k-fold Subsampling based Sequential Backward Feature Elimination

Jeonghwan Park, Kang Li and Huiyu Zhou
School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, U.K.

Keywords: Feature Selection, Appearance Model, Human Detection.

Abstract: We present a new wrapper feature selection algorithm for human detection. This algorithm is a hybrid feature selection approach combining the benefits of filter and wrapper methods. It allows the selection of an optimal feature vector that well represents the shapes of the subjects in the images. In detail, the proposed feature selection algorithm adopts the k-fold subsampling and sequential backward elimination approach, while the standard linear support vector machine (SVM) is used as the classifier for human detection. We apply the proposed algorithm to the publicly accessible INRIA and ETH pedestrian full image datasets with the PASCAL VOC evaluation criteria. Compared to other state of the arts algorithms, our feature selection based approach can improve the detection speed of the SVM classifier by over 50% with up to 2% better detection accuracy. Our algorithm also outperforms the equivalent systems introduced in the deformable part model approach with around 9% improvement in the detection accuracy.

1 INTRODUCTION

A feature is an individual measurable property of a process being observed. Using a set of features, a machine learning algorithm can perform necessary classification (Chandrashekar and Sahinn, 2014). Compared to the situation back to the early years in the pattern recognition community, the space of features to be handled has been significantly expanded. High dimensionality of a feature vector is known to decrease the machine learning performance (Guyon and Elisseeff, 2003), and directly affects applications such as human detection systems whose system performance relies heavily on both the classification speed and accuracy. A feature with no association with a class is regarded as a redundant or irrelevant feature. A redundant feature represents a feature which does not contribute much or at all to the classification task. An irrelevant feature can be defined as a feature which may only lead to decreased classification accuracy and speed. Blum (*Blum*, 1997) defined the relevant feature f as a feature which is useful to a machine learning algorithm L with respect to a subset of features $\{S\}$: the accuracy of an hypothesized algorithm using the feature set $\{f \cup S\}$ is higher than that only using $\{S\}$. In pattern recognition, the aim of feature selection is to select relevant features (an optimal subset), which can maximise the classification accuracy, from the full feature space. When a feature selection process is applied to pattern recognition, it can be seen as an embedded automated process which removes redundant or irrelevant features, and selects a meaningful subset of features. Feature selection offers many benefits in understanding data, reducing computational cost, reducing data dimensionality and improving the classifier's performance (Chandrashekar and Sahinn, 2014). In general, feature selection is distinguished from vector dimensionality reduction techniques such as Principal Component Analysis (PCA) (Alpaydin, 2004), as feature selection merely selects an optimal subset of features without altering the original information. Feature selection requires a scoring mechanism to evaluate the relevancy of features to individual classes. The scoring mechanism is also named the feature selection criterion, which must be followed by an optimal subset selection procedure. Naively evaluating all the subsets of features (2^N) becomes an NP-hard problem (Amaldi and Kann, 1998) as the number of features grows, and this search becomes quickly computationally intractable. To overcome this computation problem, a wide range of search strategies have been introduced, including best-first, branch-and-bound, simulated annealing and genetic algorithms (Kohavi and John, 1997)(Guyon and Elisseeff, 2003), etc. In terms of feature scoring, feature selection methods have been broadly categorised into filter and wrapper methods (Kohavi and John, 1997). Filter methods allow one to rank features using a proxy measure such as the distance between features and a class, and select

Park, J., Li, K. and Zhou, H.
k-fold Subsampling based Sequential Backward Feature Elimination.
DOI: 10.5220/0005688804230430
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 423-430
ISBN: 978-989-758-173-1

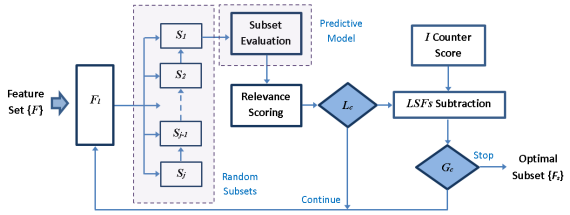

Figure 1: Proposed Feature Selection System Overview.

the most highly ranked features to be a candidate feature set. Wrapper methods score features using the power of a predictive model to select an optimal subset of features.

In this paper, we propose a hybrid feature selection approach which combines the benefits of filter and wrapper methods, and discuss the performance improvement of the human detection system achieved using the proposed algorithm. We also demonstrate how optimal features selected by the proposed algorithm can be used to train an appearance model. The rest of the report is organised as follows. In section 2, feature selection methods and their core techniques are briefly reviewed. Section 3 introduces the proposed feature selection algorithm. The experiment results of the pedestrian detection system using the proposed algorithm is described in Section 4. Finally, Section 5 concludes the paper.

2 RELATED WORK

One way for feature selection is simply evaluating features based on their information content, using measures like interclass distance, statistical dependence or information-theoretic measures (Estevez et al., 2009). The evaluation is independently performed against different features, and the evaluation result called "feature rank" is directly used to define the usefulness of each feature for classification. Entropy and Mutual information are popular ranking methods to evaluate the relevancy of features (Foithong et al., 2012)(Peng et al., 2005)(Javed et al., 2012)(Estevez et al., 2009). *Zhou et al.* (Zhou et al., 2011) used the *Rényi* entropy for feature relevance evaluation of overall inputs in their automatic scaling SVM. *Battiti* proposed the MIFS algorithm (Battiti, 1994), which selects the most relevant k feature elements from an initial set of n feature dimensions, using a greedy selection method. Many MIFS variations have been introduced since then such as the mRMR (Peng et al., 2005), which used the first-order incremental search mechanism to select the most relevant feature element at a time. *Estevez et al.* (Estevez et al., 2009) replaced the mutual information calculation in the MIFS by the minimum entropy of two features.

Wrapper methods utilise classifier's performance to evaluate feature subsets. Wrapper methods have a significant advantage over filter methods as the classifier (learning machine) used for evaluation is considered as a black box. This flexible framework was proposed in (Kohavi and John, 1997). *Gutlein et al.* (Gutlein et al., 2009) proposed to shortlist k ranked feature elements firstly, and then applied a wrapper sequential forward search over the features. *Ruiz et al.* (Ruiz et al., 2006) proposed an incremental wrapper-based subset selection algorithm (IWSS), which identified the optimal feature area before the exhaustive search was applied. *Bermejo et al.* (Bermejo et al., 2011) improved the IWSS by introducing a feature replacement approach. *Foithong et al.* (Foithong et al., 2012) used the CMI algorithm to guide the search of an optimal feature space, and applied the VPRMS as an intermediate stage before the wrapper method started. *Pohjalainen et al.* (Pohjalainen et al., 2013) proposed the RSFS which used the dummy feature relevance score as a selection criterion. *Li and Peng* (Li and Peng, 2007) introduced a fast model-based approach to select a set of significant inputs to a nonlinear system. *Heng et al.* (Heng et al., 2012) addressed the overfitting problem of the wrapper methods by proposing a shrink boost method. *Yang et al.* (Yang et al., 2012) proposed a wrapper method with the LRGA algorithm to learn a Laplacian matrix for the subset evaluation.

3 PROPOSED ALGORITHM

In this section, a novel feature selection algorithm is presented. The proposed feature selection algorithm is a wrapper method which adopts the k-fold subsampling method in the validation step. The search strategy is an exhaustive search, namely the sequential backward elimination (SBE, *Marill and Green* 1963). The proposed algorithm is a classifier dependent feature selection method, and the linear SVM is used as the classifier to evaluate the selected feature subsets.

3.1 Frame Work

First, the entire feature set $\{F\}$ is randomly divided into j small subsets where an evaluation process is performed, $S_n \in F,\ for\ n = 1,2,..,j$. Square root calculation on the size of the full feature set $\{F\}$ is used to determine the size of a subset. When the local stopping criterion L_c is satisfied, another relevance score I contributes to the computation of feature relevance scores, and considerably a large number of

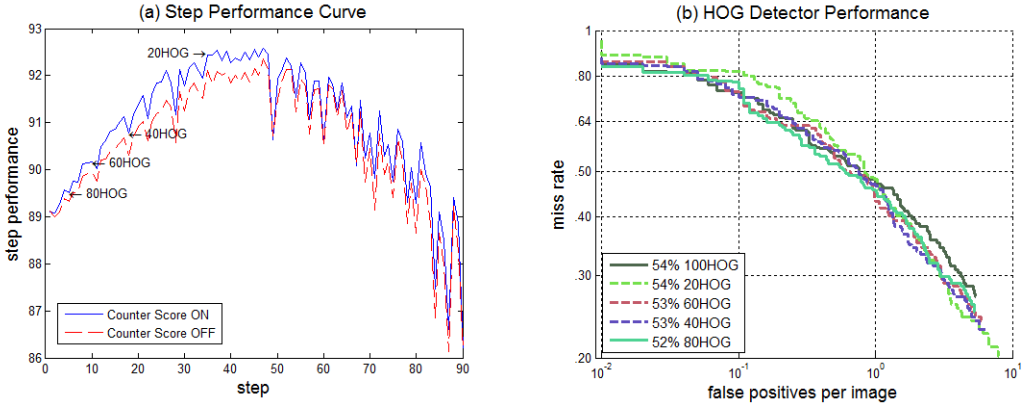

Figure 2: Performance illustration: (a) Step performance curve of the proposed feature selection algorithm. (b) Human Detection performance using different HOG detectors trained with different feature subsets. The evaluation was carried out on the first 100 INRIA Full Images. The computed miss rates between 0.01 and 1 false positives per-image (FPPI) are shown in the legend. HOG Detectors trained with optimal feature subsets perform better than or similar to 100HOG showing a small overfitting problem.

irrelevant/redundant feature elements (from the least significant one, $LSFs$) are subtracted. The process continues on until the global stopping criterion is satisfied. The whole process is sketched in Fig 1.

Random Subsets: When the remaining feature set $\{F_l\}$ is reset, a temporary ID is given to individual vector elements for backtracking which gives them an equal opportunity in the evaluation. The temporary ID is only valid within one step. A step refers to the point where the local stopping criterion L_c is satisfied, $LSFs$ are subtracted and all the iteration parameters are re-initialised. An iteration refers to that the evaluation has completed over all the subsets, and each vector element has been evaluated once. At the beginning of each iteration within one step, the feature vector elements of the remaining feature set $\{F_l\}$ is randomly re-arranged and divided into subsets, $n \times \{S\}$. The size of a subset is chosen as $\sqrt{N}$ where N is the number of the remaining features.

Relevance Score: The algorithm uses two scores, step performance score P and feature relevance score R. During each iteration in the exhaustive search stage, the relevance score of individual features is updated according to the prediction power score over the subset where the feature participates. In this paper, the Unweighted Average Recall (UAR) is used to calculate the prediction power score over the subsets:

$$P\{S_n\} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{C_{iP}}{C_{iP}+C_{iF}}\right) \tag{1}$$

where S_n is nth subset, i is the class index, C_{iP} is the class true positive (correct prediction) and C_{iF} is the class false negative (wrong prediction).

The individual feature relevance score is then updated as (Pohjalainen et al., 2013):

$$R_f = R_f + P\{S_n\} - E, \quad f \in S_n, \tag{2}$$

where E is the UAR of the cumulated C_P and C_F over a step. As the search continues, the feature score R_f represents how much the corresponding feature has contributed to the prediction.

Stopping Criterion: The local stopping criterion L_c is calculated using the standard deviation of the step's performance with a specific predictor. L_c is defined as:

$$L_c = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - \frac{1}{N}\sum_{j=1}^{N}P_j\right)^2} \tag{3}$$

where P_i is the step performance score. L_c is then compared with a supervised parameter (0.6 in this study) to decide the evaluation fairness over all the features. The global stopping criterion G_c is a supervised parameter, and the number of features to be finally selected is used here.

Counter Score: Even though an algorithm has performed a large number of iterations and the local stopping criterion L_c has been satisfied, it can not always be guaranteed that the chosen $LSFs$ are truly irrelevant/redundant features, especially in a large feature space. To overcome this problem, the proposed algorithm uses information ranks, named *counter score*, in the $LSFs$ selection to update each feature's relevance score. In this paper, the mutual information is chosen to compute the counter score. Mutual information was originally introduced in information theory. It is used for empirical estimates

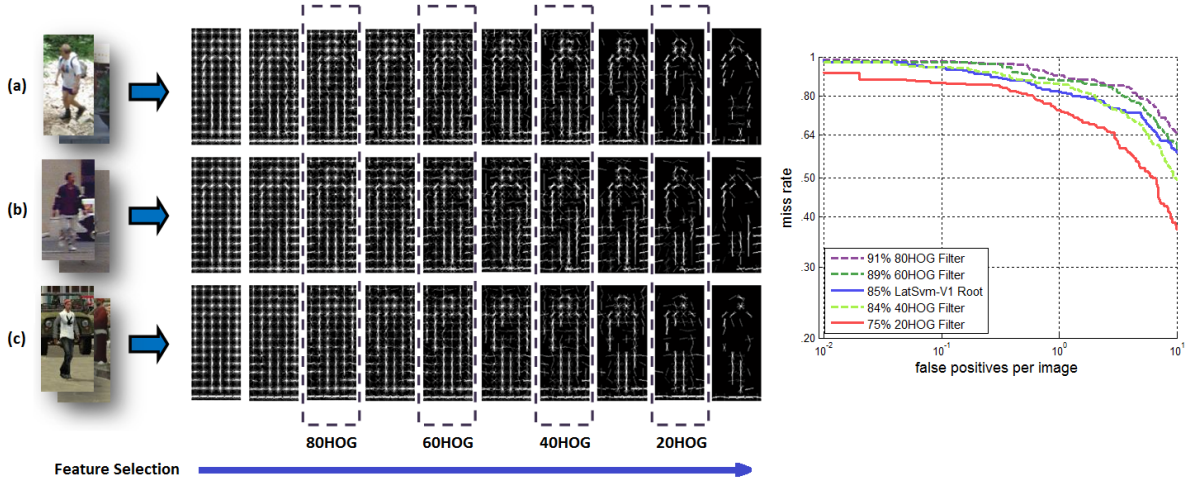

Figure 3: Appearance models trained using the proposed feature selection algorithm. (a) INRIA Dataset, (b) MIT Dataset and (c) CVC04 Dataset (Vazquez et al., 2014). The detection performances of the INRIA appearance models are evaluated on the first 100 INRIA full images. The appearance model from 20HOG shows 10% better performance than the root filter of LatSVM-V1 (Felzenszwalb et al., 2008).

between individual features and classes (Guyon and Elisseeff, 2003). Mutual information is derived from entropy. Entropy H is an uncertainty measure of event occurrence. Entropy of the discrete random variable X is described as $H(X) = -\sum_{x \in X} p(x) \log p(x)$, where $p(x)$ denotes the probability density of an event $x \in X$. The entropy of variable X can be conditioned on variable Y as $H(X|Y)$. If variable Y does not introduce any information which influences the uncertainty of X, in other word, X and Y are statistically independent, the conditional entropy is maximised (Vergara and Esteves, 2014). From this description, mutual information $IG(X;Y)$ can be derived as $H(X) - H(X|Y)$. Mutual information represents the amount of information mutually shared between variable X and Y. This definition is useful within the context of feature selection because it gives a way to quantify the relevance of a feature with respect to the class (Vergara and Esteves, 2014). Therefore, using mutual information in the wrapper approach benefits both the optimal feature space search and the selection performance enhancement. The proposed algorithm uses the mutual information to compute the counter score of each feature. The counter score I_f and its contribution to the feature relevance score are calculated as follows:

$$R_f = R_f + \alpha I_f, \quad I_f = IG_f / IG_{Max} \qquad (4)$$

where $\alpha = (R_{Max} \times FN_{Remain})/FN_{Full}$ is the counter score contribution rate, IG_f is the mutual information of feature elements. The rate is dynamically decremented as more steps are processed $\frac{FN_{Remain}}{FN_{Full}}$, which means the counter score contributes more in the large $LSFs$ subtraction. Fig.2 (a) shows that the counter score improves the performance of the proposed wrapper feature selection algorithm in terms of local prediction accuracy.

Feature Subtraction: The number of features to be removed at each step is chosen as $(5/N) \times 100$, where N is the number of the remaining features. In comparison to the original SBE algorithm, which removes only the least significant feature at a time, it is reasonable to remove more than one feature at a time, as only a small portion of features are highly relevant in many applications. In a human detection system, the most relevant features can be viewed as the features which are centred on the human contour. This can be visually demonstrated in Fig.3. When a step is completed (L_c has been satisfied), the algorithm subtracts a group of the least significant features, $m \times LSFs$, which have the lowest relevance score R_f.

3.2 Appearance Model

In a human detection system, the optimal feature elements tend to represent the human contour as illustrated in Fig.3. This was also pointed out in (Dalal and Triggs, 2006) as the most important cells are the ones that typically contain major human contours. In other words, the optimal feature elements are useful to build an appearance model. To evaluate the discriminative power of the selected feature elements, a simple appearance model is created. The model consists of a positive filter and a negative filter. The positive filter is formed with the average HOG of the selected feature elements from the positive examples.

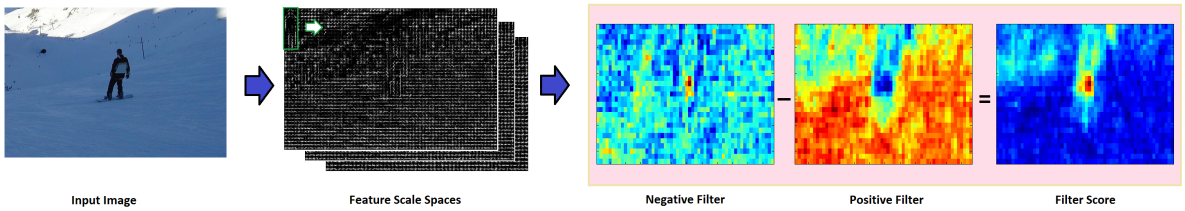

Figure 4: Filter Score Computation: The positive filter score is subtracted from the negative filter score.

The negative filter is generated with the negative examples in the same manner. The HOG scheme allows us to divide the appearance of an object into geometrical units called cells. A cell is then represented with n angles and their weighted magnitudes. The HOG feature of an example is an one-dimensional array of magnitudes whose indexes indicate the angle and the location of the cell. The proposed feature selection algorithm outputs the array indexes of an optimal feature subset. Each element f_i of a filter is computed as follows:

$$f_i = \frac{1}{n}\sum_{j=1}^{n} \beta_{ji} \tag{5}$$

where f_i is the array element of the proposed filter, β is the vector of an example of dataset. The score of region S_r is the regional similarity and computed with the euclidean distances between the ROI and the filters as shown in the following Equation:

$$S_r = d(N_s, O_s) - d(P_s, O_s) \tag{6}$$

where N_s and P_s are the optimal sub-vectors of the negative and positive filters, O_s is the sub-vector of ROI feature vector, and $d(x,y)$ is the euclidean distance.

4 EXPERIMENTAL WORK

Dollár (Dollár et al., 2012) (Dollár et al., 2009) provided 288 INRIA pedestrian full images for the benchmarking purpose. The majority of tests in this paper are carried out against 288 full images. However, some of the tests are conducted on the first 100 images, which show similar results as for the case of 288 full images. To evaluate the generalisation of the proposed approach, the test is also performed on the ETH Bahnhof sequence (Ess et al., 2008), which contains 999 street scenes. The performance of the human detection systems in terms of the detection accuracy are evaluated using the PASCAL VOC criteria (Everingham et al., 2007). All the algorithms and systems in the experiments are realised using Matlab 2014 with the relevant mex files. The test computer is of 2.49GHz Intel i5-4300U CPU running with Window 8.

4.1 Feature Vectors

Feature Vector: The feature used in the experiments is the Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2006). First of all, a linear SVM classifier is trained using the MIT pedestrian dataset. The MIT dataset consists of up-straight person images which have less dynamic poses. The positive examples shortlisted from other datasets using this classifier tends to include rather static pose examples. The dataset consisting of static pose examples is used to build an appearance model. Secondly, a subset of the INRIA dataset is selected by the classifier. The INRIA dataset offers cropped 2416 positive examples, and also allows to generate 12180 negative examples in the random selection manner for the training purpose. The classifier selects 1370 positive and 1579 negative examples from the training dataset. The negative examples include 79 false positive examples called "hard negative example". Thirdly, the feature extraction algorithm introduced in (Felzenszwalb et al., 2008) is used to compute HOG descriptors for the experiments with LatSvm-V1 (Feature Vector A). The algorithm in (Felzenszwalb et al., 2010) is also used for the tests with LatSvm-V2 (Feature Vector B).

Feature Selection: The proposed feature selection algorithm is applied to the extracted feature vectors shown above. From Feature Vector A, the algorithm selects four optimal feature subsets which have 80%, 60%, 40% and 20% elements of the full feature vector. The detection system trained with these feature subsets are referred to as 80HOG, 60HOG, 40HOG, and 20HOG respectively. 100HOG represents the system trained with the full feature vector. The algorithm also selects 20%, 15% and 10% elements from Feature Vector B. They are referred to as 20HOG-V2, 15HOG-V2 and 10HOG-V2 respectively.

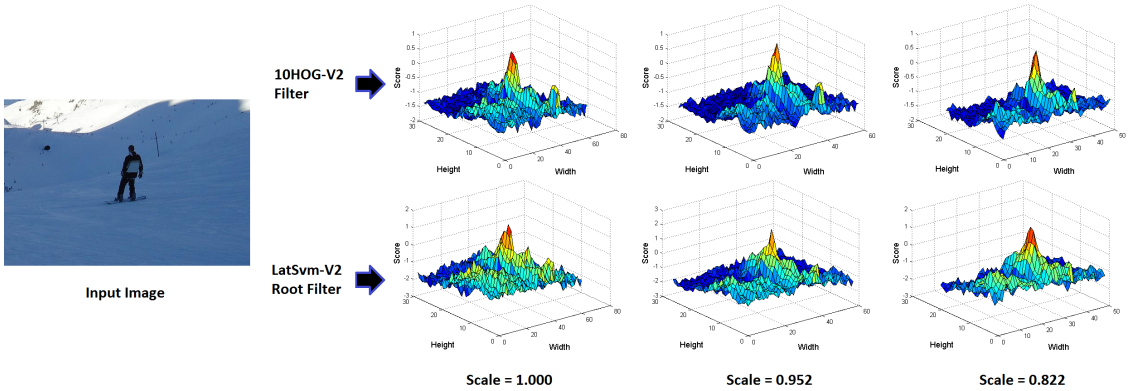

Figure 5: Localisation accuracy: The human localisation accuracy of the 10HOG-V2 Filter is compared to that of the root filter of the LatSvm-V2 Model (Felzenszwalb et al., 2010). Top - Score maps of the 10HOG-V2 filter at three scales. Bottom - Score maps from the LatSvm-V2 root filter at the identical scale. The score map from the 10HOG-V2 filter shows less noise than the one of the LatSvm-V2 root filter.

4.2 Full Image Results

Feature Vector A: To evaluate the feature selection performance, a simple human detection system using the HOG and the linear SVM (Dalal and Triggs, 2006) is created. Fig.2 (b) shows the detection accuracy of the systems trained with the selected feature subsets from Feature Vector A. The 40HOG, 60HOG and 80HOG slightly improve the accuracy up to 2% compared to the system trained with the full feature set, 100HOG. The detection accuracy is improved until the 40HOG is applied, which uses less than half of the full feature vector dimension. The 20HOG shows no improvement in the detection performance even though 20% feature vector has the best score in the local classification score curve as shown in Fig.2 (a). The results of the evaluation reveal that as the feature selection progresses the proposed algorithm gradually introduces the overfitting problem. The window sliding speed is significantly improved. The 100HOG takes 478.634s to scan 1060×605 pixels image. Compared to this, the 20HOG takes only one tenth of the search time required by the original system completing the same search within 45.895s.

Feature Vector B: The proposed appearance model is trained with the 20HOG-V2, 15HOG-V2 and 10HOG-V2. Fig.6(a) shows the evaluation results of the appearance models. The evaluation is carried out on the 288 INRIA Full Images, and compared to the LatSvm-V2 Root filter (Felzenszwalb et al., 2010) and the appearance models trained with the other feature selection algorithms. Unlike the classifier dependent systems, the appearance model shows better performance as the feature vector is better optimised, and no overfitting problem is observed with the appearance models. The 10HOG-V2 model outperforms the LatSvm-V2 Root filter scoring with a 9% better detection rate. To test the generalisation performance of the proposed approach, filters trained with the INRIA dataset are directly applied in the experiments on the ETH Bahnhof sequence (Ess et al., 2008), which consists of 999 street scenes. The 10HOG-V2 outperforms all the other filters achieving 2% better performance compared to the LatSvm-V2 root filter as shown in Fig.6 (b). The root filter of the LatSvm approach (Felzenszwalb et al., 2008)(Felzenszwalb et al., 2010) is equivalent to the model from *Dalal*'s original HOG approach (Dalal and Triggs, 2006). Therefore, the proposed appearance model can simply replace the LatSvm root filter. Fig.6 (c) illustrates the detection performance evaluation of the Deformable Part Model, LatSvm+10HOG-V2, which has the proposed appearance model as a root filter. The evaluation on the 288 INRIA Full Images shows that the improved performance of the 10HOG-V2 filter is slightly worse than the LatSvm+10HOG-V2. The performance decrease can be explained in two-folds: Firstly, the part filters of the Deformable Part Model contributes more than the root filter does. On the first 100 INRIA Full Images, the part filters scores a 44% detection rate, while the root filter achieves 91%. Secondly, there are many supervised parameters involved in the Deformable Part Model, and the supervised parameters appear to affect the performance of the LatSvm+10HOG-V2. Therefore, the further optimisation is required to make the proposed filter fit with the Deformable Part Model.

5 CONCLUSION

We have presented a feature selection algorithm which can generate the optimal feature subset. We

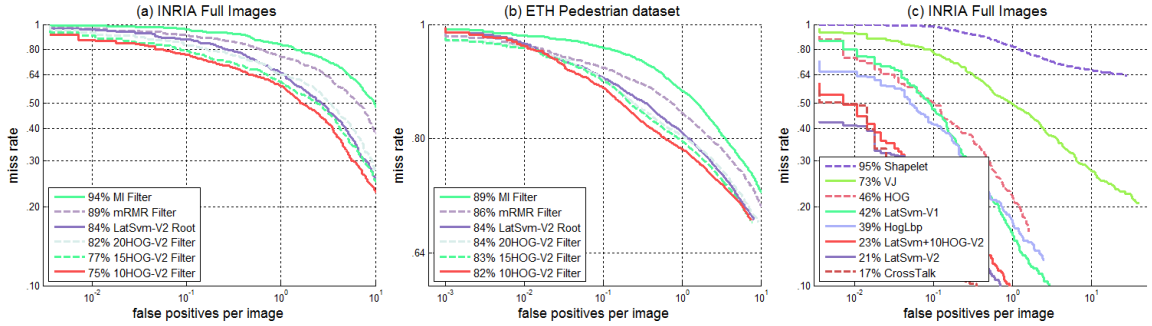

Figure 6: Performance comparison: (a) Filter detection performance comparison on the 288 INRIA Full Images. (b) Filter detection performance on the ETH Bahnhof sequence, (c) Comparison to the state-of-the-arts on the 288 INRIA Full Images.

have demonstrated the chosen feature subset can be used to improve the human detection system, which relies on the classifier performance, in both speed and accuracy. It has also been shown that the optimal features represent the object shape. Base on this observation, we have demonstrated that the optimal feature vector can be directly used to form the appearance models. This approach does not require highly accurate annotation data of objects to generate models. Therefore, it can be easily applied to a wide range of datasets.

REFERENCES

Alpaydin, E. (2004). In *Introduction to Machine Learning*. The MIT Press.

Battiti, R. (1994). In *Using Mutual Information for Selecting Features in Supervised Neural Net Learning*. IEEE Transactions on Neural Networks, Vol. 5, No. 4.

Bermejo, P., Gamez, J. A., and Puerta, J. M. (2011). In *Improving Incremental Wrapper-based Subset Selection via Replacement and Early Stopping*. International Journal of Pattern Recognition and Artificial Intelligence. Vol. 25.

Chandrashekar, G. and Sahinn, F. (2014). In *A survey on feature selection methods*. Journal of Computers and Electrical Engineering 40, P. 16-28.

Dalal, N. and Triggs, B. (2006). In *Histograms of Oriented Gradients for Human Detection*. IEEE Conference on Computer Vision and Pattern Recognition.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2009). In *Pedestrian Detection: A benchmark*. IEEE Conference on Computer Vision and Pattern Recognition.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012). In *Pedestrian Detection: An Evaluation of the State of the Art*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Ess, A., Leibe, B., Schindler, K., , and van Gool, L. (2008). In *A Mobile Vision System for Robust Multi-Person Tracking*. IEEE Conference on Computer Vision and Pattern Recognition.

Estevez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. (2009). In *Normalized Mutual Information Feature Selection*. IEEE Transactions on Neural Networks, Vol. 20, No.2.

Everingham, M., Zisserman, A., Williams, C., and Gool, L. (2007). In *The PASCAL visual obiect classes challenge 2007 results*. Technical Report, PASCAL challenge 2007.

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). In *Object Detection with Discriminatively Trained Part Based Models*. IEEE Conference on Computer Vision and Pattern Recognition.

Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). In *A Discriminatively Trained, Multiscale, Deformable Part Model*. IEEE Conference on Computer Vision and Pattern Recognition.

Foithong, S., Pinngern, O., and Attachoo, B. (2012). In *Feature Subset Selection Wrapper based on Mutual Information and Rough sets*. Journal of Expert Systems with Applications 39, P.574-584, Elsevier.

Gutlein, M., Frank, E., Hall, M., and Karwath, A. (2009). In *Large-scale attribute selection using wrappers*. IEEE Symposium Series on Computational Intelligence and Data Mining.

Guyon, I. and Elisseeff, A. (2003). In *An Introduction to Variable and Feature Selection*. Journal of Machine Learning Research 3, 1157-1182.

Heng, C. K., Yokomitsu, S., Matsumoto, Y., and Tmura, H. (2012). In *Shrink Boost for Selecting Multi-LBP Histogram Features in Object Detection*. IEEE Conference on Computer Vision and Pattern Recognition.

Javed, K., Babri, H. A., and Saeed, M. (2012). In *Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data*. IEEE Transactions on Knowledge and Data Engineering, Vol. 24, P. 465-477.

Kohavi, R. and John, G. H. (1997). In *Wrappers for feature subset selection*. Artificial Intelligence.

Li, K. and Peng, J. (2007). In *Neural Input Selection - A fast model-based approach*. Journal of Neurocomputing, Vol. 70, P. 762-769.

Peng, H., Long, F., , and Ding, C. (2005). In *Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy*.

IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8.

Pohjalainen, J., Rasanen, O., and Kadioglu, S. (2013). In *Feature Selection methods and Their combinations in High-dimensional Classification of Speaker Likability, Intelligibility and Personality Traits*. Journal of Computer Speech and Language. Elsevier.

Ruiz, R., Riquelme, J., and Aguilar-Ruiz, J. S. (2006). In *Incremental wrapper-based gene selection from microarray data for cancer classification*. Pattern Recognition, Vol. 39.

Vazquez, D., Marin, J., Lopez, A., Ponsa, D., and Geronimo, D. (2014). In *Virtual and Real World Adaptation for Pedestrian Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Vergara, J. R. and Esteves, P. A. (2014). In *A review of Feature Selection method based on Mutual Information*. Journal of Neural Computing and Applications 24, 175-186.

Yang, Y., Nie, F., Xu, D., Luo, J., Zhuang, Y., and Pan, Y. (2012). In *A Multimedia Retrieval Framework Based on Semi-Supervised Ranking and Relevance Feedback*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 4, P. 723-742.

Zhou, H., Miller, P., and Zhang, J. (2011). In *Age classification using Radon transform and entropy based scaling SVM*. British Machine Vision Conference.