Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 195941, 8 pages
http://dx.doi.org/10.1155/2014/195941

*Research Article*

# Accurate Object Recognition with Assembling Appearance and Motion Information

**Yongxin Chang,[1,2,3] Huapeng Yu,[1,2,3] Zhiyong Xu,[1] Jing Zhang,[2] and Chunming Gao[2]**

[1] *Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China*
[2] *School of Optoelectronic Information, University of Electronic Science and Technology of China, Chengdu 610054, China*
[3] *Graduate University of Chinese Academy of Sciences, Beijing 100049, China*

Correspondence should be addressed to Yongxin Chang; cyongxin@126.com

Received 11 June 2014; Accepted 1 October 2014; Published 21 October 2014

Academic Editor: Guangming Xie



How to effectively detect object and accurately give out its visible parts is a major challenge for object detection. In this paper we propose an explicit occlusion model through integrating appearance and motion information. The model combines together two parts: part-level object detection with single frame and object occlusion estimation with continuous frames. It breaks through the performance bottleneck caused by lack of information and effectively improves object detection rate under severe occlusion. Through reevaluating the semantic parts, the detecting performance of partial object detectors is largely enhanced. The explicit model enables the partial detectors to have the capability of occlusion estimation. By discarding the geometric representation in rigid single-angle perspective and applying effective pattern of objective shape, our proposed approaches greatly improve the performance and robustness of similarity measurement. For validating the performance of proposed methods, we designed a comparative experiment on challenging pedestrian frame sequences database. The experimental results on challenging pedestrian frame sequence demonstrate that, compared to the traditional algorithms, the methods proposed in this paper have significantly improved the detection rate for severe occlusion. Furthermore, it also can achieve better localization of semantic parts and estimation of occluding.

## 1. Introduction

In recent years, great progresses have been achieved for targets detection under complex scenes in the propelling of the PASCAL VOC challenge. However, the detections for arbitrary perspectives are far from satisfaction, and several existing bottleneck issues such as severe occlusions and clutter are still required to further investigate [1–3]. Among these issues, how to detect the targets effectively in the conditions of severe occlusions and recognize the targets accurately is the major challenge.

Occlusions deduction by utilizing a set of predefined blocking modes involving with the algorithm of current detection is proposed in [4–9]. Contour information is considered to promote recognizing performance in [4]. A neural network is computational models inspired by an animal's central nervous systems, in particular the brain. It is composed of a large number of highly interconnected processing elements (neurons) working to solve specific problems. The neural network generally consists of three layers in which an input layer is connected to a hidden layer, which is connected to an output layer. At present, the technology has been widely applied in machine vision and pattern recognition fields [5]. Recently, 3D information of the targets is also considered for the purpose of effective blocking deductions in [6, 7]. The fundamental issue of above several methods is that the restricted predefined model is difficultly appropriate for the complexities of the practical blockings. In addition, these approaches achieved a better capability of detection in several standard databases. However, by means of heuristic and optimized solution, the approaches had difficulty in ensuring to obtain the global optimal solution.

Alternatively, inexplicit occlusion model approaches are performed. For instance, Felzenszwalb et al. [8] applied deformable part models to detect just the visible parts of the objects. In this case, potential issue of this method is that

the detection fails to perform when the targets are severely blocked. In addition, the semantic components are not given by this model. According to the algorithm [8, 9], entire objects are recognized by a set of partial target detectors with the output data clustering. Obviously, this method does not carry out the statistical evaluation for occlusion, which is always postulated to be given the full targets. Compared with the previous methods, the semantic components are their advantages.

Recently, information complementary such as 3 dimensional information is an important research trend. Depth and motion information are introduced for occluding estimation [10]. Obtained visibility is used as weight of the each part of classifier. Obviously, it is a more natural and efficient solution to take account of utilizing depth and motion information to resolve the problems of blocking and arbitrary pose. Therefore, the classification and recognition are both involved in this work.

The explicitly block model is used and the appearance and moving information of object are integrated in this paper, which can show the occlusion evaluation and detect object in part-level. By using partial detectors, we can obtain semantic parts which play a significant role in the further processing like the evaluation of man's posture.

To obtain the moving information of object, the structure and motion adaptive regularization of optical flow method is utilized. Compared with other optical flow approaches [11, 12], the method has a better performance for motional estimation.

For proving the performance of our methods, we implement a comparative experiment on challenging pedestrian frame sequences database [10]. The experiment shows that, compared to traditional methods, the approaches proposed in this project can obviously improve the target detection rate in the blocking occasion, better locate semantic parts, and estimate occlusion.

The rest of the paper is structured as follows. The next section principally introduces the methods of part-level object detection with shared appearance. In Section 3 we mainly evaluate and recognize object by motion information. Section 4 shows the experimental results and analysis. Finally, we conclude this paper in Section 5.

## 2. Part-Level Object Detection with Shared Appearance

Generally, an object is composed of different parts by constrained geometric configuration. For example, a human body includes head, torso, arms, and legs. In [8, 9] object detecting model based on parts achieved state-of-the-art performance and surpassed other similar models on people detection.

In this work, our objective detection model is also based on parts which are semantic.

*2.1. Training Object Detector Based on Parts.* The $K$ partial target detectors are defined as $G_k$ $(k = 1 \cdots K)$. In order to train $G_k$, 300 groups (3 as a group) of similar posture images from different parts of human body are chosen as the positive samples. Images of equal quantity without objects are chosen randomly as negative samples. Figure 1 shows a few positive samples. They are 3 positive samples in the same team with similar postures. Linear SVM is chosen as the classifier and HOG as the appearance feature [13]. We train the classifier by bootstrapping method. First, an initial classifier is obtained by training positive sample sets and randomly negative sample sets. Then, we detect the images which do not include the object with this initial classifier for gaining false alarm sets. Finally for the sake of optimizing the detector performance, the false alarm sets are added to negative samples as difficult false positive sets for a second training.

Figure 1: Some examples of positive sample set. How to recognize these objects of similar posture.

It is significant for us to find that the different parts of the target have various discrimination with each other. For example, wheel is the salience part for automobile. Bourdev and Malik [9] prove that the partial detectors for upper limb of human body are better discriminative than others. The discrimination abilities of body parts, from better to worse, are, respectively, the frontal face, torso, and legs. Based on this result, partial target detectors are extended to promote the precision. The extension approaches are described in detail in Section 2.2.

*2.2. Reevaluating Objective Occlusion Based on Semantic Parts.* For the tested input image, we applied the part target detector $G_k$ $(k = 1 \cdots K)$ trained in Section 2.1 to all locations in multiple scales. Then, we cast votes and use mean-shift algorithm to cluster for the object location $x$. Finally, we can obtain $P(O \mid x)$ that denotes the probability position $x$ of the detected object $O$. In fact, only to vote and cluster for overall object is still far from the potential performance of the part target detector. On one hand, because each of the semantic parts has different discriminative capacity (see the second

Figure 2: Some examples for effective performance improvement with reevaluating. Original images are from PASCAL VOC 2010.

paragraph in Section 2.1), it would be useful in the rescoring of the semantic components. On the other hand, we can gain much more objective information by the means of voting and clustering based on the semantic pars.

In this work, we define human body which consists of three semantic parts such as head, torso, and legs. $O_h$, $O_t$, and $O_l$ denote, respectively, head, torso, and legs. $P(O_h \mid x)$, $P(O_t \mid x)$, and $P(O_l \mid x)$, which are gained by voting and clustering, are defined as the probability position $x$ of corresponding object $O_i$, $i \in \{h, t, l\}$. In formula (1), $a_k(x)$ is the score which a part target detectors is evaluated in position $x$, and $w_k$ is the weight of the partial target detector $G_k$ $(k \in \{1 \cdots K\})$ [9]. Consider

$$P(O_i x) \infty \sum_k w_k a_k(x). \tag{1}$$

According to formula (2), $P(O \mid x)$ can be obtained by calculating $P(O_i \mid x)$, $i \in \{h, t, l\}$, and $\beta_i$; $\beta_i$ is the weight of the semantic parts $i$ and represents the discrimination of the semantic parts. In this paper, we choose $\beta_i$ as a fixed set of values $\{0.68, 0.22, 0.1\}$ and satisfies $\sum \beta_i = 1$. Consider

$$P(Ox) \infty \sum_i \beta_i P(O_i x). \tag{2}$$

In practice, we need to deal with the problem of some semantic components absent due to the measured object not in the image or occluded. We postulate that high weight $O_h$ always exists (if the head is not present, we consider it undetectable) and sometimes low weight $O_t$ and $O_l$ are inexistent. In this situation, the weight of the absent semantic component is transferred to the higher weight of the most adjacent semantic part. The reason for this is that the part detectors can recognize the object in the parts which are existent and can still detect the target in case of missing parts.

Figure 2 shows some examples for effective performance improvement because of reevaluating. In Figure 2, the green bounding box denotes the correct detection, and blue bounding box represents the false alarm filtered out validly. Note that the parts have better discrimination by means of reevaluating, and therefore they can filter out false alarms effectively.

## 3. Estimating and Locating Object by Motion Information

We extend and improve the performance of partial object detector through utilizing the motion information with continuous frames. According to the experiments results of PASCAL VOC challenge, they define AP (average precision) to estimate the recognizing performance. As far as we know the current best result AP approximated to 0.4 [3, 8]. Only making use of 2D HOG appearance feature is a principal cause of bottleneck. Hence motion information among frames is introduced to estimate object occlusion and enhance the detecting capability in this paper. We believe that combining appearance feature with motion information is very important for achieving more excellent and effective detection performance in occlusion and various pose conditions.

*3.1. The Segmentation of Optical Image.* As described above, we have adopted the structure and motion adaptive regularization of optical flow method to gain optical flow image and then mean-shift algorithm is selected to segment gained images in this section. The segmentation results are defined as $\varphi_c$, $c = 1 \cdots k$, and $k$ is the number of clusters. In our experiments, a set of images with severe occlusion

Figure 3: Some examples for optic flow images and corresponding segmentation results.

and arbitrary aspects are tested. The experimental results are shown in Figure 3. The original image sequences are from [10]. And then optical flow images and corresponding segmentation results are given by the approaches [14–17]. Detailedly speaking, the first column is the last frame of the original image frame sequence, wherein the human is in motion. The second column is the corresponding results of optical flow images. The third column is the segmentation results. The experimental results demonstrate that the partial occluded pedestrian obviously displayed discontinuity in occlusion boundary. Therefore motional objects and barrier are divided into different clusters through segmenting optical flow images. And the results present that the visual parts of human basically belong to a similar cluster class which is consistent with the conclusion from [10], and optical flow method is fully efficient for motional object estimation.

*3.2. Analysis and Estimation for Occlusion.* Given the segmentation results $\phi_c$ of optical flow image, we can combine the detected results based on appearance information with the results of part detector to estimate the target occlusion. Using a similar measurement based on correlation and combining object shape information, we obtain the visible area cluster of the corresponding target. The main difference is that we discard geometric representation of rigid target and utilize a more effective target shape mode to effectively promote the performance and applicability of the method. There is no doubt that all of them are based on the above described part target detector.

Firstly, the geometric representation of rigid and single perspective target is not suitable for flexible target (such as human body) and multiviews target. Actually, the detected results $O_k$, $k \in \{h, t, l\}$, gained in Section 2.2, can play a better role.Given the bounding box of part $O_k$, we can filter out a large number of clusters, while the geometric representation of rigid target does not result in the loss of information. The consequence of $\phi_c$ filtering out by $O_k$ is denoted by $\varphi_c^k$, $c = 1 \cdots k$.

Secondly, it is too complex and not desirable that a limited number of perspectives (corresponding to the front body/back, left) are applied to represent the shape pattern of target [10]. Therefore, we have adopted a multiaspects expression of target shape model. In fact, the average shape $m_v$ of the target can be gained through the use of part target detector, the multiviews shape from the average expression in the target portion of the detector (the viewing angle) and cluster vote; different perspective of some of the target detector means sufficiently fine multiangle expression.

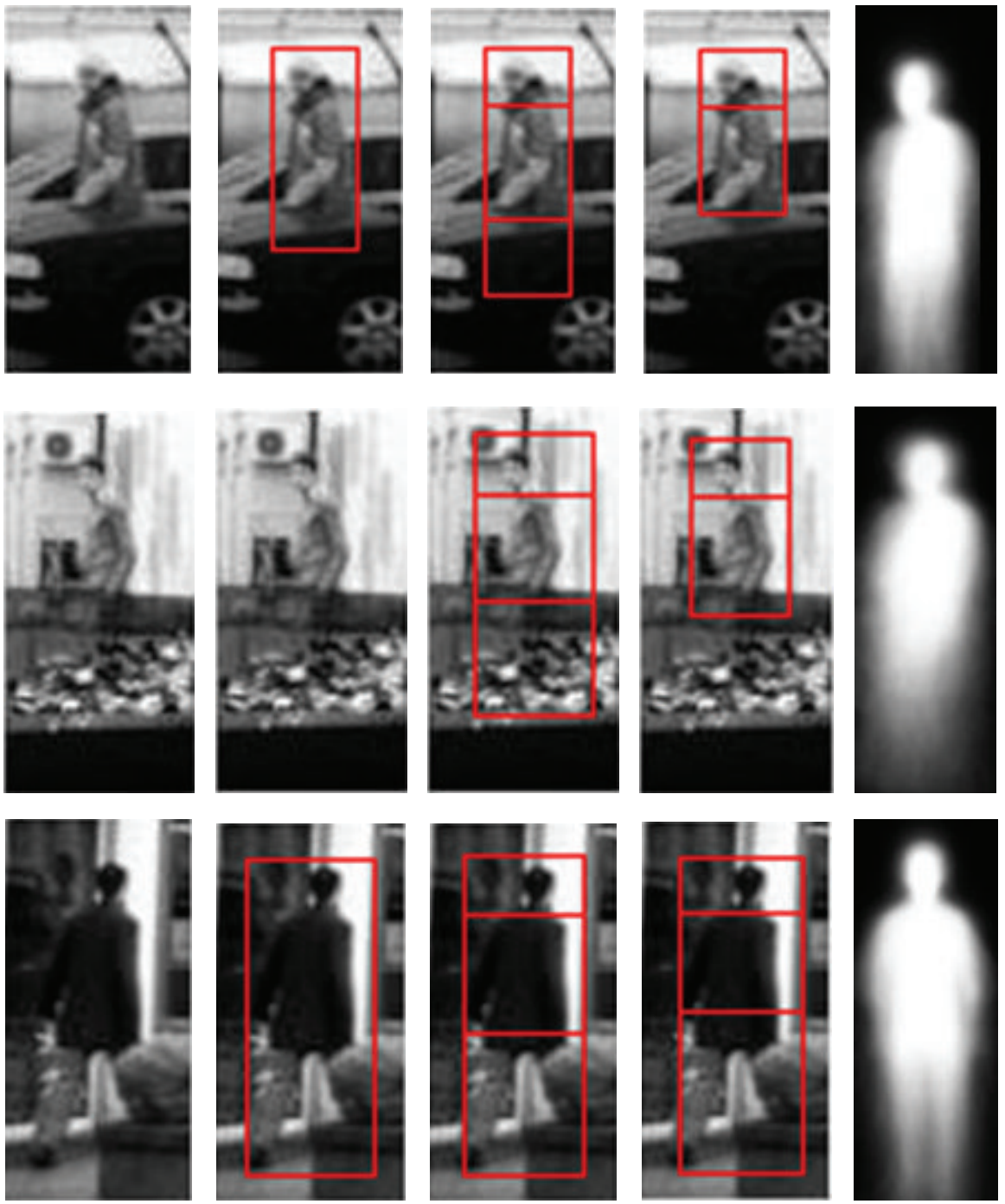


Figure 4: The recognizing results of different algorithms on pedestrian database.

The expression capacity of average shape $m_v$ is derived from the voting and clustering of the part detector depending on aspects. The expression ability of multiangle model is proportional to the number of part target detectors of different perspectives. Therefore, in this paper we are no longer limited to enumerate a few perspectives, while the $m_v$ is the most effective poses expression for object in the current specific situation. The $m_v$ can be obtained through summing each part shape recognized by the target detector. Expression (3) describes it very well, where $M_i(x, y)$ represents the binary shape information given by the part target detector $G_k$. The last column of Figure 4 shows the results of average shape $m_v$. Note that $m_v$ can be further divided into $m_v^k$, $k \in \{h, t, l\}$, in accordance with the components. Consider

$$m_v = \sum_i M_i(x, y). \tag{3}$$

From the above discussion, we use the similarity measure method based on the correlation [10] to estimate occlusion. $\Psi(\overrightarrow{\phi_c^k}, \overrightarrow{m_v^k})$ denotes a correlation on the base of similarity measure in formula (4), where the vectorized representations of $\phi_c^k$ and $m_v^k$ are, respectively, denoted by $\overrightarrow{\phi_c^k}$ and $\overrightarrow{m_v^k}$. The equation $\overrightarrow{v_v^k} = 1 - \overrightarrow{m_v^k}$ can be considered as the inverse shape of $\overrightarrow{m_v^k}$. In formula (4), the dot product of $\overrightarrow{m_v^k} \cdot \overrightarrow{\phi_c^k}$ denotes a similar probability of average shape and cluster, while the dot product of $\overrightarrow{v_v^k} \cdot \overrightarrow{\phi_c^k}$ indicates a similar correlation of cluster and inverse average shape. Consider

$$\Psi\left(\overrightarrow{\phi_c^k}, \overrightarrow{m_v^k}\right) = \overrightarrow{m_v^k} \cdot \overrightarrow{\phi_c^k} + \left(1 - \overrightarrow{v_v^k} \cdot \overrightarrow{\phi_c^k}\right). \tag{4}$$

According to formula (4), $\overrightarrow{\phi_c^k}$ corresponding to the parts $k$ and the average shape $\overrightarrow{m_v^k}$ are calculated to gain $\Psi(\overrightarrow{\varphi_c^k}, \overrightarrow{m_v^k})$. And then the maximum probable estimation $\overrightarrow{\varphi_{\text{vis}}}$ for visible region can be obtained by the following formula:

$$\overrightarrow{\phi_{\text{vis}}} = \underset{\overrightarrow{\phi_c^k}}{\text{argmax}}\left(\max_{\overrightarrow{m_v^k}}\left(\Psi\left(\overrightarrow{\phi_c^k}, \overrightarrow{m_v^k}\right)\right)\right). \tag{5}$$

Note that the $\overrightarrow{m_v^k}$ reflects the most likely perspective representation of object, which is fundamentally different from the expression of [10]. Compared with [10], our approaches can reduce the complexity at least three times while bringing in more accurate occlusion evaluation owing to the partial detectors providing more sophisticated multiple aspects expression.

## 4. The Experimental Result and Analysis

In this section, for validating the performance of our methods for occlusion and different views, we designed a comparative experiment on challenging pedestrian frame sequences database [10]. The experiments are done in the same environment. These pedestrian frame sequences are typical of complex scenes and pedestrians may be partial severe occlusion. The used detector has been described in detail in Section 2. According to [11], we have adopted the structure and motion adaptive regularization of optical flow method with its default settings to measure optical flow images of frame sequence. Optical flow image segmentation is done by the mean-shift algorithm from [17] with its default settings.

Figure 4 shows the recognizing results of different algorithms on pedestrian database. As shown in Figure 4, the first column is the last frame of the original image frame sequence, wherein the human is in motion. The second column is the detection result of [8] which fails the severe occlusion detection. The third column is the detection result of [9] which cannot handle the parts occluded. The fourth column is the detecting results of this paper presenting approach which has a capacity to effectively solve the blocked problem and accurately recognize the visible portion of object. The last column is the average shape of this detection, which is the corresponding object from different perspectives. Note that the detection results of the third column and the fourth column give the given semantic components, from top to bottom; each small bounding box of the test results is given head, torso, and legs. As shown in Figure 4, [8] will fail under severe occlusion, because [8] only detects the visible part of the target. If the visible part is too small, then it will fail to detect the object. Although [9] still can detect the target under severe occlusion, due to lack of explicit occlusion model, it is always given the assumption that the entire goal is existent. Our approach can not only detect the target under the severe occlusion condition, but also give accurate semantic parts of the visible targets.

Table 1: Statistical result on 1000 frame sequences.

| | Recognizing rate |
|---|---|
| Reference [8] | 90.8% |
| Reference [9] | 88.4% |
| Our methods | 94.2% |

We select 1000 frame sequences to gain statistical experimental results. We first complete the calibration of the visible part of the body because the original frame sequence is not calibrated. The evaluation of test results was done by precision recall curve [18, 19], and the definitions of precision and recall were written as formula (6), wherein FN is the number of false negatives, TN represents the quantity of true negatives, and FP and TP, respectively, denote false positives and true positives. Ideally, we want the precision and recall to be as high as possible. But in fact, the two parameters are contradictory in practical engineering application. When the precision is high, the recall rate is low. Consider

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}. \end{aligned} \tag{6}$$

Figure 5 shows the 1-precision/recall curve. Table 1 further shows the detection rate in the case of equal FP (false positives).

Figure 5 indicates that the method proposed in this paper has greater advantages compared with [8, 9]. The recall rate of the method proposed in [8] is slightly higher only when the demand on the accuracy is quite low. It is important to note that the performance of the method in [9] is not as good as the one in [8], which is mainly caused by the assumption of a complete object, whether the object is occluded or not. By overcoming this problem, this proposed method shows better performance than [8, 9].

## 5. Conclusion

In this paper, the explicit model for estimating the object occlusion is built up based on the appearance information of the single frame and motion information of continuous

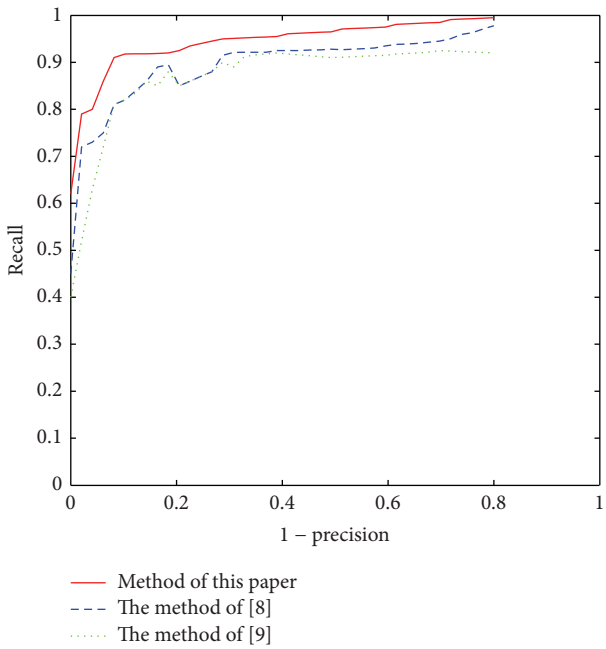


Figure 5: Statistical results on 1000 frame sequences. The precision recall curves of different methods.

frames. At first, the object detection of multiple views is obtained by combining the appearance information and the partial detectors, which provide semantic parts and average shape of the object. Then, the motion information is gained by the segmentation of optical flow image. Finally, we can estimate the objective visible region by calculating the correlation of average shape and segmentation results.

The detecting performance is largely enhanced by reevaluating the semantic parts. The explicit model enables the partial detectors to have the ability of occlusion estimation. By abandoning the geometric representation in rigid single-angle perspective and applying effective pattern of objective shape, our proposed approaches greatly improve the performance and robustness of similarity measurement. The experimental results on challenging pedestrian frame sequence prove that, compared to the traditional algorithms, the methods proposed in this paper have greatly improved the detection rate for severe occlusion. Furthermore, it also provides better semantic part localization and occlusion estimation.

## Conflict of Interests

The authors declare no conflict of interests regarding the publication of this paper.

## Acknowledgments

This research is supported by National Science Foundation of China (NSF grant number 61205004) and by the Graduate Innovation Fund of Chinese Academy of Sciences (grant number A08K001).

## References

[1] S. Savarese and L. Fei-Fei, "3D generic object categorization, localization and pose estimation," in *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV '07)*, pp. 1–8, Rio de Janeiro, Brazil, October 2007.

[2] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, no. 1, pp. 1–19, 2004.

[3] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

[4] M. Maire, P. Arbeláez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[5] M. Everingham, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge," http://www.pascal-network.org/challenges/VOC.

[6] M. Zia, M. Stark, and K. Schindler, "Explicit occlusion modeling for 3D object class representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3326–3333, Portland, Ore, USA, June 2013.

[7] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[9] L. Bourdev and J. Malik, "Poselets: body part detectors training using 3D human pose annotations," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '09)*, 2009.

[10] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 990–997, June 2010.

[11] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[12] H. Inoue, T. Tachikawa, and M. Inaba, "Robot vision system with a correlation chip for real-time tracking, optical flow and depth map generation," in *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1621–1626, Nice, France, May 1992.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 886–893, San Diego, Calif, USA, June 2005.

[14] A. Wedel, D. Cremers, T. Pock, and H. Bischof, "Structure- and motion-adaptive regularization for high accuracy optic flow," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 1663–1668, October 2009.

[15] R. Fransens, C. Strecha, and L. van Gool, "A mean field EM-algorithm for coherent occlusion handling in MAP-estimation problems," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 1, pp. 300–307, June 2006.

[16] X. Wang, T. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proceedings of the IEEE 12th International Conference on Computer Vision*, pp. 32–39, Kyoto, Japan, September-October 2009.

[17] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[18] F. J. Estrada and A. D. Jepson, "Benchmarking image segmentation algorithms," *International Journal of Computer Vision*, vol. 85, no. 2, pp. 167–181, 2009.

[19] T. Brox, L. Bourdev, S. Maji, and J. Malik, "Object segmentation by alignment of poselet activations to image contours," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 2225–2232, June 2011.