SIViP (2016) 10:1503–1510
DOI 10.1007/s11760-016-0962-x

ORIGINAL PAPER

Efficient object detection using convolutional neural network-based hierarchical feature modeling

Byungjae Lee[1] · Enkhbayar Erdenee[1] · Songguo Jin[1] · Phill Kyu Rhee[1]

Received: 26 March 2016 / Revised: 14 July 2016 / Accepted: 5 August 2016 / Published online: 1 September 2016

Abstract A hierarchical data-driven object detection framework is addressed considering a deep feature hierarchy of object appearances. The performance of many object detectors is degraded due to ambiguities in inter-class appearances and variations in intra-class appearances, but deep features extracted from visual objects show a strong hierarchical clustering property. Deep features were partitioned into unsupervised super-categories at the inter-class level, and augmented categories at the object level, to discover deep feature-driven information. A hierarchical feature model is built using a latent topic model algorithm, assembling a one-versus-all support vector machine at each node to constitute a hierarchical classification ensemble. Extensive experiments show that the proposed method is superior to state-of-the-art techniques using the PASCAL VOC 2007 and VOC 2012 datasets.

Keywords Object detection · Deep learning · Convolutional neural network · Hierarchical feature modeling

1 Introduction

Increasingly, object detection, which consists of object categorization and localization, is becoming a very challenging problem in computer vision area [1–6]. Object detection is a very complex process due to image ambiguities in inter-class appearance and deformations due to large intra-category variations. Much research has addressed improving the performance degradation of object detection by dividing training samples into multiple components and learning the components independently [2]. The decomposition of a training dataset can relieve local deformation and variations within intra-classes. Some early pioneering research investigates clustering approaches for training data in terms of object scale, pose [7,8], aspect ratio [2], and component labels [9]. However, most of them only consider intra-class variations [10,11] and do not investigate inter-class ambiguity, even though performance can be improved further by considering ambiguity between inter-classes. Some progress in detection performance is based on more general sub-category models within semantic object categories [12–14]. Gu et al. [11] partitioned the samples into components using the annotated key point and masks, and Aghazadeh et al. [14] used a similarity graph denoting intra-class information to split the data into spectral clusters. Ruan et al. [15] investigated weakly supervised multicomponent model learning for sub-categories.

Even though much of the research takes advantage of sub-category structures to improve the accuracy of object detection [16,17], most sub-categories are built based on only intra-class similarity information. However, there are many confusing objects from inter-class ambiguities [18]. Valuable inter-class information can be used to solve the confusing sub-category problem of intra-class samples. Recently, Dong et al. [1] proposed an sub-category mining approach to explore intra-class diversities. However, the performance is

Electronic supplementary material The online version of this article (doi:10.1007/s11760-016-0962-x) contains supplementary material, which is available to authorized users.

✉ Phill Kyu Rhee
pkrhee@inha.ac.kr

Byungjae Lee
jaylee@inha.edu

Enkhbayar Erdenee
enkhbayar@inha.edu

Songguo Jin
sgkim735@inha.edu

[1] Inha University, 235 Yong-Hyun Dong, Nam Ku, Incheon, South Korea

much more inferior to recently proposed deep feature-based object detection methods, such as the fast region-based convolutional network (fast R-CNN) and the spatial pyramid pooling in convolutional network (SPP) [19,20].

This paper presents a hierarchical deep feature-driven learning framework with a generalization ability instead of traditional algorithm-centered detection algorithms. This is motivated by the following observations: (1) The performance of many object detectors is degraded, due not only to large intra-class variations but also to ambiguity in inter-class differences and (2) the deep features extracted from visual objects show a strong hierarchical clustering property. This paper presents a novel object detection method using a hierarchical feature model (HFM) and a hierarchical classifier ensemble (HCE), which is characterized by a generic and flexible feature structure in terms of super-, augmented, and sub-categories. Here, the augmented category is a partition of a semantic object category considering the effect of super-categories using latent topic model (LTM, Sect. 3.1). Therefore, each augmented category corresponds to a single semantic object category. Figure 1 shows the concept of an augmented category of HFM. For example, the person category can be divided into three augmented categories such as a person who is sitting (augmented category 2), standing (augmented category 3), and riding (augmented category 4). In large-scale object detection, the classifier built using an augmented class is expected to offer better performance than one built using a semantic object category, since the category similarity error is efficiently reduced.

The proposed method is the first complete end-to-end approach, which interactively builds a hierarchical feature structure and classifier ensemble to explore generalization abilities in object categorization and localization. At each node of a data hierarchy HFM, a multi-level classification ensemble like that of Goh et al. [21] is employed, but adaptively. Instead of using a flat linear SVM for all object classes [19,20], the hierarchical SVM ensemble, HCE, is used for both inter-class and intra-class decisions. In the first step, HCE employs one-against-all SVMs to calculate the confidence factor for one class prediction made by a binary SVM classifier for each augmented category label. In the second step, multi-class confidence scores for object detection are aggregated by combining multiple detectors. In the third step, each HCE tree is trained on a different decision path of the HFM and is used to calculate the overall confidence score of a test image to minimize detection error. In the detection phase, HCE-based object detection is performed by calculating the confidence score(s) of each region proposal driven by HFM. The major contributions of this paper are summarized as follows.

- The concept of augmented object categories can resolve inter-class ambiguity and intra-class variation problems, especially in very-large-scale object detection. The method reduces computation overhead, since regions of interest (ROIs) are assigned restricted augmented categories instead of full assignment of entire semantic

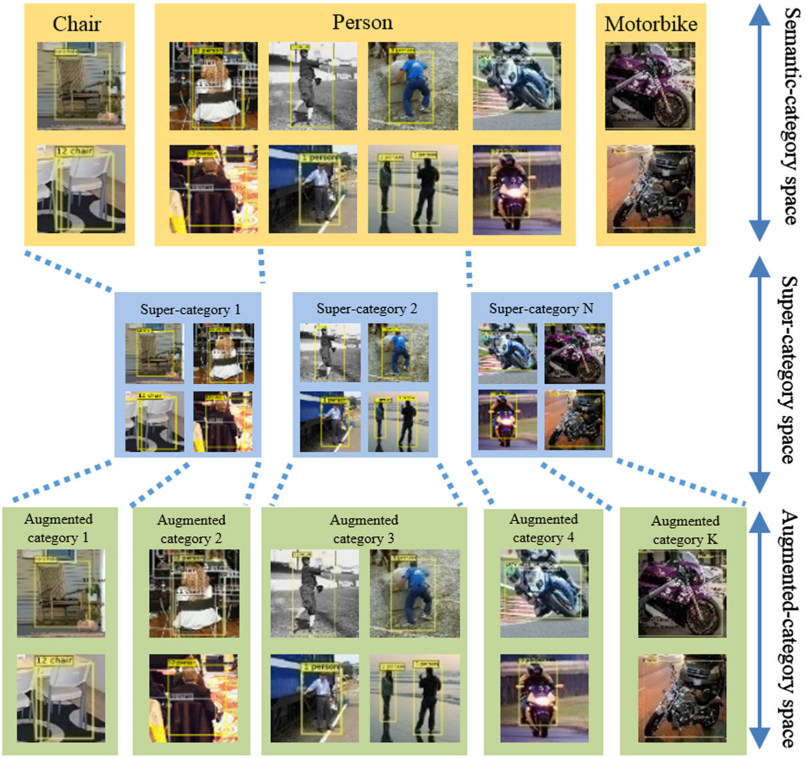

Fig. 1 The concept of augmented category compared to the original semantic object categories. For each semantic object category, training samples are partitioned into unsupervised super-categories. The augmented category is determined by partitioning a semantic object category based on super-categories. Therefore, each augmented category corresponds to a single semantic object category

categories, as can be found in state-of-the-art technologies such as SPP [19], and fast R-CNN [20].

- HFM is shown to be more effective than the flat feature model [19,20] and sub-category-based feature models [1, 16,17,22] by combining it with a hierarchical classifier ensemble, which takes advantage of the clustering quality of the deep-feature hierarchy.
- Many confusing data samples can be clustered properly into sub-categories by taking advantage of inter-class information, and overall detection accuracy can be improved by solving simplified sub-problems.

2 System overview

In general, classification performance degrades as the number of object classes increases [21]. While deployments of the flat SVM and fully connected neural network are successful for a small or moderate number of object classes [19,23], detection performance degrades as the number of object categories increases. Data imbalance is a common phenomenon due to the increment of noise and variations, contaminating the image data in the real world. The proposed HFM-driven detection method aims at providing robust object detection with a generalization ability. The novelty is the accuracy improvement based on HCE and HFM in a data-driven manner. HFM, the core of the proposed detection method, is constituted by a three-level cluster tree consisting of the super-category, augmented object category, and sub-category feature models. HFM takes advantage of feature information of unsupervised super-categories and semantic sub-categories for semantic object category recognition [23]. The region proposal algorithm EdgeBoxes is used [18] to extract the region of interest (ROI). In the learning phase, category hierarchy can be found by using LTM with extracted features from pre-trained CNN. HFM is built for the inter-class level and intra-class level by being fine-tuned on a hierarchical category. HCE is built by training a multi-classifier at each node of HFM using an SVM ensemble algorithm. The region compensation model is built using the hierarchical ridge regression algorithm.

In the detection phase, the pool of augmented object categories is predicted in terms of HCE subtree hypotheses for the ROIs generated by the region proposal algorithm. Object detection is performed based on the hypothesis, followed by ridge regression and non-maximum suppression similar to that of Girshick [20]. The scores of an HCE subtree are combined in terms of the super-category, augmented category, and sub-categories for the hypothesis ROIs. Finally, the post-processing of non-maximum suppression is executed using the combined scores and position information, and object category is determined.

3 Hierarchical feature modeling

The region proposal algorithm EdgeBoxes [18] is employed to find ROIs from an image, and ROI features are generated using 16-layer CNN [19,20]. Using normalized ROI features builds a deep-feature hierarchy HFM that consists of three different levels: H-level (inter-class), M-level (augmented class), and L-level (intra-class), as shown in Fig. 2.

The root node of HFM has the super-category nodes as children in the H-level. Each super-category has one or more augmented object categories as children in the M-level, which are original or partitioned semantic object classes according to the inter-class characteristics. Each augmented object category node has sub-category leaves as children in the L-level. HCE is built by training the multi-category classifier at each node of HFM, which is an assembly of one-versus-all SVMs [21]. One can notice that Girshicks flat-feature structure [20] is a special case of HFM that only has the root and entire semantic object categories in the M-level without augmented categories. This flat HFM structure is called the flat feature model (FFM). HFM is also thought of as a generalization of the sub-category-based approaches [1,16,17,22].

3.1 Latent topic model for category hierarchy

This section introduces an unsupervised approach to learning a data-driven hierarchical category. For the unsupervised learning step, a super-category using a latent topic model [24] is built. As a mixture model, LTM provides a novel way to represent latent mixture components for grouping data, which is advantageous for learning hierarchical structures. In the LTM, the ROIs are represented by the combinations of latent topics. These learned topics correspond to a super-category to build a category hierarchy.

For ROI representation, a feature extracted from a deep convolutional neural network is used. More specifically, the CNN model is fine-tuned on the training set using a pre-trained CNN model. Then, fixed-length feature vectors are extracted from the last fully-connected layer for each ROI. Extracted features are encoded, quantized, and scaled.

Considering that each ROI is represented by a vector, the goal is to learn a super-category by fitting a mixture model on the represented ROIs. In detail, the LTM presents each ROI in combinations of K topics. Each topic corresponds to one super-category or a few super-categories. A similar generative process to that of the original LTM [24] is used.

3.2 Hierarchical feature model

HFM is built using ROI features of the training dataset D associated with semantic object space Ω. HFM is constituted by the root nodes, super-category nodes, augmented category

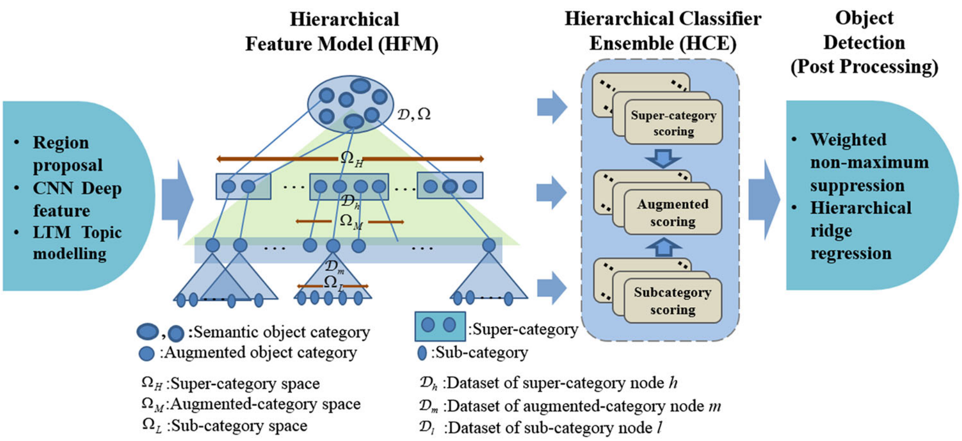
Hierarchical Feature Model (HFM)
Hierarchical Classifier Ensemble (HCE)
Object Detection (Post Processing)
• Region proposal
• CNN Deep feature
• LTM Topic modelling
$\mathcal{D}, \Omega$
Ω_H
$\mathcal{D}_h$
Ω_M
$\mathcal{D}_m$
Ω_L
Super-category scoring
Augmented scoring
Subcategory scoring
• Weighted non-maximum suppression
• Hierarchical ridge regression
:Semantic object category
:Augmented object category
:Super-category
:Sub-category
Ω_H :Super-category space
Ω_M:Augmented-category space
Ω_L :Sub-category space
$\mathcal{D}_h$:Dataset of super-category node h
$\mathcal{D}_m$:Dataset of augmented-category node m
$\mathcal{D}_l$:Dataset of sub-category node l

Fig. 2 The proposed object detection framework based on the hierarchical deep feature model (HFM), and hierarchical classifier ensemble (HCE)

nodes, and sub-leaves. The root of HFM is associated with the entire feature set D with category space Ω and is connected to the super-category nodes as children. Note that a semantic object with multiple super-categories is partitioned into several augmented categories according to the inter-class characteristics. The concept of an augmented category is introduced to reduce the effect of ambiguity at the inter-class level as well as variation at the intra-class level of the original semantic object class and can be extended to represent a multi-label category and occluded semantic object categories. A super-category node h is associated with training dataset D_h, which is a subset of D and has multiple augmented object nodes as children. LTM analysis allows one semantic object category to belong to several super-categories, since different objects can share parts with similar appearance or characteristics. At the M-level, a semantic object category that belongs to multiple super-categories is divided into multiple augmented categories. The augmented category m has a training dataset, denoted by D_m, which is partitioned from the training dataset D_h. The training set of each augmented object category is further partitioned into sub-categories at the low level using the LTM algorithm to minimize the effect of intra-class variations. The training dataset of sub-category l is denoted as D_l, which is partitioned from the augmented training dataset D_m.

3.3 Hierarchical classifier ensemble

The node confidence functions are constituted by multi-class classifiers, which are built by the assembly of binary classifiers of the individual child nodes. Confidence scores are required to keep linear relationships with the expected prediction accuracies and to increase prediction reliability with multiple considerations. Let Ω_H, Ω_M, and Ω_L denote the spaces of the super-, augmented, and sub-categories, respectively (Fig. 2). The root has $|\Omega_H|$ super-categories as children, super-category h has $|\Omega_M|$ augmented object categories, and augmented category m has $|\Omega_L|$ sub-categories. At the root is constructing an SVM ensemble that calculates the confidence score of an ROI used when traversing to a super-category node. $|\Omega_H|$ binary SVM classifiers $\phi_1, \phi_2, \ldots, \phi_{|\Omega_H|}$ are trained using D at the root node, which is used by an ROI in deciding super-category nodes. It is not possible to trust the predictions estimated by binary SVMs as being used directly for multi-class classification [21,25], so the confidence functions discussed below are introduced.

Given an ROI r, linear SVM ϕ_h for super-category h is projected to pseudo-probability $P(y=h|r)$ as follows [26]:

$$P(y=h|r)=\frac{1}{1+\exp(\alpha\times\phi_h(r)+\beta)} \tag{1}$$

where parameters α and β are determined by logistic regression, as follows [25]:

$$(\alpha^*\beta^*)=\arg\min_{\alpha,\beta}\sum\nolimits_{r\in \mathrm{D}_h} wy(\alpha^T\phi_h(r)+\beta) - \log(1+e^{\alpha^T\phi_h(r)+\beta})$$

where w_i is the weight for the ROI sample, and y is the corresponding label. Given ROI r, the multi-class prediction at the root node is begun by deciding the top-scored super-category node for ROI r in terms of the pseudo-probability, as follows:

$$h^{(1)}(r)=\arg\max_{h\in\Omega_H} P(h|r) \tag{2a}$$

Let the multi-class margin be defined as follows:

$$\xi_h^{(1)}(r)=P(h^{(1)}|r)-\max_{1\le h\le|\Omega_H|,h\ne h^{(1)}} P(h|r). \tag{2b}$$

Normalized multi-class margin φ_h is calculated based on the relationship between the pseudo-prediction $P(h^{(1)}|r)$ and multi-class margin ξ_h with the sigmoid function

$$\varphi_h^{(1)}(r) = A + \frac{B}{1 + \exp(-C \times \xi_h^{(1)}(r))} \tag{2c}$$

where parameters A, B, and C are determined through empirical fitting [21].

The confidence function $CS_h(r)$ for the first super-category prediction $h^{(1)}$ at the root is defined as follows:

$$CS_h^{(1)}(r) = \sqrt{P(h^{(1)}|\mathrm{r}) \times \varphi_h^{(1)}(r)} \tag{2d}$$

The kth confidence function is denoted by

$$CS_h^{(k)}(r) = \sqrt{P(h^{(k)}|r) \times \varphi_h^{(k)}(r)} \tag{2e}$$

where

$$h^{(k)} = \arg \max_{h \in \{\Omega_H - h^{(1)} - h^{(2)} \ldots, h^{(k-1)}\}} P(h|r)$$

$$\xi_h^{(k)}(r) = P(h^{(k)}|r) - \max_{1 \le h \le |\Omega_H|, h \ne h^{(1)}, \ldots, h^{(k-1)}} P(h|r)$$

and

$$\varphi_h^{(k)}(r) = A + \frac{B}{1 + \exp(-C \times \xi_h^{(k)}(r))}.$$

A super-category node has the set of augmented object categories, Ω_M, which is much smaller than total semantic object categories N. An SVM ensemble for each super-category node is built and calculates the confidence scores for each augmented object category. At each augmented object category, an SVM ensemble is built.

$|\Omega_M|$ binary SVM classifiers $\phi'_1, \phi'_2, \ldots, \phi'_{|\Omega_M|}$ are trained at a super-category node to decide the best node(s) in the augmented category space, Ω_M. Linear SVM ϕ'_m for augmented category m is projected to pseudo-probability $P(y = m|r)$, defined in the following:

$$P(y = m|r) = \frac{1}{1 + \exp(\alpha \times \phi'_m(r) + \beta)} \tag{3a}$$

Define the $k'th$ confidence score for each augmented category calculated at a super-category node:

$$CS_m(r) = \sqrt{P(m^{(k')}|r) \times \varphi_m^{(k')}(r)} \tag{3b}$$

where

$$\varphi_m^{(k')}(r) = A' + \frac{B'}{1 + \exp(-C' \times \xi_m'^{(k')}(r))}$$

and

$$\xi_m'^{(k')}(r) = P(m^{(k')}|r) - \max_{1 \le m \le |\Omega_M|, m \ne m^{(1)}, \ldots, m^{(k'-1)}} P(m|r).$$

Train $|\Omega_L|$ binary SVM classifiers $\phi''_1, \phi''_2, \ldots, \phi''_{|\Omega_L|}$ at an augmented category node to decide the best leaves in sub-category space Ω_L. Similarly, the k''^{th} confidence score is defined for each sub-category at an augmented-category node as follows:

$$CS_l^{(k'')}(r) = \sqrt{P(l^{(k'')}|r) \times \varphi_l^{(k'')}(r)} \tag{4}$$

where

$$P(y = l^{(k'')}|r) = \frac{1}{1 + \exp(\alpha'' \times \phi''(r) + \beta'')}$$

$$\varphi_l^{(k'')}(r) = A'' + \frac{B''}{1 + \exp(-C'' \times \xi_l''^{(k'')}(r))}$$

and

$$\xi_l''^{(k'')}(r) = P(l^{(k'')}|r) - \max_{1 \le l \le |\Omega_L|, l \ne l^{(1)}, \ldots, l^{(k''-1)}} P(l|r).$$

4 Experiments

The object detector based on HFM with HCE was evaluated on the PASCAL VOC 2007 and PASCAL VOC 2012 [27] detection tasks. Each dataset contains thousands of images of real-world scenes, and the goal is to predict the bounding boxes of all objects in an image. If a predicted bounding box overlaps by more than 50 % with ground truth, it is considered a true positive. Sixteen-layer VGG-Net [28] was used as the system's baseline. Among the state-of-the-art region proposal algorithms, EdgeBoxes [18] was employed because it is fast and provides more accurate region proposals. In the first experiment on PASCAL VOC 2007, we trained the detector on the trainval set. The second experiment is evaluated PASCAL VOC 2007 with knowledge transfer learning on HFM. Finally, we compare our method with state-of-art methods on PASCAL VOC 2012 public leaderboard.

4.1 VOC 2007 results

The experimental results on the PASCAL VOC 2007 dataset are shown in Table 1. Each method is distinguished in terms of FFM, HFM with augmented-category level (M-level HFM), and HFM with sub-category level (L-level HFM). All experiments in Table 1 were trained on the VOC 2007 trainval set and were evaluated with mean average precision (mAP) on a test set using the standard PASCAL evaluation tool [27]. Detection performance compared to a state-of-the-art detector's results is shown in Table 1.

Table 1 Detection results on the PASCAL VOC 2007 test set

Approach	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Person	mAP (%)
SPPnet VGG16 [19]	73.9	72.3	62.5	51.5	44.4	74.4	73.0	74.4	42.3	73.6	57.7	70.3	74.6	74.3	54.2	63.1
R-CNN VGG16 [23]	73.4	77.0	63.4	45.4	44.6	75.1	78.1	79.8	40.5	73.7	62.2	79.4	78.1	73.1	64.2	66.0
Fast R-CNN VGG16 [20]	74.5	78.3	69.2	53.2	36.6	77.3	78.2	62.0	40.7	72.7	67.9	79.6	79.2	73.0	69.0	66.9
FFM	74.8	76.5	65.5	54.1	40.0	78.0	78.5	81.7	44.9	74.5	66.8	79.4	81.9	74.6	66.3	67.9
M-level HFM	75.3	83.6	70.6	59.3	42.5	81.3	81.6	84.7	45.1	80.7	69.2	83.2	83.3	76.2	72.0	70.8
L-level HFM	76.6	82.4	74.1	61.4	46.9	83.7	84.4	87.5	48.1	80.6	68.5	84.1	84.4	78.5	73.6	72.3

4.1.1 FFM

FFM is a special type of HFM that only has the root and entire semantic object categories without augmented categories. FFM is used for the feature extractor in the experiment. A public VGG16 [28] CNN structure was chosen as the baseline, following the training protocols [20]. To build the FFM, an ImageNet pre-trained CNN was fine-tuned on data $(\mathcal{D}, \Omega)$ with 50K iterations at a learning rate of 0.001. After 50K iterations, the learning rate was decreased by a factor of 10 for fine-tuning with 20K iterations. During FFM fine-tuning, only the weights from *conv4_1* to *fc7* were fine-tuned, whereas the ones from *conv1_1* to *conv3_3* were fixed.

4.1.2 M-level HFM

M-level HFM was built without sub-category level on the VOC 2007 trainval set. To build a category hierarchy, K topics were set on LTM, finding K-dimensional super-category distribution θ^K for each ROI. In this experiment, K was set as 5, which was selected by a grid search over {1,2,…,9}. A disjoint HFM H-level can be easily overfitted because of data sparsity, and the HFM hierarchical structure misleads some candidates. To overcome the overfitting and misleading problems, multiple super-categories were allowed for each ROI r by determining super-categories i where $\left\{\theta_r^i\right\}_{i=1}^K > T_\theta$ instead of $\text{argmax}_i\ \theta_r^i$. T_θ was empirically determined as 0.3. To build the HFM, post hoc SVM training was implemented with hard negative mining [23] for training with a very large dataset. The HFM was constructed by following Sect. 3.2. After learning HCE (see Sect. 3.3), feasible HCE subtrees compete against each other in post-processing to localize the final object position. In post-processing, hierarchical ridge regression was adapted, as well as the weighted non-maximum suppression described in supplementary material. Table 1 shows significant improvement can be achieved by adapting HFM with 2.9 % from FFM and reached 70.8 %.

4.1.3 L-level HFM

L-level HFM was built by considering sub-category level. After M-level HFM is constructed, sub-categories were discovered by LTM. Therefore, each augmented object category node at M-level had sub-category leaves as children in the L-level. Learning process was same as M-level HFM. The improvement in L-level HFM is 4.4 % from FFM. An overall 72.3 % mAP was achieved on the VOC 2007 dataset, which is higher than state-of-the-art methods such as Fast R-CNN [20], at 66.9 %, as shown in Table 1.

4.2 VOC 2007 results with domain adaptation

The potential of the model for cross-domain transfer learning was evaluated using Microsoft's Common Objects in Context (COCO) 2014 [33] and PASCAL VOC 2007 + VOC 2012 (VOC+) as two domains. First, the following two baselines were considered without transfer learning.

4.2.1 VOC+

In order to build FFM_{VOC+}, an ImageNet pre-trained CNN was fine-tuned on data $(\mathcal{D}^{VOC+}, \Omega^{VOC+})$ with 50K iterations and a learning rate of 0.001, and then the learning rate was decreased by a factor of 10 for 20K iterations. After FFM_{VOC+} was fine-tuned, HFM was built on VOC+ using category hierarchy, which was obtained by LTM. All parameters were fixed as described in Sect. 4.1. Performance on the VOC 2007 test of the VOC+ baseline was 75.6 %, as described in Table 2. Performance improvement compared to fast R-CNN was more dramatically achieved by training with additional data (from 3.1 to 4.8 %). This result is mainly due to the HFM approach of constructing a hierarchical structure, and its ability, which is boosted on a larger dataset.

4.2.2 COCO

Even though the COCO dataset has a different number of classes compared to VOC, all 80 classes were used to train the COCO baseline, since VOC can be considered a subset of COCO. First, FFM_{COCO} was constructed, which was fine-tuned from ImageNet pre-trained CNN on data $(\mathcal{D}^{COCO}, \Omega^{COCO})$ with 200K iterations and a learning rate of 0.001. Then, the learning rate was decreased by a factor

Table 2 Detection results on the PASCAL VOC 2007 test set with an additional training dataset

Approach	Training data	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Person	mAP (%)
Fast R-CNN	VOC+	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	70.0
M-level HFM	VOC+	77.5	79.1	74.5	65.7	61.6	83.0	87.5	87.5	57.4	82.8	71.3	85.7	84.5	78.0	78.8	75.6
M-level HFM	COCO	78.2	78.1	73.4	58.6	65.6	84.0	76.9	83.4	57.8	81.8	61.9	76.3	84.5	81.8	82.9	73.9
M-level HFM	COCO→ VOC+	85.5	85.4	76.7	71.8	68.6	87.9	88.7	88.3	66.4	84.1	74.2	85.1	86.7	83.8	85.8	79.7
L-level HFM	COCO→ VOC+	85.9	84.6	78.2	74.7	70.1	88.4	89.0	88.6	66.7	84.4	76.4	86.1	86.5	84.2	86.0	80.4

All methods use VGG16

Table 3 The object detection leaderboard of the PASCAL VOC 2012 test set with additional training data

Approach	Training data	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Person	mAP (%)
M-level HFM	COCO	85.3	82.0	71.2	56.6	48.4	82.1	79.5	89.2	46.6	78.2	52.5	87.1	83.5	84.4	80.6	71.0
M-level HFM	COCO→VOC+	87.8	85.0	76.2	62.4	60.7	84.0	83.4	89.9	58.1	79.6	60.9	87.3	83.3	85.0	84.6	76.2
L-level HFM	COCO→VOC+	88.8	85.1	76.8	64.8	61.4	85.0	84.1	90.0	59.9	82.6	61.9	88.5	85.2	85.6	86.9	77.5
LocNet [29]		86.3	83.0	76.1	60.8	54.6	79.9	79.0	90.6	54.3	81.6	62.0	89.0	85.7	85.5	82.8	74.8
HRCNN		85.9	83.9	75.5	60.9	54.5	81.4	79.1	90.6	53.3	79.7	61.6	89.9	86.2	85.8	78.2	74.6
MR-CNN [30]		85.5	82.9	76.6	57.8	62.7	79.4	77.2	86.6	55.0	79.1	62.2	87.0	83.4	84.7	78.9	73.9
HyperNet_VGG [31]		84.2	78.5	73.6	55.6	53.7	78.7	79.8	87.7	49.6	74.9	52.1	86.0	81.7	83.3	81.8	71.4
Fast R-CNN & YOLO [20]		83.4	78.5	73.5	55.8	43.4	79.1	73.1	89.4	49.4	75.5	57.0	87.5	80.9	81.0	74.7	70.7

of 10 for 80K iterations. The same procedure described for VOC+ was followed. The COCO baseline achieved mAP of 73.9 % in Table 2, which is lower than the VOC+ baseline because of the domain difference.

4.2.3 COCO→VOC+

The effectiveness of knowledge transfer learning on HFM was verified. Instead of using the same domain to build a category hierarchy, an outside domain was used for prior knowledge. First, FFM_{COCO} was used for the COCO baseline. Second, a category hierarchy was constructed using LTM with FFM_{COCO}. Then VOC hierarchical category was obtained by transferring appearance from the COCO dataset. Finally, FFM_{COCO} was fine-tuned with data $(\mathcal{D}^{VOC+}, \Omega^{VOC+})$ to build the HFM. Fine-tuning options were set at 50K iterations at a learning rate of 0.001, and then, learning rate was decreased by a factor of 10 for 20K iterations. Experiment parameters were the same as those in Sect. 4.1. Table 2 shows that knowledge-transfer learning based on HFM performed better than the baselines for both COCO and VOC+ by 80.4 %.

4.3 VOC 2012 results

In this experiment, detection performance on the VOC 2012 test set is evaluated. For final results on the VOC 2012 dataset, CNN was fine-tuned on the COCO trainval set and a domain adaptation method was conducted on the VOC 2012 trainval set, which adapts the same procedure described in Sect. 4.2. Table 3 compares HFM to the entries in the VOC 2012 leaderboards, using VGG16 as their baseline and additional training data. Even without using domain data, HFM is one of the high-performing detection methods, at 71.0 %. After fine-tuning via the domain adaptation approach and constructing L-level HFM, HFM achieved a 77.5 % mAP, which is the state-of-the-art in the VOC 2012 test results.

5 Conclusion

This paper presented a novel data-driven hierarchical object-detection framework. The framework surpasses the performance of state-of-the-art results on PASCAL VOC 2007 and VOC 2012 datasets. Deep features were partitioned by building a hierarchical deep-feature model HFM via an LTM algorithm. A classifier was assembled at each node of the HFM and constituted HCE. A future research direction is to let go of the optimization problem about HFM structure to determine the optimal hierarchical structure via latent SVM.

Acknowledgments This work was supported by an Inha University research grant. A GPU used in this research was generously donated by NVIDIA Corporation.

References

1. Dong, J., Chen, Q., Feng, J., Jia, K., Huang, Z., Yan, S.: Looking inside category: subcategory-aware object recognition. IEEE Trans. Circuits Syst. Video Technol. **25**(8), 1322–1334 (2015)

2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
3. Song, Z., Chen, Q., Huang, Z., Hua, Y., Yan, S.: Contextualizing object detection and classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1585–1592 (2011)
4. Cinaroglu, I., Bastanlar, Y.: A direct approach for object detection with catadioptric omnidirectional cameras. Signal Image Video Process. **10**(2), 413–420 (2016)
5. Fusek, R., Sojka, E.: Energy transfer features combined with DCT for object detection. Signal Image Video Process. **10**(3), 479–486 (2016)
6. Takarli, F., Aghagolzadeh, A., Seyedarabi, H.: Combination of high-level features with low-level features for detection of pedestrian. Signal Image Video Process. **10**(1), 93–101 (2016)
7. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. In: Proceedings of the IEEE European Conference Computer Vision, pp. 241–254 (2010)
8. Gu, C., Ren, X.: Discriminative mixture-of-templates for viewpoint classification. In: Proceedings of the IEEE European Conference Computer Vision, pp. 408-421 (2010)
9. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: Proceedings of the IEEE European Conference Computer Vision, pp. 168–181 (2010)
10. Malisiewicz, T., Gupta, A., Efros, A. A.: Ensemble of exemplar-svms for object detection and beyond. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 89–96 (2011)
11. Gu, C., Arbelez, P., Lin, Y., Yu, K., Malik, J.: Multi-component models for object detection. In: Proceedings of the IEEE European Conference Computer Vision, pp. 445–458 (2012)
12. Divvala, S.K., Efros, A.A., Hebert, M.: How important are Deformable Parts in the Deformable Parts Model? In: Proceedings of the IEEE European Conference Computer Vision, Workshops and Demonstrations, pp. 31–40 (2012)
13. Zhu, X., Vondrick, C., Ramanan, D., Fowlkes, C.: Do We Need More Training Data or Better Models for Object Detection?. In: BMVC, vol. 3, p. 5 (2012)
14. Aghazadeh, O., Azizpour, H., Sullivan, J., Carlsson, S.: Mixture component identification and learning for visual recognition. In: Proceedings of the IEEE European Conference Computer Vision, pp. 115–128 (2012)
15. Ruan, Z., Wang, G., Xue, J.H., Lin, X.: Subcategory clustering with latent feature alignment and filtering for object detection. Signal Process. Lett. IEEE **22**(2), 244–248 (2015)
16. Ding, K., Huo, C., Xu, Y., Zhong, Z., Pan, C.: Sparse hierarchical clustering for VHR image change detection. Geosci. Remote Sens. Lett. IEEE **12**(3), 577–581 (2015)
17. Yu, X., Yang, J., Lin, Z., Wang, J., Wang, T., Huang, T.: Subcategory-aware object detection. Signal Process. Lett. IEEE **22**(9), 1472–1476 (2015)
18. Zitnick, C. L., Dollr, P.: Edge boxes: locating object proposals from edges. In: Proceedings of the IEEE European Conference Computer Vision, pp. 391–405 (2014)
19. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
20. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
21. Goh, K.S., Chang, E.Y., Li, B.: Using one-class and two-class SVMs for multiclass image annotation. IEEE Trans. Knowl. Data Eng. **17**(10), 1333–1346 (2005)
22. Wang, L., Qiao, Y., Tang, X.: Latent hierarchical model of temporal structure for complex activity classification. IEEE Trans. Image Process. **23**(2), 810–822 (2014)
23. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
24. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
25. Cheng, D., Wang, J., Wei, X., Gong, Y.: Training mixture of weighted SVM for object detection using EM algorithm. Neurocomputing **149**, 473–482 (2015)
26. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. **10**(3), 61–74 (1999)
27. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge 2012 (2012)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Gidaris, S., Komodakis, N.: LocNet: Improving Localization Accuracy for Object Detection. arXiv preprint arXiv:1511.07763 (2015)
30. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware cnn model. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1134–1142 (2015)
31. Kong, T., Yao, A., Chen, Y., Sun, F.: HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. arXiv preprint arXiv:1604.00600 (2016)
32. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: Proceedings of the IEEE European Conference Computer Vision, pp. 340–353 (2012)
33. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L.: Microsoft coco: common objects in context. In: Proceedings of the IEEE European Conference Computer Vision, pp. 740–755 (2014)