Neurocomputing 142 (2014) 256–266

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Detection of user-registered dog faces

Zhiwei Ruan [a], Guijin Wang [a,*], Jing-Hao Xue [b], Xinggang Lin [a], Yong Jiang [c]

[a] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[b] Department of Statistical Science, University College London, London WC1E 6BT, UK
[c] Canon Information Technology (Beijing) Co., LTD, China

ARTICLE INFO

Article history:
Received 14 January 2014
Received in revised form
12 March 2014
Accepted 24 March 2014
Communicated by M. Wang
Available online 27 May 2014

Keywords:
Deformable part-based model
Dog faces detection
Object detection
User-registered detection

ABSTRACT

Dog face detection is an important object detection task, widely applied in many fields such as auto-focus and image retrieval. In many applications, users only care about specific target species, which are unknown to a detection system until the users register some relevant information like a limited number of target samples. We call this scenario the detection of user-registered dog faces. Due to the great variation between different dog species, no single model can describe all the species well. Meanwhile, it is also impractical to learn individual models for every potential target species that the users may care about, given the large number of dog species. Furthermore, the registered samples are usually too few to train a robust detector directly. In this context, we propose a novel user-registered object detection framework. This framework can generate an adaptive detector, from only a limited number of user-registered target samples and a couple of off-line trained auxiliary models. In addition, we build an annotated dog face dataset, which contains 10,712 images of 32 species. Experimental results on the dataset demonstrate that the proposed framework can achieve superior detection performance to the state-of-the-art approaches.

1. Introduction

Animal detection is a hot topic in the field of object detection. It has been widely applied in auto-focus, image retrieval, multimedia contents analysis, etc. [1–4]. In animal detection, dog face detection is an interesting and especially challenging task.

One of the greatest challenges to dog face detection is that there are so many dog species. The appearance of dog faces has great diversity between different species, from long-nosed dogs to short-nosed dogs, from shaggy dogs to smooth-haired dogs, etc. Hence a single model will surely not be robust enough to the variety of dog breed.

Meanwhile, it is impractical in many application scenarios to train individual models for every dog species, given the large number of species. There will also be huge costs for the sample collection, model training and computational storage. Therefore, a practically reasonable approach is to create and model a few superordinate groups which represent groups of similar species, and then offer the corresponding superordinate models to users.

On the other hand, in many application scenarios, a user only wants to detect a specific dog species that they care about. For example, a user comes across some pictures of a specific species and she wants to find from some large datasets more images that contain the specific species. Sometimes the user even cannot tell the species name and what she can provide to the system is only to register a small number of samples of the target species. In this case, unlike traditional object detection tasks, the main issue here is that the target species are unknown to the system until the users register a few target samples.

For these unknown species, the information extracted from the registered samples is the most relevant, but the number of the provided samples can be too limited and insufficient to directly train a robust detector.

Therefore, we propose a new framework for the detection of user-registered dog faces. The framework can generate an adaptive detector from a limited number of user-registered target samples and a couple of off-line trained auxiliary superordinate models. The framework is illustrated in Fig. 1.

Our paper makes the following three main contributions. (1) We propose an novel framework to detect user-registered dog faces. The framework can generate a detector adaptive to users' demand. (2) The framework can combine the knowledge from both the off-line trained auxiliary models and the registered samples of the target species for the detection. (3) We design a strategy-selection algorithm to automatically determine when and how to appropriately utilize the auxiliary models and the registered target samples.

* Corresponding author.
E-mail addresses: rzw09@mails.tsinghua.edu.cn (Z. Ruan), wangguijin@tsinghua.edu.cn (G. Wang), jinghao.xue@ucl.ac.uk (J.-H. Xue), xglin@tsinghua.edu.cn (X. Lin), jiangyong@canon-ib.com.cn (Y. Jiang).

http://dx.doi.org/10.1016/j.neucom.2014.03.058
0925-2312/

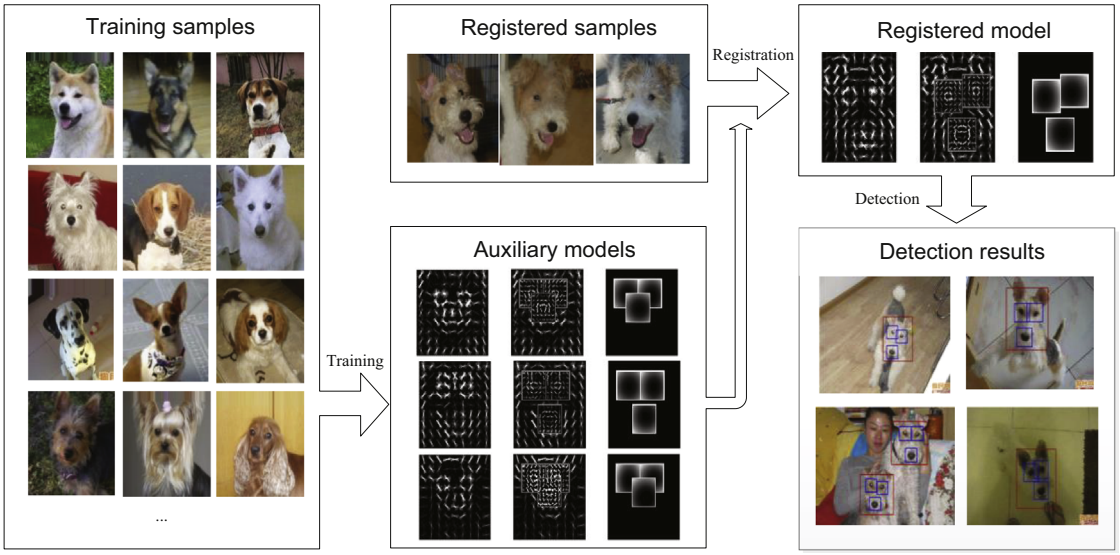


Fig. 1. User-registered detection. An adaptive detector is generated from the user-registered samples and the off-line trained auxiliary models.

In addition, we built an annotated dataset of near-frontal dog faces, which contains 10,712 images of 32 species. The dog images were collected from a web site where dog owners can upload the species information and images of their dogs. Experimental results on the dataset demonstrate that the proposed framework is superior to the state-of-the-art methods. Although our research focuses on the detection of user-registered dog faces, the proposed framework can be extended readily to other tasks of animal detection or more generic object detection.

2. Related work

In our proposed framework, we exploit both off-line trained auxiliary models and user-registered samples to generate an adaptive detector of dog faces. Hence, the works most relevant to ours are the off-line training methods and the adaptive model learning methods for object detection.

Most of the recent work on animal detection [1–4] focuses on the off-line training. Kozakaya et al. [1] cascade a coarse model trained with AdaBoost [5] and a fine model trained with a linear SVM classifier. Zhang et al. [3] design a joint detection algorithm with two global templates, one to describe the shape and the other to describe the texture. Both of these approaches employ global templates and are not robust enough for deformable objects. Aytar and Zisserman [6] and Azizpour and Laptev [4] build their algorithms on the deformable part model (DPM) [7] and get better performance on highly deformable objects such as dog faces.

In the recent literature, to generate an adaptive model with only a limited number of samples of the target object, mainly three directions are explored. The first direction is to employ online learning [8–10], which updates the model continuously as new samples arrive. This approach can only adjust the off-line trained model gradually and only be implemented on a single model. The second direction is to perform sample selection and re-weighting [11–13] for the off-line training. Samples similar to the registered samples are selected from the dataset previously used by the off-line training and are re-weighted, in order to help the training of a new model for the target object. However, given the large number of potential species that users may care about, it may happen that few similar samples can be found for a reliable detection of certain user-registered species. The third direction is to conduct model sharing and transferring [6,14]. This approach leverages similar templates from the auxiliary models as prior knowledge when constructing a model for the target object. However, using the model transferring only may hurt the performance of learning when the auxiliary models are far away from reasonably representing the user-registered species.

In this context, we adopt DPM due to its excellent performance to represent the deformable objects and the sharable knowledge. Meanwhile, in order to generate a model adaptive to users' demand, we consider how and when to utilize the knowledge from both the auxiliary models and the user-registered samples.

3. Detection of user-registered dog faces

The flow diagram of the proposed framework is illustrated in Fig. 2. The framework comprises three main modules: training, registration and detection.

First, in the training stage, the pre-annotated images are used to train several detectors for the superordinate species. The learnt detectors make up a pool of auxiliary models. Then, in the registration stage, a new detector will be generated for the user-registered samples, with the help of some auxiliary models. A strategy-selection algorithm is designed to decide how and when to utilize both the prior knowledge stored in the auxiliary model pool and the knowledge newly extracted from the user-registered samples. Finally, in the detection stage, the new detector will be employed to detect the target objects in the test samples.

The detection stage follows the common slide-window detection approach and this paper will not go into much of its details. We shall mainly present the training stage in Section 3.1 and the registration stage in Section 3.2.

3.1. Training stage

In the training stage, the main task is to learn the auxiliary models that can be utilized as the prior knowledge for the registration stage. There are so many dog species that it will be a heavy burden to train and store models for all the species. Hence the training images are categorized into several superordinate groups based on their appearance, and models are only trained for these superordinate species as auxiliary models. Examples of six superordinate groups are shown in Fig. 3.

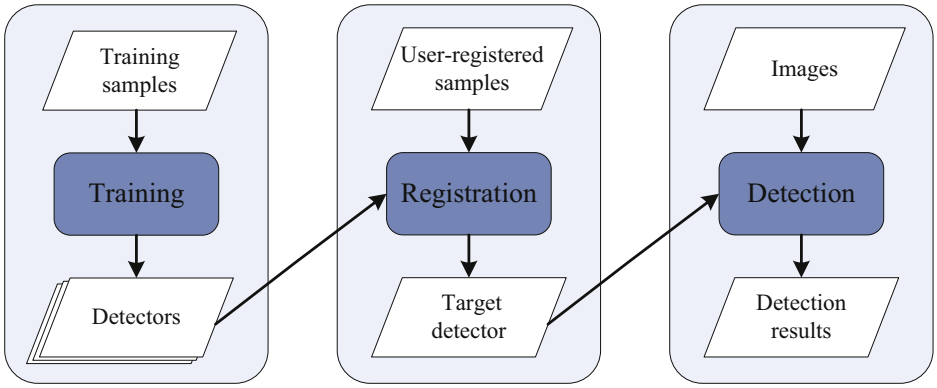


Fig. 2. Three main modules of the proposed framework.

Fig. 3. Examples of six superordinate groups from the dog face dataset.

To model a superordinate species, we adopt DPM, one of the most successful detection approaches on the PASCAL Visual Object Class (VOC) challenge [15]. A DPM consists of one root filter and several part filters. The root filter represents the global appearance and the part filters capture the local features.

For a model with P part filters, each example x is given a score by using the filter parameters ω and the feature vector $\psi(H,z)$. The score function is defined as

$$f_\omega(x) = \max_z \{\omega \cdot \psi(H,z)\}. \quad (1)$$

Here $\omega = (F_0, \ldots, F_P, d_1, \ldots, d_P, b)$ is a vector of filter parameters, in which F_0 is for the root filter, F_i is for the ith part filter, d_i is a four-dimensional vector that specifies coefficients of a quadratic function of penalty to moving the ith part far away from its supposed location v_i, and b is a bias term.

The function $\psi(H,z)$ is defined as

$$\psi(H,z) = (\phi_a(H,z_0), \ldots, \phi_a(H,z_P), -\phi_d(dz_1), \ldots, -\phi_d(dz_P), 1). \quad (2)$$

It is a feature vector extracted from the feature pyramid H with a specific latent spatial configuration $z = (z_0, z_1, \ldots, z_P)$, in which $z_i = (z_{xi}, z_{yi}, l_i)$ specifies the location and scale level of the ith part filter. The local appearance feature $\phi_a(H,z_i)$ describes the image area covered by the ith part; the deformation feature is $\phi_d(dz_i) = (dz_{xi}, dz_{yi}, dz_{xi}^2, dz_{yi}^2)$ with $(dz_{xi}, dz_{yi}) = (z_{xi}, z_{yi}) - (2(z_{x0}, z_{y0}) + v_i)$.

Here for semantic consideration, three part filters are learnt: two for eyes and one for nose. Examples are shown in Fig. 4.

3.2. Registration stage

In the registration stage, users only need to register a small number of images of the dog species that they care about, and the system will generate a suitable detector for the registered species.

The suitable detector has to cope well with three different situations regarding the relationship between the registered target samples and the off-line learnt species. (1) If the registered species has coincidentally been learnt already in the off-line training stage, picking the best fit auxiliary model can already meet the user's demand. (2) If the difference between the registered species and any of the previously learnt species is too large, the auxiliary models are helpless so only the registered samples should be used to learn the new model. (3) If some species that are similar to the registered species have been learnt, these auxiliary models can be utilized to generate a suitable model adapting to the registered samples. Such a transfer learnt model, with the help of these auxiliary models, can perform better than the directly trained model, as the latter can only be trained from a limited number of registered samples.

Therefore, we design a strategy-selection algorithm to cope with the three cases above. Given the user-registered samples $\{x_i, i = 1, \ldots, n\}$ and the auxiliary models $\{\omega_i, i = 1, \ldots, m_0\}$ with score function $f_{\omega_i}(\cdot)$ in (1), the task is to decide what kind of knowledge from them is useful and how to use the knowledge. The key point to accomplish the task is to measure each model's compatibility with the target object, and here we employ the average confidence $\overline{s_i} = (1/2n)\sum_{j=1}^{n}(f_{\omega_i}(x_j)+1)$ of each auxiliary model over all the registered samples as the measure. The strategy-selection algorithm is illustrated in Fig. 5 and Algorithm 1.

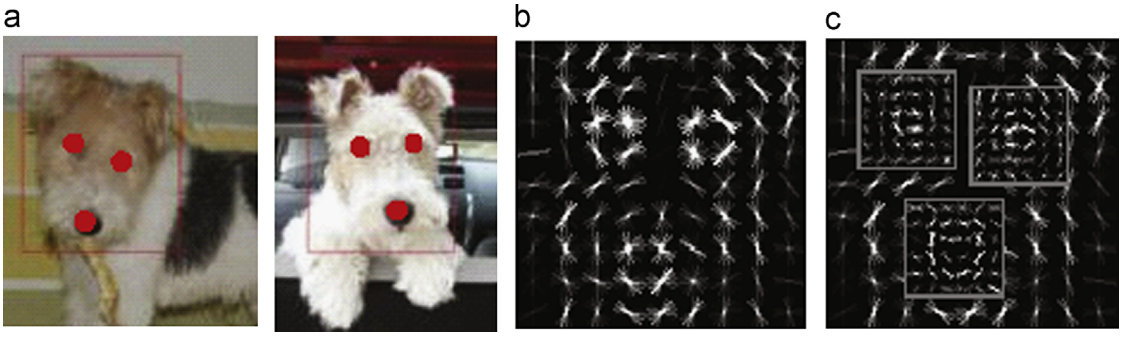


Fig. 4. Example of a learnt DPM: (a) annotated training samples, in which each dog face is annotated with a tight bounding box around the face and three points, two for eyes and one for nose; (b) the root filter; (c) three part filters.

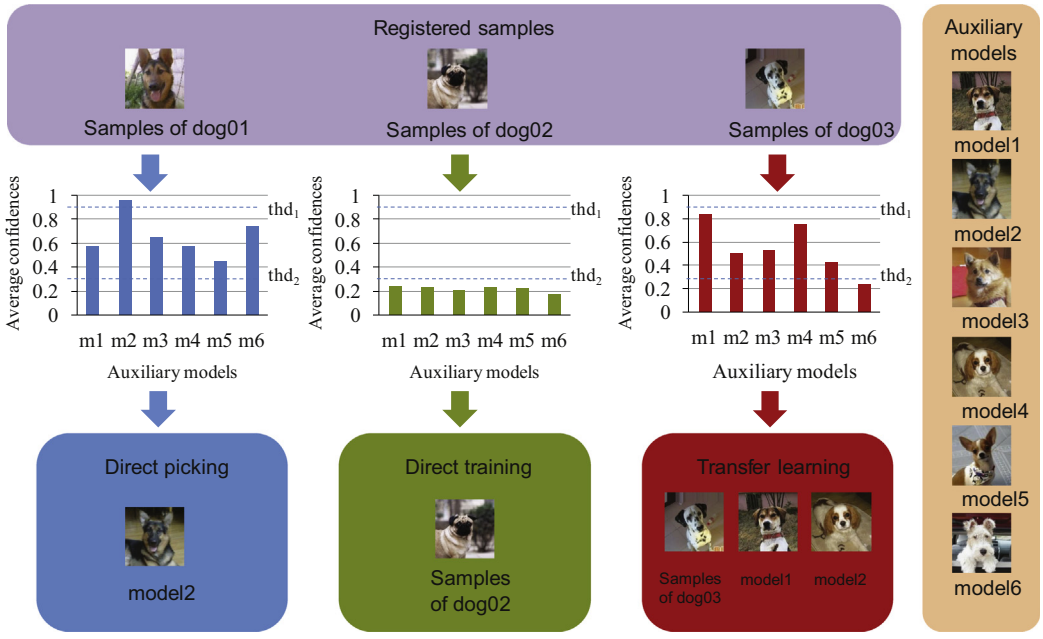


Fig. 5. The strategy-selection algorithm. Dog01's best auxiliary model has an average confidence above thd_1, hence the algorithm directly picks this model. Dog02's responses on all the auxiliary models are below thd_2, hence the algorithm directly trains a new detector by using the registered samples only. Dog03 does not have a high confident auxiliary model, but has two models with moderate confidences, hence the algorithm generates a new detector by transfer learning.

Algorithm 1. The strategy selection algorithm.

Input:
The auxiliary models $\Omega = \{\omega_i, i = 1, \ldots, m_0\}$
The registered samples $X = \{x_j, j = 1, \ldots, n\}$
Output:
The target object model ω
$\overline{s_i} \leftarrow \frac{1}{2n}\sum_{j=1}^{n} f_{\omega_i}(x_j), \quad i = 1, \ldots, m_0$
if $\max_i(\overline{s_i}) > thd_1$ **then**
  $I \leftarrow \text{argmax}_i \overline{s_i}$
  $\omega \leftarrow \omega_I$
else
  if $\max_i(\overline{s_i}) < thd_2$ **then**
    $\omega \leftarrow \text{DPMTrain}(X)$
  else
    $I_{sel} \leftarrow \{i, \overline{s_i} > \max(\lambda \cdot \max_i(\overline{s_i}), thd_2)\}$
    $\Omega_{sel} \leftarrow \{\omega_i, i \in I_{sel}\}$
    $\omega \leftarrow \text{DPMTransfer}(X, \Omega_{sel})$
  end if
end if

- *Case* 1 (*Direct Picking*): If one of the off-line trained auxiliary models is already able to perform well on the registered samples (i.e., the best auxiliary model has the average confidence $\overline{s_i}$ above a threshold thd_1), our framework will directly pick it as the suitable detector.
- *Case* 2 (*Direct Training*): If no auxiliary model can provide helpful prior knowledge (i.e., all the auxiliary models have low average confidences $\{\overline{s_i}\}$ below threshold thd_2), our method will directly train a new detector by using the registered samples only.
- *Case* 3 (*Transfer Learning*): If the auxiliary models are not suitable enough to be directly picked but some still share certain similarities with the registered species (i.e., all the models' average confidences $\{\overline{s_i}\}$ are below threshold thd_1 but some of them are above threshold thd_2 ($thd_1 > thd_2$)), then a transfer-learning algorithm will be applied to generate a new detector.

For Case 3, any transfer-learning algorithm that leverages similar auxiliary models for the construction of a new model can be applied. Here we extend the transfer learning algorithm that we proposed in [16] from utilizing the single best auxiliary model to utilizing multiple auxiliary models.

Since the transfer-learning algorithm transfers prior knowledge from auxiliary models, if we use only the auxiliary models of high confidences, the auxiliary models that may hurt the performance can be removed. In addition, often the more the auxiliary models

of high confidences, the better the transfer learning. Therefore here we enhance our transfer-learning algorithm of [16] by using a model-selection algorithm. The latter chooses as many as auxiliary models whose average confidences are close to that of the best model (i.e., each model whose average confidence $\overline{s_i} > \max(\lambda \cdot \max_i(\overline{s_i}), thd_2)$ is selected). The enhanced transfer-learning algorithm is described briefly as follows.

Given labeled examples $X = (\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle)$ where $y_i \in \{-1, 1\}$, and given DPMs $\Omega = \{\omega_i^s, i = 1, \ldots, m_0\}$ where the ith model ω_i^s has P_i parts, the task is to learn a new DPM ω for the target species.

First the model-selection algorithm selects a subset $\Omega_{sel} = \{\omega_i, i \in I_{sel}\}$ from the give auxiliary model pool Ω, where $I_{sel} = \{i, \overline{s_i} > \max(\lambda \cdot \max_i(\overline{s_i}), thd_2)\}$. With these m selected auxiliary models, we try to assemble the root and part filters of them and adapt the filters to the registered samples to generate a new DPM.

If all the auxiliary part filters are considered as uncorrelated filters, the number of the part filters that the target model ω will have is equal to $P = \sum_{i=1}^{m} P_i$. If the auxiliary part filters correspond to $P'(P' <= \sum_{i=1}^{m} P_i)$ parts which have same filter sizes and similar locations and semantic meanings (e.g., two part filters for the noses of two dog species correspond to the same part of ω), the target model ω has $P = P'$ part filters.

Subsequently, the set of m selected DPMs $\Omega_{sel} = \{\omega_i, i \in I_{sel}\}$ can be converted to a set of auxiliary filters $\{\omega_i', i = 1, \ldots, m'\}$ 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 learnt by optimizing a new objective function,

$$L_{Z_p}(\omega, \alpha) = \left\| \omega - \sum_{i=1}^{m'} \beta_i \omega_i' \right\|^2 + \alpha \sum_{i=1}^{m'} \overline{\beta_i}^{-2} + C \sum_{i=1}^{n} \max(0, 1 - y_i g_w(x_i, z_i)),$$

$$= \|\overline{\omega}\|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(\overline{\omega} \cdot \overline{\psi}_i)), \tag{3}$$

where

$$Z_P = \{z_i, i = 1, \ldots, n | z_i = \underset{z}{\operatorname{argmax}}\, g_w(x_i, z)\}, \tag{4}$$

$$\overline{\beta_i} = \beta_i \| \omega_i' \|, \tag{5}$$

$$\overline{\omega} = \{\Delta\omega, \sqrt{\alpha}\overline{\beta}_1, \ldots, \sqrt{\alpha}\overline{\beta}_{m'}\}, \quad \Delta\omega = \omega - \sum_{i=1}^{m'} \beta_i \omega_i', \tag{6}$$

$$\overline{\psi_i} = \left\{ \psi(H_i, z_i), \frac{1}{\sqrt{\alpha}\| \omega_1' \|} \omega_1' \cdot \psi(H_i, z_i), \ldots, \frac{1}{\sqrt{\alpha}\| \omega_{m'}' \|} \omega_{m'}' \cdot \psi(H_i, z_i) \right\}. \tag{7}$$

In (3), the first term represents the distance between the learnt model ω and the assembled filters; the second term indicates the re-weighting of the auxiliary filters; the third term is the standard hinge loss. The parameters α and C control the relative weights between the two regularization terms, and the set Z_p includes the highest latent values for the positive samples. It can be optimized by using a two-step EM iterative procedure, as detailed in [16].

4. Experimental results

4.1. Experimental settings

We evaluate the proposed user-registered detection framework on a dataset of dog faces. The dataset contains 10,712 images of near-frontal dog faces from 32 species. Each dog face is annotated with a tight bounding box around the face and three points, two for the eyes and one for the nose. The positive training samples of dog faces are normalized so that the distance between the two eyes is 48 pixels. The negative samples come from the dog-free images in the PASCAL VOC 2007 dataset [15]. For learning auxiliary models, six superordinate groups, each of 3 similar dog species, are generated based on the similarity in appearance (Fig. 3). Each auxiliary model is obtained by training a DPM from more than 200 positive examples of a superordinate group.

For object description, features based on the histogram of oriented gradient (HOG) descriptors [17] are employed, as with [7]. In all experiments, we fix parameter C to 0.002 (the default value of [7]) and set α simply to 1. For the strategy-selection algorithm and the model-selection algorithm, we empirically fix thd_1 to 0.94, thd_2 to 0.37 and λ to 0.87.

We use the detection rate and the false positives per image (FPPI) 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.

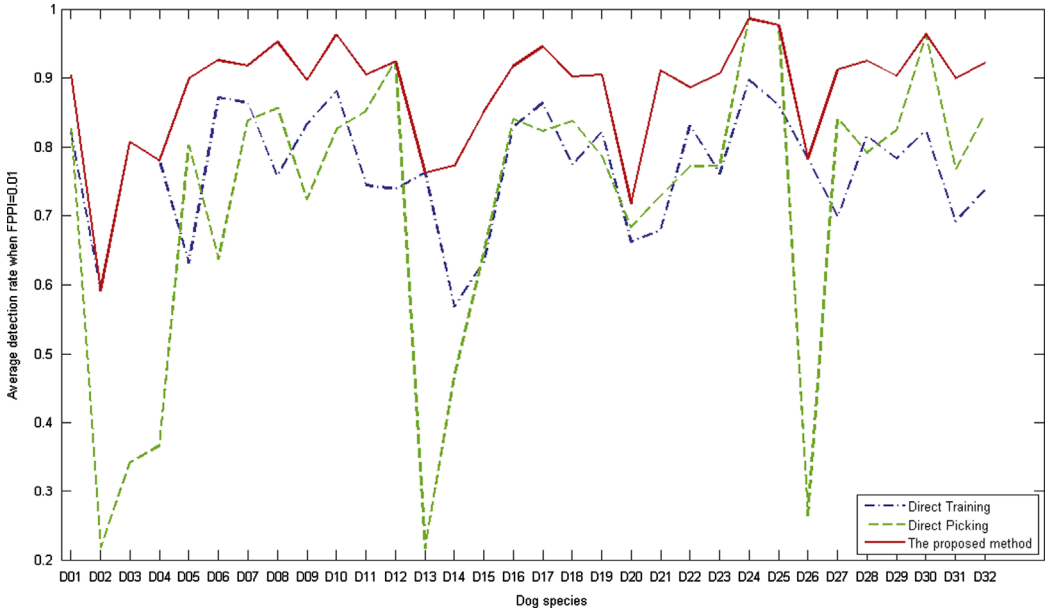


Fig. 6. Average detection rate (%) of different detector-generating methods on 32 dog species with 10 registered samples when FPPI=0.01.

4.2. Overall performance of the proposed method

The dataset of each species is divided into a training set and a test set. From the training set, only ten positive examples of a species are randomly selected as the user-registered samples. On the test set, the testing performs five trials and their detection rates are averaged.

For overall performance evaluation, three detector-generating methods are compared: (1) 'Direct Training' only explores knowledge in the registered samples to train a new DPM as the detector; (2) 'Direct Picking' only explores knowledge in the auxiliary models and adopts the best scored auxiliary model to detect the target object; (3) the proposed method utilizes both the registered samples and the auxiliary models to generate a suitable detector.

The average detection rates of these three methods when FPPI=0.01 are presented in Fig. 6. We can observe that, for all 32 species, the proposed method performs the best.

We believe such an excellent performance can be significantly attributed to the strategy-selection algorithm in the proposed framework. Remember in the registration stage we employ three model-generating strategies, 'Direct Picking', 'Direct Training' and 'Transfer Learning', and the strategy-selection algorithm will automatically choose a suitable strategy for the target object. As illustrated in Fig. 7, the strategy-selection algorithm applies 'Direct Training' to 5 species, 'Direct Picking' to 4 species and 'Transfer Learning' to the rest 23 species.

To verify whether the strategy-selection algorithm selects the right strategy, we choose three species, German shepherd, Pug and Dalmatian, to which, respectively, the three strategies are automatically selected. The performance of the three strategies on these three species are shown in Fig. 8. We can observe that the strategy-selection algorithm has selected the right strategy for each species.

- For German shepherd, there is one auxiliary model whose average confidence is above thd_1, so the strategy-selection method adopts 'Direct Picking' which directly picks the best model. As shown in Fig. 8(a), indeed in this case 'Direct Picking' is the best strategy, while with only limited number of registered samples 'Direct Training' performs the worst.
- For Pug, there are no auxiliary models having average confidences above thd_2. The strategy-selection method thus adopts 'Direct Training' for this species. As shown in Fig. 8(b), even the best scored auxiliary model has poor performance (22% of the detection rate when FPPI=0.01) for this species. Hence, the prior knowledge in the auxiliary models is not much relevant,

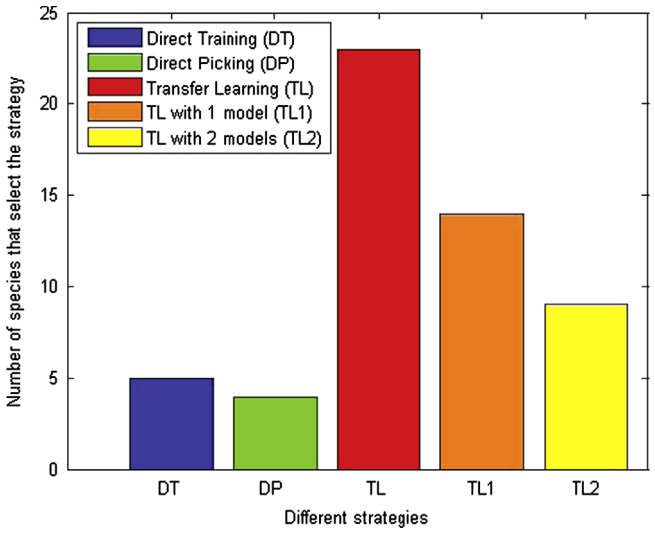


Fig. 7. Count of species selected by the strategy-selection algorithm vs. different strategies.

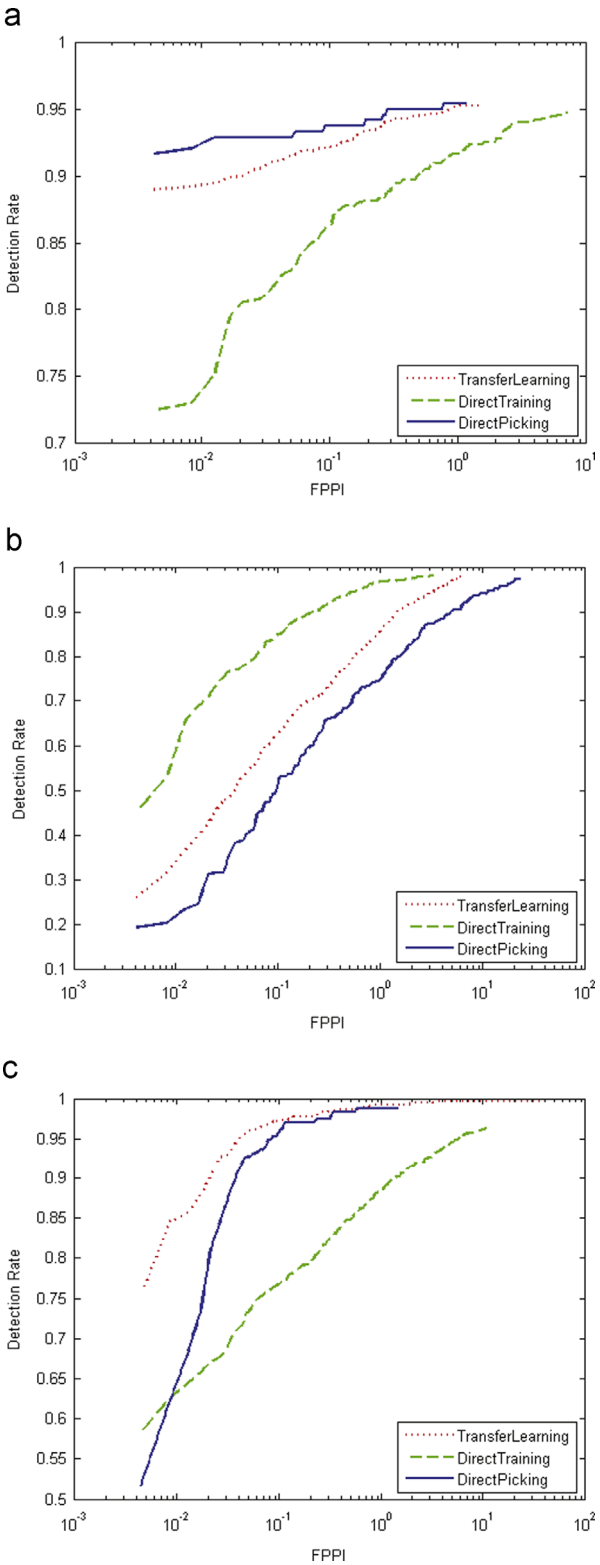


Fig. 8. Comparison of three strategies on 3 dog species. The strategy-selection algorithm selects 'Direct Picking' for German shepherd, 'Direct Training' for Pug and 'Transfer Learning' for Dalmatian, according to the average confidences of the registered samples on the auxiliary models. As illustrated for each species, the strategy that the strategy-selection algorithm automatically selected is the optimal one. (a) Results on German shepherd. (b) Results on Pug. (c) Results on Dalmatian.

and 'Direct Training' is optimal, better than 'Direct Picking' and 'Transfer Learning'.

- For Dalmatian, there are no auxiliary models having average confidences above thd_1 but there are a couple of models having average confidences above thd_2. Hence, the strategy-selection method selects 'Transfer Learning', to exploit the prior knowledge from these auxiliary models of moderate confidences. As shown in Fig. 8(c), in this case, the transfer-learning method (85% of the detection rate when FPPI=0.01) outperforms both the direct trained model (63% when FPPI=0.01) and the direct picked model (65% when FPPI=0.01). Although for the target object the selected auxiliary models are not strong since the confidence $< thd_1$, they still can help improving the performance of the detector by transfer learning. Therefore, the algorithm correctly selects 'Transfer Learning' as the best strategy for this species.

For the strategy-selection algorithm, two thresholds thd_1 and thd_2 are used to select appropriate strategies. To investigate their impact, we vary their values (± 0.05) around the values we set empirically. As shown in Fig. 9, the average detection rate over all the species has less than 2% of change. This indicates that the framework is rather stable to the two thresholds.

In addition, we can observe that, among the 32 species, 23 species selected 'Transfer Learning'. As illustrated in Fig. 7, among the 23 species that select transfer learning, 14 species employ one best auxiliary model ('TL1') and 9 species employ two best auxiliary models ('TL2') for transfer. This indicates the effect of the proposed transfer-learning and model-selection algorithms.

To further investigate the benefits of using our proposed transfer-learning and model-selection methods, respectively, we carry out more experiments on Dalmatian, one of the 9 species that employ two best auxiliary models for transferring.

4.2.1. Benefit from the proposed transfer-learning algorithm

To investigate the advantage of the proposed transfer-learning method, we compare four transfer-learning methods on Dalmatian: (1) 'None-Trans' directly learns a new DPM from the registered target samples without transfer; (2) 'Root-Trans' transfers similar part-like patches only from the root filters of auxiliary models to generate a new rigid template [14]; (3) 'ASB-Trans' assembles a new DPM by using the root and part filters of auxiliary models without

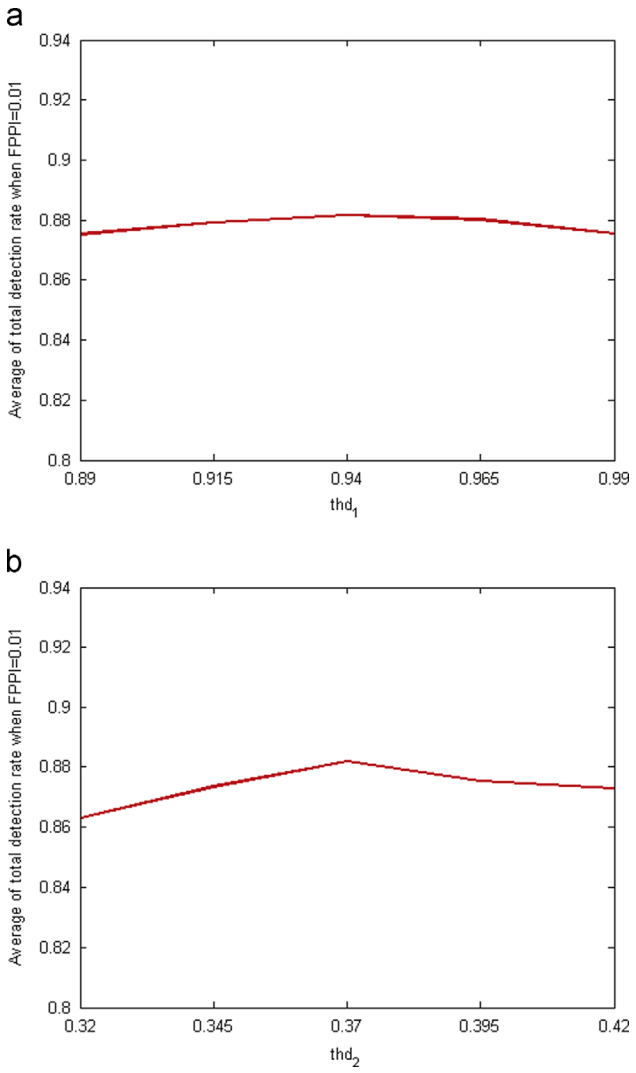


Fig. 9. Average detection rate over all the species versus thresholds thd_1 and thd_2.

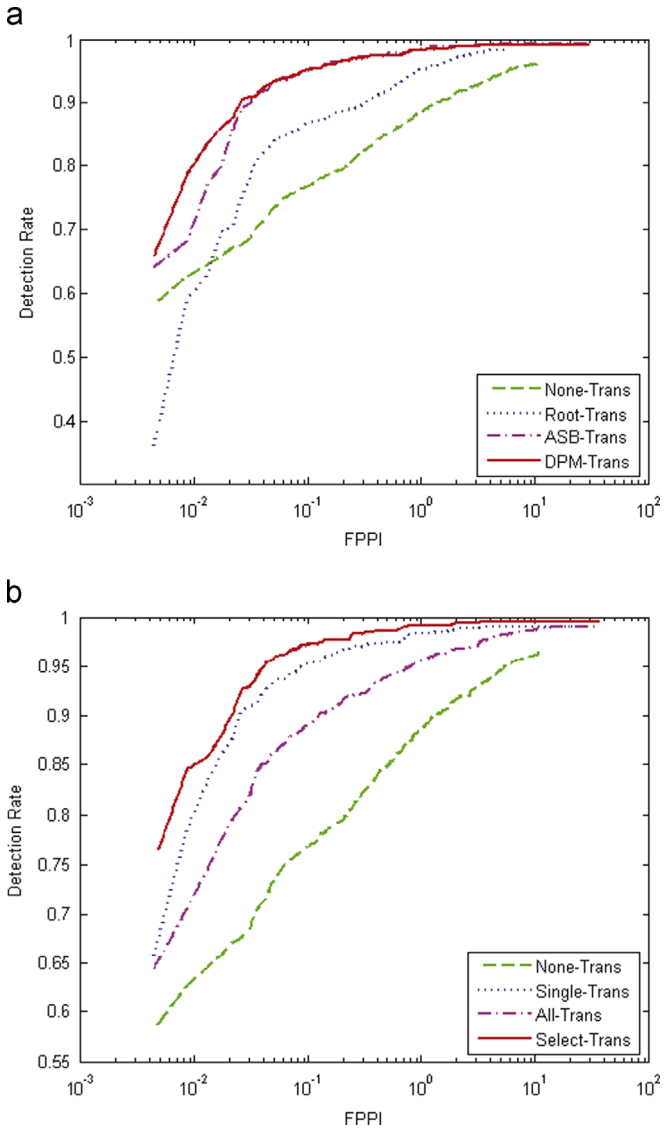


Fig. 10. Comparison of different transfer-learning methods and different model-selection methods on Dalmatian. As shown in 10(a), the proposed transfer-learning algorithm ('DPM-Trans') performs best among the four transfer-learning algorithms. As shown in 10(b), based on the proposed transfer-learning algorithm, the proposed model-selection method ('Select-Trans') can further improve the performance. (a) Comparison of transfer-learning methods. (b) Comparison of model-selection methods.

adaption [6]; (4) 'DPM-Trans' is our proposed method of transferring and adapting all the filters of auxiliary models.

The evaluation results are presented in Fig. 10(a). We can observe that the proposed transfer-learning method performs the best. It demonstrates the importance of transferring and adapting the part filters.

4.2.2. Benefit from the proposed model-selection algorithm

Based on the proposed transfer-learning method, we want to see how the proposed model-selection method can further improve the performance. Four model-selection methods are compared on Dalmatian: (1) 'None-Trans' learns DPM from the registered target samples without transfer; (2) 'Single-Trans' transfers only the single best-scored auxiliary model; (3) 'All-Trans' transfers all the auxiliary models in the pool; (4) 'Select-Trans' transfers the selected auxiliary models by our proposed model-selection method.

As shown in Fig. 10(b), the models learnt from the auxiliary models selected by the proposed model-selection method perform the best. When there is more than one suitable auxiliary models, the proposed model-selection method can choose the auxiliary models that can really help and avoid the auxiliary models that will hurt the performance. Therefore, it outperforms the transfer results which use single best-scored model or use all the models.

In short, with the help of the proposed strategy-selection, model-selection and transfer-learning algorithms, the user-registered detection framework can utilize the knowledge from both the registered

Fig. 11. Detection results of various dog species.

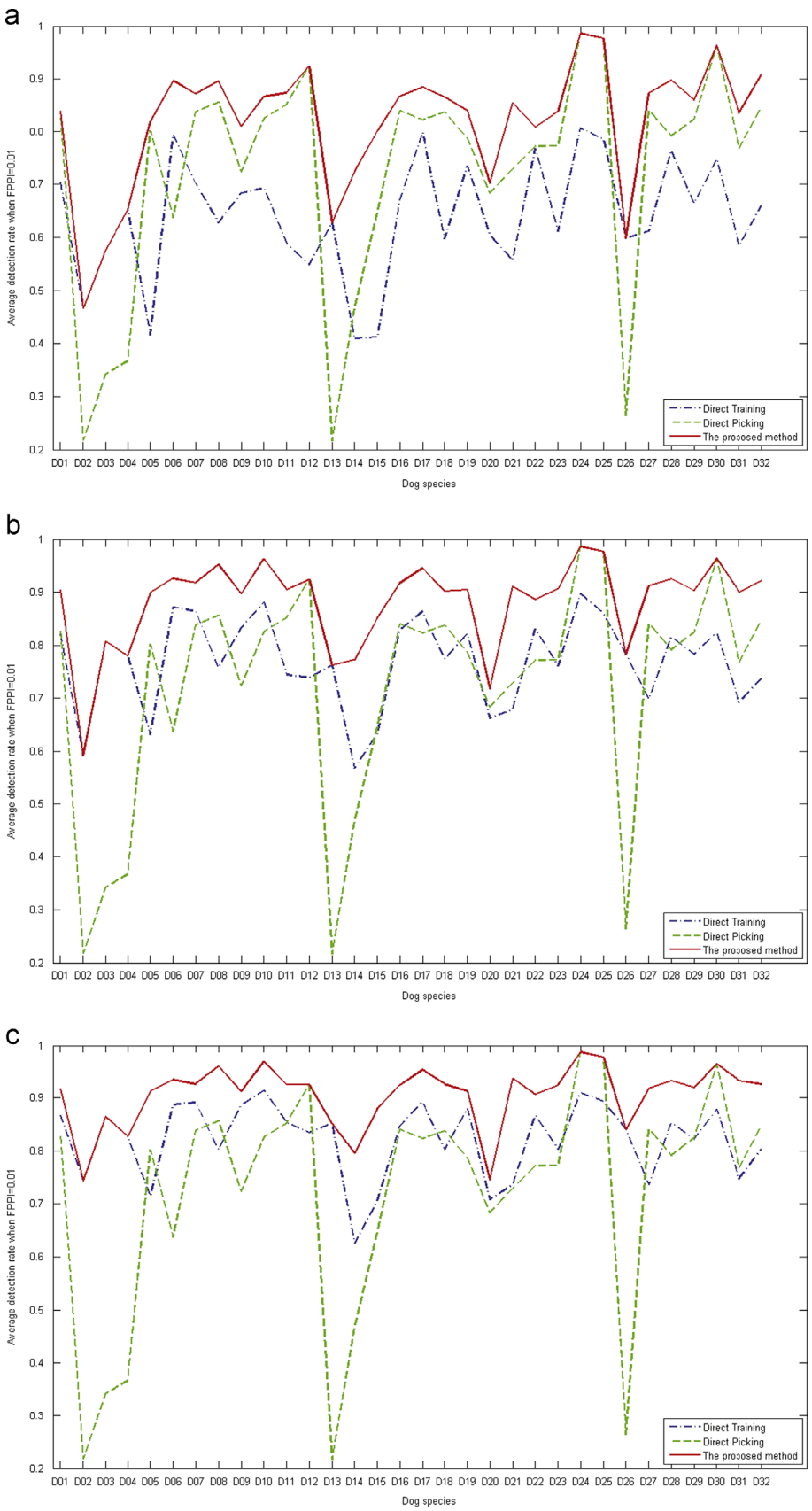


Fig. 12. Average detection rate (%) of different detector-generating methods on 32 dog species with 5, 10 and 20 registered samples when FPPI=0.01. The proposed user-registered detection framework remains the best with different numbers of registered samples. (a) With 5 registered samples. (b) With 10 registered samples. (c) With 20 registered samples.

samples and the auxiliary models rationally, generating a detector fit for the target object.

For illustrative purposes, more detection result are shown in Fig. 11. As we can see, although the different dog species shown in the images have great diversity in appearance, such as different colors, lengths of hair, lengths of nose, shapes of ear, etc., the proposed registered detection framework can generate suitable detectors for the registered species. For instance, the dog faces with different poses and expressions are all well detected, and the eyes and noses are also well located.

4.3. Evaluation with different numbers of registered samples

In the following experiments, we investigate how the numbers of registered samples affect the performance of our proposed framework. Each experiment is repeated five times with 5, 10 and 20 randomly picked registered samples. For each of these three cases, we compare the same three methods 'Direct Training', 'Direct Picking' and our proposed user-registered detection framework.

As shown in Fig. 12, in all three cases, the proposed method outperforms both 'Direct Training' and 'Direct Picking'. Compared with the results with 10 registered samples, the performance of 'Direct Training' drops significantly with 5 samples and improves to certain extent with 20 samples. The performance of 'Direct Picking' shows little change with different numbers of registered samples. The performance of the proposed method improves gradually as the number of samples grows and is much stable than 'Direct Training'.

To see how stable the proposed strategy-selection method is when the number of registered samples varies, the three strategies ('Direct Picking', 'Direct Training' and 'Transfer Learning') are compared on the three species: German shepherd, Pug and Dalmatian, as with Section 4.2. The results are shown in Fig. 13.

- For German shepherd, the strategy-selection algorithm always choose 'Direct Picking'. As shown in Fig. 13(a), although 'Direct Training' and 'Transfer Learning' improve their performances as the number of registered samples grows, they are always worse than 'Direct Picking'. That is, the strategy selection is stably correct.
- For Pug, the strategy-selection algorithm always choose 'Direct Training'. As shown in Fig. 13(b), 'Direct Training' outperforms both the best scored auxiliary model and the transfer learnt models, even when there are only 5 registered samples. It not only indicates that the strategy selection is stable and reliable, but also demonstrates that the prior knowledge from irrelevant auxiliary models can hurt the transferring performance greatly.
- For Dalmatian, the strategy-selection algorithm always choose 'Transfer Learning'. We can see from Fig. 13(c) that 'Transfer Learning' is indeed the best strategy for this species for different numbers of registered samples, compared with 'Direct Training' and 'Direct Picking'.

In short, with different numbers of registered samples, the proposed strategy-selection stably selects the optimal strategy and the proposed user-registered detection framework remains the best.

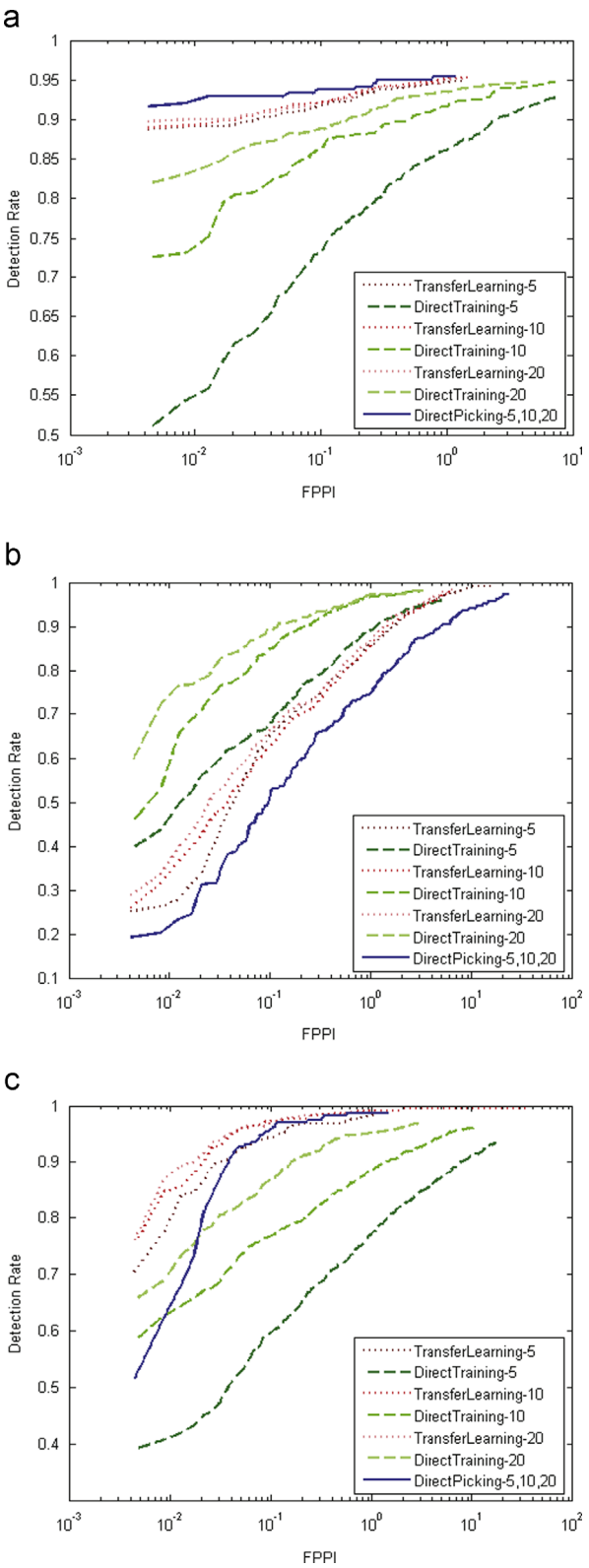


Fig. 13. Comparison of different detector-generating methods with 5, 10 and 20 registered samples on 3 dog species. The performance of 'Direct Picking' shows little change with different numbers of registered samples, so we only plot a single curve 'DirectPicking-5, 10, 20' here. As shown in the results, the proposed strategy-selection algorithm can select the optimal strategy with different numbers of registered samples. (a) Results on German shepherd. (b) Results on Pug. (c) Results on Dalmatian.

5. Conclusion

In this paper, we have proposed a user-registered object detection framework. The framework can offer users customizable detectors after the users register a small number of samples of the target object that they are interested in. The framework can effectively leverage the off-line trained auxiliary models and the user-registered samples, through using a strategy-selection algorithm. The experiments on the dataset of dog faces demonstrated that this method can lead to a detector superior to

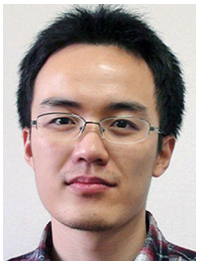
the state-of-the-art methods. The proposed method can adapt to a great deal of dog species, and can be easily extended to other object categories, such as cat, bird and motor vehicle.

Acknowledgments

We are grateful to the reviewers for their constructive comments and suggestions. The work was partially sponsored by National Natural Science Foundation of China (No. 61271390).

References

[1] T. Kozakaya, S. Ito, S. Kubota, O. Yamaguchi, Cat face detection with two heterogeneous features, in: ICIP, IEEE, 2009, pp. 1213–1216.
[2] O.M. Parkhi, A. Vedaldi, C. Jawahar, A. Zisserman, The truth about cats and dogs, in: ICCV, IEEE, 2011, pp. 1427–1434.
[3] W. Zhang, J. Sun, X. Tang, From tiger to panda: animal head detection, IEEE Trans. Image Process. 20 (6) (2011) 1696–1708.
[4] H. Azizpour, I. Laptev, Object detection using strongly-supervised deformable part models, in: ECCV, Springer, Florence, Italy, 2012, pp. 836–849.
[5] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[6] Y. Aytar, A. Zisserman, Enhancing exemplar SVMs using part level transfer regularization, in: BMVC, 2012, pp. 1–11.
[7] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[8] H. Grabner, H. Bischof, On-line boosting and vision, in: CVPR, vol. 1, IEEE, 2006, pp. 260–267.
[9] Z. Qi, Y. Xu, L. Wang, Y. Song, Online multiple instance boosting for object detection, Neurocomputing 74 (10) (2011) 1769–1775.
[10] C. Liu, G. Wang, J. Fan, X. Lin, Online HOG method in pedestrian tracking, IEICE Trans. Inf. Syst. 93 (5) (2010) 1321–1324.
[11] W. Dai, Q. Yang, G.-R. Xue, Y. Yu, Boosting for transfer learning, in: ICML, ACM, 2007, pp. 193–200.
[12] M. Wang, W. Li, X. Wang, Transferring a generic pedestrian detector towards specific scenes, in: CVPR, 2012, pp. 3274–3281.
[13] B. Zeng, G. Wang, Z. Ruan, X. Lin, Data level object detector adaptation with online multiple instance samples, in: ICASSP, IEEE, 2012, pp. 1397–1400.
[14] K. Yang, M. Wang, X.-S. Hua, S. Yan, H.-J. Zhang, Assemble new object detector with few examples, IEEE Trans. Image Process. 20 (12) (2011) 3341–3349.
[15] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, URL: ⟨http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html⟩.
[16] Z. Ruan, G. Wang, X. Lin, J.-H. Xue, Y. Jiang, Deformable part-based model transfer for object detection, IEICE Trans. Inf. Syst. 97-D (5) (2014) 1394–1397.
[17] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, 2005, pp. 886–893.

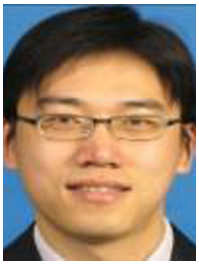


Zhiwei Ruan received the B.S. degree in Information and Electronics Engineering from the Department of Electronic Engineering, Tsinghua University, China, in 2009. He is currently a Ph.D. candidate in the Department of Electronic Engineering, Tsinghua University. His research interests are in the area of object detection and tracking and intelligent surveillance.

Guijin Wang received the B.S. and Ph.D. degree in signal and information processing (with honors) from the Department of Electronics Engineering, Tsinghua University, China, in 1998 and 2003, respectively. From 2003 to 2006, he was with Sony Information Technologies Laboratories as a researcher. Since 2006, he has been with the Department of Electronics Engineering at, Tsinghua University, China, as an associate professor. He has published over 50 International journal and conference papers and holds several patents. He is the session chair of IEEE CCNC'06. His research interests are focused on wireless multimedia, image and video processing, depth imaging, pose recognition, intelligent surveillance, industry inspection, object detection and tracking and online learning.

Jing-Hao Xue received the B.Eng. degree in telecommunication and information systems in 1993 and the Dr.Eng. degree in signal and information processing in 1998, both from Tsinghua University, the M.Sc. degree in medical imaging and the M.Sc. degree in statistics, both from Katholieke Universiteit Leuven, in 2004, and the Ph.D. degree in statistics from the University of Glasgow, in 2008. He has worked in the Department of Statistical Science at University College London as a Lecturer since 2008. His research interests include statistical and machine-learning techniques for pattern recognition, data mining and image processing, in particular supervised, unsupervised and incompletely supervised learning for complex and high-dimensional data.

Xinggang Lin received his B.S. in electronics engineering, Tsinghua University, China, in 1970; an M.S. in 1986 and a Ph.D. in 1982, both in information science from Kyoto University, Japan. He joined the Department of Electronics Engineering at Tsinghua University in 1986 where he has been a full professor since 1990. He received a "Great Contribution Award" from the Ministry of Science and Technology of China, and "Promotion Awards of Science and Technology" from Beijing Municipality. He was a General co-chair of the second IEEE Pacific-Rim Conference on Multimedia, an associate editor of IEEE T. on CSVT, and a technical/organizing committee member of many international conferences. He is a fellow of the China Institute of Communications, and has published over 140 referred conference and journal papers in diversified research fields.



Yong Jiang received his B.S. in electronics engineering, Zhengzhou University, China, in 2001; and received his M.S. and Ph.D. in Computer vision and image processing, Nanjing University of Aeronautics and Astronautics, China, in 2004 and 2007. From 2006 to the present, he was working in Canon information Technology (Beijing) Co., LTD as an intern, researcher, senior researcher and project manager, and applied more than 10 patents in US, Japan, and China as the first inventor. His research interests are focused on image and video processing, intelligent surveillance, industry inspection, object detection and tracking and online learning.