LOCATION-AWARE OBJECT DETECTION VIA COHERENT REGION GROUPING

Shen-Chi Chen[1], Kevin Lin[2], Chu-Song Chen[2], Yi-Ping Hung[1]

[1]National Taiwan University, [2]Academia Sinica, Taiwan

ABSTRACT

We present a scene adaptation algorithm for object detection. Our method discovers scene-dependent features discriminative to classifying foreground objects into different categories. Unlike previous works suffering from insufficient training data collected online, our approach incorporated with a similarity grouping procedure can automatically gather more consistent training examples from a neighbour area. Experimental results show that the proposed method outperforms several related works with higher detection accuracies.

Index Terms— Object detection, visual surveillance.

1. INTRODUCTION

A main difficulty of applying a pre-trained object detector to visual surveillance is to collect the training data suitable for various unknown situations. To learn an effective classifier, a fundamental issue is to keep the training and testing data drawn from the same distribution. However, a surveillance camera could be mounted in any environments. The objects may have various appearances in different environments. Hence, object detectors pre-trained [1, 2, 3, 4, 5] usually perform worse for a surveillance camera even though a large-scale dataset has been used for off-line learning.

Collecting on-site training data and doing re-training or transfer-learning [6, 7] from the data is a solution to this problem. However, the object appearance could be highly variant in the entire scene, and thus the data are often inappropriate to learn a classifier available for the whole scene. Fortunately, the target objects seen from a road-side camera usually have some local regularities in their sizes, shapes, and/or moving speeds. For example, when objects (eg., vehicle, bike, or pedestrian) move in a scene, it is often that the foreground blobs sharing the same moving directions have a consistent object orientation. Examples include pedestrians going across a zebra crossing (Fig. 1) and cars driving on roads. Such scene-regularities is helpful in gathering consistent data for better online training to learn an effective object detector adaptive to locations.

Since the object movement is regularized by the local structure in a scene, different locations would be suitable to have different classifiers applied. To do this, several approaches subdivide the input image into grids (or size-fixed subregions) and gather the foreground samples passing through the same grid for online training [8, 9, 10, 11]. A limitation is that they are apt to get immature solutions as the number of objects passing by a grid could be small and therefore the training data collected per grid is insufficient to conduct a non-superficial detector.

Fig. 1: Pedestrians cross the road with similar appearance and moving directions.

In this paper, we propose a superpixel-based approach to employ the scene regularity. Our method segments the scene image into nonuniform regions (i.e., superpixels), and the object samples passing by each region have similar location-dependent features (such as coherent object size, shape, moving direction and speed). Our method thus avoids the difficulty caused by constant-shape grids. Each superpixel contains a larger number of feature-consistent training samples, making the online learning more efficient. The proposed approach does not rely on pre-specified object size or grid size that is crucial to determine. We use the MASS (referred to as "Moving direction, Aspect ratio, Size, Speed") features extracted from the passing-by objects per location for grouping a region. In addition to the location-aware object detectors learned, the MASS features can serve as further prior knowledge for object detection in the scene. In our work, they are combined to conduct an effective scene-dependent object detection algorithm.

2. METHODOLOGY

We introduce a pixel-driven and grouping approach in the prior-scene understanding phase. Our approach then segments the scene into homogeneous regions based on several semantically meaningful and location-adaptive features. Figure 2 shows an overview diagram. The proposed method consists of four stages and is described in detail as follows.

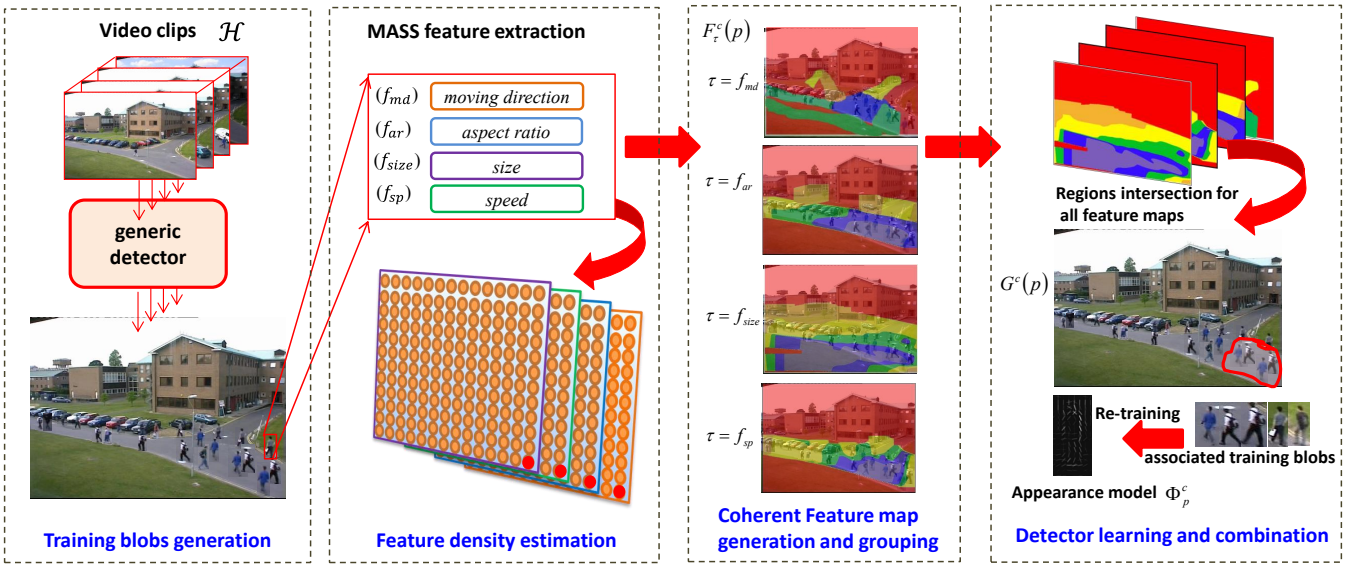


Fig. 2: Overview of learning the location-aware object detectors in a specific scene

2.1. Training blobs generation

At first, we assume that several video clips were gathered for the target scene, constituting the training dataset $\mathcal{H}$.

Given a training image $I \in \mathcal{H}$, we apply a generic detector [1] (learned offline) of category c $(c = 1...C)$ to the foreground blobs extracted from background subtraction [12]. If a foreground blob is classified as positive by the c-th generic detector, we then put this blob into the training bag of the category c. After dealing with all the training images, we obtain C training bags $\mathbf{B_c}$ $(c = 1...C)$ as follows:

$$\mathbf{B}_c = \{o_k^c \mid k = 1...K_c\}, \tag{1}$$

where $o_k^c \in \mathbf{B_c}$ are the foreground blobs found by the generic detector of category c. The categories considered in our study include pedestrian, vehicle, and bicycle/scooter (we treat both as the same) with $C = 3$. Our method is also applicable to detect a single category of objects, eg., by setting $C = 1$ for learning pedestrian detectors only.

2.2. Feature density estimation

Let $(x, y) \in \mathbb{Z}^2$ $(1 \le x \le W, 1 \le y \le H)$ be a pixel location in the input image. What we are interested are the MASS features of the foreground blobs passing by (or nearby) the position (x, y). Given the training blobs, we estimate the Gaussian distribution $\mathcal{N}(\mu_{\tau;p}^c, \sigma_{\tau;p}^c)$ of the MASS feature $\tau \in \{md, ar, size, sp\}$ on the site $p = (x, y)$ for the category c.

Note that the training blobs could be noisy because both the generic detector and background subtraction methods are non-perfect. However, we need only a rough estimation of very few parameters (μ and σ) at this stage. Thus the parameters estimated still work well for such a preliminary purpose. The parameters are easy to be refined incrementally too.

2.3. Coherent feature map generation and grouping

To collect the training data online, a significant issue is that it causes easily the failure of a learning algorithm when the training dataset is composed of exemplars of various orientations or viewing angles. Collecting these orientation-similar objects online is thus helpful to learning an effective object detector adaptive to locations. Similarly, features such as size and aspect ratio of the foreground blob are also useful to collecting appearance-similar objects and ease the online learning process.

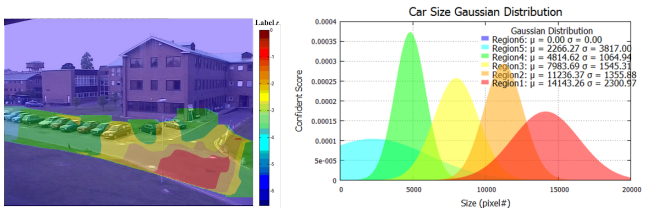
We consider the MASS features $\tau \in \{md, ar, size, sp\}$ that reflect the location-based features of object moving direction, aspect ratio, size, and speed, respectively. As the samples associated only with location p are insufficient for training, given a feature τ, we group the positions with similar feature values in terms of τ by segmenting the image of into feature coherent regions. The segmentation is achieved by using the multi-label graph cuts algorithm in [13] that minimizes an objective function E of the following form:

$$E(l) = \sum_{p \in \mathbf{P}} E_{data}(l_p) + \sum_{p,q \in \mathbf{N}} E_{smooth}(l_p, l_q), \tag{2}$$

with $l_p \in L$. L is the set of labels, $\mathbf{P}$ is the set of all pixels and $\mathbf{N}$ is a set of all pixel pairs of a neighbourhood system. We define the data term based on the probability distribution map, $\{\mathcal{N}(\mu_{\tau;p}^c, \sigma_{\tau;p}^c) \mid p \in \mathbf{P}\}$ as follows:

$$E_{data}(l_p) = \left| l_p - \mu_{\tau;p}^c \right|. \tag{3}$$

To formulate the smoothness term, we hope that neighbor positions tend to have the same labels. When two positions in the neighbor system $\mathbf{N}$ have similar means but different labels



(a) (b)

Fig. 3: An example of the size coherence map and its Gaussian distribution of each region for category car. There are other speed, aspect ratio, moving direction coherence maps generated by the process in Sec. 2.3 as well.

l_p and l_q, we set a large penalty weight to this pair of positions; besides, we decrease the penalty weight if their variances are large. The smoothness term is defined as

$$E_{smooth}\left(l_p, l_q\right) \propto \frac{1-\delta\left(l_p - l_q\right)}{\left(\mu^c_{\tau;p} - \mu^c_{\tau;q}\right)\left[\left(\mu^c_{\tau;p}\right)^2 + \left(\mu^c_{\tau;q}\right)^2\right]}, \quad (4)$$

with $\delta(\cdot)$ denoting the delta function.

Given a feature τ and category c, we call the segmentation result the *feature coherence map* of the feature and category, denoted as $F^c_\tau\left(p\right), p \in \mathbf{P}$. The number of regions $|L|$ can be set according to the complexity of scenes. In our implementation, we choose $|L|$= 6. Fig. 3 shows an example of the feature coherence map obtained for the vehicle category from the MASS features. It can be seen that the distributions of the feature coherence map are roughly consistent with the perspective effect and the road areas of the scene.

2.4. Detector learning and combination

Once the feature coherence maps are constructed, we first intersect the regions for all MASS features,

$$G^c\left(p\right) = \cap\{F^c_\tau\left(p\right)|\tau \in \{md, ar, size, sp\}\}, p \in \mathbf{P}. \quad (5)$$

Each region in the obtained map G^c then contains similar values for all MASS features. For every region γ in G^c, we train an appearance detector Φ^c by using the training blobs associated with this region, and employ Φ^c as the appearance detector for every location $p \in \gamma$ (i.e, we set $\Phi^c_p = \Phi^c\ \forall p$). Compared to using the grid regions of a fixed size, our approach discovers automatically the non-uniform-shape regions with coherent characteristics of the moving objects. Hence our method is more suitable for on-site learning than previous methods.

We can use either transfer learning or re-training for constructing Φ^c from the associated training blobs, where the former migrates the generic detector to a domain-transferred detector, and the later just trains a new detector. Without loss of generality, we choose the later strategy but the former is also applicable in our framework. We re-train a detector in the HOG feature space by using SVM with probabilistic output to obtain Φ^c. Finally, for each location p, we combine both low level features and appearance-based detectors for classification. Consider a foreground blob B in the test image with the center position $p = C\left(B\right)$. We classify B by computing

$$\hat{c} = \arg\max_c \left(\prod_{\tau \in \{md, ar, size, sp,\}} \Theta^c_{\tau;p}(B)\right) \Phi^c_p\left(B\right), \quad (6)$$

where $\Phi^c_p\left(B\right)$ is the probability of classifying B as category c by the appearance detector Φ^c_p , and $\Theta^c_{\tau;p}\left(\cdot\right)$ is the prior probability drawn at the location p of the MASS feature τ,

$$\Theta^c_{\tau;p}\left(\cdot\right) \sim \mathcal{N}(\mu^c_{\tau;p}, \sigma^c_{\tau;p}). \quad (7)$$

If the probability in eq. (6) is larger than a threshold ε, we then classify B as category c ; otherwise B is determined as not belonging to any of the categories $c \in \{1, ..., C\}$.

3. EXPERIMENTAL RESULTS

We use the proposed algorithm for pedestrian detection and multi-class object detection/classification. The results are presented in Sections 3.1 and 3.2, respectively.

3.1. Pedestrian detection

The evaluation is conducted on the public dataset *Central Pedestrian Crossing Sequences* [14]. In the experiment, the training and testing sequences contain 301 frames with 560 pedestrians and 601 frames with 340 pedestrians, respectively, consistent to the setting of [10] for fair comparison.

We compare our method with both generic (off-line trained) detectors and scene/location-based approaches. The generic detectors in the comparison include the cascaded Haar-like features [15], HOG+SVM [1], deformable part-based model (DPM) [3] and the Fastest Pedestrian Detector in the West (FPDW) [2]. A method employing only simple blob features [4] is also compared. The location-adapted methods [16, 10] are used for comparison. The method in [16] learns a gird-based detector with fixed updating rules. An improved version is then presented in [10], which uses online boosting for feature selection. Both methods are scene and location dependent with fixed grids employed.

To quantify the detection performance, we adopt the F-measure [17] that is the harmonic mean between precision and recall as follows:

$$\mathrm{FM} = \frac{2 \times P \times R}{P + R} \quad (8)$$

The results are shown in Table 1. As discussed, it is difficult to train a generic detector that can work robustly for every scene. Hence, the location-adaptive methods (including both ours

Table 1: Pedestrian detection results sorted by F-measure.

Method	R	P	FM
Ours	0.98	0.91	0.95
Sternig *et al.* [10]	0.63	0.96	0.76
Grabner *et al.* [16]	0.55	0.88	0.68
FPDW [2]	0.53	0.72	0.61
BGM [4]	0.93	0.24	0.38
DPM [3]	0.26	0.32	0.29
Viola and Jones [15]	0.31	0.13	0.18
Dalal and Triggs [1]	0.19	0.15	0.17

Table 2: Specification of the training and testing data. Dataset A is only for training the generic detector.

		A	B	PETS01
Train	PED#	59719	2799	31178
	Bike#	5450	1207	1432
	Car#	33382	8810	2734
Test	PED#		3528	44992
	Bike#		897	1358
	Car#		4071	6166

and those in [16, 10]) perform better than the generic detectors. Among the generic detectors, FPDW performs the best as can be seen in Table 1. DPM [3] performs the second best, but its F-measure drops fast as there is a large gap between the accuracies of them. Due to the problem of appearance inconsistency, the overall performance of generic detectors is unsatisfied. The detector using only simple foreground-blob features [4] performs better than many generic detectors.

Among the location-adaptive methods, it can be seen that our method performs better than the others [16, 10]. It is because our method employs a prior scene segmentation step to discover and group feature-similar regions, which is useful for more discriminant training.

3.2. Multi-class object classification

To evaluate the performance of multi-class object detection, we conduct experiments on two datasets. One is the publicly available dataset PETS2001. In this experiment, we have $C = 3$ categories, namely, pedestrian, vehicle, and bicycle/scooter (treated the same). We select the training sequences from camera#1 in dataset 1, 2 and 4 and testing sequences from camera#2 in dataset 3, respectively. The other is our own datasets **A** and **B** collected in a campus shown in Fig. 4. The amount of training and testing data are depicted in Table 2.

Because most location-based approaches [8, 9, 10] are designed for 2-class problems, in this experiment, we compared our approach to other two methods, generic object detector [1] and location-based method [11] for multi-class classification. In this experiment, the generic detectors are learned by using the images from dataset A as the training samples. Then, the obtained genetic detector is used for the other datasets, PETS2001 and dataset B. As shown in Table 3 and Table 4, the overall accuracy can be significantly improvement by the location-adaptive approach of both our and Zhang *et al.* [11] methods. In comparison of ours and Zhang *et al.* [11], our method achieves better detection performance than the method in [11].

(a) (b) (c)

Fig. 4: Experimental datasets: (a) PETS2001, (b) dataset A, and (c) dataset B

Similar to the pedestrian detection experiments, our approach can build category-dependent coherent regions. Hence, unlike the method that subdivides the scene into size-fixed subregions for all the categories [11], our method can conduct location-adaptive detectors more effectively.

Table 3: Comparison of the confusion matrices on PETS2001.

	Our Method			Zhang *et al.* [11]			generic detector [1]		
	Ped	Bike	Car	Ped	Bike	Car	Ped	Bike	Car
Ped	97.8	1.55	0.64	78.89	19.57	1.52	91.3	8.3	0.3
Bike	26.41	70.68	2.90	22.75	40.94	36.30	34.9	51.0	13.9
Car	4.32	3.78	91.89	55.10	13.72	31.17	3.6	7.3	89.0
Overall	86.79 %			50.33 %			77.14%		

Table 4: Comparison of the confusion matrices on dataset B.

	Our Method			Zhang *et al.* [11]			generic detector [1]		
	Ped	Bike	Car	Ped	Bike	Car	Ped	Bike	Car
Ped	93.8	4.66	2.33	99.17	0.50	0.30	86.8	13.1	0
Bike	40.20	51.90	7.89	34.62	37.41	27.96	52.9	47.0	0
Car	7.04	6.65	86.29	48.25	28.39	23.35	3.4	5.4	91.0
Overall	77.06 %			53.31 %			74.95%		

4. CONCLUSION

We have introduced an approach to conduct location-adaptive object detectors for specific scenes. By segmenting a scene into nonuniform superpixels based on scene-dependent features, better location-aware object detectors can be learned. Experimental results show that our approach performs better than both generic detectors and other location-based object detectors with more semantically consistent training samples discovered and employed in the proposed scheme.

5. REFERENCES

[1] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE CVPR*, 2005, vol. 1, pp. 886–893.

[2] Piotr Dollár, Serge Belongie, and Pietro Perona, "The fastest pedestrian detector in the west.," in *BMVC*. Citeseer, 2010, vol. 2, p. 7.

[3] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE CVPR*, 2008, pp. 1–8.

[4] Nigel JB McFarlane and C Paddy Schofield, "Segmentation and tracking of piglets in images," *MVA*, vol. 8, no. 3, pp. 187–193, 1995.

[5] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele, and Tomaso Poggio, "Full-body person recognition system," *Pattern Recognition*, vol. 36, no. 9, pp. 1997–2006, 2003.

[6] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE CVPR*, 2009, pp. 951–958.

[7] Meng Wang, Wei Li, and Xiaogang Wang, "Transferring a generic pedestrian detector towards specific scenes," in *IEEE CVPR*, 2012, pp. 3274–3281.

[8] Peter M Roth, Sabine Sternig, Helmut Grabner, and Horst Bischof, "Classifier grids for robust adaptive object detection," in *IEEE CVPR*, 2009, pp. 2727–2734.

[9] Sabine Sternig, Peter M Roth, and Horst Bischof, "Online inverse multiple instance boosting for classifier grids," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 890–897, 2012.

[10] Sabine Sternig, Peter M Roth, Helmut Grabner, and Horst Bischof, "Robust adaptive classifier grids for object detection from static cameras," in *Proceedings Computer Vision Winter Workshop*, 2009.

[11] Zhaoxiang Zhang, Kaiqi Huang, Yunhong Wang, and Min Li, "View independent object classification by exploring scene consistency information for traffic scene surveillance," *Neurocomputing*, vol. 99, pp. 250–260, 2013.

[12] Chris Stauffer and W Eric L Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE CVPR*, 1999, vol. 2.

[13] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *IEEE TPAMI*, vol. 23, no. 11, pp. 1222–1239, 2001.

[14] Bastian Leibe, Konrad Schindler, and Luc Van Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *IEEE ICCV*, 2007, pp. 1–8.

[15] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE CVPR*, 2001, vol. 1, pp. I–511.

[16] Helmut Grabner, Peter M Roth, and Horst Bischof, "Is pedestrian detection really a hard task," in *Proc. IEEE Intern. Workshop on PETS*, 2007, pp. 1–8.

[17] Shivani Agarwal, Aatif Awan, and Dan Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE TPAMI*, vol. 26, no. 11, pp. 1475–1490, 2004.