Journal of Electronic Imaging

SPIEDigitalLibrary.org/jei

Human object annotation for surveillance video forensics

Muhammad Fraz
Iffat Zafar
Giounona Tzanidou
Eran A. Edirisinghe
Muhammad Saquib Sarfraz

SPIE

IS&T imaging.org

Human object annotation for surveillance video forensics

Muhammad Fraz*
Iffat Zafar*
Giounona Tzanidou
Eran A. Edirisinghe
Loughborough University
Department of Computer Science
Loughborough LE11 3TU, United Kingdom
E-mail: m.fraz@lboro.ac.uk

Muhammad Saquib Sarfraz
Karlsruhe Institute of Technology
Computer Vision for Human Computer Interaction Lab
76131 Karlsruhe, Germany

Abstract. *A system that can automatically annotate surveillance video in a manner useful for locating a person with a given description of clothing is presented. Each human is annotated based on two appearance features: primary colors of clothes and the presence of text/logos on clothes. The annotation occurs after a robust foreground extraction stage employing a modified Gaussian mixture model-based approach. The proposed pipeline consists of a preprocessing stage where color appearance of an image is improved using a color constancy algorithm. In order to annotate color information for human clothes, we use the color histogram feature in HSV space and find local maxima to extract dominant colors for different parts of a segmented human object. To detect text/logos on clothes, we begin with the extraction of connected components of enhanced horizontal, vertical, and diagonal edges in the frames. These candidate regions are classified as text or nontext on the basis of their local energy-based shape histogram features. Further, to detect humans, a novel technique has been proposed that uses contourlet transform-based local binary pattern (CLBP) features. In the proposed method, we extract the uniform direction invariant LBP feature descriptor for contourlet transformed high-pass subimages from vertical and diagonal directional bands. In the final stage, extracted CLBP descriptors are classified by a trained support vector machine. Experimental results illustrate the superiority of our method on large-scale surveillance video data.* © *2013 SPIE and IS&T [DOI: 10.1117/1.JEI.22.4.041115]*

1 Introduction

The use of closed-circuit television (CCTV) cameras for the surveillance of important public and private places has significantly increased recently. Networks of CCTV cameras are in operation to assist security agencies to keep an eye on suspicious individuals and record abnormal events. With this increase, need for an effective management of growing collections of video data has also increased. It is a significantly difficult task to index such data by manually assigning meaningful annotations in order to enable rapid search. In this case, at the most basic level, an event of interest is described either as a vehicle or a person passing a camera.

Humans are the most important subjects for monitoring in any typical video forensics application. They are usually described in terms of appearance, size, and behavior/action. For example, it may be required to identify a person wearing a particular colored top (appearance) with certain height (size) running (action) from the scene. However, it may also be useful to perform further detailed annotations such as "a person wearing a short-sleeved, white and blue top that includes text (or a logo)" and black trousers. Annotation of human objects at different levels of detail can help in forensic applications, which will involve the search for people with a known description as seen by a witness.

Processing videos obtained from CCTV cameras for forensic purposes is challenging. They are often of low resolution and frame rate, vary in quality such as environmental conditions or lighting, change over time, contain noise, and vary in direction of view. No standard exists for defining what information a personal description should contain in video forensic applications. In most cases, the description mimics a witness statement and what the witness remembers. Therefore, a limited number of research exists in the application of computer vision and pattern recognition approaches to CCTV video forensics. In addition, most of the standard techniques used in video analytics, which may be potential candidates for the use in video forensics (see Sec. 2), do not work effectively when applied to typical low-medium-quality CCTV videos of moderate-to-crowded public scenes.

In this work, we aim at an automatic annotation system for surveillance videos. The paper focuses initially on the classification of foreground moving objects into humans and nonhumans and the subsequent detailed annotation of these objects based on a number of appearance measures. These measures include color of clothes on different parts of the human body and the presence of any text/logo information.

The paper is organized as follows: After a brief review of related work in Sec. 2, the proposed methodology is

*The authors contributed equally and assert joint authorship for this work.

Paper 13211SSP received Apr. 17, 2013; revised manuscript received Jul. 20, 2013; accepted for publication Jul. 30, 2013; published online Aug. 29, 2013.

explained in Sec. 3. Experimental results and a discussion are presented in Sec. 4, followed by the conclusion and future enhancements in Sec. 5.

2 Related Work

Any annotation process begins with the classification of foreground segmented blobs into humans and nonhumans. Our literature review revealed that automated detection of humans is a very well-researched area. Papageorgiou and Poggio[1] used Haar wavelets as descriptors of a human figure. This method is invariant to changes in color and texture. Viola and Jones[2] integrated image intensity information with motion using Haar-like wavelets and applied this method to human movement detection. Lowe[3] used scale-invariant feature transform (SIFT) descriptors to describe local features in images. As SIFT features are local and are based on the appearance of an object at particular points of interest, it follows that they are invariant to image scale and rotation. Dalal and Triggs[4] used the locally normalized histograms of oriented gradient (HOG) descriptors for human detection. Chen and Chen[5] used a combination of intensity-based rectangular and gradient-based features. Wang et al.[6] combined HOG with cell-structured local binary pattern (LBP) as the feature set. LBP descriptors are invariant to monotonic gray-level changes and they are computationally efficient. However, these feature descriptors have a problem of containing a high-dimensional feature space. To overcome this problem, Zheng et al.[7] used center-symmetric LBPs (CS-LBPs) for pedestrian detection. Kim et al.[8] proposed an approach based on the combination of a wavelet-based CS-LBP (WCS-LBP) with a cascade of random forests. Three types of WCS-LBP features are extracted from the scanning window of wavelet-transformed subimages to reduce the feature dimension. The extracted WCS-LBP descriptors are then applied to a cascade of random forests. Our work presented in this paper is closely related to that presented in Ref. 8.

There is no existing standard on which descriptors should be used to annotate humans. Therefore, existing techniques vary in the type and number of features used to annotate humans. The work by Hansen et al.[9] is more closely related to our work. The system proposed by them is based on four annotations associated to humans:primary color of the clothing, the height, and focus of attention.

We discuss existing literature based on individual appearance measures (color extraction and text/logo detection) used in the proposed research as follows.

Color is an important attribute for efficient visual processing. According to Schettini et al.,[10] the problem of color extraction is based on either predefined conversant colors or a query illustration. Our work addresses only the latter case and therefore requires a query input. Brown[11] proposed a very similar system for retrieval of predefined familiar object color. Their framework accumulates a histogram of colored pixels for a small number of human perceived colors. Parameterization of this discretization is performed to determine the dominant color of the object. Swain and Ballard[12] provided the initial idea of color recognition based on color histograms, which are matched by histogram intersection. Modifications of this idea contain improvements upon histogram measurements, incorporating information about the spatiotemporal relationships of the color pixels. Our previous approach[13] is based on a histogram binning technique to extract the dominant color pixels in an object.

Wui et al.[14] addressed the task of color classification into prespecified colors for tracked objects. Weijer et al.[15] and Zhang et al.[16] proposed probabilistic latent semantic analysis-based approaches for object color categorization in videos. These methods rely on complex features like SIFT[3] and maximally stable extreme regions[17] to articulate the objects into various parts, i.e., tires and windscreens for vehicles, and separate them in order to reduce the effect of their color in the categorization of vehicle's main color. These methods require extensive processing, which makes them less suitable for real-time applications.

Text/logo detection serves as another important and detail-level appearance measure for human annotation. A significant amount of literature exists on detection of text/logo from images and videos.

Jin and Geman[18] established composition machines for constructing probabilistic hierarchical image models that hold contextual relationships. The approach allows reusability of parts among multiple entities and non-Markovian distributions. Berg et al.[19] presented a Markov chain framework for parsing images. Weinman and Miller[20] proposed a technique that combines language information (such as bi-grams and letter case) and image features in a single model and integrates dissimilarity information between character images.

A few researchers have used cascaded detectors in order to identify the candidate region in an image. In Ref. 21, Chen and Yuille employed the AdaBoost algorithm and the joint probabilities of the features s (X and Y derivatives, histogram of intensity and edge linking) to detect text in natural scenes. This approach was initially proposed by Viola and Jones[22] for face detection. It is also mentioned in Ref. 21 that simple features like Haar-like features are not sufficiently capable to effectively deal with text regions in images due to variations (e.g., font and illumination) that typically exist on text regions. For this reason, they used block-based informative features. However, some of the features they used are not very cost effective in terms of the processing time required. Lalonde and Gagnon[23] detected text in documentary films using the same technique but with a different feature set.

3 Proposed Methodology

This section describes the proposed method for human annotation in CCTV videos. As in other image processing systems, the proposed system includes an image preprocessing stage that reduces the noise and enhances the quality of input video for obtaining improved accuracy in subsequent processing stages. The proposed approach to detailed annotation of human objects in CCTV video comprises a number of processing stages. In the first stage, we extract foreground objects from a video frame. This is followed by examining the foreground for various features to enable their categorization into humans or nonhumans. Once objects have been classified as humans, they are further analyzed to extract details such as separation of the human body into head, torso, and leg regions, extraction of dominant colors from these regions, and the detection of text/logo that may be present on the trousers or the shirt of the person being detected (see Fig. 1). The operational details of all the above-mentioned stages are presented in the following sections.

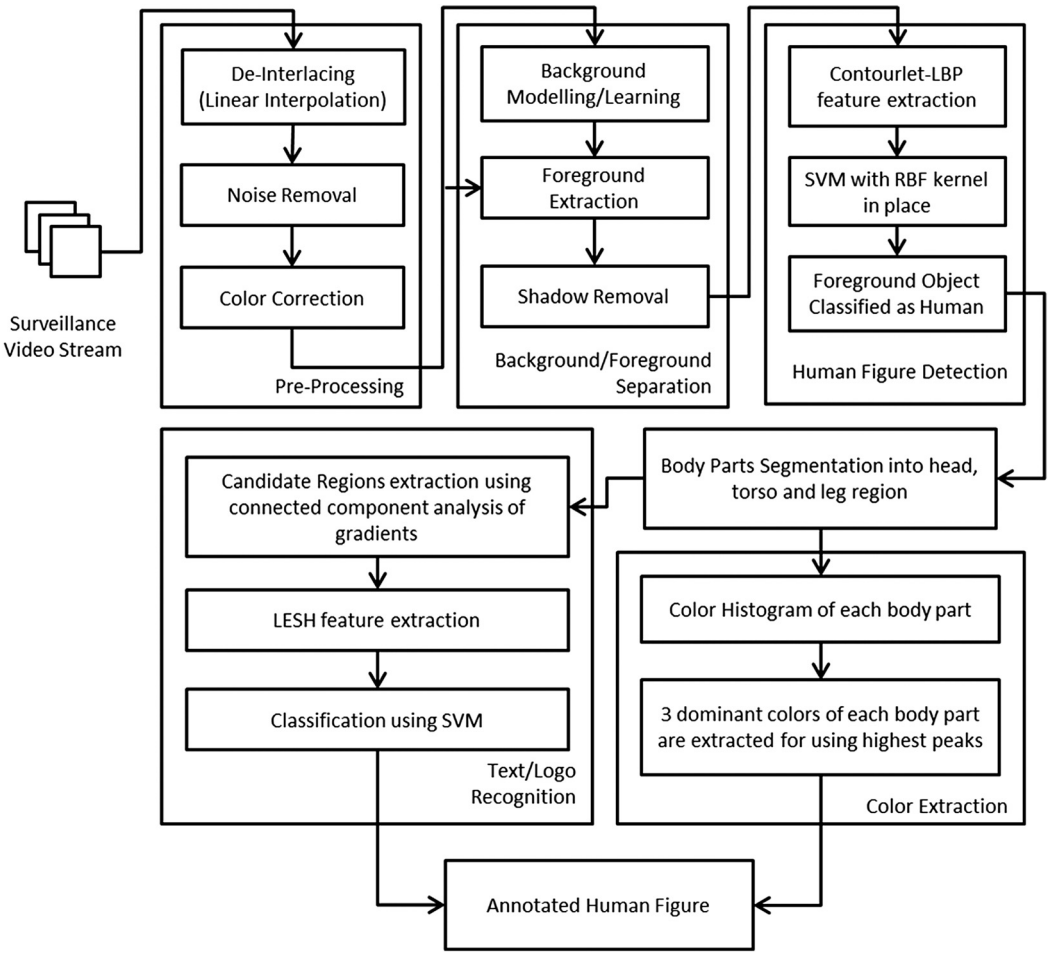

Fig. 1 Overview of the proposed human object annotation framework.

3.1 *Preprocessing*

CCTV videos are often recorded at very low resolution. Pixel interlacing and compression up to a maximum possible level is applied so that these videos acquire the least possible space on storage media and/or require minimal bandwidth in the event of transmission to remote locations. These factors significantly affect the ability to capture important details of videos and pose a challenge for video analytics and retrieval systems in reaching an acceptable level of performance accuracy. To address the poor quality of CCTV videos, we have proposed the use of an image preprocessing pipeline (see Fig. 1) that includes a multitude of enhancement operations. Image preprocessing pipeline begins with deinterlacing input frames. Deinterlacing is performed through linear interpolation of the values of neighboring pixels. Deinterlaced frames are processed through a median filter to plummet the effect of nonlinear noise, resulting in a better quality frame with clearer details.

3.2 *Color Correction*

We propose the use of our previously proposed color correction approach[13] that consists of a conventional color constancy algorithm followed by a novel set of postprocessing procedures to obtain true colors of objects present in CCTV video frames. The detailed operation of these two stages can be presented as follows.

3.2.1 *Color constancy*

Color constancy is extremely important to reduce the effect of changing light sources. It is impossible for a color recognition system to perform well without the aid of a good color constancy framework. We have adopted conventional gray world (GW)[24] computational color constancy technique to reduce the effect of variable illumination and camera calibrations. The GW algorithm[24] is based on the assumption that the color in each sensor channel averages to gray over the entire image. The GW algorithm estimates the deviation from the assumptions and is given by a simple expression

$$l_r = \text{mean}(E_R), \quad l_g = \text{mean}(E_G), \quad l_b = \text{mean}(E_B), \quad (1)$$

where l_r, l_g, l_b are the mean values of pixel intensity in each channel, respectively, and E_R, E_G, E_B are individual image channels. The GW algorithm[24] provides a constancy solution independent of the illuminant color by dividing each color channel by its average value.[25]

$$\begin{bmatrix} R_{\text{out}} \\ G_{\text{out}} \\ B_{\text{out}} \end{bmatrix} = \begin{bmatrix} \frac{1}{l_r} & 0 & 0 \\ 0 & \frac{1}{l_g} & 0 \\ 0 & 0 & \frac{1}{l_b} \end{bmatrix} \begin{bmatrix} R_{\text{in}} \\ G_{\text{in}} \\ B_{\text{in}} \end{bmatrix} \quad (2)$$

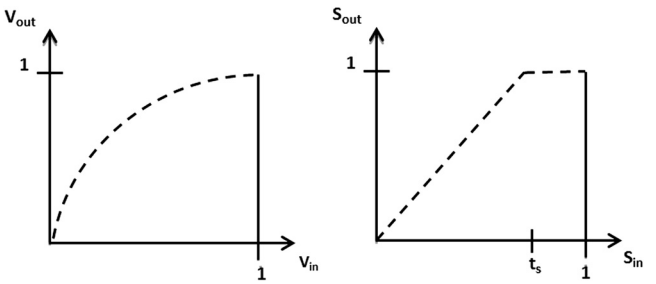

Fig. 2 Modified value (V) and saturation (S) components.

3.2.2 *Post-CC enhancements*

Every frame of an input video is passed through a further stage of color enhancement where the output of color constancy stage is processed in the HSV color space to boost contrast and brightness. CCTV videos are often so poor in quality that even after color constancy procedures the actual colors of objects appear dull and indistinguishable. The proposed method applies modifications to Saturation and Value components to enhance color information and further applies gamma correction to improve illumination as detailed below.

The original Value component represented as V_{in} (of all pixels) is scaled up in a way that the lower V_{in} values get a higher scaling factor while the scaling factor for higher values decreases gradually. This can be represented as a form of gamma correction[26] [see Eq. (3)]. For all those pixels that have their saturation component less than threshold t_s, the saturation component is scaled up by a factor of f_s. The saturation value of the remaining pixels is increased up to a maximum [Eq. (4)], i.e., 1.

$$V_{\text{out}} = V_{\text{in}}^{\gamma} \tag{3}$$

$$S_{\text{out}} = \begin{cases} S_{\text{in}} * f_s & \text{if } S_{\text{in}} < t_s \\ 1 & \text{if } S_{\text{in}} \geq t_s \end{cases}, \tag{4}$$

where $\gamma = 0.333$, $f_s = 1.25$, and $t_s = 0.8$ has worked best in our experiments Fig. 2 illustrates the effect of proposed modifications on input Value and Saturation components.

The use of proposed modifications improve the contrast (and the appearance) of the image, hence making the frames better prepared for subsequent processing.

3.3 *Foreground Extraction*

Before any annotation can commence, moving foreground objects should be segmented from the video frames. To this end, we have applied an improved version of the foreground object segmentation method initially proposed by Stauffer and Grimson.[27] The background is modeled by a mixture of Gaussian (MoG) distributions; i.e., each pixel in the video frames is modeled by a weighted MoG whose parameters and weights are continuously updated throughout the time. Based on probability, this allows the prediction of the pixel values that belong to the background in consecutive frames. However, Stauffer and Grimson's original proposal[27] has some shortcomings, namely, its inefficiency of dealing with shadows, spurious objects, illumination changes, and similarities in foreground-background regions. This has led to the proposal of a significant number of new approaches aimed at improving the original concept. In our attempt to utilize the positive contributions of the use of the Gaussian mixture model and at the same time overcome its weaknesses, we combined some of the most successful improvements to address the aforementioned issues.

The most common approach adopted by researchers[28–31] to eliminate shadows is based on intensity analysis. This method suggests that the Stauffer–Grimson algorithm is applied on the normalized R_nG_nI color space, where I is the intensity component calculated as the average of the R, G, and B components. The pixels belonging to shadow or to highlighted regions are found by thresholding the ratio of $r = I_c/I_m$, where I_c is the current frame's intensity and I_m is the mean intensity over the sequence of frames. However, we found that the utilization of the normalized R_nG_nI color space removes the pixels with low intensity (e.g., gray or dull white color) along with the shadows [Fig. 3(b)]. Thus to avoid the loss of regions of interest, we decided to utilize the RGB color space [Fig. 3(c)].

Inspired by the work of Javed et al. in Ref. 33, we attempt to recover the missing foregrounds due to similar background and remove the shadows and other spurious objects by exploiting the image gradient information. As in Ref. 33, we apply the Stauffer–Grimson algorithm on the gradient magnitude image I_g to get the foreground edge information $E_{\text{forg}}(I_g)$) [Fig. 4(c)]. Regarding the detected spurious objects that are not part of the true foreground and to address the illumination variation problem, we apply region-level processing as proposed in Ref. 33. However, the authors in Ref. 33 do not use the foreground edges $E_{\text{forg}}(I_g)$ to define the detected region except in the case of illumination changes. In our application, we utilize $E_{\text{forg}}(I_g)$ to its full extent.

The primary goal of the proposed approach is the accurate mapping of the contours of foreground regions. If the RGB colored frame is I, then let the extracted foreground using Stauffer–Grimson method[27] be $F(I)$ and the binary edge map of the image I be $E(I)$. Morphological dilation and closing operations are applied to achieve continuous edges in $E_{\text{forg}}(I_g)$ and $E(I)$.

(a) (b) (c)

Fig. 3 The effect of shadow removal on foreground segmentation. Image (b) illustrates the extracted foreground with shadow removal based on intensity. Image (c) illustrates the results of MoG on *RGB* color space as implemented in Ref. 32.

Fig. 4 The procedure of accurate contour definition of the foreground regions. (a) Original frame. (b) Dilated and closed edges $M[E(I)]$. (c) Foreground of gradient image $G(I_g)$. (d) Edges corresponding to foreground regions $C(I)$. (e) Final output.

At the next stage, a convex hull is computed to enclose $E_{\text{forg}}(I_g)$. The convex is used as a reference to discard all the pixels in $F(I)$ that lay out of the convex, thus achieving the removal of shadow regions that do not have strong gradients. Conclusively, the smooth contour of the foreground region C [Fig. 4(d)] is the pixels of $E(I)$ present in the sum of $F(I)$ and $E_{\text{forg}}(I_g)$. The accurately defined smooth contour is

$$C = E(I) \times [F(I) + E_{\text{forg}}(I_g)]. \tag{5}$$

The regions enclosed in the contour C are considered to belong to the foreground [Fig. 4(e)]. Afterward, median filtering is applied to the detected foreground to smooth the edges. The advantage of the above method is the detection of well-defined contours of the foreground, as the original output of MoG-based algorithms most of the time fails to deliver an accurately defined region. On the other hand, the convex hull does not always enclose the whole object, resulting in broken regions.

3.4 *Feature Extraction for Human Object Detection*

In this section, we discuss the proposed methodology of extracting features from the segmented object, which can discriminate between humans and nonhumans.

3.4.1 *Contourlet transform*

Contourlet transform was proposed by Do and Vetterli[34] in the discrete form as a simple directional extension for wavelets. In addition to multiscale and time-frequency localization, contourlet transform also offers a high degree of directionality and anisotropy. It provides improvements to two-dimensional separable wavelet transforms for representing images with smooth contours in all directions (see Fig. 5). The contourlet transform as explained by Pan et al.[35] allows for different and flexible number of directions at each scale. It decomposes an image into several directional sub-bands at multiple scales in the frequency domain by first applying a Laplacian pyramidal multiscale decomposition to capture the point discontinuities, and this is followed by a critically sampled directional filter bank applied on each undecimated high-frequency band to link point discontinuities into linear structures. The overall result is an image expansion using basic elements like contour segments.

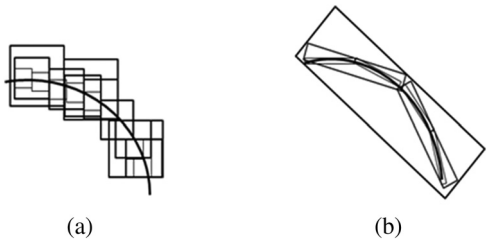

Fig. 5 Illustration of successive refinement by (a) wavelet and (b) contourlet near a smooth contour.[34]

Humans being objects of nature (i.e., not man-made) are made up of a large collection of curved features, e.g., edges. Therefore, contourlet transform that provides a near optimal representation of objects with curves are better suited to represent humans than other transforms such as wavelets, which are limited to using vertical and horizontal line segments to represent an edge.

3.4.2 *Local binary patterns*

LBP belongs to the class of nonparametric local image features that spatially exploits the geometric properties of a pattern and provides gray scale and rotation invariant texture features. The initial version of LBP introduced by Ojala et al.[36] labels the pixels of an image by thresholding a 3×3 neighborhood of each pixel with center value, where 0 is assigned for a negative difference between the center pixel and the neighboring pixel and 1 is assigned for a positive difference. A histogram that gives the distribution of $2^8 = 256$ binary patterns obtained from this procedure can be used as texture feature for subsequent analysis. The value of an $\text{LBP}_{P,R}$ code that takes P sample points with radius R around a pixel (x_c, y_c) is given by

$$\text{LBP}_{P,R} = \sum_{p=0}^{P-1} s(v_p - v_c)2^p \quad s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where v_c and v_p are the gray scale values of the center pixel (x_c, y_c) and surrounding pixel (x_p, y_p). Ojala later made an extension of the original operator to make LBP gray scale and rotation invariant[37] by considering a circular neighborhood and bilinear interpolation. Gray scale invariance is achieved by subtracting the gray value of the center pixel with a gray value of the circular-symmetric neighborhood and by only considering the sign instead of their exact gray values. The required rotation invariance of LBP ($\text{LBP}^{ri}_{P,R}$) is obtained by considering the minimum value of binary pattern. The minimum value is produced by shifting binary structure to put a maximum number of zeros at the beginning of the binary pattern. It is observed that certain binary patterns that have 0 to 1 or vice versa transitions are not more than a certain limit n contains most of the texture information and often their frequency of occurrence is >90%. These patterns are called uniform patterns $\text{LBP}^{\text{ruin}}_{P,R}$ and provide the reduced dimensionality of the feature vector.

Hence a small subset of the whole 2^P patterns is sufficient to describe the texture of images. We have taken advantage of this technique to represent the texture in the contourlet sub-bands.

3.4.3 *Contourlet-based local binary patterns: proposed feature descriptor*

The human body in a standing position has strong vertically oriented contours and edges along its boundaries. Further, the spatial structure of the human body has bilateral symmetry. Therefore, details in the contour are one of the most relevant features for identifying a human body. LBP is suitable for modeling repetitive textures, which means these features are sensitive to random noise in uniform image areas. LBP extraction in contourlet domain helps to analyze contours while reducing the effect of that noise.

We make use of contourlet transform as a multiscale and multidirectional global image contour representation and local binary pattern ($\mathrm{LBP}^{\mathrm{riu2}}_{P,R}$) to describe texture around these contours. We have named this object representation as contourlet local binary patterns (CLBP). For a segmented foreground object image (x, y), CLBP is constructed as explained below.

First, we perform contourlet transform to $I(x, y)$ with a prescribed decomposition level and directional decomposition at each scale level to obtain the transformed image $I_c(x, y)$ in the contourlet domain. Band selection plays an important role in extracting features that are more discriminative and computationally less expensive. A total of 14 bands with information in vertical and diagonal orientations are selected (i.e., second half of bands from each high-frequency decomposition levels). The reason is that human body in its standing position exhibits more strong vertical and diagonal edges along the contour of body than horizontal edges. This characteristic makes humans distinguishable from nonhuman objects. The high-frequency contourlet bands of a human foreground image are shown in Fig. 6.

We apply rotation invariant uniform LBP with $P = 8$ and $R = 1$ to obtain a 59-dimensional feature vector for each of these sub-bands. Finally, these feature vectors are concatenated to produce a $59 \times 14 = 826$-dimensional feature vector for an image $I(x, y)$. Figure 6 depicts the steps of extracting the CLBP feature vector.

3.4.4 *Human and nonhuman classification using support vector machine*

In order to classify the foreground objects, we have used a nonlinear radial basis function (RBF) kernel to train the support vector machine (SVM) model using a large number of CLBP descriptors that have been computed as positive and negative training images. The best combination of C and γ has been selected by a grid search with exponentially growing sequences of C and γ.

3.5 *Human Figure Annotation*

The annotation of a detected human figure is carried out in a few stages to extract various details. At the initial stage, three main parts (i.e., head, torso, and limbs) of the body are separated. This is followed by a color extraction stage where the dominant colors in each portion of the body are extracted to analyze the clothes. Finally, clothes of human figures are analyzed for the presence of text/logo.

3.5.1 *Body parts classification*

A simple approach has been adopted here to categorize the various parts of human body. The human figure is divided into three parts based on the assumption that human figures

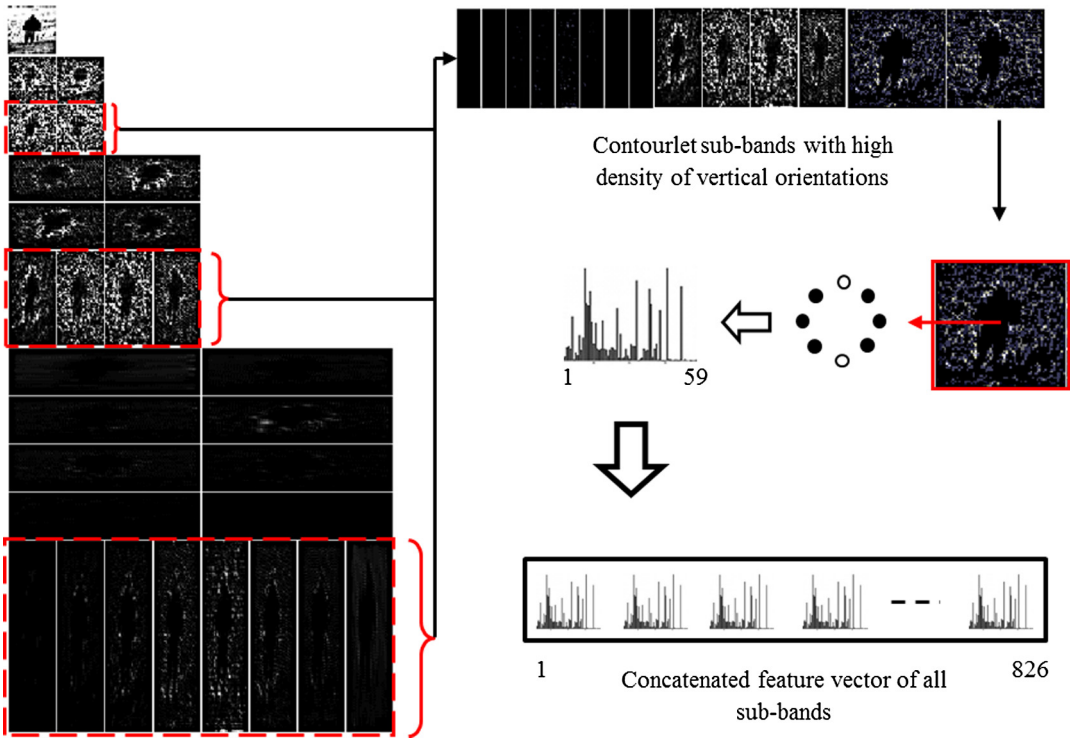

Fig. 6 Contourlet sub-bands and CLBP descriptor calculation.

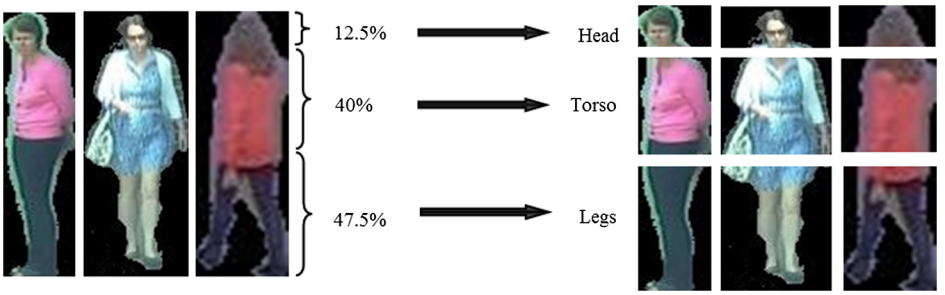

Fig. 7 Illustration of body parts categorization.

are in their standing position (i.e., foreground humans are those who are walking). The upper 12.5% of the figure is categorized as head region. The next 40% is extracted as the torso, and the remaining 47.5% of the body is taken as the leg region of the detected human. These values have been selected by empirical analysis of a large number of extracted human figures. The categorized human body parts are passed on to the color extraction stage for color analysis, and the torso region is further analyzed for text/logo detection and recognition. Figure 7 illustrates the implementation of the explained body parts classification technique.

3.5.2 *Color extraction*

Color extraction is carried out by histogram quantization. The image is transformed from *RGB* into HSV color space. HSV representation is perceptually more relevant as compared to Cartesian representation due to the fact that it can be quantized into meaningful colors. *RGB* colors are difficult to define and distinguish because of their nonabsolute space. To compute the color histogram, HSV space is divided into six hue, five saturation, and five value bins. The angular cut-offs in the hue separates six primary colors, whereas saturation and value bins distinguish between various shades of primary colors. The framework is able to distinguish between 150 ($6 \times 5 \times 5$) shades of colors, which is a reasonable range to specify the clothes' color of a human in surveillance videos. Every frame has to pass through a color correction framework where the effect of variable illumination and camera calibrations are reduced and the actual object colors are enhanced to achieve better color extraction.

The highest histogram peak (local maximum) is considered as the dominant color of the object. However, we are not relying only on the highest peak for the color extraction of an object. The probability of false color extraction becomes high due to inaccurate foreground segmentation or the presence of occlusion. Therefore, we choose the three highest peaks as the dominant colors of an object [by object we mean a single part (i.e., head, torso, or leg region) of the human body]. Three colors for each part result in nine colors associated with a human figure.

3.5.3 *Text/logo detection*

In Ref. 38, we presented a local energy–based shape histogram (LESH) feature-based technique to detect text and logo in natural scene images. The same technique has been adopted here to detect text/logo present on the shirts or tops of human figures to use it as another parameter to annotate a given human figure. Text/logo detection framework operates by extracting connected components of enhanced horizontal, vertical, and diagonal edges in the video frames. LESH feature vectors of these candidate regions are calculated using the technique presented in Ref. 7 and are classified using an SVM model.

Candidate extraction. The candidate extraction procedure starts by edge information enhancement in the input torso region window using the Sobel edge detector. Image dilation is applied at four major angles, 0, ± 45, ± 90, and ± 135 deg, to make these orientations stronger as text information contains a high frequency of edges in these orientations. The dilated image is further processed to acquire connected components as candidate regions for text/logo. The extracted candidate regions are passed through an initial screening to filter out straightforward nontext regions. The screening procedure checks the following conditions to filter out nontext regions:

a. If the number of rows or columns of a connected component is $>75\%$ of the image rows or columns, respectively, then that particular connected component is less likely to be a text/logo.
b. If the block height or width is too small, then there is a high possibility of nontext information in that connected block.
c. If edge density of a certain block is less than a specific threshold, then that block is not a promising candidate of text/logo.

Feature extraction. After initial screening, successful candidate regions are represented using their LESH description. Each candidate region is resized to a 32×32 segment and the LESH descriptor is calculated.

LESH features have been developed by Sarfraz et al.[39] for facial feature extraction. We have selected these features for text/logo detection due to their insensitivity to illumination and other common variations that occur frequently in text regions. Since local energy signifies the underlying corners, edges, or contours, this information is significantly helpful for the identification of text regions. A local histogram accumulating the local energy along each filter orientation on different subregions of the image is generated. The local histograms are extracted from different subregions of the image and then concatenated together. An orientation label map is obtained where each pixel is assigned the label of the orientation at which it has the largest energy across all scales. Figure 8 illustrates the LESH vector extraction of

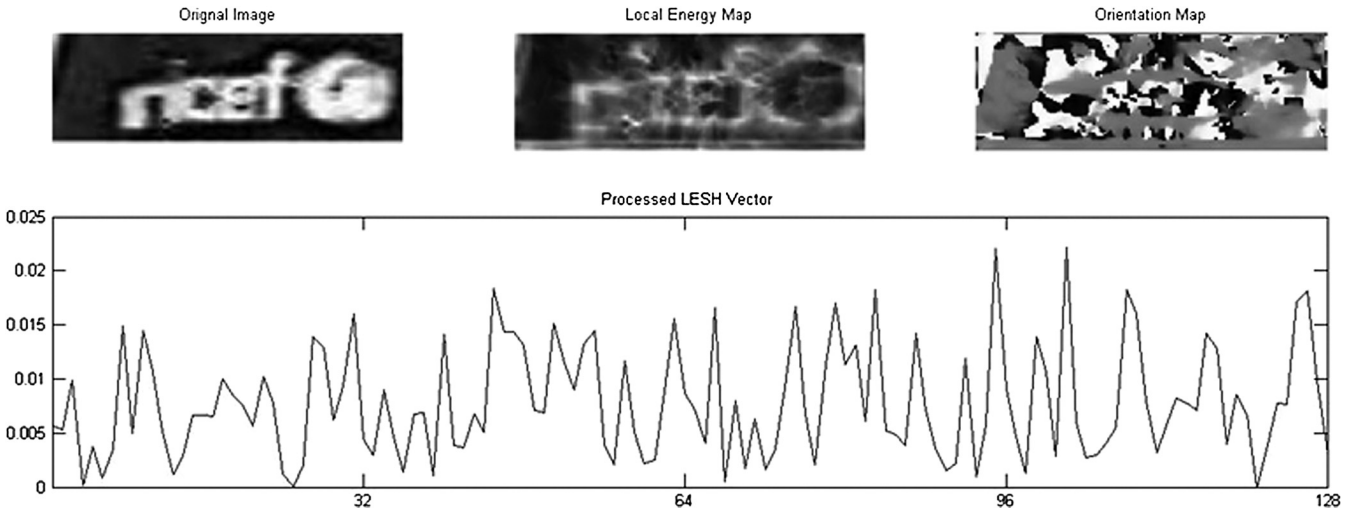

Fig. 8 LESH vector extraction of a candidate image along with local energy map and orientation map.

a candidate region along with the local energy maps and orientation map.

Classification. The candidates are classified as text/logo or nontext/logo using SVM. An SVM model is trained using the LESH descriptions of a training image set. A nonlinear RBF kernel has been used to train the SVM model using the CLBP descriptor of positive and negative images. The best combination of C and γ has been selected by a grid search with exponentially growing sequences of C and γ. The decision about an input test region is finalized by the probability estimate calculated by SVM.

4 Experiments, Results, and Analysis

This section presents the experimental results and the analysis of proposed methodologies for human object annotation presented in the previous sections.

4.1 *Foreground Object Extraction*

In our experiments, we have compared our method with the one proposed by KaewTraKulPong and Boweden in Ref. 32, the implementation of which is publicly available in OpenCV 2.2. According to them, the initialization of the background requires the provision of four parameters: the number of Gaussians in the mixture (K), the portion of distributions accounted for by the background (T), the initial standard deviation (σ), and the learning rate (α). The values of parameters in our experiments were assigned as follows: $K = 4$, $T = 0.55$, $\sigma = 30$, $\alpha = 0.004$ for the color frame processing and $K = 2$, $T = 0.45$, $\sigma = 40$, $\alpha = 0.03$ for the gradient frame processing. The learning rate values for the two background models were set in such a way that if a foreground object is not in motion any more, it is absorbed by the background of the two models at the same time. The foreground segmentation process has been performed on frames that have *RGB* components ranging from 0 to 255. It should be noted that the same parameter values ($K = 4$, $T = 0.55$, $\sigma = 30$, $\alpha = 0.004$) were used in the implementation, and blob detection was applied in both cases to remove significantly small particles corresponding to noise.

In this section, we present the experimental results conducted on the CAVIAR[40] dataset, CCTV footage from DIRG dataset, and ChangeDetection.net dataset.[41] To compute the metrics that measure the accuracy of the system, we have randomly selected frames from DIRG dataset and manually labeled the ground truth. The frames selected from the ChangeDetection.net dataset belong to cases "pedestrians" and "backdoor." As depicted in Tables 1 and 2, the proposed method improves all the metrics including the F-measure that shows the amount of overall improvement. The false positive rate (FPR) and false negative rate (FNR) have been decreased showing that a large amount of shadow has been removed and at the same time the incomplete foreground has been enclosed in the detected contour. High recall and specificity values indicate that the true positive rate (TPR) and true negative rate (TNR) have been improved without affecting the precision. The results in Table 2 are statistically significant as they are derived from a large number of sample frames. However, we selected to display the results from the other two datasets to show the extent of improvement achieved on the videos that are used for further processing and human annotation.

A visual comparison of the methods in Fig. 9 reveals that the implementation of the method in Ref. 32 is prone to incomplete and inaccurate foreground segmentation and

Table 1 Comparative metrics for the proposed method and the method presented in Ref. 32.

	Prec.	Rec.	F-measure	Spec.	FPR	FNR
Proposed method	0.9658	0.9812	0.8110	0.0188	0.0029	0.8817
Method as proposed in Ref. 32	0.8391	0.9798	0.7429	0.0202	0.011	0.7881

Table 2 Comparative metrics for the proposed method and method proposed in Ref. 32 on "ChangeDetection.net datasets."

		Prec.	Rec.	F-measure	Spec.	FPR	FNR
Proposed method	Pedestrians Frames: 302 to 1048	0.9647	0.9995	0.9435	0.0005	0.0003	0.9540
	Backdoor Frames: 3 to 1998	0.9126	0.9994	0.9659	0.0006	0.0015	0.9385
	Average	**0.9387**	**0.9995**	**0.9547**	**0.00055**	**0.0009**	**0.9463**
Method as proposed in Ref. 32	Pedestrians Frames: 302 to 1048	0.9500	0.9997	0.9594	0.0003	0.0004	0.9546
	Backdoor Frames: 3 to 1998	0.9073	0.9864	0.5317	0.0136	0.0016	0.6705
	Average	**0.9287**	**0.9931**	**0.7456**	**0.0070**	**0.001**	**0.8126**

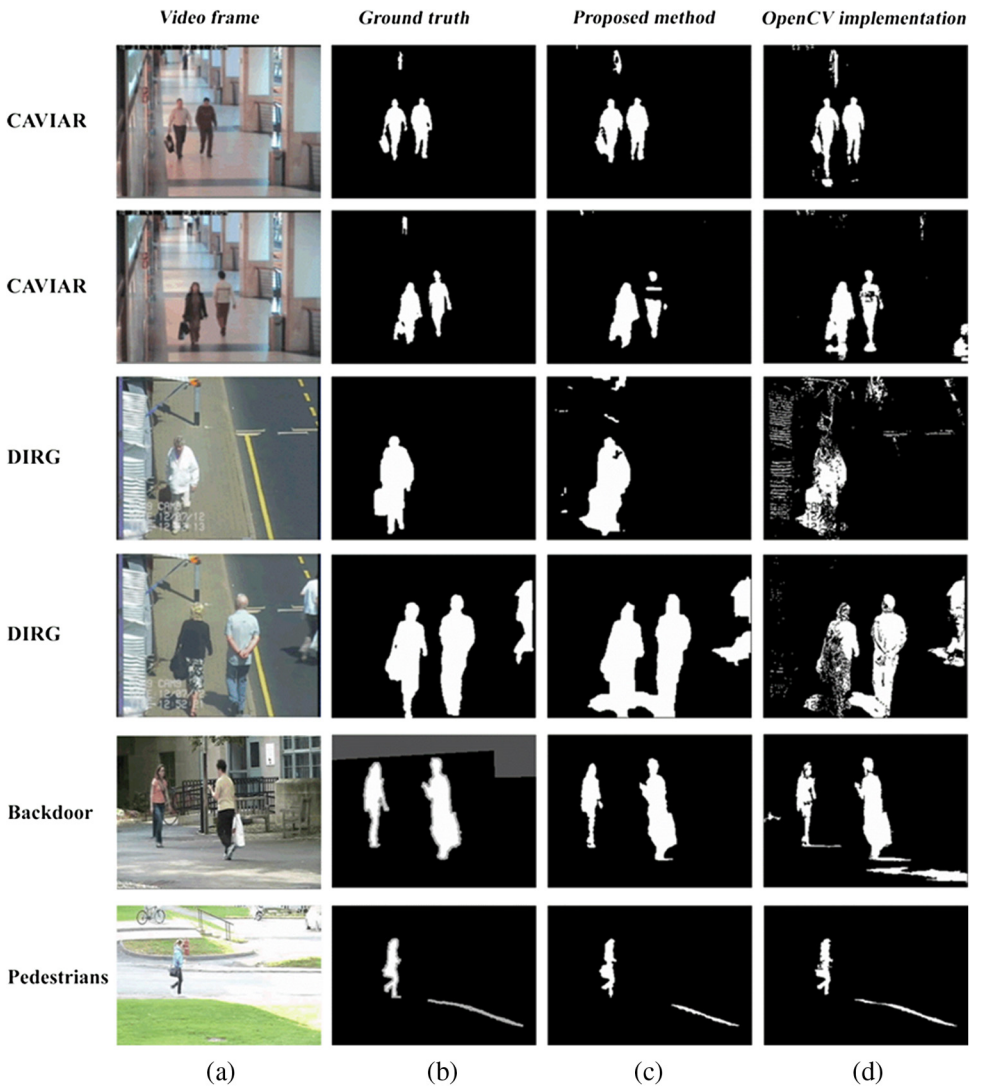

Fig. 9 Visual comparison of the proposed foreground segmentation method with the method proposed in Ref. 32. (a) Video frames. (b) Ground truth. (c) Proposed method. (d) Method presented in Ref. 32.

shadow detection. In the first two examples, it is shown that the gradient-based method is capable of removing shadows which do not form a distinctive outline. On the other hand, in cases such as two rows within the range of shadow edges, the removal is not possible with the proposed method. The examples from the "backdoor" sequence (see Fig. 9) demonstrate elimination of shadows and false foreground regions that are caused by illumination changes. At the same time, the objects appear connected and the gaps are filled. As our major objective is to achieve complete foreground

segmentation, false detections have to be tolerated up to some extent. Another weakness of the proposed system is presented in the second example where the lack of gradient information (i.e., similarity of the foreground-background) results in incomplete foreground segmentation.

4.2 *Human Identification*

In order to test the human identification module, we have carried out two sets of experiments. In the first experiment, we study the effect of selecting various sub-bands from contourlet decomposition on the performance of CLBP descriptor. In comparison to the baseline detector (HOG), the second set of experiments show the performance of contourlet-LBP feature descriptor on the publically available CAVIAR dataset. We have used the DIRG dataset for training and the CAVIAR dataset for testing the performance of the proposed human detection framework. Thev DIRG dataset has been collected from outdoor surveillance videos acquired by a CCTV surveillance center and is not publically available at present.

4.2.1 *Contourlet-LBP detector*

We study the effects of different choices of orientations and decomposition levels of the contourlet coefficients. The density of vertical edges is significantly high in a human figure; therefore we give more emphasis to those sub-bands that highlight vertically oriented edges in the image. Choosing the right set of sub-bands is essential for the efficient performance of contourlet-LBP. In the proposed method, we decompose the image with 4, 8, and 16 directional sub-bands at three decomposition levels, respectively.

As also noted in the original contourlet proposal,[31] an l-level directional filter tree decomposition can be viewed as a 2^l parallel channel filter bank with equivalent filters and sampling matrices. This corresponds to separable sampling matrices grouped in two ranges of the l-level directional sub-bands in each pyramid decomposition level. Accordingly, the two sets correspond to 0 to 2^{l-1} and 2^{l-1} to 2^l sub-bands in each pyramid decomposition level. These two sets convey the mostly horizontal and mostly vertical set of directions. In our case, this translates that at each pyramid level, the second half of the corresponding directional sub-bands would convey the mostly vertical directions. Apart from sub-bands, the size selection of the test image is also important for the improvement of detection results. Our experimental results confirm this and show that the best performance is achieved when the second half of sub-bands from the three levels contourlet decomposition are selected using an image size of 256×256 (see Fig. 10).

4.2.2 *Detection results with contourlet-LBP feature*

We have used CLBP as the feature vector ($14 \times 59 = 826$ dimensions) and SVM (RBF kernel with parameter values $C = 103$, $g = 0.625$) as the classifier for detection of human figures in the video clips of the CAVIAR dataset. The system has been trained using 2000 training images (1000 positive and 1000 negative) taken from DIRG dataset.

Receiver operating characteristics curves in Fig. 11 show that CLBP-SVM has outperformed the baseline (HOG-SVM) technique. As shown in Table 3, the proposed method has achieved high figures of precision, recall, specificity, and f-measure, which indicate its superior performance. The FPR and FNR have been decreased, showing the reduced number of misses and false alarms. High recall and specificity values indicate that the TPR and TNR have been improved without affecting the precision.

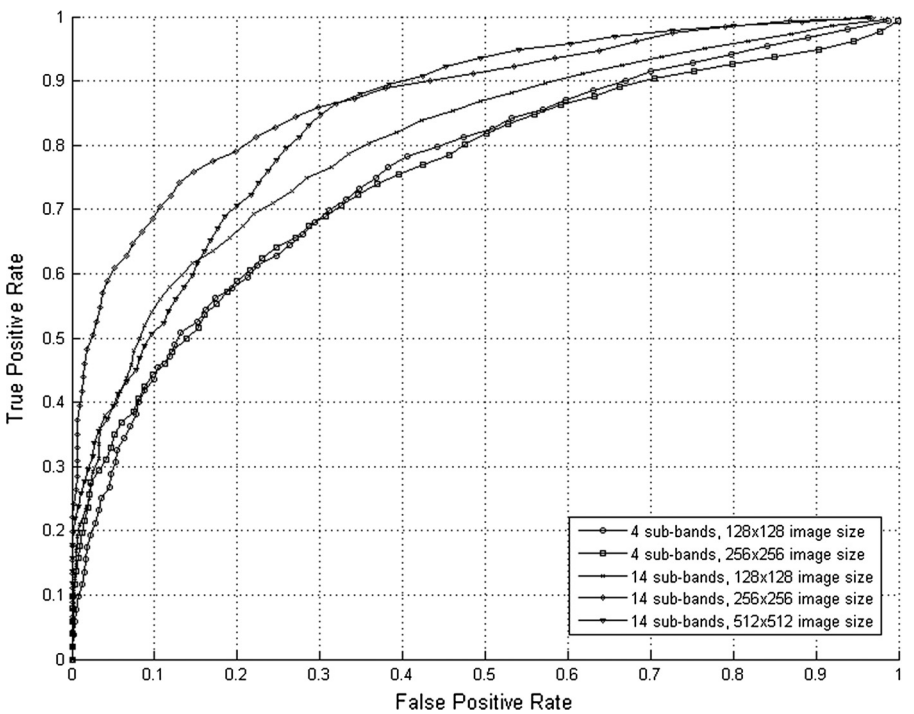

Fig. 10 The performance comparison of various band selections and sizes of input candidates. For the first two graphs, four sub-bands were selected (two from first contourlet decomposition level and two from third contourlet decomposition level), which contain the majority of vertical contours. For the rest of the graphs, all vertical sub-bands were selected from first, second, and third contourlet decomposition levels.

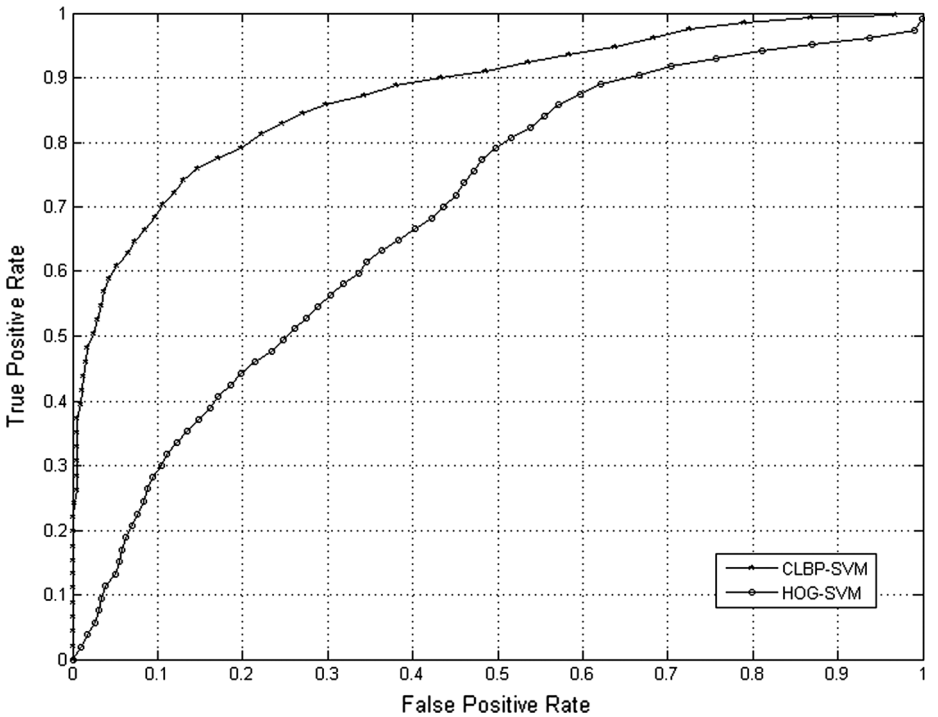

Fig. 11 The performance comparison between proposed human detection technique (CLBP-SVM) and baseline technique (HOG-SVM).

Table 3 Comparative metrics for the proposed method and the baseline.

	Precision (PPV)	Sensitivity (recall)	F-measure	Accuracy	Specificity	FPR	FNR
HOG-SVM (baseline)	0.433	0.810	0.942	0.5932	0.479	0.190	0.567
CLBP-SVM (proposed method)	0.784	0.818	0.970	0.8716	0.788	0.182	0.216

Fig. 12 Color extraction results for various human figures.

Table 4 Confusion matrix for different color classifications.

	Black	Blue	Green	Gray	Red	Pink	White	Yellow	Brown
Black	**0.9693**	0.0563	0.0434	0.106	0.0226	0.0127	0.0123	0.0343	0.0722
Blue	0.0047	**0.8502**	0.0221	0.0318	0	0	0.0076	0	0
Green	0.0097	0.0474	**0.8727**	0.0167	0	0	0.0054	0	0
Gray	0.0163	0.0461	0.0316	**0.807**	0.0310	0	0.0148	0.0021	0.0084
Red	0	0	0	0	**0.8932**	0.0269	0	0	0.0012
Pink	0	0	0	0	3.24	**0.9183**	0	0	0
White	0	0	0.0223	0.085	2.08	0.0421	**0.9518**	0.61	0.0143
Yellow	0	0	0.0079	0	0	0	0.0081	**0.8957**	0.0359
Brown	0	0	0	0	0	0	0	0.0618	**0.8680**

4.3 *Color Extraction*

The color extraction module has been tested on a set of 227 human figure images extracted from surveillance videos of DIRG dataset using the proposed foreground extraction technique. Head, torso, and leg regions have been labeled using the technique explained in Sec. 3.5.2, and the proposed color extraction technique is applied on labeled regions of the human figure. The use of HSV histogram quantization has worked well along with the idea of extraction of three dominant colors for each body part. Surveillance videos have poor brightness and contrast, and the accuracy of color extraction has been improved by applying the proposed color enhancement technique. Figure 12 demonstrates the results of color extraction after the color correction stage.

Table 4 presents a confusion matrix demonstrating the accuracy of the technique for the recognition of various colors. The accuracy for the recognition of all the colors is noticeably good.

The idea of extracting three dominant colors instead of one for each region of human body significantly reduces the chances of missing the real color of that region. Table 5 presents the accuracy at first, second, and third dominant color extraction for each part of a human figure. It is evident from the results that the accuracy is not high if the biggest cluster is always considered as the actual color of object. The chances for achieving the actual color of object in the second and third biggest clusters are significantly high.

Table 5 Relative color extraction accuracy of head, torso, and legs region as first, second, and third dominant color.

	First dominant color	Second dominant color	Third dominant color
Head	0.40	0.86	0.927
Torso	0.66	0.926	1.00
Legs	0.66	0.86	0.927

4.4 *Text/Logo Detection*

For the evaluation of our approach, we collected a new dataset from a collection of surveillance videos containing people with shirts and tops having text/logos on them. The dataset contains 178 test images with human figures in them wearing various colored plain shirts or shirts having text/logos on them. The dataset further contains 500 positive and 500 negative training images that have been extracted from surveillance videos using the connected component-based candidate extraction technique explained in Ref. 38. In order to increase the size of training set, we combined our training data with the ICDAR2003[42] training dataset. Training images of ICDAR2003 have been processed through connected component-based candidate extraction technique to extract positive and negative training images.

The performance of text/logo detection modules has been analyzed using standard metrics. ROC curves in Fig. 13 illustrates true acceptance and true rejection rates that have been achieved using proposed technique. Table 6 shows that the proposed text/logo detection technique has achieved a recall rate of 54.1%, which demonstrates a high miss rate. The key reason for achieving such a low sensitivity is the poor quality of the test videos and the presence of significantly small-sized text/logos on the subject of interest, which is human in this case. Test videos have been recorded using an ordinary CCTV camera at a resolution of 800×600 from a distant, high point. Human subjects appear as occupying a small portion of the frame and the presence of any text/logo on their clothes is difficult to perceive even by naked eyes. Apart from the size, the text/logo in most scenarios is either occluded by wrinkles or amalgamated with the texture of the clothes, which makes it extremely difficult for the system to distinguish between a text/logo region and the texture. A few samples of detection results have been presented in Fig. 14.

Another reason of achieving low precision and recall for text/logo detection is the processing of text and logo as a combined, single class. Text and logo regions have different visual properties, and thus putting them in one class makes the classification model less specific to any one of them.

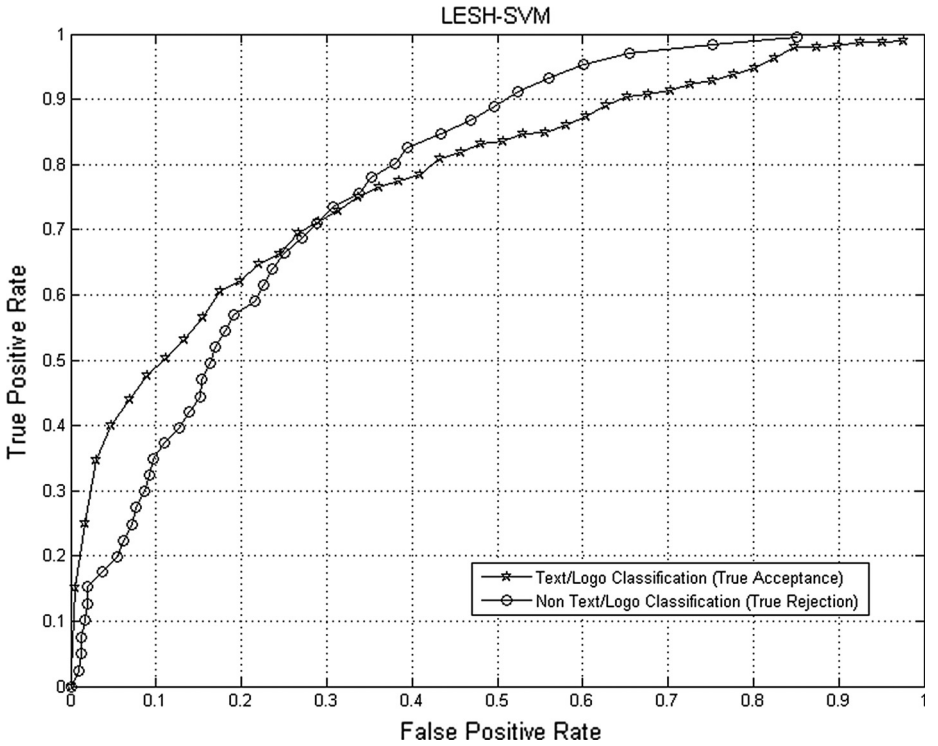

Fig. 13 ROC curve showing TPR versus FPR for text/logo detection.

Table 6 Performance metrics for text/logo detection on clothes of human figures.

	Precision (PPV)	Sensitivity (recall)	F-measure	Specificity
LESH-SVM (proposed method)	0.541	0.860	0.349	0.424

Experimental results also show that the connected component-based region extraction fails to locate a text candidate in cases where clothes have wrinkles near text/logo regions, the color contrast between the text/logo region and the background is low, the background of the shirts have textured print, and lighting at text/logo regions is inconsistent.

5 Conclusion

We have presented an image processing pipeline for human detection and detailed annotation for the purpose of CCTV forensics. The proposed system initially makes use of a novel combination of contourlet transform and LBP features to detect humans passing through the surveillance camera. Subsequently, the human figures are annotated on the basis of the color of clothes and presence of text/logo on them. A color correction pipeline followed by histogram quantization in the HSV color space has been applied to extract color information of the subjects. The presence of text/logo has been detected by making use of LESH features. The proposed system showed good performance for identification and annotation of human subjects within surveillance videos when tested on a comprehensive set of test CCTV footage. The accuracy of human detection is affected in crowded scenarios where subjects move in groups and are occluded. This

Fig. 14 Text/logo detection results. (a) to (e) Successful detections. (f) and (g) False detections.

could be improved by incorporating sophisticated segmentation algorithms on connected foreground blobs to separate them. In addition, tracking information can help to overcome issues related to occluded humans. The performance of text/logo modules can be further improved by applying a sliding window-based approach to candidate region extraction. The low quality of CCTV videos poses a challenge and demands more research in this area for accurate annotation and indexing of these videos.

References

1. C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.* **38**(1), 15–33 (2000).
2. P. Viola and M. Jones, "Detecting pedestrians using patterns of motion and appearance," in *IEEE Int. Conf. on Computer Vision*, Vol. 2, pp. 734–741 (2003).
3. D. G. Lowe, "Distinctive image features from scale invariant keypoints," *Int. J. Comput. Vis.* **60**(2), 91–110 (2004).
4. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Vol. 1, pp. 886–893 (2005).
5. 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).
6. X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *IEEE Int. Conf. on Computer Vision,* pp. 32–39 (2009).
7. Y. Zheng, C. Shen, and X. Huang, "Pedestrian detection using center-symmetric local binary patterns," in *IEEE Int. Conf. on Image Processing*, pp. 3497–3500 (2010).
8. D. Kim et al., "Human detection using wavelet-based CS-LBP and a cascade of random forests," in *IEEE Int. Conf. on Multimedia and Expo*, pp. 362–367 (2012).
9. D. M. Hansen et al., "Automatic annotation of humans in surveillance videos," in *Proc. IEEE Int. Conf. on Computer and Robot Vision* **63**(2), pp. 103–115 (2005).
10. R. Schettini et al., "A survey on methods for color image indexing and retrieval in image databases," in *Color Imaging Science: Exploiting Digital Media*, pp. 1–9, Media, John Wiley (2001).
11. L. M. Brown, "Color retrieval for video surveillance," in *Proc. of 5th IEEE Int. Conf. on Advanced Video and Signal Surveillance*, pp. 283–290 (2008).
12. M. Swain and D. Ballard, "Color indexing," *Comput. Vis.* **7**(1), 11–32 (1991).
13. M. Fraz, I. Zafar, and E. A. Edirisinghe, "Object color extraction for CCTV video annotation," in *Proc. of 8th Int. Conf. on Computer Vision Theory and Applications*, pp. 455–459 (2013).
14. G. Wui et al., "Identifying color in motion in video sensors," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 561–569 (2006).
15. J. V. Weijer, C. Schmid, and J. Verbeek, "Learning color names from real-world images," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8 (2007).
16. Y. Zhang et al., "Object color categorization in surveillance videos," in *Proc. of IEEE 18th Int. Conf. on Image Processing*, pp. 2913–2916 (2011).
17. J. Matas et al., "Robust wide baseline stereo from maximally stable extremal regions," *Proc. of British Machine Vision Conf.*, pp. 384–396 (2002).
18. Y. Jin and S. Geman, "Context and hierarchy in a probabilistic image model," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2145–2152 (2006).
19. A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondence," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 26–33 (2005).
20. J. Weinman and E. Miller, "Improving recognition of novel input with similarity," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 308–315 (2006).
21. X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Vol. 2, pp. 366–373 (2004).
22. P. Viola and M. J. Jones, "Robust real-time face detection," *Comput. Vis.* **57**(2), 137–154 (2004).
23. M. Lalonde and L. Gagnon, "Key-text spotting in documentary videos using adaboost," *Proc. SPIE* **6064**, 60641N (2006).
24. G. Buchsbaum, "A spatial processor model for object color perception," *J. Franklin Inst.* **310**(1), 1–26 (1980).
25. J. M. Buenaposada and L. Baumela, "Variations of grey world for face tracking," *Image Process. Commun.* **7**(3–4), 51–62 (2001).
26. C. A. Poynton, *Digital Video and HDTV: Algorithms and Interfaces*, pp. 260–630, Morgan Kaufmann Publishers, San Francisco (2003).
27. C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Vol. 2, pp. 246–252 (1999).
28. H. Wang and D. Suter, "Tracking and segmenting people with occlusions by a sample consensus based method," in *Proc. of IEEE Int. Conf. on Image Processing*, Vol. 2, pp. 410–413 (2005).
29. K. Schindler and H. Wang, "Smooth foreground-background segmentation for video processing," in *Proc. of the 7th Asian Conf. on Computer Vision*, pp. 581–590 (2006).
30. A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *IEEE Frame-Rate Workshop*, pp. 751–767 (2000).
31. H. Wang and D. Suter, "A re-evaluation of mixture of Gaussian background modeling [video signal processing applications]," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 1017–1020 (2005).
32. P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proc. of European Workshop on Advanced Video Based Surveillance Systems*, Vol. 2, pp. 135–134 (2001).
33. O. Javed et al., "A hierarchical approach to robust background subtraction using color and gradient information," in *Proc. of the Workshop on Motion and Video Computing, MOTION*, pp. 22, IEEE Computer Society, Washington, DC (2002).
34. M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Trans. Image Process.* **14**(12), 2091–2106 (2005).
35. H. Pan et al., "Contourlet-based feature extraction for object recognition," *Proc. SPIE* **7495**, 749522 (2009).
36. T. Ojala, M. Pietikainen, and T. Maenpaa, "A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification," in *2nd Int. Conf. on Advances in Pattern Recognition*, pp. 397–406 (2001).
37. T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patters," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002).
38. M. Fraz and M. S. Sarfraz, "Text detection in still images and CCTV videos using Local Energy based Shape Histogram (LESH)," in *Proc. of 6th Int. Conf. on Computer Vision Theory and Applications*, pp. 433–436 (2011).
39. M. S. Sarfraz and O. Hellwich, "Head pose estimation in face recognition across pose scenarios," in *Proc. of Int. Conf. on Computer Vision Theory and Applications*, pp. 235–242 (2008).
40. R. Fisher, "PETS04 surveillance ground truth data set," in *IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 319–336 (2004).
41. N. Goyette et al., "Changedetection.net: a new change detection benchmark dataset," in *Proc. IEEE Workshop on Change Detection (CDW-12) at CVPR-12*, pp. 16–21 (2012).
42. S. M. Lucas et al., "ICDAR 2003 robust reading competitions: entries, results and future directions," *Int. J. Doc. Anal. Recogn.* **7**(2–3), 105–122 (2005).

Muhammad Fraz is a PhD research student in the Department of Computer Science at Loughborough University, United Kingdom. He received his BScEng (hons) degree in computer engineering in 2006 from COMSATS Institute of Information Technology (CIIT), Lahore, Pakistan, and his MSc degree in embedded digital systems from Sussex University, United Kingdom, in October 2008. His PhD research focuses on automatic content-based surveillance video annotation. Before starting PhD, he was a part of Computer Vision Research Group at CIIT, where his research significantly focused on detection and recognition of text in images and videos.

Iffat Zafar Iffat Zafar is working as a senior researcher at Apical Ltd., United Kingdom. She completed her MSc in multimedia and Internet computing in October 2004 and PhD in computer science in 2008 from Loughborough University. Before joining Apical, she was a part of Digital Imaging Research Group at Computer Science Department of Loughborough University as a postdoctoral research associate, where she worked on various industrial sponsored

projects. Her research interests include machine learning, pattern recognition, visual tracking, and automated surveillance.

Giounona Tzanidou is a PhD research student in the Department of Computer Science at Loughborough University. She has studied applied informatics at the University of Macedonia, Greece, and received her diploma in 2009. She obtained her MSc degree in Internet computing and network security from Loughborough University in October 2010, and her MSc project focused on baggage detection. Afterward, carried object detection in videos became a considerable part of her PhD research.

Eran A. Edirisinghe obtained his BScEng (hons) degree from University of Moratuwa, Sri Lanka, in 1994. He completed his MSc and PhD degrees at Loughborough University in 1996 and 1999, respectively. He joined Loughborough University as a lecturer of computer science in July 2000 and was promoted to a senior lecturer in 2004. He was awarded the title of Reader in Digital Imaging, Loughborough University in April 2008. He was appointed as the head of Department of Computer Science, Loughborough University in August 2011 and was promoted to a chair in July 2012. His research has been funded by the Engineering and Physical Science Research Council (EPSRC), Technology Strategy Board (TSB), and United Kingdom (UK) industry.

Muhammad Saquib Sarfraz is a postdoc at Computer Vision for Human Computer Interaction at Karlshruhe Institute of Technology, Germany. He completed his MS in the field of electrical and computer engineering from National University of Sciences and Technology, Pakistan, in 2003 and his PhD degree in computer vision at Technische Universität Berlin, Germany, in 2008. Before joining Karlshruhe Institute of Technology, he worked as head of Computer Vision Research Group, CIIT, Lahore, Pakistan. His research interests include pattern recognition, statistical machine learning, face recognition, and multimodal biometrics.