mapping takes advantage of. VFH is a feature adapted specifically to the task of viewpoint estimation from point cloud data. No prior point cloud smoothing was done to filter out depth holes and so its performance suffered.

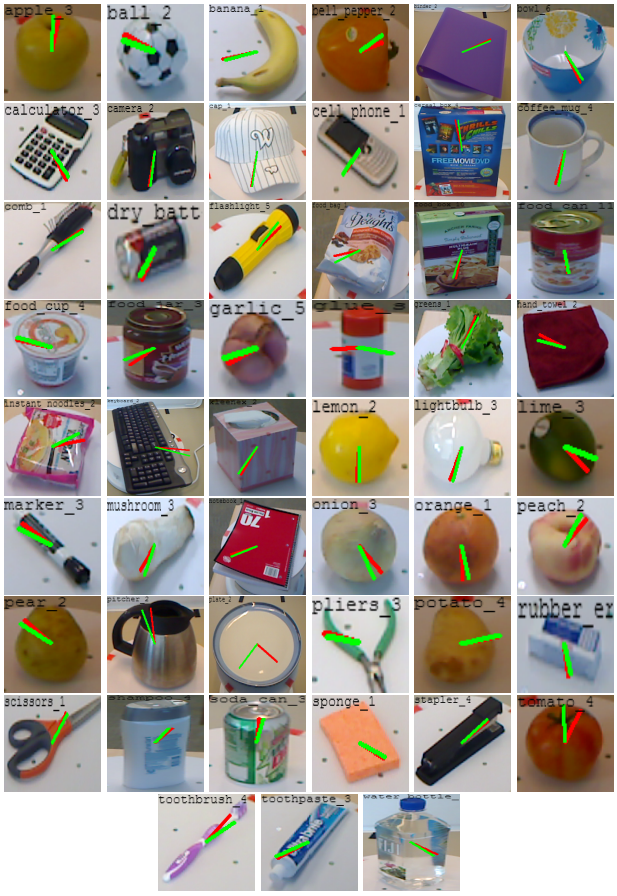

Figure 3: Sample correct results for object and pose recognition on RGB-D dataset. Black text: category name and instance number. Red line: estimated pose. Green line: ground truth pose.

Table-top Object Category Recognition System

Using our approach we built a near real-time system for category recognition of table-top objects. Our system was trained and tested on a subset of 10 different categories from the RGB-D dataset. The category recognition runtime per object in one frame is <2 seconds. Our MATLAB implementation was not optimized for real-time processing but despite this, the potential for real-time capability can be seen. The system was tested on videos provided in the RGB-D dataset that contain cluttered scenes with occlusion, much wider variation of viewpoints and varying scales. Our system achieved $>62\%$ category recognition accuracy. An interesting observation was that depth-only recognition outperformed visual-only recognition in cluttered scenes; intuitively due to the fact that background texture around objects introduces visual noise. On the other hand in the depth mode, large depth discontinuities separate objects from background clutter. We also tested our system on never-seen-before objects. Depth segmentation is performed on point clouds sensed by a Kinect sensor in real-time using (Rusu and Cousins 2011). Our framework is then able to perform category recognition on the detected objects. A video demo showing our system running on never-seen-before objects and objects from the videos provided in the RGB-D dataset is shown in the accompanying supplementary video.

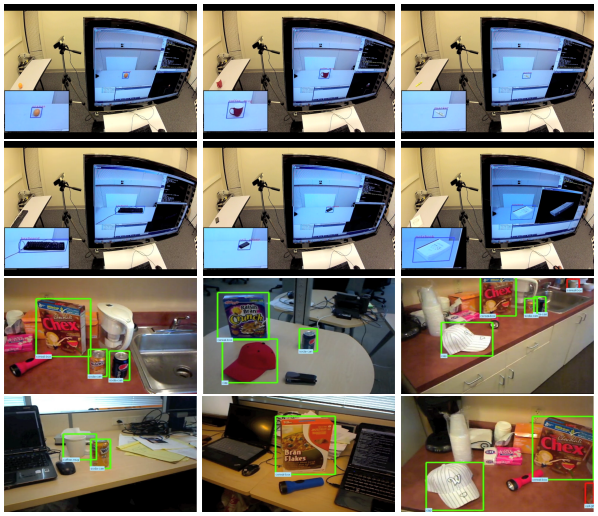

Figure 4: Near real-time system running on single table-top objects (first 2 rows) and the RGBD video dataset (last 2 rows)

6 Conclusion

We have presented a unified framework based on homeomorphic mapping between a common manifold representation and different object manifolds to solve the subproblems of object recognition. Extensive experiments on recent datasets validates the strength of this approach. We significantly outperform state-of-the-art in pose recognition. For category and instance recognition we achieve similar performance to state-of-the-art. We also outperform the state-of-the-art in two challenging multi-view visual-only datasets. We have also built a working system for application to the field of AI and robotic visual reasoning that performs table-top object detection and category recognition using the Kinect sensor.

Acknowledgments: This work was partially supported by the Office of Navel Research grant N00014-12-1-0755.

References

Arulampalam, M.; Maskell, S.; Gordon, N.; and Clapp, T. 2002. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing* 50(2):174–188.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition 2005*, volume 1, 886–893.

El-Gaaly, T.; Zhang, H.; Torki, M.; Elgammal, A.; and Singh, M. 2012. Rgbd object pose recognition using local-global multi-kernel regression. In *International Conference on Pattern Recognition 2012*.

Elgammal, A., and Lee, C.-S. 2004. Separating style and content on a nonlinear manifold. In *IEEE Conference on Computer Vision and Pattern Recognition 2004*, volume 1, I–478–I–485.

Felzenszwalb, P. F., and Huttenlocher, D. P. 2005. Pictorial structures for object recognition. *IJCV* 61(1):55–79.

Ham, J. H.; Lee, D. D.; and Saul, L. K. 2005. Semisupervised alignment of manifolds. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 120–127.

Jolliffe, I. T. 1986. *Principal Component Analysis*. Springer-Verlag.

Kimeldorf, G. S., and Wahba, G. 1970. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* 41(2):495–502.

Kimeldorf, G., and Wahba, G. 1971. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* 33(1):82–95.

Kinect. 2010. Microsoft kinect. www.xbox.com/en-us/kinect.

Lai, K.; Bo, L.; Ren, X.; and Fox, D. 2011a. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE International Conference on Robotics and Automation (ICRA)*, 1817 –1824.

Lai, K.; Bo, L.; Ren, X.; and Fox, D. 2011b. A scalable tree-based approach for joint object and pose recognition. In *In Twenty-Fifth Conference on Artificial Intelligence (AAAI)*.

Lamdan, Y., and Wolfson, H. 1988. Geometric hashing: A general and efficient model-based recognition scheme.

Lathauwer, L. D.; de Moor, B.; and Vandewalle, J. 2000. A multilinear singular value decomposiiton. *SIAM Journal On Matrix Analysis and Applications* 21(4):1253–1278.

Lawrence, N. D. 2004. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Liebelt, J., and Schmid, C. 2010. Multi-view object class detection with a 3d geometric model. In *IEEE Conference on Computer Vision and Pattern Recognition 2010*, 1688–1695.

Murase, H., and Nayar, S. 1995. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision* 14:5–24.

Ozuysal, M.; Lepetit, V.; and Fua, P. 2009. Pose estimation for category specific multiview object localization. In *IEEE Conference on Computer Vision and Pattern Recognition 2009*, 778–785.

Payet, N., and Todorovic, S. 2011. From contours to 3d object detection and pose estimation. In *IEEE International Conference on Computer Vision 2011*, 983–990.

Poggio, T., and Girosi, F. 1990. Networks for approximation and learning. *Proceedings of the IEEE* 78(9):1481–1497.

Rusu, R. B., and Cousins, S. 2011. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*.

Rusu, R.; Bradski, G.; Thibaux, R.; and Hsu, J. 2010. Fast 3d recognition and pose using the viewpoint feature histogram. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2155–2162.

Savarese, S., and Fei-Fei, L. 2007. 3d generic object categorization, localization and pose estimation. In *IEEE International Conference on Computer Vision 2007*, 1–8.

Seung, H. S., and Lee, D. D. 2000. The manifold ways of perception. *Science* 290(5500):2268–2269.

Sivic, J.; Russell, B. C.; Efros, A. A.; Zisserman, A.; and Freeman, W. T. 2005. Discovering objects and their location in images. In *ICCV*.

Sun, M.; Su, H.; Savarese, S.; and Fei-Fei, L. 2009. A multi-view probabilistic model for 3d object classes. In *IEEE Conference on Computer Vision and Pattern Recognition 2009*, 1247–1254.

Tenenbaum, J. B., and Freeman, W. T. 2000. Separating style and content with bilinear models. *Neural Computation* 12(6):1247–1283.

Tenenbaum, J. B. 1998. Mapping a manifold of perceptual observations. In *Advances in Neural Information Processing Systems 10*. Cambridge, MA: MIT Press.

Thomas, A.; Ferrari, V.; Leibe, B.; Tuytelaars, T.; Schiel, B.; and Van Gool, L. 2006. Towards multi-view object class detection. In *IEEE Conference on Computer Vision and Pattern Recognition 2006*, volume 2, 1589–1596.

Tipping, M. E., and Bishop, C. M. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(3):611–622.

Torki, M., and Elgammal, A. 2011. Regression from local features for viewpoint and pose estimation. In *IEEE International Conference on Computer Vision 2011*, 2603–2610.

Vasilescu, M. A. O., and Terzopoulos, D. 2002. Multilinear analysis of image ensebles: Tensorfaces. In *Proc. of ECCV, Copenhagen, Danmark*, 447–460.

Willamowski, J.; Arregui, D.; Csurka, G.; Dance, C. R.; and Fan, L. 2004. Categorizing nine visual classes using local appearance descriptors. In *IWLAVS*.