MPLBOOST-BASED MIXTURE MODEL FOR EFFECTIVE HUMAN DETECTION WITH DEFORMABLE PART MODEL

Chaoran Gu, Luntian Mou, Yonghong Tian, Tiejun Huang[1]

National Engineering Laboratory for Video Technology, School of EE & CS, Peking University

ABSTRACT

The Deformable Part Model has shown high accuracy in tackling certain occlusion or deformations of objects such as cars and bikes. However, as for human category characterized by a larger number of articulated parts and more significant appearance variations, its performance gain is not so remarkable. To address this issue, we propose an MPLBoost-based mixture model which splits data into coherent groups and trains one root classifier for each, resulting in automated selection of discriminative root models and better representation of intra-class variations through visual feature clustering. Based on this boosting framework, multiple complementary features are combined to capture shape, texture and color information. Experimental results demonstrate that the proposed model can achieve an impressive performance improvement, especially in handling larger variations of human poses and viewpoints.

Index Terms— MPLBoost, mixture model, deformable part model, feature combination, human detection

1. INTRODUCTION

Over the last few years, human detection in still images or videos has received a lot of attention. It is widely used in visual surveillance, robotics, automotive safety and advanced human machine interfaces. But as indicated by [1, 3], despite significant progress, its performance still has much room for improvement. The recall of the best methods is still below 50% for 0.1 false positive per image in most frequently used datasets. Presence of occlusion, pose changes, viewpoint and other intra-class variation render human detection a still challenging task.

Existing approaches with the best performance typically follow a sliding window paradigm which entails feature extraction, binary classification (such as boosting or support vector machines) and dense multi-scale scanning of detection windows followed by non-maximum suppression.

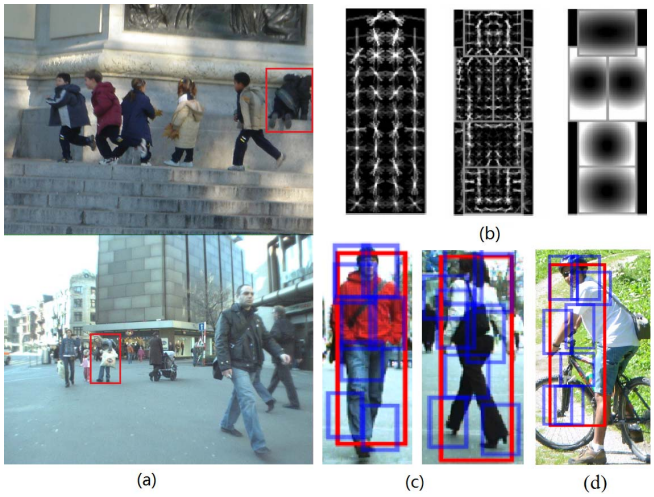

Fig. 1. (a) Objects in red rectangle with huge deformations can't be detected by DPM (b) The model is defined by a coarse root filter (left) , several higher resolution part filters (middle), a spatial model for the location of each part relative to the root (right). (c) Detections obtained with a single component person model. (d) A false positive.

Deformable part models (DPM) [4, 5] which accord with above schema have recently got a lot of publicity for tackling this challenge. DPM incorporates a higher degree of learnt invariance by partitioning the object model into a set of local parts which are allowed to move around subject to soft spatial constraints. Take the human in Fig. 1(c) as an example. The score of a hypothesis is given by the scores of a root filter plus all part filters and minus a deformation cost that depends on relative position of each part with respect to the root. But for human with larger variation as shown in Fig. 1(a) and (d), the result is not satisfactory. Although the part of left-head-shoulder is visible in Fig. 1(d), its part detection score is relatively low because its visual cue in the image is meaningless and does not fit the part detector well. It's not reasonable to detect this human using model Fig. 1(b). DPM method adopts the idea of a mixture model to capture larger intra-variation in appearance that an object

[1] This work is partially supported by grants from the Chinese National Natural Science Foundation under contract No.61035001 and No.61072095, and National Basic Research Program of China under contract No.2009CB320906. Contact the authors via yhtian@pku.edu.cn.

category may exhibit. Specially, the mixture model segregates object instances into disjoint groups according to aspect-ratio, and then learns a separate model per group. However, their experiments show that their mixture models are important for the car or bike category but not for person category of PASCAL VOC dataset. Besides, assignment of positive samples to components in LSVM formulation is a non-convex optimization and thus sensitive to initialization. Therefore, we propose a robust and discriminative mixture model suitable for human category without heuristic initialization and make it more data-driven than before.

The performances of sliding-window-based detectors are mainly determined by two factors: the feature and the underlying classification algorithm. In this work, we also aim at improving the performance of human detectors from the aspect of the feature based on DPM's data-mining hard examples method. Augmenting feature set to include both contrast sensitive and contrast insensitive features is usually a common way to improve performance. The importance of feature combination has been widely recognized by the computer vision community [12, 14, 15]. More discriminative feature type can capture more effective information about shape, texture, color etc. We extend this idea by seeking complementary features via making full use of the boosting framework's nature feature fusion schema.

Our approach builds on the state-of-the-art human detector of [5], which we extend in two ways. First, a new mixture model is proposed that separates data into coherent groups to train discriminative root classifiers and second, complementary feature combination form for DPM is presented. The experimental results show that our new model performs well on challenging datasets with multiple viewpoints and poses, showing an improvement to [5] by more than 2.1% and 6.1% in recall at 0.1 FPPI on INRIA and ETH datasets respectively.

2. RELATED WORK

Felzenszwalb points out in [5] that their work allow for exploration of additional latent structure such as mixture models with many components. However, no remarkable work has been done until [6] demonstrates that the conceptually simpler notion of mixture model is indeed an equally important contribution in the DPM detector. They performed an appearance-based clustering by unsupervised method (K-means algorithm using Euclidean distance function) to initialize the component instead of aspect-ratio and left-right flipping heuristic. However, similar to [5], we observe that their methods bring no performance gain or perform even worse on human category. It's also observed in [20, 21] that the initialization of DPM components significantly influences detection performance and the standard way to initialize the components based on the bounding box aspect ratio does not appear to be effective because people with different poses often have similar Bounding boxes. Thus, an alternative initialization strategy is explored by clustering the images according to the relative displacement of the 2D joint locations with regard to the fixed body joint. But the method of [20, 21] relies on manual annotations and prior knowledge.

When training a model for object detection, we often face the asymmetric problem that negative training set is much larger than positive one. To resolve this problem, a "representative" set of negative examples can be collected during training. The most common method is bootstrapping [16] which selectively adds images to the training set as training progresses instead of collecting the images before training is started. [5] proposes a new data-mining algorithm motivated by the bootstrapping idea. In each iteration, the algorithm removes easy feature vectors which contain latent information from the cache and adds new hard feature vectors and then retrain the model. Note that a feature vector is not considered easy even if there is another easy feature vector belonging to the same example. Although [12, 13, 14, 15] have already pointed out that the combination of multiple and complementary feature types can help improve performance, they didn't take into account finding "hard negative instances" procedure during training. However, [8, 11] indicate that the number of bootstrapping rounds and choice of the classifier-feature combination are crucial to reach final optimal performance. Different and improper iterative bootstrapping may lead to incorrect conclusions about the performance of different feature sets. Our method is to seek for complementary features based on DPM's data-mining hard examples methods and make use of boosting fusion schema for combination. In this sense, our work can be seen as an extension of [11], using different features which have recently been proved to be effective.

3. PROPOSED METHOD

3.1. MPLBoost and LatSVM review

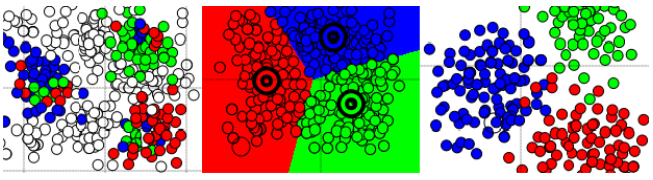

Fig. 2. Illustration of XOR problem. The positive class (colorized circle) Comprising three sub-clusters and the negative class (white circle) in background make the XOR configuration (Left). Conventional clustering methods (e.g. k-means) can't solve this problem (middle). MPLBoost construct different clusters which have their own distinct expertise. The method learns three classifiers that nicely settle into desired clusters and decision boundaries (right).

Human positive samples appear in multiple clusters often arrange in a XOR layout [9], see Fig. 2 for details. MPLBoost solve this problem by assuming there are K latent variables $y_i^k \in \{-1, +1\}$ associated with each sample. Each latent variable defines membership to one of the K

groups. A sample is considered positive if it belongs to at least one of these groups, which can be expressed as follows:

$$y_i = \max_k (y_i^k)$$

Our goal is to simultaneously split the positive data into K groups and train K classifiers $h^1, \ldots, h^K$, one per group, so that $\max_k (h^k(x_i)) = y_i$. This problem is called Multiple Pose Learning in [3]. MPLBoost optimize $\mathcal{L}(h^1,\ldots,h^K)$ by coordinate descent, cycling through k, where in each phase we train and add a weak classifier to h^k while keeping all other weak classifiers fixed. The whole algorithm is shown in Line 2-9 of Fig. 3. We refer the reader to [3, 9] for more details.

Felzenszwalb et al. [5] introduce a coordinate-descent algorithm for learning the model parameters: the Latent-SVM. The method simultaneously learns the object detector and parts detectors without part-level training annotation. The problem of part discovery is casted as a multiple instance SVM learning problem. It is possible to reformulate the DPM learning as a structural SVM learning problem. In the standard (fully supervised) framework for training of an object detector, positive images are annotated with the locations of object bounding boxes, but the part locations are treated as latent information. The LSVM learning procedure acquires part appearance and layout parameters by alternating between making assignments to latent variables (part locations in training images) given the model parameters, and re-optimizing the model parameters given the latent variable assignments. This optimization framework has been very successful at discovering useful latent part structure in highly deformable categories with large intra-class appearance variability.

3.2. Mixture Model

MPLBoost can be used for wide-applications in perceptual data exploration [9]. It generally solves a new co-clustering problem of a data set (e.g. a set of face images) and a feature set (e.g. simple visual features) in a way to maximize discrimination of the data set from another data set (e.g. a set of random images). Inspired by this idea, we model the human as an aggregation of visual features. Each strong classifier in MPLBoost selects a series of visual features. Our mixture model composed of subsets of visual features. Experiment shows that it works well on object categories that have several distinct views. See Figure 3 for the pseudo code of our MPLBoost-based Mixture model. The procedure is outlined below:

The outmost loop implements a fixed number of iterations of coordinate decent on latent "component".

MPLBoost feature learning: Line 2-9. MPLBoost learn multiple strong classifiers in parallel, allowing individual classifiers to focus on specific regions of the feature space without degrading the overall classification performance.

Positive Latent Fixation: Line 10-13. Bounding boxes are treated as "partially latent" and allowed to move in a small neighborhood of the initial position to compensate for noisy annotation. We allow sliding around in the neighborhood of bounding box to maximize the model score. The detection window of a root filter specified by a hypothesis overlaps with bounding box B by at least 70%.

Hard negatives Harvest: Line 14-17. Similar to [5], we make use of the data mining methods to select a relatively small number of representative negative examples and converge to the exact solution of the training problem defined by a large training set.

Initializing Root Filters: Line 19-21. We use the learnt feature and SVM to train root filter of each component which has extreme intra-class variation.

Input :
Positive examples $P = \{(I_1, B_1), \ldots, (I_n, B_n)\}$
Negative examples $N = \{J_1, \ldots, J_m\}$
Initialization :
Intialize F_p *using bounding boxes of positive examples from P*
Initialize F_n *using negative examples sampled sequently from N*
Initialize weights w_i, *set* $w_i^k = 1/(N_{pos} \times K)$ *if* $x_i \in k$
and $w_i^k = 0$ *otherwise; for negative examples* $w_i^k = -1/N_{neg}$
Training :
1 *for relable* := 1 *to num-relable do*
2 *for* $t = 1$ *to* T *do*
3 *for* $k = 1$ *to* K *do*
4 *Compute weights* $\omega_i^k = -\dfrac{\partial \mathcal{L}}{\partial h_i^k}$
5 *Train weak classifier* h_t^k *using weights* $|\omega_i^k|$
$h_t^k = \underset{h}{argmin} \sum_i \mathbf{1}(h(x_i) \neq y_i) |w_i^k|$
6 *Find* α_t *via line search to minimize* $\mathcal{L}(\cdot, h^k, \cdot)$
$\alpha_t = \underset{\alpha}{argmin} \mathcal{L}(\cdot, h^k + \alpha h_t^k, \cdot)$
7 *Update strong classifier* $h^k \leftarrow h^k + \alpha_t^k h_t^k$
8 *end*
9 *end*
10 $F_p := \phi$
11 *for* $i := 1$ *to* n *do*
12 *Add highest scoring hypothesis that overlaps* B_i *in* I_i *to* F_p^k
13 *end*
14 *for* $j := 1$ *to* m *do*
15 *if* $|F_n| \geq$ *memory - limit then break*
16 *if* $\max_k(y_j^k) < t$ *Add* f_j *to* F_n
17 *end*
18 *end*
19 *for* $c := 1$ *to* K *do*
20 *train each component root using a standard SVM*
21 *end*

Fig. 3. MPLBoost-based Mixture Model Learning Algorithm

The rest work is similar to [5]. We skip the merging components procedure; initialize part classifiers from the root classifier. We lift the restriction that parts should be symmetric along the vertical axis; otherwise it will degrade discrimina-

tive performance of different component. Then we use latent SVM to train part classifiers and geometry relationship between parts simultaneously.

3.3 Implementation details

Weak Classifier: Different from [3] which uses decision stump as weak classifier, we choose the separating hyperplane computed using a linear SVM. We select the linear kernel because of its fast training and classification speed.

Variable size blocks: We construct variable-size blocks in our method, which is similar to the one proposed in [18]. We increase our feature space to include blocks of different sizes, locations and aspect ratios. We consider all blocks whose size ranges from 16 x 16 to 32 x 48 in a 64 x 128 detection window. The ratio between block width and block height includes (1:1), (1:2) and (2:1). The step-size includes 4, 6 and 8 pixels. As a result, we have 383 blocks to choose from, compared to the 105 blocks used in the Dalal-Triggs algorithm, each of which contains a 36-D histogram vector of HOG.

MPLBoost initialization: Inspired by [6] which choose the K-means clustering method using Euclidean distance function to provide a good initialization, we perform soft K-means on the initial data set to get quick convergence for MPLBoost. K-means describes the ownership of a data point belonging to a group by using a responsibility vector $\boldsymbol{r_n}$, a binary vector with length k, where the $\boldsymbol{k}$th element $\boldsymbol{r_n}$ is 1 if data point n belongs to group k, and 0 otherwise. Soft K-means is a variation of K-means: instead of giving each data point a hard assignment, it assigns each data point to more than one group with some probability. In this case, $\boldsymbol{r_n}$ is not necessarily a binary vector. Let the $\boldsymbol{k}$th element of $\boldsymbol{r_n}$ be the probability of data point n belonging to the group k, so $\sum_{i=1}^{k} r_{n,k} = 1$ and $r_{n,k} \geq 0$. Soft k-means is useful when we do not want to assign a point into any certain group since the point may have some floating. It helps in finding optimal data clusters and associated feature sets in an unsupervised manner.

Number of Parts and K: The optimal number of parts depends on the variability of an object class and may be significantly different between classes. We get the best performance with 8 parts, so we use this number in all the subsequent experiments. Experiment was also designed to identify the optimal K, the number of groups required for the pedestrian models. We trained 10 models with different K using the exact same training set and tested them on the testing set. Experiment shows the best performance using the group number of 9.

4. FEATURES AND COMBINATION

By carefully observing the part filters chosen by DPM in Fig. 1(c), we can draw three conclusions roughly:

(1)The selected features mainly focus on the silhouette of human such as head and shoulder, flank of legs and so on;
(2)The part filter contains both the silhouette area and color consistent area;
(3) The part filter encodes the abundant texture where have similar color. So, it implicitly encodes shape, color and texture information that boost its discriminative ability.

Lots of Features including edge of histogram, haar features, covariance features, motion information, integral channel features and so on have been proposed to encode different information, providing us with a much richer descriptor set. We focus on three of them which recently proved to be effective as follows.

HOG: Histograms of oriented gradients is a most popular feature for object detection. They collect gradient information in local cells into histograms using trilinear interpolation, and normalize overlapping blocks composed of neighboring cells. Interpolation, local normalization and histogram binning make the representation robust to changes in lighting conditions and small variations in pose.

CSLBP: Cell-Structured LBP [17] feature combines the human shape information with local texture information, and encodes spatial relationship in different granularity to improve its discriminative ability. The motivation lies that HOG can only capture the human silhouette information and ignores some informative texture, so combination of both can capture extra information to improve performance. Many previous work demonstrate that only subset of LBP features are effective for localization while others not. **CSS** We encode color self-similarities [11] in HSV space within the descriptor window, i.e. similarities between colors in different sub-regions. To leverage the robustness of local histograms, we compute D local color histograms over 8 x 8 pixel blocks, using trilinear interpolation as in HOG to minimize aliasing. We use histogram intersection to computed pairwise similarities between histograms. We apply L2-normalization to the (D x (D-1) / 2)-dimensional vector of similarities. In our implementation D = 128 blocks, so CSS has 8128 dimensions.

Combination: To combine features we L2-normalize each feature type and concatenate all vectors. Since high dimension features are usually noisy, there are two common methods to find the most discriminative features: PCA used by [5] and boosting. Boosting is born to be suitable for solving multi-feature fusion problems. MPLBoost decreases the number of features, removes irrelevant or noisy data, and increases mining performance such as predictive accuracy and comprehensibility. We do little modification to use MPLBoost for selecting effective visual features.

5 EMPIRICAL RESULTS

We evaluate the proposed method on the publicly available INRIA (easier one) and ETH (harder one) datasets which are widely used as benchmark datasets for human detection. As our baseline system, we use the latest release

of the DPM detector [4] (without the bounding-box prediction and context-rescoring steps).

For training we use INRIA's positive training data which has 2416 human annotations and extend this to include 4500 annotations in total, using labeled data from [19] to cover a wide range of shape and pose variations. This expansion is necessary to guarantee each component will get enough and balanced training data. Otherwise we observe that an unequal distribution of training samples among different components occurs and leads to performance decrease. All experiments are conducted using this synthetic training dataset.

Testing is done under a state-of-the-art and unified evaluation framework [1]. A detected bounding box (BB) and a ground truth BB form a potential match if their overlapping area exceeds 50%. As a recent popular reference point, we will report the obtained recall at 0.1 FPPI which result in an acceptable number of false positives. We standardize all ground truth BB to have a width of 0.41 times the height during evaluation and all evaluation results are reported on pedestrians 50 pixels and up as described in [1].

We begin our analysis with the INRIA dataset. Detection results are shown in Fig. 4 as miss rate vs. false positives per image curves. As our main contribution is a modified mixture model of DPM, a straightforward comparison is applying the original LatSVM baseline and our detector to human detection task. Performance gain is observed by 2.1% improvement over LatSVM baseline. However, the improvement is limit partly because INRIA consists of images of standing or walking (upright) pedestrians with little occlusion. In addition, several detections showing our advantage like Fig. 1(a) are not labeled as ground truth. Besides, our detector belongs to state-of-the-art detectors such as FPDW, ChnFtrs, VeryFast [1].

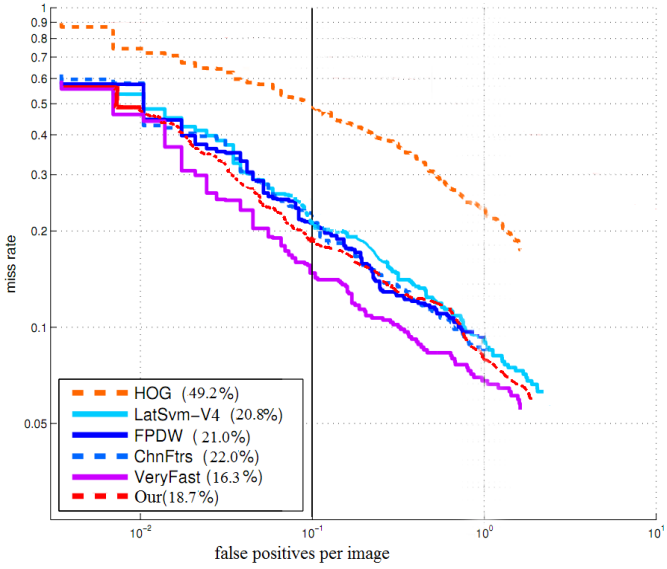

Fig. 4. Comparison to the state-of-the-art detectors on the INRIA dataset. The miss rate at 0.1 FPPI is reported in the legend.

We evaluate on a more complicated dataset-ETH pedestrian. Three sets (999 frames "BAHNHOF" sequence; 451 frames "JELMOLI" sequence; 354 frames "SUNNY DAY" sequence) are evaluated together and the annotations used are the updated annotations provided by Christian Wojek in [14]. Our new mixture model achieves 4.6% at 0.1FPPI and 7.0% at 1FPPI performance gain .We make trivial modification on LatSVM baseline, and our performance gains exactly come from an underlying change in the mixture model during learning.

Besides, Fig. 5 also illustrates miss rate vs. FPPI curves obtained with different feature combination. All these three block-based features contribute to boost in performance. CSS produces consistent and robust improvements: 1.2% at 0.1FPPI and 5.1% at 1FPPI compared to using HOG only, which indicates that color self-similarity is indeed complementary to gradient information. CSLBP improves performance at high precision region but gradually losing its impact as FPPI increases. Empirically it's not as robust as CSS.

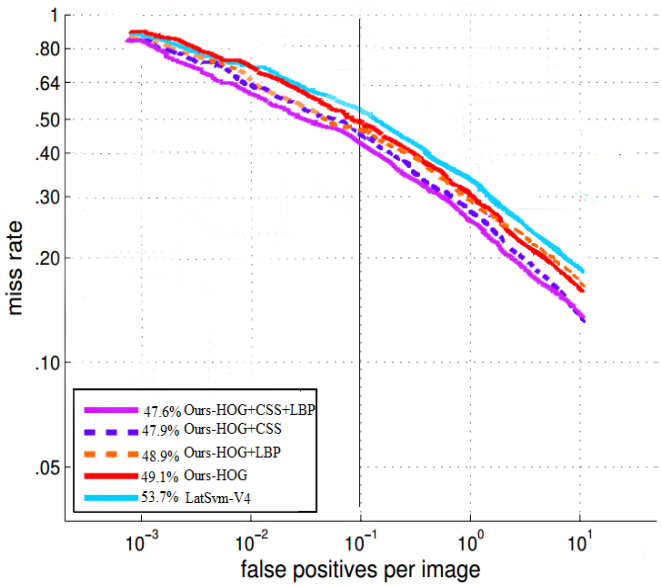

Fig. 5. Comparison of different features combinations on the ETH dataset. The miss rate at 0.1 FPPI is reported in the legend.

HOG+CSS+CSLBP achieve the best with 52.4% recall at 0.1 FPPI and outperform other state-of-the-art detectors.

In summary, we can state that the combination of different features is successful to improve state-of-the-art human detection performance. Several conclusions can be drawn from these results: 1) the performance of DPM is still limited by the robust low-level feature representation. 2) A combination of HOG features and CSS is able to achieve about 5% better recall and is a good choice which balances speed and accuracy. 3) Combination strategy is different from previous work [11], demonstrating meaningful research due to DPM's data-mining hard examples method.

Some qualitative results of our detectors on the INRIA and ETH test images are shown in Fig. 6. The detector using new mixture model outperforms original DPM on the basis of high baseline, especially in handling larger pose and viewpoint variation, as shown in Fig. 6.

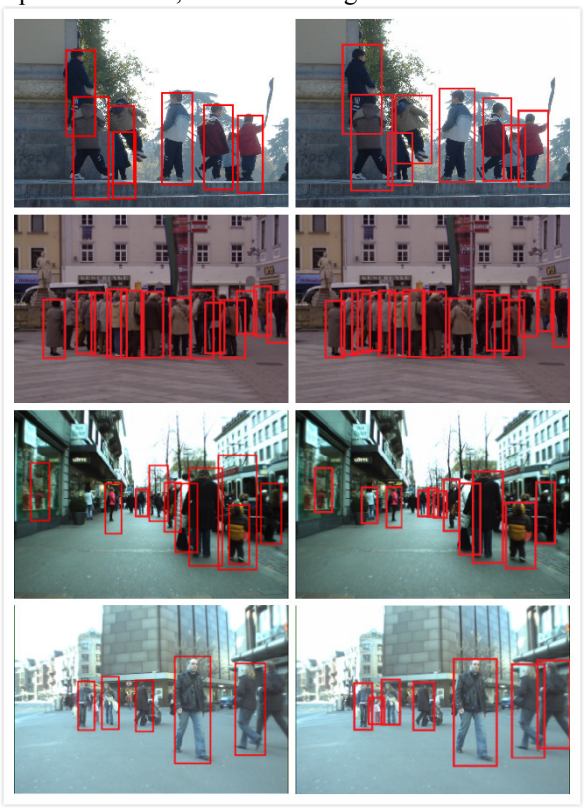

Fig. 6. Qualitative results of LatSVM baseline (left column) and our method (right column). The top two test images are from INRIA dataset and bottom two from ETH.

6. CONCLUSIONS

Our study aims to address the problem of DPM's performance bottleneck caused by larger intra-class variation such as viewpoint, pose etc., focusing on challenging human category. For this purpose, we introduce a novel approach to learning mixture models using MPLBoost-based visual feature selection methods which only needs trivial modification compared to original DPM. In order to make full use of shape, texture and color information, we combine block-based complementary features based on above boosting framework. We evaluate our approach on two popular benchmarks and get more than 4% performance gain on average at 0.1 FPPI and report competitive performance with respect to current leading techniques.

7. REFERENCES

[1] Dollar, P., et al., "Pedestrian Detection: An Evaluation of the State of the Art," Pattern Analysis and Machine Intelligence (PAMI) 2011, IEEE Transactions on, 2012.

[2] Dollar, P., et al., "Pedestrian detection: A benchmark," CVPR, 2009.

[3] Boris Babenko, et al., "Simultaneous Learning and Alignment: Multi-Instance and Multi-Pose Learning," Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, 2008.

[4] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively trained deformable part models," release4 http://www.cs.brown.edu/~pff/latent-release4/.

[5] Felzenszwalb,P.F., et al., "Object Detection with Discriminatively Trained Part-Based Models," PAMI 2010, IEEE Transactions on, 2010. 32(9): p. 1627-1645.

[6] SK Divvala, et al., "How important are 'Deformable Parts' in the deformable part model," CVPR, 2012.

[7] Pandey, M. and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," ICCV, 2011.

[8] Benenson, R., et al., "Pedestrian detection at 100 frames per second," CVPR, 2012.

[9] T.-K. Kim and R. Cipolla, "MCBoost: Multiple Classifier Boosting for Perceptual Co-clustering of Images and Visual Features," Advances in Neural Information Processing Systems (NIPS), 2008.

[10] Schwartz, W.R., et al., "Human detection using partial least squares analysis," ICCV, 2009.

[11] Walk, S., et al., "New Features and Insights for Pedestrian Detection," CVPR, 2010.

[12] Wojek, C., S. Walk and B. Schiele, "Multi-Cue Onboard Pedestrian Detection," CVPR, 2009.

[13] Gavrila, D.M. and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," International journal of computer vision (IJCV), 2007.

[14] Wojek, C. and B. Schiele, "A performance evaluation of single and multi-feature people detection," Pattern Recognition, 2008: p. 82-91.

[15] Wu, B. and R. Nevatia, "Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection," CVPR, 2008.

[16] Rowley, H.A., et al., "Human face detection in visual scenes," 1995: School of Computer Science, Carnegie Mellon University.

[17] Wang, X., T.X. Han and S. Yan., "An HOG-LBP human detector with partial occlusion handling," ICCV, 2009.

[18] Zhu, Q., et al., "Fast human detection using a cascade of histograms of oriented gradients," CVPR, 2006.

[19] Overett, G., et al., "A new pedestrian dataset for supervised learning," Intelligent Vehicles Symposium, 2008.

[20] Pishchulin, L., et al., "Articulated people detection and pose estimation: Reshaping the future," CVPR, 2012.

[21] Lopez Sastre, R.J., T. Tuytelaars and S. Savarese. "Deformable part models revisited: A performance evaluation for object category pose estimation". IEEE International Conference of Computer Vision Workshops(ICCV Workshops), 2011