Joint Object Recognition and Pose Estimation using a Nonlinear View-Invariant Latent Generative Model

Amr Bakry* Tarek Elgaaly* Mohamed Elhoseiny Ahmed Elgammal
Rutgers University
{amrbakry, elgaaly, elhoseiny, elgammal}@cs.rutgers.edu

Abstract

Object recognition and pose estimation are two fundamental problems in the field of computer vision. Recognizing objects and their poses/viewpoints are critical components of ample vision and robotic systems. Multiple viewpoints of an object lie on an intrinsic low-dimensional manifold in the input space (i.e. descriptor space). Different objects captured from the same set of viewpoints have manifolds with a common topology. In this paper we utilize this common topology between object manifolds by learning a low-dimensional latent space which non-linearly maps between a common unified manifold and the object manifold in the input space. Using a supervised embedding approach, the latent space is computed and used to jointly infer the category and pose of objects. We empirically validate our model by using multiple inference approaches and testing on multiple challenging datasets. We compare our results with the state-of-the-art and present our increased category recognition and pose estimation accuracy.

1. Introduction

Visual object recognition and pose estimation are fundamental problems in the field of computer vision. Recognizing objects and their poses/viewpoints are critical components of vision and robotic systems. With the pervasiveness of robots today, they are required to not only visually recognize objects but also grasp and interact with them. For this reason simultaneously recognizing objects as well as their poses is of utmost importance. The difficulty of the problem lies in the large variation in appearance within object categories and between varying poses of the same objects.

Recent research in the field of generic object categorization and pose estimation can be divided into four tracks, depending on how the models deal with different views and different categories. First, models ignore estimating object pose and learn discriminative object models from training data to categorize objects. The assumption here is that the representation and classifier become view-invariant. With limited size training data, this assumption is hard to meet in reality. Second, learn view-specific models for object category recognition. These approaches discretize the view space into a small number of canonical views (e.g., front, right, back, rear-right) and learn classifiers for each of them. Thirdly, there are approaches that learn category-specific models, with the aim of estimating the viewpoint at a finer granularity, e.g. [23, 13, 21]. Finally, few recent approaches aim at learning a joint representation of object categories and poses [10, 25].

In this paper, we tackle the problems of category recognition and pose estimation simultaneously through a joint representation. The intuition behind our model follows from [25], where a common topology is used as a central representation of the multiple views of all objects (e.g. a unit-circle manifold for views of an object rotating on a turn-table). All objects are assumed to share the same topology and that there is a homeomorphism between these manifolds and the feature/input space for each set of multiple...
views per object. The space of all mappings (between common topology and input/feature space) for the objects encodes the variation in appearance and shape. The model is then factorized into a bi-linear generative model over two latent factors: \textit{content} and \textit{style}, representing the parameterization on the common topology and parameterization across different mappings, respectively.

To address the limitations and leverage the model proposed in [25], we first factorize the manifold-mapping space in a supervised way to utilize the available class labels. Secondly, we propose a generative model that is nonlinear in both category and pose latent variables. This helps to capture the nonlinearities in the latent factors. We also address scalability limitation of [25] by proposing a hierarchical model where a discriminative model is used to classify super-categories and the generative models can be used on these super-categories.

There are three main contributions of this work: (1) we propose a novel framework for view-invariant category recognition and pose estimation. This framework presents a nonlinear method for separating the manifold parameterization (referred to as \textit{style}) and pose variations over the manifold (referred to as \textit{content}) in sets of images. Our generative model is a purely non-linear approach which represents both category and pose using nonlinear latent space embedding. This is in contrast to previous approaches that do not represent the non-linearities across object categories. For this reason our approach has the advantage of being more robust to within-category and pose variations. (2) Our framework of \textit{style/content} factorization moves the inference of the category and pose from the very high-dimensional feature space into two orthogonal low-dimensional spaces, one for category and the other for pose. Inference in a low-dimensional space guarantees increased accuracy and computational performance. Our framework uses supervised manifold embedding in a low-dimensional space and thus increases the point clustering, and in turn the classification accuracy. (3) We present the use of different distance metrics, different optimization techniques and compare the results of these configurations through extensive experimentation.

To validate our approach, we present theoretical derivations of the generative model and also test our model on three large challenging datasets. We compare our results with state-of-the-art approaches and show the advantages of our model.

We first describe the related work in section 2. Our latent generative model is described in section 3. Learning of the model is described in section 4 and inference is presented in section 5. Section 6 describes the datasets, experiments and results.

2. Related Work

We refer the reader to [20] for a comprehensive review of recent work on object recognition and pose estimation. We will highlight the most relevant research in this section.

Successful work have been done in estimating the object pose of a known category of objects [5, 13, 21, 23]. These models have the limitation of being category-specific. This stipulates that independent models are learnt for different object. This does not scale well to many categories with high intra-class variation.

Recently, category recognition and pose estimation have been solved simultaneously (e.g. [19, 10]). In [19], multiple-view object model is addressed by linking distinct parts of objects from different (discrete) views. This model belongs to the category of limited-pose (discrete-pose) object recognition since it uses a classification approach to deal with pose estimation. Very few works formulate the problem of pose estimation as a regression problem over a continuous space. In [10], a semantic structured tree is built for doing hierarchical inference (object category, instance and pose recognition). This work involves a classification strategy for pose recognition which results in coarse pose estimates and does not fully utilize the information present in the continuous distribution of descriptor spaces. Work presented in [25] and [23] explicitly model the continuous pose variations of objects. In [25], the authors do not solve the problem of large category confusion which we explicitly solve for using our supervised embedding and hierarchical clustering model. In addition we are able to learn non-linearities in the category parameterizations which [25] do not handle.

Solving object recognition and pose estimation using a manifold-based representation is not novel. The pioneering work of Murase and Nayar explored this idea in [14]. However, the recognition in [14] is for object instances and not for generic categories. The model used is a linear model based on PCA, while our model is nonlinear in both the pose and the category. Manifold based representations have also been recently used in [13], however for object-specific view estimation on video sequences.

3. Dual Latent Generative Model

Images of an \textit{object instance} consists of a set of images of the same object from different viewpoints. Let the \(k\)-th instance \(S^k = \{(x^k_i, v^k_i, y^k)\}, i = 1, \cdots, N_k\) be the set of images \((x^k_i \in \mathbb{R}^D)\) for the object labeled by its class \(y^k \in \{1, \cdots, C\}\). \(v^k \in \mathbb{R}^v\) is the viewpoint label of the image \(x^k_i\). We propose a generative model that generates points in the image space from two independent latent spaces: pose \(\theta \in \mathbb{R}^{d_1}\) and category \(t \in \mathbb{R}^{d_2}\).

The images of each instance are assumed to lie on low-dimensional manifold, which we call \textit{instance manifold}. Inference in a low-dimensional space guarantees increased accuracy and computational performance. Our framework uses supervised manifold embedding in a low-dimensional space and thus increases the point clustering, and in turn the classification accuracy. (3) We present the use of different distance metrics, different optimization techniques and compare the results of these configurations through extensive experimentation.
(\mathcal{M}^k \in \mathbb{R}^D). For the case of view manifold, we assume that all instance manifolds have the same topology. Under the assumption that all the input object instances are captured using the same degrees of freedom between the camera and object, the corresponding instance manifolds are topologically equivalent but have different geometry in \( \mathbb{R}^D \). We can find a unified manifold that is topologically equivalent to all these instance manifolds. The dimensionality of the unified manifold depends on the degrees of freedom allowed for the viewing conditions. In the case of views of an object rotating on a turn-table, and assuming no degeneracy, a viewing circle manifold (one-dimensional unit circle) is used. This can be extended to a viewing sphere (two-dimensional unit sphere). All manifolds are internally parameterized by a latent variable \( \theta \) lying on the unified manifold, representing the viewpoint. This unified manifold is homeomorphic to the input/feature space of each object instance.

As a result of the homeomorphism, each manifold geometry can be parameterized by its geometric deformation from the unified manifold. This parameterization space is view-invariant. The large dimensionality of the manifold parameterization space makes the inference hard and non-robust. Therefore, we learn low-dimensional representation for the deformation space. We use supervised kernel-based partial least squares (K-PLS [18]) to discover this low-dimensional latent space. As a result, the latent variable \( t \) is a view-invariant category representation, and encodes the appearance and geometric characteristics of object instances.

The generative model can be formalized as

\[
\hat{x}(t, \theta) = \mathbf{A} \times_1 \phi(t) \times_2 \psi(\theta)
\]

where \( \phi \) is the nonlinear mapping of the latent space and \( \psi \) is a Radial-Basis Function (RBF) over the viewpoints \( \theta \).\(^1\) This model is based on nonlinear regression over two factors: the pose \( \theta \) and latent category representer \( t \). Figure 1 shows graphical representation of the used generative model. This figure shows that for each point in the image space it corresponds to a point on the unified manifold (\( \Theta \)) and a point in the low-dimensional parameterization space (\( \mathbf{T} \)). A new point in the image space can be computed using the tensor \( \mathbf{R} \) in Eq. 1.

4. Learning the Model

As shown in Figure 1 and described in [25], each object instance is parameterized using a nonlinear mapping from a unified manifold representation to the input feature space (Section 4). This manifold parameterization is then embedded in a low-dimensional latent space using a supervised dimensionality reduction (DR) method called K-PLS (Section 4). We then learn the nonlinear mapping from the latent space to the parameterization space to complete the generative model (Section 4). The details of each step is provided in this section.

**Manifold Parameterization** The first step is to find a parameterization for each instance manifold. We adapt the approach in [25]. Let \( \{x_i^k \in \mathbb{R}^D, i = 1, \cdots, n_k\} \) be a set of points on instance manifold \( \mathcal{M}^k \). Let \( \{z_i^k \in \mathbb{R}^c, i = 1, \cdots, n_k\} \) be the corresponding points on the unified manifold \( U \).

We learn a regularized mapping functions \( \gamma^k(\cdot) : \mathbb{R}^c \to \mathbb{R}^D \), which maps from \( U \) to each instance manifold \( \mathcal{M}^k \). The mapping can be written in the following matrix form.

\[
\gamma^k(z) = C^k \psi(z)
\]

where \( C \) is a \( D \times n \) matrix, the vector \( \psi(z) = [k(z, w_1), \cdots, k(z, w_N)]^\top \) represents a nonlinear kernel map from the embedded representation to a kernel induced space, given a set of RBF centers \( \{w_1, \cdots, w_N\} \). \( k(\cdot, \cdot) \) is an RBF kernel. The solution of Eq 2 is shown in [16] to have a closed form solution:

\[
C^k = (A^k \mathbf{A}^k + \lambda \mathbf{G})^{-1} A^k \mathbf{X}^k, \quad (3)
\]

where \( A^k \) is a \( n_k \times n \) matrix with \( A_{ij} = k(z_i, w_j), i = 1, \cdots, n_k, j = 1 \cdots n \) and \( \mathbf{G} \) is a \( n \times n \) matrix with \( G_{ij} = k(w_i, w_j); i, j = 1 \cdots n \). \( X^k \) is the \( n_k \times D \) data matrix for manifold instance \( k \).

**Supervised Manifold Embedding** The second phase of our framework is to learn good\(^2\) low-dimensional embedding for the parameterizations. Consider Eq 2, if we ignore the superscript \( k \), and if we have a test image \( x \), then the objective is to find the best \( C^* \) and \( z^* \) that minimizes the reconstruction error: \( \delta(x - \gamma^k(C^*, \psi(z^*))) \). However, the generated parameterization (\( C \)'s from Eq 3) populate a high-dimensional space. This makes solving the objective hard and non-robust. Therefore, there is a need to find a suitable low-dimensional latent space for those parameterizations.

Subspace analysis is used in [7] to obtain a latent representation of the manifold parameterization space. However, these approaches do not benefit from available class labels. Alternatively we propose a supervised way to achieve a low-dimensional latent manifold parameterization space, which benefits from the class labels. We use Partial Least Squares (PLS) [24, 18, 2].

PLS compromises between PCA and LDA dimension reduction methods by generating orthogonal components (in

\(^1\)For convenience: we use bold small letters for vectors, bold capital for matrices, calligraphic capital for tensors and regular small for scalars. \( \times_i \) denotes mode-i tensor vector multiplication as defined in [11].

\(^2\)Good in sense that balance between encoding the spatial relationship between points in the original space and separate points with different labels.
the latent space) using the already existing correlations between observation variables (in the input space) and corresponding labeling, while keeping most of the variance of the points in the input space [17]. A good interpretation for PLS and its relationship with PCA can be found in [12, 2]. PLS has been proven to be useful in situations where the dimensionality of the input space significantly exceeds the number of observations (sparse case) and/or multicollinearity\(^3\) is high among the explanatory variables. PLS embeds the input points \(X\) into low-dimensional latent space \(T\), so that it satisfies: \(\arg \min T \|XW - T\|\) and \(\arg \max_T \text{cov}(T, y)\). \(W\) is the learned projection matrix and \(y\) is the set of labels. A kernelized version of PLS (K-PLS) [18] is more common in the case of coefficient matrices (C’s). First, we need to define a kernel on the parameterization space: since \(C\) is \(D \times N\) matrix, and \(D \gg N\), then \(C\) represents \(N\)-dimensional subspace in \(\mathbb{R}^D\). Therefore, we can use Cosine-Similarity kernel (CSK) or Grassmannian kernels [8]. In this work, we use CSK, for its efficiency, defined by

\[
K(C_i, C_j) = \frac{tr(C_i C_j^\top)^2}{\|C_i\|_F \|C_j\|_F},
\]

where \(\|\cdot\|_F\) is matrix Frobenius norm.

Given the instance manifold parameterization \(C^k\) and parameterization kernel and label \(y^k\), we use K-PLS for embedding parameterizations space into low-dimensional latent space. The points in this latent space satisfy the two objectives of the PLS.

K-PLS maps the point \(C^k\) to latent points \(\{t^k \in \mathbb{R}^m\}, \text{for each } k = 1 \cdots N\), by

\[
t = K(C, \cdot)W
\]

where \(W\) is a non-linear projection matrix. Like PCA, the choice of latent dimensionality (\(m\)) is a crucial step, since small and large values of \(m\) lead to under-fitting and over-fitting of the training data, respectively.

**Nonlinear Back Map from Latent Space to Parameterization Space** Almost all linear DR techniques, such as (linear) PLS and PCA, provide 2-way mapping from input space to latent space and vice-versa. In contrast, almost all nonlinear DR techniques do not provide mapping back from latent to input space. K-PLS (nonlinear DR) does not provide a closed-form mapping from the latent space \(T\) to the parameterization space \(C\). We need this reverse mapping to complete the generative model in Eq 1 and illustrated in Figure 1. In this work, we learn non-linear Gaussian RBF mapping \(\beta : T \rightarrow C\).

\[
C = \beta(t) = A \times_1 \phi(t)
\]

where \(C\) is an \(n \times D\) matrix, \(t\) is a \(m\)-D vector and \(A\) is a \(h \times n \times D\) tensor, and

\[
\phi(t) = K(t, \cdot) = \exp(\sigma \|t - u_i\|)
\]

where \(u_i; i = 1 \cdots h\) are the centers of the RBF kernel. Those centers are the \(h\)-means of the training points in the latent space.

The tensor \(A\) is computed as follows. Let \(c\) (\(nD\)-dimensional vector) is the vectorization of \(C\) (\(n \times D\) matrix).

\[
c = A^\top \phi(t)
\]

Then we can find the mapping \(A\) by the same way we computed \(C\) in Eq 3. Finally, \(A\) (\(h \times n \times D\) tensor) is the reshaped version of \(A\) (\(h \times nD\) matrix).

Substituting Eq 6 in Eq 2 results in the full bi-linear generative model: \(x(t, \theta) = A \times_1 \phi(t) \times_2 \psi(\theta)\).

**5. Inference**

This section shows how to use the proposed generative model for inferring the best pose \((v^*)\) and class \((y^*)\) for a given test image \((x)\). This is done in two steps: first, find the embeddings in the two latent spaces \((t^*, \theta^*)\) that optimizes

\[
(t^*, \theta^*) = \arg \min_{(t, \theta)} \delta(x, \hat{x}(t, \theta))
\]

where \(\delta(\cdot)\) is a distance metric in the feature space, and \(\hat{x}\) is the predicted image. Then, process \((t^*, \theta^*)\) to get the labels \((v^*, y^*)\). In this work, we use two distance metrics to measure the difference between the test image and the predicted image \(\hat{x}^\top(x, \hat{x})\): Euclidean distance \(\|x - \hat{x}\|\) and Normalized Cross-Correlation (NCC) \((1 - \frac{\sum \hat{x}^\top(x, \hat{x})}{\|x\| \|\hat{x}\|})\). We adopt two optimization techniques for solving Eq 8: gradient-based method and sampling-based methods.

**5.1. Optimization**

**Gradient-based Method** Our gradient-based optimization uses the second order BFGS quasi-Newton optimizer with cubic polynomial line search for optimal step size selection [1].

To use this algorithm, for each value of \(t\) and \(\theta\), we need to compute the distance \(\delta(x, \hat{x}(t, \theta))\), and its derivative. The derivation is shown in the supplementary material.

**Sampling-based method** In this method we use MCMC sampling [3] to solve Eq 8. We use simulated annealing [4] to enhance resampling of particles. The details of the algorithm are provided in Algorithm 1.

**5.2. Classification**

At this point, for a given test image, we found the best match in both the category latent space \((t^*)\) and the pose latent space \((\theta^*)\). We need to infer the label for object category \(y^*\) and for pose \(v^*\).
For pose, we have two cases: the actual pose label is discrete, then we use $v^* = k - N(\theta^*)$, and if the pose label domain is continuous we set $v^* = \theta^*$.

For category, we use SVM in the category latent space, and also we use Regression for classification (RfC), i.e., we use the regression results of K-PLS [18]

$$y^* = t^\top T^\top y$$ (9)

where $T$ is the set of embedded training points in the K-PLS latent space (Eq 5), and $y$ is the corresponding labels of the training points.

5.3. Hierarchical Model

Dealing with large datasets containing object images with large visual and semantic similarity - e.g. tableware, fruits, etc. found in the RGBD dataset [9] is quite challenging. To solve this we extend our framework to a hierarchical model which performs recursive spectral clustering [17] to discover clusters of similar objects (as seen in 2). For each super-category we use the same generative model described to learn the latent space and perform inference. A Support Vector Machine (SVM) classifier is trained to classify a test image into each super-category and then the generative model for that particular super-category is used to infer the category and pose.

6. Experiments and Results

We experiment on three challenging multi-view datasets: 3DObjects [19], RGB-D [9] and EPFL [15]. In this section, we outline these experiments.

We use HOG features [6] as our input image representation and we use a unit circle as the unified manifold, i.e, $\theta \in [0, 2\pi]$.

For inference (Section 5), we use gradient-based optimization with Euclidean distance and sampling-based optimization with NCC distance metric. The results here are reported based on these best configurations.

6.1. 3DOObjects

We show the experimental results of our work applied to 3D Objects dataset and compare our results to [19, 25]. The 3D Objects dataset has images of 10 objects categories. Every category has 10 different instances, differing in brand, color and shape. Every instance has images captured at 3 heights and 3 scales. Every image sequence has 8 poses covering all views: back, back-right, right, front-right, front, front-left, left, back-left.

We show our results for 2 different configurations: 1) 8 classes excluding the farthest scale and 2) All classes (details in supplementary material). Pose accuracy is reported even if the object is incorrectly classified. We use the first configuration for comparison purposes because this is the configuration in prior work ([19],[25]). For all experiment, we train over 7 instances (brands/shapes) and test on remaining 3 brands. The final results are the average of several folds.

Table 1 and 2 show both our sampling and gradient based methods compared against the baselines. We can see an improvement of about 38% in pose recognition (Table 1(right)). Table 2 shows the comparison of the gradient optimization and sampling optimization in both category and pose recognition using different classification techniques (as described in Section 5.2). For pose, we computed the percentages of poses that are less than 22.5° and 45° away from groundtruth. The gradient-based optimization technique achieves the highest accuracy.

Table 1 shows the recognition rate of every category compared to the baselines. We used a 13-dimensional K-
PLS latent space. Inference can be done efficiently on very low-dimensional latent spaces instead of the very high-dimensional coefficient mapping space. From the table, it is clear that our model outperforms state-of-the-art results in most categories. We achieve an overall increase of up to about 10% in accuracy using sampling-based inference and 11.8% using gradient-based inference.

### 6.2. EPF Cars

To compare with previous work ([15],[22] and [23]), we use the same experiment configuration, we train over the sequences of first ten cars (cars from 1 to 10) and test over the last ten cars (cars from 11 to 20). In this experiment test images are for car instances that are not present in the training data. Our framework finds the closest instance to the query car instance and estimates the pose.

Table 3 shows the results of this configuration compared with previous work. The result are: 90.81% have error less than 22.5 degrees compared to 41.69% reported in [15], 70.31% reported in [23] and 78.1% reported in [22]. This means that we achieved 16.27% improvement. This shows the power of our framework for continuous pose estimation. We also reduce the mean absolute error (MAE) by about 45%. Leave-one-out has significant improvement over 50% split, the reason behind this is that our algorithm has a wider space (more car instances) to search for the closest point to the query car. More discussion of the results and figures are in the supplementary material.
Table 4. Category recognition and pose estimation accuracy (%) on the RGB-D dataset. We report the RGB-only accuracy of [25]. [10] only report their multi-modal RGB+D accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Category</th>
<th>Avg. Pose (C)</th>
<th>Avg. Pose (All)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ours (visual only)</td>
<td>85.00</td>
<td>77.31</td>
<td>73.78</td>
</tr>
<tr>
<td>visual only baseline [25]</td>
<td>92.00</td>
<td>80.01</td>
<td>61.57</td>
</tr>
<tr>
<td>visual + depth baseline [10]</td>
<td>94.30</td>
<td>86.80</td>
<td>53.50</td>
</tr>
</tbody>
</table>

7. Discussion

The proposed model is supervised and nonlinear, in contrast to the linear-unsupervised previous approaches. The use of supervision to achieve the category latent space is fundamental to retain the discriminative ability, while reducing the dimensionality for better inference in this space. This resulted in significantly better results in 3DObjects (category + pose), EPFL-Cars (pose) and RGB-D (pose). For the classification in RGB-D dataset, although the predicted images are notably similar to the corresponding test image, deficit in the recognition accuracy is remarked. We believe that this is because K-PLS fails to find the best compromise between minimizing reconstruction of the coefficient mapping space and maximizing the correlation with labels. (see the supplementary material)

While the use of supervision to achieve latent spaces has been studied before, this has not been addressed in the context of the space of manifold deformations as presented here. Adding supervision is not trivial for high-dimensional parameterization spaces, and results in a nonlinear latent space, which mandates the use of a nonlinear model as we have presented. The nonlinearity in the model requires developing suitable inference methods; we compare between sampling and gradient-based inference methods.

Unfortunately there is no large multi-view dataset to really evaluate scalability. Category-specific models for pose estimation (e.g. [13, 21]) are not scalable, since one model per category is needed, and these models do not allow sharing knowledge among similar classes. In contrast, we aim for a model that solves simultaneously for category and pose, where there is one common representation for all categories, hence the scalability potentials. Interestingly, we achieved better pose estimation over many category-specific models, even though our model combines all objects. Compared to [25], our approach is scalable in two fundamental ways. First, in [25], the category latent space was the subspace spanned by all training data, which is still high dimensional for inference. [25] used 300 dimensional space for the RGB-D dataset. We only use 20 dimensions and achieve comparable results. In addition, [25] used a k-NN classifier because of the use of the full subspace. This is not scalable and not robust, therefore, supervised DR is needed to achieve a low dimensional representation that retains the discriminative power. Second, we propose a hierarchical Model, which is essential for scalability.

Supervision and nonlinearity are coined together in our use of K-PLS. Comparing our sampling results with [25] (also used sampling, although no algorithm was provided) shows the value of the unsupervised-linear/supervised-nonlinear (88.4% vs 80% Table 1).

8. Conclusion

We present a novel framework for recognizing object category and pose estimation. The framework uses a generative model that is based on homeomorphic manifold analysis, supervised manifold embedding into a latent space and nonlinear mapping. The advantage of our model is that the inference procedure is moved from the very high-dimensional coefficient mapping space to two low-dimensional orthogonal pose and category spaces, which makes the inference more accurate and computationally easier. We also incorporate this model into a hierarchical structure to deal with large intra-class variation. We show theoretical basis of our framework and compare our results with the state-of-the-art. We show that our framework both achieves higher performance than state-of-the-art in many configurations and is comparable in others.

References


