Using Non-symmetry and Anti-packing Representation Model For Object detection

Chuanbo Chen, Guangwei Wang, Xiaochen Wang

School of Computer Science and Technology
Huazhong University of Science and Technology
Wuhan, China
wgw1949@gmail.com

Abstract

In this paper, we present a non-symmetry and anti-packing object pattern representation model (NAM) for object detection. A set of distinctive sub-patterns (object parts) is constructed from a set of sample images of the object class; object pattern are then represented using sub-patterns, together with spatial relations observed among the sub-patterns. Many feature descriptors can be used to describe these sub-patterns.The NAM model codes the global geometry of object category, and the local feature descriptor of sub-patterns deal with the local variation of object. By using Edge Direction Histogram (EDH) features to describe local sub-pattern contour shape within an image, we found that richer shape information is helpful in improving recognition performance. Based on this representation, several learning classifiers are used to detect instances of the object class in a new image. The experimental results on a variety of categories demonstrate that our approach provides successful detection of the object within the image.

Keywords: object detection, NAM, edge direction histogram, pattern representation, SVM classifier

1. Introduction

In this paper we consider the problem of detecting and localizing object of a generic category, such as horse or car in static images. This is a difficult problem because objects in a category can vary greatly in shape and appearance. Variation arise not only from changes in illumination, occlusion, background clutter and view point, but also due to non-rigid deformations, and intra-class variation in shape and other visual properties among objects in a rich category.

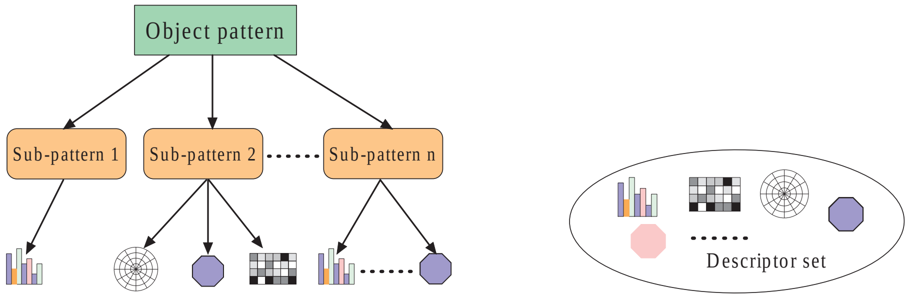
How do we deal with the variation, especial the intra-class and pose variability of object? Most of the current researches have focused on modeling object variability, including several kinds of deformable template models [1][2], and a variety of part-based, fragment-based models [3, 4, 5, 6, 7, 8, 9].

The method of Leibe et al. [10] give a highly flexible learned representation for object shape that can combine the information observed on different training examples. Opelt, et al. [9] explore a similar geometric representation to that of Leibe et al. [10] but use only the boundaries of the object, both internal and external (silhouette). The pictorial structure models [5][11] represent an object by a collection of parts arranged in a deformable configuration, where the deformable configuration is represented by spring-like connections between pairs of parts. Crandall et al. propose k-fans model [12] to study the extent to which additional spatial constraints among parts are actually helpful in detection and localization. The patchwork of parts model from [6] is similar, but it explicitly considers how the appearance model of overlapping parts interacts to define a dense appearance model for images. It is proved that adding spatial constraints gives better performance.

Our approach has two methods to deal with the variation of object, both global and local. Firstly, we propose a non-symmetry and anti-packing object pattern representation model (NAM) to represent an object category. The NAM object model consist of several local parts, we call it sub-patterns. The model codes the global geometry of generic visual object categories with spatial relations linking object pattern to sub-patterns.

Secondly, the descriptors of sub-pattern can deal with the local variation of object. Shape based information have been selected as a key component of local features. We introduce the edge direction histogram (EDH) to describe the contour shape of sub-pattern. Contour shape have been used in object recognition to a certain extent: Shatton et al. [21] and Opelt et al. [9] use boundary fragments to represent a object and use boundary matching method to detect object.

The proposed framework can be applied to any object that consists of distinguishable parts arranged in a relatively fixed spatial configuration. Our experiments are performed on images of side views of horses; therefore, this object class will be used as a running example throughout the paper to illustrate the ideas and techniques involved.



(a) (b)

Figure 1: (a)Our hierarchical object model, (b)Descriptor set.

The rest of the paper is organized as follow. Section 2 describes the non-symmetry and anti-packing model. Section 3 introduces the sub-pattern descriptor. Section 4 present the framework of our approach. In section 5, experiments on real images show that the model is effective for object categorization.

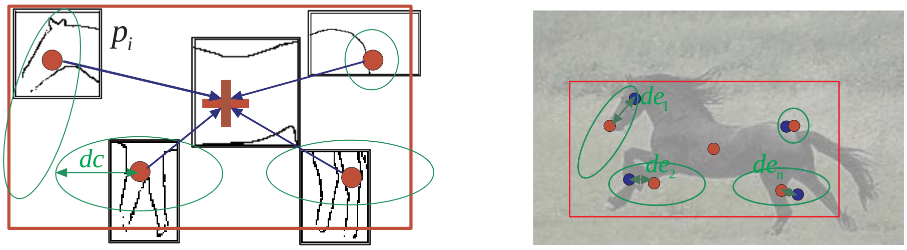
2. Description of non-symmetry and anti-packing pattern representation model

The non-symmetry and anti-packing object pattern representation model (NAM) is an anti-packing problem. The idea of the NAM can be described as follows: Given a packed pattern and n predefined sub-patterns $\{p_1, p_2, ..., p_n\}$, pick up these sub-patterns from the packed pattern and then represent the packed pattern with the combination of these sub-patterns.

The object pattern representation method is a hierarchical model that codes the global geometry and local appearance of generic visual object categories with spatial relations linking object pattern (top level) to sub-patterns (second level), and local feature cues linking sub-pattern (second level) and local feature class (third level). See Fig.1 (a), the object pattern is at the top level, the second and third level are the sub-patterns and local feature descriptors of sub-pattern respectively.

Global Spatial relation: The spatial relations between top level and the second level can be described by global spatial structure. Fig.2 presents an example of the global spatial structure for horse class.

Local feature encoding: Sub-patterns can be described by a rich set of cues (such as shape, color and texture) inside them. Between second level



(a) (b)

Figure 2: (a)Global spatial structure. Sub-patterns p_i (black box) are arranged within the bounding box (red). Small red circles show the optimal position, and blue ellipses the spatial uncertainty dc. (b)To code the pose variation of an side view horse using global spatial structure. Green arrows show the deviation distance de.

and third level, we capture different sub-pattern cues from the sub-pattern window, each type of cue is encoded by using an appropriate descriptor, and these encoded information concatenated into a feature vector. In this paper, we use shape information to represent the sub-patterns and edge direction histogram (EDH) to describe the shape information. The detail of encoding sub-pattern feature have presented in Section 3.

Here, we use an object pattern Γ to describe an object category. The object pattern that consists of n sub-patterns p_i can be defined by the following expression:

$$\Gamma = \bigcup_{i=1}^{n} p_i(x, r, de, w, \phi(x, r) | \phi = \{f_1, f_2, \cdots, f_{m_i}\})$$

Where x is a two-dimensional vector specifying an "anchor" position for sub-pattern p_i relative to the object pattern position; r represent the scale of the sub-pattern; de is a deviation vector; w shows the discriminative weight of the sub-pattern. $\phi(x, r)$ denote a feature vector for the i^{th} sub-pattern and $f_j (1 \leq j \leq m_i)$ is one of the feature descriptors.

The non-symmetry relationship between sub-patterns describes the global structure information of object category and is designed to decouple variations due to affine warps, pose variability and other forms of shape deformations. Anti-packing is a procedure that finding sup-patterns in query images, combining them into a object pattern and classifying.

3. Sub-pattern description

Sub-pattern can be described by a rich set of cues inside them, such as shape, color and texture. Based on the observation that for a wide variety of common object categories, shape matters more than local appearance. In this paper we use shape information as a key component for object detection.

Since edge points are related to shape information closely, edge direction histogram (EDH) is a very simple and direct way to characterize shape information of an object. It has been applied successfully to image retrieval [13, 14, 15], and classification [16]. In addition, Kim used EDH to watermarking text document images [18] based on the idea that sub-images have similar-shaped EDHs. EDH is usually normalized to be scaling invariant, but Zhang et al. [19]compute the 1-D FFT of the normalized EDH to obtain rotation invariance and take it as the final signature of image.

EDH is computed by grouping the edge pixels which fall into edge directions and counting the number of pixels in each direction. Edge map are extracted by edge detection operator (we use Canny edge detector) and each of edge points can be represented with the vector $\overrightarrow{D}_{i,j} = \{dx_{i,j}, dy_{i,j}\}$ where $dx_{i,j}$ and dy are, respectively, horizontal and vertical differences of the point. Each point's edge direction (i.e., gradient direction) is calculated with the equation $\theta_{i,j} = \arctan(\frac{dy_{i,j}}{dx_{i,j}})$. We then divide direction into bins (e.g. 20° per bin) and calculate the orientation histogram over some region.

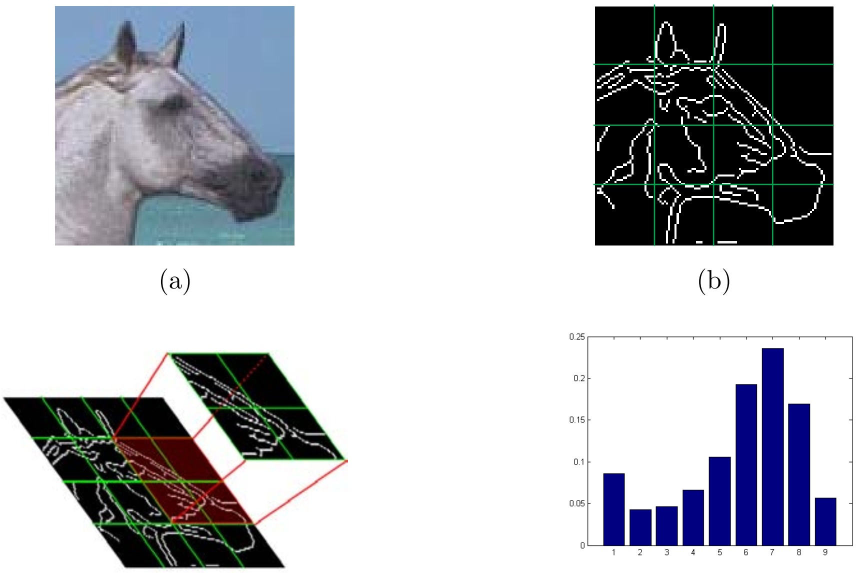
A global direction histogram of a sub-pattern would average too much spatial information to infer pose. We describe a sub-pattern window by dividing evenly its bounding box into $n \times n$ grid, and accumulating a local 1-D histogram of edge direction over the edge pixels within the 2×2 grid, as illustrated in Fig.3. In the experiments reported, we use $n = 4$. The combination of these histograms then represents the descriptor.

4. Detection

The pipeline of our detection framework as follow: first, training a classifier for each sub-pattern. Next, using one classifier to detect hypothesis of object location, i.e., initial detection. After that, a verification scheme is applied to the hypothesis to obtain final detection.

4.1. Learning classifiers

The task of learning is to establish n classifier $\{Cf_1(\cdot), Cf_2(\cdot), \cdots, Cf_n(\cdot)\}$ for an object pattern with n sub-patterns, each classifier is corresponding to



(a) (b) (c) (d)

Figure 3: The contour shape descriptor. (a)Original image, (b)Edge map, (c)4x4 grid on edge map, (d)Edge direction histogram of one block containing 4 grids.

a sub-pattern. Take a classifier for example, given a set of training image windows labeled as positive (object) or negative (nonobjective), each image window is converted into a feature vector as described above. These vectors are then fed as input to a supervised learning algorithm that learns to classify an image window as member or nonmember of the object pattern. In our experiments, the learning algorithm is a two-class SVM.

4.2. Detection hypothesis using one of the learned classifiers

The initial detection problem is to determine whether the query image contains instances of sub-pattern and where it is. Having trained a SVM window classifier, we can detect and localize novel object instances in a test image using a simple sliding-window mechanism [24][25]. Here, we select the j^{th} sub-pattern p_j as an initial detected sup-pattern. The classifier $Cf_j(\cdot)$ corresponding to sub-pattern p_j is applied to fix-sized windows at various locations in the feature pyramid, each window being represented as a feature vector $\phi(x, r)$, where x specifies the position of the window in the image, and r specifies the level of the image in pyramid. The following expression represents the classifier $Cf_j(\cdot)$ at one of the sliding windows.

$$s_{p_j} = Cf_j(\phi(x,r)) \tag{1}$$

A threshold α is introduced to determine whether the window is positive or contain a instance. If $s_{p_j} > \alpha$, then,the window is positive, $h_j = (x,r)$ is a hypothesis and we put the h_j into the sub-pattern hypothesis set $H = \{h_{j,1}, h_{j,2}, \cdots, h_{j,k}\}$. Lowering the threshold increase the correct detections but also increases the false positives; raising the threshold has the opposite effect. In our experiment, we use $\alpha = 0.5$.

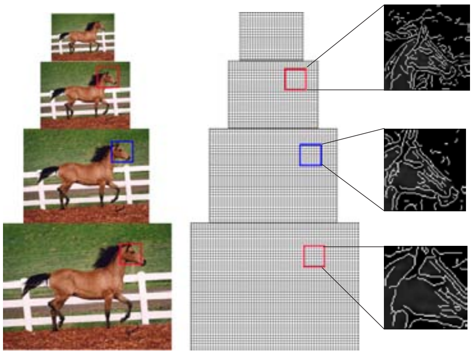


Figure 4: Image pyramid (Left), feature pyramid (Middle) and example of feature windows (Right).

The feature pyramid illustrated in Figure 4, which is similar to [17] , specifies a feature map for a finite number of scales in a fixed range. In practice we compute feature pyramid by computing a standard image pyramid via repeated soothing and subsampling, and then computing a feature map from each level of the image pyramid. A test image is scaled to sizes ranging from 0.48 to 1.2 times the original size, each scale differing from the next by a factor of 1.2.

4.3. Verification

These hypothesis are then refined through a verification scheme to obtain final detection result. The first step is to generate a hypothesis h_Γ of the object pattern Γ by applying a transformation T$(\cdot)$ to h_j and Γ . Fig. 5 (b) illustrate the transformation procedure. h_j is one of hypothesis in set H . The transformation T$(\cdot)$ exploits the rough localization provided by the spatial relationship between the sub-patterns and the object pattern. Then the transformation for the hypothesis h_j and object pattern Γ is characterized by:

$$h_\Gamma = \mathrm{T}(h_j, \Gamma) \tag{2}$$
$$h_\Gamma = (L_1, L_2, \cdots, L_n)$$

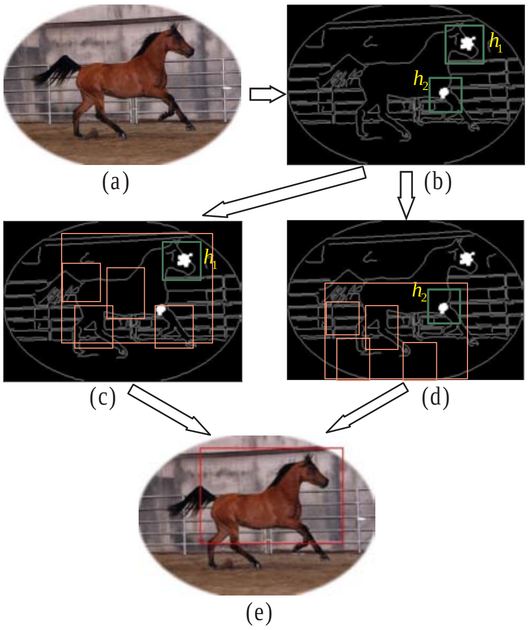


Figure 5: Initial detection and verification. (a)Query image, (b)Hypothesis set of initial detection $\{h_1, h_2\}$, (c)Process of verification for h_1,(d)Process of verification for h_2 and (e)Final result.

Where $L_i = (x_i, r_i, d_i)(1 \leq i \leq n)$ is the expected location of sub-patterns beside h_j. This transformation provides not only position x_i , scale estimation r_i but also deviation d_i of sub-patterns in the object pattern.

Next, classifier $Cf_i(\cdot)$ is applied to the corresponding window at location L_i.

$$s_{p_i} = Cf_i(\phi(L_i)) \tag{3}$$

Where s_{p_i} is to determine that whether location L_i contains sub-pattern p_i . The overall verification score $S_{ver}(h_\Gamma)$ for object pattern Γ is a combination of the sub-patterns detection result s_{p_i}:

$$S_{ver}(h_\Gamma) = \sum_{i=1}^{n} (w_i \cdot s_{p_i} - de_i) \tag{4}$$

Where w_i is the discriminative weight of sub-pattern p_i , de_i is the deviation of sub-pattern from the optimal position. The verification of object pattern was illustrated in Fig. 5 (c) and (d). When the value of the S_{ver} is above a threshold β, the hypotheses position h_Γ contains an instance of the object pattern.

5. Experiments

We present extensive experimental evaluation, involving several existing data sets covering eight diverse shape-based object classes for a total of more than 1400 test images.

5.1. Evaluation Criteria

In this section, we investigate the performance of our system under the PASCAL criterion. For a detection to be marked as correct, its inferred bounding box b_{inf} must agree with the ground truth bounding box b_{gt} based on an overlap criterion as $\frac{area(b_{inf} \bigcap b_{gt})}{area(b_{inf} \bigcup b_{gt})} > 0.5$. Each b_{gt} can match to only one b_{inf} , and so spurious detections of the same object count as false positives.

When a detection system is put into practice, we are interested in knowing how many of the objects it detects, and how often the detections it makes are false. This trade-off is captured more accurately by a variation of the recall-precision curve, where

$$Recall = \frac{TP}{nP} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

where TP is the number of true positives; FP is the number of false positives. nP is the total number of positives in data set.The first quantity of interest, namely, the proportion of objects that are detected, is given by the recall. The second quantity of interest, namely, the number of false detections relative to the total number of detections made by the system, is given by

$$1 - Precision = \frac{FP}{TP + FP} \tag{7}$$

Plotting recall versus (1-precision), therefor, expresses the trade-off.

Performance is also evaluated by plotting detection rate (DR) versus the incidence of false positives (false positives per image (FPPI)) while varying the detection threshold, where

$$Detection\ rate = \frac{TP}{nP} \tag{8}$$

$$False\ positives\ per\ image = \frac{FP}{nN} \tag{9}$$

where nN is the total number of images in data set. Comparison between different methods is mainly based on two points on the DR/FPPI plot, at 0.3 and 0.4 FPPI.

5.2. INRIA horse and Weizmann-Shotton horse

INRIA horse [20]. This challenging data set consists of 170 images containing one or more horses, seen from the side, and 170 images without horses. Horses appear at several scales and against cluttered backgrounds. Weizmann-Shotton horse [21]. Shotton et al. [21] propose another horse detection data set, composed of 327 positive images containing exactly one horse each and 328 negative images. The INRIA and Weizmann are very challenging data sets of horse images, containing different breeds, colors, and textures, with varied articulations, lighting conditions, and scales.

Table 1: Performance of our detection system on INRIA horse data set, containing 170 positive images and 170 negative images.

Threshold β	NO. of correct detections,TP	No. of false detections,FP	Recall,R TP/170	Precision,P TP/(TP+FP)
20	152	110	0.8941	0.5802
30	138	107	0.8118	0.5633
40	130	98	0.7647	0.5702
50	126	86	0.7412	0.5942
60	121	74	0.7118	0.6205
80	110	53	0.6471	0.6748
100	103	45	0.6059	0.6959
120	87	39	0.5118	0.6905
150	67	26	0.3941	0.7204
180	41	14	0.2412	0.7455
210	23	7	0.1353	0.7667

Table 2: Performance of our detection system on Weizmann horse data set, containing 328 images .

Threshold β	NO. of correct detections,TP	No. of false detections,FP	Recall,R TP/170	Precision,P TP/(TP+FP)
20	299	36	0.9144	0.8925
30	294	34	0.8991	0.8963
40	289	31	0.8838	0.9031
50	283	28	0.8654	0.9099
60	266	24	0.8135	0.9172
80	229	17	0.7003	0.9309
100	206	11	0.6299	0.9493
120	185	8	0.5657	0.9585
150	149	6	0.4557	0.9612
180	106	4	0.3242	0.9636
210	62	2	0.1896	0.9688

We present our results in Table 1 and 2. The different detection results are obtained by varying the threshold parameter β as described in Section 4.3. Figure 6 show the output of our detector on some sample test images.

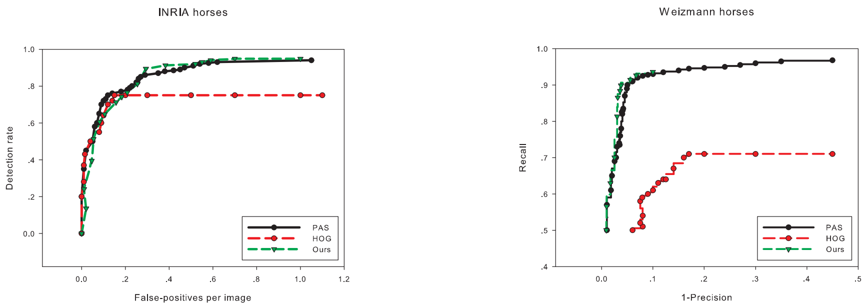
We compare our method with Dalal et al. [24] and Ferrari et al. [23] on the INRIA and Weizmann horses data sets. We randomly select 100 training images per category in Caltech101 and Google images to train classifiers. Dalal's method is currently the state of the art in human detection and has proven very competitive on other classes as well. The object detection method by Ferrari et al. achieved considerable gains on many object categories. Like ours, their object detectors is based on sliding a window subdivided into tile but uses different feature descriptors.

The results are displayed in Figure 8. Our detector achieves a substantially higher performance than HOG.

5.3. ETHZ Shape classes

The ETHZ shape database(collected by V. Ferrari et al. [22]) consists of five distinctive shape categories (apple logos, bottles, giraffes, mugs and swans) in a total of 255 images. All categories have significant intra-class variations, scale changes, and illumination changes. Moreover, many objects are surrounded by extensive background clutter and have interior contours.

We compare to [22] on the ETHZ shape database using the same detection system with the same settings. Experiments are conducted in 5-fold cross-validation. We split the entire set into half training and half test for each



(a) (b)

Figure 6: Comparison between our detector, Ferrari et al. [23] PAS-based one and Dalal et al. [24] HOG-based one.

Table 3: Comparison of detection performance with Ferrari et al. [22] on the ETHZ shape data set at 0.4 FPPI.

	Applelogos	Bottles	Giraffes	Mugs	Swans	average
Our method	88.6%(3.7)	83.4%(4.8)	83.9%(4.9)	83.5%(5.3)	87.5%(7.2)	85.4%
Ferrari	83.2%(1.7)	83.2%(7.5)	58.6%(4.6)	83.6%(8.6)	75.4%(13.4)	76.8%

category, and average performance from 5 random splits is reported. This is consistent with the implementation in [22] which reported the state-of-the art detection performance on this data set. Table 3 show our comparison to [22] on each of the categories. Average over all categories we improve the performance of [22] by 8.6% to 85.4%. On applelogos, giraffes and swans, we improve the performance by 5.4%, 25.3% and 12.1% respectively. On bottles and mugs, our approach performs comparable. We account the performance on the bottles and mugs to the shape which is less discriminant with respect to the background. As the data set was designed to test shape-based approaches, the improvements obtained by our approach underlines the versatility and adaptively of the hierarchical representation.

6. Conclusion

We have proposed a non-symmetry and anti-packing object pattern representation model (NAM) to represent a object category. This model can effectively codes global structure of object. The object pattern model consists of several part sub-patterns. We selected appropriate feature descriptor for the sub-pattern to deal with the local variation. In our work, the edge direction histogram descriptor introduced to describe the shape information of

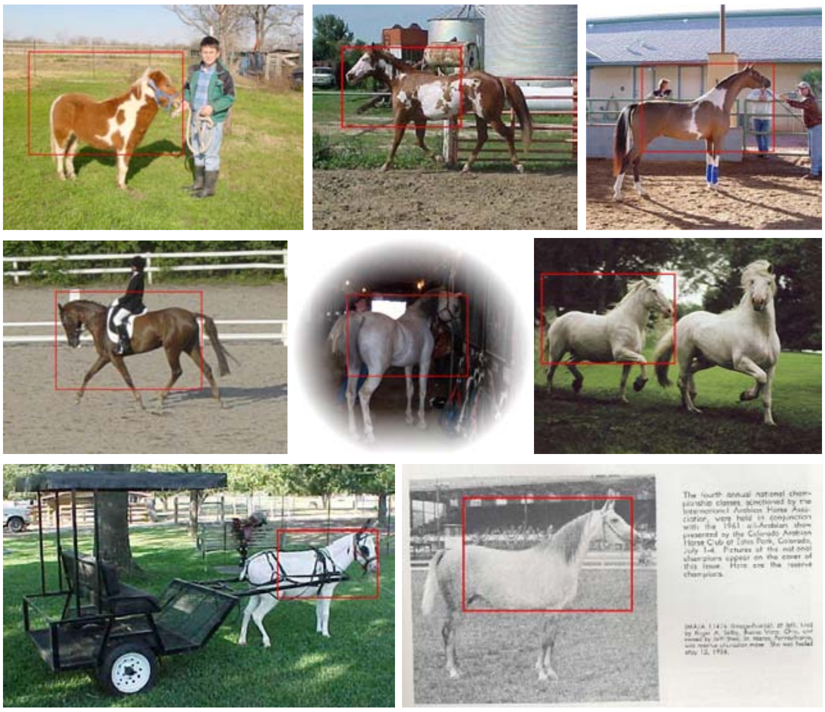


Figure 7: Detections for INRIA and Weizmann horses data sets when the value of threshold β is 50. In the middle row, the rightmost image shows a missed detection.

sub-patterns. Based on this representation, several learning classifiers were trained to detect sub-pattern instances. The proposed framework can be applied to any object category that consists of distinguishable parts arranged in a relatively fixed spatial configuration.

Acknowledgment

The authors wish to acknowledge the support of the National Natural Science Foundation of China under Grant No. 60973085.

References

[1] T.F. Cootes, G.J. Edwards and C.J. Taylor, “Active appearance models”. IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 681-685, 2001.6.

[2] J. Matthews and S. Baker, “Active appearance models revisited”. International Journal of Computer Vision, pp. 135-164, 2004.

[3] R. Fergus, A. Perona and A. Zisserman, “A sparse object category model for efficient learning and exhaustive recognition”. IEEE conf. on Computer Vision and Pattern Recognition, 2005.

[4] S. Agarwal, A. Awan and D. Roth, “Learning to detect objects in images via a sparse, part-based representation”. IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 1475-1490, 2004.

[5] P.F. Felzenszwalb, and D.P. Huttenlocher, “Pictorial structures for object recognition”. International Journal of Computer Vision, pp. 55-79, 2005.

[6] Y. Amit, and A. Trouve, “POP: Patchwork of parts models for object recognition”. International Journal of Computer Vision, pp. 267-282, 2007.

[7] L. Kyoung-Mi, “Component-based face detection and verification”. Pattern Recognition Letters, pp. 200-214, 2008.

[8] H. Schneiderman, and T. Kanade, “Object detection using the statistics of parts”. International Journal of Computer Vision,, pp. 151-177, 2004.

[9] A. Opelt , A. Pinz and A. Zisserman, “A boundary-fragment-model for object detection”. European conf. on Computer Vision, 2006.

[10] B. Leibe, A. Leonardis and B. Schiele, “Robust object detection with interleaved categorization and segmentation”. International Journal of Computer Vision, pp. 259-289, 2008.

[11] P. Felzenszwalb, D. McAllester and D. Ramanan, “A discriminatively trained, multiscale, deformable part model”. IEEE conf. on Computer Vision and Pattern Recognition, 2008.

[12] D. Crandall, P. Felzenszwalb and D. Huttenlocher, “Spatial priors for part-based recognition using statistical models”. IEEE conf. on Computer Vision and Pattern Recognition, 2005.

[13] F. Mahmoudi and J.Shanbehzadeh, “Image retrieval based on shape similarity by edge orientation autocorrelogram”. Pattern Recognition, pp. 1725-1736, 2003.

[14] D. Tao, X. Tang, X. Li and X. Wu, “Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval”. IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 1088-1099, 2006.

[15] T. Dacheng, T. Xiaoou and L. Xuelong, “Which components are important for interactive image searching?”. IEEE Transactions on Circuits and Systems for Video Technology, 2008.

[16] G. Xinbo, X. Bing, T. Dacheng and L. Xue long, “Image categorization: Graph edit distance + edge direction histogram”. Pattern Recognition, pp. 3179-3191, 2008.

[17] Pedro F. Felzenszwalb, Ross B. Girshick, D. McAllester and D.Ramanan, “Object detection with discriminatively trained part based models”. IEEE Trans. Pattern Analysis and Machine Intelligence, accepted, 2010.

[18] K. Young-Woo and O. Il-Seok, “Watermarking text document images using edge direction histograms”. Pattern Recognition Letters, pp. 1243-1251, 2004.

[19] M. Leordeanu, M. Hebert and R. Sukthankar, “Beyond local appearance: category recognition from pairwise interactions of simple features”. IEEE conf. on Computer Vision and Pattern Recognition, 2007.

[20] A.Opelt, A. Pinz, M.Fussenegger, and P.Auer, “Generic object recognition with boosting”. IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 516-431, 2006.

[21] J. Shotton, A. Blake and R. Cipolla “Multi-Scale categorical object recognition using contour fragments ”. IEEE Trans. Pattern Analysis and Machine Intelligence, 2008.

[22] V. Ferrari, L. Fevrier and C. Schmid “Accurate object detection with deformable shape models learnt fromimages ”. In CVPR, 2007

[23] V. Ferrari, L. Fevrier, F.Jurie and C. Schmid “Groups of adjacent contour segments for object detection ”. IEEE Transaction on MAMI, 2008

[24] N. Dalal and B. Triggs “Histgrams of oriented gradients for human detection ”. Proc. Conf. Computer Vision and Pattern Recognition, 2005

[25] P. Viola and M.jones “Rapid object Detection using a boosted cascade of simple features ”. Proc. Conf. Computer Vision and Pattern Recognition, 2001.