Gender Classification in Low-Resolution Surveillance Video: In-depth Comparison of Random Forests and SVMs

Christopher D. Geelen[a], Rob G.J. Wijnhoven[b], Gijs Dubbelman[c] and Peter H.N. de With[d]

[ab]ViNotion BV, Horsten 1, Eindhoven, The Netherlands;
[cd]Eindhoven University Of Technology, Den Dolech 2, Eindhoven, The Netherlands

ABSTRACT

This research considers gender classification in surveillance environments, typically involving low-resolution images and a large amount of viewpoint variations and occlusions. Gender classification is inherently difficult due to the large intraclass variation and interclass correlation. We have developed a gender classification system, which is successfully evaluated on two novel datasets, which realistically consider the above conditions, typical for surveillance. The system reaches a mean accuracy of up to 90% and approaches our human baseline of 92.6%, proving a high-quality gender classification system. We also present an in-depth discussion of the fundamental differences of SVM and RF classifiers. We conclude that balancing the degree of randomization in any classifier is required for the highest classification accuracy. For our problem, an RF-SVM hybrid classifier with the combination of HSV and LBP features results in the highest classification accuracy of 89.9±0.2%, while classification computation time is negligible compared to the detection time of pedestrians.

1. INTRODUCTION

Video surveillance is an important topic in modern-day society. In the last decade, the number of installed video surveillance cameras has reached the point where the vast majority cannot be manually monitored anymore by security personnel. This results in an increasing demand for automatic and intelligent video content analysis systems. Such a system enables more efficient monitoring, by only presenting interesting footage to the security personnel. This is achieved by characterizing objects by their attributes, such as gender, age and clothing.

Gender is semantically a very interesting characteristic and one of the first mentioned attributes when describing people. Gender classification can be applied to several domains, such as in surveillance for efficient search and retrieval of persons from video footage. Also, gender information is very useful for marketing purposes, where advertisements and offers can be automatically adapted to the audience.

The aim of this research is to develop a gender classification system. First, people are detected in a video stream using an existing pedestrian detection system, similar to the systems described in the survey of Dollar *et al.*[1] Next, gender classification is performed on the detected persons, using the methods developed in this research. The process from video capture to classification is shown in Figure 1. In this paper, we specifically focus on the classification stage.

We develop our research along two conceptual lines. First, we evaluate several design choices that influence the performance of our gender classification system for surveillance environments. To this end, different features (HOG, LBP and HSV) are extracted from still images and compared, using accuracy and computational complexity. Particularly interesting is the combination of several features. These features will be used in two classification methods: Support Vector Machines (SVMs), which is the current baseline in classification, and Random Forests (RFs), which can describe complex decision boundaries, while being computationally efficient. The second research question considers the fundamental differences between these two classification methods. We evaluate and discuss the two classifiers and show that each classifier has a different field of operation.

Further author information: (Send correspondence to Rob G.J. Wijnhoven)
Rob G.J. Wijnhoven: E-mail: rob.wijnhoven@vinotion.nl, Telephone: +31 40 2366761
Christopher D. Geelen: E-mail: christopher.geelen@vinotion.nl, Telephone: +31 40 2366761

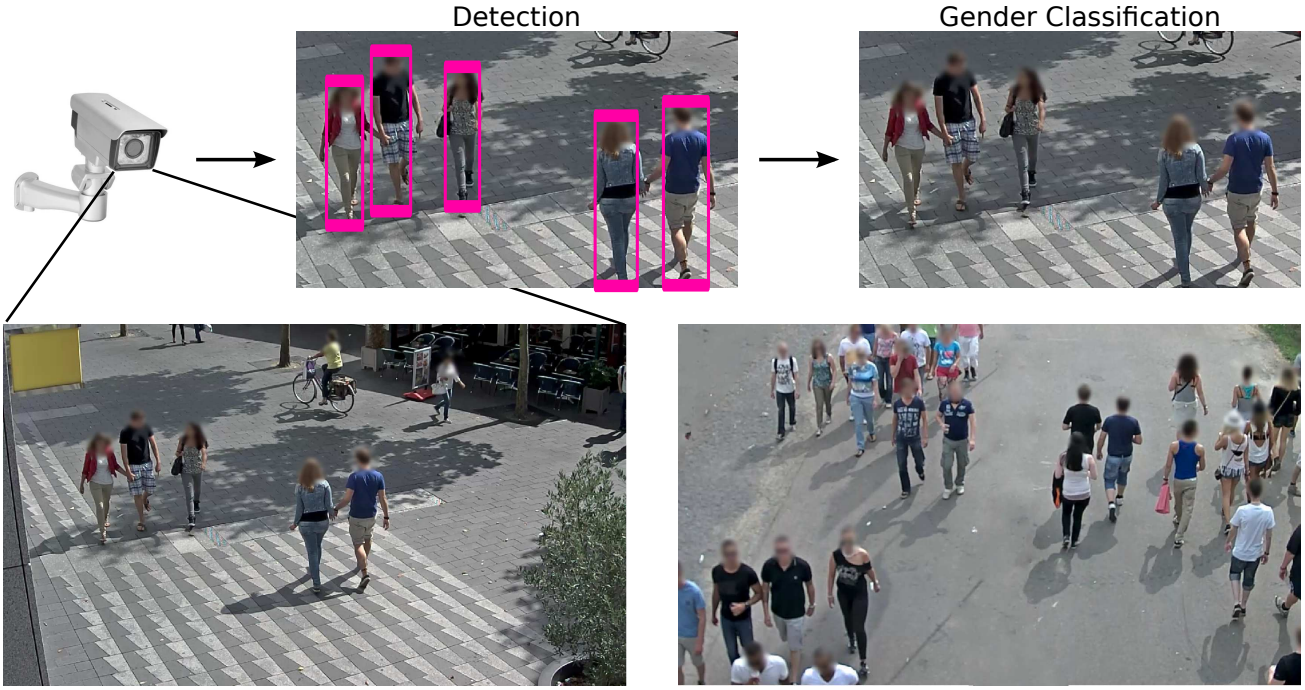

Figure 1. Overview of the three stages in the system process as shown at the top, with a focus on gender classification. Video footage used for dataset construction is shown at the bottom. Faces are blurred for privacy reasons.

The remainder of the paper is organized as follows. Our gender classification system is described in Section 2, including the different feature descriptors and classifiers. Section 3 describes the used datasets and the experimental evaluations. The main experiment is the comparison between the Support Vector Machine (SVM) classifier and Random Forest (RF) classifier. To address the second research question, Section 4 contains an in-depth discussion of the fundamental differences between SVM and RF, followed by the conclusions in Section 5.

1.1 Related Work

Three distinctive fields in gender classification can be distinguished in literature: classification using human faces, gait and full body. A complete overview on gender classification is presented in the paper of Ng *et al.*[2]

Face classification was addressed in 1990 by Golomb *et al.* , who presented SEXNET,[3] the first work performing gender classification. An alternative is based on analysis of the gait. Gait in gender classification is considered as the coordinated, cyclic combination of movements that result in human walking. An interesting example is proposed by Lu *et al.*,[4] which tackles the problem of people changing walking direction during the periodic cycle.

Face and gait classification both reach very high accuracies of up to 98%. Face classification only results in high accuracy when applied on high-resolution, frontal, up-close face images. Also, gait requires a clean capture of a full-body periodic movement. Both are very difficult to retrieve in surveillance footage. Therefore, full body classification is addressed next.

In 2008, Cao *et al.*[5] were the first to propose full-body gender recognition using low-resolution images. They developed a Patch-Based Gender Recognition (PBGR) approach using the Histogram of Oriented Gradients (HOG) feature descriptor.[6] They obtained 75.0% overall classification accuracy* on the MIT CBCL dataset,[7] originally created for pedestrian detection but annotated by Cao *et al.* for gender classification. Collins *et al.*[8] continued on this work and developed a classification system which combined HOG and Hue-Saturation-Value (HSV) features, reaching 76.0% on the MIT dataset (frontal only). Guo *et al.*[9] performed gender classification

*In literature, overall accuracy is used. However, this accuracy is biased towards the class with the highest number of training samples. Therefore, throughout this paper we use the mean accuracy: the average of male and female accuracy.

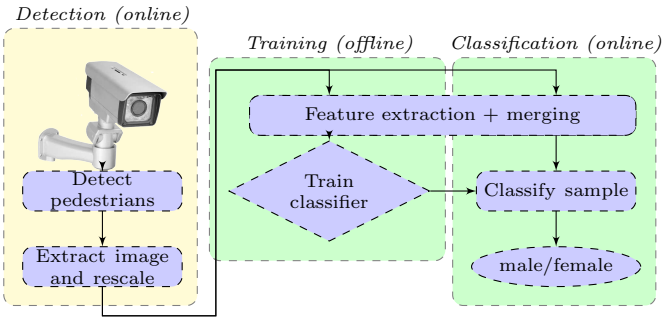

Figure 2. Current framework of the system implementation. This research assumes prior detection information, highlighted in yellow. Our research develops the training and classification stages of the actual person image, colored with green.

using a different approach, by utilizing Biologically-Inspired Features (BIF) and manifold learning. Their results are the current state-of-the-art, obtaining 80.6% on the MIT dataset. Finally, Bourdev *et al.*[10] also provide extensive work on gender classification on their own dataset, using poselets to implicitly decompose the viewpoints and poses. However, detailed annotations of many keypoints of the human body are required for each image.

Currently, common feature descriptors are HOG, HSV and Local Binary Patterns (LBP).[11] BIF is used by the current state-of-the-art, but preliminary experiments have shown that we could not reproduce these results. Preliminary experiments have also been performed using Transformed Color (TC), because TC is more color invariant than HSV.[12] However, we have experimentally found that this does not contribute to a higher accuracy. Therefore, the BIF and TC feature descriptors are not used in the remainder of our research.

After feature extraction, many different classifiers can be chosen. The baseline in classification has become Support Vector Machines (SVMs)[13] and is also used by Collins *et al.*,[8] Guo *et al.*[9] and Bourdev *et al.*[10] Linear SVMs provide a simple model and result in good performance in complex problems.[14] SVM can be used as a linear classifier, although it is also regularly used in combination with non-linear kernels, to exploit non-linear relations between feature dimensions. Collins *et al.* investigated several kernels and concluded that a linear kernel SVM performed best for full-body gender classification.

Random Forests (RFs)[15] have gained significant attention in the last years, because they can describe complex decision boundaries, while being computationally efficient.[16] RFs can also be used to simultaneously provide additional information, such as class probabilities and image clustering.[17] Cao *et al.*[5] evaluated the use of AdaBoost and RF, after which they combined the two principles and introduced PBGR. Verikas *et al.*[18] provides an extensive literature survey about articles comparing both SVM and RF. They state that although correlations are low, RF seems to perform better when there are strong single feature dimensions, and when there is a high number of samples per feature dimension. Do *et al.*[19] evaluated RF with high-dimensional data and proposed to combine SVM and RF. They show that this results in an increased classification accuracy. The main conclusion of Verikas *et al.* is that the choice of the classifier is application-specific.

2. APPROACH

The goal of the developed system is to analyze pedestrian images and label them as male or female. Rectangular bounding boxes are received from the preceding pedestrian detection system. These bounding boxes determine the locations of the pedestrians in each captured video frame. The developed system assumes perfect detection.

The framework is shown in Figure 2. First, the pedestrian image will be preprocessed by rescaling to a fixed size. Second, the pedestrian image will be transformed to extract information, e.g. texture. All information is then merged in a feature vector, and several feature vectors can be concatenated together. Third, a classifier is trained that uses these feature vectors to discriminate between male and female samples. Last, the classifier outputs a score, where a positive score labels the image as male and a negative score labels the image as female.

2.1 Feature Descriptors

Three different features are chosen: Shape, color and texture. To extract these features, the following feature descriptors are discussed and evaluated: Histogram of Oriented Gradients (HOG) as a shape descriptor, the Hue-Saturation-Value (HSV) color descriptor, and a novel feature descriptor to the field of gender classification, Local Binary Patterns (LBP) as a texture descriptor.

Histogram of Oriented Gradients. We have chosen Histogram of Oriented Gradients (HOG) as the shape descriptor. HOG is introduced by Dalal and Triggs[6] and used for gender classification by Cao *et al.*,[5] Collins *et al.*[8] and Bourdev *et al.*[10] The image is divided in square cells and per cell, image gradients are calculated by filtering with $[1, 0, -1]$ filters in both spatial dimensions. These gradients are then sorted into orientation bins, weighted by their magnitude and interpolated between bins and between cells. The gradient orientation sign can be used to further subdivide the bins. Lastly, all cell histograms are normalized, concatenated and stored as a feature vector.

Local Binary Patterns. First introduced by Ojala *et al.*,[11] LBP is a currently popular texture descriptor and shows good results for example in human detection.[20] However, to our best knowledge, LBP has not been applied to the case of full-body gender classification. Per cell in the image, a square sliding window is used as seen in Figure 3. Per window, each pixel is thresholded with respect to the center pixel, resulting in a bit. Then, a binary number is constructed by concatenating each bit clockwise, starting from the left-center pixel. The binary number is considered uniform, when there are at most two transitions from zero to one or from one to zero. Non-uniform values are collected in a single histogram bin, uniform values are stored in their respective unique bin. These histograms are collected for each image cell and then concatenated, resulting in a high-dimensional feature vector. For further accuracy, LBP is calculated within a circular radius around the center pixel, using interpolated pixel values. The radius of the circle, and the number of points on the circle can be varied.

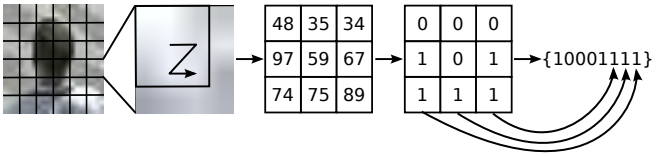

Figure 3. Schematic representation of the calculation of LBP features.

The number of LBP dimensions is much higher than the number of HOG dimensions. Therefore, also CS-LBP - a variation of LBP - is examined. Symmetrically-paired pixels are formed around the center, after which pixel pairs are thresholded with respect to each other. This reduces the number of dimensions with a factor of four.

Hue-Saturation-Value color system. Collins *et al.*[8] combined HOG features with a color descriptor, based on the HSV color system. They show that the addition of color improves the classification score. For a useful color descriptor, one needs to aim at photometric invariance,[12] as it results in independence of illumination variance or change in camera viewpoint, and thus in better overall descriptors. HSV is scale-invariant and shift-invariant with respect to light intensity.[21] Each pixel is sorted in hue bins and weighted by its saturation value (to eliminate hue instability around the gray axis[21]). Furthermore, each value is interpolated between bins and between cells. Each cell histogram is concatenated and then stored as the feature vector.

2.2 Classifiers

The main challenge of the gender classification system is to form a decision boundary, which perfectly separates the two classes (male/female). Although humans make distinct classifications of gender, the decision boundary is not well-defined, due to the large intraclass variation and interclass correlation. Features and attributes are not uniquely assignable to one of the two classes. For example, long hair is more stereotypical for females, but is also present among males. Besides these challenges, the classifier also has to tackle common problems, such as varying illumination and occlusions.

Two classifiers are being discussed in this work. Support Vector Machine (SVM) is a binary classifier that uses all feature dimensions to classify samples. Random Forest (RF) combines many weak classifiers (trees), of which each individual classifier is trained on a random subset of the total feature dimensions.

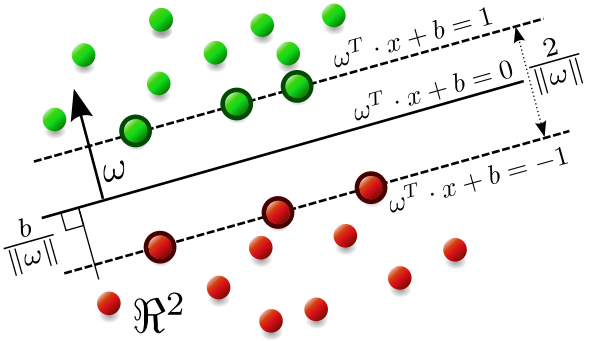

Figure 4. Schematic representation of an SVM. Red and green samples represent class zero and one respectively. Expression $||\boldsymbol{\omega}||^{-1}$ is the margin to each class. Samples with bold lining are support vectors. Image after.[14]

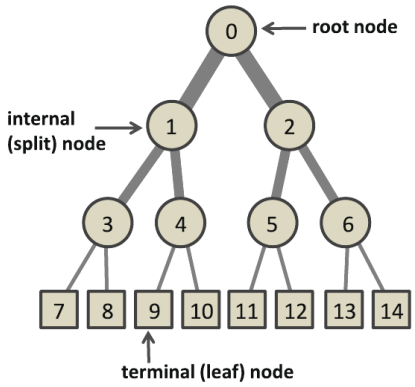

Figure 5. Schematic representation of a decision tree. All training samples enter at the root node and propagate through the tree. After traversing one or several split nodes, each sample ends in exactly one leaf node. There, the distribution of the classes is stored as the probability of each class. Image after.[22]

2.2.1 Classifier: Support Vector Machine

SVMs[13] are binary classifiers which assign new samples to one of the two categories (male/female). The two categories are separated by a decision boundary. This boundary is formed by the support vectors, the most informative samples for classification. SVM uses all feature dimensions to construct this boundary.

The model used for training is constructed by mapping the samples to a different feature space (typically high-dimensional). The mapping is done using kernels, which measure the similarity between samples. Instead of mapping the entire feature space, SVM uses a kernel-trick which implicitly maps samples using the kernel function $\mathbf{K}(\mathbf{x}, \mathbf{x_i})$, where $\mathbf{x}$ is a test sample and $\mathbf{x_i}$ is a support vector.

We now define the margin $||\boldsymbol{\omega}||^{-1}$ as the distance between the support vectors and the decision boundary. A large margin results in good classification. The margin is shown in Figure 4. The margin optimization for classification can be solved using quadratic programming. More details can be found in the paper of Cortes and Vapnik.[13]

Because the training data is generally not perfectly separable, the optimization process has to balance between misclassification (accuracy) and a large margin (generality). This results in a soft margin decision boundary. This optimization problem is solved by adding a regularization parameter C.[13] Different classification problems require different values of C.

The choices for kernels are numerous. We will only consider the (Gaussian) Radial Basis Function (RBF) kernel, which is represented by $K(\mathbf{x}, \mathbf{x_i}) = \exp(\gamma||\mathbf{x} - \mathbf{x_i}||_2^2)$. The size of the kernel, γ, requires careful evaluation. The superscript in the norm indicates a quadratic measure, and the subscript the norm basis. A small kernel will label each training sample as a support vector, resulting in a very complex decision boundary (over-training). Training with a large kernel will result in few distinctive samples (support vectors), leading to a coarser decision boundary. The size of the kernel γ, together with the regularization parameter C, are the main parameters that influence classification accuracy.

At test time, SVMs solve a kernel-based metric equation to determine the class label of the test sample. The resulting score determines the object class (if the metric is negative, the sample is classified as female, otherwise as male).

Linear classifier. We now consider a linear kernel (linear SVM), where the decision boundary is represented as a hyperplane in the original feature space. The SVM defines a hyperplane that has maximum distance to each support vector (max-margin). Figure 4 illustrates the case of a linear SVM, for clarity shown in the two-dimensional Euclidean space.

In our main experiments, we use a linear SVM. The kernel equation $\mathbf{K}(\mathbf{x}, \mathbf{x_i})$ now becomes a dot product, resulting in $y'(\mathbf{x}) = \boldsymbol{\omega}^T \cdot \mathbf{x} + b$.

2.2.2 Classifier: Random Forest

A Random Forest (RF) is an ensemble classification method. RFs are a combination of multiple decision trees, where each tree is constructed by evaluating a random subset of feature dimensions. In this research, we only consider binary trees, where each node is recursively split in two children, as shown in Figure 5. An introduction to RFs is given in the following. More details can be found in the extensive work by Criminisi and Shotton.[22]

During training, a collection of samples propagates through the tree. At each node, the training set is split in two new sets and each set is transferred to one of the two child nodes. This process is repeated until a stopping criterion is reached, such as the maximum node depth, or the minimum number of samples in a node. The split function maximizes the homogeneity after the split, thereby implicitly maximizing the homogeneity in the leaf nodes. A random subset of feature dimensions is tested at each split, after which the best single or subset of feature dimension(s) is used for splitting. Because of this randomization, RF requires strong single feature dimensions and a large number of training samples per dimension. Consequently, high-dimensional data is challenging. Also, bagging[15] is used, where the initial training set is randomly sampled with replacement.

During testing, each test sample traverses all decision trees in the forest. In each tree, the unknown sample reaches a single leaf node. The tree returns the class distribution of this leaf, normalized to the prior distribution of the training set of the tree. The final classification score is constructed by combining the distributions of the individual trees.

Forest structure. The structure of the forest can be changed by altering the maximum depth of each tree and the maximum number of trees in the forest. Increasing the depth of each tree results in stronger trees. However, the maximum depth of the tree should be set in relation to the problem complexity, as shallow trees cannot model the required complex boundary, whereas allowing too many splits can result in over-training. By increasing the number of trees, more feature dimension combinations are tested and the generalization of the forest increases. Adding more trees always increases classification accuracy, until it converges to the maximum achievable accuracy. However, computation time also increases linearly with the number of trees.

Randomization. The *degree of randomization* (ρ) is the main characteristic of the RF and needs to be carefully chosen. The degree of randomization is defined as the number of feature dimensions tested at each split. By growing ρ, the strength of each split increases, but the correlation between the trees also increases. When ρ is reduced, individual trees become weaker, but the correlation between the trees is lower. RF has the best performance when individual trees are strong and the correlation amongst trees is low, thus the key is to balance both, using the degree of randomization. Breiman[15] suggests to set the value of ρ as the square root of the total number of dimensions, but the survey of Verikas *et al.*[18] concludes that the choice of ρ is very dependent on the problem complexity.

Split function and hybrid RF forms. The amount of randomization in the RF is also influenced by the split function. The split function optimizes the homogeneity of the trainset, commonly performed by maximizing the Gini index or the Information Gain per split. The most common split functions distribute samples using a single feature dimension. More information can be injected into the tree by evaluating more dimensions in each node, for example by training an SVM on a subset of feature dimensions.[19] We will call this an RF-SVM hybrid classifier. The resulting SVM scores are used to maximize the Gini Index or Information Gain. Evaluating more dimensions benefits the analysis of high-dimensional feature descriptors, but also significantly increases the computation time. This RF-SVM hybrid classifier will be further explored in this paper.

3. EVALUATION

In the previous section, we have identified a number of important parameters and design choices. For each feature descriptor and classifier, we evaluate the effect of these parameters on the classification performance. Two performance criteria are used. First, the overall accuracy is specified by $\frac{\text{correct \# subjects}}{\text{total \# subjects}}$, commonly used in literature. However, this measurement is biased towards the gender with the highest number of images, which is undesirable. Therefore, unless stated otherwise, the reported performance is always the mean accuracy,

Table 1. Overview of the public MIT CBCL pedestrian dataset and our two novel datasets, A and B.

Dataset	Males	Females	Body part	Comment
MIT CBCL	600	288	Body	
Dataset A	4,269	2,994	Head	Street, Music event
Dataset B	1,120	1,170	Head, Body	Street

defined by $\frac{1}{2}(\frac{\text{correct \# males}}{\text{total \# males}} + \frac{\text{correct \# females}}{\text{total \# females}})$. Five-fold cross-validation is used in all our experiments[†]. For benchmarking the system performance, we have constructed two novel datasets, introduced in Section 3.1.

Experiments are performed to evaluate the system performance on multiple aspects. First we begin with establishing a baseline for classification accuracy by evaluating classification by human subjects. Second, several feature descriptor parameters are evaluated. Third, a linear and RBF kernel SVM are compared and the feature descriptors and their combinations are tested with a linear SVM. Fourth, several design choices of RF are discussed: the forest structure and the degree of randomization. Then the feature descriptors and their combinations are tested with an RF. Fifth, an RF-SVM hybrid classifier is discussed for classifying the high-dimensional feature vector. Sixth, a short validation on the MIT CBCL dataset is presented, using a linear SVM, to give a comparison with the state-of-the-art.

3.1 Datasets

Three datasets are considered in this research. First we describe the public MIT CBCL dataset and then our two novel datasets, A and B. An overview of the datasets is given in Table 1. Because there is no dedicated dataset for gender classification, Cao *et al.*[5] have annotated gender in the MIT CBCL dataset,[7] originally designed for pedestrian detection. Examples of this dataset are shown in Figure 6. However, this dataset is not suitable. First, the dataset only contains front and back facing people, whereas uncontrolled environments involve multiple viewpoints. Second, the number of training samples is too low when high-dimensional features are used for classification. Third, occlusions are not represented in this dataset. Therefore we conclude that this dataset is not representative for gender classification in uncontrolled environments.

Figure 6. Eight examples from the MIT CBCL dataset. Image size is 64×128 pixels. From left to right: two males from the back, two males from the front, two females from the back and two females from the front.

We introduce two novel datasets which are better suited for the targeted application. Footage is used from surveillance cameras, positioned at two different environments, as shown in Figure 1. The first environment is a shopping street, containing a lot of viewpoint variation. However, occlusions are rare and small. The second is a music event. The event is crowded, resulting in lots of occlusions. However, viewpoint variation is low.

The first novel dataset, Dataset A, is constructed to be a good representation of the application. The preceding pedestrian detection system identifies persons and returns the head-region of each person. This region is extended to include the head-shoulder region. From each randomly selected frame, each visible head is annotated, including all occurring occlusions and viewpoints. This results in 866 male and 990 female images obtained from the shopping street footage and 3,403 male and 2,004 female images obtained from the music event footage, resulting in a total of 4,269 male and 2,994 female images.

The second dataset, Dataset B, is constructed for semantically analyzing gender classification. Therefore, this dataset only contains images without occlusion and with full-body visibility. Each person is annotated twice,

[†]The train set is split up in five equally sized parts. The classifier is trained on four parts and validated on the remaining part. The process is repeated five times for different validation parts.

one head-shoulder annotation and one full-body annotation. All annotations are obtained from the shopping street footage. This results in a dataset of 1,120 male and 1,170 female images. Figure 7 presents average images of Dataset B, showing a clean full-body capture and no visible occlusions. The averaged image becomes blurry below the knee area, resulting from the many different legs and feet poses.

Figure 7. Average images of males ('M') and females ('F') from Dataset B. The images clearly show a clean capture and no visible occlusions. The images are shown for front view, back view and mixed view, from left to right, respectively.

Both datasets are constructed by manually annotating pedestrian images. The annotations of Dataset A are constructed by drawing a box around the head. The bounding boxes are converted to square boxes by equaling the width to the height of the bounding box. Lastly, the boxes are extended with 20% on each side and 40% to the bottom, resulting in a head-shoulder annotation. The images are extracted and rescaled to 24×24 pixels, composed of 6×6 cells of 4×4 pixels. The full-body images of Dataset B are constructed similarly, resulting in an image size of 24×72 pixels, divided in 6×18 cells of 4×4 pixels.

3.2 Human baseline

To provide a baseline of gender classification, a small experiment is performed to determine the human classification accuracy. Humans are not perfect in male/female recognition, but are regarded as optimal classifiers for uncontrolled environments, which include occlusions, illuminations, etc. The experiment is performed with 15 subjects. From Dataset B, 100 pedestrians are selected, which are displayed to the human subjects in three stages. First, the head-shoulder image, then the full-body image and last the full-view frame to also provide full context information. This results in 300 different classifications made by each human subject. The ground-truth is determined by analyzing the video stream, which contains significantly more information than the still (cropped) images. The result of this experiment is a 92.6±2.8% classification accuracy for head-shoulder images, 96.0±1.7% for full-body images and 97.6±1.2% for full-view images. Our system only focuses on the head-shoulder images, therefore the 92.6% classification accuracy can be regarded as the maximum achievable classification performance for our problem.

3.3 Individual feature descriptors

The most significant feature descriptor parameters are evaluated on classification accuracy and computational complexity, using a linear SVM classifier. The tests are performed on Dataset A, the set best resembling the aimed application. Normalization of the feature descriptors is fixed for all experiments. Each cell is normalized individually, where HOG is normalized using L2 normalization and LBP and HSV are normalized by a fixed factor, defined by the number of pixels in a cell. For the HSV descriptor, we evaluate the number of hue bins from 2 to 32 bins with linear bin spacing. For the HOG descriptor, the same amount of bins and spacing is used, and we evaluate the use of the orientation sign. For the LBP descriptor, we evaluate the number of points on the circle from 2 to 16 points, and evaluate using unity radius, radius 2 or both radii combined.

For HSV, the highest mean accuracy of 82.9% is obtained using 8 hue bins, resulting in 288 dimensions. The maximum mean accuracy obtained for HOG is 86.3%, using 16 bins while employing the orientation sign. However, 8 bins with the use of orientation sign are chosen by balancing performance and computational complexity. This also results in 288 dimensions. The maximum accuracy with LBP features is 87.7%, using 8 points and both radii. Balancing performance and complexity results in 8 points on a unity circle, with a mean accuracy of 86.5%. This results in 2,124 dimensions. These parameter settings will be used for the remaining experiments.

We also consider the constrained version of LBP, CS-LBP. This feature descriptor is tested using both an RF and an SVM, and is compared to LBP and the combinations with HOG and LBP. These experiments show that the low-dimensional features do not outweigh the reduction in accuracy, so that it will not be further considered.

3.4 Support Vector Machine (SVM)

We evaluate the SVM for different feature descriptors and their combinations on Dataset A. The learning rate λ, analogous to the regularization parameter C, is evaluated from 10^{-5} to 100 in multiples of 10. The following experiments are based on a learning rate $\lambda = 0.001$, $\eta_0 = 1$ and 10 epochs. The cost ratio is the factor used to increase the step size of the update with the current sample and is set to the ratio of male and female samples.

The results of the linear SVM classification of all feature descriptors and their combinations are shown in Figure 8. From these results we remark the following. First, adding HSV features to other features always increases the classification score, while individual features have a low accuracy. This indicates that color is a distinctive feature and adds information to edges or texture features. Second, the combination of HOG and LBP features scores lower than using only LBP features. Because HOG and LBP both are the same type of feature descriptor (describing object shape/texture), adding HOG does not add additional information, but only adds dimensions and is thus redundant. We conclude that the combination of LBP and HSV features results in the highest classification accuracy of 89.3±0.2%.

Next we evaluate the performance of a non-linear kernel SVM and its classification accuracy. A Radial Basis Function (RBF) kernel SVM is tested and the resulting classification accuracy is compared with the linear SVM performance. Two parameters of the RBF are evaluated: the size of the kernel γ and the regularization parameter C, which controls the balance between maximizing the margin and minimizing misclassifications. We evaluate γ between 10^{-4} and 10^2, and C between 10^{-2} and 10^2. Both are sampled in multiples of 10. Using an RBF kernel SVM results in 87.7±0.3%, a slightly worse performance than a linear SVM (confirming the result of Collins[8]).

The experiments show that the gender classification problem is well separable using a linear hyperplane. Modeling a non-linear decision boundary using an RBF-kernel SVM does not aid in the classification task for our experiments.

3.5 Random Forest (RF)

Random Forests offer an alternative method to model non-linear decision boundaries. We carefully evaluate the most important parameters: the forest structure (the maximum allowed depth and the maximum number of trees), the degree of randomization and the split function.

3.5.1 Forest structure

We evaluate the influence of the forest structure on the classification accuracy. We vary the maximum depth from 4 to 20 and the maximum allowed trees from 10 to 100. We evaluate the performance using LBP-HSV features, since this combination results in the highest accuracy with a linear SVM. The results are shown in Figure 10. The accuracy converges at a depth of 14 and a maximum of 50 trees, which are the default settings in the following experiments. It should be noted that the result stays constant when increasing the depth or maximum number of trees beyond the values depicted in the graph. Literature states that increasing the number of trees will increase the generalization accuracy until a maximum is reached, which is confirmed by our experiments. In contrast to literature, we have found that increasing the depth of the trees also converges. According to Criminisi and Shotton,[22] increasing the depth implies over-training, which does not seem to occur in our experiments. Results show that the average number of leafs in the RF is 245, resulting in an average amount of samples of 24 per leaf (of the 5,800 total training samples). Furthermore, the RF is trained with an average depth of 14. This indicates that the maximum depth is not reached, thus no over-training occurs.

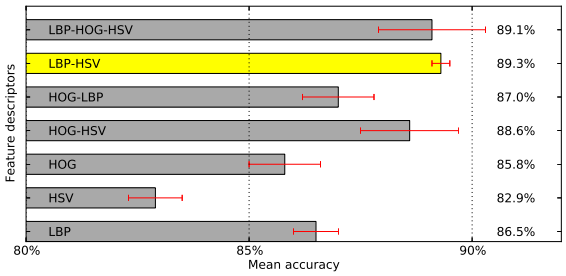

Figure 8. SVM classification using different feature descriptors and their combinations. The red error bars indicate the standard deviation.

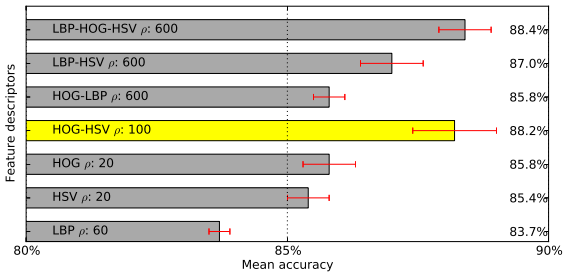

Figure 9. RF classification using different feature descriptors and their combinations. The red error bars indicate the standard deviation.

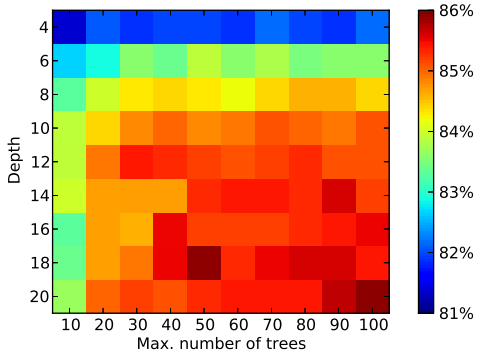

Figure 10. Evaluation of different tree structures for RF. Tests are performed on Dataset A using LBP-HSV features. Figure is best viewed in color.

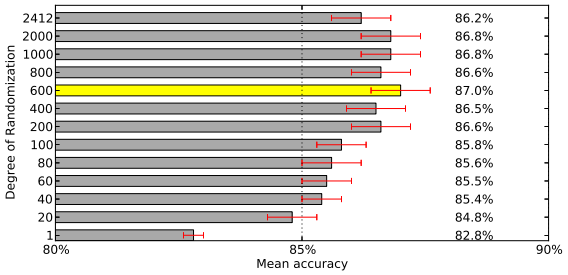

Figure 11. The degree of randomization for RF is varied by changing the number of feature dimensions tested at each split. A combination of HSV and LBP features is used. The red error bars indicate the standard deviation.

3.5.2 Degree of randomization

To test the required degree of randomization for our problem, we classify the combination of LBP and HSV features, varying the number of tested dimensions from 1 to 2,412. Results are presented in Figure 11. The highest classification accuracy is obtained using the square root of the total number of dimensions for individual feature descriptors, in line with literature.[15] However, we have found that a significantly higher number of dimensions (±500) is required when combining multiple feature descriptors. We conclude that when feature descriptors are combined, the number of dimensions required for optimal classification accuracy is much higher.

3.5.3 Feature descriptors and their combinations

We evaluate the RF for different feature descriptors and their combinations on Dataset A. Figure 9 presents the results. First, using HSV features results in a higher performance compared with an SVM classifier (shown in Figure 8). Individual HOG features result in the same accuracy, while other descriptors and their combinations result in a lower accuracy. Second, the LBP feature descriptor is too noisy, since adding LBP never increases classification accuracy. We conclude that LBP only has weak single dimensions and that the number of training samples is too low with respect to the high dimensionality of the LBP descriptor.

3.5.4 RF-SVM hybrid

The previous experiment questions how an RF can cope with high-dimensional features. The data is too rich with information to model the splitting in a binary decision function. As discussed earlier, therefore multiple feature dimensions can be exploited to provide more information during node splitting. This approach is implemented by training a linear SVM on a random subset of LBP and HSV features, so that each node will thus contain a separate SVM classifier. This leads to the concept of a hybrid RF-SVM classifier. Figure 12 shows the obtained results when training a linear SVM on a completely randomized subset of feature dimensions. Combining multiple features in the split function does indeed increase the performance and even outperforms a linear SVM, resulting in an accuracy of 89.9±0.2%. Examples of correctly classified images are shown in Figure 13.

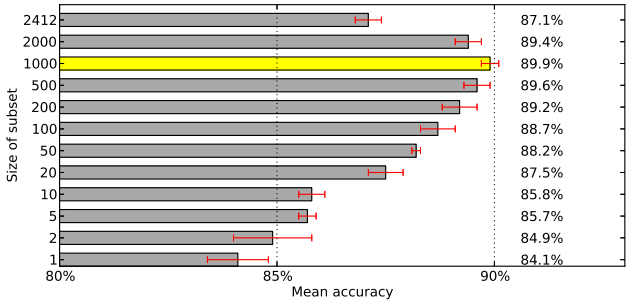

Figure 12. Results of a linear SVM, trained in each node on a random subset of LBP and HSV features. The size of the subset is evaluated. Tested with a depth of 14, a maximum of 50 trees, $\rho = 20$ for size 1-20, $\rho = 3$ for size 50 and higher.

Figure 13. Examples of correctly classified males (left) and females (right) under challenging conditions (e.g. face occluded by hands or background clutter). Images are classified using an RF-SVM hybrid classifier with LBP and HSV features. Figure is best viewed in color.

In additional experiments, we evaluate two different classifier architectures. First, we train an SVM on the training set, after which a separate RF is trained on both resulting subsets. Second, we train a linear SVM at each root node of all trees. Results show no increased accuracy compared to the previous experiments.

3.5.5 Computation time

Although training a linear SVM in each node of the RF results in the highest classification performance, it is also computationally more complex. A linear SVM has a computational complexity in the order of $\mathcal{O}(W)$, where W is the size of the weight vector dimensionality. A standard RF implementation has a complexity in the order of $\mathcal{O}(N \cdot D)$, where N is the number of trees and D is the average depth of the trees. When training a linear SVM in each node, this complexity increases to $\mathcal{O}(N \cdot D \cdot W)$. Comparing our best performing RF-SVM hybrid with $D = 14$, $N = 50$ and $W = 1,000$ to a linear SVM, with $W = 2,412$, using an RF is 290 times more complex.

However, the calculations have to be performed per pixel only, not per frame. For the typical size of surveillance camera frames (1,280×960 pixels), operating a state-of-the-art object detector[23] consumes 344 msec per frame. Our feature extraction of LBP and HSV features consumes 0.3 msec per object. Classification using an RF-SVM hybrid classifier consumes 1.4 msec per object, with $D = 14$, $N = 50$ and $W = 1000$. Assuming an average of 7 objects per frame (the average of our datasets), the total is 11.9 msec per frame for classification, which is 3% of the total time required for detection and classification. We conclude that the higher computation time of the RF-SVM hybrid is negligible and is thus attractive for gender classification.

3.6 Performance on MIT CBCL dataset

We compare our system using the MIT CBCL pedestrian dataset, the only public dataset. Because the dataset is not suited for our application (see Section 3.1), no extensive parameter evaluations have been performed. Since the number of training samples present in the MIT CBCL pedestrian dataset is limited, only a linear SVM classifier is used. Table 2 presents the results for different feature descriptors and their combinations. From these results, we conclude that we outperform the state-of-the-art results, obtaining an overall accuracy of 80.9±2.4% and a mean accuracy of 76.6±0.9% on the mixed-view set. Our proposal is only outperformed by the system based on BIF (Biologically-Inspired Features)[9] while using back-view images. However, after careful implementation of this feature descriptor, we could not reproduce the reported results from literature.

We conclude that our system obtains state-of-the-art performance on this dataset, and also obtains good performance on the two larger, more realistic datasets.

Table 2. Several feature descriptors and their combinations are compared with literature using the MIT CBCL dataset for frontal-, back- and mixed-view pictures. Features are classified using a linear SVM.

Feature Descriptor(s)	Mixed view		Frontal view		Back view	
	Overall accuracy	**Mean accuracy**	**Overall accuracy**	**Mean accuracy**	**Overall accuracy**	**Mean accuracy**
HOG	78.9±1.7%	**75.9±3.0%**	**81.2±1.7%**	76.9±1.6%	77.5±3.4%	72.6±4.3%
LBP	76.1±3.2%	68.5±4.1%	78.4±1.6%	73.5±1.7%	77.7±2.7%	71.9±3.3%
HSV	71.3±2.8%	64.8±3.3%	69.4±6.4%	65.0±3.1%	71.5±3.9%	63.5±0.8%
LBP+HSV	77.6±2.4%	73.7±2.4%	77.6±2.5%	73.9±3.4%	78.3±3.8%	72.7±4.5%
HOG+HSV	**80.9±2.4%**	75.3±3.6%	**81.6±1.6%**	**79.0±2.2%**	80.5±1.5%	74.1±2.9%
HOG+LBP	79.8±1.2%	**76.6±0.9%**	81.2±1.5%	76.6±1.2%	80.3±1.2%	75.5±1.6%
HOG+LBP+HSV	80.1±1.7%	76.7±1.7%	81.0±2.5%	73.9±3.3%	**82.7±1.8%**	**79.3±1.8%**
Cao 2008[5]	75.0±2.9%		76.0±1.2%		74.6±3.4%	
Collins 2009[8‡]			76.0±8.13%			
Guo 2010[9]	**79.2±1.4%**		**79.5±2.6%**		**84.0±3.9%**	

4. DISCUSSION

Let us now discuss the fundamental differences between linear SVM and Random Forest, resulting from the evaluations in Section 3. The main difference between a linear SVM and an RF is the amount of randomization inserted in the classifier. A linear SVM uses all feature dimensions and all training samples and contains no randomization. This enables the SVM to find a good generalization of the decision boundary, resulting in a good separation of the average male and female. However, the decision boundary of gender classification is inherently complex, due to the large interclass correlation. This is ignored by the generic nature of the SVM.

With RF, different amounts of randomization are injected through bagging, changing the split function, and the degree of randomization. Therefore, RF is able to construct a complex decision boundary, resulting in the ability to classify non-stereotypical males or females. However, it has difficulties in exploiting the relations between feature dimensions and over-trains on specific dimensions. By using an RF-SVM hybrid classifier, the strengths of both classifiers are combined. The relations between dimensions are modeled by SVM, and the RF structure allows the analysis of several randomized subspaces of the feature space, enabling the mapping of complexer boundaries.

However, experiments indicate that the number of training samples (i.e. 5,800 samples) is too limited to fully extract all relevant information from the feature space. First, the accuracy of RF converges when increasing the maximum depth of the trees (see Section 3.5.1), in contrast to results from literature. Second, using an RBF kernel in SVM also results in a low accuracy (see Section 3.4). Third, Figure 9 shows a decreased RF classification performance when using LBP features in comparison with a linear SVM. This is likely due to the high dimensionality of the LBP feature descriptor. These results indicate that not all information is extracted from the training samples. The number of samples is sufficient for a linear SVM to find a generalized decision boundary, but to construct the correct complex decision boundary with RF, more data (samples) about non-stereotypical pedestrians are needed.

5. CONCLUSION

This research addresses the problem of automatic gender classification in surveillance scenes, which is inherently difficult, due to the large interclass correlation. We have developed a gender classification system for surveillance environments and successfully evaluated the system on two realistic datasets. Compared to the public MIT CBCL dataset, the two introduced datasets contain eight times more samples and better resemble challenges of surveillance environments, such as a large amount of viewpoint variations and occlusions. Given the location of pedestrians in the image, the proposed classification system extracts features and classifies these as male or female. The system reaches a mean accuracy of up to 90% and approaches the human baseline of 92.6%, proving

‡Note that Collins *et al.* use a different annotation set.

a high-quality gender classification system. Furthermore, the system has a balanced male/female accuracy performance and outperforms the state-of-the-art on the public MIT CBCL dataset with an overall accuracy of 81%.

We have shown that a linear SVM outperforms an RBF-kernel SVM. When evaluating HOG, HSV and LBP feature decriptors, the best performance is obtained with a combination of HSV and LBP features. An RF classifier converges with a maximum depth of 14 and a maximum of 50 trees, with a single dimension split function. When classifying individual feature descriptors, setting the degree of randomization ρ to the square root of the maximum feature dimensions, results in the highest accuracy. However, combining descriptors requires a larger ρ. Comparing SVM and RF shows that RF outperforms SVM using HSV features, but cannot cope with the high-dimensional LBP descriptor. We conclude that RF requires strong single feature dimensions and/or a large number of training samples.

Using the SVM classifier results in an accuracy of 89.3±0.2%, outperforming the RF with an 87.9±0.6% accuracy. An RF-SVM hybrid classifier is proposed by training a linear SVM on a random subset of features, so that each node will thus contain a separate SVM classifier. This RF-SVM hybrid classifier results in an accuracy of 89.9±0.2%. Balancing the degree of randomization in any classifier is required for the highest classification accuracy. We conclude that for our gender classification problem, an RF-SVM hybrid classifier with the combination of HSV and LBP features results in the highest classification accuracy, while computation time is negligible compared to the detection time of pedestrians.

6. FUTURE WORK

To further address the problem of correctly classifying non-stereotypical pedestrians (see Figure 14 for examples of currently misclassified persons), we recommend to increase the total number of training samples. An interesting research topic will be to find the mathematical relationship between the number of feature dimensions, the amount of information per dimension, and the number of training samples required for the highest classification accuracy.

Figure 14. Examples of misclassified males (blue border) and females (red border). The images show several challenging situations, such as difficult poses (left), poor illumination (right) and an example of interclass correlation (female with short hair length). Figure is best viewed in color.

Three additions can be made when using the developed system in an industrial application. First, the RF can be extended to include multi-attribute classification, such as clothing, accessories or age. Second, the temporal domain can be exploited to cope with occlusions and unknown poses or viewpoints. Third, the system is designed for surveillance applications, where a high classification accuracy is required. However, marketing applications require the actual ratio of males and females to be correct, which should lead to a reconsideration of different trade-offs regarding performance.

REFERENCES

[1] Dollar, P., Wojek, C., Schiele, B., and Perona, P., "Pedestrian detection: An evaluation of the state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **34**(4), 743–761 (2012).

[2] Ng, C. B., Tay, Y. H., and Goi, B. M., "Vision-based human gender recognition: A survey," *arXiv preprint arXiv:1204.1611* (2012).

[3] Golomb, B. A., Lawrence, D. T., and Sejnowski, T. J., "Sexnet: A neural network identifies sex from human faces.," in [*NIPS*], 572–579 (1990).

[4] Lu, J., Wang, G., and Huang, T. S., "Gait-based gender classification in unconstrained environments," in [*Pattern Recognition (ICPR), 2012 21st International Conference on*], 3284–3287, IEEE (2012).

[5] Cao, L., Dikmen, M., Fu, Y., and Huang, T. S., "Gender recognition from body," in [*Proceedings of the 16th ACM international conference on Multimedia*], 725–728, ACM (2008).

[6] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*], **1**, 886–893, IEEE (2005).

[7] Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., and Poggio, T., "Pedestrian detection using wavelet templates," in [*Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*], 193–199, IEEE (1997).

[8] Collins, M., Zhang, J., Miller, P., and Wang, H., "Full body image feature representations for gender profiling," in [*Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*], 1235–1242, IEEE (2009).

[9] Guo, G., Mu, G., and Fu, Y., "Gender from body: A biologically-inspired approach with manifold learning," in [*Computer Vision–ACCV 2009*], 236–245, Springer (2010).

[10] Bourdev, L., Maji, S., and Malik, J., "Describing people: A poselet-based approach to attribute classification," in [*Computer Vision (ICCV), 2011 IEEE International Conference on*], 1543–1550, IEEE (2011).

[11] Ojala, T., Pietikainen, M., and Maenpaa, T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**(7), 971–987 (2002).

[12] Gevers, T., Gijsenij, A., Van de Weijer, J., and Geusebroek, J.-M., [*Color in computer vision: Fundamentals and applications*], vol. 24, The Wiley-IS&T (2012).

[13] Cortes, C. and Vapnik, V., "Support-vector networks," *Machine learning* **20**(3), 273–297 (1995).

[14] Wijnhoven, R. G. and de With, P. H., "Fast training of object detection using stochastic gradient descent," in [*Pattern Recognition (ICPR), 2010 20th International Conference on*], 424–427, IEEE (2010).

[15] Breiman, L., "Random forests," *Machine learning* **45**(1), 5–32 (2001).

[16] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM* **56**(1), 116–124 (2013).

[17] Shi, J. and Malik, J., "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(8), 888–905 (2000).

[18] Verikas, A., Gelzinis, A., and Bacauskiene, M., "Mining data with random forests: A survey and results of new tests," *Pattern Recognition* **44**(2), 330–349 (2011).

[19] Do, T.-N., Lenca, P., Lallich, S., and Pham, N.-K., "Classifying very-high-dimensional data with random forests of oblique decision trees," in [*Advances in knowledge discovery and management*], 39–55, Springer (2010).

[20] Mu, Y., Yan, S., Liu, Y., Huang, T., and Zhou, B., "Discriminative local binary patterns for human detection in personal album," in [*Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*], 1–8, IEEE (2008).

[21] Van De Sande, K. E., Gevers, T., and Snoek, C. G., "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **32**(9), 1582–1596 (2010).

[22] Criminisi, A. and Shotton, J., [*Decision forests for computer vision and medical image analysis*], Springer (2013).

[23] Felzenszwalb, P. F., Girshick, R. B., and McAllester, D., "Cascade object detection with deformable part models," in [*Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*], 2241–2248, IEEE (2010).