Robust Object Pose Estimation via Statistical Manifold Modeling

Liang Mei, Jingen Liu, Alfred Hero, Silvio Savarese
Department of EECS
University of Michigan, Ann Arbor
{mliang,liujg,hero,silvio}@umich.edu

Abstract

We propose a novel statistical manifold modeling approach that is capable of classifying poses of object categories from video sequences by simultaneously minimizing the intra-class variability and maximizing inter-pose distance. Following the intuition that an object part based representation and a suitable part selection process may help achieve our purpose, we formulate the part selection problem from a statistical manifold modeling perspective and treat part selection as adjusting the manifold of the object (parameterized by pose) by means of the manifold "alignment" and "expansion" operations. We show that manifold alignment and expansion are equivalent to minimizing the intra-class distance given a pose while increasing the inter-pose distance given an object instance respectively. We formulate and solve this (otherwise intractable) part selection problem as a combinatorial optimization problem using graph analysis techniques. Quantitative and qualitative experimental analysis validates our theoretical claims.

1. Introduction

The ability to accurately estimate the pose of generic object categories from videos or images is crucial in many applications such as robotic manipulation, human-object interaction and image indexing. A large literature in this area has mostly focused on detecting and/or estimating object poses from the single instance of a rigid object [25, 20, 3, 13, 14, 16, 26, 5, 11]. In this class of problems, the object, whose pose one wants to recognize, is already observed in a training stage. Although these methods have demonstrated competitive results, the extension to pose classification of object categories is not trivial. As Fig. 1 shows, two main issues must be addressed in category-level pose estimation: (A) Intra-class variability: the appearance of object instances from a particular viewpoint may change dramatically because of changes in illumination conditions, occlusions, shape and appearance properties; (B) Inter-pose variability: different poses of a specified object share similar appearance wherein only a small portion of the object carries key information for pose discrimination. Obviously, it is critical to simultaneously minimize the intra-class variability while maximizing the intra-pose distance for category-level pose estimation. Recent works have leveraged machine learning methods to classify object pose at categorical level from single images [32, 7, 19, 29, 31, 12, 23] or videos [21]. While most of these works mainly focus on the problem of minimizing intra-class variability (issue A), little attention has been put to simultaneously tackle both issue A and issue B.

In this paper, we propose a novel statistical manifold modeling approach that is capable to classify poses of object categories from video sequences by simultaneously minimizing intra-class variability and maximizing inter-pose variability. We use [21] as a starting point for our work. In [21] authors show that a more compact and descriptive statistical manifolds can be learnt from short video sequences rather than still images, and that these manifolds enable accurate pose classification of object categories. Unlike in [21], however, where an holistic object pose representation is used, we follow the intuition that a part based representation (or pictorial structure or constellation model) [9, 18, 10] is capable of delivering more design flexibility to handle intra-class variability, occlusion and background clutter. As illustrated in Fig. 1(b), careful part selection is indeed the critical ingredient that allows us to tame intra-class confusion (issue A) and pose ambiguities (issue B) at the same time. We formulate the informative part selection problem from a statistical manifold modeling perspective, and treat part selection as adjusting the manifold structure of the object poses by means of an "alignment" and "expansion" operation. As demonstrated in Fig.2(a), wherein each trajectory corresponds to one object instance with varied poses, manifold alignment and expansion are equivalent to minimizing the intra-class distance given a pose while increasing the inter-pose distance given an object instance respectively. We formulate and solve this (otherwise intractable) part selection problem as a combinatorial optimization problem

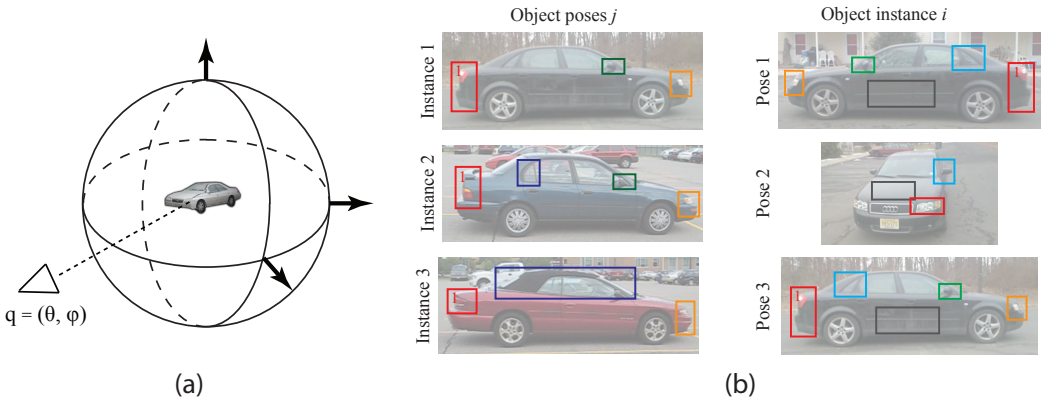

Figure 1. **(a)** Object pose is defined as the camera location (θ, ϕ) on the viewing sphere from which the object is observed. **(b)** Two main challenges in categorical object pose estimation. **Issue A**(Left column): Due to intra-class variability, different object instances have distinctive appearances given the same pose. We aim at representing the object pose using parts (e.g., orange part, red part and dark-green part) that minimize the intra-class variability. **Issue B**(Right column): Objects observed from different poses may share similar appearance. We wish to find parts (e.g., red part and light-blue part) that maximize the inter-pose distance. The dark-green parts minimize intra-class variability but not inter-pose distance, thus should not be used to represent the object.

using graph analysis techniques.

Notice that our problem is conceptually different from traditional (object) classification problems. In these cases, inter-class variability must be maximized so as to increase the margins between a discrete number of classes in the feature space. In our case, maximizing inter-pose distance means spreading (inflating) the structure of the manifold which is constructed as a continuous (rather than discrete) function of the pose parameters. Thus, we argue that traditional part (feature) selection methods such as LDA/Fishers Discriminant Analysis [24] (wherein the between-class scatter matrix is constructed over a discrete number of classes) are actually not adequate for solving our problem. Also, our work differs conceptually from previous work on manifold learning such as [17, 33, 27, 6] where the primary goal is to recover the structure of the manifold (so as to facilitate classification and visualization) rather than modifying the actual structure of the manifold as we seek to do.

A number of experimental results (on a public dataset [21] and on an extension of [21]) demonstrate that our methods increases discrimination power in pose classification even in presence of large intra-class variability, background clutter, and occlusions. We show that our method achieves superior pose classification rate than state-of-the art holistic approaches [21, 28] as well as benchmark methods based on feature selection such as LDA/FDA [24].

2. Overview of Our Approach

As Fig. 1 (a) shows, an object pose is defined as a tuple (θ, ϕ) corresponding to the azimuth and zenith angles. To estimate the object pose, our input is a video sequence that captures an object under a certain range of poses, rather than still images. The video sequence is split into a set of short video segments and each of them is associated with a unique pose. Then each video segment is represented by a collection of parts, which are characterized by the appearance of the surrounding spatial-temporal volumes, called Spatial-Temporal (ST) parts (Fig.2(b)). In fact, each ST part contains a sequence of patches, from which we can estimate a PDF associated with the ST part. This PDF captures the appearance and geometrical location distribution of patches within the video. Thus, each patch within the ST part is a realization of the ST-Part PDF. Then an object pose is modeled as a joint distribution of the parts. Specifically, each frame of the video segment corresponding to a pose is regarded as a realization of the joint distribution of the ST parts. Consequently, each object PDF will be uniquely associated with one pose.

In order to distinguish object poses, we design a distance function based on the Kullback-Leibler divergence between two pose PDFs. Since an object is represented by a collection of parts, the KL-distance between two object poses is based on the joint distributions of parts (**Sec**.3.1). We then use the Multi-Dimensional Scaling (MDS) technique [4] to embed object pose PDFs into a Euclidian space (**Sec**.4). Video slices sampled from smooth trajectories on the viewing sphere will generally induce smooth manifolds in the embedding space (Fig.3). In the ideal case, where intra-class variability, illumination changes and occlusions are neglected, manifolds from different instances would perfectly align with each other, resulting in a 2D manifold parameterized only by pose (Fig.2(a)). Furthermore, if there are no appearance ambiguities across different poses (i.e., all of the poses are clearly distinguishable), manifolds will occupy the "largest hypervolume" - i.e., without inward

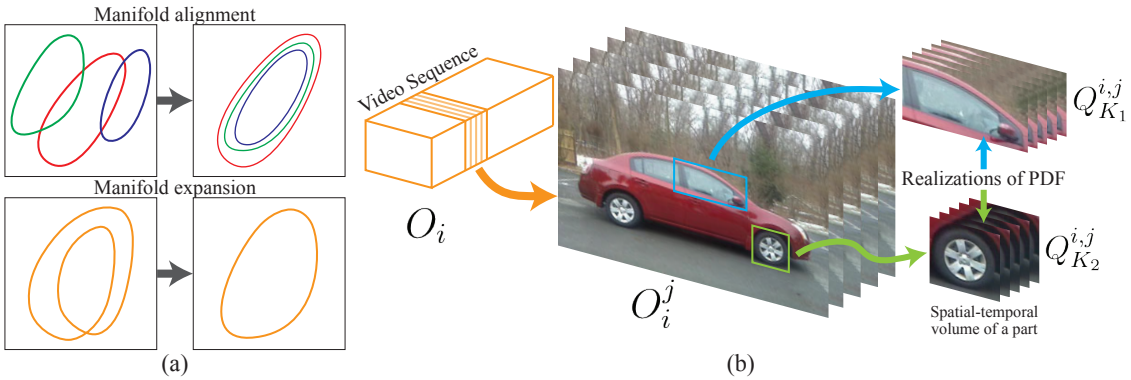

Figure 2. **(a)** Manifold alignment and expansion. Each manifold (a closed trajectory) corresponds to an object instance and it is parameterized by pose. **Up:** Manifold of different object instances should be aligned according to poses to minimize intra-class distance. **Down:** Quasi-symmetric poses usually induce loops in the manifold space. Expanding manifold of object instance to the largest hyper-volume will facilitate pose classification. **(b)** Part-based object pose representation. A long video sequence (left) is decomposed into video segment (center) that is associated with a unique pose. Each video segment consists of a set of discriminative spatial-temporal parts (right). Each video segment is the realization of a certain object PDF and each ST part is a realization of certain part PDF.

bends or loops. Unfortunately, this does not happen in practice (a typical pose manifold is shown in Fig.3, Left).

Observing that there is a strong correlation between the manifold structure and pose estimation accuracy (namely having good alignment and expansion of the manifold is equivalent to decreasing intra-class variability and increasing inter-pose distance), we propose a methodology that allows us to *model* the manifolds so as to increase the discrimination power of our classification scheme. In practice, we model the manifold structure by selecting ST parts, such that (1) manifolds of different object instances are best "aligned" at any given pose; (2) manifolds associated with each object instance are "expanded" as much as possible. This is well illustrated in Fig.2(a). The process of "alignment" and "expansion" are formulated mathematically as a combinatorial optimization problem by jointly optimizing two cost functions, which correspond to the problems of decreasing intra-class variability and increasing inter-pose distance respectively (**Sec**.4.2). We solve this otherwise untractable optimization problem using graph analysis techniques Sec 4.3).

3. Probabilistic Modeling of Pose Dissimilarity

Let V_i be a video sequence capturing an object instance i of a given category as the camera moves along a trajectory on the viewing sphere. As Fig.2(b) shows, this video sequence is temporally split into N video segments, and each of them is associated with one unique object pose j, thus $V_i = \{O_i^j\}_{j=1}^N$. Furthermore, each O_i^j is decomposed into a collection of ST-parts $O_i^j = \{Q_k^{i,j}\}_{k=1}^K$, and each ST-part Q consists of two variables (A, X) modeling part appearance (e.g., raw pixel intensity values, SIFT descriptor, etc.) and geometry (e.g., part position, size and aspect ratio) respectively. As a result, we define the probability density function (PDF) of object i observed at pose j as a joint distribution of the appearance and geometry of ST parts. Formally, it is defined as,

$$P_i^j = P_i^j(\{Q_k\}_1^K) \quad (1)$$

$$= P_i^j(A_1, \ldots, A_K, X_1, \ldots, X_K) = P_i^j(\mathbf{A}, \mathbf{X}), \quad (2)$$

where $\mathbf{A} = (A_1, \ldots, A_K)$ and $\mathbf{X} = (X_1, \ldots, X_K)$. The collection of object pose PDFs P_i^j for $j = 1, ..., N$ forms a trajectory T_i in a high dimensional manifold parameterized by object pose j.

3.1. Object Pose Dissimilarity Measure

Given two object pose realizations O_i^j and O_m^n (associated with object instances i and m respectively), as well as their corresponding PDFs P_i^j and P_m^n, we define the dissimilarity between two poses as the Kullback-Leibler divergence between two distributions P_i^j and P_m^n,

$$D(O_i^j, O_m^n) = D_{KL}(P_i^j || P_m^n) = D_{KL}(P_i^j(\mathbf{A}, \mathbf{X}) || P_m^n(\mathbf{A}, \mathbf{X})) \quad (3)$$

Claim 1. Given (3) the distance between two object poses O_i^j and O_m^n can be computed as,

$$D(O_i^j, O_m^n) = E_X(D_{KL}(P_i^j(\mathbf{A}|\mathbf{X}) || P_m^n(\mathbf{A}|\mathbf{X}))) \quad (4)$$
$$+ D_{KL}(P_i^j(\mathbf{X}) || P_m^n(\mathbf{X})).$$

Proof: see [22] for details.

We further assume that part appearance $\mathbf{A}$ is conditionally independent given the geometry $\mathbf{X}$ of the parts, and that part correspondence between O_i^j and O_m^n is known. Thus the mapping between $Q_k^{i,j}$ to $Q_k^{m,n}$ is known. Then we can derive the following proposition,

Claim 2. If $A_1, \ldots, A_K$ are independent given X and a one-to-one part correspondence $f_{map}(Q_k^{i,j}) = Q_k^{m,n}$ exists,

the following equation holds,

$$E_X(D_{KL}(P_i^j(\mathbf{A}|\mathbf{X})||P_n^m(\mathbf{A}|\mathbf{X}))) \quad (5)$$
$$= E_X(\sum_{k=1}^{K} D_{KL}(P_i^j(A_k|\mathbf{X})||P_m^n(A_k|\mathbf{X}))),$$

where $P_i^j(A_k|\mathbf{X})$ is the conditional PDF of part A_k given $\mathbf{X}$. Proof: see [22] for details.

Integrating Claim 1 and 2, the distance between two object poses is,

$$D(O_i^j, O_m^n) = E_X\{\sum_{k=1}^{K} D_{KL}(P_i^j(A_k|\mathbf{X})||P_m^n(A_k|\mathbf{X}))\} \quad (6)$$
$$+D_{KL}(P_i^j(\mathbf{X})||P_m^n(\mathbf{X})).$$

The second term in (6) ("geometry" term) is modeled as a $4 \times K$ dimensional Gaussian distribution corresponding to the normalized location, size and aspect ratio of all the parts in the image. The KL-divergence between them can be solved analytically. The PDFs in the first term ("appearance" term) is modeled non-parametrically. It can be estimated from a sequence of patches of the corresponding ST-part using kernel density estimation techniques similar to [21]. The pairwise KL-divergence is then calculated accordingly.

In practice, since the cardinality of parts may vary across two instances i and m, we utilize the "normalized" symetric KL-Divergence to measure the distance,

$$\hat{D}(O_i^j, O_m^n) = \frac{1}{S} D_{KL\_sym}(P_i^j, P_m^n) \quad (7)$$

where $D_{KL\_sym}(P,Q) = D_{KL}(P||Q) + D_{KL}(Q||P)$, and S is the number of parts shared by O_i^j and O_n^m.

3.2. Object Parts Extraction

Given a video segment, different segmentation procedures, such as [1][8], can provide salient regions which can be used as parts. In this work we adopt a simpler strategy and extract parts by dividing the video segments (frame) into a regular grid and associate each grid element to an ST part. This strategy has a number of advantages: i) it solves the problem of establishing the corresponding parts across different video segments; parts are in correspondence if they have the same grid index (i.e.,spatial location). ii) It is very efficient; iii) we can decouple the effect of segmentation from our part selection algorithm. Notice, however, that our proposed approach is general and can handle parts extracted by any segmentation algorithm (as long as part correspondence across video segments is provided). Note that given an ST-part Q_k^{ij}, A_k is the appearance of the corresponding grid element, X captures the grid index (i.e., part location) and grid element aspect ratio.

4. Manifold Modeling for Pose Estimation

Given a number of training object instances $i = 1, ..., M$, we form a set of trajectories T_i in the manifold parameterized by pose j. Our goal is to align the trajectories such that the pose distance $D(O_i^j, O_m^j)$ is minimized for each pair of object instances $i \neq m$ and for each pose $j = 1, ..., N$. This will result in reducing intra-class variability. Meanwhile, we seek to expand each trajectory T_i such that $D(O_i^j, O_i^n)$ is maximized for each $i = 1, ..., M$ and $j \neq n$. We argue this will result in increasing inter-pose distance for each object instance.

4.1. Euclidean Space Embedding

Due to the high dimensionality of the object pose representation $P(O_i^j)$, we propose to embed each trajectory into a Euclidean space $\mathbb{R}^d$ where d is much smaller than the original dimensionality. This has two key benefits: 1) the computation of each $D_{KL}(P,Q)$ distance becomes much more efficient; 2) It is much easier to visualize results. Moreover, it is possible to observe the topological properties of each trajectory as a function of the object pose (see fig. 3). Note that the actual degree of freedom is essentially regularized by the parameters θ, φ. We use MDS [4] to embed object pose PDFs into a Euclidean space. As demonstrated in [2] such embedding is able to preserve the KL divergence in the original embedding. We indicate this embedding by $f_e(*)$.

4.2. Manifold Alignment and Expansion

Given a part based representation as introduced in Sec. 3, our goal is to select *descriptive* parts so as to align the trajectories in the manifold (reduce intra-pose variability) as well as *discriminative* parts so as to expand the trajectories (increase inter-pose distance). Note that unlike general feature selection algorithms (such as Fisher's Discriminant Analysis), our optimization must be applied on the smooth manifold parameterized by pose, as opposed to over a discrete number of k classes.

Specifically, given a training set of object poses $O = O_i^j, i = 1, ..., M, j = 1, ..., N$, we extract a set of ST-parts F from them. We seek to select an informative part set $\hat{F}$ from F. The part selection is formulated as follows,

$$\underset{\hat{F} \subset F}{\operatorname{argmin}} \sum_{j=1}^{N} \sum_{(i_1,i_2)\in\mathcal{P}_1} ||x_{i_1}^j - x_{i_2}^j|| - \lambda \sum_{i=1}^{M} \sum_{(j_1,j_2)\in\mathcal{P}_2} ||x_i^{j_1} - x_i^{j_2}|| \quad (8)$$

where $x_i^j = f_e(P_i^j)$ is an embeded image of PDF P_i^j in the embedded space, $\lambda > 0$ is a factor that indicates the relative importance of these two terms, $\mathcal{P}_1$ contains all pairs of object instances for pose j ($i_1, i_2 = 1, ..., M$) except $i_1 = i_2$, and $\mathcal{P}_2$ contains all pairs of object poses of object instance i ($j_1, j_2 = 1, ..., N$) except $j_1 = j_2$. $\| \cdot \|$ is a l_2-norm in the $\mathbb{R}^d$ Euclidean space. Notice it approximates the KL

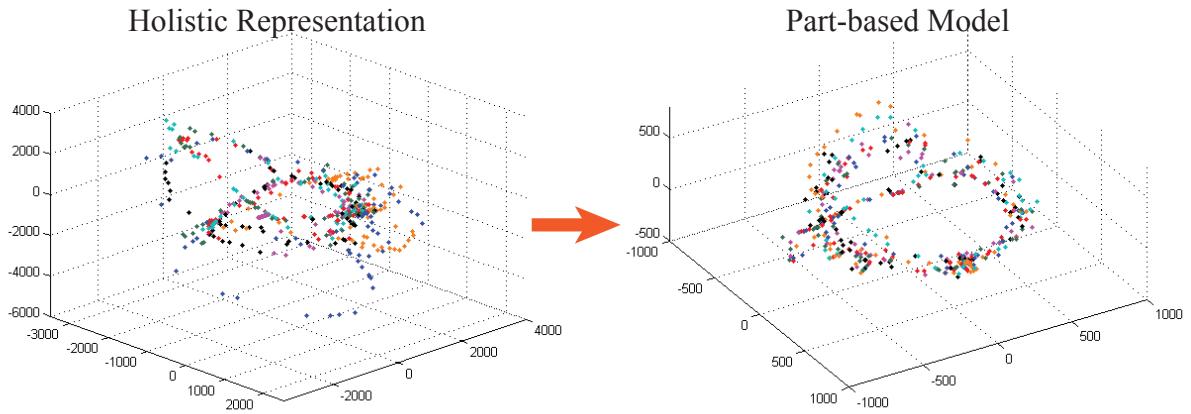

Figure 3. **Reconstructed manifolds of the 7-car dataset.** Each trajectory, which consists of a sequence points corresponding to object pose PDFs, is associated with one object instance (identified by a unique color). The trajectories demonstrates the smooth transition between neighbor poses as the camera moves(here, a video sequence, taken by moving a camera around one object instance, is uniformly split into short segments corresponding to different poses).**Left:** Manifold reconstructed using holistic pose model [21]. Notice that the trajectories are not aligned. **Right:** Manifold reconstructed based on part-based model with part selection. After part selection, the trajectories from different object instances are well aligned and expanded.

divergence in the original embedding. The left term minimizes the intra-category variability. Specifically, points corresponding to different object instances under similar poses will be close to each other. This term guarantees the generalization power of our method as it penalizes parts that do not generalize well across different object instances. The right term attempts to maximize the inter-pose distance (i.e., distance between points on the same trajectory). This property guarantees the discriminative power of our algorithm and prevents the manifold from collapsing into a trivial single cluster. It penalizes non-discriminative trivial parts.

The objective function (8) is a multi-objective combinatorial optimization problem. Following the lexicographic ordering formulation [15], we approximate the multi-objective optimization problem as a set of sub-problems, namely,

$$\underset{F' \subset F}{\operatorname{argmin}} \sum_{j=1}^{N} \sum_{(i_1,i_2)\in\mathcal{P}_1} ||x_{i_1}^j - x_{i_2}^j|| \tag{9}$$

$$\underset{F'' \subset F'}{\operatorname{argmin}} - \sum_{i=1}^{M} \sum_{(j_1,j_2)\in\mathcal{P}_2} ||x_i^{j_1} - x_i^{j_2}|| \tag{10}$$

In fact, we solve the problem of (8) by sequentially solving its the sub-problems (9) and (10).

4.3. Part Selection by Graph-Based Ranking

We further decompose the part selection problem in (9) into N independent local optimization subproblems corresponding to N poses. Specifically, assuming F_j to be the part set extracted from all M object instances under the same pose j, we aim to solve the following problem,

$$\underset{F_j' \subset F_j}{\operatorname{argmin}} \sum_{(i_1,i_2)\in\mathcal{P}_1}^{M} ||x_{i_1}^j - x_{i_2}^j|| \tag{11}$$

Consequently, we obtain $F' = \{F_j'\}_{j=1}^N$. Similarly, we decompose the problem of (10) into M local optimization subproblems corresponding to M instances as follows,

$$\underset{F_i'' \subset F_i}{\operatorname{argmin}} \sum_{(j_1,j_2)\in\mathcal{P}_2}^{N} -||x_i^{j_1} - x_i^{j_2}|| \tag{12}$$

where $F_i \subset F'$ contains all parts which are extracted from all N poses of object instance i and selected by (11) (i.e., in F'). Thus, the final selected parts $F'' = \{F_i''\}_{i=1}^M$, which is an approximate set of $\hat{F}$ in Eq. 8. We observe that for the two local combinatorial optimization problems (11) and (12), the search space is $2^{|F_i|}$ and $2^{|F_j|}$ respectively, which increases exponentially with the number of instance/pose labels. Therefore, searching for an exact optimal answer is still intractable when $|F_i|$ and $|F_j|$ become large.

The search space can be reduced by ranking parts according to their "descriptiveness" and "discrimination" power. Specifically, given a pose j, we want to select the most descriptive parts, which are shared across different object instances; while given a object instance i, we expect to select most discriminative parts (i.e., uncommon parts across poses), which can help distinguish poses. Both problems can be solved by the same ranking scheme described bellow. The idea is to seek to select "good" parts greedily by local parts ranking. This greedy strategy works well for our problem.

For a given pose j (object instance i), we build a Part Similarity Graph (PSG) $G_j = (V_j, E, W)$ ($G_i =$

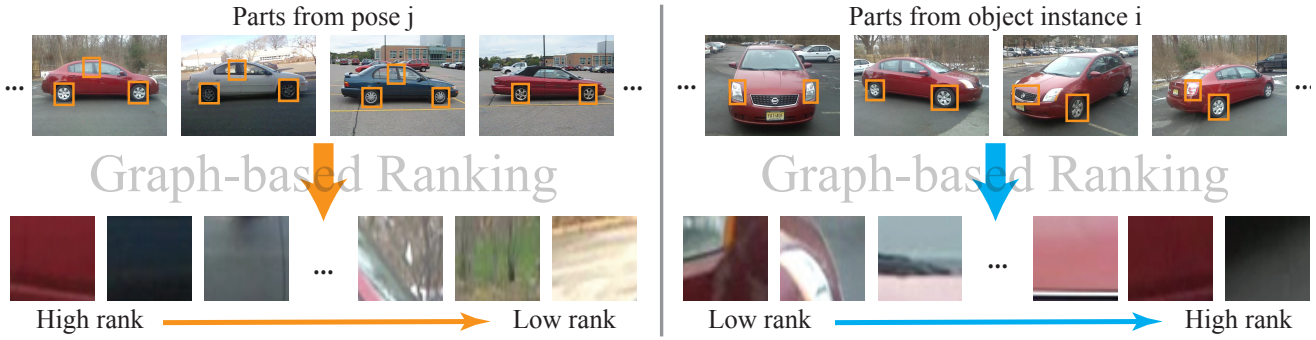

Figure 4. **Part Similarity Graph. Left:** PSG ranks all the parts from varied instances given a fixed pose. High-ranked nodes are shared across different instances and useful to minimize intra-class variability. **Right:** PSG ranks all the parts from varied poses given a certain object instance. Low-ranked nodes are less shared across different pose, so they can help maximize the inter-pose distance.

(V_i, E, W)), where V_j is a set of vertices corresponding to parts extracted from all object instances under pose j (V_i is the part set for all poses of instance i). E is an edge set and W is the associated weight matrix, where $w_{st} = exp(-\alpha D_{KL}(Q_s, Q_t))$ for parts Q_s and Q_t. Given such a graph G, we would like to measure the relative "importance" of the vertices (i.e., the centrality of vertex within a graph). Intuitively, an "important" vertex is the one that is more similar to other vertices (thus having higher centrality). For graph G_j, the importance of a vertex is consistent with the descriptiveness of its corresponding part. Hence, by ranking the importance of vertices we can select descriptive parts from V_j. On the other hand, for graph G_i, an "important" vertex is the one that is less discriminative since it is similar to more parts from different poses. The ascending ranking of importance also provides a way to select most discriminative parts from V_i.

To rank the importance of vertices within a graph, we apply Eigenvector Centrality, which is defined as the principal eigenvector of a stochastic adjacency matrix derived from weight matrix. Let W' be the column normalized matrix of W (which measures the similarity between parts), then the PartRank (PR) can be iteratively defined as, $W' * PR = PR$ (PR is a vector containing an importance value for each vertex). Since, it requires a strong connected graph (to ensure that the principal eigenvalue be 1). Thus, a dumping factor d ($d > 0.8$) is introduced. Thus, PR can be computed as, $W' * PR + (1-d)v = PR$ where $v = [1/n]_{n\times 1}$,n is the number of parts.

Fig.4 demonstrates the process of parts ranking for part selection. The left panel shows a selection of the most descriptive parts (which appear consistently across different object instances) for good generalization. By conducting this process for all poses $j = 1, \ldots, M$, we obtain a subset of parts F' that meet criteria in (11). The right panel illustrates the selection of discriminative parts to better distinguish poses of an object instance. We select less important parts that are highly discriminative for pose distinguishing. We repeat this process for all object instances and find F'', which is then used to estimate the pose in testing. We use PageRank technique [30] to greedily obtain the F'' in Eq. 12.

4.4. Recognizing novel object poses

After part selection, we obtain a descriptive and discriminative training part set $F' = \{Q_i\}_1^T$ with each $Q_i = \{(A_i, X_i), C_i\}$ where C_i is the pose label. Now, given a novel video segments O^t with unknown object pose, we first extract ST-parts $\{Q_s^t\}_1^S$ by deviding the video segments into a grid as explained in sec 3.2. Then the pose of O^t is estimated by K-Nearest Neighbor classifier as, $\hat{C} = argmax_C \sum_{s=1}^{S} \sum_{k=1}^{K} I_s^k$, where I_s^k is an indicator function which is equal to 1 if the pose label of the k-th nearest part in F' is C. Note that once we have the manifold, many machine learning techniques can be adopted to do the classification/estimation task. Here we focus on the 1-NN classifier, because although simple, 1NN's performance is directly related to the manifold structure, which makes it more interesting in our case.

5. Experiments

5.1. Synthetic Car Dataset

We first test our proposed algorithm on the synthesized car dataset available from [21]. This dataset contains 10 computer generated car instances with variation in shape and texture. Images with pose labels are available for different azimuth ($0^\circ - 360^\circ$) and zenith ($0^\circ - 40^\circ$) angles. We generate our object parts by dividing object into even sized blocks and regard each block as a part. As a baseline comparison, we test [21] with the best reported feature configuration (Edge map + SIFT). In both cases, we use the object bounding box provided by [21] to localize the object, and normalize our images into unit size to remove the

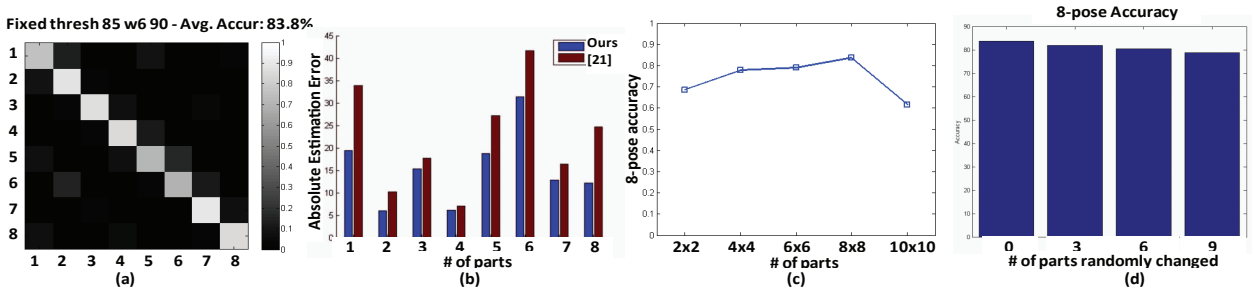

Figure 5. **(a)** 8-pose confusion table for the new car dataset. **(b)** Pose estimation error for different poses. **(c)** 8-pose accuracy for the new car dataset with different part configuration. **(d)** 8-pose accuracy w.r.t occlusion parts.

bias in the bounding box size. We use a 1 Nearest Neighbor classifier and consolidate the result as a 1 v.s. 32 pose (8 azimuth by 4 zenith poses) classification problem. By adopting our proposed part selection procedure, we are able to increase the overall accuracy of [21] (88.9%) by 10% and achieve nearly perfect classification (99.7%). See [22] for the confusion table and comparison of the manifolds.

5.2. Pose estimation on real dataset

We then test our algorithm on the real world car dataset introduced in [21]. This dataset contains video sequences capturing different object instances following a semi-sinusoidal trace on the viewing sphere. We also introduce one new PC mouse and stapler datasets, along with three new car instances with higher intra-class variability, all with pose annotations. To the best of our knowledge, this is the most complicated video dataset for categorial object continuous pose recognition from videos. We follow the scheme proposed in [21] to build the training/testing datasets by equally dividing the video sequences into short slices. We then use the leave one instance out cross validation scheme to estimate novel poses.

Our classification accuracy is shown in Fig. 6 and in Fig.5, where the reconstructed manifold is shown in Fig.3. As shown in Fig. 6, our algorithm achieves $5\%-20\%$ better performance and outperforms [21] consistently under different classification granularity. This improvement in accuracy is indeed related to the quality of the reconstructed manifold. As shown in Fig.3, the original manifold is able to capture trajectories indicating the pose transition foro single object instance; however, looked as a whole, manifolds from different object instances are quite scattered and orderless. The manifold constructed by our proposed method shows clearly the common structure shared with different object instances.

Fig.5(a) shows the confusion table for the new car dataset on 8-pose classification. As shown in the figure, most errors occur in neighboring poses and opposite poses (due to the intrinsic symmetry of car). Fig.5(b) compares our pose estimation error with [21], which demonstrates that our part-based model achieves less estimation error for all given poses. Fig.5(c) shows the performance of our model with different part decompositions. As shown in the figure, classification accuracy goes up as we divide the whole object into finer parts, and when the parts becomes too small, it looses descriptive power and the performance goes down again. Another common issue with object detection and pose estimation algorithms is the occlusion of object parts. We test this by replacing some parts inside the object bounding box with random background patches and show the accuracy in Fig.5(d). As shown in the figure, our algorithm is robust to part occlusions.

Comparison with FDA. To distinguish our problem from general feature selection problems and illustrate the power or our algorithm, we discretize our video segments into 8 poses and treat our problem as an 8-class classification problem. The *within-class scatter matrix* and *between-class scatter matrix* is calculated as

$$S_W = \sum_{\phi}((D^{\phi,\phi}_{p,m}))_{p,m}, S_B = \sum_{\phi}\sum_{\theta}((D^{\phi,\theta}_{p,m}))_{p,m},$$

. Features are selected by solving the generalized eigenvalue problem $\max_u \frac{u^T \Lambda u}{u^T \Sigma u}$. The best 8-pose classification accuracy for FDA is 55%, which is worse than both the Holistic model(77.4%) and our part-based model(83.7%). This indicates the challenges in our part-selection problem. The reason that FDA is not working well is: (1) FDA assumes that classes are discretized (and data are clustered in each class); whereas in our case the transition between neighboring poses is smooth; (2) FDA's learned optimal part configuration is fixed for all poses, while our algorithm is able to select different configurations according to different poses.

Comparison with Spatial Pyramid Matching. To compare with the state-of-the-art classification algorithms on single frames, we adopted the Spatial Pyramid Matching framework proposed in [28]. We generate a 4-level pyramid with 100 codewords (the best configuration in our case) for

(%)	8-pose		12-pose		16-pose		32-pose	
Dataset	[21]	Ours	[21]	Ours	[21]	Ours	[21]	Ours
Cars [6]	84.5	**90.4**	75.7	**82.3**	71.8	**78.0**	51.5	**62.6**
Cars [6] + new	77.4	**83.7**	68.9	**73.2**	65.1	**67.3**	43.9	**47.0**
Mouse new	61.8	**68.6**	56.4	**62.9**	46.1	**62.9**	31.8	**45.4**
Staple new	64.0	**82.6**	50.6	**72.1**	45.9	**69.8**	27.9	**51.8**

Figure 6. Pose classification accuracy. Comparison with [21] shows that our method outperforms [21] consistently under different classification granularity.

individual frames and perform leave one instance out testing on each frame. Histogram intersection kernel is used as distance measure. The 1-Nearest Neighbor classification result for 8-pose is 72.5%, worse than the Holistic model(77.4%) and our part-based model(83.7%). This justifies our motivation of using videos and information divergence as opposed to images.

6. Conclusion and Future Work

In this paper, we propose a novel framework for object pose estimation on a category level. We treat an object pose as a collection of spatial-temporal parts. By modeling each ST part as a probabilistic PDF, we further represent an object pose as a joint distribution of part PDFs. We then use statistical manifold to model all the object poses in the pose space by MDS embedding. By adjusting the structure of the manifolds, we demonstrate we can simultaneously maximize the inter-pose distance and minimize the intra-class variability. In order to adjust the structure of the manifolds on the training data set, we choose a graph-based technique to rank all the parts and select the informative ones, which can produce well aligned and expanded manifolds. Finally, the selected informative parts are used to recognize the pose of an unknown instance. Our method achieves the state-of-the-art results on a publicly available data set.

7. Acknowledgements

This research was partially supported by ARO grant W911NF-09-1-0310 and NSF grant #0931474.

References

[1] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *PAMI*, 24(8):1026–1038, 2002. 4

[2] K. Carter, R. Raich, W. Finn, and A. Hero. Fine: Fisher information nonparametric embedding. *PAMI*, pages 2093–2098. 4

[3] A. Collet, D. Berenson, S. Srinivasa, and D. Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *ICRA*, 2009. 1

[4] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, 2000. second edition. 2, 4

[5] S. Dickinson, A. Pentland, and A. Rosenfeld. 3-d shape recovery using distributed aspect matching. *IEEE Transactions on PAMI*, 14(2):174–198, 1992. 1

[6] P. Dollár, V. Rabaud, and S. Belongie. Non-isometric manifold learning: Analysis and an algorithm. In *ICML*, 2007. 2

[7] A. Farhadi, M. Kamali, I. Endres, and D. Forsyth. A latent model of discriminative aspect. In *ICCV*, 2009. 1

[8] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. 4

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:2005, 2003. 1

[10] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, 71(3):273–303, 2007. 1

[11] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation from single or multiple model views. *IJCV*, 67(2):159–188, 2006. 1

[12] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010. 1

[13] E. Hsiao, A. Collet, and M. Hebert. Making specific features less discriminative to improve point-based 3d object recognition. In *CVPR*, 2010. 1

[14] D. Huttenlocher and S. Ullman. Object recognition using alignment. In *ICCV*, 1987. 1

[15] K. D. J. Branke and K. Miettinen. *Multiobjective Optimization: Interactive and Evolutionary Approaches*. Spinger. 5

[16] D. Jacobs and R. Basri. 3-d to 2-d pose determination with regions. *IJCV*, 34:123C145, 1992. 1

[17] M. V. John A. Lee. *Nonlinear dimensionality reduction*. Springer, 2007. 2

[18] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, 2004. 1

[19] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3d feature maps. In *CVPR*, 2008. 1

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2:91–110, 2004. 1

[21] L. Mei, K. Carter, M. Sun, A. Hero, and S. Savarese. Unsupervised object pose classification from short video sequences. In *BMVC*, 2009. 1, 2, 4, 5, 6, 7, 8

[22] L. Mei, J. Liu, A. Hero, and S. Savarese. Robust object pose estimation via statistical manifold modeling. In *Technical Report, the University of Michigan*, 2011. 3, 4, 7

[23] M. Ozuysal, V. Lepetit, and P. Fua. Pose Estimation for Category Specific Multiview Object Localization. In *CVPR*, 2009. 1

[24] D. G. S. Richard O. Duda, Peter E. Hart. *Pattern classification*. Wiley, 2001. 2

[25] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *IJCV*, pages 231–259, 2006. 1

[26] C. Rothwell, A. Zisserman, D. Forsyth, and J. Mundy. Canonical frames for planar object recognition. In *ECCV*, 1992. 1

[27] S. Roweis and L. Saul. Non-linear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323C2326, 2000. 2

[28] C. S. S Lazebnik. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2, 7

[29] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. 2007. 1

[30] B. Sergey and L. Page. The anatomy of a large-scale hyper-textual web search engine. In *WWW*, 1998. 6

[31] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *ICCV*, 2009. 1

[32] M. Sun, H. Su, S. Savarese, and L. Fei-Fei. A multi-view probabilistic model for 3d object classes. In *CVPR*, 2009. 1

[33] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319C2323, 2000. 2