Subcategory Clustering with Latent Feature Alignment and Filtering for Object Detection

Zhiwei Ruan, Guijin Wang, Jing-Hao Xue, and Xinggang Lin

***Abstract*—For objects with large appearance variations, it has been proved that their detection performance can be effectively improved by clustering positive training instances into subcategories and learning multi-component models for the subcategories. However, it is not trivial to generate subcategories of high quality, due to the difficulty in measuring the similarity between positive instances. In this letter we propose a new weakly supervised clustering method to achieve better sub-categorization. Our method provides a more precise measurement of the similarity by aligning the positive instances through latent variables and filtering the aligned features. As a better alternative to the initialization step of the latent-SVM algorithm for the learning of the multi-component models, our method can lead to a superior performance gain for object detection. We demonstrate this on various real-world datasets.**

***Index Terms*—Latent-SVM, multi-component models, object detection, subcategory clustering.**

I. Introduction

ONE of the major challenges to object detection is the significant diversity of the positive training instances due to view points, clutter, intra-class variation, etc. An intuitive and effective approach to tackling the challenge is to cluster the positive instances into subcategories, as described in many previous works [1]–[12]. For example, the instances of near-frontal dog faces can be clustered into several subcategories as illustrated in Fig. 1 so that the instances in each subcategory share more similarities. Such subcategories are referred to as components in the multi-component modeling. The components are trained separately with their clustered instances, and representing the instances with the learned models is more accurate than representing them with only a monolithic model.

The premise of constructing multi-component models of high performance is to identify each positive instance with its best fit subcategory. The identification of the positive instances can be achieved through two key steps. The first step is to cluster the positive instances into appropriate subcategories as an initialization [2]–[7]. The second step is to re-identify the instances with the best fit components (i.e. subcategories) via a designed optimization function of model learning [9], [11]. In this letter, we focus on the first key step. This step is important and will directly affect the optimization result in the second step. For the second step of the multi-component models learning, we employ the optimization method described in the latent-SVM algorithm [10].

For the first step, the main challenge is how to robustly and precisely measure the similarity between the positive instances with large appearance variations and clutter. In the recent literature mainly three directions are explored. The first direction is to introduce richer sets of annotations and utilize them for clustering. Gu and Ren [3] segregate instances utilizing viewpoint annotations. Bourdev *et al.* [4] generate models of poselets with the help of keypoint annotations of configuration. Gu *et al.* [5] pick several seed instances and align the rest instances to the seeds based on keypoint and mask annotations to generate clusters. However, the annotation always requires prior knowledge of the object and datasets. The performance will drop with only weak supervision. The second direction is based on the exemplar SVM [8]. The exemplar SVM treats each object instance as a component and trains a linear SVM model for it. Lan *et al.* [6] use the exemplar-SVMs to form the components by the top-scored detections of them. However, the size of the generated component is small and this limits the generalization ability of each component. The computational burden of the algorithm is also heavy for practical use. The third direction is to cluster the instances based on their properties and features that are of more general applicability. The latent-SVM algorithm [10] clusters positive instances on the basis of the aspect ratio of each instance and then splits each cluster bilaterally into the left view and the right view. However, this strategy is invalid for the instances that have similar aspect ratios but large intra-class variations. Park *et al.* [7] and Zeng *et al.* [12] cluster the instances into far- and near-scale subcategories to adapt to different resolutions but it ignores other variations. Divvala *et al.* [2] perform k-means clustering on the warped histogram of oriented gradient (HOG) features [13] of the positive instances. However, warping can be considered as a coarse alignment for the instances and the similarity between the warped instances still cannot be precisely measured. Besides, the similarity measurement on the warped instances is vulnerable with clutter background.

In this letter, we propose a new weakly supervised clustering algorithm to sub-categorize the positive instances and improve the performance of object detection. Our proposed method only

Manuscript received May 05, 2014; revised August 10, 2014; accepted August 14, 2014. Date of publication August 20, 2014; date of current version September 19, 2014. This work was supported in part by the National Natural Science Foundation of China under Grants 61132007 and 61271390. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Giuseppe Scarpa.

Z. Ruan, G. Wang and X. Lin are with the Department of Electronic Engineering, Tsinghua University, Beijing, China (e-mail: rzw09@mails.tsinghua.edu.cn; wangguijin@tsinghua.edu.cn; xglin@tsinghua.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London, U.K. (e-mail: jinghao.xue@ucl.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2349940

Fig. 1. First we group the positive instances that share similarity in appearance into subcategories. Each color indicates a subcategory. Then we learn multi-component models to describe the subcategories.

uses the weak annotations of bounding boxes around the instances and falls in the third direction. Unlike aspect ratio-based clustering [10], our method considers the appearance variation of the instances with similar aspect ratios and carries out deeper clustering. Rather than warping the positive instances to be of a uniform size as done by Divvala *et al.* [2], we seek latent location and scale variables of the instances for alignment. The aligned features are further filtered to remove the impact from the clutter background so that the instances can be better clustered. The clustering algorithm can be used as a better alternative to the initialization step of the latent-SVM algorithm for the learning of the multiple components. We evaluate our algorithm on various datasets and the experiments demonstrate that it can lead to a superior performance gain to state-of-the-art methods.

II. Subcategory Clustering

A. Mixture Components for Object Detection

In order to deal with the variations that can not be tackled by a monolithic model, approaches to learning models for multiple components have been introduced. Each component corresponds to and is learned from a subset of the training instances that share similar properties and features. The learning framework consists of two steps: positive instances initial clustering and multi-component models optimization.

In the initial clustering step, when given the desired number of components m and the positive instances $X_p = \{x_i, b_i\}_{i=1}^n$, where b_i indicates the position and the size of the annotated bounding box around the instance in the sample x_i, each instance needs to be assigned to a component $c_i \in \{1, \ldots, m\}$. In the latent-SVM algorithm [10], positive instances are clustered into k groups based on an aspect ratio heuristic and then each group is split bilaterally into the left view and the right view. This approach is effective for the instances that have big differences in aspect ratio and the number of components $m = 2k$. However, for many datasets, there are still large intra-class variations in the instances that have similar aspect ratios, and these instances still need to be further clustered.

In the multi-component models optimization step, given the clustered positive instances $X_{p,c} = \{(x_i, y_i, b_i, c_i)\}_{i=1}^n$ with label $y_i = 1$ and the negative samples $X_n = \{x'^i, y'^i\}_{i=1}^{n'}$ with label $y'_i = -1$, the target is to learn a set of linear models $\omega = (\omega_1, \ldots, \omega_m)$ to describe the multiple components. In the latent-SVM algorithm, each ω_i is first separately initiated with the corresponding cluster of the positive training data using the standard SVM algorithm. Then the multi-component models are optimized by minimizing the objective function

$$L(\omega) = \sum_{i=1}^{m} \|\omega_i\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i f_\omega^+(x_i, b_i)) + C \sum_{i=1}^{n'} \max(0, 1 - y'_i f_\omega^-(x'_i)), \quad (1)$$

where

$$f_\omega^+(x, b) = \max_{c,z \in Z(x,b)} g_\omega(x, c, z), \quad (2)$$

$$f_\omega^-(x) = \max_{c,z \in Z(x)} g_\omega(x, c, z), \quad (3)$$

$$g_\omega(x, c, z) = \omega_c \cdot \psi(H(x), z), \quad (4)$$

in which $H(x)$ is a pyramid of the features built on differently scaled x, $\psi(H, z)$ is a feature vector extracted from H with a specific latent spatial configuration z that specifies the location and scale level, $Z(x)$ indicates all the spatial configuration candidates in x, and $Z(x, b)$ indicates the candidates overlapping more than 50% with the annotated bounding box b.

B. Feature Alignment with Latent Variables

The aspect ratio-based heuristic clustering algorithm, which was introduced in the latent-SVM, is invalid when the positive instances have similar aspect ratios but large intra-class variation. We first utilize this algorithm as a coarse clustering method to group the instances based on aspect ratios. Then each coarse cluster is further split based on visual appearance.

Appearance feature alignment is important for visual similarity measurement. Rather than warping the appearance features to the same detection window size before clustering [2], we propose a new clustering method utilizing latent variables for feature alignment.

We utilize the multiple-component models optimization method in the latent-SVM learning framework to seek informative latent variables. During the optimization, the best location and scale for feature extraction on each samples, which maximize the scoring function (4), are estimated. This latent variables estimation step can be applied as an appropriate feature alignment approach.

Given the coarse clustering result with m_0 coarse components, we first learn the multi-component models $\omega_0 = (\omega_{0,1}, \ldots, \omega_{0,m_0})$ as described in subsection II-A. Then, for each positive instance x_i, the re-identified component membership c_i^r and the latent variable z_i that specify the location and scale can be obtained through

$$(c_i^r, z_i) = \operatorname*{argmax}_{c,z \in Z(x_i,b_i)} g_\omega(x_i, c, z). \quad (5)$$

With the estimated latent variables, an aligned feature set Ψ_j corresponding to the jth coarse component can be presented as

$$\Psi_j = \{\psi_a(x_i) | c_i^r = j, 1 \le i \le n\}, j = 1, \ldots, m_0, \quad (6)$$

where

$$\psi_a(x_i) = \psi(H(x_i), z_i). \tag{7}$$

These aligned instances in each Ψ_j can be further clustered into k_j fine subcategories as previously determined and the total number of the fine components $m = \sum_{j=1}^{m_0} k_j$. With the help of feature alignment, the similarity between instances can be more accurately estimated so the instances can be better clustered and the learned fine multi-component models can describe the object more precisely.

Here we only need the alignment of global features of objects for clustering. Hence, only global templates are trained for the coarse components, which are described as root filters in the latent-SVM algorithm.

C. Feature Filtering for Clustering

Although each instance is aligned via its latent variables as in subsection II-B, its appearance can still be affected by the foreground and background clutter. The irrelevant clutter needs to be disregarded such that the similarity between instances can be measured more robustly.

We utilize the weights of the learned linear models which represent the coarse components as a measure of the clutter degrees. Each weight in the linear model stands for the importance of its corresponding feature element in representing the object. Large clutter in a feature element usually depresses its weight. Meanwhile, with the HOG features employed, the negative weights of the learned DPM are more likely associated with negative samples. Therefore, we are motivated to disregard the feature elements with small or negative weights in the clustering of positive instances.

Let $\omega_{0,j} = (w_{j,1}, \ldots, w_{j,l_j})$ denote the learned linear model of the jth coarse component and $\psi_a(x_i) = (f_1, \ldots, f_{l_j})$ denote an aligned feature vector assigned to the jth coarse component, where l_j indicates the length of the vector which is determined by the size of the learned model $\omega_{0,j}$. Likewise, the linear models of the coarse components are global templates. The filtered feature vector can be represented as

$$\psi_f(x_i) = (\varepsilon(w_{j,1}) \cdot f_1, \ldots, \varepsilon(w_{j,l_j}) \cdot f_{l_j}), \tag{8}$$

in which $\varepsilon(\cdot)$ is the step function at a threshold. Here for simplicity we set the threshold to zero, and thus the operation in (8) filters out some irrelevant features that are more prone to clutter and negative samples. Consequently, by filtering out these feature elements, the clustering results can be less affected by the background and negative samples, and the performance of the learned models can be enhanced.

In this letter, we aim to stress the importance of feature alignment and filtering. After the features are processed, any clustering algorithm can be applied. Here we employ k-means clustering to identify the fine components. After that, we follow the multi-component models optimization step of the latent-SVM algorithm to learn the fine multiple components as described in subsection II-A for object detection. Besides the global templates, local templates described as part filters [10] are also used in the detector to best represent the object.

III. Experiments

A. Experimental Settings

We evaluate our proposed clustering method on two typical datasets with different variation factors.

One dataset is the PASCAL VOC 2007 dataset [14]. There are 20 object categories in the dataset and each category has great variations in both aspect ratio and appearance.

The other is an annotated dataset of near-frontal dog faces [15]. All the aspect ratios of the annotation boxes equal one. Hence, we mainly focus on the appearance variations in this dataset. For evaluation, we pick 3400 samples from 25 species in the dataset. The dataset of each species is divided into a training subset and a test subset. All the training (test) subsets are gathered together as a training (test) set for the multi-component models that address a generic dog-face detection task. The 25 species can be divided into six groups, which are terrier-like, chihuahua-like, shepherd-like, spitz-like, beagle-like and spaniel-like. With the group labels, we can use the cluster purity [9], [16] to measure the clustering results.

To compare the performance of different clustering methods, multi-component models are trained with the clustering results and we use the detection performance of the learned models as a measurement. The positive samples are left-right flipped and we train two coarse components of the left and right views for each aspect ratio-based cluster. The appearance-based clustering methods will perform further clustering based on the coarse components. The negative samples come from the object-free images in the PASCAL VOC 2007 dataset [14]. Each final component model learned by the latent-SVM algorithm has one root filter and eight part filters as default for object detection.

For object description, features based on the HOG descriptors [13] are employed, as with [10]. We use the average precision (AP) as in the PASCAL toolkit to evaluate the performance of detectors. A detection result is regarded correct if the area of its intersection with the ground truth covers more than 50% of their union.

B. Experimental Results

In the experiment on PASCAL VOC 2007 dataset, we aim to investigate how the clustering methods affect the performance of the object detection on various object categories with large variations in aspect ratio and appearance. We use the release 4 version of the DPM detector (vocR4) [17] as the baseline. With three aspect ratio-based clusters and two flipped components for each cluster, the DPM detector has six components as default. In this experiment, we compare six methods with respect to the baseline without any post-processing (contextual rescoring or bounding box prediction): 1) 'Warp' warps the instances into a uniform size and performs k-means clustering to generate subcategories [2]. It has 15 components as the optimal set for this dataset; its model learning is based on the release 4 version of the DPM; 2) 'vocR5' is the release 5 version of the DPM detector [18] with an improved optimization method; 3) 'AFC12' is our new method which both aligns and filters the instances before performing fine clustering. We utilize the six coarse components learned by vocR4 and learn two fine components for

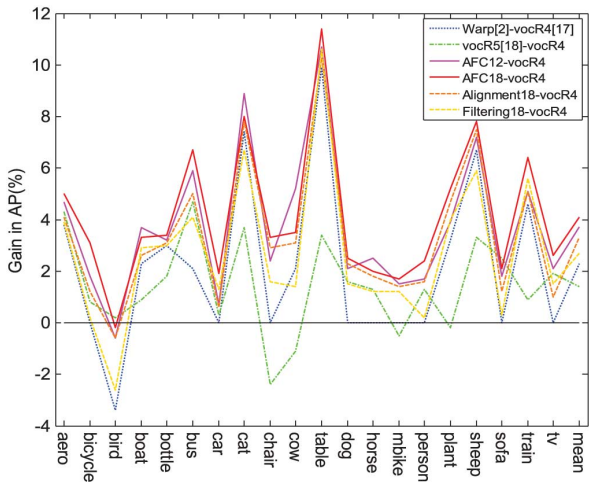

Fig. 2. Gains in AP(%) of different learning methods with respect to the baseline [17] on 20 VOC object categories.

each coarse component. Hence we have 12 final components, learned by using the optimization method of vocR4; 4) Similar to 'AFC12', 'AFC18' is our new method which learns three fine components for each coarse component and has 18 final components; 5) 'Alignment18' is our method which only aligns the instances without filtering, with 18 final components; and 6) 'Filtering18' is our method which filters the warped features in each coarse component without alignment, and it generates 18 final components.

The comparative results are illustrated in Fig. 2, from which we can observe that our proposed method 'AFC' performs better than any other methods on 18 of the 20 categories, either with 12 or 18 components. On average, our 'AFC18' gets 4% gains in AP over the baseline while 'Warp' gets 2% gains. This demonstrates the benefits of feature alignment and filtering for subcategory clustering. Our method also outperforms the 'vocR5', indicating the effectiveness of subcategory clustering for improving the object detection performance. With only 'Alignment', the learned components still outperform the models learned by 'Warp'; this indicates the advantage of feature alignment over instance warping. With only 'Filtering', the performance of the learned components is close to that of the models learned by 'Warp'; this shows that the filters for the coarse components cannot always match the warped features. In contrast, with the aligned features, the filtering process can achieve a much higher performance gain.

In the experiment on the near-frontal dog face dataset, all the aspect ratios of the annotation boxes equal one. Hence, aspect ratio-based clustering methods can only get one cluster with two coarse components for the left-right views. The appearance-based clustering methods can further generate fine components. In this experiment, the selected dog species can be broadly grouped into six categories and we aim to investigate whether the appearance-based clustering methods can generate meaningful components under such conditions. Hence, each coarse component is further clustered into six subcategories; that is, except the 'vocR4' method with two final components, both 'Warp' and 'AFC' have 12 final fine components. Here we

TABLE I
AP AND CLUSTER PURITY FOR THE DOG FACE DATASET

Method	vocR4 [17]	Warp [2]	AFC
AP (%)	90.2	91.8	**95.7**
Purity (%)	19.7	58.1	**64.5**

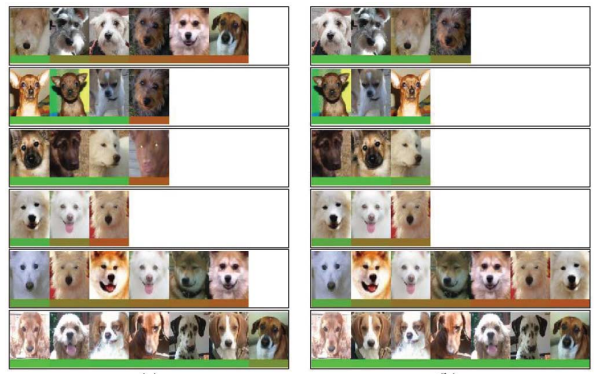

Fig. 3. Six dog subcategories clustered by (a) warping instances and (b) our proposed method. Each row represents a subcategory. Each exemplar stands for the species that has samples identified with the subcategory, and the color bar below it indicates the percentage of the species' samples identified, which ranges from 100% (green) to 0% (red). Only the species with the identification percentages above 30% are shown. (a) By 'Warp'. (b) By 'Alignment+Filtering'.

compare AP and the cluster purity of three methods: 'vocR4' [17], 'Warp' [2] and our method 'AFC'. The results are shown in Table I and our proposal 'AFC' performs the best in both AP and the cluster purity.

Moreover, for visual comparison, Fig. 3 displays the subcategory clustering results in units of species. Compared with the subcategories clustered by warping instances, our proposed method generates subcategories with higher purity especially for the first three subcategories. The first subcategory is mainly composed of the terrier-like species. Our proposed method admits most of the terrier-like species' samples for the first subcategory while the 'Warp' method admits three terrier-like species with low percentages (in red in Fig. 3) and also admits some other species. For the second (chihuahua-like) and third subcategories (shepherd-like), our method also avoids interference from other dissimilar species.

In summary, the experiments of both the PASCAL VOC object categories detection and the near-frontal dog face detection demonstrate the benefits of using our proposed clustering method for object detection.

IV. CONCLUSIONS

We have proposed a new subcategory clustering method for object detection. Our method categorizes the instances by latent feature alignment and filtering. Our experiments demonstrated that this method can lead to a superior performance gain than state-of-the-art methods. Our main future work is to investigate how to design a deeper hierarchical model for better clustering and detector learning.

REFERENCES

[1] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes, "Do we need more training data or better models for object detection?," in *BMVC 2012*, 2012, pp. 1–11.

[2] S. K. Divvala, A. A. Efros, and M. Hebert, "How important are deformable parts in the deformable parts model?," in *ECCV 2012*, 2012, pp. 31–40.

[3] C. Gu and X. Ren, "Discriminative mixture-of-templates for viewpoint classification," in *ECCV 2010*, 2010, pp. 408–421.

[4] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *ECCV 2010*, 2010, pp. 168–181.

[5] C. Gu, P. Arbeláez, Y. Lin, K. Yu, and J. Malik, "Multi-component models for object detection," in *ECCV 2012*, 2012, pp. 445–458.

[6] T. Lan, M. Raptis, L. Sigal, and G. Mori, "From subcategories to visual composites: A multi-level framework for object detection," in *ICCV 2013*, 2013, pp. 369–376.

[7] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *ECCV 2010*, 2010, pp. 241–254.

[8] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *ICCV 2011*, 2011, pp. 89–96.

[9] M. Hoai and A. Zisserman, "Discriminative sub-categorization," in *CVPR 2013*, 2013, pp. 1666–1673.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[11] G.-T. Zhou, T. Lan, A. Vahdat, and G. Mori, "Latent maximum margin clustering," *Adv. Neural Inf. Process. Syst.*, pp. 28–36, 2013.

[12] B. Zeng, G. Wang, X. Lin, and C. Liu, "A real-time human detection system for video," *IEICE Trans. Inf. Syst.*, vol. 95-D, no. 7, pp. 1979–1988, 2012.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005*, 2005, pp. 886–893.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge 2007(VOC2007) results," [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[15] Z. Ruan, G. Wang, J.-H. Xue, X. Lin, and Y. Jiang, "Detection of user-registered dog faces," *Neurocomputing*, vol. 142, pp. 256–266, 2014.

[16] M. Meilă, "Comparing clusterings–an information based distance," *J. Multivar. Anal.*, vol. 98, no. 5, pp. 873–895, 2007.

[17] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively trained deformable part models, release 4," [Online]. Available: http://people.cs.uchicago.edu/pff/latent-release4/

[18] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," [Online]. Available: http://people.cs.uchicago.edu/ rbg/latent-release5/