A Structural Filter Approach to Human Detection

Genquan Duan[1], Haizhou Ai[1], and Shihong Lao[2]

[1] Computer Science & Technology Department, Tsinghua University, Beijing, China
ahz@mail.tsinghua.edu.cn
[2] Core Technology Center, Omron Corporation, Kyoto, Japan
lao@ari.ncl.omron.co.jp

Abstract. Occlusions and articulated poses make human detection much more difficult than common more rigid object detection like face or car. In this paper, a Structural Filter (SF) approach to human detection is presented in order to deal with occlusions and articulated poses. A three-level hierarchical object structure consisting of words, sentences and paragraphs in analog to text grammar is proposed and correspondingly each level is associated to a kind of SF, that is, Word Structural Filter (WSF), Sentences Structural Filter (SSF) and Paragraph Structural Filter (PSF). A SF is a set of detectors which is able to infer what structures a test window possesses, and specifically WSF is composed of all detectors for words, SSF is composed of all detectors for sentences, and so as PSF. WSF works on the most basic units of an object. SSF deals with meaningful sub structures of an object. Visible parts of human in crowded scene can be head-shoulder, left-part, right-part, upper-body or whole-body, and articulated human change a lot in pose especially in doing sports. Visible parts and different poses are the appearance statuses of detected humans handled by PSF. The three levels of SFs, WSF, SSF and PSF, are integrated in an embedded structure to form a powerful classifier, named as Integrated Structural Filter (ISF). Detection experiments on pedestrian in highly crowded scenes and articulated human show the effectiveness and efficiency of our approach.

1 Introduction

Human detection has attracted much attention and significant progresses have been achieved in [1][2][3][4][5][6][7][8][9][10][11][12]. However, highly accurate and real time human detection is still far from reality. There are mainly two difficulties for human detection: 1) Humans are highly articulated objects which change a lot in view, pose, size, position, etc. 2) Lots of things, including all around, may cause occlusions, like accessories (backpacks, briefcases, bags, etc.), or other persons. Especially in crowded scenes, humans always obscure each other.

Various algorithms are proposed for object detection to deal with occlusions or articulated poses. Deformable part model based on HOG features combined with a latent SVM was proposed in [2] for object detection, in which a root filter and several parts models are learned for each object category that can

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part VI, LNCS 6316, pp. 238–251, 2010.

Fig. 1. Some combined results of Structural Filter approach to detect occluded pedestrian and articulated human

detect objects with some pose changes. A generic approach based on pictorial structure model was proposed in [7] to estimate human poses where a classifier is learned for each part and it infers locations of each part by a graph model. "Bags-of-words" method is widely applied to category [13] and detection [4][14] in computer vision. The approach in [14] is able to represent objects sparsely. Implicit Shape Model was proposed in [4] to detect pedestrians in crowed scenes in a bottom up way by a collection of visual words.

Holistic detectors are often limited when some parts are missing and it is even impractical to learn a holistic detector for objects with very large deformations. Therefore some approaches turn to parts/components to handle occlusions. Multiple occluded humans in [3] were detected by a Bayesian combination of part detectors where three types of body parts, head-shoulder, torsos and legs, are used. This approach was extended in [6] where a part hierarchy of an object class is defined and each part is a sub-region of its parent. There are also some component based methods to detect object through integrating part detectors by matching isomorphic graphs [15]. This kind of approach is more robust to occlusions where holistic object detectors will fail. But a critical issue here is how to integrate part detectors because parts tend to be less discriminative and part detectors are prone to producing more false positives. Some approaches rely on geometrical constraints of parts to handle false positives. But parts are easily missed due to occlusions, which often makes the constraints invalid.

Inspired by previous works in [3][6][15], in considering the relations among local regions, we propose a novel way to integrate part detectors, named as Structural Filter (SF) for object detection. Our aim is to handle occluded and articulated human detection in one framework and some combined results are shown in Fig. 1. A SF is a set of detectors which is able to infer what structures a test window possesses, where the structures could be *words*, *sentences* or *paragraphs*, corresponding to Word Structural Filter (WSF), Sentence Structural Filter (SSF) and Paragraph Structural Filer (PSF) respectively. A test window is positive if at least one detector in PSF provides a positive decision at last. We carry out some experiments on partially occluded pedestrian detection and articulated (multi-pose) human detection to demonstrate the effectiveness and efficiency of our approach.

The rest of this paper is organized as follows. The following section gives related work; Section 3 and Section 4 presents hierarchical structures of objects

and Structural Filter separately; some experiments are carried out on pedestrian detection in crowded scenes and multi-pose human detection in section 5; and the discussion and the conclusion are given in the last two section.

2 Related Work

The first thing for human detection with occlusions and articulated poses is to model humans. Various models have been proposed to represent humans such as pictorial structure model [8], star model [7], multiple tree model [9], non-tree model [10] and part hierarchy model [6].

Pictorial structure [8] was proposed to represent humans by a joint configuration of parts in which an articulated model with 14 joints and 15 body parts was used and classifiers for each part using simple image features (first and second Gaussian derivatives) were learned. More discriminative detector for each part was proposed based on star model in [7]. In order to capture additional dependencies between body parts, multiple tree models was used in [9] to alleviate the limitations of a single tree-structured model. Non-tree model was proposed in [10] to enforce any type of constraints. These four typical models are proposed for pose estimation problem.

A part hierarchy model was proposed in [6] for detection and segmentation of partially occluded objects, in which parts are placed in specific locations and each part is a sub-region of its parent. Placing parts in specific locations is a convenient method for detection problem and provides the potential for sharing weak features. Detector ensemble [15] was proposed for face detection in heavy occlusions where sub-structures are applied to make each part more discriminative.

Following the works in [3][6][15], we build up a hierarchical structure of human and propose a Structure Filter approach to integrate part detectors to handle occlusions and articulated poses in one framework. The proposed hierarchical structure contains three levels, *words*, *sentences* and *paragraphs*, which combines the strengths of the approaches in [3][6][15]. The main differences are: 1) Parts are totally independent in [3] and each part is a sub-region of its parent in [6]. While in our method, *words* are basic units of objects. *Sentences*, consisting of *words*, are common sub structures of objects. *Paragraphs* are also composed of *words* and cover a set of *sentences*. *Paragraphs* correspond to the appearance statuses of detected objects, for example visible parts or particular poses in human detection. 2) Sub structures are also mentioned in [15] where a detector is learned for each sub structure and a detector ensemble which consists of a set of sub-structures gives a positive decision if at least one sub-structure is positive. While in our method, in addition to the two level structures, *words* and *sentences*, which are similar to [15], we add a *paragraph* level structure to learn a more robust detector to handle occlusions and articulated poses. 3) In our framework, a *word* is a general concept, which is a component of an object in a specific position and it can be a part, a component or a block. Furthermore, we propose a Structural Filter (SF) approach to integrate part detectors.

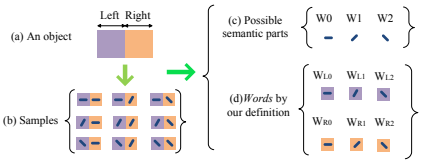

Fig. 2. The *words* and general semantic parts of an object. (a) An object with two regions, left and right. (b) Some samples of this object. (c) Possible general semantic parts. (d) *Words* defined in this paper.

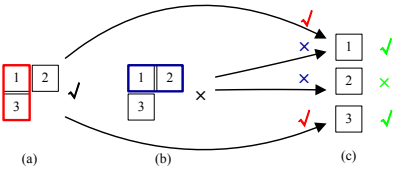

Fig. 3. SSF works interdependently. The red/blue blocks are two *sentences* shown in (a)/(b). Suppose a test window possesses the red structure but not the blue one, then the test result is that *word* "1" and word "3" are possessed as shown in (c).

Our contributions are summarized in three folds:

1) A three-level hierarchical object structure consisting of *words*, *sentences* and *paragraphs* in analog to text grammar is proposed for object detection.

2) A Structural Filter approach is proposed to integrate part detectors.

3) The proposed Structural Filter approach is a more general framework for object (rigid/non rigid) detection based on *words* (/parts/regions).

3 Three-Level Object Structure

3.1 Three-Level Object Structure

Words. A *word* is a component of an object in a specific block. In fact, the instance of *word* can be a part, a component or a block (of an area), which is similar as in [3][6][15]. Fig. 2 illustrates the difference of our *word* from general semantic part. It is worth mentioning that: 1) For humans, semantic "part" like head, leg, torso etc. may appear in different blocks due to no-rigid movement; 2) One block may contain several different parts of an object. In this paper, location is used as the first priority. One block may contain several parts for the Structural Filter approach to handle.

Sentences are sub-structures of an object which consist of *words*. A *word* is relatively less discriminative. Some of the *words* form a sub-structure which will be more discriminative as in [15]. Fig. 3 shows how SSF works interdependently.

Paragraphs corresponding to the appearance statuses of detected objects, are composed of *words* and cover a subset of *sentences*. Objects may show different statuses in different scenes. For example, parts of a pedestrian may be invisible in crowded scenes. The statuses of detected pedestrians can be head-shoulder, left-part, right-part, upper-body or whole-body.

3.2 Problem Formulation

Suppose an object O consists of N_W *words* which are denoted as a set $W = \{w_1, w_2, ..., w_{N_W}\}$. The *sentences* are $S = \{s_1, s_2, ..., s_{N_S}\}$ where N_S is the total

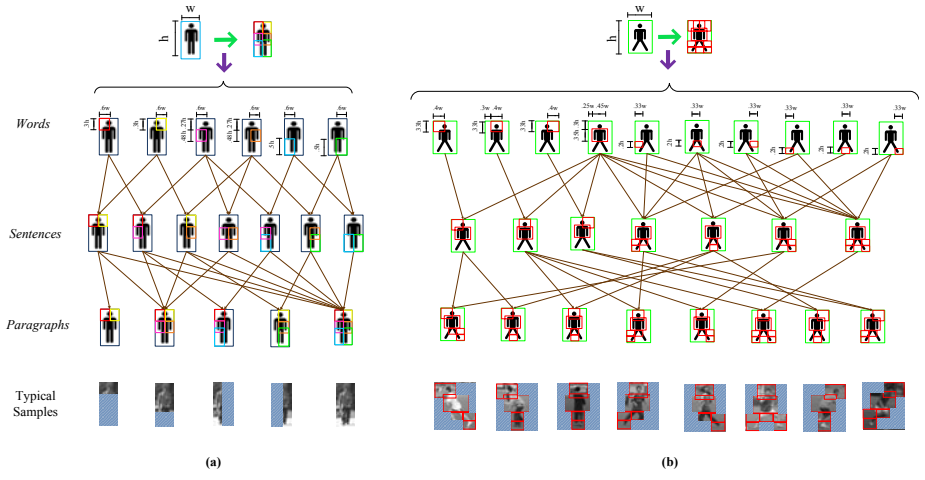

Fig. 4. Hierarchical structures of pedestrian and articulated human. (See Section 3.3 for details.)

number and each element $s_i (1 \leq i \leq N_S)$ is a subset of W. Similarly, *paragraphs* are represented as a set $P = \{p_1, p_2, ..., p_{N_P}\}$ where N_P is the number of the appearance statuses of detected objects and $p_i (1 \leq i \leq N_P)$ cover a set of S. *Sentences* are common sub-structures of an object which make *words* more discriminative and are used for inferences of *paragraphs*.

Each structure ϕ, either at *word* level, or *sentence* level or *paragraph* level, is associated with a detector with the detection rate $d(\phi)$ and the false positive rate $f(\phi)$. Our problem is to use a Structural Filter (SF) approach to integrate all these detectors. Each structure ϕ also has a missing tolerance parameter of parts, denoted as σ_ϕ, for integration.

3.3 Hierarchical Structures of Pedestrian and Articulated Human

As in [3][6], the simplest way to achieve *words* is to partition the sample space into some blocks according to heuristic knowledge. Hierarchical structures of pedestrian and articulated human are shown in Fig. 4 (a) and (b): 1^{st} row shows a sample space of pedestrian or articulated human; $2^{nd}/3^{rd}/4^{th}$ row shows *words*/*sentences*/*paragraphs* designed by prior knowledge; and 5^{th} row shows typical examples. The arrows between *words* and *sentences* show that *sentences* consist of *words*. Similarly, the arrows between *sentences* and *paragraphs* show that *paragraphs* cover a set of *sentences*.

Hierarchical structures of pedestrian. Pedestrians are relatively in strong cohesiveness. So we just evenly partition the sample space into six *words* shown in Fig. 4 (a). To deal with occlusions, we have defined five *paragraphs* of pedestrians, head-shoulder, upper-body, left-body, right-body and whole-body.

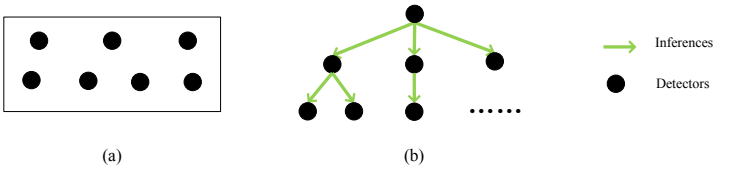

Fig. 5. Two typical methods to organize detectors, set method in (a) and tree method in (b)

Hierarchical structures of articulated human. Articulated (multi-pose) humans are more flexible than pedestrians. As a detection problem, all poses of humans as a whole are too difficult to deal with. We pay attention to a subset of poses where humans stand up on ground like walk, run etc. Mainly taking into account the varieties of heads and feet, we partition articulated human sample space into 10 *words* and define 8 *paragraphs* as shown in Fig. 4 (b).

4 Structural Filter Approach

4.1 The Definition of Structural Filters

A Structural Filter (SF) is a set of detectors which is able to infer what structures a test window possesses. Word Structural Filter (WSF) is composed of all the detectors for *words*, Sentence Structural Filter (SSF) is composed of all detectors for *sentences*, and so as Paragraph Structural Filer (PSF).

4.2 Three Level SFs: WSF/SSF/PSF

We adopt Real Adaboost [16] and Associated Pairing Comparison Features (APCFs) [1] to learn a cascade detector [17] for each structure (*word*, *sentence* or *paragraph*). 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 is a Boolean color comparison of two granules and a PCG is a Boolean gradient comparison of two granules in which a granule is a square window patch. See [1] for details.

There are typically two methods to organize detectors in each SF of different levels: 1) The set method, where detectors are organized as a set and give decisions separately as shown in Fig. 5 (a). With the set method, all detectors involved are processed. 2) The tree method, where detectors are organized as a tree as shown in Fig. 5 (b). With the tree method, child nodes will be processed only if their parent node gives a negative decision. The tree method is much faster than the set method in decision making since only parts of its detectors are used. The tree method also provides the possibilities of sharing of weak features. For example, if the whole-body is visible, there is no need to test on head-shoulder detector or other detectors.

WSF and SSF tend to describe parts of objects. Each detector in WSF or SSF gives a decision independently. So detectors in WSF or SSF are organized as a set method. Organizing detectors in PSF as a tree method or a set method depends on the object to be detected. Fig. 4 (a) shows 5 paragraphs of pedestrian where left-part, right-part and upper-body are sub-regions of whole-body and head-shoulder is a sub-region of upper-body, so detectors in PSF for pedestrians are organized with a tree method. Fig. 4 (b) shows 8 paragraphs of articulated human where there is no any paragraph which is a sub-region of another one. So detectors in PSF for articulated human are organized with a set method.

4.3 Integrated Structural Filter

To construct a final human detector, the three level SFs, WSF, SSF and PSF, are integrated together to form a powerful classifier, which is called Integrated Structural Filter (ISF). The integration can be represented as sequences of WSF, SSF and PSF, for example, WSF$\Longrightarrow$SSF$\Longrightarrow$PSF, PSF$\Longrightarrow$SSF$\Longrightarrow$WSF or PSF$\Longrightarrow$WSF$\Longrightarrow$PSF$\Longrightarrow$SSF$\Longrightarrow$PSF. Each SF (WSF, SSF or PSF) in a sequence is called a *stage*.

Structural Filter inference is the inference by one stage of ISF, which can be summarized as three steps:

Step 1. Suppose that η is currently the stage to be dealt with, where η is one of WSF, SSF and PSF. Let Ω denote the set containing all passed *words* before the processing of η. Note that at the very beginning, Ω contains all words.

Step 2. A detector in η will be carried out if $|\omega| \leq \sigma_\phi$, in which ϕ is the structure associated to this structure, σ_ϕ is the missing tolerance and $\omega = \{w|w \in \phi, w \notin \Omega\}$. If this detector gives a positive decision, then push the structure ϕ into the passed structure set κ. Note that: 1) If the detectors in η are organized as set method, all detectors will be considered. 2) Else the detectors in η are organized as tree method. If the root detector gives positive decision, then its child detectors will be ignored and otherwise they will be considered.

Step 3. After the processing of η , update the passed *word* set $\Omega = \{\alpha|\alpha \in \phi, \phi \in \kappa\}$.

After each stage, we can obtain the passed *word* set. Actually we concern about the final decision of a test object's appearance statuses, so the last stage is always PSF and the statuses of a test object can be easily inferred by the passed structure set of the last stage.

Integration of SFs. Two simple methods for the integration are: 1) Bottom-Up method, which may be implemented by WSF$\Longrightarrow$SSF$\Longrightarrow$PSF, is similar to sub-structure in [15]. Bottom-Up method can depict parts of objects well, and is particularly efficient to deal with occlusion and share weak features. But it needs more time to discard negatives. 2) Top-Down method, which may be implemented by containing only one stage, PSF, gives the last decision directly. Top-Down method can discard negatives fast. But it is not easy to share features in Top-Down method.

In order to take advantages of both Bottom-Up method and Top-Down method, three level SFs, WSF, SSF and PSF, are integrated in an embedded structure

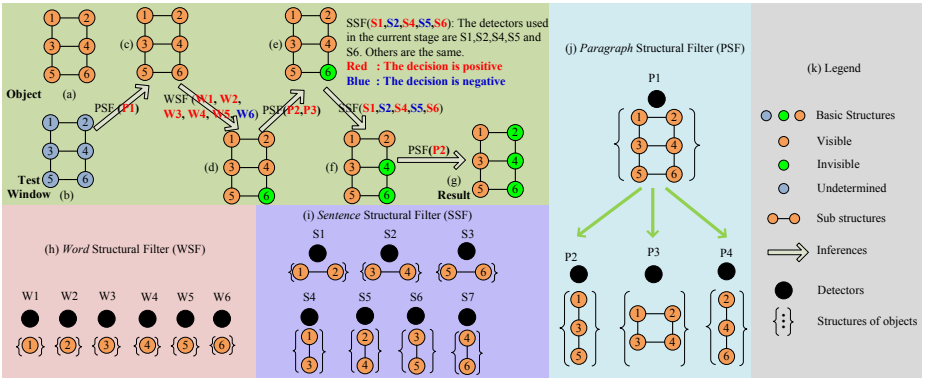

Fig. 6. An example of ISF. (See Section 4.3 for details.)

through five stages, PSF$\Longrightarrow$WSF$\Longrightarrow$PSF$\Longrightarrow$SSF$\Longrightarrow$PSF, to form an ISF for both pedestrian detection and articulated human detection. Here the three PSFs are different sets of detectors in different stages. PSF in the 1^{st} stage is to discard negatives quickly. PSF in the 3^{rd} stage is to integrate the detection results of WSF and to provide passed *words* for SSF. PSF in the 5^{th} stage gives the final decision, positive or negative. Missing tolerances of *words* are applied when two consecutive stages of SFs are integrated. A detector in a SF will be involved only if missing *words* are within tolerance. A testing sample is positive if at least one detector in PSF gives a positive decision at last. **The learning algorithm** for ISF is summarized in Table 1.

An example of ISF is shown in Fig. 6. An illustrative "object" is shown in (a) which consists of six *words*. The missing tolerance for each word, sentence and paragraph is assumed to be zero. (h) (i) and (j) show WSF, SSF and PSF respectively, where detectors in WSF and SSF are organized by set mothod and PSF are organized by tree method. A test window (b) is processed by ISF with the flow shown in (c)-(g). For example, WSF (W1, W2, W3, W4, W5, W6) are applied from (c) to (d) where the used detectors are W1, W2, W3, W4, W5 and W6. **Red/Blue** means that a detector gives **positive/negative** decision.

Table 1. Learning algorithm for ISF

Input: *Word* set **W**; *Sentence* set **S**; *Paragraph* set **P**; Sample set $R = \{(x_i, y_i)|x_i \in \chi, y_i = \pm 1\}$ where χ is instance space; Five stages of ISF, PSF$\Longrightarrow$WSF$\Longrightarrow$PSF$\Longrightarrow$SSF$\Longrightarrow$PSF.
Initialize: Each detector in each stage of ISF is NULL.
For each stage ψ in ISF (ψ is WSF, SSF or PSF)
* The structure set for ψ is denoted as ζ (ζ is **W**, **S** or **P**)
* **For** each structure ϕ in ζ
  – Select $R^{'}$ ($R^{'} \subseteq R$). Enumerate each sample $\mathbf{x} \in R$. The passed *word* set Ω of $\mathbf{x}$ is inferred by all the previous stages. If missing *words* are within tolerance, add $\mathbf{x}$ into $R^{'}$.
  – Learn detector ρ_ϕ on sample set $R^{'}$ by algorithm in [1] and add ρ_ϕ to ψ.
Output: The learned ISF.

Fig. 7. Positives for pedestrian detection and articulated human detection. Images in (a) and (b) are from INRIA dataset and our collected dataset respectively. Both (a) and (b) are for pedestrian detection. (c) shows positives of articulated human.

The final decision is that the test window (b) is positive and its structure is the structure associated to P2 which is shown in (j).

5 Experiment

Experiments are done for partially occluded pedestrian detection and articulated human detection in cluttered backgrounds. We compare our ISF, which contains five stages, PSF$\Longrightarrow$WSF$\Longrightarrow$PSF$\Longrightarrow$SSF$\Longrightarrow$PSF with other state-of-the-art algorithms. In our experiments, the missing tolerance of *words* is set to 0 for PSF$\Longrightarrow$WSF and PSF$\Longrightarrow$SSF for both pedestrian detection and articulated human detection, while for WSF$\Longrightarrow$PSF and SSF$\Longrightarrow$PSF it is set to 1 for pedestrian detection and to 2 for articulated human detection. During the training for cascade classifiers, the detection rate is set to 0.998 and false positive rate is set to 0.33 for each layer of the detector associated to each structure, *word*, or *sentence* or *paragraph*, which guarantees that the ISF achieves high detection rate and low false positive rate. All experiments are conducted on an Intel Core(TM)2 2.33GHz PC with 2G memory.

5.1 Occluded Pedestrian Detection

INRIA [5] dataset is a popular public dataset for pedestrian detection. The database has 2416 64×128 people images for training and 1126 64×128 for testing. They are downscaled to 24×58 in our experiment. Some positives are shown in Fig. 7 (a). We compare ISF with other state-of-the-art algorithms by False Positive Per Window (FPPW). The ROC curve is given in Fig. 8, in which the x-axis is False Positives Per Window (FPPW), that is, FalsePos/(TrueNeg+FalsePos); and the y-axis is the detection rate, that is, TruePos/(FalseNeg+TruePos) or

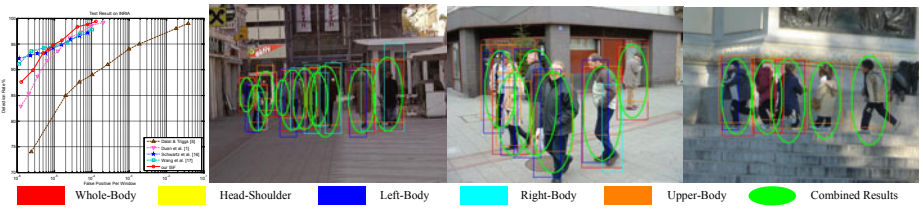

Fig. 8. Evaluation of ISF on INRIA dataset

1-missing rate. The result achieved by ISF improves the whole body detector [1] about 5% at FPPW=10^{-6} which is comparable to the results achieved in [18][19].

ETHZ [20] dataset consists of four video sequences (640×480 pixels at 15 frames/second), one for training and three for testing and only the three testing ones are used in our experiment. We collect 18474 positive samples of 24×58 for learning a robust ISF, which contains 9594 front/rear, 4440 left profile and 4440 right profile samples. Some positives are shown in Fig. 7 (b). We compare ISF with the methods in [20] and [18] by False Positive Per Image (FPPI) which is a better criterion for evaluating detector performance pointed out in [21]. In order to show the efficiency and effectiveness of ISF, we also train one stage PSF which is a Top-Down method mentioned in Section 4.3 on the same positive set.

When the intersection between a detection response and a ground-truth 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. We obtain the ROC curves and some results shown in Fig. 9. Our ISF achieves better results than [20] and [18] on the first two sequences (Seq.#1 and Seq.#2), but the method in [18] achieves better results than ours on the third sequence (Seq.#3). The main reason is perhaps due to significant light changes in Seq.#3 for which our used features (APCFs) are somewhat sensitive. After the comparison of ISF and PSF, we can find that ISF achieves more accurate results with less false positives than PSF in general. The average cost time of ISF on ETHZ dataset is about 1.4s but that of PSF is about 2.6s. So ISF is much faster than PSF.

Furthermore, there are two things should be mentioned: One is that we do not use any additional cues like depth maps, ground-plane estimation, and occlusion reasoning, which are used in [20]. The other one is that there are some problems existed in ETHZ dataset which may affect the evaluation result as shown in Fig. 10. Some shadows of pedestrian are regarded as non-positives which are very hard for any pedestrian detector to identify and some pedestrians no longer exist in a scene are still labeled as positives.

More experiment results on USC SET B [3], Dataset S1 of PETS2009 [22] and our own collected dataset are given in Fig. 12.

5.2 Articulated Human Detection

We have labeled 11482 positive samples of 58×66 for articulated human detection. Typical positives are shown in Fig. 7 (c). Since currently there is no

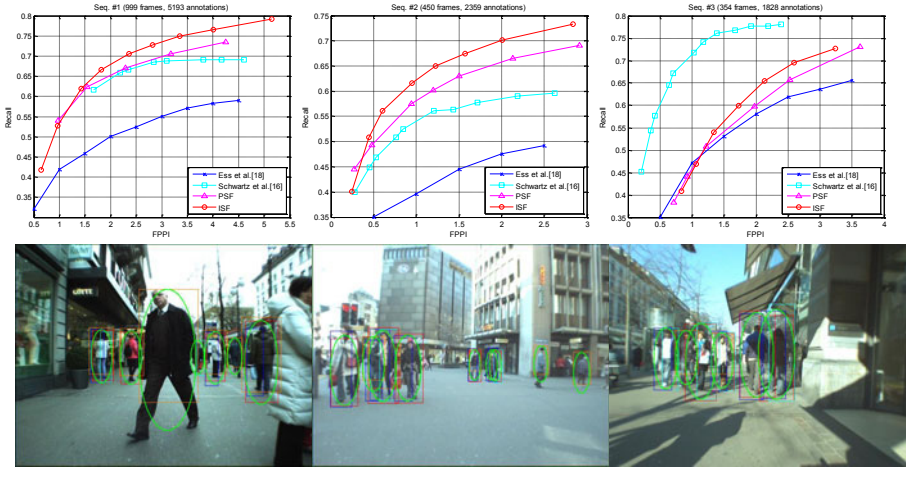

Fig. 9. Evaluation of ISF on ETHZ dataset

Fig. 10. Some groundtruths of ETHZ (in the 1^{st} and 3^{rd} columns) and our results (in the 2^{nd} and 4^{th} columns)

public available dataset for articulated human detection, we have labeled 170 images of 816×612 size with 874 humans for evaluation. Most of them are doing sports (playing football or basketball), so their poses differ a lot and are complex enough.

To compare with ISF, we have also trained PSF. The ROC curve and some results are shown in Fig. 11. This figure shows that ISF achieves more accurate results with less false positives than PSF. The average cost time of ISF is 1.8s and that of PSF is 9.2s. ISF is much faster than PSF.

6 Discussion

Feature sharing. One holistic detector is rather limited to handle occlusions of pedestrians. It is also difficult or impractical to train a usable holistic detector for articulated human due to the diversity. In our experiment, we show that our proposed SF approach is faster and more accurate than the approaches in which part detectors or specific poses detectors are fused simply. The intrinsic reason

Fig. 11. Evaluation of ISF on our own collected dataset

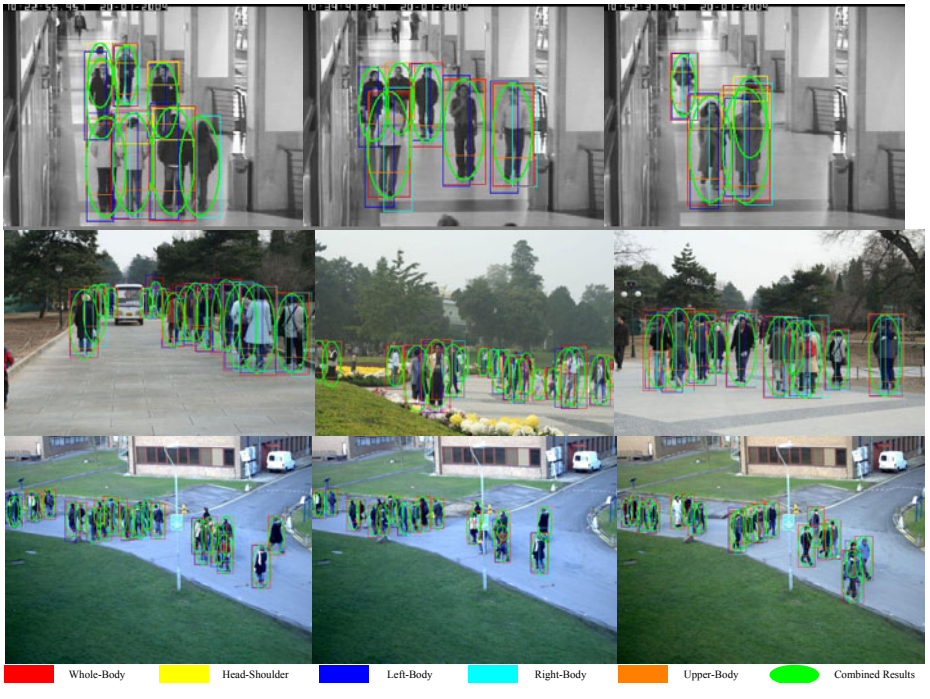

Fig. 12. Results of pedestrian detection on USC SET B(1^{st} line), our own collected dataset (2^{nd} line) and Dataset S1 of PETS 2009 (3^{rd} line)

lies on feature sharing. To explicitly define feature sharing, we first suppose two regions A and B, and region C is the shared area of A and B. Feature sharing means that A and B share the weak features in C. In our designed hierarchical object structures, a significant advantage of *words*, *sentences* and *paragraphs* is that they provide the potential to share weak features. For example, the weak features in head-shoulder can be shared with upper-body and whole-body.

Take the detectors learned for pedestrian detection as an example. There are 13417 weak features in PSF without any feature sharing, while there are 10714 weak features in ISF with feature sharing. Mainly due to the number of features in ISF is less than that in PSF, our ISF is faster than PSF. Feature sharing is of great benefit to our SF indeed. The experiment in the previous section also proves that ISF is more accurate than PSF, which in other words means that our SF approach has explored more discriminative features.

Relation with discriminative models (DM) and generative models (GM). We have proposed hierarchical structures and SF for object detection. In one hand, the detectors are learned by Boosting algorithm. From this point, our model is of DM. Detectors of different parts or poses in a traditional DM are independent but they are related to each other in our model. In another hand, the proposed hierarchical structures formulate one kind of object. From this point, our model is of GM. In fact, we have fused the DM of parts and GM of body structure in our approach.

7 Conclusion

In this paper, we present a SF approach to human detection. The three level SFs are WFS, SSF and PSF which correspond to a hierarchical structure of object, *words*, *sentences* and *paragraphs*. The approach can deal with occlusions and non rigid object detection. In a sense, it is a general framework for object (rigid/non rigid) detection based on *words* (/parts/regions). Experiment results on pedestrian detection in highly crowded scenes and articulated human detection demonstrate its effectiveness and efficiency.

There are some further works to be done to improve our SF approach. Currently *words* and *sentences* of an object are manually designed according to heuristic knowledge. It is hard to generalize this method for more complex objects therefore automatically learning of *words* and *sentences* is expected in the future.

Although the approach is proposed for human detection, we argue that it can be easily extended to other object detection problem, and also to multiple object categorization problems.

Acknowledgements

This work is supported in part by National Basic Research Program of China (2006CB303102), National High-Tech Research and Development Plan of China (2009AA01Z336), and it is also supported by a grant from Omron Corporation.

References

1. Duan, G., Huang, C., Ai, H., Lao, S.: Boosting associated pairing comparison features for pedestrian detection. In: 9th Workshop on Visual Surveillance (2009)
2. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)

3. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: ICCV (2005)
4. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR (2005)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
6. Wu, B., Nevatia, R., Li, Y.: Segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. In: CVPR (2008)
7. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR (2009)
8. Ronfard, R., Schmid, C., Triggs, B.: Learning to parse pictures of people. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 700–714. Springer, Heidelberg (2002)
9. Wang, Y., Mori, G.: Multiple tree models for occlusion and spatial constraints in human pose estimation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 710–724. Springer, Heidelberg (2008)
10. Jiang, H., Martin, D.R.: Global pose estimation using non-tree models. In: CVPR (2008)
11. Lin, Z., Hua, G., Davis, L.: Multiple instance feature for robust part-based object detection. In: CVPR (2009)
12. Dollr, P., Babenko, B., Belongie, S., Perona, P., Tu, Z.: Multiple component learning for object detection. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 211–224. Springer, Heidelberg (2008)
13. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR (2005)
14. Agarwal, S., Roth, D.: Learning a sparse representation for object detection. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 113–127. Springer, Heidelberg (2002)
15. Dai, S., Yang, M., Wu, Y., Katsaggelos, A.: Detector ensemble. In: Computer Vision and Pattern Recognition, CVPR (2007)
16. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning 37, 297–336 (1999)
17. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
18. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: ICCV (2009)
19. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV (2009)
20. Ess, A., Leibe, B., Gool, L.V.: Depth and appearance for mobile scene analysis. In: ICCV (2007)
21. Tran, D., Forsyth, D.: Configuration estimates improve pedestrian finding. In: NIPS (2007)
22. PETS 2009 (2009), http://www.cvg.rdg.ac.uk/PETS2009/