2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

A Hierarchical Approach for Joint Multi-view Object Pose Estimation and Categorization

Mete Ozay[1], Krzysztof Walas[1,2] and Aleš Leonardis[1]

Abstract— **We propose a joint object pose estimation and categorization approach which extracts information about object poses and categories from the object parts and compositions constructed at different layers of a hierarchical object representation algorithm, namely Learned Hierarchy of Parts (LHOP) [7]. In the proposed approach, we first employ the LHOP to learn hierarchical part libraries which represent entity parts and compositions across different object categories and views. Then, we extract statistical and geometric features from the part realizations of the objects in the images in order to represent the information about object pose and category at each different layer of the hierarchy. Unlike the traditional approaches which consider specific layers of the hierarchies in order to extract information to perform specific tasks, we combine the information extracted at different layers to solve a joint object pose estimation and categorization problem using distributed optimization algorithms. We examine the proposed generative-discriminative learning approach and the algorithms on two benchmark 2-D multi-view image datasets. The proposed approach and the algorithms outperform state-of-the-art classification, regression and feature extraction algorithms. In addition, the experimental results shed light on the relationship between object categorization, pose estimation and the part realizations observed at different layers of the hierarchy.**

I. Introduction

The field of service robots aims to provide robots with functionalities which allow them to work in man-made environments. For instance, the robots should be able to categorize objects and estimate the pose of the objects to accomplish various robotics tasks, such as grasping objects [14]. Representation of object categories enables the robot to further refine the grasping strategy by giving context to the search for the pose of the object [15].

In this paper, we propose a joint object categorization and pose estimation approach which extract information about statistical and geometric properties of object poses and categories extracted from the object parts and compositions that are constructed at different layers of the Learned Hierarchy of Parts (LHOP) [7], [8], [9].

In the proposed approach, we first employ LHOP [7], [8] to learn hierarchical part libraries which represent object parts and compositions across different object categories and views as shown in Fig. 1. Then, we extract statistical and geometric features from the part realizations of the objects in the images in order to represent the information about the object pose and category at each different layer of the hierarchy. We propose two novel feature extraction algorithms, namely Histogram of Oriented Parts (HOP) and Entropy of Part Graphs. HOP features measure local distributions of global orientations of part realizations of objects at different layers of a hierarchy. On the other hand, Entropy of Part Graphs provides information about the statistical and geometric structure of object representations by measuring the entropy of the relative orientations of parts. In addition, we compute a Histogram of Oriented Gradients (HOG) [5] of part realizations in order to obtain information about the co-occurrence of the gradients of part orientations.

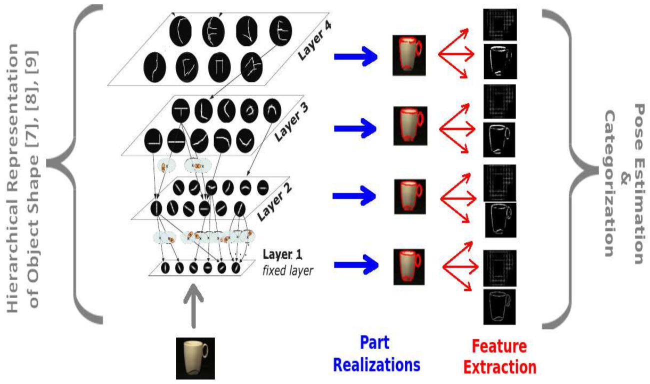
Hierarchical Representation of Object Shape [7], [8], [9]
Layer 4
Layer 3
Layer 2
Layer 1
fixed layer
Part Realizations
Feature Extraction
Pose Estimation & Categorization

Fig. 1: Combination of features extracted from part realizations detected at different layers of LHOP.

Unlike traditional approaches which extract information from the object representations at specific layers of the hierarchy to accomplish specific tasks, we combine the information extracted at different layers to solve a joint object pose estimation and categorization problem using a distributed optimization algorithm. For this purpose, we first formulate the joint object pose estimation and categorization problem as a sparse optimization problem called Group Lasso [19]. We consider the pose estimation problem as a sparse regression problem and the object categorization problem as a multi-class logistic regression problem using Group Lasso. Then, we solve the optimization problems using a distributed and parallel optimization algorithm called the Alternating Direction Method of Multipliers (ADMM) [1].

In this work, we extract information on object poses and categories from 2-D images to handle the cases where 3-

This work was supported in part by the European Commission project PaCMan EU FP7-ICT, 600918.

[1]Mete Ozay, Krzysztof Walas and Aleš Leonardis are with School of Computer Science, University of Birmingham, Edgbaston B15 2TT Birmingham, United Kingdom {m.ozay,walask,a.Leonardis}@cs.bham.ac.uk

[2]Krzysztof Walas is also with Department of Electrical Engineering, Poznan University of Technology, ul. Piotrowo 3a, 60-965 Poznan, Poland krzysztof.walas@put.poznan.pl

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

D sensing may not be available or may be unreliable (e.g. glass, metal objects). We examine the proposed approach and the algorithms on two benchmark 2-D multiple-view image datasets. The proposed approach and the algorithms outperform state-of-the-art Support Vector Machine and Regression algorithms. In addition, the experimental results shed light on the relationship between object categorization, pose estimation and the part realizations observed at different layers of the hierarchy.

In the next section, related work is reviewed and the novelty of our proposed approach is summarized. In Section II, a brief presentation of the hierarchical compositional representation is given. Feature extraction algorithms are introduced in Section III. The joint object pose estimation and categorization problem is defined, and two algorithms are proposed to solve the optimization problem in Section IV. Experimental analyses are given in Section V. Section VI concludes the paper.

A. Related Work and Contribution

In the field of computer vision the problem of object categorization and pose estimation is studied thoroughly and some of the approaches are proliferating to the robotics community. With an advent of devices based on PrimeSense sensors, uni-modal 3-D or multi-modal integration of 2-D and 3-D data (e.g. rgb-d data) have been widely used by robotics researchers [13]. However, 3-D sensing may not be available or reliable due to limitations of object structures, lighting resources and imaging conditions in many cases where single or multiple view 2-D images are used for categorization and pose estimation [3], [4], [20]. In [20], a probabilistic approach is proposed to estimate the pose of a known object using a single image. Collet et al. [3] build 3D models of objects using SIFT features extracted from 2D images for robotic manipulation, and combine single image and multiple image object recognition and pose estimation algorithms in a framework in [4].

A promising approach to the object categorization and the scene description is the use of hierarchical compositional architectures [7], [9], [15]. Compositional hierarchical models are constructed for object categorization and detection using single images in [7], [9]. Multiple view images are used for pose estimation and categorization using a hierarchical architecture in [15]. In the aforementioned approaches, the tasks are performed using either discriminative or generative top-down or bottom-up learning approaches in architectures. For instance, Lai et al. employ a top-down categorization and pose estimation approach in [15], where a different task is performed at each different layer of the hierarchy. Note that, a categorization error occurring at the top-layer of the hierarchy may propagate to the lower layer and affect the performance of other tasks such as pose estimation in this approach. In our proposed approach, we first construct generative representations of object shapes using LHOP [7], [8], [9]. Then, we train discriminative models by extracting features from the object representations. In addition, we propose a new method, which enables us to combine the information extracted at each different layer of the hierarchy, for joint categorization and pose estimation of objects. We avoid the propagation of errors of performing multiple tasks through the layers and enable the shareability of parts among layers by the employment of optimization algorithms in each layer in a parallel and distributed learning framework.

The novelty of the proposed approach and the paper can be summarized as follows;

1) In this work, the Learned Hierarchy of Parts (LHOP) is employed in order to learn a hierarchy of parts using the shareability of parts across different views as well as different categories [7], [8].
2) Two novel feature extraction algorithms, namely Histogram of Oriented Parts (HOP) and Entropy of Part Graphs, are proposed in order to obtain information about the statistical and geometric structure of objects' shapes represented at different layers of the hierarchy using part realizations.
3) The proposed generative-discriminative approach enables us to combine the information extracted at different layers in order to solve a joint object pose estimation and categorization problem using a distributed and parallel optimization algorithm. Therefore, this approach also enables us to share the parts among different layers and avoid the propagation of object categorization and pose estimation errors through the layers.

II. Learned Hierarchy of Parts

In this section, Learned Hierarchy of Parts (LHOP)[7], [8] is briefly described. In LHOP, the object recognition process is performed in a hierarchy starting from a feature layer through more complex and abstract interpretations of object shapes to an object layer. A learned vocabulary is a recursive compositional representation of shape parts. Unsupervised bottom-up statistical learning is encompassed in order to obtain such a description.

Shape representations are build upon a set of compositional parts which at the lowest layer use atomic features, e.g. Gabor features, extracted from image data. The object node is a composition of several child nodes located at one layer lower in the hierarchy, and the composition rule is recursively applied to each of its child nodes to the lowest layer Γ_1. All layers together form a hierarchically encoded vocabulary $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \ldots \cup \Gamma_L$. The entire vocabulary Γ is learned from the training set of images together with the vocabulary parameters [8].

The parts in the hierarchy are defined recursively in the following way. Each part in the l^{th} layer represents the spatial relations between its constituent subparts from the layer below. Each composite part $\mathcal{P}_k^l$ constructed at the l^{th} layer is characterized by a central subpart $\mathcal{P}_{central}^{l-1}$ and a list of remaining subparts with their positions relative to the center as

$$\mathcal{P}_k^l = (\mathcal{P}_{central}^{l-1}, \{(\mathcal{P}_j^{l-1}, \boldsymbol{\mu}_j, \Sigma_j)\}j), \qquad (1)$$

where $\boldsymbol{\mu}_j = (x_j, y_j)$ denotes the relative position of the subpart $\mathcal{P}_j^{l-1}$, while Σ_j denotes the allowed variance of its position around (x_j, y_j).

III. Feature Extraction from Learned Parts

LHOP provides information about different properties of objects, such as poses, orientations and category memberships, at different layers [7]. For instance, the information on shape parts, which are represented by edge structures and textural patterns observed in images, is obtained using Gabor features at the first layer L_1. In the second and the following layers, compositions of parts are constructed according to the co-occurrence of part realizations that are detected in the images among different views of the objects and across different object categories. In other words, a library of object parts and compositions is learned jointly for all object views and categories.

In order to obtain information about statistical and geometric properties of parts, we extract three types of features from the part realizations detected at each different layer of the LHOP.

A. Histogram of Orientations of Parts

Histograms of orientations of parts are computed in order to extract information on the co-occurrence of orientations of the parts across different poses of objects. Part orientations are computed according to a coordinate system of an image I whose origin is located at the center of the image I, and the axes of the coordinate system are shown with blue lines in Figure 2.

If we define $p_k^l, \forall k = 1, 2, \dots, K, \forall l = 1,, 2 \dots, L$ as the realization of the k^{th} detected part in the l^{th} layer at an image coordinate (x_k, y_k) of I, then its orientation with respect to the origin of the coordinate system is computed as

$$\theta_{k,l} = \arctan(\frac{y_k}{x_k}).$$

Then, the image I is partitioned into M cells $\{I_m\}_{m=1}^{M}$, and histograms of the part orientations $\{\theta_{k,l}\}_{k=1}^{K'}$ of the part realizations $\{p_{k,l}\}_{k=1}^{K'}$ that are located in each cell I_m are computed. The aggregated histogram values are considered as variables of a D_p dimensional feature vector $\mathbf{f}_{hop}^l \in \mathbb{R}^{D_p}$.

B. Histogram of Oriented Gradients of Parts

In addition to the computation of histograms of orientations of part realizations $p_k^l, \forall k = 1, 2, \dots, K, \forall l = 1, 2, \dots, L$, we compute histogram of oriented gradients (HOG) [5] of p_k^l in order to extract information about the distribution of gradient orientations of $p_k^l, \forall k, l$. We denote the HOG feature vector extracted using $\{p_k^l\}_{k=1}^{K}$ in the l^{th} layer as $\mathbf{f}_{hog}^l \in \mathbb{R}^{D_h}$, where D_h is the dimension of the HOG feature vector. The details of the implementation of HOG feature vectors are given in Section V.

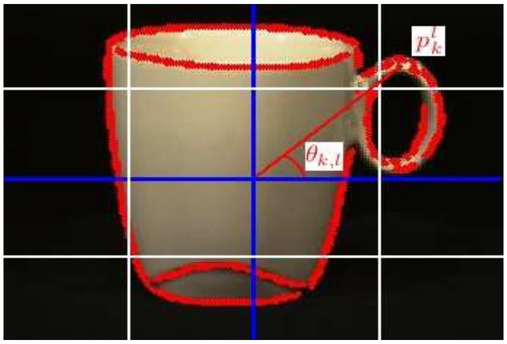

Fig. 2: An image is partitioned into cells for the computation of histograms of orientations of parts. A part realization p_k^l is depicted with a red point and associated to a part orientation degree $\theta_{k,l}$.

C. The Entropy of Part Graphs

We measure the statistical and structural properties of relative orientations of part realizations by measuring the complexity of a graph of parts. Mathematically speaking, we define a weighted undirected graph $G_l := (E_l, V_l)$ in the l^{th} layer, where $V_l := \{p_k^l\}$ is the set of part realizations, $E_l := \{e_{k',k}\}_{k',k=1}^{K}$ is the set of edges, where each edge $e_{k',k}$ that connects the part realizations $p_{k'}^l$ and p_k^l is associated to an edge weight $w_{k',k}$, which is defined as

$$w_{k',k} := \arccos(\frac{\mathbf{pos}_{k'} \cdot \mathbf{pos}_k}{\|\mathbf{pos}_{k'}\|_2 \|\mathbf{pos}_k\|_2}),$$

where $\mathbf{pos}_k := (x_k, y_k)$ is the position vector of $p_{k'}^l$, $\| \cdot \|_2$ is the ℓ_2 norm or Euclidean norm, and $\mathbf{pos}_{k'} \cdot \mathbf{pos}_k$ is the inner product of $\mathbf{pos}_{k'}$ and $\mathbf{pos}_k$. In other words, the edge weights are computed according to the orientations of parts relative to each other.

We measure the complexity of the weighted graph by computing its graph entropy. First, we compute the normalized weighted graph Laplacian $\mathcal{L}$ [6], [16] as

$$\mathcal{L} = \frac{1}{K(K-1)}(\mathcal{D} - \mathcal{W}),$$

where $\mathcal{W} \in \mathbb{R}^{K \times K}$ is a weighted adjacency matrix or a matrix of weights $w_{k',k}$, and $\mathcal{D} \in \mathbb{R}^{K \times K}$ is a diagonal matrix with members $\mathcal{D}_{k,k} := \sum_{k'=1}^{K} w_{k',k}$. Then, we compute the von Neumann entropy of G_l [6], [16] as

$$\begin{aligned} S(G_l) &= -\operatorname{Tr}(\mathcal{L} \log_2 \mathcal{L}) && (2) \\ &= -\sum_{k=1}^{K} \nu_k, && (3) \end{aligned}$$

where $\nu_1 \geq \nu_2 \geq \dots \geq \nu_k \geq \dots \geq \nu_K = 0$ are the eigenvalues of $\mathcal{L}$, $\operatorname{Tr}(\mathcal{L} \log_2 \mathcal{L})$ is the trace of the matrix product $\mathcal{L} \log_2 \mathcal{L}$ and $0 \log_2 0 = 0$. We use $S(G_l)$ as a feature variable $f_{ent}^l := S(G_l)$.

IV. Combination of Information Obtained at Different Layers of LHOP for Joint Object Pose Estimation and Categorization

In hierarchical compositional architectures, a different object property, such as object shape, pose and category, is represented at a different layer of a hierarchy in a vocabulary [15]. According the structures of the abstract representations of the properties, i.e. vocabularies, recognition processes have been performed using either a bottom-up [7], [8] or top-down [15] approach. It's worth noting that the information in the representations are distributed among the layers in the vocabularies. In other words, the information about the category of an object may reside at the lower layers of the hierarchy instead of the top layer. In addition, lower layer atomic features, e.g. oriented Gabor features, provide information about part orientations which can be used for the estimation of pose and view-points of objects at the higher layers. Moreover, the relationship between the pose and category of an object is bi-directional. Therefore, an information integration approach should be considered in order to avoid the propagation of errors that occur in multi-task learning and recognition problems such as joint object categorization and pose estimation, especially when only one of the bottom-up and top-down approaches is implemented.

For this purpose, we propose a generative-discriminative learning approach in order to combine the information obtained at each different layer of LHOP using the features extracted from part realizations. We represent the features defining a $D_p + D_h + 1$ dimensional feature vector $\mathbf{f}^l = (\mathbf{f}^l_{hop}, \mathbf{f}^l_{hog}, f^l_{ent})$. The feature vector $\mathbf{f}^l$ is computed for each training and test image, therefore we denote the feature vector of the i^{th} image I_i as $\mathbf{f}^l_i$, $\forall i = 1, 2, \dots, N$, in the rest of the paper.

We combine the feature vectors extracted at each l^{th} layer for object pose estimation and categorization under the following Group Lasso optimization problem [19]

$$\text{minimize} \quad \|\mathcal{F}\boldsymbol{\omega} - \mathbf{z}\|_2^2 + \lambda \sum_{l=1}^{L} \|\boldsymbol{\omega}_l\|_2, \tag{4}$$

where $\|\cdot\|_2^2$ is the squared ℓ_2 norm, $\lambda \in \mathbb{R}$ is a regularization parameter, $\boldsymbol{\omega}_l$ is the weight vector computed at the l^{th} layer, $\mathcal{F} \in \mathbb{R}^{N \times L}$ is a matrix of feature vectors $\mathbf{f}^l_i$, $\forall i = 1, 2, \dots, N$, $\forall l = 1, 2, \dots, L$ and $\mathbf{z} = (z_1, z_2, \dots, z_N)$ is a vector of target variables $z_i \in \mathbb{R}$, $\forall i = 1, 2, \dots, N$. More specifically, $z_i \in \Omega$ where Ω is a set of object poses, i.e. object orientation degrees, in a pose estimation problem.

We solve (4) using a distributed optimization algorithm called Alternating Direction Method of Multipliers [1]. For this purpose, we first re-write (4) in the ADMM form as follows

$$\begin{aligned} &\text{minimize} && \|\mathcal{F}\boldsymbol{\phi} - \mathbf{z}\|_2^2 + \lambda \sum_{l=1}^{L} \|\boldsymbol{\omega}_l\|_2 \\ &\text{subject to} && \boldsymbol{\omega}_l - \hat{\boldsymbol{\phi}}_l = \mathbf{0} \ , l = 1, 2, \dots, L \ , \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\phi}}_l$ is the local estimate of the global variable $\boldsymbol{\phi}$ for $\boldsymbol{\omega}_l$ at the l^{th} layer. Then, we solve (5) in the following three steps [1], [18],

1) At each layer l, we compute $\boldsymbol{\omega}_l^{t+1}$ as

$$\boldsymbol{\omega}_l^{t+1} := \underset{\boldsymbol{\omega}_l}{\operatorname{argmin}} \Big(\rho \|\boldsymbol{\mu}_l^t\|_2^2 + \lambda \|\boldsymbol{\omega}_l\|_2 \Big), \tag{6}$$

where $\boldsymbol{\mu}_l^t = \mathcal{F}_l(\boldsymbol{\omega}_l - \boldsymbol{\omega}_l^t) - \bar{\boldsymbol{\phi}}^t + \mathbf{a}^t + \overline{\mathcal{F}_l\boldsymbol{\omega}_l}^t$, $\rho > 0$ is a penalty parameter, $\overline{\mathcal{F}_l\boldsymbol{\omega}_l}^t = \frac{1}{L}\sum_{l=1}^{L} \mathcal{F}_l\boldsymbol{\omega}_l^t$, $\bar{\boldsymbol{\phi}}^t$ is the average of $\boldsymbol{\phi}_l^t$, $\forall l = 1, \dots, L$, and $\mathbf{a}^t$ is a vector of scaled dual optimization variables computed at an iteration t.

2) Then we update $\hat{\boldsymbol{\phi}}_l$ as

$$\hat{\boldsymbol{\phi}}_l^{t+1} := \frac{1}{L + \rho} \Big(\mathbf{z} + \rho \overline{\mathcal{F}_l\boldsymbol{\omega}_l}^{t+1} + \rho \mathbf{a}^t \Big). \tag{7}$$

3) Finally, $\mathbf{a}$ is updated as

$$\mathbf{a}^{t+1} := \mathbf{a}^t + \overline{\mathcal{F}_l\boldsymbol{\omega}_l}^t - \hat{\boldsymbol{\phi}}_l^{t+1}. \tag{8}$$

These three steps are iterated until a halting criterion, such as $t \geq T$ for a given termination time T, is achieved. Implementation details are given in the next section.

In a C class object categorization problem, $z_i \in \{1, 2, \dots, c, \dots, C\}$ is a category variable. In order to solve this problem, we employ *1-of-C* coding for sparse logistic regression as

$$P(z_i^c = 1 | \mathbf{f}_i) \quad = \quad \frac{\exp(h_j(\mathbf{f}_i))}{1 + \exp(h_c(\mathbf{f}_i))}, \tag{9}$$

where $h_c(\mathbf{f}_i) = \mathbf{f}_i \cdot \boldsymbol{\omega^c}$, $\boldsymbol{\omega^c}$ is a weight vector associated to the c^{th} category, $z_i^c = 1$ if $z_i = c$, $\forall i = 1, 2, \dots, N$. Then, we define the following optimization problem

$$\text{minimize} \quad - \sum_{l=1}^{L} \sum_{i=1}^{N} loss_l(i) + \lambda \|\boldsymbol{\omega^c}\|_1, \tag{10}$$

where $loss_l(i) = z_i^c h_c(\mathbf{f}_i) - \log \Big(\exp(h_c(\mathbf{f}_i)) + 1 \Big)$. In order to solve (10), we employ the three update steps given above with two modifications. First, we solve (6) for the ℓ_1 norm in the last regularization term $\lambda\|\boldsymbol{\omega}_l\|_1$ instead of the ℓ_2 norm. Second, we employ the logistic regression loss function in the computation of $\hat{\boldsymbol{\phi}}_l$ as

$$\hat{\boldsymbol{\phi}}_l^{t+1} := \underset{\boldsymbol{\phi}_l}{\operatorname{argmin}} \Big(\rho \|\boldsymbol{\phi_l} - \overline{\mathcal{F}_l\boldsymbol{\omega}_l}^{t+1} - \mathbf{a}^t\|_2 + \log(1 + \exp - (L\boldsymbol{\phi}_l)) \Big). \tag{11}$$

In the training phase of the pose estimation algorithm, we compute the solution vector $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \dots, \boldsymbol{\omega}_L\}$ using training data. In the test phase, we employ the solution vector $\boldsymbol{\omega}$ on a given test feature vector $\mathbf{f}_i$ of the part realizations of an object to estimate its pose as

$$\hat{z}_i = \mathbf{f}_i \cdot \boldsymbol{\omega}.$$

In the categorization problem, we predict the category label $\hat{z}_i$ of an object in the i^{th} image as

$$\hat{z}_i = \underset{c}{\operatorname{argmax}} \ \hat{z}_i^c.$$

V. Experiments

We examine our proposed approach and algorithms on two benchmark object categorization and pose estimation datasets, which are namely the Amsterdam Library of Object Images (ALOI) [10] and the Columbia Object Image Library (COIL-100) [17]. We have chosen these two benchmark datasets for two main reasons. First, images of objects are captured by rotating the objects on a turntable by regular orientation degrees which enable us to analyze our proposed algorithm for multi-view object pose estimation and categorization in uncluttered scenes. Second, object poses and categories are labeled within *acceptable* precision which is important to satisfy the statistical stability of training and test samples and their target values. In our experiments, we also re-calibrated labels of pose and rotation values of the objects that are mis-recorded in the datasets.

We select the bin size ($bSize$) of the histograms and cell size M of HOP (see Section III-A) and HOG features (see Section III-B) by greedy search on the parameter set $\{8, 16, 32, 64\}$, and take the *optimal* $b\hat{Size}$ and $\hat{M}$ which minimizes pose estimation and categorization errors in pose estimation and categorization problems using training datasets, respectively. In the employment of optimization algorithms, we compute $\lambda = \alpha\lambda_{\max}$, where $\lambda_{\max} = \|\mathcal{F}\boldsymbol{\omega}\|_{\infty}$, $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_L)$, $\|\cdot\|_{\infty}$ is ℓ_{∞} norm and α parameter is selected from the set $\{10^{-6}, 10^{-5}, \ldots, 10^{1}\}$ using greedy search by minimizing training error of object pose estimation and categorization as suggested in [1]. In the implementation of LHOP, we learn the compositional hierarchy of parts and compute the part realizations for $L = 1, 2, 3, 4$ [7].

In the experiments, pose estimation and categorization performances of the proposed algorithms are compared with state-of-the-art Support Vector Regression (SVR), Support Vector Machines (SVM) [2], Lasso and Logistic regression algorithms [12] which use the state-of-the-art HOG features [5] extracted from the images as considered in [11]. In the results, we refer to an implementation of SVM with HOG features as SVM-HOG, SVM with the proposed LHOP features as SVM-LHOP, SVR with HOG features as SVR-HOG, SVR with the proposed LHOP features as SVR-LHOP, Lasso with HOG features as L-HOG, Logistic Regression with HOG features as LR-HOG, Lasso with LHOP features as L-LHOP, Logistic Regression with LHOP features as LR-LHOP.

We use RBF kernels in SVR and SVM. The kernel width parameter σ is searched in the interval $\log(\sigma) \in [-10, 5]$ and the SVR cost penalization parameter ϵ is searched in the interval $\log(\epsilon) \in [-10, 5]$ using the training datasets.

A. Experiments on Object Pose Estimation

We have conducted two types of experiments for object pose estimation, namely *Object-wise* and *Category-wise* Pose Estimation. We analyze the sharability of the parts across different views of an object in Object-wise Pose Estimation experiments. In Category-wise Pose Estimation experiments, we analyze incorporation of category information to sharability of parts in the LHOP and to pose estimation performance.

1) Experiments on Object-wise Pose Estimation: In the first set of experiments, we consider the objects belonging to each different category, individually. For instance, we select $\aleph^{o}_{tr} = 4$ objects for training and $\aleph^{o}_{te} = 1$ objects for testing using objects belonging to **cups** category. The ID numbers of the objects and their category names are given in Table I. For each object, we have 72 object instances each of which represents an orientation of the object $z_i = \Theta_i$ on a turntable rotated with $\Theta_i \in \Omega$ and $\Omega = \{0°, 5°, 10°, \ldots, 355°\}$.

In the experiments, we first analyze the variation of part realizations and feature vectors across different orientations of an object. We visualize the features $\mathbf{f}^{l}_{hop}$, $\mathbf{f}^{l}_{hog}$ and f^{l}_{ent} in Figure 3 for a cup which is oriented with $\Theta \in \{20°, 60°, 120°, 180°, 240°, 280°, 340°\}$ and for each $l = 1, 2, 3, 4$. In the first row at the top of the figure, the change of f^{l}_{ent} is visualized $\forall l$. In the second row, the original images of the objects are given. In the third to the sixth rows, $\mathbf{f}^{l}_{hop}$ are visualized by displaying the part realizations with pixel intensity values $\|\mathbf{f}^{l}_{hop}\|_2^2$ for each $l = 1, 2, 3, 4$. $\mathbf{f}^{l}_{hog}$ features are visualized in the rest of the rows for each l.

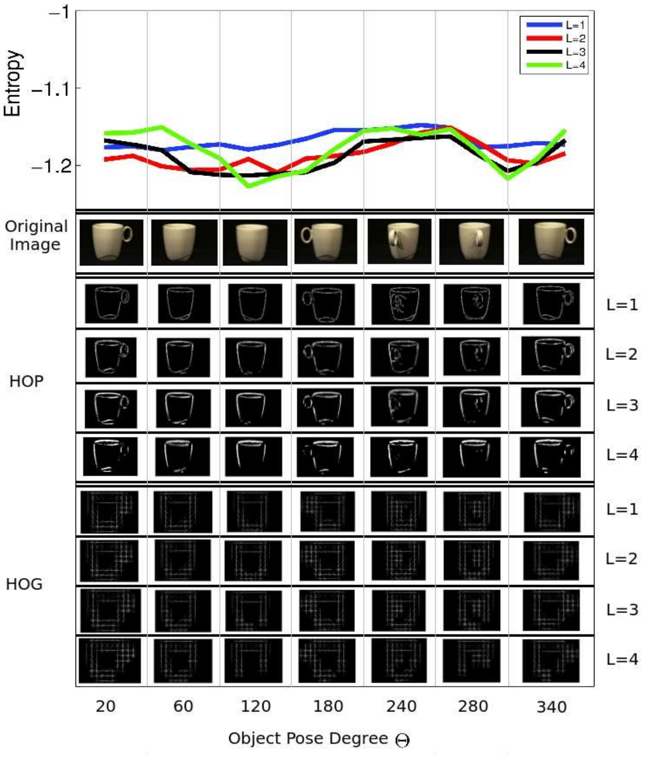

Fig. 3: Visualization of features extracted from part realizations for each different orientation of a cup and at each different layer of LHOP.

In Figure 3, we first observe that $f^{l=1}_{ent}$ values of the object change discriminatively across different object orientations Θ. For instance, if the handle of the cup is not seen from the front viewpoint of the cup (e.g. at $\Theta = 60°, 120°$), then we observe a smooth surface of the cup and the complexity of the part graphs, i.e. the entropy values, decrease. On the other

TABLE I: The samples that are selected from ALOI dataset and used in Object-wise Pose Estimation Experiments

Category Name	Apples	Balls	Bottles	Boxes	Cars	Cups	Shoes
Object IDs for Training	82	103	762	13	54	157	9
Object IDs for Testing	363, 540, 649, 710	164, 266, 291, 585	798, 829, 831, 965	110, 26, 46, 78	136, 138, 148, 158	36, 125, 153, 259	93, 113, 350, 826

hand, if the handle of the cup is observed at a front viewpoint (e.g. at $\Theta = 240°, 280°$), then the complexity increases. In addition, we observe that the difference between f^l_{ent} values of the object parts across different orientations Θ decreases as l increases. In other words, the discriminative power of the generative model of the LHOP increases at the higher layers of the LHOP since the LHOP captures the *important* parts and compositions that are co-occurred across different views through different layers.

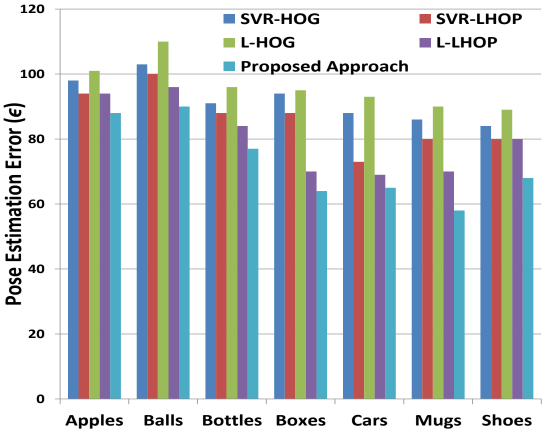

Fig. 4: Comparision of Object-wise Pose estimation errors (ϵ) of the proposed algorithms.

Given a ground truth Θ and an estimated pose value $\hat{\Theta}$, the pose estimation error is defined as $\epsilon = ||\Theta - \hat{\Theta}||_2^2$. Pose estimation errors of state-of-the-art algorithms and the proposed Hierarchical Compositional Approach are given in Figure 4. In these results, we observe that the pose estimation errors of the algorithms which are implemented using the symmetric objects, such as apples and balls, are greater than that of the algorithms that are implemented on more structural objects such as cups.

In order to analyze this observation in detail, we show the ground truth Θ and the estimated orientations $\hat{\Theta}$ of some of the objects from Apples, Balls, cups and Shoes categories in Figure 5. We observe that some of the different views of the same object have the same shape and textural properties. For instance, the views of the ball at the orientations $\Theta = 10°$ and $\Theta = 225°$ represent the same pentagonal shape patterns. Therefore, similar parts are detected at these different views and the similar features are extracted from these detected parts. Then, the orientation of the ball, which is rotated by $\Theta = 10°$, is incorrectly estimated as $\hat{\Theta} = 225°$.

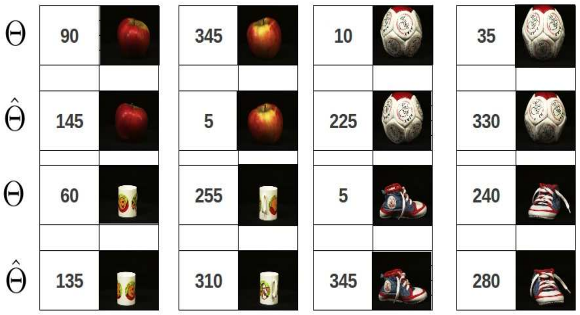

Fig. 5: Results for some of the objects from Apples, Balls, Cups and Shoes categories obtained in Object-wise Pose estimation experiments.

2) *Experiments on Category-wise Pose Estimation:* In Category-wise Pose Estimation experiments, we select different $\aleph^o_{tr}$ number of objects from different C number of categories as training images to estimate the pose of test objects, randomly. We employ the experiments on both ALOI and COIL datasets.

In the ALOI dataset, we randomly select $\aleph^o_{tr} = 1, 2, 3, 4$ number of training objects and $\aleph^o_{te} = 1$ test object which belong to Cups, Cow, Car, Clock and Duck categories. We repeat the random selection process two times and give the average pose estimation error for each experiment. In order to analyze the contribution of the information that can be obtained from the parts to the pose estimation performance using the part sharability of the LHOP, we initially select Cups and Cow categories ($C = 2$) and add new categories (Car, Clock and Duck) to the dataset, incrementally. The results are given in Table II. The results show that the pose estimation error decreases as the number of training samples, $\aleph^o_{tr}$, increases. This is due to the fact that the addition of new objects to the dataset increases the statistical representation capacity of the LHOP and the learning model of the regression algorithm. In addition, we observe that the pose estimation error observed in the experiments for $C = 2$ decreases when the objects from Car category are added to a dataset of objects belonging to Cups and Cow category in the experiments with $C = 3$. The performance boost is achieved by increasing the shareability of co-occurred object parts in different categories. For instance, the parts that construct the rectangular silhouettes of cows and cars can be shared in the construction of object representations in the LHOP (see Figure 6.

We employed two types of experiments on COIL dataset, constructing balanced and unbalanced training and test sets,

TABLE II: Category-wise Pose estimation errors (ϵ) of SVR-HOG/SVR-LHOP/L-HOG/L-LHOP/Proposed Approach for different number of categories (C) and training samples ($\aleph^o_{tr}$) selected from ALOI dataset.

$\aleph^o_{tr}$	C=2	C=3	C=4	C=5
1	133/103/140/97/91	116/99/110/97/89	110/95/102/95/88	102/94/99/95/88
2	130/100/133/95/85	108/93/104/88/81	105/91/95/88/80	100/94/100/91/85
3	105/91/104/86/75	93/83/87/83/70	99/86/94/84/75	95/81/93/75/70
4	94/86/90/73/68	90/79/84/73/65	92/77/86/72/64	95/75/88/71/60

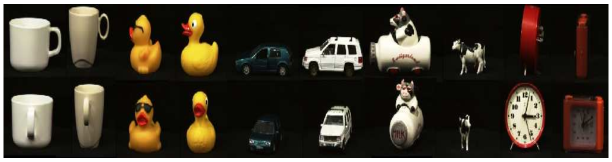

Fig. 6: Sample images of the objects that are used in Category-wise Pose Estimation experiments.

in order to analyze the effect of the unbalanced data to the pose estimation performance. In the experiments, the objects are selected from Cat, Spatula, Cups and Car categories which contain 3, 3, 10 and 10 objects. Each object is rotated on a turntable by 5° from 0° to 355°.

In the experiments on balanced datasets, images of $\aleph^o_{tr}$ number of objects are initially selected from Cat and Spatula categories (for $C = 2$), and then images of the objects selected from Cups and Car categories are incrementally added to the dataset for $C = 3$ and $C = 4$ category experiments. More specifically, $\aleph^o_{tr}$ objects are randomly selected from each category and the random selection is repeated two times for each experiment. The results are shown in Table III. We observe that the addition of new objects to the datasets decreases the pose estimation error. Moreover, we observe a remarkable performance boost when the images of the objects from the categories that have similar silhouettes, such as Cat and Cups or Spatula and Car, are used in the same dataset.

TABLE III: Category-wise Pose estimation errors (ϵ) of SVR-HOG/SVR-LHOP/L-HOG/L-LHOP/Proposed Approach for different number of categories (C) and training samples ($\aleph^o_{tr}$) selected from COIL dataset.

$\aleph^o_{tr}$	C=2	C=3	C=4
1	125/109/120/95/85	120/85/103/77/68	110/79/95/71/62
2	120/95/114/89/77	93/77/81/63/59	104/76/92/69/51

We prepared unbalanced datasets by randomly selecting the images of $\aleph^o_{te} = 1$ object from each category as a test sample and the images of the rest of the objects belonging to the associated category in the COIL dataset as training samples. For instance, the images of a randomly selected cat are selected as test samples and the images of the remaining two cats are selected as training samples. This procedure is repeated two times in each experiment and the average values of pose estimation errors are depicted in Figure 7. The results show that SVR is more sensitive to the balance of the dataset and the number of training samples than the proposed approach. For instance, the difference between the pose estimation error of SVR given in Table III and Figure 7 for $C = 4$ is approximately 10°, while that of the proposed Hierarchical Compositional Approach is approximately 5°.

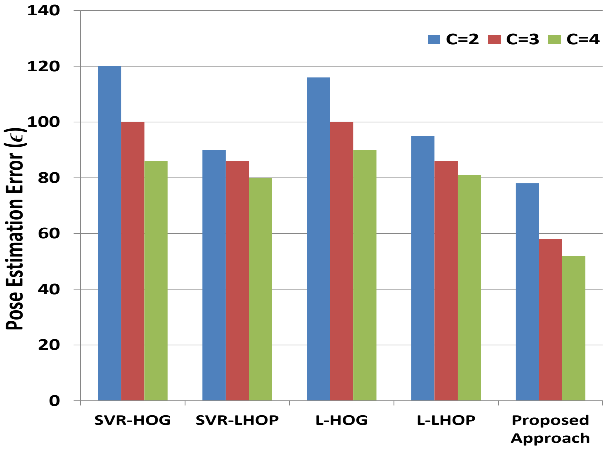

Fig. 7: Category-wise Pose estimation errors (ϵ) of the state-of-the-art algorithms and the proposed Hierarchical Compositional Approach in the experiments on COIL dataset.

In the next subsection, the experiments on object categorization are given.

B. Experiments on Object Categorization

In the Object Categorization experiments, we use the same experimental settings that are described in Section V-A.2 for Category-wise Pose Estimation.

TABLE V: Categorization performance (%) of SVM-HOG/SVM-LHOP/LR-HOG/LR-LHOP/Proposed Approach using COIL dataset.

$\aleph^o_{tr}$	C=2	C=3	C=4
1	94/93/92/95/100	89/88/91/91/97	81/79/80/81/84
2	97/97/96/97/100	89/91/90/93/97	84/86/83/87/90

The results of the experiments employed on ALOI dataset and balanced subsets of COIL dataset are given in Table IV and Table V, respectively. In these experiments, we observe that the categorization performance decreases as the number of categories increases. However, we observe that the pose estimation error decreases as the number of

TABLE IV: Categorization performance (%) of SVM-HOG/SVM-LHOP/LR-HOG/LR-LHOP/Proposed Approach for different number of categories (C) and training samples ($\aleph^{o}_{tr}$) selected from ALOI dataset.

$\aleph^{o}_{tr}$	C=2	C=3	C=4	C=5
1	88/89/91/93/100	85/88/84/92/98	85/85/84/85/90	81/81/81/83/90
2	88/91/92/94/100	88/91/87/93/98	87/87/86/88/92	81/83/81/84/91
3	95/98/94/98/100	91/93/91/95/99	90/90/90/91/93	83/85/83/88/91
4	97/98/98/99/100	93/96/93/97/100	90/91/90/91/94	87/91/89/95/96

categories increases in the previous sections. The reason of the observation of this error difference is that the objects rotated on a turn table may provide similar silhouettes although they may belong to different categories. Therefore, addition of the images of new objects that belong to different categories may boost pose estimation performance. On the other hand, addition of the images of these new objects may decrease the categorization performance if the parts of the object cannot be shared across different categories and increase the data complexity of the feature space.

VI. Conclusion

In this paper, we have proposed a compositional hierarchical approach for joint object pose estimation and categorization using a generative-discriminative learning method. The proposed approach first exposes information about pose and category of an object by extracting features from its realizations observed at different layers of LHOP in order to consider different levels of abstraction of information represented in the hierarchy. Next, we formulate joint object pose estimation and categorization problem as a sparse optimization problem. Then, we solve the optimization problem by integrating the features extracted at each different layer using a distributed and parallel optimization algorithm.

We examine the proposed approach on benchmark 2-D multi-view image datasets. In the experiments, the proposed approach outperforms state-of-the-art Support Vector Machines for object categorization and Support Vector Regression algorithm for object pose estimation. In addition, we observe that shareability of object parts across different object categories and views may increase pose estimation performance. On the other hand, object categorization performance may decrease as the number of categories increases if parts of an object cannot be shared across different categories, and increase the data complexity of the feature space. The proposed approach can successfully estimate the pose of objects which have view-specific statistical and geometric properties. On the other hand, the proposed feature extraction algorithms cannot provide information about the view-specific properties of symmetric or semi-symmetric objects, which leads to a decrease of the object pose estimation and categorization performance. Therefore, the ongoing work is directed towards alleviating the problems with symmetric or semi-symmetric objects.

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[2] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, May 2011.

[3] A. Collet, D. Berenson, S. Srinivasa, and D. Ferguson, "Object recognition and full pose registration from a single image for robotic manipulation," in *Proc. IEEE Conf. Robotics and Automation*, 2009, pp. 48–55.

[4] A. Collet, M. Martinez, and S. S. Srinivasa, "The moped framework: Object recognition and pose estimation for manipulation," *Int. J. Rob. Res.*, vol. 30, no. 10, pp. 1284–1306, Sep 2011.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1. Washington, DC, USA: IEEE Computer Society, 2005, pp. 886–893.

[6] W. Du, X. Li, Y. Li, and S. Severini, "A note on the von neumann entropy of random graphs." *Linear Algebra Appl.*, vol. 433, no. 11-12, pp. 1722–1725, 2010.

[7] S. Fidler and A. Leonardis, "Towards scalable representations of object categories: Learning a hierarchy of parts," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[8] S. Fidler, M. Boben, and A. Leonardis, *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press, 2009, ch. Learning Hierarchical Compositional Representations of Object Structure.

[9] ——, "A coarse-to-fine taxonomy of constellations for fast multi-class object detection," in *Proceedings of the 11th European Conference on Computer Vision: Part V*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 687–700.

[10] J.-M. Geusebroek, G. Burghouts, and A. Smeulders, "The amsterdam library of object images," *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[11] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich, "Viewpoint-aware object detection and continuous pose estimation," *Image Vision Comput*, vol. 30, pp. 923–933, 2012.

[12] T. Haštie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[13] Y. Jiang, M. Lim, C. Zheng, and A. Saxena, "Learning to place new objects in a scene," *Int. J. Rob. Res.*, vol. 31, no. 9, pp. 1021–1043, Aug 2012.

[14] G. Kootstra, M. Popović, J. A. Jørgensen, K. Kuklinski, K. Miatliuk, D. Kragic, and N. Krüger, "Enabling grasping of unknown objects through a synergistic use of edge and surface information," *Int. J. Rob. Res.*, vol. 31, no. 10, pp. 1190–1213, Sep 2012.

[15] K. Lai, L. Bo, X. Ren, and D. Fox, "A scalable tree-based approach for joint object and pose recognition," in *Proc. The 25th AAAI Conf. Artificial Intelligence*, Aug 2011.

[16] A. Mowshowitz and M. Dehmer, "Entropy and the complexity of graphs revisited." *Entropy*, vol. 14, no. 3, pp. 559–570, 2012.

[17] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia Object Image Library (COIL-100)," Department of Computer Science, Columbia University, Tech. Rep., Feb 1996.

[18] M. Ozay, I. Esnaola, F. Vural, S. Kulkarni, and H. Poor, "Sparse attack construction and state estimation in the smart grid: Centralized and distributed models," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1306–1318, 2013.

[19] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J Comput Graph Stat*, vol. 10, pp. 231–245, 2012.

[20] D. Teney and J. Piater, "Probabilistic object models for pose estimation in 2d images," in *Pattern Recognition*, ser. Lecture Notes in Computer Science, R. Mester and M. Felsberg, Eds. Springer Berlin Heidelberg, 2011, vol. 6835, pp. 336–345.