Optical Engineering

SPIEDigitalLibrary.org/oe

Three-level cascade of random forests for rapid human detection

Byoung Chul Ko
Deok-Yeon Kim
Ji-Hoon Jung
Jae-Yeal Nam

SPIE

Three-level cascade of random forests for rapid human detection

Byoung Chul Ko
Deok-Yeon Kim
Ji-Hoon Jung
Jae-Yeal Nam
Keimyung University
Department of Computer Engineering
1000 Shindang-dong Dalseo-gu, Daegu, 704-701
Republic of Korea
E-mail: niceko@kmu.ac.kr

Abstract. We propose a novel human detection approach that combines three types of center symmetric local binary patterns (CS-LBP) descriptors with a cascade of random forests (RFs). To detect human regions in a low-dimensional feature space, we first extract three types of CS-LBP features from the scanning window of a downsampled saliency texture map and two wavelet-transformed subimages. The extracted CS-LBP descriptors are applied to a three-level cascade of RFs, which combines a series of RF classifiers as a filter chain. The three-level cascade of RFs with CS-LBPs delivers rapid human detection with higher detection accuracy, as compared with combinations of other features and classifiers. The proposed algorithm is successfully applied to various human and nonhuman images from the INRIA dataset, and it performs better than other related algorithms. [DOI: 10.1117/1.OE.52.2.027204]

Subject terms: human detection; center symmetric local binary patterns; saliency texture map; random forest; wavelet transform; three-level cascade.

Paper 121330 received Sep. 18, 2012; revised manuscript received Nov. 18, 2012; accepted for publication Jan. 2, 2013; published online Feb. 4, 2013.

1 Introduction

Human detection in images and videos is an essential step in dynamic computer vision for many applications, including video surveillance, human action recognition, and content-based image/video retrieval. As a result, human detection and face detection has received widespread interest during the last decade, based on the visual features of humans and several pattern classifiers. However, human detection in images is a more challenging task than face detection because of the following problems.[1]

- The wide variability in appearance due to human clothing.
- Variations in the illumination of images due to day and night lighting, light reflections, and shadows.
- Wide range of human poses and partial occlusion.
- Cluttered background objects resembling a human body.

There are many methods for the automatic detection of humans in still images, which can be classified into two main themes based on the types of features and decision classifiers used for human verification. Table 1 summarizes the representative categories of human detection algorithms based on visual features and their classification algorithms.

Many researchers have tried to extract efficient and accurate features that are suitable for human detection. Papageorgiou and Poggio[2] used Haar wavelets as input descriptors. This method is invariant to changes in color and texture and has been used to robustly define a rich and complex class of objects, including people. Viola et al.[3] integrated image intensity information with motion using Haar-like wavelets and applied this method to human movement detection. Lowe[4] used scale-invariant feature transform (SIFT) descriptors to describe local features in images. SIFT features are local and based on the appearance of an object at particular points of interest, which means they are invariant to image scale and rotation. Dalal and Triggs[5] used the locally normalized histograms of oriented gradient (HOG) descriptors for human detection. A dense overlapping HOG grid provided good results for human detection, and it had a lower false positive rate compared with Haar wavelet-based descriptors. HOG is the most popular feature used for human detection, but its heavy computation demand is the one of its drawbacks. Maji et al.[6] proposed multilevel histograms of oriented edge energy and showed that this method yields better classification performance than the original HOG. This feature computes the oriented edge energy responses in eight directions using the magnitude of the odd elongated oriented filters at a fine scale ($\sigma = 1$), with nonmax suppression performed independently in each orientation. Chen and Chen[7] used a combination of intensity-based rectangle features and gradient-based features. Local binary pattern (LBP) descriptors were used as gradient-based features to detect humans in still images because they are invariant to monotonic gray level changes, and they are computationally efficient. This method has been used widely in various applications, especially facial recognition. Zhang et al.[8] used LBP as a texture feature with color information. In this method, the image was divided into M small nonoverlapping cells before the LBP histograms were extracted from each cell and concatenated into a single, spatially enhanced feature vector. Wang et al.[9] combined HOG with cell-structured LBP as the feature set. The scanning windows were divided into nonoverlapping cells measuring 16×16. LBPs were extracted from the cells and concatenated into a cell-structured LBP.

A common problem with these feature descriptors is their high-dimensional feature space. For example, the dimension of HOG is 3780 while LBP is 15,104 per scan window. In such high-dimensional spaces, classical machine learning

Table 1 Representative categories of human detection algorithms based on visual features and their classification algorithms.

Human detection theme	Methods and related works
Visual features	Haar (like) wavelet[2,3] scale-invariant feature transform (SIFT),[4] histograms of oriented gradient (HOG),[5] multilevel histograms of oriented edge energy,[6] local binary pattern (LBP),[7,8] HOG + LBP,[9] center-symmetric (CS)-LBP,[10] Hu-moment[11]
Classification methods	Support vector machine (SVM) cascade,[1] SVMs,[2,5] AdaBoost classifiers,[3] histogram intersection kernel SVMs,[6] LogitBoost cascade,[10] L1-norm minimization learning (LML),[12] LML cascade (CLML)[13]

algorithms such as SVM are almost intractable for training and testing. Color is also not useful information because humans wear a variety of clothing colors. Therefore, Zheng et al.[10] used center-symmetric local binary patterns (CS-LBP) for pedestrian detection to reduce the feature dimensions. Hu-moments used to identify a pattern of rotated object and its position in the three-dimensional (3-D) space.[11]

Relatively few algorithms have been proposed for human classification other than feature extraction. Support vector machines (SVMs) are a representative classifier used for human detection.[1,2,5,6] An SVM classifier is a reasonable choice for general classification because of its high performance and accuracy. However, SVM is not suitable for testing and training when the feature has high dimensionality. Zhu et al.[1] constructed a strong classifier from several weak classifiers by applying a linear SVM to each level of the cascade. Viola et al.[3] used a cascade of AdaBoost classifiers to train a chain of progressively more complex region rejection rules based on Haar wavelet descriptors. This method can reduce the computation time using an approach that combines a cascade of rejection rules. Some researchers have proposed variants of the AdaBoost algorithm. In Ref. 6, it was shown that an approximation of the histogram intersection kernel SVM classifier can be built with the same human-detection performance but a constant runtime with a number of support vectors, as opposed to the linear runtime with a standard approach. Xu et al.[12] proposed L1-norm minimization learning (LML), which is applied widely in the field of signal compression to extract compact feature representations, and they designed a harmonious linear classifier for human detection via L1-norm minimization. In Ref. 13, a cascade of LML classifiers was proposed to provide higher detection rates by training a series of weak-classifiers using L1-LML to construct a strong classifier. Yao and Odobez[14] used a cascade of LogitBoost classifiers with covariance features as human descriptors. The LogitBoost algorithm iteratively learns a set of weak classifiers by minimizing the negative binomial log-likelihood of the training data.

In this study, we extracted more effective and compact features by extending our initial method[15] in several ways to speed up the computation and improve the detection performance as following ways:

1. We generated additional saliency texture map to boost texture of human region by comparing the texture contrast between the human body and the background.
2. One type of CS-LBP features from the scanning window of a saliency texture map and two types of CS-LBP feature from wavelet transformed the subimages are extracted to capture the human texture and reduce the feature dimensions.
3. We proved that the wavelet based CS-LBP extracted from LH subimage is more important feature than HH subimage and HL subimage because humans have strong vertical edges along the body boundaries.
4. Three types of random forests (RFs) are built during the training phase by assembling weak decision trees to model the distribution of each feature using positive and negative classes.
5. The three RFs are rearranged as a cascade.
6. During the testing phase, three types of CS-LBP descriptors are applied to each cascade of RFs. The cascade of RFs combined with three types of CS-LBP descriptors acts as a filter chain, which can increase the detection accuracy by removing negative windows at each level, and it allows human detection to be performed in near real-time.
7. A RF using CS-LBP extracted from LH subimage is located in the second filter, while a RF using CS-LBP extracted from HH subimage is located in the third filter without using HL subimage.

The remainder of this paper is organized as follows. Section 2 describes the feature extraction algorithm for human detection, i.e., three types of CS-LBP from saliency texture map and three wavelet subimages. Section 3 introduces our human verification method using a cascade of RFs. Section 4 presents an experimental evaluation of the accuracy and applicability of our proposed human detection method. Section 5 presents our conclusions and discusses the scope for future work.

2 Human Representation Using CS-LBP Descriptors

In a standing position, the human body can be distinguished from other objects based on the following characteristics:[8] (1) humans have strong vertical edges along the boundaries of the body; (2) the spatial structure of the human body has bilateral symmetry; (3) clothing textures are different from nature textures. Therefore texture is the most pertinent feature for identifying a human body.

HOG and LBP have been used successfully in many human detection studies[8,9] to improve the detection accuracy. However, a common problem with these feature descriptors is their high-dimensional feature space. Therefore, to consider the characteristics of human body and reduce the computation time, we extract the center-symmetric LBP (CS-LBP), which results in a smaller dimension with a similar performance to LBP and HOG using a saliency texture map and wavelet transformed subimages.

2.1 *Center-Symmetric LBP*

CS-LBP[16] uses a modified scheme to compare neighboring pixels in the original LBP, which simplifies the computation

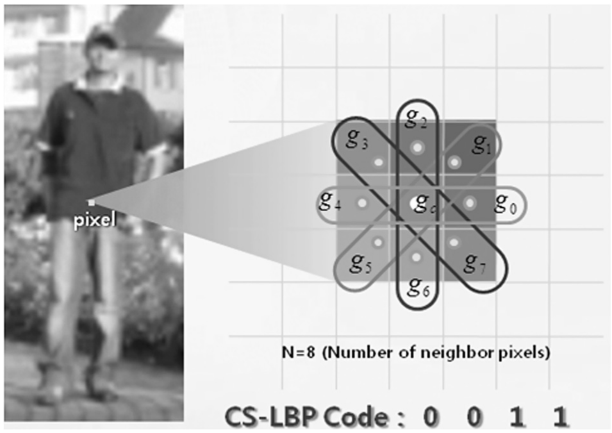

Fig. 1 Example of CS-LBP features for a neighborhood of eight pixels.

while maintaining certain characteristics such as tolerance of illumination changes and robustness against monotonic gray-level changes.[17] CS-LBP differs from LBP because it compares center-symmetric pairs of pixels with a central pixel (g_c), rather than comparing each pixel with the center, as shown in Fig. 1.

LBP produces 256 different binary patterns, whereas CS-LBP only produces 16 (2^4) different binary patterns. Furthermore, robustness is maintained in flat image regions by thresholding the gray-level differences using a small value T with Eqs. (1) and (2),[17] as follows:

$$s(x) = \begin{cases} 1 & x > T \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

$$\mathrm{CS-LBP}_{R,N}(x,y) = \sum_{i=0}^{N/2} s(n_i - n_{i+(N/2)})2^i, \tag{2}$$

where n_i and $n_{i+(N/2)}$ correspond to the gray values of the center-symmetric pairs of pixels for N equally spaced pixels in a circle with radius R.

The concept is similar to a gradient operation, because it calculates the difference between pairs of opposite pixels in a neighborhood.[18]

2.2 *CS-LBP Feature Extraction Using a Saliency Texture and Wavelet-Transformed Subimages*

Humans tend to have strong vertical edges, wear different textural clothing, and they have weak or strong gradients depending on the background intensity, so it is necessary to boost the human boundary by comparing the textural contrast of the inner clothing area with the outer background. To achieve this, we extract CS-LBP descriptors from a saliency texture map and wavelet-transformed subimages of the candidate human region.

Itti et al.[19] proposed a saliency-based visual attention model and selected the most salient area based on a winner-take-all competition. In Ref. 20, the saliency feature maps were drawn from the complementary work of Itti et al.,[19] including the downsampling ratio, color model, and orientation model used for visual search and attention. However, a human's identity generally cannot be determined using color and luminance observation, so texture or geometric properties are better feature for identifying humans than color or luminance. Thus we only estimate a saliency texture map using a simple wavelet transform by modifying the concept proposed in Ref. 20. After a one-level wavelet transform, a saliency texture map $T(c,s)$ is produced from the three high-pass subimages (HH, HL, and LH), rather than a low-pass subimage. Then, only one filter, s, with filter size 9×9 is applied to the $1/4$ down-sampled HH, HL, and LH subimages, c, to reduce the computational requirements. The filter estimates the center-surround difference between the center point and the surrounding points within the filter scale s and this difference yields the feature map. Using Eq. (3), three contrast maps are produced from the three subimages and one filter.

$$\bar{T} = \frac{1}{3}\left[\sum_{c\in\{HH,HL,LH\}}\sum_{s\in\{9\times 9\}} T(c,s)\right]. \tag{3}$$

These maps are then summed and normalized into a saliency texture map $\bar{T}$ without upsampling.

After generating the saliency texture map, a saliency-based CS-LBP (SCS-LBP) descriptor is estimated using the following three steps. First, the saliency texture of half the size of the original scanning window (32×64) is determined. Second, a downsampled scanning window is divided into 4×4 nonoverlapping cells and the CS-LBP codes are calculated using $P = 8$ for the saliency texture map. We represent each SCS-LBP descriptor cell in a histogram with 16 bins. Third, an L1-sqrt normalization step is conducted within each cell. The final SCS-LBP histogram of the saliency texture map is generated by concatenating the local histograms of the 16 cells. There are 16 cells, which means that we generate $16 \times 16 = 256$ histogram bins in the saliency texture map. Figure 2 shows examples of extracting the SCS-LBP descriptor.

As a second feature, we extract a CS-LBP descriptor from wavelet-transformed subimages of candidate human region. Wavelet transforms have a good spatial frequency localization property and they can preserve the spatial and gradient information of an image. LBP pattern extraction from the wavelet domain can also reduce noise because LBP and CS-LBP are suitable for modeling repetitive textures, which means they are sensitive to random noise in uniform image areas.[18] Thus several researchers[18,19] have tried to extract LBP and CS-LBP features from the wavelet-transformed domain. Therefore we extract a CS-LBP descriptor from two wavelet-transformed subimages, rather than LBP descriptors, which reduces the feature dimension. We exclude low-pass and HL high-pass filtered subimages, after one-level Daubechies wavelet decomposition.

In general, human images have a strong edge distribution in the vertical and diagonal directions, but a relatively weak edge distribution in the horizontal direction. Thus, the two high-pass filtered subimages (LH, HH) have important properties when detecting human regions. We prove that an HL subimage is not an appropriate property for human detection in Sec. 4.

The extraction of CS-LBP descriptors from wavelet subimages (WCS-LBPs) consist of three steps. First, a wavelet transform is applied to a 64×128 window that contains a

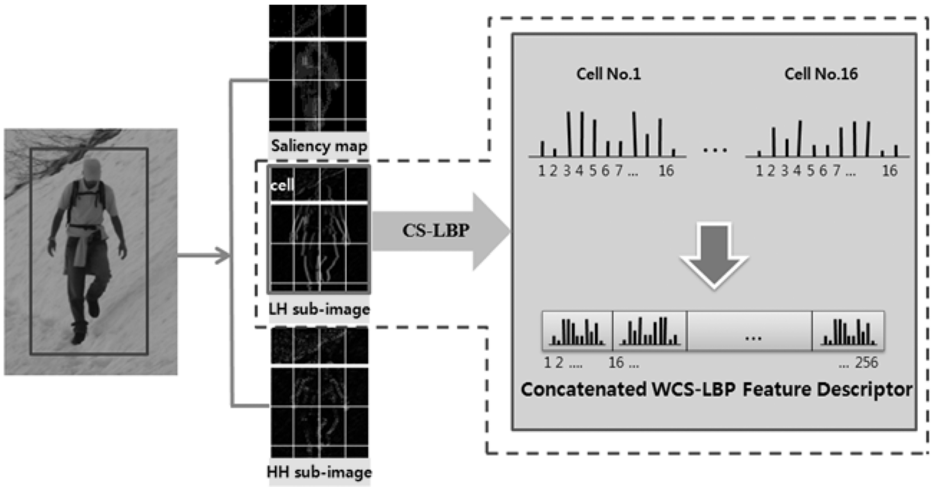

Fig. 2 Steps for extracting the three types of CS-LBP descriptors generated from a scan window: (a) a scan window; (b) downsampled saliency texture map of the LH subimage and HH subimage; (c) feature concatenation of the final WCS-LBP descriptors in the LH subimage. SCS-LBP descriptors are generated using the same method with WCS-LBP.

human. Down sampled high-frequency subimages (i.e., LH) are divided into 4×4 nonoverlapping cells. We calculate the WCS-LBP codes using $P = 8$ for each high-frequency subimage. We represent each WCS-LBP distribution for one cell of the subimage on a histogram with 16 bins. An L1-sqrt normalization step is then conducted within each cell.

The final WCS-LBP histogram for each subimage is generated by concatenating the local histograms of 16 cells. There are 16 cells, which means we generate a total of $16 \times 16 = 256$ histogram bins for each subimage. Figure 2 shows the steps for extracting the WCS-LBP descriptors.

3 Three-Level Cascade of Random Forest for Human Classification

Before extracting the three types of CS-LBPs, square root gamma correction is applied to the input image to map the image luminance into a more perceptually uniform domain, based on the results of Ref. 5. The scanning windows are classified into human and nonhuman classes using the three types of CS-LBPs and pattern classifiers. An SVM classifier is a reasonable choice for general classification because of its high performance and accuracy. Many human detection algorithms have used SVM classifiers.[1,2,5,6] However, SVM is not suitable when the feature has high dimensionality. A cascade of AdaBoost classifiers[3] and variants of the AdaBoost algorithm[1,8] have also been used for human classification. However, despite the great success of AdaBoost (and its descendant algorithms) in theory and applications, it is difficult to use AdaBoost to classify a target class with significant intra-class variation against a large background class.[21]

In this study, we modified the original version of RFs proposed by Breiman.[22] A RF is a decision tree ensemble classifier, where each tree is grown using some form of randomization. RFs have the capacity to process huge amounts of data at high training speeds.

A RF is a collection of T binary structured trees rather than a single decision tree.

$$\mathrm{RF} = [h_t(\mathbf{x}, \Theta_t)], \qquad t = 1, \ldots , T, \tag{4}$$

where h_t is the t'th individual tree and $h(\cdot)$ is the tree's prediction. $\mathbf{x}$ represent human and nonhuman window samples, $\mathbf{x} = \{S_n\}, n = 1, \ldots , N_{\mathrm{samples}}$, where each S_n is a database sample. Θ_t are independent identically distributed random vectors, and each tree casts a unit vote for the most popular class at input $\mathbf{x}$.

In this study, the two classes of RFs were defined as human and nonhuman. These classes have corresponding probability values and their probabilities indicate the likelihood of being human at a particular time. The basic characteristics of the three types of CS-LBPs were different, so we trained three RFs using each CS-LBP descriptor extracted from one saliency texture map and two wavelet-transformed subimages in a cascading manner.

Our proposed three-level cascade of RFs is similar to a series of AdaBoost classifiers,[1] but with the following modifications to reduce the computation time. We combined a series of RF classifiers as a filter chain, as shown in Fig. 3. Each filter is a set of strong classifiers (decision trees) having a number of n weak classifiers. Since each node of a decision tree has a respective split function, we regard split functions as week classifiers. In this study, we generated a three-level cascade of filters (RFs) CRF1, CRF2, and CRF3 separately, using the three types of CS-LBP descriptors.

For each level of the cascade, we train the individual decision tree until predefined quality requirements were met by modifying original RF. After training the individual tree, these trees form a RF and each RF is concatenated to produce a three-level cascade of RFs sequentially.

During training, 950 training images were randomly selected (450 positive samples and 500 negative samples) from the INRIA Person dataset.[5] The maximum false positive rate for an I-level RF was 0.7 in each stage. An I-level RF cascade for human detection consisted of the following learning procedures (Algorithm 1).

The important parameters of the RF are the depth of trees and the number of trees. In this study, we set a maximum tree depth of 20 and the number of tree sets (weak classifiers) in each stage was 120, based on the experimental results of Ref. 18.

Algorithm 1 Training the I-level cascade of a random forest.

```
  Input: $F_{\text{target}}$: the maximum acceptable false positive rate per I-level random forest
             $T_{\text{target}}$: the maximum number of trees to grow
             $D_{\text{target}}$: the maximum depth of trees to extend
         $S_n$: the number of training sets, including positive and negative samples
  Initialize: $i = 0$, $j = 0$, $k = 0$; $F_i = 1.0$
Assign $n$ bootstrap samples and their I-th CS-LBP descriptors to training set $S_n$
Loop: $F_i > F_{\text{target}}$
        $i = i + 1$
    Loop: $T_i < T_{\text{target}}$
          $j = j + 1$
          Select $n$ new bootstrap samples from training set $S_n$
            Loop: $D_i < D_{\text{target}}$
                $k = k + 1$
              1. Grow an unpruned tree using the n bootstrap samples.
              2. Each internal node randomly selects $p$ variables and determines the best split function using only these variables. Using
                 different $p$'th variables, the split function $f(v_p)$ iteratively splits the training data into left ($I_l$) and right ($I_r$) subsets using
                 equation. $I_l = [p \in I_n | f(v_p) < t], I_r = I_n \setminus I_l$. The threshold $t$ is randomly chosen by the split function $f(v_p)$ in the range
                 $t \in [\min_p f(v_p), \max_p f(v_p)]$.
            Loop end
            Add the $j$'th decision tree (week classifiers) to the RF (strong classifier) Loop end
        Evaluate positive and negative using the current N decision trees and compute $F_i$.
          If $F_i > F_{\text{target}}$
        Next, evaluate the current I-level RF for the negatives (i.e., nonhuman images), and add misclassified samples
        to the negative samples
   Loop end
   Output: An I-level RF
         Each level has a boosted RF classifier that consists of decision trees.
```

After training three-level RFs, the test candidate windows are applied to the rejection cascade of RFs. In the proposed three-level cascade of RFs, each RF (CRF1, CRF2, and CRF3) only used its assigned three types of CS-LBP descriptors for classification. Humans tend to have a strong gradient contrast compared with the background intensity, so a RF using the SCS-LBP extracted from the saliency texture map is located in the first filter. In addition, humans have strong vertical edges along the body boundaries, so a RF using the WCS-LBP extracted from LH is located in the second filter, while a RF using WCS-LBP extracted from HH is located in the third filter. The order of the RFs is determined by the performance of each filter, as shown in Fig. 4.

In the first RF using SCS-LBP, the majority of scanning windows are discarded according to the passing Eq. (6). The remaining windows are applied to the second RF using the WCS-LBP extracted from LH. Low-probability windows are discarded using the same method. The final windows that pass through the third RF using the WCS-LBP extracted from HH are declared as human regions, if they also satisfy the passing Eq. (6).

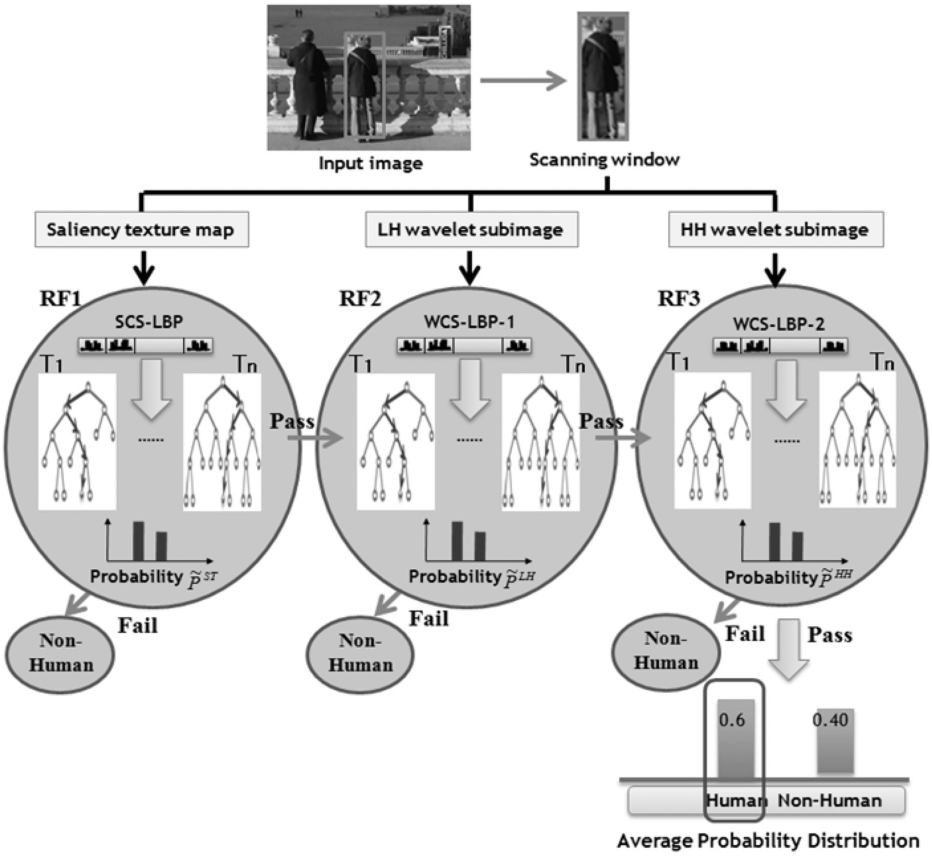

Fig. 3 Classification process using separate CS-LBPs with a trained cascade of random forests. In this example, the test image was classified into the human class, and it had an average posterior probability of 0.6.

The probability distribution of the I'th RF is generated by the arithmetic averaging of each distribution of all trees $L = (l_1, l_2, \ldots, l_T)$, using Eq. (5):

$$\tilde{P}(c_{i\in\{\text{human,non-human}\}})=\frac{1}{T}\sum_{t=1}^{T}P(c_{i\in\{\text{human,nonhuman}\}}|lt). \qquad (5)$$

In Eq. (5), T is the number of trees. Then, according to Eq. (6), the scanning window is passed to the next RF if the average probability of a human class $(\tilde{P}(c_{i\in\{\text{human}\}}))$ is greater than the average probability of a nonhuman class $[\tilde{P}(c_{i\in\{\text{nonhuman}\}})]$.

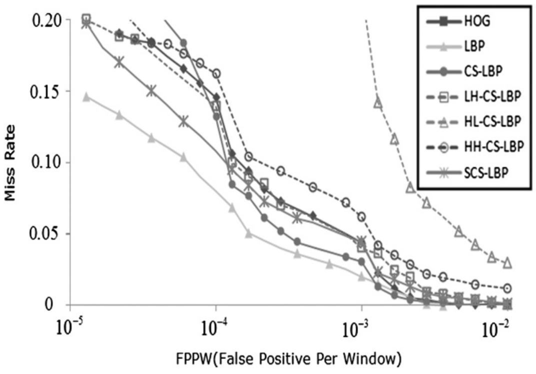

Fig. 4 FPPW versus MR for the three proposed types of CS-LBPs with four different features using a one-level random forest and the INRIA dataset.

$$\begin{cases}\text{Pass} & \text{if}[\tilde{P}(c_{i=\{\text{human}\}}) > \tilde{P}(c_{i=\{\text{nonhuman}\}})] \\ \text{Fail} & \text{otherwise}\end{cases} \qquad (6)$$

Figure 3 shows that if the scanning window passes through the three filters, it is declared as a human and the final probability of the input window $(\tilde{P}(c_i))$ is estimated by averaging the three level probabilities (SCS-LBP:$\tilde{P}^{\text{ST}}(c_i|l_t)$, WCS-LBP:$\tilde{P}^{\text{LH}}(c_i|l_t)$, $\tilde{P}^{\text{HH}}(c_i|l_t)$) using Eq. (7).

$$\tilde{P}(c_i) = \left[\frac{1}{T}\sum_{t=1}^{T}\tilde{P}^{\text{ST}}(c_i|lt) + \tilde{P}^{\text{LH}}(c_i|lt) + \tilde{P}^{\text{HH}}(c_i|lt)\right]/3 \qquad (7)$$

4 Experimental Results

We performed experiments using the INRIA person dataset,[5] which includes a wide variety of human body and background scenes for training and testing systems. This dataset was collected as part of INRIA's research work on detection of upright people in images and video without particular camera setting. Many people are bystanders taken from the backgrounds of these input photos, so ideally there is no particular bias in their pose. This database provides a training dataset containing 2418 positive and negative samples of 64- × -128 pixels, as well as dynamic background images containing no humans that can be used as negative exemplars. For testing, we used 1380 human samples of 70- × -134 (a margin of pixels around each side) pixels and 253 images containing nonhumans of 320-× -240 pixels.

During testing, we applied three samplings of the one-level wavelet-transformed image (160 × 120): upsampling by a ratio of 1.2 from the one-level wavelet resolution to 144- × -192 pixels, downsampling by a ratio of 0.8 from the one-level wavelet resolution to 96- × -154 pixels, and downsampling by a ratio of 0.6 from the one-level wavelet resolution to 72- × -96 pixels. In addition, the shifting step-size of the scanning window was four pixels, and it was reduced by three pixels in 0.8 and by two pixels in 0.6 down-sampled images.

To evaluate the experiments for human detection, false positive per window (FPPW) and false positive per image (FPPI) methodologies are generally used instead of traditional confidence level.[23] In the FPPW, the detector is evaluated by classifying cropped humans versus nonhumans crops. In the FPPI, the detector scans the image by a sliding window approach and evaluate the correspondence between the detected bounding box and the ground truth.

In this study, we used a FPPW versus a miss rate (MR) as performance criteria as follows because most researches are evaluated the performance using FPPW and INRIA dataset is optimized to evaluate FPPW.

- FPPW versus MR curves, where FPPW was defined as $\text{FalsePos}/(\text{TrueNeg} + \text{FalsePos})$ while MR was defined as $\text{FalseNeg}/(\text{FalseNeg} + \text{TruePos})$. The main purpose of the human detection problem is to minimize the miss rate with a very low false positive rate. *FalsePos* is obtained by testing all the windows of the test data that overlapped by less than 50% with any ground truth object. By contrast, *TruePos* was obtained by testing all windows in the test data that overlapped by over 50% with any ground truth object.

Experiments to detect humans in the test data were performed using an Intel Core 2 Quad processor PC with a Windows 7 operating system.

4.1 *Performance Evaluation of the Feature Sets*

To validate the effectiveness of the three proposed types of CS-LBP features, we compared the human detection performance using five different features with one-level RF classifiers: HOG,[5] LBP,[17] CS-LBP,[16] CS-LBPs from LH wavelet subimages (LH-CS-LBP), CS-LBPs from HL wavelet subimages (HL-CS-LBP), CS-LBPs from HH wavelet subimages (HH-CS-LBP), and CS-LBP from a saliency texture map (SCS-LBP).

Figure 4 shows the results of the FPPW and MR curves. As shown in Fig. 4, we confirmed that each SCS-LBP, LH-CS-LBP, and HH-CS-LBP produced similar good detection performance compared with the original HOG and CS-LBP (i.e., approximately 0.11, 0.14, and 0.16 with a FPPW of 10^{-4}). LBP had the best detection performance, but the computation time required for detection was 1.5 times higher than SCS-LBP because of its high dimensionality. SCS-LBP produced the second best detection performance of the other five features with a FPPW of 10^{-4}. This method provided 0.04 lower MR than HOG features, which have been used in many previous studies.[5,9] HL-CS-LBP produced the worst detection results, showing that HL subimages are not appropriate properties for human detection.

4.2 *Performance Comparison with Related Studies*

To evaluate the performance of the proposed algorithm, HOG with Adaboost (HOG + Adaboost),[1] HOG with SVM (HOG + SVM),[5] and CS-LBP with SVM (CS-LBP+ SVM),[10] which provide the best performance of existing algorithms, were compared with the three proposed types of CS-LBP with a three-level cascade of RFs. The experiments were performed using the same INRIA dataset. As shown in Fig. 5, we confirmed that our proposed algorithm produced better human detection performance than the other three methods. With an FPPW rate of 10^{-4}, our method achieved a 0.03MR, which was 0.11 lower than the HOG-SVM method, 0.09% lower than the CS-LBP + SVM method, and 0.07 lower than the HOG+ Adaboost method.

The main reason for the lower MR with our proposed method compared with related methods was that our algorithm found many potential candidate human regions during the first cascade using the SCS-LBP feature, while it eliminated a large amount of false positives in the last two cascades using two different wavelet subimages.

In addition, the RF classifier has the capacity to process huge amounts of data with high training speeds and better performance than SVM,[18] because it is based on decision trees, so our proposed cascade of RFs also had better performance than SVM-based detection methods.

4.3 *Performance Evaluation of Computation Time Requirements*

One of the main advantages of our method is a reduction in the computation time, so we compared the computation time required for human detection using our proposed method and three related approached [i.e., HOG with Adaboost (HOG + Adaboost)[1]], HOG with SVM (HOG + SVM),[5] and CS-LBP with SVM (CS-LBP + SVM).[10]

The computational complexity was evaluated by applying the detectors to test data and measuring the average processing time in per image instead of operation number because false data can be rejected in the first or second level. The average processing speeds for HOG + SVM, CS-LBP+ SVM, HOG + Adaboost, and our proposed method were 2.52, 0.93, 0.45, and 0.28 s per image, respectively, using

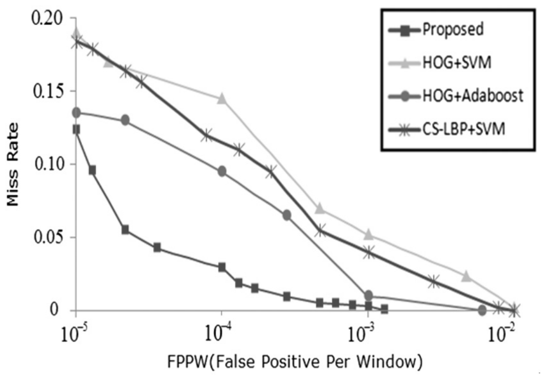

Fig. 5 FPPW versus MR for the proposed detection algorithm with three different methods using the same INRIA dataset.

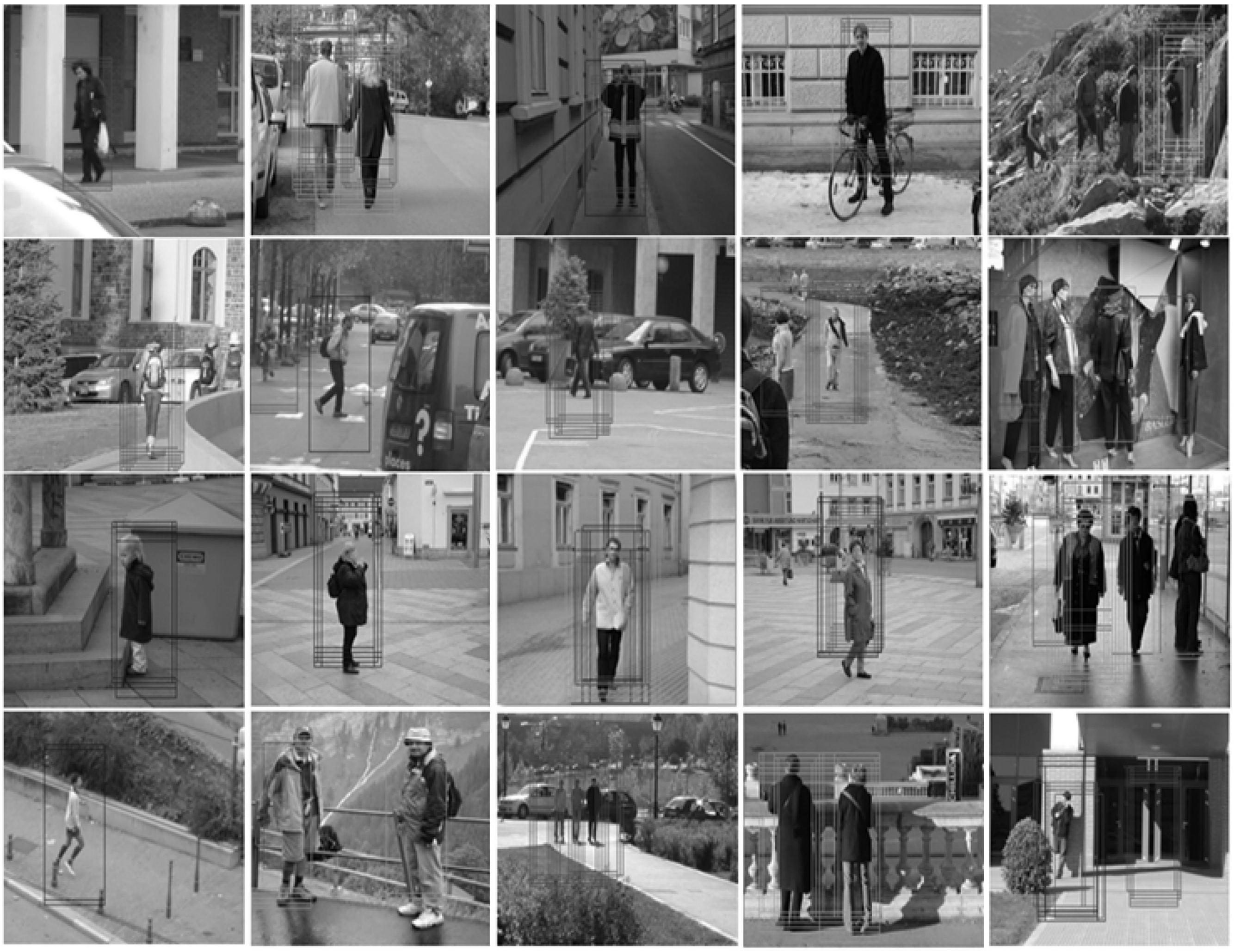

Fig. 6 Sample human detection results using INRIA images with the proposed method. The final column shows some false and misidentified results. Red box represents detected results from one-level wavelet-transformed image, blue box represents detected results from downsampling by a ratio of 0.8, green box represents detected results from downsampling by a ratio of 0.6, and yellow box represents detected results from upsampling by a ratio of 1.2.

the same system environment. Thus our proposed method was faster than the other three methods because it could reject false windows using a cascade method, and it had low-dimensional SCS-LBP and WCS-LBPs features. In particular, RFs reduced the computation time for testing regardless of any increase in the dimensions of test images. However, the computation time requirement of the SVM-based method increased linearly as the dimension of the test image increased.

Figure 6 shows some human detection results with our proposed method using the INRIA test set. As shown in Fig. 6, our proposed method detected humans correctly in test images containing humans of different sizes and in backgrounds rich in texture information. However, our method yielded some false or misdetection (last column) results when a human was occluded by background objects or when a background object had a similar structure to humans.

5 Conclusion

In this study, we developed a three-level cascade of RFs using three types of CS-LBP descriptors to improve human detection performance, which significantly reduced the time required for human detection.

To detect human regions, we extracted two types of CS-LBP descriptors from the scan window of a saliency texture map and wavelet-transformed subimages (i.e., LH and HH, but not HL). The two CS-LBP descriptors were then applied to a corresponding cascade of RFs, which were ensembles of random decision trees. The experimental results with INRIA images showed that our algorithm improved the human detection performance when compared with other feature descriptors and other classification methods.

In the future, we plan to modify our algorithm to handle partial occlusion and the articulated deformation of humans in image and videos. Moreover, we plan to apply our approach to thermal videos for the nighttime detection of humans.

Acknowledgments

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation.

References

1. Q. Zhu et al., “Fast human detection using a cascade of histograms of oriented gradients,” in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Washington, DC, pp. 1491–1496 (2006).
2. C. Papageorgiou and T. Poggio, “A trainable system for object detection,” *Int. J. Comput. Vis.* **38**(1), 15–33 (2000).
3. P. Viola and M. Jones, “Detecting pedestrians using patterns of motion and appearance,” in *Proc. IEEE Conf. Computer Vision*, Nice, France, pp. 734–741 (2003).
4. D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.* **60**(2), 91–110 (2004).
5. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, San Diego, CA, pp. 886–893 (2005).
6. S. Maji, A. Berg, and J. Malik, “Classification using intersection Kernel support vector machines is efficient,” in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Anchorage, Alaska, pp. 1–8 (2008).
7. Y. T. Chen and C. S. Chen, “Fast human detection using a novel boosted cascading structure with meta stages,” *IEEE Trans. Image Process.* **17**(8), 1452–1464 (2008).
8. M. Zhang, J. Lv, and J. Yang, “Human detection using relational color similarity features,” *Opt. Eng.* **50**(9), 097201 (2011).
9. X. Wang, T. X. Han, and S. Yan, “An HOG-LBP human detector with partial occlusion handling,” in *Proc. IEEE Conf. Computer Vision*, Kyoto, Japan, pp. 32–39 (2009).
10. Y. Zheng et al., “Pyramid center-symmetric local binary/trinary patterns for effective pedestrian detection,” *LNCS* **6495**, 281–292 (2011).
11. J. A.r Muñoz-Rodríguez, “Shape connection by pattern recognition and laser metrology,” *Appl. Opt.* **47**(20), 3590–3608 (2008).
12. R. Xu et al., “Human detection in Images via L1-norm minimization learning,” in *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing*, Dallas, pp. 3566–3569 (2010).
13. R. Xu et al., “Cascaded L1-norm minimization learning (CLML) classifier for human detection,” in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, San Francisco, pp. 89–96 (2010).
14. J. Yao and J. M. Odobez, “Fast human detection from videos using covariance features,” in *Proc. European Conf. on Computer Vision Visual Surveillance Workshop*, pp. 1–8, Springer, Marseille, France (2008).
15. D. Y. Kim et al., “Human detection using wavelet-based CS-LBP and a cascade of random forest,” in *Proc. IEEE Conf. on Multimedia and Expo*, Melbourne, Australia, pp. 362–367 (2012).
16. M. Heikkilä, M. Pietikäinen, and C. Schmid, “Description of interest regions with local binary patterns,” *Pattern Recogn.* **42**(3), 425–436 (2009).
17. T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002).
18. B.C. Ko, S. H. Kim, and J. Y. Nam, “X-ray image classification using random forests with local wavelet-based CS-local binary patterns,” *J. Digital Imag.* **24**(6), 1141–1151 (2011).
19. L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998).
20. B. C. Ko and J. Y. Nam, “Object-of-interest image segmentation using human attention and semantic region clustering,” *J. Opt. Soc. Amer.* **23**(10), 2462–2470 (2006).
21. O. M. Danielsson, B. Rasolzadeh, and S. Carlsson, “Gated classifiers: boosting under high intra-class variation,” in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Colorado Springs, pp. 2673–2680 (2011).
22. L. Breiman, “Random forests,” *Mach. Learn.* **45**(1), 5–32 (2001).
23. J. E. Freund and B. M. Perles, *Modern Elementary Statistics*, 12th ed., Pearson (2006).

Byoung Chul Ko received a BS from Kyonggi University, Korea, in 1998, and an MS and PhD in computer science from Yonsei University, Korea, in 2000 and 2004. He was a senior researcher of Samsung Electronics from 2004 through 2005. He is currently an associate professor in the Department of Computer Engineering, Keimyung University, Daegu, Korea. His research interesting includes content-based image retrieval, fire detection, and robot vision.

Deok-Yeon Kim received a BS from Keimyung University, Korea, in 2011. He is currently a master student of Keimyung University, Korea. His research interesting includes human detection.

Ji-Hoon Jung received a BS from Keimyung University, Korea, in 2012. He is currently a master student of Keimyung University, Korea. His research interesting includes object detection.

Jae-Yeal Nam received BS and MS degrees from Kyongbuk National University, Korea, in 1983 and 1985. He received a PhD in electronic Engineering from University Texas at Arlington, USA, in 1991. He was a researcher of ETRI from 1985 through 1995. He is currently a professor in the Department of Computer Engineering, Keimyung University, Daegu, Korea. His research interesting includes video compression and content-based image retrieval.