IEICE TRANS. INF. & SYST., VOL.E97–D, NO.5 MAY 2014

LETTER

Deformable Part-Based Model Transfer for Object Detection

Zhiwei RUAN[†a)], ***Nonmember***, **Guijin WANG**[†b)], ***Member***, **Xinggang LIN**[†], **Jing-Hao XUE**[††], ***and*** **Yong JIANG**[†††c)], ***Nonmembers***

SUMMARY The transfer of prior knowledge from source domains can improve the performance of learning when the training data in a target domain are insufficient. In this paper we propose a new strategy to transfer deformable part models (DPMs) for object detection, using offline-trained auxiliary DPMs of similar categories as source models to improve the performance of the target object detector. A DPM presents an object by using a root filter and several part filters. We use these filters of the auxiliary DPMs as prior knowledge and adapt the filters to the target object. With a latent transfer learning method, appropriate local features are extracted for the transfer of part filters. Our experiments demonstrate that this strategy can lead to a detector superior to some state-of-the-art methods.

key words: *deformable part model, object detection, transfer learning*

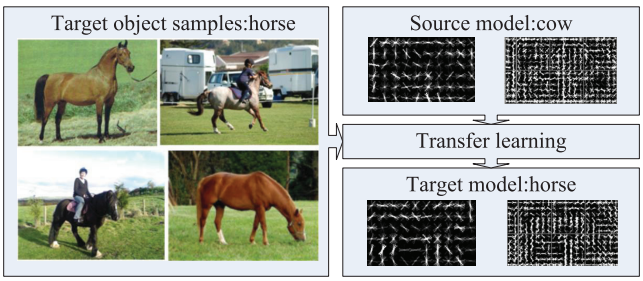


Fig. 1 Transfer learning of the DPM. The learnt DPM for a cow is used as the source model for learning a horse detector with few training samples.

1. Introduction

Object category detection is one of the most important tasks in computer vision, widely applied in many fields such as security monitoring and intelligent navigation. Most state-of-the-art approaches learn a binary object/non-object classifier trained with positive and negative samples. Recently, the most successful approach in the PASCAL Visual Object Class (VOC) challenge [1] has been the deformable part model (DPM) proposed by Felzenszwalb et al. [2]. However, it demands a large number of training samples from the domain of the target objects to be detected, requiring a substantial effort to collect and annotate data in advance. Its performance drops dramatically when there is only a limited number of training samples available in practice.

Such a demand for substantial training samples from the target domain can be lessened, if we can transfer some knowledge from other domains called source domains. The performance of the detector can also be improved in this case [3]–[8]. Transfer learning approaches to object detection differ in the form of the knowledge transferred, which ranges from similar samples [3], [9], locally structured priors [4], visual attributes [5] to sharable part templates [6]–[8]. Among them, sharing part templates can make more use of the previously trained auxiliary models and make transfer learning more flexible. Aytar et al. [6] use part-like patches from previously learnt models to enhance exemplar SVMs. However, the algorithm is only executed on rigid templates, and this limits the performance of the detector. Ott et al. [7] extend the DPM to share part filters among multiple models. It can make efficient use of training data by learning shared part filters from different object classes, but the training processes for different classes are integrated hence a new detector cannot be learnt directly based on previously trained models. Yang et al. [8] assemble a new detector with root and part filters from auxiliary detectors. However, the part filters are used without adjusting parameters, and this may suppress its performance.

In this paper, we propose a new DPM-based transfer learning method for object detection, using both the root and part filters as prior knowledge to transfer. We transfer deformable part filters, rather than only transferring rigid templates as done by Aytar et al. [6]. Unlike Yang et al. [8], our method also adjusts the parameters of the part filters. We also design a new regularisation formulation for training a discriminative latent SVM, in which we not only share the root and part filters between models but also adapt the parameters of the filters to the target object. As illustrated in Fig. 1, a detector of a target category can be learnt using some source models, with few training samples. Our experiments demonstrated that the new detector is superior to some state-of-the-art methods.

2. Model Transfer of DPM

Suppose we need to train a detector for a target category which only has limited positive training samples, and we

Manuscript received October 7, 2013.
Manuscript revised January 20, 2014.

[†]The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
[††]The author is with the Department of Statistical Science, University College London, London, WC1E 6BT, UK.
[†††]The author is with Canon Information Technology (Beijing) Co., LTD, China.
a) E-mail: rzw09@mails.tsinghua.edu.cn
b) E-mail: wangguijin@tsinghua.edu.cn
c) E-mail: jiangyong@canon-ib.com.cn

DOI: 10.1587/transinf.E97.D.1394

Copyright © 2014 The Institute of Electronics, Information and Communication Engineers

have several detectors for visually similar source categories trained with sufficient training samples. We aim to transfer knowledge from the source categories to the target category to learn a good detector of target objects.

2.1 Transfer of Rigid Model

An original idea introduced by Aytar et al. [10] is to regularise the training of the target object's rigid template with the previously learnt source rigid template. They further this method to use patches of various source rigid templates [6].

Given n labelled examples $X = \{x_i, y_i\}_{i=1}^{n}$, where $y_i \in \{-1, 1\}$, and m global template models $\{\omega_i^s\}_{i=1}^{m}$ learnt in a source domain, the task is to learn a linear target model ω with a scoring function $f_\omega(x) = \omega \cdot \psi(x)$, where $\psi(x)$ is the feature vector. Let $\{u_i\}_{i=1}^{m'}$ denote patches matching the target object from the source models and padded with zeros to be a template of the same size as ω. The target model can be trained by minimising an objective function

$$L(\omega, \alpha) = \|\omega - \sum_{i=1}^{m'} \alpha_i u_i\|^2 + \gamma \sum_{i=1}^{m'} \alpha_i^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i f_\omega(x_i)) , \tag{1}$$

where α_is are the transfer weights for u_i, $\max(0, 1-y_i f_\omega(x_i))$ is the standard hinge loss, and parameters γ and C control the relative weights of the regularisation term and the loss term, respectively. The patches borrowed from well trained source templates can bring more discriminative power to enhance the target detector. However, this algorithm only transfer rigid templates, and this limits its performance.

Here we propose a learning method to transfer the deformable part filters, which are better than the rigid patches of global templates for learning the part filters of the target.

2.2 Deformable Part Model (DPM)

A DPM consists of one root filter and several part filters [2]. The root filter describes the global appearance of an object; the part filters represent the local appearance with twice the resolution of the feature in the root filter and have flexible positions relative to the root filter.

For a model with P parts, each example is scored as

$$f_\omega(x) = \max_z g_\omega(x, z) = \max_z \{\omega \cdot \psi(H, z)\} , \tag{2}$$

where

$$\omega = (F_0, \ldots, F_P, d_1, \ldots, d_P, b) , \tag{3}$$

$$\psi(H, z) = (\phi_a(H, z_0), \ldots, \phi_a(H, z_P) , -\phi_d(dz_1), \ldots, -\phi_d(dz_P), 1) , \tag{4}$$

in which ω is a vector of filter parameters, F_0 is a root filter, F_i is the ith part filter, d_i is a four-dimensional vector that specifies the coefficients of a quadratic function incurring a penalty to moving the ith part too far from its supposed location v_i, b is a bias term, and $\psi(H, z)$ is a feature vector extracted from the feature pyramid H with a specific latent spatial configuration $z = (z_0, z_1, \ldots, z_P)$, where $z_i = (z_{xi}, z_{yi}, l_i)$ specifies the location and scale level of the ith part filter, $\phi_a(H, z_i)$ are local appearance features which describe the image area covered by the ith part, and $\phi_d(dz_i) = (dz_{xi}, dz_{yi}, dz_{xi}^2, dz_{yi}^2)$ are deformation features with $(dz_{xi}, dz_{yi}) = (z_{xi}, z_{yi}) - (2(z_{x0}, z_{y0}) + v_i)$. The model parameters can be learnt in a latent SVM framework.

Such a deformable configuration is fully exploited by our DPM-based transfer learning method, as below.

2.3 Transfer of DPM

Given m deformable part models $\{\omega_i^s\}_{i=1}^{m}$ learnt from a source domain, where the ith model ω_i^s has P_i parts, our task is to learn a new DPM ω for the target domain.

The models learnt from similar source categories are specified with the same root-filter size and the same part-filter size, such that the filters are easy to be shared between different categories. We transfer the filters while omitting d_is from the auxiliary models. If all the part filters are assumed independent, the number of the part filters in the target model equals $P = \sum_{i=1}^{m} P_i$. As ω for a DPM has the form of (3), each filter F_{p_i} (either a root or part filter) in ω_i^s that corresponds to filter F_p in ω can be rewritten in a sparse form $u = (0, \ldots, 0, F_{p_i}, 0, \ldots, 0)$ to match the form of ω and the position of F_p in ω.

In this way, we can convert the set of DPMs $\{\omega_i^s\}_{i=1}^{m}$ into a set of auxiliary filters $\{u_i\}_{i=1}^{m'}$ with $m' = m + \sum_{i=1}^{m} P_i$ (i.e. m root filters and $\sum_{i=1}^{m} P_i$ part filters). The ω is then trained by minimising an objective function

$$L_{Z_P}(\omega, \alpha) = \|\omega - \sum_{i=1}^{m'} \alpha_i u_i\|^2 + \gamma \sum_{i=1}^{m'} \bar{\alpha}_i{}^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i g_w(x_i, z_i)) , \tag{5}$$

where

$$\bar{\alpha}_i = \alpha_i \|u_i\| , \tag{6}$$

$$Z_P = \left\{ z_i, i = 1, \ldots, n \,|\, z_i = \operatorname*{argmax}_z g_w(x_i, z) \right\} . \tag{7}$$

The first term in (5) represents the distance between the learned model ω and the assembled filters. The second term offers re-weighting of the auxiliary filters. In [6] the patches from root filters are treated indiscriminately as in (1), but here we consider the different sizes of the root and part filters in the DPM and thus let $\bar{\alpha}_i$ be u_i-adaptive as in (6).

As $\gamma \to \infty$, α_is will be forced to be zero because of the infinite penalisation, and (5) will converge to the standard latent-SVM that transfers no knowledge from the source domain. As $\gamma \to 0$ and $C \to 0$, ω will be forced to be equal to $\sum_{i=1}^{m'} \alpha_i u_i$ and the learnt model is a linear combination of auxiliary filters. The set Z_P includes the highest latent scores for the positive examples. With latent values z_is included, our method is more adaptive to the diversity of the object

than the method in [6].

2.4 Optimisation Algorithm

The objective function (5) can be transformed to an equivalent latent-SVM formulation:

$$\begin{aligned} &\|\omega - \sum_{i=1}^{m'} \alpha_i u_i\|^2 + \gamma \sum_{i=1}^{m'} \bar{\alpha}_i{}^2 \\ &\quad + C \sum_{i=1}^{n} \max(0, 1 - y_i(\omega \cdot \psi(H_i, z_i))) \qquad (8) \\ &= \|\bar{\omega}\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i((\bar{\omega} \cdot \bar{\psi}_i))) \,, \end{aligned}$$

where

$$\bar{\omega} = \{\triangle\omega, \sqrt{\gamma}\bar{\alpha}_1, \ldots, \sqrt{\gamma}\bar{\alpha}_{m'}\} \;, \triangle\omega = \omega - \sum_{i=1}^{m'} \alpha_i u_i \;, \quad (9)$$

$$\begin{aligned} \bar{\psi}_i = \Bigg\{ &\psi(H_i, z_i), \frac{1}{\sqrt{\gamma}\|u_1\|} u_1 \cdot \psi(H_i, z_i), \ldots, \\ &\frac{1}{\sqrt{\gamma}\|u_{m'}\|} u_{m'} \cdot \psi(H_i, z_i) \Bigg\} \;. \qquad (10) \end{aligned}$$

With latent values Z_P, (5) leads to a semi-convex optimisation problem, which can be minimised with an Expectation Maximization-like iterative process of two steps:

1. Latent value estimation: select the highest latent scores for each positive example by fixing ω and α.

2. Parameter optimisation: minimise the objective function in (5) over ω and α with selected latent values Z_P. This can be solved by minimising (8) over $\bar{\omega}$ with a stochastic gradient descent method as described in [2]. By using the parameters $\triangle\omega$ and α_is in $\bar{\omega}$, we can get $\omega = \triangle\omega + \sum_{i=1}^{m'} \alpha_i u_i$.

In short, Step 1 searches over the space of latent spatial values, and Step 2 searches over the space of model parameters, until convergence is reached.

3. Experiments

For performance evaluation, we conduct two types of experiments as in [10]: (i) *between-category transfer* where transfer is conducted from a similar class to the target class (e.g. from cow to horse); and (ii) *superior-to-subordinate transfer* where transfer is conducted from the superior class to the target subclass (e.g. from quadruped to horse).

3.1 Experimental Settings

In both cases, evaluation is performed on the PASCAL VOC 2007 dataset [1], and we restrict it to the side-views of the categories that share similar appearances, such as horse vs cow and motorbike vs bicycle. We select the positive training examples on the basis of the view attributes annotated and the aspects of the bounding boxes around the objects. For testing, we build a test set which consists of all the positive samples of the side-view target class and a random collection of up to 500 samples from other classes.

In the training stage, we learn two components as mirror of each other (e.g. right view and left view for horse) for each detector, as the viewpoints of the objects in the selected samples do not vary much. Each DPM has one root filter and eight part filters. We use features based on the histogram of oriented gradient (HOG) descriptors [11] as with [2]. The HOG is calculated from a dense grid of cells and normalised with overlapping local contrast.

In the test stage, the detection performance is evaluated on the test set by using the average precision (AP) as in the PASCAL toolkit. A detection result is considered correct if the area of its intersection with the ground truth is greater than 50% of their union. For evaluation we only randomly select ten positive examples for transfer learning or direct DPM training. We repeat experiments five times and take their average AP. In all experiments, parameter C is fixed to 0.002 as the default value of [2] and γ is simply set to 1.

3.2 Experimental Results

Regarding the *between-category transfer*, we aim to train a horse detector by transferring from a cow detector, and train a bicycle detector by transferring from a motorbike detector. We compare four methods: 1) 'Direct training' learns the DPM from a limited number of positive target samples without transfer; 2) 'RT-TRANS' learns a target rigid template by transferring part-like patches from the rigid templates of source models [6]; 3) 'ASB-TRANS' transfers the root filters but only assembles without adapting the part filters of source models [8]; this can be done by modifying the first term in (5) to constrain the parameters adaption only on the root filter; and 4) 'DPM-TRANS' is our new method of transferring and adapting entire DPM models.

The results are listed in Table 1. For both target classes (horse and bicycle), our new transfer learning approach performs the best. For example, for the horse-cow case, our method (57%) outperforms both the source models (26%) and the model directly trained on the given positive samples only (48%); this indicates the benefit from transfer learning. Our method performs dramatically better than RT-TRANS (38%), a method transferring only root filters, which suggests the importance of transferring part filters. Then the superiority of our method to ASB-Trans (47%), a method assembling part filters but not adjusting them, demonstrates the importance of adjusting the part filters.

The DPMs learnt from different methods are illustrated in Fig. 2, from which we can make the following three observations. 1) If we directly train a DPM from only a limited number of target samples (Fig. 2 (a)), the resulting model

Table 1 AP (%) of different methods for between-category transfer.

Method	horse (cow: 26.3)	bicycle (motorbike: 52.5)
Direct training	47.6 ± 6.7	72.6 ± 3.7
RT-TRANS [6]	38.3 ± 9.7	69.1 ± 2.3
ASB-TRANS [8]	46.8 ± 8.6	61.3 ± 4.4
DPM-TRANS	**56.5 ± 6.0**	**76.7 ± 3.3**

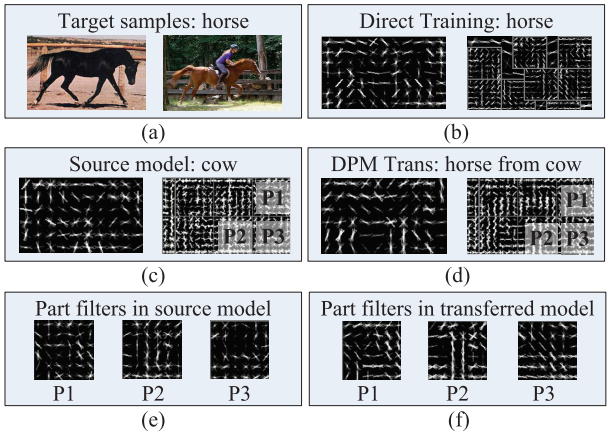


Fig. 2 Different DPMs in the horse-from-cow transfer.

Table 2 AP (%) of different methods for superior-to-subordinate transfer.

Method	horse (quadruped: 65.1)	cow (quadruped: 56.8)	sheep (quadruped: 38.7)
Direct training	47.6 ± 6.7	39.8 ± 8.4	36.3 ± 3.4
RT-TRANS [6]	40.3 ± 8.7	37.5 ± 9.1	31.3 ± 3.3
ASB-TRANS [8]	65.0 ± 8.4	57.1 ± 8.0	39.2 ± 3.8
DPM-TRANS	**69.6 ± 6.0**	**58.7 ± 7.3**	**40.8 ± 2.9**

(Fig. 2 (b)) looks very poor for the target object. The shape of the object in the root filter is distorted and obscure. Some part filters deviate from their supposed positions, and the captured local features are over-fitting to the limited samples. 2) With the help of the prior knowledge from the similar source model (the cow model in Fig. 2 (c)), the transfer learnt model in Fig. 2 (d) has a well-shaped root filter and well-located part filters. 3) The head filter (P1) and two foreleg filters (P2, P3) in Fig. 2 (f) are more horse-shaped than the corresponding source filters in Fig. 2 (e). This demonstrates that the parameters adaption in the proposed method help the transfer learnt filters to fit the target object better.

The *superior-to-subordinate transfer* experiments are performed on the horse, cow and sheep categories. The superior category 'quadruped' is defined as a union of the three categories. The model of quadruped is trained from 100 randomly selected side-view samples of the three categories. Then we aim to train a specific subordinate detector by transferring from the quadruped detector.

The evaluation results are presented in Table 2. For all the three subordinate categories, our approach performs the best and also consistently better than the superior category detector. It indicates that a generic detector can be further adapted to a specific subordinate by the proposed method.

In summary, the experiments for both the *between-category transfer* and *superior-to-subordinate transfer* clearly demonstrate the benefit of performing the DPM-based transfer learning for object detection.

4. Conclusions

We have proposed a new DPM-based transfer learning method for object detection. Our method can transfer and adapt both the root and part filters to the target samples. Our experiments demonstrated that this strategy can lead to a detector superior to some state-of-the-art methods.

Our main future work is to investigate how to select appropriate auxiliary models to further enhance our method.

Acknowledgment

The work is partially sponsored by National Natural Science Foundation of China (No.61132007 and No.61271390).

References

[1] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results." http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

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

[3] M. Wang, W. Li, and X. Wang, "Transferring a generic pedestrian detector towards specific scenes," 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3274–3281, 2012.

[4] T. Gao, M. Stark, and D. Koller, "What makes a good detector?–Structured priors for learning from few examples," ECCV 2012, pp.354–367, 2012.

[5] C.H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.951–958, 2009.

[6] Y. Aytar and A. Zisserman, "Enhancing exemplar SVMs using part level transfer regularization," BMVC, pp.1–11, 2012.

[7] P. Ott and M. Everingham, "Shared parts for deformable part-based models," 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1513–1520, 2011.

[8] K. Yang, M. Wang, X.S. Hua, S. Yan, and H.J. Zhang, "Assemble new object detector with few examples," IEEE Trans. Image Process., vol.20, no.12, pp.3341–3349, 2011.

[9] B. Zeng, G. Wang, Z. Ruan, and X. Lin, "Data level object detector adaptation with online multiple instance samples," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1397–1400, IEEE, 2012.

[10] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," 2011 IEEE International Conference on Computer Vision (ICCV), pp.2252–2259, 2011.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.886–893, 2005.