Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-10728-z

Skin detection in video under uncontrolled illumination

Biplab Ketan Chakraborty[1] · M. K. Bhuyan[2] · Karl F. MacDorman[3]

Received: 5 January 2020 / Revised: 22 January 2021 / Accepted: 16 February 2021 /
Published online: 05 April 2021

Abstract
Many vision-based human-computer interaction (HCI) applications require skin detection. However, their performance relies on accuracy in detecting skin regions in video, which is difficult under uncontrolled illumination. The chromatic appearance of skin changes because of shading, often caused by body movement. To address this, we propose a dynamic adaptation method to detect skin regions affected by local color deformations. Static and dynamic skin regions are detected by a corresponding module. The static module includes a facial skin distribution model (FSDM) and a fusion-based background distribution model (FBDM). The FBDM is obtained from a local background distribution model (LBDM) and a global background distribution model (GBDM). The LBDM is obtained by comparing a frame pixel distribution model with the FSDM and GBDM. Next, the FBDM is derived from the LBDM and the GBDM. The dynamic module includes a moving skin distribution model (MSDM), derived from a set of moving skin samples. Initially, moving skin regions are detected using a modified double frame-difference method and then modeled using a Gaussian mixture model. To avoid misidentifying background regions as skin, the final MSDM is obtained by comparing the initial moving skin model to the FSDM and FBDM. Finally, the static and the dynamic models are fused by applying a maximization rule. Experimental results shows that the proposed method can detect skin regions more accurately than state-of-the-art methods.

Keywords First keyword · Second keyword · More

1 Introduction

Skin detection is an important step in many human-computer interaction (HCI) applications like facial expression recognition [32], facial feature point detection [9], gesture recogni-

✉ M. K. Bhuyan
mkb@iitg.ernet.in

Biplab Ketan Chakraborty
biplab.ketan@sankhyasutralabs.com

[1] Sankhyasutra Labs, Karnataka, India

[2] Department of Electronics and Electrical Engineering, IIT Guwahati, Guwahati, 781039, India

[3] Indiana University School of Informatics and Computing, 535 West Michigan St., Indianapolis, IN 46202 USA

tion [24, 25, 31, 39], real-time heart rate monitoring [10, 36], forensics [26], and medical diagnosis [11, 28]. Chyad et al. [8] conducted an extensive review and analysis of different methods for skin detection. According to Chyad et al., accuracy of skin detection depends on multiple factors such as variations in skin pigmentation among different races, scene illumination, shadows, presence of skin-like colors in the background, and camera characteristics. A major challenge is the effect of varying illumination on apparent skin color. Illumination can vary globally or locally. A global variation occurs when the characteristics of the illuminating source vary with time. However, a local illumination variation occurs when illumination becomes non-uniform over the exposed skin regions. A frequent cause of non-uniform illumination is the curvature of the skin surface, which results in form shadows. In addition, a body part occluding the illumination of another body part results in cast shadows. For directional light sources, the extent of illumination on a skin patch also depends on its orientation with respect to the light source. In most HCI applications, local illumination variations occur more frequent than global illumination variations. Applications such as human-robot interaction [21], medical systems and assistive technology [16], and gesture-controlled computer games [38] are some of the real-life examples where both the background and the scene illuminant are almost static.

In recent years, much research [3, 5, 6, 8, 14, 20, 22] has been reported on skin detection in images under unconstrained environments. However, research has seldom addressed skin detection in video under varying illumination conditions. Soriano et al. [35] investigated the effect of static but non-uniform illumination on skin color: A skin locus is introduced to describe a chromatic constraint on skin color appearance. However, derivation of a skin locus requires a calibrated camera and the face must be captured for different illuminants. Also, the skin locus is camera specific and, hence, must be recalculated for every unknown video. For varying illumination conditions, Sigal et al. [34] proposed a second-order Markov model with dynamic histogram adaptation. Their method assumes illumination changes gradually and globally. This approach is not suitable for local illumination changes because of their unpredictability. Habili et al. [12] used motion and color information to detect skin regions in video. However, their method assumes the background contains no skin-colored regions. Awad et al. [2] make the same assumption in proposing a fusion-based model. However, they further assume uniform illumination of skin regions. Therefore, under unconstrained illumination and background conditions, the skin pixels cannot be located accurately. Also, their method requires a set of labeled initial frames to train a support vector machine (SVM) classifier and determine the initial positions of skin-colored objects. Han et al. [13] proposed a skin segmentation and tracking algorithm for sign language recognition by using SVM active learning. The SVM is trained with a set of initial frames. However, its major drawback is an inability to handle varying illumination. In addition, the SVM must be relearned at every frame, which is computationally expensive. Liu et al. [23] proposed a face detection-based model update scheme for varying illumination. However, only variations in global illumination were considered. The method is not suitable for local illumination changes, which occur mainly because of moving body parts.

In recent years, deep learning-based object classification methods became a new trend due to their higher accuracies as compared with classical methods. In [22], Lei et al. used stacked autoencoders for skin detection under different illumination conditions. Zuo et al. [40] showed that skin color segmentation can be treated as a semantic segmentation problem. Here, layers of recurrent neural networks (RNNs) are integrated into a fully convolutional neural network (FCN) module. This improves skin segmentation by using spatial correlation among skin pixels through RNN blocks. However, identification of color whether it belongs to the skin or non-skin region is image-specific. Therefore, the accuracy

of a skin detector, even though it uses deep networks, is limited by the skin and non-skin region statistics provided by the training data. Dynamic illumination change is a temporal phenomenon in a video and it can be modeled using Markov models [34] if it follows some pattern. Presently, LSTM-RNNs are widely used for temporal sequence predictions such as activity recognition [4, 30], gesture recognition [1], etc. However, using RNNs to model temporal variation of local illumination may be difficult due to the excessive randomness associated with the change in the illumination pattern. Also, the training of deep networks requires a large training dataset, which is not available for the the problem to be addressed. However, the presented algorithm does not require a large training dataset, and its skin detection model updates itself by using moving skin samples.

A survey of the literature reveals a need to explore further the effects of local color and shading deformation of skin regions in video. To detect these regions, we propose a dynamic adaption scheme, which employs a Bayes classifier. It has two modules: a static module for the detection of static skin regions and a dynamic module for the detection of moving skin regions. The static module has two components: a skin distribution model (SDM) and a background distribution model (BDM). The skin distribution model, termed *facial skin distribution model* (FSDM), is derived from a set of facial skin samples of initial frames of a video. The background distribution model must adapt to the background characteristics of the video. To obtain it, a *global background distribution model* (GBDM) is first created from a set of randomly collected background samples. However, skin colors could be present in the background of a video frame. Therefore, a *local background distribution model* (LBDM) needs to be derived for the video frame. To obtain the LBDM, we follow a similarity match-based algorithm by using pixel distribution information [6]. The similarity between two distribution functions can be measured by using the Bhattacharyya distance [7]. The local deformations in skin color may create multiple modes in the skin distribution. The Bhattacharyya distance treats these modes as different distribution functions and produces finite distances accordingly. Therefore, a new distance metric is needed to discriminate the distribution functions belonging to similar regions of an image (*e.g.*, skin regions) from those belonging to dissimilar regions (*e.g.*, background regions).

In this paper, we propose to modify the Bhattacharyya distance to reduce the distance between two distribution functions if they belong to the same skin region. The LBDM is derived from the FSDM and the GBDM using the modified Bhattacharyya distance (MBD) as a metric. The final background model, termed as *fusion-based background distribution model* (FBDM), is obtained by fusing the LBDM and the GBDM. The main component of the dynamic model is a moving skin distribution model (MSDM). The MSDM reflects the distribution of pixels belonging to moving skin regions having chromatic deformations. The regions are detected by using a modified double frame-difference method. An initial distribution model for these moving skin regions is obtained by using a GMM. In addition to the skin regions, some background regions can be falsely detected as skin regions during the moving object detection process. The final MSDM is obtained by performing a filtering procedure based on similarities of the initial moving skin model to the FSDM and the FBDM. Finally, the static and dynamic modules are fused by following a maximization rule. The proposed method is detailed in the following sections.

2 Proposed method

The block diagram of the proposed method is shown in Fig. 1. The proposed skin detection algorithm has two modules: a static module (FSDM and FBDM) and a dynamic model

(MSDM). The background distribution model as shown in Fig. 1 is common to both modules. Each module produces a skin probability map (SPM). The two maps are fused together to obtain the final skin probability map for a video frame. The proposed skin detection method is explained in detail in the following sections.

2.1 Static module

The components of the static module are as follows:

2.1.1 Facial skin distribution model

Facial skin tone can be used as a person's reference skin tone. At first, the Viola and Jones algorithm [37] is applied over a set of initial frames $\mathbf{I}_{TF} = I_1, I_2, ...I_{N_{TF}}$ to locate faces in the frames. Here, N_{TF} is the number of initial training frames. A set of sample pixels is obtained from the localized facial regions. The distribution of reference skin pixels extracted from the facial regions or the FSDM is modeled as a single multivariate Gaussian function G^f:

$$G^f = N\left(\mu^f, \mathbf{\Sigma}^f\right) \tag{1}$$

2.1.2 Background distribution model

The proposed fusion-based background distribution model has two components: a global background distribution model (GBDM) and a local background distribution model (LBDM). To obtain the GBDM, a set of sample background pixels obtained from a standard dataset is used. The global background model is expressed as

$$G^{gb} = \sum_{k=1}^{K_{gb}} \omega_k^{gb} N\left(\mu_k^{gb}, \Sigma_k^{gb}\right) \tag{2}$$

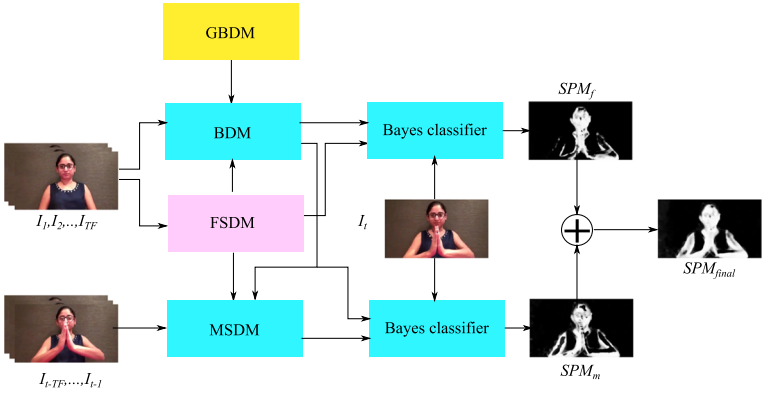

Fig. 1 Block diagram of the proposed method

To obtain the LBDM, a local distribution model adaptation scheme [6] is followed. At first, a pixel distribution model G^I for pixels of a set of initial frames $\mathbf{I}_{TF}$ is derived from a Gaussian mixture model (GMM). The expression for G^I is given as

$$G^I = \sum_{k=1}^{K_I} \omega_k^I N\left(\mu_k^I, \Sigma_k^I\right) \tag{3}$$

The number of Gaussian components K_{gb} and K_I are selected by using the method given in [6].

The model G^I gives a color distribution of frame pixels, which includes both skin and non-skin pixels. The Gaussian components of G^I, reducing to the real skin regions, should be statistically more similar to the FSDM, and they should be more dissimilar to the Gaussian components of the GBDM. Now, the Bhattacharyya distance is a well-known metric to find distance between two probability distribution functions. The closed form expression for Bhattacharyya distance between two multivariate Gaussian distribution functions is given by

$$\begin{aligned} d_{Bh}\left(G_i, G_j\right) = & \tfrac{1}{8}\left(\mu_i - \mu_j\right)^T \left[\tfrac{\Sigma_i + \Sigma_j}{2}\right]^{-1} \left(\mu_i - \mu_j\right) \\ & + \tfrac{1}{2} \ln \tfrac{\left|\left(\Sigma_i + \Sigma_j\right)/2\right|}{\sqrt{\left|\Sigma_i\right| \cdot \left|\Sigma_j\right|}}, \ \forall i, j \end{aligned} \tag{4}$$

and the overlap between G_i and G_j is given by

$$\varepsilon_{i,j} = \exp\left[-d_{Bh}\left(G_i, G_j\right)\right] \ \forall i, j \tag{5}$$

In this paper, the standard Bhattacharyya distance is termed *original Bhattacharyya distance* (OBD).

Now, let us consider two Gaussian clusters of pixels in RGB space: a reference cluster C_r $(\boldsymbol{\mu}_r, \Sigma_r)$ and one modified to account for local illumination changes C_l (μ_l, Σ_l). Here, illumination change on skin regions can be approximated as translation of a Gaussian distribution of skin samples along its mean vector. Hence, the intended distance metric should produce a smaller distance between two Gaussian distributions if one of them is a translated copy of the other along the mean vector. Also, skin tone distribution of a person should be independent of illumination changes. Hence, Gaussian distributions derived from skin samples of different body parts (subjected to local illumination variations) should have covariance matrices with the same eigenvectors. Therefore, the distance between C_l and C_r should be zero as both correspond to similar image regions, contrary to the OBD, which gives non-zero distance between C_l and C_r. To overcome this problem, the OBD is modified so that $d_{MBh}\left(G_l, G_r\right) < d_{Bh}\left(G_l, G_r\right)$ for the distribution functions corresponding to similar image regions. We termed the proposed distance measure $d_{MBh}\left(G_l, G_r\right)$ the *modified Bhattacharyya distance* (MBD). Local variations of skin colors are approximated as a combination of two independent parameters: the orientation of the centroid vector $\boldsymbol{\mu}_l$ with respect to $\boldsymbol{\mu}_r$ and the orientation of C_l with respect to C_r, as shown in Fig. 2. Let ϕ be the angle between $\boldsymbol{\mu}_l$ and $\boldsymbol{\mu}_r$. The two clusters will be perfectly aligned if $\left[1 - \frac{tr\left(\mathbf{V}_l^T \mathbf{V}_r\right)}{3}\right] = 0$, where $\mathbf{V}_l$ and $\mathbf{V}_r$ are the eigenvector matrices of Σ_l and Σ_r, respectively. However, the

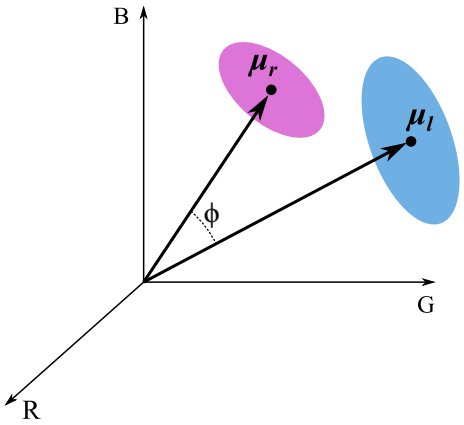

Fig. 2 Similarity between two clusters

smaller the value of ϕ, the more chromatically similar the clusters. Hence, the closed form of the MBD is expressed as

$$d_{MBh}\left(G_i, G_j\right) = \xi\left(\mu_i - \mu_j\right)^T \left[\frac{\Sigma_i + \Sigma_j}{2}\right]^{-1} \left(\mu_i - \mu_j\right) + \gamma \ln \frac{\left|\left(\Sigma_i + \Sigma_j\right)/2\right|}{\sqrt{\left|\Sigma_i\right| \cdot \left|\Sigma_j\right|}}, \quad \forall i, j \tag{6}$$

where

$$\xi = \frac{1}{8}\left[1 - \exp\left(-\frac{\phi}{\phi_{\max}}\right)\right], \gamma = \frac{1}{2}\left[1 - \frac{tr\left(\mathbf{V}_i^T \mathbf{V}_j\right)}{3}\right]$$
$$\phi = \cos^{-1}\left(\frac{\langle \mu_i, \mu_j \rangle}{\|\mu_i\| \cdot \|\mu_j\|}\right) \tag{7}$$

Here, ϕ_{max} is the maximum allowed deviation in centroid orientation. The following rule is framed on the basis of (6):

$$d_{MBh} \rightarrow d_{Bh} : (\phi \gg \phi_{\max}) \wedge \left\{\frac{tr\left(\mathbf{V}_i^T \mathbf{V}_j\right)}{3} \ll 1\right\} \equiv true \tag{8}$$

Figure 3 shows the effect of using the MBD as compared with the OBD. Three patches are extracted from an image: $Patch_1$, $Patch_2$, and $Patch_3$ (Fig. 3a). $Patch_1$ and $Patch_3$ belong to skin regions, whereas $Patch_2$ belongs to a non-skin region. We consider $Patch_1$ has the true skin tone and $Patch_3$ has a deformed skin tone due to local illumination variation. Figure 3b shows the distribution of pixels belonging to these patches in RGB space. In Fig. 3c, distance ratios calculated with the OBD and MBD are shown for these patches. The analysis shows that the relative distance between an unknown and reference skin distribution is less for the MBD than the OBD.

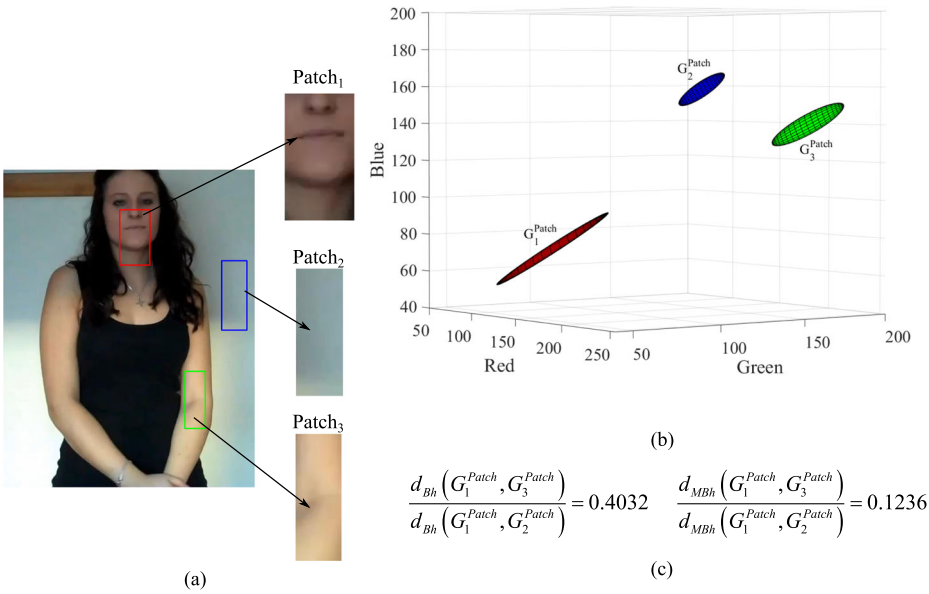

Fig. 3 Effect of using the MBD as a distance measure: **a** the original image with three patches extracted from different regions; **b** distribution of pixels in RGB space for each patch; the red, green, and blue cluster correspond to Patch_1, Patch_2, and Patch_3, respectively; and **c** distance ratios for different patches

The overlap between Gaussian components of G^I and FSDM is then given by

$$\varepsilon_i^f = \exp\left[-d_{MBh}\left(G_i^I, G^f\right)\right] \text{ for } i = 1, ...K_I \tag{9}$$

and the overlap between Gaussian components of G^I and G^{gb} is given by

$$\varepsilon_i^{gb} = \exp\left[-\min_{\forall j} d_{MBh}\left(G_i^I, G_j^{gb}\right)\right] \text{ for } j = 1, ...K_{gb} \tag{10}$$

A Gaussian component in G^I should belong to the background if it overlaps less with the FSDM than with the GBDM. Also, background regions may be chromatically similar to the skin regions in some cases. Therefore, the local background distribution model (LBDM) G^{lb} should include only those Gaussian components that follow the inclusion criterion defined below:

$$G_i^I \in \left\{G^{lb} : \left(\varepsilon_i^f \leq \varepsilon_i^{gb}\right) \vee \left(\varepsilon_i^f \leq \tau_1\right) \equiv true\right\} \tag{11}$$

The weights of the Gaussian components in G^{lb} are derived from their weights in G^I after normalization. The GBDM exceled at discriminating skin colors from other colors, whereas the LBDM exceled at discriminating skin colors from those in the background. Hence, a fusion of GBDM and LBDM should provide better discrimination between skin pixels and

non-skin pixels than either GBDM or LBDM. Accordingly, the fusion-based background distribution model (FBDM) G^{fb} is expressed as

$$G^{fb}(\mathbf{X}) = max\left\{G^{lb}(\mathbf{X}), G^{gb}(\mathbf{X})\right\} \tag{12}$$

where $\mathbf{X} = [R\ G\ B]'$ is the color vector of a pixel.

2.2 Dynamic module

As explained earlier, the moving body parts cause local illumination variations in the scene even if the global illumination is kept constant. This non-uniform illumination may change the chromatic appearance of moving skin regions in such a manner that they may become undetectable by the static module. Therefore, the skin model should adapt to local variations of skin colors due to non-uniform illumination. However, updating a skin model for every frame is too computationally intensive for real-time applications. Therefore, the skin model should be updated only during a significant change in scene color due to local illumination variations. We propose a keyframe detection technique to update the skin model only for the keyframes.

2.2.1 Keyframe selection

For a given keyframe, the next keyframe to be selected is the one with a significant change in chromaticity. Thus, an entire video clip is transformed into a small number of representative keyframes. The chromatic changes between a frame and a reference frame can be determined from their chromatic entropies. Let I_t and I_{ref} represent the frame at t and the reference frame, respectively. The change in chromatic entropy in I_t with respect to I_{ref} is given by

$$\Delta E_t = \frac{\left|E_t - E_{ref}\right|}{E_{ref}} \tag{13}$$

where

$$E_t = -\frac{\sum\limits_{\forall \mathbf{X} \in I_t} n_{\mathbf{X}} \log\left(\frac{n_{\mathbf{X}}}{w \times h}\right)}{w \times h} \tag{14}$$

E_t and E_{ref} represent the chromatic entropy of I_t and I_{ref}, respectively; $n_{\mathbf{X}}$ is the number of occurrence of the color vector $\mathbf{X} = [RGB]'$ in I_t; w is the frame width, and h is the frame height. A frame I_t is designated as a keyframe I_{key} if its ΔE_t is greater than some threshold value θ_E.

2.2.2 Moving skin distribution model

Assuming the background is static, the frame difference method is the easiest way to detect moving objects in a frame. However, the frame difference algorithm suffers from two major limitations: the occurrence of ghost foreground regions and foreground apertures, as shown in Fig. 4. Ghost foreground regions are caused by the motion of the objects. During frame differencing, ambiguity may occur between real foreground regions and ghost foreground regions. The other drawback of frame differencing is foreground object aperture (FOA). FOA is likely to occur in moving skin regions because of their low texture and intensity gradient. To avoid the occurrence of ghost foreground regions, Kameda and Michihiko [17] proposed a double-difference of frame (DDF) method. In this method, three frames at time

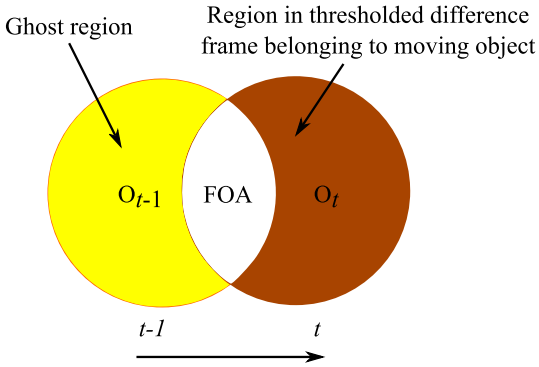

Fig. 4 Drawback of single frame difference method

$t-2$, $t-1$, and t are selected. The DDF method performs a logical AND operation over thresholded difference frames between frames at $t-2$ and $t-1$ and frames at $t-1$ and t. The result of the DDF algorithm in the presence of FOA is shown in Fig. 5. In this, O_{t-2}, O_{t-1}, and O_t show object positions at time $t-2$, $t-1$, and t, respectively.

The DDF algorithm produces a narrow region for a moving object if the object has a low texture and/or intensity gradient. Morphological operations can reduce FOA in a difference frame, as shown in Fig. 6. A dilation along the direction of motion of an object can reduce FOA with an inclusion of ghost foreground regions. Motivated by this fact, a morphological enhancement-based double difference frame method is proposed to detect moving skin regions. In our proposed method, morphological dilation is applied to each of the thresholded difference frames. Subsequently, a logical OR operation is performed over the dilated difference frames. The OR operation helps to include more moving skin regions between the consecutive frames. However, this operation significantly increases the occurrence of ghost foreground regions. To reduce their inclusion in a thresholded difference frame, dilation should be performed in the direction of motion of the foreground objects, as shown in Fig. 6. However, for articulated objects like hands, the foreground motion could be complex. We approximate complex movements as a combination of motions in four directions, 0°, 45°, 90°, and 135°, with respect to the horizontal direction. Directional opening can be used to select a region in a particular direction. After directional opening, a dilation in the perpendicular direction of the opening process grows a region in the direction of its motion,

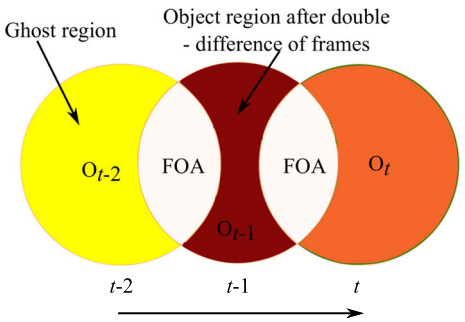

Fig. 5 Results of DDF method in presence of FOA

as shown in Fig. 7. Finally, a logical OR operation is performed on the two morphologically enhanced thresholded difference frames to obtain a moving object mask $BW_{\Delta I}^{final}$. Figure 8 shows the flowchart of our proposed morphology-based moving object localization method. An example of the proposed moving skin region localization approach is shown in Fig. 9.

Algorithm 1 Proposed algorithm to derive the moving object pixel set Q_{motion}.

Data: I_t where $t = 1, 2, ...N_f$
Result: Moving object pixel set Q_{motion}
for $N_{TF} < t \leq N_f$ **do**
  if $t = N_{TF} + 1$ **then**
    for $3 \leq k \leq t$ **do**
      Obtain:
        $\Delta I_{k-1} \leftarrow |I_{k-1} - I_{k-2}|$
        $\Delta I_k \leftarrow |I_k - I_{k-1}|$
        $BW_{\Delta I_j} \xleftarrow{binirization} \Delta I_j$ where $j = k, k-1$.
      Perform: Morphological region enhancement on $BW_{\Delta I_k}$ and $BW_{\Delta I_{k-1}}$.
      Obtain:
        $BW_{\Delta I}^{final} \leftarrow BW_{\Delta I_{k-1}} \cup BW_{\Delta I_k}$
      Obtain:
        $Q_{k-1} \leftarrow I_{k-1}(x) : BW_{\Delta I}^{final}(x) = 1$
      Add Q_{k-1} to the bottom of Q_{motion}.
    end
    Perform:
      $I_{key} \leftarrow I_t$
  else
    Calculate ΔE_t
    if $\Delta E_t > \theta_E$ **then**
      Remove $Q_{t-N_{TF}}$ from the top of Q_{motion}.
      Obtain: Q_t
      Add Q_t to the bottom of Q_{motion}.
      Update:
        $I_{key} \leftarrow I_t$
    else
      Do not update I_{key}
    end
  end
end
return Q_{motion}

Ideally Q_{motion} should only contain pixels belonging to moving skin regions. So, the pixel set Q_{motion} can be used to derive a color distribution model for moving skin regions, *i.e.*, the moving skin distribution model (MSDM). However, some pixels from non-skin foreground (e.g., clothing) and/or background regions may be included in Q_{motion} due to the presence of ghost regions and the proposed morphological region enhancement. Therefore, a filtering process is necessary to exclude these background pixels to derive the MSDM. To obtain the MSDM, a GMM is used to model the distribution of all the pixels in Q_{motion} as

$$G_m^{init} = \sum_{k=1}^{K_m^{init}} \omega_m^{init} N\left(\mu_k^{init}, \Sigma_k^{init}\right) \tag{15}$$

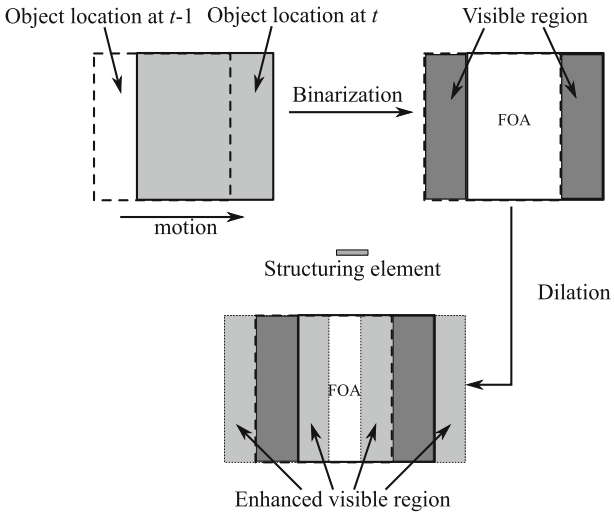

Fig. 6 Reduction in FOA using directional dilation

where G_m^{init} represents an initial pixel distribution model for moving objects. Some Gaussian components in G_m^{init} correspond to background and/or non-skin moving object regions. These Gaussian components should be nearer to the background model FBDM than to FSDM. The proposed MBD measure is used to determine the overlapping of G_m^{init} components with FSDM and FBDM. The overlap between Gaussian components of G_m^{init} and G^f is given by

$$\varepsilon_i^{f,m} = \exp\left[-d_{MBh}\left(G_{m,i}^{init}, G^f\right)\right] \text{ for } i = 1, ...K_m^{init} \tag{16}$$

Similarly, the overlap between Gaussian components of G_m^{init} and the global background model G^{gb} is given by

$$\varepsilon_i^{gb,m} = \exp\left[-\min_{\forall j} d_{MBh}\left(G_{m,i}^{init}, G_j^{gb}\right)\right] \text{ for } j = 1, ...K_{gb} \tag{17}$$

and the overlap between Gaussian components of G_m^{init} and the local background model G^{lb} is given by

$$\varepsilon_i^{lb,m} = \exp\left[-\min_{\forall j} d_{MBh}\left(G_{m,i}^{init}, G_j^{lb}\right)\right] \text{ for } l = 1, ...K_{lb} \tag{18}$$

The overall overlapping of G_m^{init} with FBDM is expressed as

$$\varepsilon_i^{fb,m} = max\left(\varepsilon_i^{gb,m}, \varepsilon_i^{lb,m}\right), \forall i = 1, ...K_m^{init} \tag{19}$$

Finally, an inclusion criterion is proposed for Gaussian components of G_m^{init} to the final moving skin distribution model (MSDM) *i.e.*, G_m^{skin}. The inclusion criterion is formulated as follows:

$$G_{m,i}^{init} \in \left\{G_m^{skin} : \left(\varepsilon_i^{f,m} > \varepsilon_i^{fb,m}\right) \wedge \left(\varepsilon_i^{f,m} > \tau_2\right) \equiv true\right\} \tag{20}$$

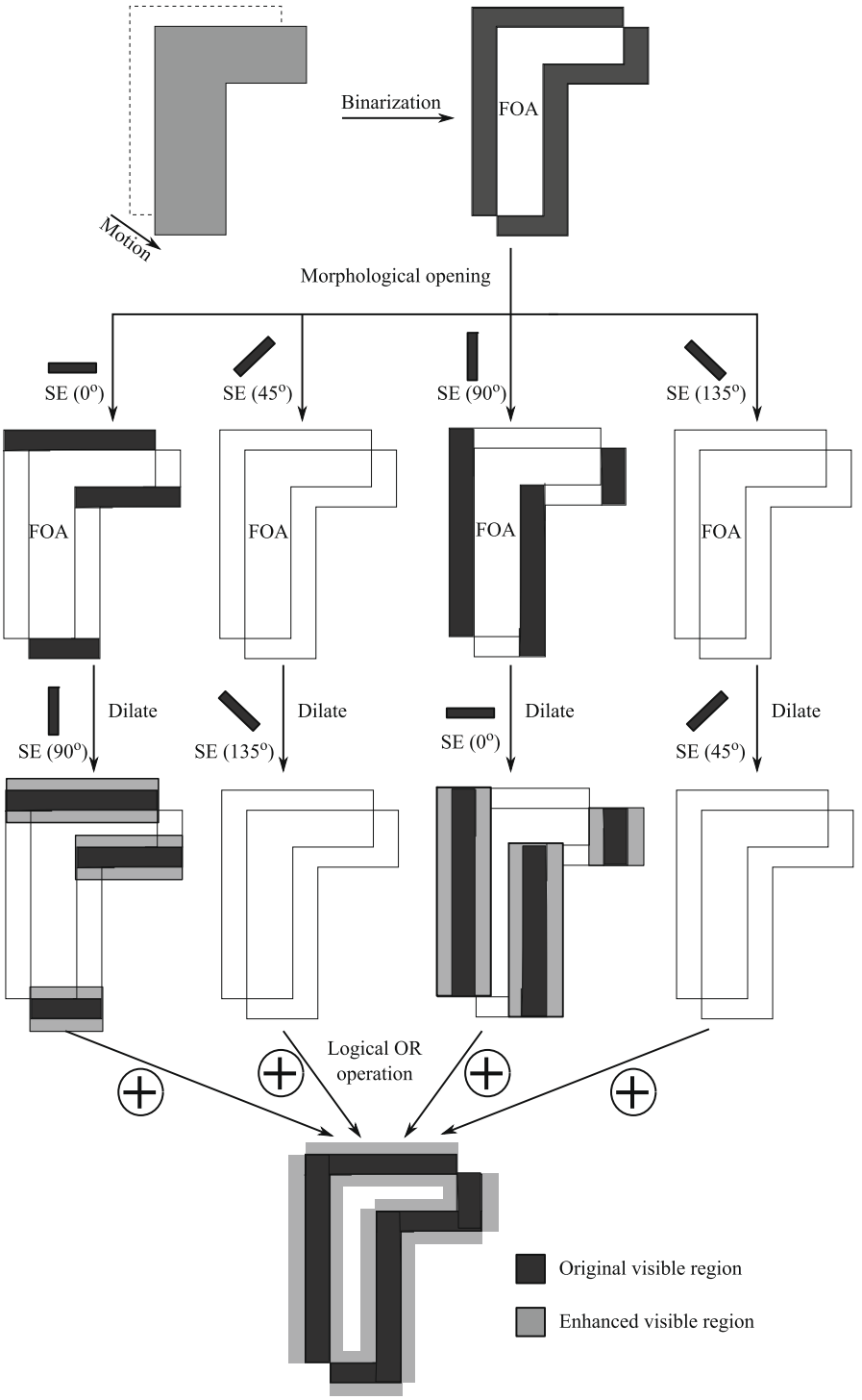

Fig. 7 Proposed method for reducing FOA using directional opening followed by directional dilation

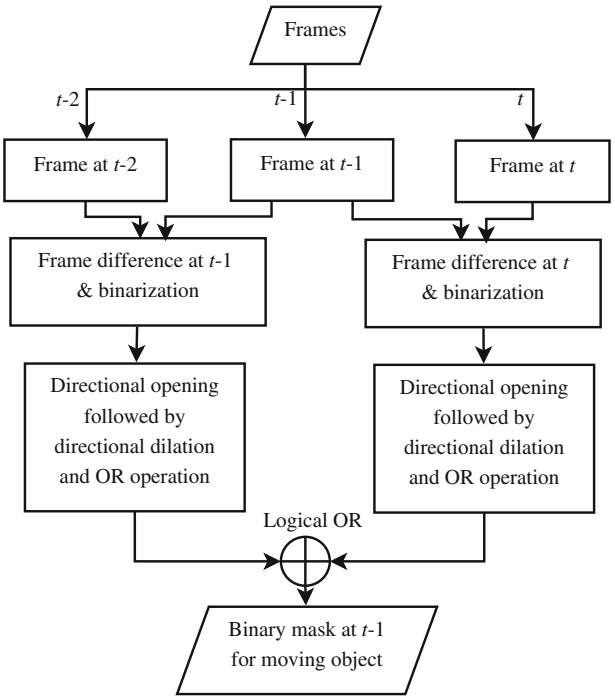

Fig. 8 Flowchart of the modified DDF algorithm

2.3 Derivation of a Skin Mask

The pixels of each frame are classified into skin and non-skin pixels by using a Bayes classifier [15]. The Bayes classifier provides a SPM for a frame at t. The SPM at a location x represents a posteriori probability of a pixel $\mathbf{X}$ belonging to skin region at that location, and it is defined as

$$SPM(x) = \frac{P(S) \cdot P(\mathbf{X}\,|S)}{P(S) \cdot P(\mathbf{X}\,|S) + P(NS) \cdot P(\mathbf{X}\,|NS)} \tag{21}$$

where $P(\mathbf{X}\,|S)$ and $P(\mathbf{X}\,|NS)$ are likelihoods, $P(S)$ and $P(NS)$ are priors for skin (S) and non-skin (NS) pixels, respectively. Hence, the SPM derived using FSDM and FBDM is given by

$$SPM_f(x_t) = \frac{P(S)G^f(\mathbf{X}_t)}{P(S)G^f(\mathbf{X}_t) + P(NS)G^{fb}(\mathbf{X}_t)} \tag{22}$$

The SPM value derived using MSDM and FBDM is given by

$$SPM_m(x_t) = \frac{P(S) \cdot G_m^{skin}(\mathbf{X}_t)}{P(S) \cdot G_m^{skin}(\mathbf{X}_t) + P(NS) \cdot G^{fb}(\mathbf{X}_t)} \tag{23}$$

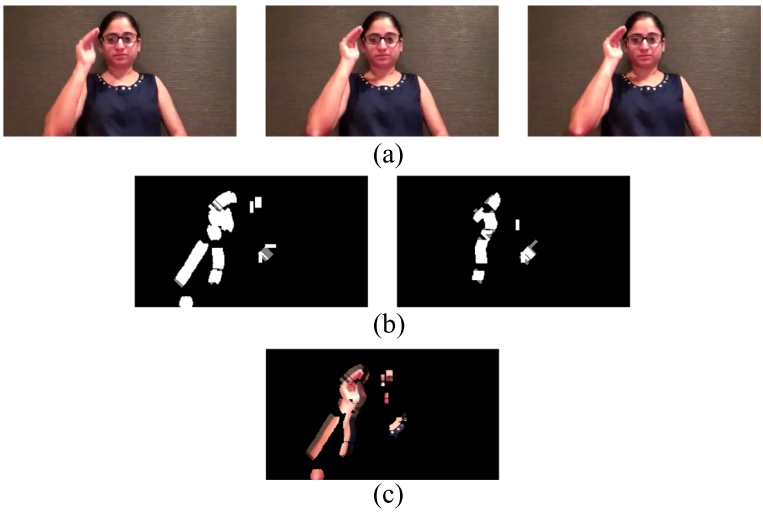

Fig. 9 Example of morphology-based moving object localization: **a** frames at $t-2, t-1$ and t, **b** morphologically enhanced binarized difference frames at $t-1$ and t, and **c** final localized moving object regions

It is observed that SPM_f can capture static skin regions more accurately, whereas SPM_m is more responsive to moving skin regions. Therefore, the final SPM at time t is expressed as

$$SPM_{final}(x_t) = \max\left\{SPM_f(x_t), SPM_m(x_t)\right\} \tag{24}$$

The final skin mask $Mask$ for a frame at t is obtained by thresholding SPM_{final} with an appropriate threshold θ_{th} as

$$Mask(x_t) = \begin{cases} 1 \text{ if } SPM_{final}(x_t) \geq \theta_{th} \\ 0 \text{ otherwise.} \end{cases} \tag{25}$$

3 Experimental analysis

3.1 Experimental setup

To obtain a global background model G^{gb}, background samples were extracted from a set of images selected at random from the ECU dataset [29]. To validate our proposed algorithm experimentally, a set of sign language videos collected from the web were used. The selected videos were captured in unconstrained illumination and varying background conditions. Each video has a duration of 10 seconds with varying frame rate, spatial resolution and illumination conditions. For quantitative performance analysis, these videos are manually annotated for skin and non-skin regions. From the annotated videos, it is observed that $P(S) \sim [0.1, 0.2]$ for standard definition (SD) video (4:3 aspect ratio), and $P(S) \sim [0.05, 0.15]$ for high definition (HD) video (16:9 aspect ratio). Hence, we judiciously choose $P(S) = 0.15$ for SD videos and $P(S) = 0.1$ for HD videos. The respective values for $P(NS)$ are derived as $P(NS) = 1 - P(S)$. Four quantitative measures, namely, *detection accuracy (Acc.)*, *false alarm rate* (δ_{fp}), *miss rate* (δ_{fn}), and *total detection error rate* (δ_t) are selected for evaluation, where $\delta_t = \delta_{fp} + \delta_{fn}$. All the detection errors, are

obtained by thresholding SPM_{final} with a threshold θ_{th}. For simplicity of implementation, we select $\tau_1 = \tau_2 = \tau$.

3.1.1 Determination of τ

The parameter τ controls the inclusion of Gaussian components in G^{lb} and G_m^{skin}. A smaller value of τ results in the inclusion of Gaussian components into G_m^{skin} which belong to skin colors in the background. This increases the chance of false alarms. However, a larger value of τ results in inclusion of some Gaussian components into G^{lb}. This increases the chance of misses. Hence, we have proposed a selection scheme for the parameter τ as follows:

Let I_t be the frame at t. The chromatic randomness of an image can be determined from its chromatic entropy. Our previous work [6] found that the chromatic entropy of a skin-masked image, also termed as skin-chroma entropy (E_s), approximates the false alarm rate. The parameter E_s can be obtained by

$$E_s = -\frac{\sum\limits_{\forall y \in \mathbf{Y}} n_{\mathbf{X}(y)} \log\left(\frac{n_{\mathbf{X}(y)}}{N_{\mathbf{Y}}}\right)}{N_{\mathbf{Y}} \cdot \log(255^3)} \tag{26}$$

where

$$y \in \left\{\mathbf{Y} : P_{rgb}^{skin}(y) \geq \theta_{sp}\right\} \tag{27}$$

and

$$P_{rgb}^{skin}(y) = \frac{P(S) \cdot G^f(\mathbf{X})}{P(S) \cdot G^f(\mathbf{X}) + P(NS) \cdot G^{gb}(\mathbf{X})} \tag{28}$$

Here, $n_{\mathbf{X}(y)}$ is the number of count of color vector $\mathbf{X}(y)$ and $N_{\mathbf{Y}}$ is the total number of pixel locations in $\mathbf{Y}$. The threshold θ_{sp} is obtained by using Otsu's method [27].

Similarly, a face-masked version of I_t *i.e.*, I_{face} is obtained by applying a face mask derived by using the Viola and Jones algorithm [37]. Let E_f be the chromatic entropy of I_{face}. Ideally, E_s should be equal to E_f. However, in practice I_{skin} may contain skin colored background regions along with some regions that contain skin. So, $E_s > E_f$ in presence of skin colors in the background. Thus, we express the parameter τ as

$$\tau = \left[\sum_n \left(\frac{E_f}{E_s}\right)^{n/2}\right]^{-1} \simeq \left[1 + \sqrt{\frac{E_f}{E_s}} + \frac{E_f}{E_s}\right]^{-1} \tag{29}$$

where $n = 0, 1, 2, \ldots$. As the ratio $\frac{E_f}{E_s} < 1$, higher order powers ($n > 2$) are neglected.

3.2 Experimental validation

At first, we examine the effect of ϕ_{max} on detection results. For this, detection errors are calculated by varying ϕ_{max}, and the results are given in Table 1. It is observed that when $\phi_{max} = 0°$, the average false alarm rate $\delta_{fn,avg}$ is maximal. For $\phi_{max} = 0°$, the parameter $\xi = \frac{1}{8}$. This implies that some clusters corresponding to the true skin pixels in G^I are included in LBDM, and/or excluded from MSDM. In either condition, clusters whose centroids are very close to that of FSDM are only selected as skin clusters. An increase in ϕ_{max} value causes more skin colored clusters to be excluded from the background model

Table 1 Detection results for different values of ϕ_{max}

ϕ_{max}	$\delta_{fp,avg}$ (%)	$\delta_{fn,avg}$ (%)	$\delta_{t,avg}$ (%)	Avg. Acc. (%)
0°	6.41	14.34	20.75	92.50
10°	6.04	9.48	15.52	93.62
20°	**5.78**	**8.19**	**13.97**	**94.04**
30°	7.00	8.29	15.29	93.01
40°	11.09	8.53	19.62	89.59
50°	11.52	7.92	19.44	89.34

Table 2 Comparative analysis for the OBD and MBD

Method	$\delta_{fp,avg}$ (%)	$\delta_{fn,avg}$ (%)	$\delta_{t,avg}$ (%)	Avg. Acc. (%)
SPM$_{final}$ using OBD	6.13	11.28	17.41	93.26
SPM$_{final}$ using MBD	**5.78**	**8.19**	**13.97**	**94.04**

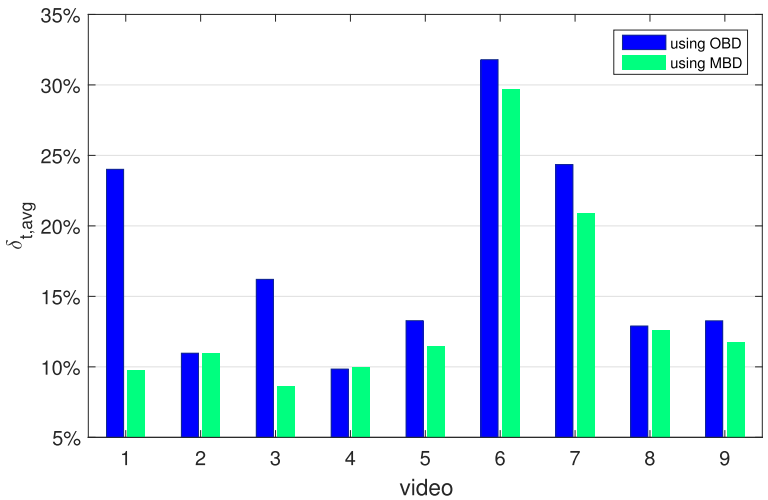

Fig. 10 Video comparison between the OBD and MBD

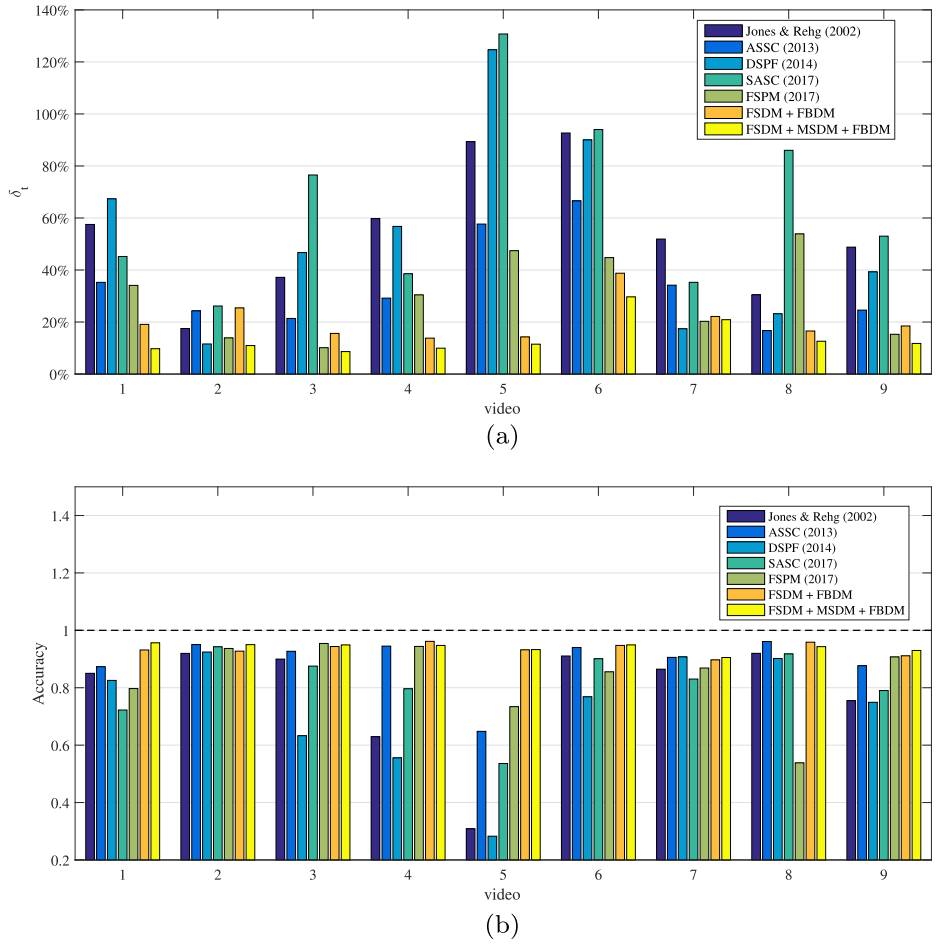

Fig. 11 Video comparative bar charts for minimum δ_t values: **a** δ_t for different videos and **b** accuracy for different videos

and/or included in MSDM. Thus, false alarms reduce with an increase in the value of ϕ_{max}. However, this increases the chance of misses. Some of the clusters of skin-like background pixels can be included in MSDM and/or excluded from LBDM if the ϕ_{max} is increased further. For example, the average false alarms $\delta_{fp,avg}$ increases significantly for an increase in ϕ_{max} from 20° to 40°. However, the average false alarm rate $\delta_{fn,avg}$ decrease significantly for an increase in ϕ_{max} from 0° to 20°. It is also observed that for $\phi_{max} \geq 20°$, the average total detection error $\delta_{t,avg}$ for all the videos becomes the lowest. Hence, ϕ_{max} is fixed at 20° for the remaining analysis.

A comparative analysis was also performed between the original Bhattacharyya distance and the proposed modified Bhattacharyya distance. Detection errors were calculated by both OBD and MBD. As mentioned in (8), the maximum value of the MBD was fixed by OBD. The comparison between the OBD and MBD in calculating SPM is shown in

Table 3 Comparative analysis of different methods

Method	for minimum δ_t				for maximum Accuracy			
	$\delta_{fp,avg}$ (%)	$\delta_{fn,avg}$ (%)	$\delta_{t,avg}$ (%)	Avg. Acc.	$\delta_{fp,avg}$ (%)	$\delta_{fn,avg}$ (%)	$\delta_{t,avg}$ (%)	Avg. Acc. (%)
Bayes classifier [15]	19.69	34.22	53.91	78.42	14.05	48.31	62.36	81.44
SDDMA [33]	4.71	58.08	62.80	0.8710	4.71	58.08	62.80	87.10
ASSC [19]	8.57	25.85	34.42	89.19	6.53	30.12	36.65	90.22
DSPF [20]	21.67	22.37	44.04	0.7832	16.23	32.31	48.54	81.78
SASC [22]	11.70	45.15	56.84	0.8470	11.70	45.15	56.84	84.70
FSPM [6]	16.85	13.17	30.02	0.8373	4.20	34.27	38.47	92.10
FSDM + BDM	**5.33**	15.14	20.47	0.9345	2.74	22.90	25.64	94.45
FSDM + MSDM + BDM	5.78	**8.19**	**13.97**	**0.9404**	**2.33**	**16.68**	**19.01**	**95.80**

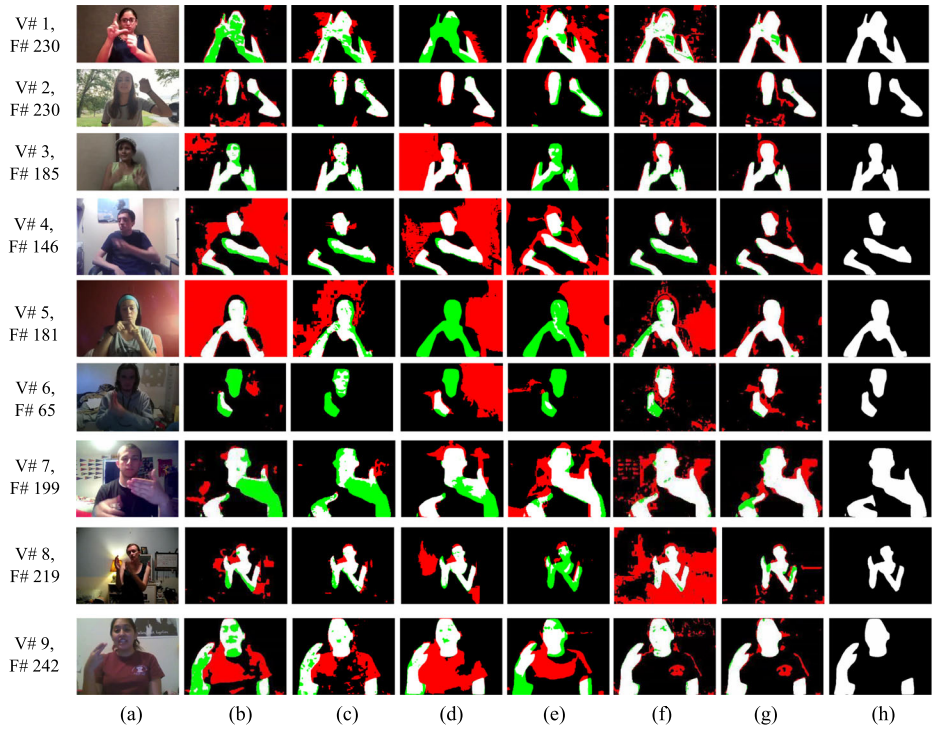

Fig. 12 Comparative detection results for different videos: **a** original frames, **b** Jones and Rehg [15], **c** ASSC [19], **d** DSPF [20], **e** SASC [22], **f** FSPM [6], **g** proposed method, and **h** groundtruth. Here, white represents hits, black represents correct rejections, red represents false alarms, and green represents misses

Table 2. In contrast to OBD, detection errors reduced due the application of the MBD. This validates the efficacy of the proposed MBD in skin detection. In Fig. 10, two comparative bar charts are given for different videos. The results show that $\delta_t|_{MBD} \leq \delta_t|_{OBD}$. The total detection errors were calculated at the maximum attainable accuracy values for different videos. Figure 11 give the corresponding bar charts showing the δ_t and accuracy values for different videos.

Finally, the proposed method is compared with state-of-the-art methods, such as Bayes classifiers [15], fast propagation-based skin segmentation (FPSS) [18], adaptive seed-based skin classification (ASSC) [19], skin detection by dual maximization of detectors agreement (SDDMA) [33], discriminative skin-presence features (DSPF) [20], stacked autoencoders-based skin classification (SASC) [22], and fusion-based skin probability map (FSPM) [6]. The detection results obtained by different methods are given in Table 3. The Bayes classifier method proposed by Jones and Rehg [15] is considered as a benchmark. For training, the method needs a set of skin and non-skin pixel samples, which are obtained globally. However, the accuracy of this method depends on the training sample set. The Jones and Rehg's method produces more misses as compared with our proposed method. To compare the SDDMA method with the proposed method, it is trained with N_{TF} number of labelled initial frames of nine videos (In total, $9 \times N_{TF}$ frames). However, SDDMA largely fails to detect true skin regions. It produces the highest rate of misses among all the benchmark methods. The DSPF method can give better results than the standard SPM, and it gives a

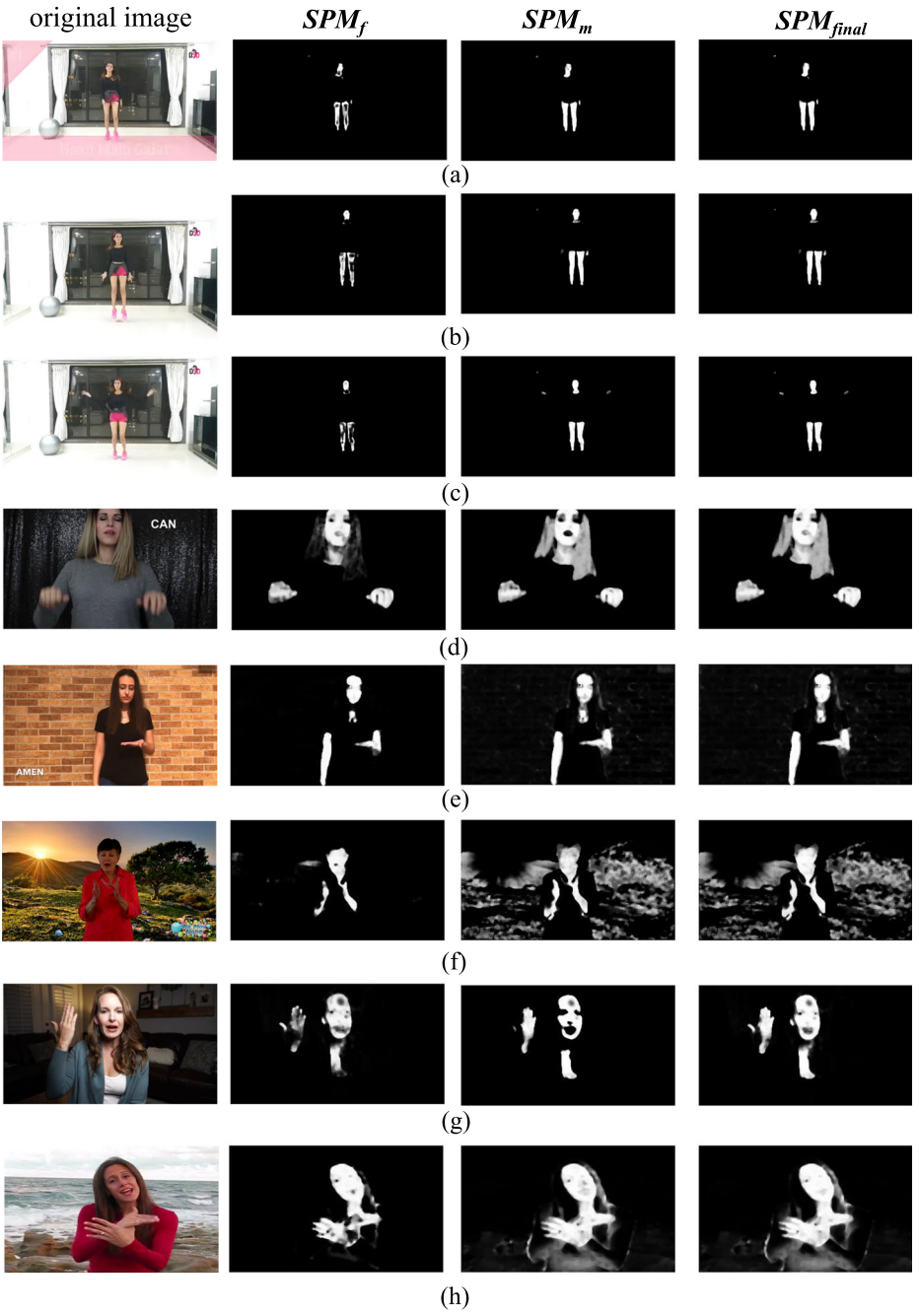

Fig. 13 Detection results for videos with different characteristics: **a**-**c** video of dance where exposed skin regions are small, **d**-**g** videos with textured background, and **h** video with dynamic background (sea)

discriminative space-based skin map. The DSPF method mostly relies on the SPM derived from global training samples, and thus it is not adaptive to local environmental conditions of an image.

As the skin color of the face is similar to that of other body parts, face detection-based skin model adaptation can perform well. The ASSC method uses adaptive seeds for growing skin regions. The adaptive seeds are derived from a local skin model of facial skin pixels. The ASSC method relies on a standard SPM for region growing; hence, it is unable to detect many skin pixels, resulting in more misses than the proposed method. The FSPM-based method gives better detection accuracy by using image pixel distribution information to derive a local skin probability map (LSPM). The LSPM is later fused with the original or global SPM to get the FSPM. The FSPM can detect more skin pixels than other benchmark methods. In our method, the combination of FSDM and FBDM gives a static skin detection model for video. The incorporation of MSDM makes the skin detection model adaptive to local illumination changes, and a dynamic skin detection model for video is obtained. The static model is prone to produce more misses than its dynamic counterpart. The dynamic adaptation of the proposed MSDM to local illumination changes makes the proposed system more robust under unconstrained illumination and background conditions. In Fig. 12, the detection results are shown for all the test videos. The experiential results show that the proposed method can detect skin regions in a video more accurately than current state-of-the-art methods even in the presence of local color deformations. Some additional detection results are shown in Fig. 13 where videos of different characteristics are tested. Figure 13a-c show that the detection ability of the proposed algorithm is not dependent on skin region sizes. However, Fig. 13d-g show that skin detection is robust to background texture variation. Also, a video with dynamic background was tested, and the detection result is shown in Fig. 13h. Therefore, it is evident from the results shown in Fig. 13 that the proposed skin detection method can work flawlessly under different background conditions.

4 Conclusion

Many vision-based human-computer interaction (HCI) applications require the segmentation of skin regions in video. This is challenging, however, when the shading of other body parts causes local chromatic variations. We propose a skin detection method that dynamically adapts to local skin color variations. The proposed method has three main components: a facial skin pixel distribution model for user-specific skin modeling, a video-specific local background distribution model, and a moving skin-pixel distribution model to detect skin regions affected by local color deformations caused by moving body parts. The FSDM is derived from a set of facial pixels from the initial frames of a video. The LBDM is derived from the FSDM and a global background distribution model. A modified Bhattacharyya distance is employed to measure similarity between two distribution models. Subsequently, a fusion-based background distribution model is derived from the GBDM and LBDM. The MSDM gives the distribution of moving skin pixels. The final skin detection model is derived from the FSDM, the MSDM, and the FBDM. As the MSDM is updated at every keyframe, the proposed skin model adapts to local illumination changes. The experimental results show that the MBD produces fewer detection errors than the benchmark methods.

One research direction is to detect skin regions affected by both local and global illumination variations. Global illumination variation may change the chromaticity of a scene

(background and skin color). Therefore, a temporal model for global illumination change needs to be estimated along with a moving scene color distribution model.

References

1. Avola D, Bernardi M, Cinque L, Foresti GL, Massaroni C (2019) Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. IEEE Trans Multimed 21(1):234–245
2. Awad G, Han J, Sutherland A (2006) A unified system for segmentation and tracking of face and hands in sign language recognition. In: Proc. Int. Conf. Pattern Recogn., vol 1, pp 239–242
3. Bianco S, Gasparini F, Schettini R (2015) Adaptive skin classification using face and body detection. IEEE Trans Image Process 24(12):4756–4765
4. Carrara F, Elias P, Sedmidubsky J (2019) Lstm-based real-time action detection and prediction in human motion streams. Multimed Tools Appl 78:27309–27331
5. Chakraborty BK, Bhuyan M (2020) Image specific discriminative feature extraction for skin segmentation. Multimed Tools Appl:1573–7721
6. Chakraborty BK, Bhuyan MK, Kumar S (2017) Combining image and global pixel distribution model for skin colour segmentation. Pattern Recogn Lett 88:33–40
7. Chung JK, Kannappan PL, Ng CT, Sahoo PK (1989) Measures of distance between probability distributions. J Math Anal Appl 138(1):280–292
8. Chyad MA, Alsattar HA, Zaidan BB, Zaidan AA, Al Shafeey GA (2019) The landscape of research on skin detectors: Coherent taxonomy, open challenges, motivations, recommendations and statistical analysis, future directions. IEEE Access 7:106536–106575
9. Danyang C, Menggui Z, Yanchao Z (2018) Research on a facial feature points detection method based on skin color model. In: 2018 Chinese Control And Decision Conference (CCDC), pp 5688–5693
10. Fouad RM, Omer OA, Aly MH (2019) Optimizing remote photoplethysmography using adaptive skin segmentation for real-time heart rate monitoring. IEEE Access 7:76513–76528
11. George Y, Aldeen M, Garnavi R (2017) A pixel-based skin segmentation in psoriasis images using committee of machine learning classifiers. In: Int. Conf. Digit. Image Comput.: Tech. and Appl. (DICTA), pp 1–8
12. Habili N, Lim C-C, Moini A (2004) Segmentation of the face and hands in sign language video sequences using color and motion cues. IEEE Trans Circ Syst Video Technol 14(8):1086–1097
13. Han J, Awad G, Sutherland A (2009) Automatic skin segmentation and tracking in sign language recognition. IET Comput Vis 3(1):24–35
14. Hettiarachchi R, Peters JF (2016) Multi-manifold-based skin classifier on feature space voronoï regions for skin segmentation. J Vis Commun Image R 41:123–139
15. Jones MJ, Rehg J (2002) Statistical color models with application to skin detection. Int J Comput Vis 46(1):81–96
16. Kakkoth SS, Gharge S (2018) Real time hand gesture recognition its applications in assistive technologies for disabled. In: Int. Conf. Comput. Comm. Control and Auto. (ICCUBEA), pp 1–6
17. Kameda Y, Michihiko M (1996) A human motion estimation method using 3-successive video frames. In: Proc. Conf. on Virtual Systems and Multimedia, pp 135–140. http://ci.nii.ac.jp/naid/10024346616/en/
18. Kawulok M (2013) Fast propagation-based skin regions segmentation in color images. In: Proc. 10th IEEE Int. Conf. and Workshops Autom. Face and Gesture Recogn. (FG), pp 1–7
19. Kawulok M, Kawulok J, Nalepa J, Papiez M (2013) Skin detection using spatial analysis with adaptive seed. In: Proc. IEEE ICIP, pp 3720–3724
20. Kawulok M, Kawulok J, Nalepa J (2014) Spatial-based skin detection using discriminative skin-presence features. Pattern Recogn Lett 41:3–13
21. Lei Q, Zhang H, Xia Z, Yang Y, He Y, Liu S (2019) Applications of hand gestures recognition in industrial robots: a review. In: Verikas A, Nikolaev DP, Radeva P, Zhou J (eds) Int. Conf. Machine Vis. (ICMV 2018), vol 11041. International Society for Optics and Photonics, SPIE, pp 455–465
22. Lei Y, Yuan W, Wang H, Wenhu Y, Bo W (2017) A skin segmentation algorithm based on stacked autoencoders. IEEE Trans Multimed 19(4):740–749
23. Liu L, Sang N, Yang S, Huang R (2011) Real-time skin color detection under rapidly changing illumination conditions. IEEE Trans Consum Electron 57(3):1295–1302

24. McBride TJ, Vandayar N, Nixon KJ (2019) A comparison of skin detection algorithms for hand gesture recognition. In: 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA), pp 211–216
25. Mo T, Sun P (2019) Research on key issues of gesture recognition for artificial intelligence. Soft Comput 24:5795–5803
26. Moreira DC, Fechine JM (2018) A machine learning-based forensic discriminator of pornographic and bikini images. In: 2018 Int. Joint Conf. Neural Net. (IJCNN), pp 1–8
27. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66
28. Pennisi A, Bloisi DD, Nardi D, Giampetruzzi AR, Mondino C, Facchiano A (2016) Skin lesion image segmentation using delaunay triangulation for melanoma detection. Comput Med Imag Graph 52:89–103
29. Phung SL, Bouzerdoum A, Chai S (2005) Skin segmentation using color pixel classification: Analysis and comparison. IEEE Trans Pattern Anal Mach Intell 27(1):148–154
30. Ramasinghe S, Rajasegaran J, Jayasundara V, Ranasinghe K, Rodrigo R, Pasqual AA (2019) Combined static and motion features for deep-networks-based activity recognition in videos. IEEE Trans Circ Syst Vid Tech 29(9):2693–2707
31. Rautaray S, Agrawal A (2012) Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intell Rev:1–54
32. Sandbach G, Zafeiriou S, Pantic M, Yin L (2012) Static and dynamic 3d facial expression recognition: A comprehensive survey. Image Vis Comput 30(10):683–697
33. SanMiguel JC, Suja S (2013) Skin detection by dual maximization of detectors agreement for video monitoring. Pattern Recogn Lett 34(16):2102–2109
34. Sigal L, Sclaroff S, Athitsos V (2004) Skin color-based video segmentation under time-varying illumination. IEEE Trans Pattern Anal Mach Intell 26(7):862–877
35. Soriano M, Martinkauppi B, Huovinen S, Laaksonen M (2003) Adaptive skin color modeling using the skin locus for selecting training pixels. Pattern Recogn 36(3):681–690
36. Tang C, Lu J, Liu J (2018) Non-contact heart rate monitoring by combining convolutional neural network skin detection and remote photoplethysmography via a low-cost camera. In: 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recogn. Work. (CVPRW), pp 1390–13906
37. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proc. IEEE Comp. Society Conf. Comp. Vis. and Pattern Recognition (CVPR), vol 1, pp 511–518
38. Wachs JP, Kölsch M, Stern H, Edan Y (February 2011) Vision-based hand-gesture applications. Commun ACM 54(2):6071
39. Yang J, Wang Y, Lv Z, Jiang N, Steed A (2018) Interaction with three-dimensional gesture and character input in virtual reality: Recognizing gestures in different directions and improving user input. IEEE Consum Electron Mag 7(2):64–72
40. Zuo H, Fan H, Blasch E, Ling H (2017) Combining convolutional and recurrent neural networks for human skin detection. IEEE Signal Process Let 24(3):289–293