Human Centric Object Detection in Highly Crowded Scenes

Genquan Duan, Haizhou Ai
Computer Science and Technology Department,
Tsinghua University, Beijing 100084, China
ahz@mail.tsinghua.edu.cn

Takayoshi Yamashita[1], Shihong Lao[2]
Development Center,
[1]OMRON Corporation,SHIGA 525-0025, JAPAN
[2]OMRON Social Solutions Co., LTD., Kyoto 619-0283, JAPAN

***Abstract*—In this paper, we propose to detect human centric objects, including face, head shoulder, upper body, left body, right body and whole body, which can provide essential information to locate humans in highly crowed scenes. In the literature, the approaches to detect multi-class objects are either taking each class independently to learn and apply its classifier successively or taking all classes as a whole to learn individual classifier based on sharing features and to detect by step-by-step dividing. Different from these works, we consider two issues, one is the similarities and discriminations of different classes and the other is the semantic relations among them. Our main idea is to predict class labels quickly using a Salient Patch Model (SPM) first, and then do detection accurately using detectors of predicted classes in which a Semantic Relation Model (SRM) is proposed to capture relations among classes for efficient inferences. SPM and SRM are designed for these two issues respectively. Experiments on challenging real-world datasets demonstrate that our proposed approach can achieve significant performance improvements.**

***Index Terms*—object detection, multi-classes, crowded scenes**

I. INTRODUCTION

Object detection is extensively studied in computer vision because of its wide applications in practical system like visual surveillance and traffic monitoring, where real time and high accuracy performance are required. The main challenges are occlusions, varying illumination, background clutter and viewpoint changes, etc. Many approaches have been proposed to detect one object category like human [1][2][3][4][5] or multiple object categories [6][7]. Apparently, object categorization is more general and useful in image and video analysis since it provides more information of the scenes.

In general, object categorization deals with the detection of many rarely related object classes, for example, human, bike, dog and others. To cope with this problem, there are mainly two kinds of approaches. One is to take each class independently and detect objects by the relevant classifiers. Felzenszwalb et al. [6] proposed a deformable part model for each class. They adopted HOG [1] as weak features and latent SVM as learning algorithm to learn a root filter and six part filters for each category. The other one is to take all classes as a whole and detect multi-class objects based on sharing features and step-by-step dividing. Torralba et al. [7] proposed a joint boosting procedure for multi-class object detection which reduced both the computational and sample complexity by finding common features shared across the classes. In such researches, the relations of different object categories are usually not considered due to the difficulties in modeling their relations, only very recently Yao and Fei-Fei [8] considered this problem in human-object interaction activities at a very high level, where they proposed to model the mutual context of object and human pose.

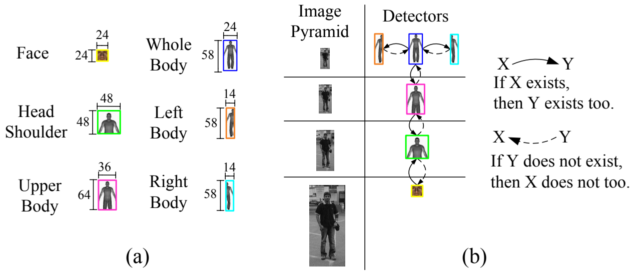

Fig. 1. Human centric classes in (a) and some semantic relations in (b).

In this paper, we address a restricted problem named as *human centric object* detection, where multiple classes are derived from human, such as face, head shoulder, upper body, left body, right body and whole body as shown in Fig. 1(a). In the literature, similar problems have been discussed in [2][5][9]. Multiple human parts are detected independently, which are combined based on a MAP formulation for human detection in [5], and used for the inference of part constellations for pose estimation in [9]. But it is well known that parts are less discriminative than the whole when parts and whole are from the same scale, which causes more difficulties into learning. Furthermore, compact relations are proved to be efficient in [2]. It is easy to define such relations for parts from the same scale of human but hard for those from different scales. In our problem, semantic relations among parts such as Fig. 1(b) play an important role, where detectors of different classes may work on different scales of the image pyramid during scanning search, but they are related in the scale space. Therefore, a good algorithm should consider both discriminations of parts from different scales and relations among them.

Considering the above mentioned issues, in this paper we deal with human centric object detection and make the following contributions: 1) Salient Patch Model (SPM) to predict class labels for human centric classes from different scales of human sub-samplings; 2) Semantic Relation Model (SRM) to capture semantic relations among instances of human centric classes that may appear in different scales for efficient inferences; and 3) our problem formulation provides

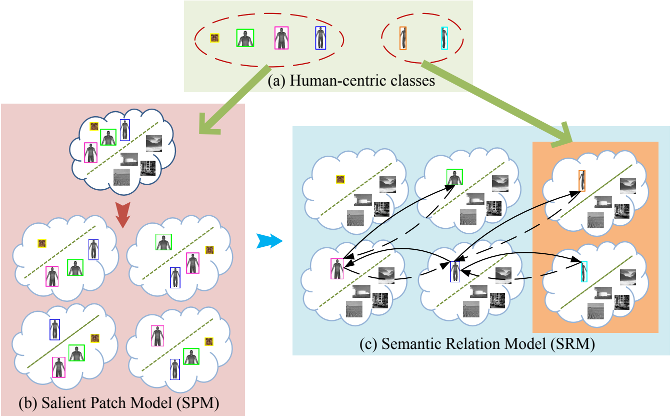
(a) Human-centric classes
(b) Salient Patch Model (SPM)
(c) Semantic Relation Model (SRM)

Fig. 2. Overview of our approach. (a)Human centric classes derived from same or different scales of human sub-samplings (As in Fig. 1(a), left body, right body and whole body are from the same scale, but face, head shoulder and upper body are from larger and different scales). (b)SPM to predict class labels through two steps, rejecting easy negatives first and then classifying one class from others. (c) SRM to detect objects accurately by detectors of predicted classes in which relations among classes are considered. Due to the large difference between face and other classes, the relation between face and head shoulder in Fig. 1(b) is ignored.

a simple but effective way to combine multi-class detectors without retraining on large datasets. An overview is illustrated in Fig. 2, which will be detailed in Sec.II; experiments are presented in Sec.III; and conclusions are given in Sec.IV.

II. OUR APPROACH

A. Problem Formulation

Boosting [10] provides a simple way to fit additive models sequentially of the form:

$$H(x,c)=\sum_{t=1}^{T}h_t(x,c)-b \tag{1}$$

where x, c, h_t and b are respectively a sample, a class label, a weak classifier and a threshold. T is the number of boosting rounds. The final decision $\langle H(x,c)\rangle$ is 1 (positive) if $H(x,c)>0$, or 0 (negative) otherwise.

As discussed in the previous section, in order to handle the proposed human centric object detection problem, our main idea is to predict class labels quickly first, and then do detection accurately using detectors of predicted classes where semantic relations among classes should be considered for efficient inferences. Formally, this problem is generally formulated based on [10] as:

$$H(x,c)=\begin{cases}H_2(x,c), & H_1(x,c)>0, H_2(x,\phi), \phi\in\psi_c\\ -\infty, & \text{otherwise.}\end{cases} \tag{2}$$

$H_1(x,c)$ predicts whether c is a possible class label or not and $H_2(x,c)$ is the strong classifier of class c. ψ_c is the class label set whose elements are related with class c. Given a sub-image, for $\phi\in\psi_c$ if c exists, then ϕ exists too, and if ϕ does not exist, then c does not exist too. In this formulation, only H_1 should be trained, while learned detectors (H_2) can be applied directly without retraining. Our intension is to mainly consider the similarities and discriminations of different classes in H_1 and the semantic relations among them in H_2 for both speedup and performance improvements.

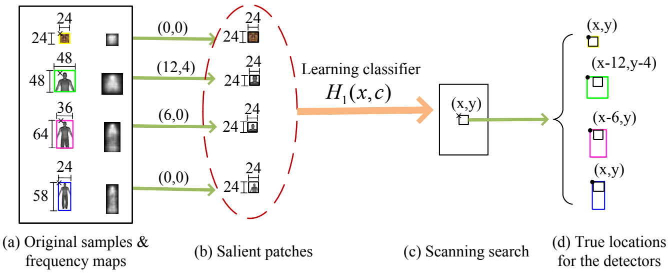

Fig. 3. An example of SPM. Salient patches in (b) are extracted from original samples in (a) at a reference point (0,0), (12,4), (6,0) and (0,0) respectively. The corresponding locations for their detectors are at (x,y), $(x-12,y-4)$, $(x-6,y)$ and (x,y) in (d), when checking $H_1(x,c)$ at (x,y) in (c).

B. Salient Patch Model (SPM)

SPM is proposed for the prediction of class labels, in which similarities and discriminations of different classes should be considered. In order to achieve this goal, a direct way is to extract and select some features, but such a process need to be elaborate. Our approach is based on salient patches, which should be able to represent their origins and preserve as much discriminative power as possible.

Extraction of salient patches. A salient patch is defined as a rectangle $P(sx,sy,ex,ey)$, where (sx,sy) is the left-top and (ex,ey) is the right-bottom. For a class c, that is to find the most salient one:

$$P^*=\arg\max_P E(P;c) \tag{3}$$

where E is a saliency measuring function. E is usually defined to evaluate the significance from the regions around it, however, it is defined on a *frequency map* (FR) in our paper, where $FR(x,y)$ denotes the frequency of location (x,y) containing weak features in a detector. As a detector is learned on thousands of labelled positives and negative images, $FR(x,y)$ reveals the importance of (x,y) to this class. Some frequency maps are illustrated in Fig. 3(a) right. Given the frequency map FR_c of class c, we explicitly define:

$$E(P;c)=\sum_{x=sx}^{ex}\sum_{y=sy}^{ey}FR_c(x,y) \tag{4}$$

In implementation, we assume that the salient patches are in the same size, $w\times h$, where $w=\arg\min w_i$, $h=\arg\min h_i$ and $w_i\times h_i$ is the sample size for class c_i. In other words, $ex=sx+w-1$ and $ey=sy+h-1$. Then, for each positive class, we can find the best starting point (sx,sy) by enumerating locations. Note that the salient patch of background is meaningless and negatives with the size $w\times h$ are extracted from negative images.

Class label prediction. Salient patches extracted from original samples form a new sample set as shown in Fig. 3(b). They are utilized to predict class labels in two steps.

The first is to reject easy negatives. Our idea is to mine the similarities of all classes by mixing all positive samples as the positive set, taking background samples as the negative set, and then learning a detector $H_{1,1}(x) = \sum_{t=1}^{T_1} h_{1,t}(x)$.

The second is to classify one class from others. Our idea is to mine the discrimination of each class by taking samples of this class as the positive set and samples of other classes as the negative set, and then learn a detector $H_{1,2}(x,c) = \sum_{t=1}^{T_2} h_{2,t}(x,c)$.

The two steps correspond to the top and bottom of Fig. 2(b) respectively. Then we can achieve $H_1(x,c) = H_{1,1}(x) + H_{1,2}(x,c)$. If $H_1(x,c)$ gives a positive decision, then c is a possible class label.

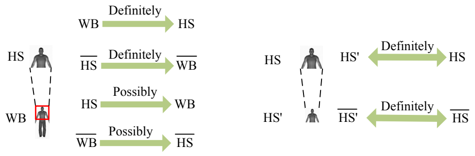

Fig. 4. Examples of Directional Relation (left) and Bidirectional Relation (right). Please see the text for details.

C. Semantic Relation Model (SRM)

After obtaining possible classes, the next step is to do detection accurately by involved detectors, where their semantic relations can be utilized for further speedup and performance improvements. Our model is Semantic Relation Model.

There are four **basic relations** between two classes. The 1st is *Independent Relation* (IR), meaning that classes are independent, which is the most common one by ignoring relations among them. The 2nd is *Mutex Relation* (MR), meaning that two classes will not exist at the same time. For example, if a regions is classified as dog, then the test of bicycle can be neglected because of so much difference between dog and bicycle. MR exists in any two unrelated classes. Compared to IR, it is possible to ignore some detectors of MR. But as the class label is unknown before testing, it is likely to test all of them. The 3rd is *Directional Relation* (DR), e.g. whole body (WB) and head shoulder (HS) in Fig. 4(left), denoted as WB→HS, meaning that if the possible class is WB, then it definitely is HS, and if it is not HS, it is definitely not WB. Due to occlusions, if it is HS, it is possibly WB and if it is not WB, it is possibly not HS. Generally, DR exists widely between part and whole. Compared to IR and MR, detectors of DR can speed up detection processes by being arranged in a proper order. The last is *Bidirectional Relation* (BR), meaning that one class can definitely predict the other one. An example of BR is shown in Fig.4 (right) with two scales of head shoulders. Assume that their detectors are accurate enough and the only difference is that they work on different scales of an image pyramid. BR can be taken as two DRs and detectors of BR can help each other during detection. Furthermore, as all detectors are learned separately, the consideration of these semantic relations can speed up detection without loss of performances.

These relations are **applied** as follows. The relations among face, head shoulder, upper body and whole body in Fig. 2(b) is MR, since they differ a lot in appearance. The solid and dash arrows in Fig. 1(b) and Fig. 2(c), and the relations in Equ. 2 are all DRs. Except relations mentioned above, other relations among classes in Fig. 2 are IRs in our paper. It is easy to apply detectors of MR and IR, but an efficient usage of DR needs the correspondences between detectors of DR. Given two detectors D_i and D_j for class c_i and c_j with sample sizes of $w_i \times h_i$ and $w_j \times h_j$, assume their relation is DR, $D_i \rightarrow D_j$. In a K-layer image pyramid $\{s_1, s_2, \ldots, s_K\}$, s_k is the scale of the k^{th} layer. Corresponding to location (x,y) of scale s_k applying D_j, the most approximate scale and location to possibly apply D_i are s_{i^*} and (x',y'), where $i^* = \arg\min_u s_u - s_k \times \frac{w_i}{w_j}$, $x' = x \times \frac{s_{i^*}}{s_k}$ and $y' = y \times \frac{s_{i^*}}{s_k}$. Our strategy is:

```
Algorithm 1. Learning algorithm of our approach.
Given Sample set $S = \{(x_i, y_i)|x_i \in \chi, y_i \in C\}$, where $\chi$ is sample space and $C$ is class label set.
    • Generate a new set $S' = \{(x_{i,P_{y_i}}, y_i)\}$, where $P_{y_i}$ is the salient patch for class $y_i$ and $x_{i,P_{y_i}}$ corresponds to the patch $P_{y_i}$ in $x_i$. Note that negatives $\{x_{i,P_{y_i}}|y_i = -1\}$ are extracted from negative images.
    • Learn $H_{1,1}(x)$ on positive set $S'_{1,+} = \{(x_{i,P_{y_i}}, y_i)|y_i \neq -1\}$ and negative set $S'_{1,-} = \{(x_{i,P_{y_i}}, y_i)|y_i = -1\}$.
    • Learn $H_{1,2}(x,c)$ for each class $c$, on positive set $S'_{2,+} = \{(x_{i,P_{y_i}}, y_i)|y_i = c\}$ and negative set $S'_{2,-} = \{(x_{i,P_{y_i}}, y_i)|y_i \neq c, y_i \neq -1\}$.
    • Learn $H_2(x,c)$ for each class $c$ independently on $S$.
Output: $H_1(x,c) = H_{1,1}(x) + H_{1,2}(x,c)$ and $H_2(x,c)$.
```

If D_j gives a *negative* decision, (x',y') of scale s_{i^*} is ignored for D_i;

If D_j gives *positive* decision, (x',y') of scale s_{i^*} is considered by D_i. In order to make it more robust, D_i is applied at a larger range, i.e. at location (x'',y'') of scale s_z, where $x' - T_1 \leq x'' \leq x' + T_1$, $y' - T_1 \leq y'' \leq y' + T_1$ and $i^* - T_2 \leq z \leq i^* + T_2$. T_1 and T_2 are set to 2 and 1 separately in our experiment. Fig. 5 shows a diagram, where (x,y) is in blue, (x',y') and possible (x'',y'') are in red.

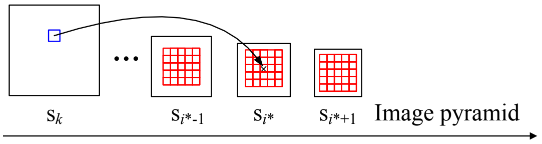

Fig. 5. Modeling scales and locations in the image pyramid. Please see the text for details.

D. Learning and post processing

We adopt Associated Pairing Comparison Features (APCFs) [3] based on Real Adaboost to learn detectors for prediction and detection. APCF describes invariance of color and gradient of an object to some extent and it contains two essential elements, Pairing Comparison of Color (PCC) and Pairing Comparison of Gradient (PCG). A PCC/PCG is a Boolean color/gradient comparison of two granules in which a granule is a square window patch. Please refer to [3] for more details. The entire learning algorithm of our approach is shown in Algorithm 1.

The detection results are extended to the corresponding human size for combination as shown in Fig. 7. Wu et al. [5] proposed a Bayesian combination method based on Maximum a Posteriori (MAP) formulation to combine part detection

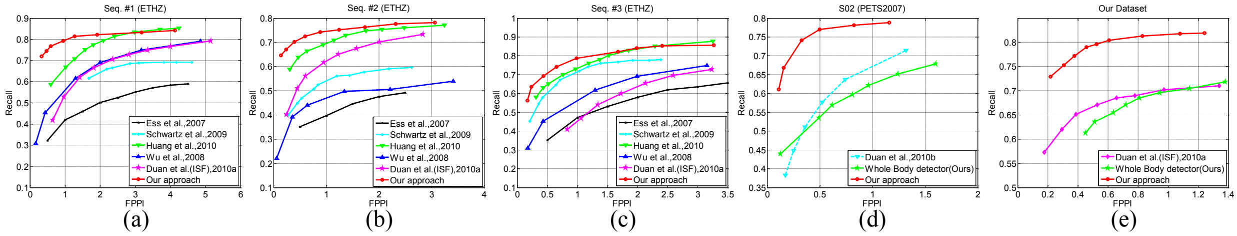

Fig. 6. Evaluation of our approach on multiple real world datasets.

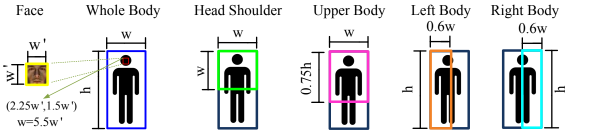

Fig. 7. Post processing. Detection results are extended to corresponding human sizes for combinations.

results with whole body detections, where they adopted a greedy method to seek the optimal state of multiple humans that best explains the detection responses. We use this efficient algorithm for combinations.

III. EXPERIMENTS

A. *Training*

We have collected $61,825$ faces and $27,938$ human parts for training. The target false positive rate of face detector is set to $1E10^{-7}$ and those of human parts detectors are set to $6E10^{-7}$. In order to use SPM to predict class labels, $H_1(x,c)$ is learned by at most two rounds of boosting containing at most 5 weak classifiers for each round with detection rate of 0.998. All experiments are conducted on an Intel Core(TM)2 2.33GHz PC with 2G memory.

B. *Evaluation*

Frequent occlusions and different viewpoints in real world crowded scenes rise many difficulties. We choose three challenging datasets, ETHZ [11], PETS2007 [12] and our collected dataset to compare our approach with several state-of-the-art algorithms. All these sequences are processed frame by frame, without usage of any temporal information. We select False Positive Per Image (FPPI) as the evaluation criterion. When the intersection between a detection response and a groundtruth box is larger than 50% of their union, we consider it to be a successful detection. Only one detection per annotation is counted as correct. Test datasets are independent from our training set.

1) ETHZ: ETHZ dataset [11] consists of four 640×480 video sequences at 15 frames/second (one for training and three for testing) captured on a moving platform in very cluttered environments. Only images from the left camera are used in our experiment. Seq.#1 contains 999 frames with $5,193$ humans; Seq.#2 contains 450 frames with $2,359$ humans; Seq.#3 contains 354 frames with $1,828$ humans. We do not use scene geometry like [11], or ground plane estimation method like [4][13]. The ROC curves are shown in Fig. 6(a-c). Our approach outperforms [2][11][14][13] on all three videos. Particularly, the significant improvement of our approach than [2] in which similar human parts were used is because: 1) we use larger sizes of head shoulder and upper body which can reduce false positives; 2) our efficient SRM models the scale and location in the image pyramid; 3) we use a Bayesian combination method to combine the part results. [4] achieved the highest performance most recently on this dataset. Our approach outperforms [4] at less than 1.5 FPPI and achieves similar results as [4] at larger than 1.5 FPPI on this dataset. In particular, compared with [4] at FPPI=1, our approach increases the detection rate by 8%, 4.9% and 5.3% on the three videos respectively.

2) PETS 2007: PETS2007 dataset [12] contains 9 sequences S00$\sim$S08 (720×576 pixels at 30 fps) and each sequence has 4 fixed cameras. There are 3 scenarios, loitering, attended luggage removal and unattended luggage. We only use S02 of the third camera of the attended luggage removal scene for evaluation, which is labelled every five frames manually and the groundtruths of the internal unlabelled frames are achieved through interpolation. There are $4,500$ frames and $17,067$ labelled humans in total. Because face is always invisible in this scene, we apply a simpler version of our approach in which face is ignored, and the ROC curves shows in Fig. 6(d). Although Duan et al. [15] utilized both appearance and motion information but we do not, our approach is significantly better than [15], which is mainly because many humans are occluded, but Duan et al. [15] only detected whole bodies and did not deal with occlusions specifically. Our approach improves the detection rate by 10% than [15] at FPPI=1.

3) Our dataset: Our dataset is recorded at a resolution of 720×576 pixels at 30 fps in a very crowded scene using a hand held camera. All frames are labelled manually, resulting in 400 frames and $3,846$ annotations. A large number of trees and grasses may result in false positives for object detection. Slight motion blur may also lead detectors to fail. Furthermore, there are much heavier occlusions in our dataset than ETHZ dataset. *We intend to release it soon*. We compare our approach with our whole body detector and ISF [2] in Fig. 6(e). Our approach outperforms our whole body detector and ISF [2], and improves the detection rate by 11% than [2] at FPPI=1.

C. *Speed Comparison*

The average speed is evaluated on 640×480 images, for which the default detection configuration is to search human-centric classes from 24 to 256 pixels wide. Although auxiliary data structures for implementations of SPM and SRM need extra time, our program using one thread costs about 1.62s, while running all detectors independently needs about 3.17s,

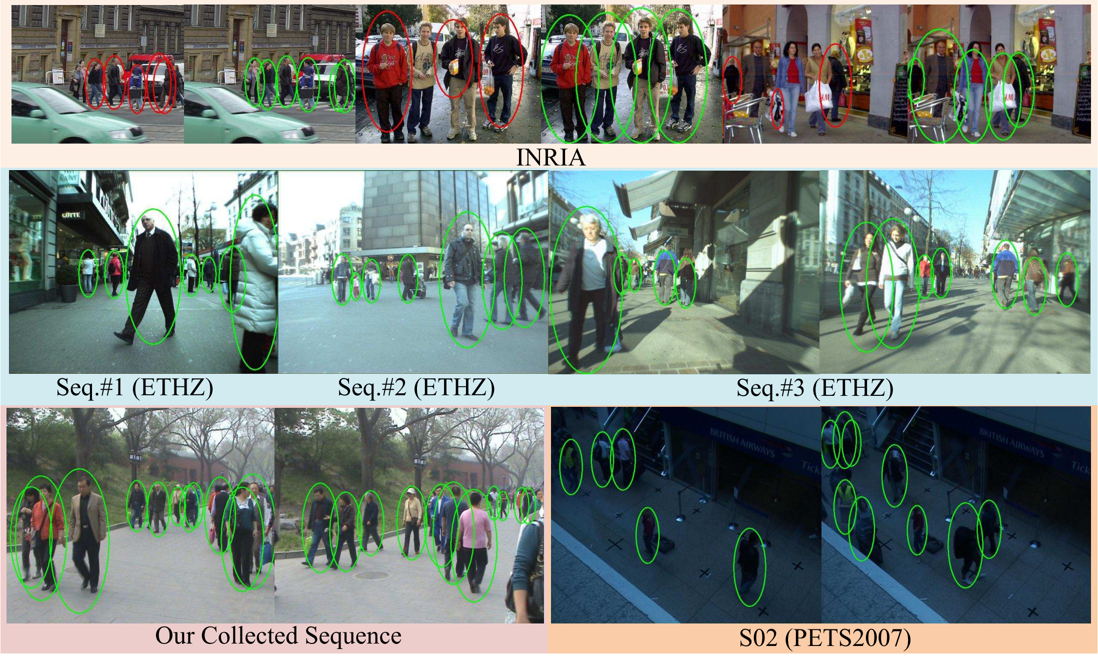

Fig. 8. Detection results of our approach on INRIA, ETHZ, PETS2007 and our collected sequence. The first row compares ISF [2] in red and ours in green, which shows that our approach is more robust to occlusions, illuminations changes and image blur.

as shown in Tab. I. Our approach is a little slower than 1.4s of ISF [2], but more accurate than ISF. Furthermore, our approach is much faster than 2.5s of [13] using ground plane estimation and multi-threads. Fig. 8 shows some results of our approach.

TABLE I
SPEED COMPARISON ON 640 × 480 IMAGES (MS).

FA	HS	UB	LB	RB	WB	IND	Our
140	700	440	610	610	670	3,170	1,620

IV. CONCLUSION

In this paper, our target is to handle human centric object detection in highly crowded scenes. We propose a detector based Salient Patch Model to predict class labels first and then Semantic Relation Model to capture the semantic relations among classes for fast and accurate detection. Experiments on challenging real-world datasets demonstrate that our proposed approach is robust to occlusions and different viewpoints, and can achieve significant performance improvements.

To extend our approach for surveillance scenes, such as entrances of buildings or aisles of subways, some future work should be done. First, a new combination algorithm could be developed to combine better the detection responses to prevent unexpected merging that results in missing detections. Second, more information can be integrated, such as ground plane estimation [13] which can speed up the current system, and motion information which can improve both speed and accuracy. Moreover, scene structure analysis [11] can be combined for further improvement.

ACKNOWLEDGEMENTS. Prof. Haizhou AI is partially supported by National Science Foundation of China under grant No.61075026. Mr. Shihong LAO is partially supported by "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Special Coordination Fund for Promoting Science and Technology of MEXT, the Japanese Government.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[2] G. Duan, H. Ai, and S. Lao, "A structural filter approach to human detection," in *ECCV*, 2010.

[3] G. Duan, C. Huang, H. Ai, and S. Lao, "Boosting associated pairing comparison features for pedestrian detection," in *9th Workshop on Visual Surveillance*, 2009.

[4] C. Huang and R. Nevatia, "High performance object detection by collaborative learning of joint ranking of granule features," in *CVPR*, 2010.

[5] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *ICCV*, 2005.

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008.

[7] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *CVPR*, 2004.

[8] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *CVPR*, 2010.

[9] R. Ronfard, C. Schmid, and B. Triggs, "Learning to parse pictures of people," in *ECCV*, 2002.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," 1998, technical report, Dept. of Statistics, Stanford University.

[11] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," in *ICCV*, 2007.

[12] "Pets 2007 dataset," http://www.cvg.rdg.ac.uk/PETS2007/. [Online]. Available: http://www.cvg.rdg.ac.uk/PETS2007/

[13] B. Wu, R. Nevatia, and Y. Li, "Segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," in *CVPR*, 2008.

[14] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *ICCV*, 2009.

[15] G. Duan, H. Ai, and S. Lao, "Human detection in video over large viewpoint changes," in *ACCV*, 2010.