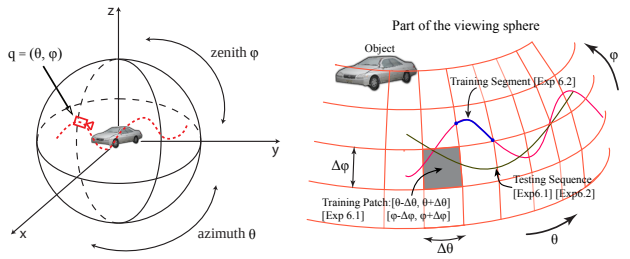
Unsupervised Object Pose Classification from Short Video Sequences

Liang Mei
alexmei@umich.edu
Min Sun
sunmin@umich.edu
Kevin M.Carter
kmcarter@umich.edu
Alfred O. Hero III
hero@umich.edu
Silvio Savarese
silvio@eecs.umich.edu

Department of EECS,
University of Michigan
Ann Arbor, USA

We address the problem of recognizing the pose of an object category from video sequences capturing the object under small camera movements. This scenario is relevant in applications such as robotic object manipulation or autonomous navigation. We introduce a new algorithm where we model an object category as a collection of non parametric probability densities capturing appearance and geometrical variability within a small area of the viewing sphere for different object instances. By regarding the set of frames of the video as realizations of such probability densities, we cast the problem of object pose classification as the one of matching probably density functions in testing and training. Our experimental results on both synthesized and real world data show promising results toward the goal of accurate and efficient pose classification of object categories from video sequences.



(a) (b)

Figure 1: 1(a) Pose Estimation as a pair of azimuth and zenith angles $q=(\theta,\varphi)$ on the viewing sphere. The object is assumed to lie at the center of the viewing sphere. 1(b) Trajectories showing the camera movement.

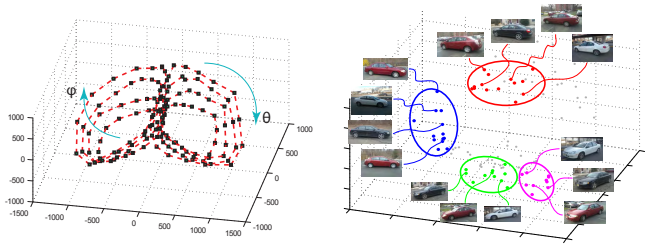
Our work starts by observing that a video sequence (portraying an object as the camera position and view point changes) can be used to parameterize a trajectory of positions on the viewing sphere, where each position corresponds to the azimuth and zenith angle coordinates describing the pose of the object (Fig.1(a)). Our key idea is to decompose the video sequence into pockets of frames (video segments). Thus, each video segment can be associated to a location on the viewing sphere that captures the average pose within the video segments. By regarding images as low-dimensional (non-linear) manifolds embedded in the high-dimensional image space, manifold learning is designed to analyze the low-dimensional structure which underlies a collection of high-dimensional data. Recent studies in statistical manifold learning [1] define information divergence as a metric of distance between probability densities and apply common dimensionality reduction techniques for visualization.

By viewing images as realizations of probability distributions, we are able to formulate our problem of object pose classification within a statistical manifold learning framework. Specifically, assuming the object lies at the center of the viewing sphere, our observation X is generated according to

$$P(X|\mathscr{C},T,\rho,\theta,\varphi), \tag{1}$$

where $\mathscr{C}$ is the object category; T is the texture, which captures the appearance of an object instance; ρ is the distance between the object and the camera, which affects the object scale; θ and φ are azimuth and zenith angles representing the viewpoint, respectively. By assuming all the objects belong to the same category, and ρ is fixed (small scale variations can be accommodated by normalizing the object bounding box to unit length, the probability density function (1) can be rewritten as

$$P(X|\mathscr{C},T,\rho,\theta,\varphi)=P(X|T,\theta,\varphi) \tag{2}$$



(a) (b)

Figure 2: 2(a) Embedding of estimated PDFs from a single instance of the synthesized car data. Each point in the figure corresponds to a PDF, which is estimated from images taken from a $10^\circ\times10^\circ$ small patch on the viewing sphere (Refer to Fig.1(b)). Trajectories in the manifold show the two main parameterizations of the learned PDFs, which corresponds to two intrinsic degrees of freedom (θ,φ) in the data. 2(b) The manifold can be naturally used to discover clusters for unsupervised pose estimation.

Suppose we are given a video sequence $\mathscr{V}^i$ capturing the object instance i as view point $q=(\theta,\varphi)$ varies on the viewing sphere. We divide the video into segments of length K, and regard frames in segment j ($j=1,2,\ldots,\lceil N^i/K\rceil$, where N^i is the number of frames in $\mathscr{V}^i$) as generated according to

$$P^i_j=P(X|T=t^i,\theta\in[\theta_j-\Delta\theta_j,\theta_j+\Delta\theta_j],\varphi\in[\varphi_j-\Delta\varphi_j,\varphi_j+\Delta\varphi_j]) \tag{3}$$

where $[\theta_j-\Delta\theta_j,\theta_j+\Delta\theta_j],[\varphi_j-\Delta\varphi_j,\varphi_j+\Delta\varphi_j]$ defines the angular support of the segment j on the viewing sphere (Fig.1(b)).

The PDFs (3) are estimated through kernel density estimation. Then the KL-divergence between all possible pairs of PDFs are calculated. We use classical multidimensional scaling (cMDS) to reduce dimensionality and reconstruct the statistical manifold. This gives rise to a manifold which consists $\sum_i N^i/K$ points, where each point corresponds to a probability density (3). Fig.2(a) shows an example of the embedded PDFs from the synthesized car dataset in a 3D space. Each point in the figure is estimated from images taken from a $10^\circ\times10^\circ$ small patch on the viewing sphere (E.g. See Fig.1(b)). Trajectories in the manifold in Fig.2(a) show the two main parameterizations of the learned probability models, which corresponds to two intrinsic degrees of freedom (θ,φ) in the data.

We demonstrate the recognition accuracy of the proposed algorithm on both synthesized and real datasets. Supervised classification results show that our method achieve an overall accuracy of 86.4% on a real car dataset and 85.4% on a real PC mouse dataset. Comparison with state-of-the-art spatial pyramid matching framework [2] shows that our algorithm outperforms the spatial pyramid matching consistently, with a notable $10\%-20\%$ lead when the detected location of the object is corrupted by noise. We also test our unsupervised learning algorithm and obtain an accuracy of 72.1% and 57.7% for these two datasets respectively.

[1] K. Carter, R. Raich, W. Finn, and A. Hero. Fine: Fisher information nonparametric embedding. *IEEE Transactions on Pattern Recognition and Machine Learning*, 2009.

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