IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 8, 2020, accepted November 13, 2020, date of publication November 23, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3039740

Modeling Temporal Visual Salience for Human Action Recognition Enabled Visual Anonymity Preservation

SALAH AL-OBAIDI, HIBA AL-KHAFAJI, AND CHARITH ABHAYARATNE, (Member, IEEE)
Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S1 3JD, U.K.

Corresponding author: Charith Abhayaratne (c.abhayaratne@sheffield.ac.uk)

ABSTRACT This paper proposes a novel approach for visually anonymizing video clips while retaining the ability to machine-based analysis of the video clip, such as, human action recognition. The visual anonymization is achieved by proposing a novel method for generating the anonymization silhouette by modeling the frame-wise temporal visual salience. This is followed by analysing these temporal salience-based silhouettes by extracting the proposed histograms of gradients in salience (*HOG-S*) for learning the action representation in the visually anonymized domain. Since the anonymization maps are based on the temporal salience maps represented in gray scale, only the moving body parts related to the motion of the action are represented in larger gray values forming highly anonymized silhouettes, resulting in the highest mean anonymity score (MAS), the least identifiable visual appearance attributes and a high utility of human-perceived utility in action recognition. In terms of machine-based human action recognition, using the proposed *HOG-S* features has resulted in the highest accuracy rate in the anonymized domain compared to those achieved from the existing anonymization methods. Overall, the proposed holistic human action recognition method, *i.e.*, the temporal salience modeling followed by the *HOG-S* feature extraction, has resulted in the best human action recognition accuracy rates for datasets DHA, KTH, UIUC1, UCF Sports and HMDB51 with improvements of 3%, 1.6%, 0.8%, 1.3% and 16.7%, respectively. The proposed method outperforms both feature-based and deep learning based existing approaches.

INDEX TERMS Visual anonymization, human action recognition, histogram of gradients in salience (*HOG-S*), temporal visual salience estimation, privacy, video-based monitoring, assisted living.

I. INTRODUCTION

Vision-based human action recognition (HAR) plays an important role in surveillance [1], [2], human computer interaction [3], human object interactions [4], healthcare monitoring [5], assisted living [6], [7], smart homes [8], [9] and etc., since vision sensors are informative [10]–[12]. Such fusion between vision sensors and the computer vision has become essential for monitoring the daily human actions in ambient assisted living (AAL) [6], [7], [13], [14], although human action recognition is a challenging task [15]. However, exploring vision sensors for in-home monitoring has often found concerns in protecting visual privacy [6], [16]–[18]. Current solutions to address visual privacy concerns in video are mainly based on processing the pixel intensity values spatially to cover the identity details. These include face or the whole body, by means of masking [19], blurring [20] and pixelation [21]. However, after visually anonymizing, the utility of such sequences in visual analysis, such as, action recognition, is severely affected. Some applications, such as, assisted living require analyzing such visually anonymized video for tasks like human action recognition. Therefore, new algorithms that can visually anonymize monitoring video while retaining the utility of the video for automated analysis are required. In this paper, we propose a new method for visual anonymization of video while retaining important salient features for human activity recognition in the visual anonymity domain.

Visual anonymization in monitoring applications usually adopt the image processing techniques, such as, Gaussian blurring [20], pixelation [21], blocking [22], cartooning [23] and masking with sold silhouette [19], to obfuscate the

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

sensitive information. However, these methods require to consider the trade-off between the visual anonymity and the utility of the anonymized sequences for monitoring tasks [24]. Achieving this trade-off is one of the major challenges associated with using the video camera in AAL. In the case of privacy concealment, the existing filtering-based models lose the accuracy of low level features for modeling the most dominant human body parts that are responsible for representing the action. Thus, discrimination among the actions tends to be inaccurate from the perspective of both the human vision and computer vision. Therefore, exploring the spatial content to obfuscate the identity leads to inaccurate modeling and misses the discrimination among the actions in HAR.

Recently, visual saliency detection for video has been proposed to highlight the most dynamic salience content in video sequences [25]–[30]. The outcome of video saliency is a useful abstract for the most dominant visual information in the scene without showing the details since the salient content is represented through highlighting the essential content, simulating perception in the human vision system (HVS). Visual saliency can be due to the spatial attentive cues as in images as well as due to the temporal saliency due to the motion in a video sequence. Although, salience estimation for video has become a widely addressed topic recently, all methods consider joint spatial and temporal salience modeling. However, since our focus is in the utility, such as HAR, in this paper we propose a novel temporal salience estimation and demonstrate the use of such salience maps for visual anonymization and HAR. The temporal saliency also seems to be a useful tool for addressing the challenges, such as background clutter often seen in computer vision, since the spatial content is excluded in modeling the temporal salience. Also, we aim to compute the temporal salience as a map in gray scale highlighting from the least salient to the the most salient regions using 0 to 255 gray values, respectively.

Our proposal is to replace the video sequences with the computed temporal salience map sequences and then explore the salience sequence for utility tasks, such as, HAR. The computed temporal salience sequences not only capture the temporal events, as in emerging neuromorphic (event-based) cameras [31], but also records significance of those events by means of recording the magnitude of pixel-wise salience in a 0-255 range. Early results of our work was presented as conference papers [17], [32]. This paper extends the model with analysis for new HAR descriptors, extension to visual anonymizations and evaluation of visual anonymity using both objective and subjective metrics. The main contributions of this work are:

1) A new methodology for estimating the temporal saliency based on modeling the intensity changes between successive frames.
2) Exploring the temporal saliency maps for achieving visual anonymity addressing privacy concerns in video-based monitoring.
3) A methodology of exploring the anonymity domain by extracting new Histogram of Oriented Gradients in Salience (*HOG-S*) features for HAR.

The rest of this paper is organized as follows: Section II reviews the related work in the literature. Section III presents the proposed method for extracting the temporal visual salience maps for visual anonymizing and extracting features in the anonymized domain for HAR. The performance evaluation of the proposed methodology in terms of both visual anonymization and anonymized domain HAR is presented and discussed in Section IV followed by the conclusions in Section V.

II. RELATED WORK

In this section, we briefly present the recent work on both privacy preservation and HAR.

A. PRIVACY PRESEVATION

Besides the work in this paper, other anonymity methods have been presented and emerged, which are valuable efforts to preserve privacy. However, these methods are mostly focused on covering the identity silhouette using image processing in spatial domain [19]–[23] or the use of low-resolution visual sensors [33]–[36], where less information for visual recognition is present. Using low-resolution sensors adopts a network of extremely low-resolution cameras [33]–[35] or low-resolution colour sensors [36] to capture low-resolution visual images. These sensors have been successfully exploited in the applications of activity recognition [33], behaviour understanding [34] and object localisation [36]. However, these sensors are more sensitive to the local changes in the light conditions [34], [36], which affects the reliability in HAR.

The second category of solutions is to adopt the image processing techniques, such as, blocking [22], cartooning [23], blurring [20], pixelation [21], to obfuscate the sensitive information. Their main characteristics are summarised in TABLE 1. These image filtering based methods destroy the original intensity magnitudes and destroying the valuable features. Therefore, exploring the anonymity domains of these methods for HAR affects the accuracy rates of recognition. Furthermore, the trade-off between the privacy protection and utility of the anonymized sequences for monitoring tasks has to be considered [24]. Often, a higher level of privacy protection means a low level of utility and vice versa. This trade-off is one of the major challenges associated with using video-based vision sensors in the application of AAL.

TABLE 1. The main characteristics of image processing-based anonymization.

Algorithms	Domain	Visual anonymization	Utility	Computation cost
Blurring [20]	*Spatial*	*Medium*	*Medium*	*Low*
Pixelation [21]	*Spatial*	*Medium*	*Medium*	*Low*
Cartooning [23]	*Spatial*	*Medium*	*Medium*	*Low*
Blocking [22]	*Spatial*	*High*	*Low*	*Low*

Therefore, our proposed approach is a valuable contribution to the development of algorithms to preserve privacy while enabling the subsequent analysis utility tasks, such as HAR.

B. ACTION RECOGNITION USING HAND-CRAFTED FEATURES

Several recent works have been reported to represent the actions based on hand-crafted feature extraction. One of the most considered algorithms is the local dense trajectories representation using Histogram of Oriented Gradients (HOG) [37] due to its robustness [38]. The existing works on HOG-based HAR are categorised into two themes: 2D HOG [39]–[41] and 3D HOG [42]–[44] representations. In the first category, the dense features are extracted from a single image/frame to show the motion history. In the second category, a volumetric representation in space-time is exploited to represent the action. However, in both categories, redundant data, such as, the background, is exploited to extract the features to describe the actions. This redundancy affects the discriminating power of the descriptor and increases the storage requirements for this information and makes the complexity higher. Mainly, it is interesting to address these problems based on determining candidate local interest points [45], although interested point-based learning has also many problems. All existing methods for HAR are based on the raw data domain, such as, colour video. However, those algorithms do not perform well on image processing-based visually anonymized sequences.

Recently, saliency estimation has attracted much attention in image and video processing [25]–[30]. The visual saliency estimation algorithms highlight the most important visual content, *i.e.*, foreground, and attenuate others, *i.e.*, background. This representation substitutes the intensities with the salience magnitudes and reduces the redundancy through modeling the saliency map. Thus, the visual saliency offers a tool for addressing the problems mentioned above of visual information [46], [47], and makes the saliency-based representation useful and accurate for the feature learning applications.

All video salience algorithms focus on joint spatio-temporal salience. However, for our work we intend to use temporal salience only. Hence we propose a new approach for temporal salience estimation for video. There is also added advantage of exploring temporal salience maps for HAR, as such maps have already abstracted the original sequences to a motion-driven event map sequence with highlighted significance of the events in gray scale.

III. THE PROPOSED METHOD

Our proposed method is two-fold: 1) Temporal visual salience mapping for visually anonymizing the video sequences and 2) human action recognition in the visually anonymized domain. For the former, we propose a novel method for estimating temporal visual saliency as detailed in Section III-A. For the latter we propose the Histograms of Gradients in Salience (*HOG-S*) features extracted from the anonymity domain, *i.e.*, the temporal visual salience map sequence as presented in Section III-B. It must be also noted that many traditional HAR methods [48] begin with temporal shot segmentation [49]–[51]. However, our proposed method detailed in this paper mainly focuses on action recognition from temporal salience maps from a given temporal window of video frames.

A. TEMPORAL VISUAL SALIENCE MODELING FOR VISUAL ANONYMIZING

Let $C = \{s_z^F, q_z\}_{z=0}^{V-1}$ be the action dataset with V video sequences and Q set of action classes, where s_i is the sequence with index i containing F frames and $q \in Q$ action label. The proposed algorithm starts by calculating the frame difference, D'_t between each two consecutive frames, f_t and $f_t - 1 \in s_i$, where t is the frame index, to define the change in the pixel intensity over time, as

$$D'_t(x, y) = f_t(x, y) - f_{t-1}(x, y), \tag{1}$$

for all (x, y) spatial coordinates. The difference at a given pixel can occur for several reasons, for example, illumination change and global motion. Therefore, the frame difference is compared with a user-defined threshold, τ, in order to eliminate the small changes and maintain the dominated moving pixels as follows:

$$D_t(x, y) = \begin{cases} D'_t(x, y) & \text{if } |D'_t(x, y)| > \tau, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $D_t(x, y)$ is the frame difference at location (x, y) with respect to the threshold τ. Note that $|\cdot|$ denotes the absolute value.

Next, for each pixel location (x, y), we compute the Shannon's Entropy $\mathcal{E}(x, y)$ of the normalised power spectral density (PSD) of D_t values considering an $N \times N$ pixel window block centred at (x, y). Let $b_m \in B$ be the corresponding $N \times N$ block with $B = \{b_1, b_2, \cdots, b_M\}$, M is the total number of blocks and m is the block index. In order to make up the blocks for pixel at the frame borders, the frame borders are padded with relevant number of zero values according to the chosen N. The PSD for each block, S_{b_m}, is defined as

$$S_{b_m}(u, v) = \frac{1}{N^2} \mathcal{A}_{b_m}(u, v)^2, \tag{3}$$

where $\mathcal{A}_{b_m}(u, v)$ is the magnitude of the 2D Fast Fourier Transform (2DFFT) coefficient at frequency location (u, v) in block b_m. S_{b_m} is normalised to suppress the high variation among those in different blocks. This is achieved by normalising with respect to the sum of all PSD components of a given block. This is followed by the computation of Shannon's entropy ($\mathcal{E}_{b_m}$) of the normalised PSD of b_m in order to get $\mathcal{E}_t(x, y)$. The computation of $\mathcal{E}_t(x, y)$ captures the contribution of the D_t values in the neighbourhood of $D_t(x, y)$. The entropy $\mathcal{E}_t(x, y)$ is proportional to the amount of variation of magnitudes of the corresponding S_{b_m}. For example the higher the variation in magnitudes in S_{b_m} the higher the value of $\mathcal{E}_t(x, y)$.

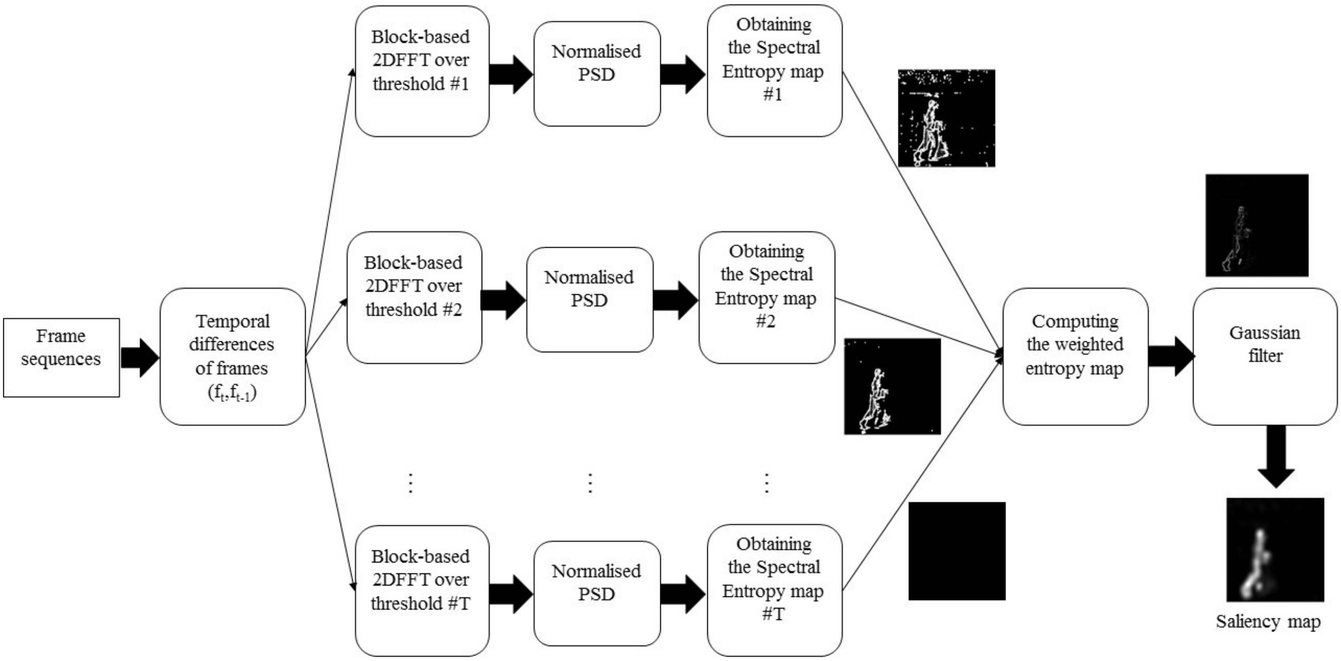

Frame sequences
Temporal differences of frames (f_t,f_{t-1})
Block-based 2DFFT over threshold #1
Normalised PSD
Obtaining the Spectral Entropy map #1
Block-based 2DFFT over threshold #2
Normalised PSD
Obtaining the Spectral Entropy map #2
Block-based 2DFFT over threshold #T
Normalised PSD
Obtaining the Spectral Entropy map #T
Computing the weighted entropy map
Gaussian filter
Saliency map

FIGURE 1. Visually anonymizing silhouette generation based on the proposed visual temporal salience estimation.

This local spectral entropy value, $\mathcal{E}_t(x, y)$, fairly captures the variations in D_t to identify the temporal salience in a frame. It exploits the source of the most dominant intensity changes to model the underlying motion (with respect to the action). Most of the time, it is difficult to determine the perfect value of τ in Eq. (2) to maintain the desired changes and suppress other noisy changes because the motion levels vary according to the actions in sequences. To make this representation more robust and generalised, we further vary τ by defining a set of thresholds, $\tau_h = 2^h$, where $h = 1, \cdots, H$, with maximum number of user defined threshold levels, H. For each pixel location (x, y), a set of entropy values, $\mathcal{E}_t^{\tau_h}(x, y)$ for the corresponding block, b_m, considering all τ_h is computed. Finally, the weighted entropy, $\hat{\mathcal{E}}_t(x, y)$, across all entropy maps, $\mathcal{E}_t^{\tau_h}(x, y)$, over all H thresholds is computed as

$$\hat{\mathcal{E}}_t(x, y) = \frac{\sum_{h=1}^{H} \tau_h \mathcal{E}_t^{\tau_h}(x, y)}{\sum_{h=1}^{H} \tau_h}. \tag{4}$$

This entropy map is then normalised to be in the range of gray level values in the range [0, 255] and smoothed by applying a 2D Gaussian kernel in order to fill in the small holes and obtain the final temporal visual salience map based silhouette, S_t. It links the neighbouring pixels that are close to each other to construct the temporal silhouette region. The generation of the silhouette of the human in action based on the proposed temporal salience estimation algorithm is shown as a block diagram in FIGURE 1 and summarized in Algorithm 1.

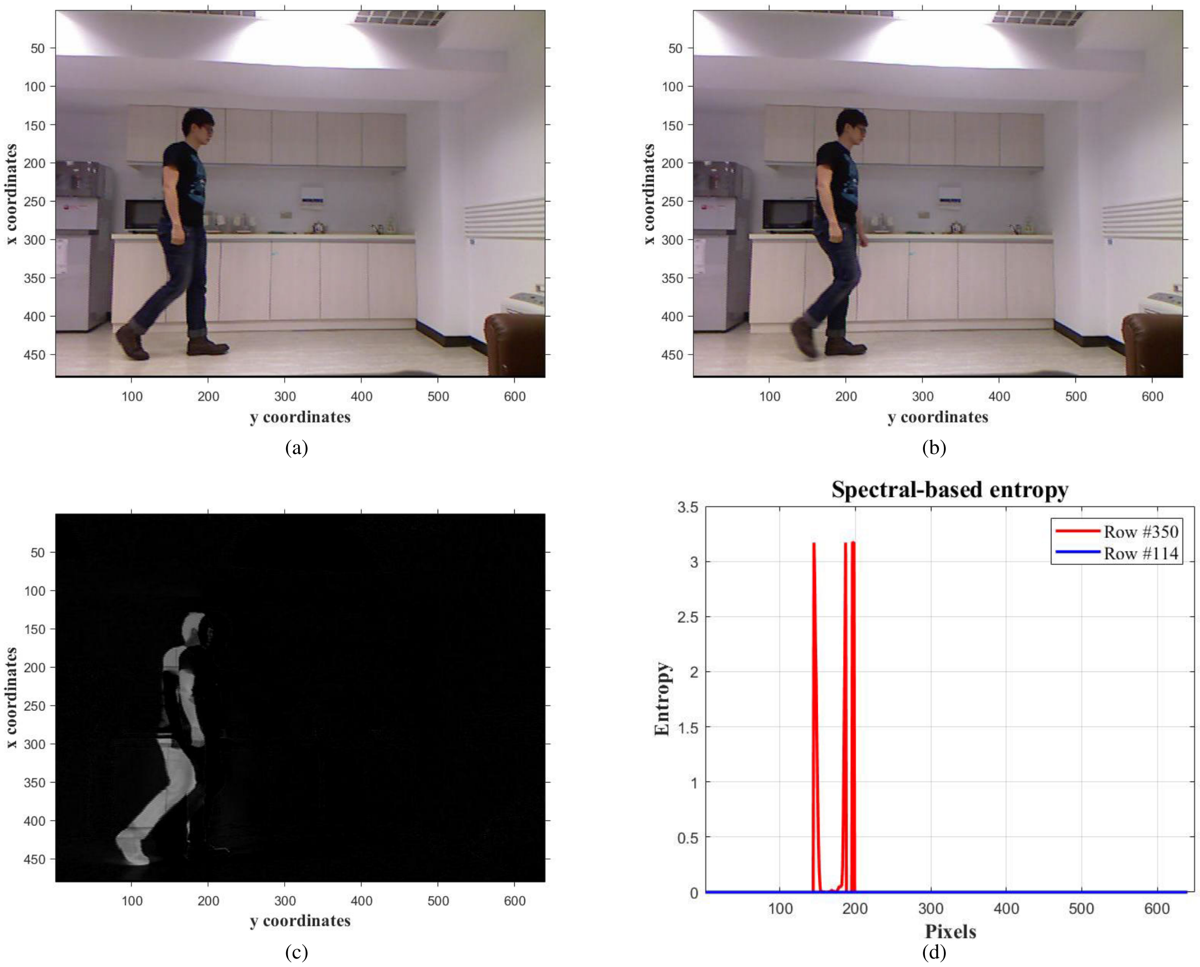
FIGURE 2 illustrates how $\mathcal{E}_t^{\tau_h}(x, y)$ captures the temporal salience. FIGURE 2(a) and FIGURE 2(b) show an example of two consecutive frames f_{t-1} and f_t, respectively.

Algorithm 1 Temporal Visual Salience Modeling for Visually Anonymized Silhouette Generation

```
Consider 2 consecutive frames, $f_t$ and $f_{t-1}$.
Find the difference map $D'_t$.
for Each user-defined threshold $\tau_h$ do
    Filter $D'_t$ to get $D_t$ using Eq. (2).
    for each location $(x, y)$ do
        Consider $N \times N$ pixel block centred on $(x, y)$.
        Compute 2DFFT of block $b_m$.
        Compute PSD $S_{b_m}$ of $b_m$ using Eq. (3).
        Normalise $S_{b_m}$.
        Compute the entropy $\mathcal{E}_t^{\tau_h}(x, y)$.
    end for
end for
Compute the weighted entropy, $\hat{\mathcal{E}}_t(x, y)$, using Eq. (4).
Map $\hat{\mathcal{E}}_t(x, y)$ into [0, 255].
Gaussian filter the map to get the salience map, $S_t$.
Output $S_t$.
```

FIGURE 2(c) shows D'_t for a chosen threshold, τ_h. FIGURE 2(d) shows the $\mathcal{E}_t^{\tau_h}(x, y)$ for pixels along two lines for $x = 114$ (in blue) and $x = 350$ (in red). There is no temporal activity along $x = 114$, hence $\mathcal{E}_t^{\tau_h}(114, y)$ values are zero. On the other hand, $\mathcal{E}_t^{\tau_h}(350, y)$ consists of non-zero values at pixels corresponding to locations where temporal activity is present.

The distribution of the temporal visual salience magnitudes on a frame is essentially based on the magnitude of the changes in the intensities of the pixels caused by the motion present in the action. If the intensity is changed significantly,



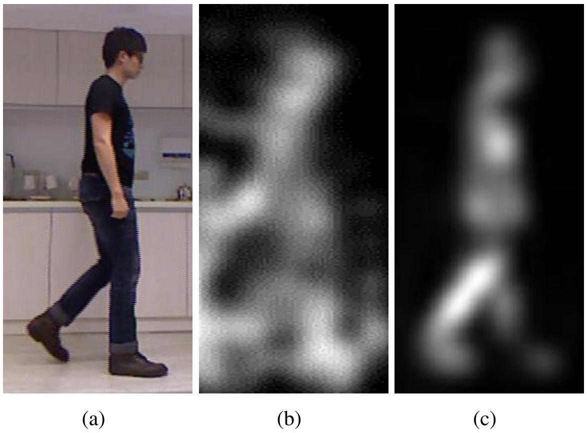
(a)

(b)

(c)

(d)

FIGURE 2. Illustration of $\mathcal{E}_t^{\tau_h}(x, y)$: (a) Frame f_{t-1}; (b) Frame f_t; (c) Frame differenceD'_t with threshold τ_h; (d) $\mathcal{E}_t^{\tau_h}(114, y)$ (in blue) and $\mathcal{E}_t^{\tau_h}(350, y)$ (in red).



(a)

(b)

(c)

FIGURE 3. An example of a generated silhouettes: (a) shows an original frame; (b) shows the silhouette using $\mathcal{E}_t^{\tau_h}(x, y)$ with $\tau_h = 4$; and (c) shows the silhouette using $\hat{\mathcal{E}}_t(x, y)$.

this produces a temporal saliency with strongly highlighting and vice versa. Furthermore, proposing Eq. (4) has another essential goal of suppressing the global changes, *i.e.*, global motion, that can come from the background objects of camera motion. FIGURE 3 shows an example of generated silhouettes using the proposed method. It demonstrates the benefit of using multiple thresholds to compute the weighted entropy, $\hat{\mathcal{E}}_t(x, y)$. It can be seen in FIGURE 3(c) that the generated silhouette further highlights the most dynamic body parts used in the action compared to the rest since the moving parts are represented with high temporal visual salience magnitude values.

B. HUMAN ACTION RECOGNITION IN THE VISUALLY ANONYMIZED DOMAIN

Our proposed silhouette generation for visually anonymizing in Section III-A produces a gray scale map corresponding to the temporal visual salience due to the motion in the sequence. In this section, we present the proposed methodology for analysing these silhouette maps for HAR. Our approach aims to construct a compact descriptor by exploiting the the temporal visual salience captured in the silhouettes. Most current HAR descriptors are based on the original or

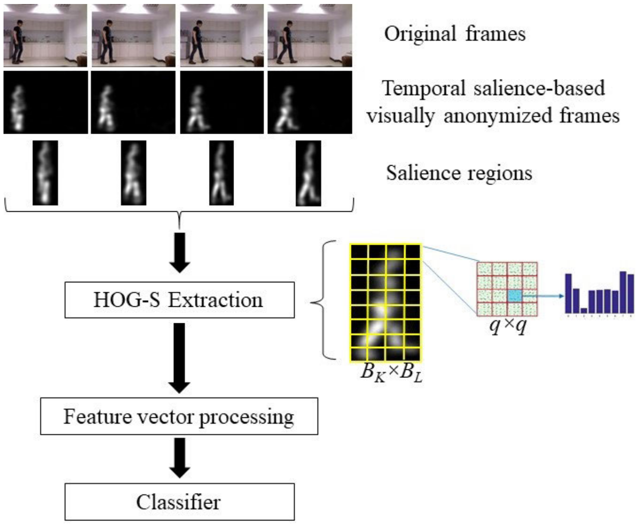


FIGURE 4. Proposed approach for HAR in the visually anonymized domain (saliency-based silhouettes).

raw video data and estimated motion from video for extracting important features. Since motion information is already encapsulated in our silhouettes, our approach can effectively the analyze the video without needing to access to the original visually non-anonymized video or without computing complex motion estimations. To achieve this, we propose histograms of gradients in salience (*HOG-S*), which is a local descriptor exploring the temporal visual salience captured in our visually anonymizing silhouettes.

The *HOG-S* focuses on the salience region, R_t, spanning in a rectangular bounding box of $K \times L$ pixels, from the silhouette in frame t. Major steps of our approach include *HOG-S* feature vector extraction from the bounding boxes, *HOG-S* feature vector processing and training a classifier as illustrated in the block diagram in FIGURE 4. We start by computing gradients, $\nabla R_t = (d_x, d_y)$ for each pixel in the region R_t, where d_x and d_y represent the horizonatla and vertical components approximated by finite differences. The gradient magnitude, G_t, and the direction, θ_t, are computed as follows:

$$G_t = \sqrt{d_x^2 + d_y^2}, \tag{5}$$

$$\theta_t = arctan\left(\frac{d_y}{d_x}\right). \tag{6}$$

R_t is partitioned into $B_K \times B_L$ blocks, each containing $pn \times pn$ pixels. Then each block is further partitioned into $p \times p$ patches, with each patch containing $n \times n$ pixels. The gradient magnitudes and the corresponding directions in each patch are formed into 9-bin histograms and all histograms are concatenated into a single feature vector, $\vec{v}_t$, of length $9p^2 B_K B_L$. This is followed by normalizing the vector as follows:

$$\hat{v}_t = \frac{\vec{v}_t}{\|\vec{v}_t\|_2^2}, \tag{7}$$

where $\|\cdot\|_2$ denotes l^2-norm. However, just considering individual $\hat{v}_t$ for individual frames cannot perfectly marginalise among features from other frames in accordance with the variations inside the action itself and similarities among other actions. This is addressed by considering the accumulated temporal changes to the feature vectors, $\hat{V}_t = \{\hat{v}_0, \hat{v}_1, \hat{v}_2, \cdots, \hat{v}_t\}$ up to frame t to compute the final feature vector, $\tilde{v}_t$, at the time instant, t, as follows:

$$\tilde{v}_t = \left| \sum_{k=0}^{\lfloor (t-1)/2 \rfloor} \hat{v}_{t-2k} - \sum_{k=0}^{\lfloor (t-1)/2 \rfloor} \hat{v}_{t-2k-1} \right|, \tag{8}$$

where $|\cdot|$ denotes the absolute value of the vector elements.

This is followed by applying the principle component analysis (PCA) on the set of feature vectors $\tilde{v}_t$ of the sequence in order to reduce the dimensionality of the *HOG-S* descriptor and to maximise the variance leading to improving the discrimination of the *HOG-S* descriptors. Finally a classifier is trained using these feature vectors to recognise the human actions in the video. We have considered two classifiers, support vector machine (SVM) and K-nearest neighbour (KNN) for evaluating our proposed method as presented in Section IV.

IV. PERFORMANCE EVALUATION

In this section, we present the evaluation of the proposed method in terms of its performance in both visual anonymization and human action recognition in the visually anonymized domain. The datasets used and the experimental parameters are shown in Section IV-A. Firstly, for the completion of evaluation, we evaluate the performance of our proposed temporal visual salience modeling and compare with the existing video salience modeling to justify the suitability of our approach for the considered application in Section IV-B. Then, we evaluate the effectiveness of the proposed anonymization method by evaluating the recognizability of the humans in video sequences and the utility of such anonymized video by recognizing the activities they do. We evaluate both these objectives firstly using human observers[1] by conducting subjective evaluations as shown in Section IV-C. Finally, the performance of HAR using the proposed *HOG-S* features in the visually anonymized domain is presented in Section IV-D.

A. EXPERIMENTAL SETUP

Six publicly available HAR datasets, namely Weizmann [52], KTH [53], DHA [54], UIUC1 [55], UCF Sports [56] and HMDB51 [57], are used to evaluate the proposed work. Each sequence in these datasets comprises of a single action.

The Weizmann dataset contains $V = 93$ low resolution (144×180) 50 frame per second (fps) video sequences showing nine different people. Each of them performing

[1] This research has received The University of Sheffield ethics approval under application No 024404.

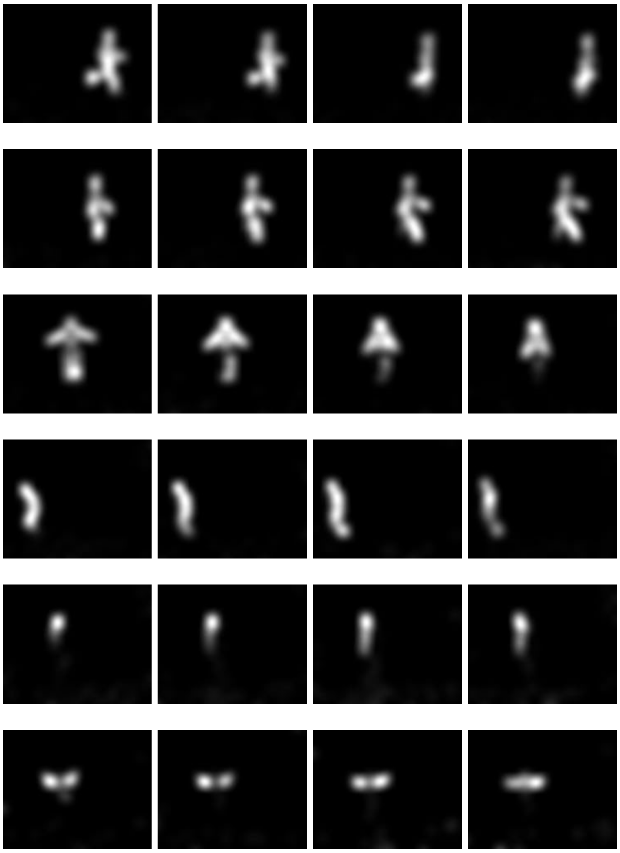


FIGURE 5. Temporal salience based silhouettes for six actions from Weizmann dataset: ***Row 1***: run, ***Row 2***: walk, ***Row 3***: jack, ***Row 4***: jump, ***Row 5***: One hand waving and ***Row 6***: Two hands waving. Each row shows how the silhouettes are changed for different frames over time.

$Q = 10$ different actions, *e.g.*, bend, run, walk, skip, jack, jump, pjump, side, one hand wave and two hands wave. This dataset is recorded using a single static camera.

The KTH dataset contains $V = 597$ video sequences showing $Q = 6$ action classes, *e.g.*, boxing, handwaving, handclapping, jogging, running and walking. There are 25 different subjects performing the actions in four different scenarios, *e.g.*, three are outdoor and one is indoor. This dataset is recorded with four different cameras to capture the action of the subject in the scene from different views. There three static cameras and another one to record the actions with zooming. The sequences are captured over a homogeneous background with a static camera recording 25 frames per second. Each video has a resolution of 160 × 120.

Depth-included Human Action (DHA) dataset contains $Q = 23$ action classes performed by participating 21 different individuals (12 males and 9 females). It is recorded using a static Kinect camera in three different scenes with 480 × 640 resolution. The RGB versions of videos are used in the experiments.

UIUC1 dataset includes $V = 532$ sequences (1024 × 768, 15 fps) showing $Q = 14$ human actions, *i.e.*, walking, running, jumping, waving, jumping jacks, clapping, jump from situp, raise one hand, stretching out, turning, sitting to standing, crawling, pushing up and standing to sitting. These actions are performed by 8 persons and recorded using a single static camera.

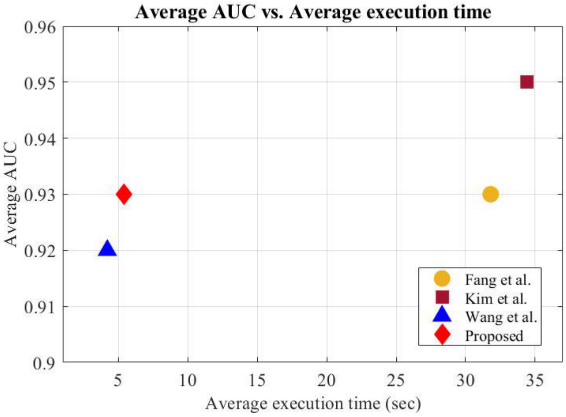


FIGURE 6. Average AUC and the execution time per frame measured by seconds for the exiting video salience modeling algorithms and the proposed method.

TABLE 2. Average AUC and the corresponding execution time of the proposed method and the existing work for video salience modeling.

Datasets	*Fang* [25]	*Kim* [28]	*Wang* [26]	*Proposed*
DHA - AUC	0.87	0.92	0.82	0.91
Weizmann - AUC	0.95	0.96	0.95	0.95
UIUC1 - AUC	0.98	0.98	0.98	0.92
Average AUC	0.93	0.95	0.92	0.93
Average time (sec)	31.8	34.4	4.21	5.4

UCF Sports dataset includes a total of $V = 150$ sequences with the resolution of 720 × 480 represents $Q = 10$ actions. This dataset represents a natural collection of actions including a wide variation in the scenes and viewpoints. The actions included in this dataset are: Diving, Golf Swing, Kicking, Lifting, Riding Horse, Running, Skate Boarding, Swing-Bench, Swing-Side, Walking.

Finally, Human Motion Database (HMDB51) dataset, which is one of the largest datasets used in HAR, contains $V = 6849$ clips distributed in $Q = 51$ action classes. Each video clip has around $20 - -1000$ frames. The action categories of this dataset can be grouped into five types based on the body movements. This dataset is considered challenging due to containing clips collected from the Internet and YouTube. Thus, this dataset can be considered as a real-world video clip collection.

In the experiments, we use $N = 3$ and $h = 7$ for evaluating the proposed visual anonymization algorithm. The weighted entropy maps $\hat{\mathcal{E}}_t(x, y)$ are smoothed using a 2D Gaussian kernel with $\sigma = 6$. All maps are resized to the resolution 256 × 256 to apply the same parameters on all datasets. We adopt a bounding box approach with $K = 168$

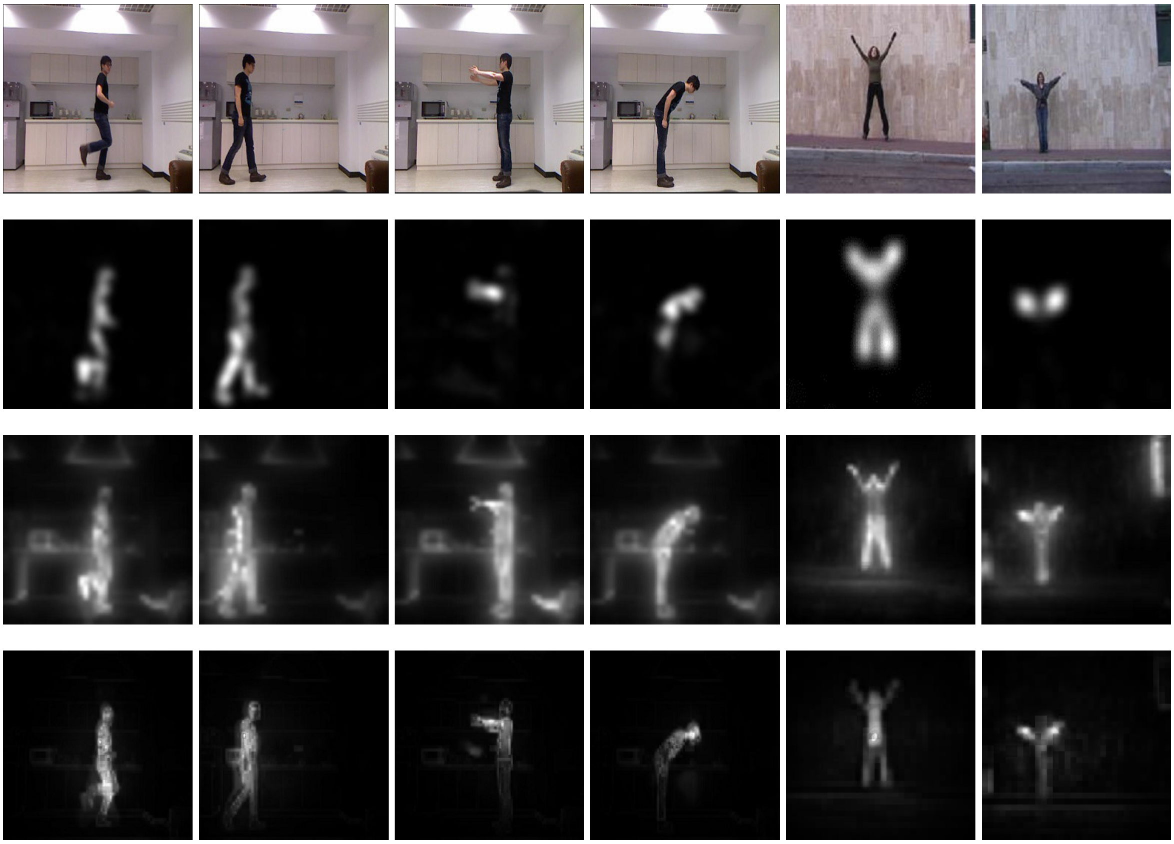


FIGURE 7. Comparison salience maps for different actions using the proposed metod and the existing methods. *Row 1*: original RGB frames, *Row 2*: corresponding temporal salience maps using our proposed method, *Row 3*: corresponding salience maps using Kim *et al.* [28] and *Row 4*: corresponding salience maps using Fang *et al.* [25]. *Column 1-4*: four actions from the DHA dataset and *Column 5-6*: two actions from the Weizmann dataset.

$L = 72$ resolution to crop the salient contents of the silhouette. We have used the parameters $n = 4$ and $p = 4$ in *HOG-S* extraction resulting in 144 length descriptors for each block in our experiments. The final length of the *HOG-S* descriptor for each frame is 23040.

B. EVALUATION OF TEMPORAL VISUAL SALIENCE MODELING

FIGURE 5 shows the generated temporal visual salience maps for visually anonymizing a few sequences of actions in the Weizmann dataset. As we can see that the silhouette for a specific action is changed for different frames over time as the motion content due to the action varies. For instance, in the case of jacking action, third row in FIGURE 5, the silhouette has a different pattern every time, as some parts are attenuated and others gain extra highlighting. In addition, the algorithm generates different salience maps for one-hand waving and two-hands waving actions, as we can see in *row 5* and *row 6*, respectively, since the patterns of these two actions are different. This representation is crucial to create a useful abstract at each frame to extract an efficient action description, which accurately identifies the variation between the actions, while the video is visually anonymized.

Although our method focuses only on temporal visual salience, for completeness of this paper, we compare our temporal salience modeling with three other existing work Fang *et al.* [25], Kim *et al.* [28] and Wang *et al.* [26], which are mainly full video visual salience modeling considering both temporal and spatial salience cues. However, our algorithm just considers temporal salience cues only. In this way, we can make sure that the full video is visually anonymized (using black pixels for salient areas) while showing only the gray scale salience map silhouette corresponding to the temporal salient regions related to the action. TABLE 2 and FIGURE 6 show the Area Under Curve (AUC) values measuring the accuracy of salient region detection and the average time take for computation for our proposed method and the exiting work considering three datasets. These results show that the proposed method, which only models temporal salience, has comparable accuracy in terms of AUC with the existing methods, while taking low computational time. Examples of salience maps for various action sequences from DHA and Weizmann datasets using our proposed method and existing work are shown in FIGURE 7. It is evident that our proposed salience maps only captures the body parts relevant to to the action, where as, other methods capture other spatial

TABLE 3. Evaluation group details.

Group No.	Dataset	No. of video sequences	No. of participants (Male - Female)
1	DHA	30	8 (7 - 1)
2	KTH	24	7 (5 - 2)
3	Weizmann	24	7 (7 -0)
4	UIUC1	30	8 (6 - 2)

information and the full body which are not relevant to the action.

C. EVALUATION OF THE PROPOSED VISUAL ANONYMIZATION ALGORITHM

We evaluated the effectiveness of the proposed visual anonymization using human observers. A survey with 30 individuals participants was conducted to evaluate the proposed method and state-of-the-art filtering algorithms for visual anonymization. In this survey, the participants were divided into four groups, where each group evaluated a specific dataset anonymized using the proposed methods and the existing methods. The datasets of DHA, KTH, Weizmann, and UIUC1 were used in this subjective evaluation.

In total, 108 anonymized video sequences for different actions were selected equally from five existing methods (blurring with $\sigma = 5$, blurring with $\sigma = 8$, pixelation, solid silhouette and binary silhouette) and the proposed method. These sequences have been spread out into four groups and each group was allocated to separate a set of participants for evaluation. TABLE 3 shows information of each group of evaluation and the number of sequences that have been assigned to each group. FIGURE 8 shows a few example frames from the sequences used in the survey and their corresponding anonymized frames using the existing methods and the proposed method.

The purpose of the survey is two-fold. Firstly it aims to find out the effectiveness of the proposed method visually anonymization. Secondly, to evaluate whether the utility of the video is affected due to the anonymization. In this case the utility was considered as the ability for an observer to accurately recognize the action present in the sequence. Three questions, shown in TABLE 4, were included in the survey to achieve these two purposes.

The first question aims to evaluate the level of visual anonymization achieved by a particular method as perceived by the observer. They are asked to score the level of anonymity on a discrete scale from 0 (no anonymization) to 5 (perfect anonymization). The score is regarded to which one they thought that could provide enough protection and reducing the concern about privacy protection. The second question collects the identity attributes, such as, gender, apparent age, facial features, clothes, hair and race, that can be recognised by the participants. These attributes are considered sensitive information that has to be protected by a visual privacy preservation model. The unmeasurable attributes were not considered due to the difficulty to determine them in the visual domain. The response to this question needs to be compatible with that for the first question. For example, a score of 5 for the anonymization level means none of the

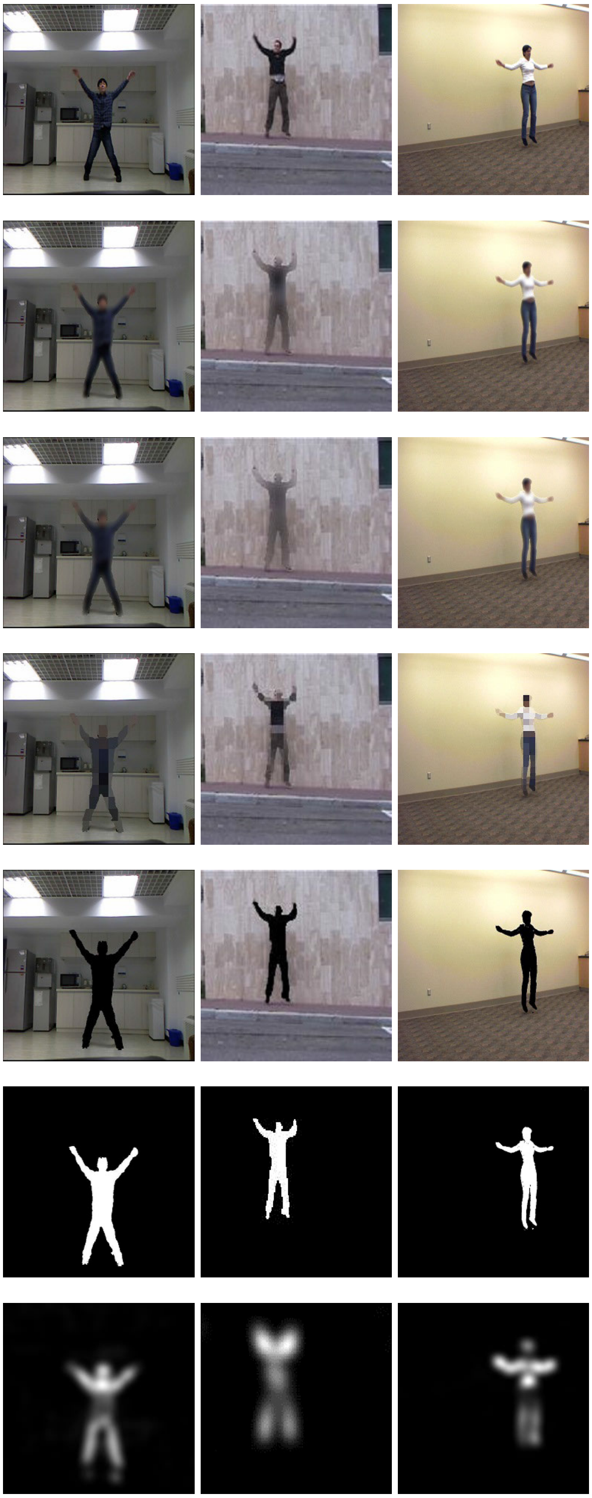


FIGURE 8. Example frames from different datasets with corresponding visually anonymized frames. *Row 1*: the original frame, *Row 2*: blurring with $\sigma = 5$, *Row 3*: blurring with $\sigma = 8$, *Row 4*: pixelation, *Row 5*: solid silhouette, *Row 6*: binary and *Row 7*: the proposed method. *Column 1*: DHA, *Column 2*: Weizmann and *Column 3*: UIUC1 datsets.

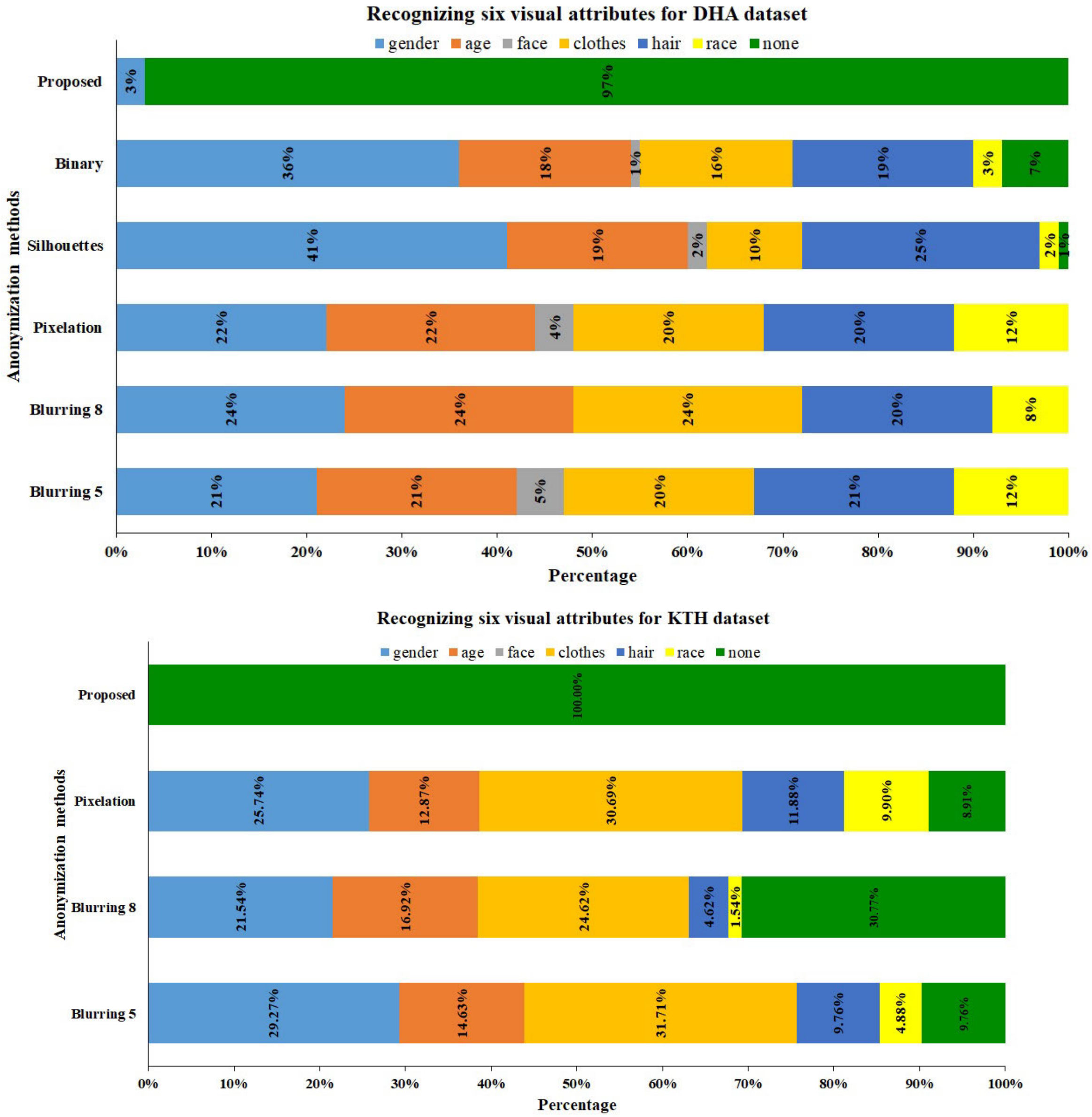


FIGURE 9. Percentages of appearance attribute recognisability for different visual anonymization methods for DHA and KTH datasets.

identity clues can be recognized from the anonymized video. Finally, the third question estimates the ability of anonymization method to retain useful information that can be used to identify the human action present in the video. This quality relies on the level of anonymity. In other words, if we need to increase the anonymity, the quality of the information has to be discarded and vice versa. The participants were asked to label the action presented in the obfuscated sequence using the information that was retained in the concealment model.

At the the beginning of a survey session, the purpose of the evaluation is conveyed to the survey participants. The region of anonymity of a scene is restricted to the human in the scene, but not for the background. The test video set used in the survey consists of various people performing various actions. We aimed to minimize the repeat of the same person doing different actions. Using the same video sequences with versions can help the participants to use their memory to recall the missed details and/or biased to the same answer ignoring the difference between the models. However, in few cases we use two different models for the same sequence in order to analyse the ability of the participants to recognise between them and if the method can make the difference for the participant or not.

As shown in TABLE 3, the number of the video sequences in this evaluation is 108 sequences, distributed as follows: DHA=30, UIUC1=30, KTH=24 and Weizmann=24.

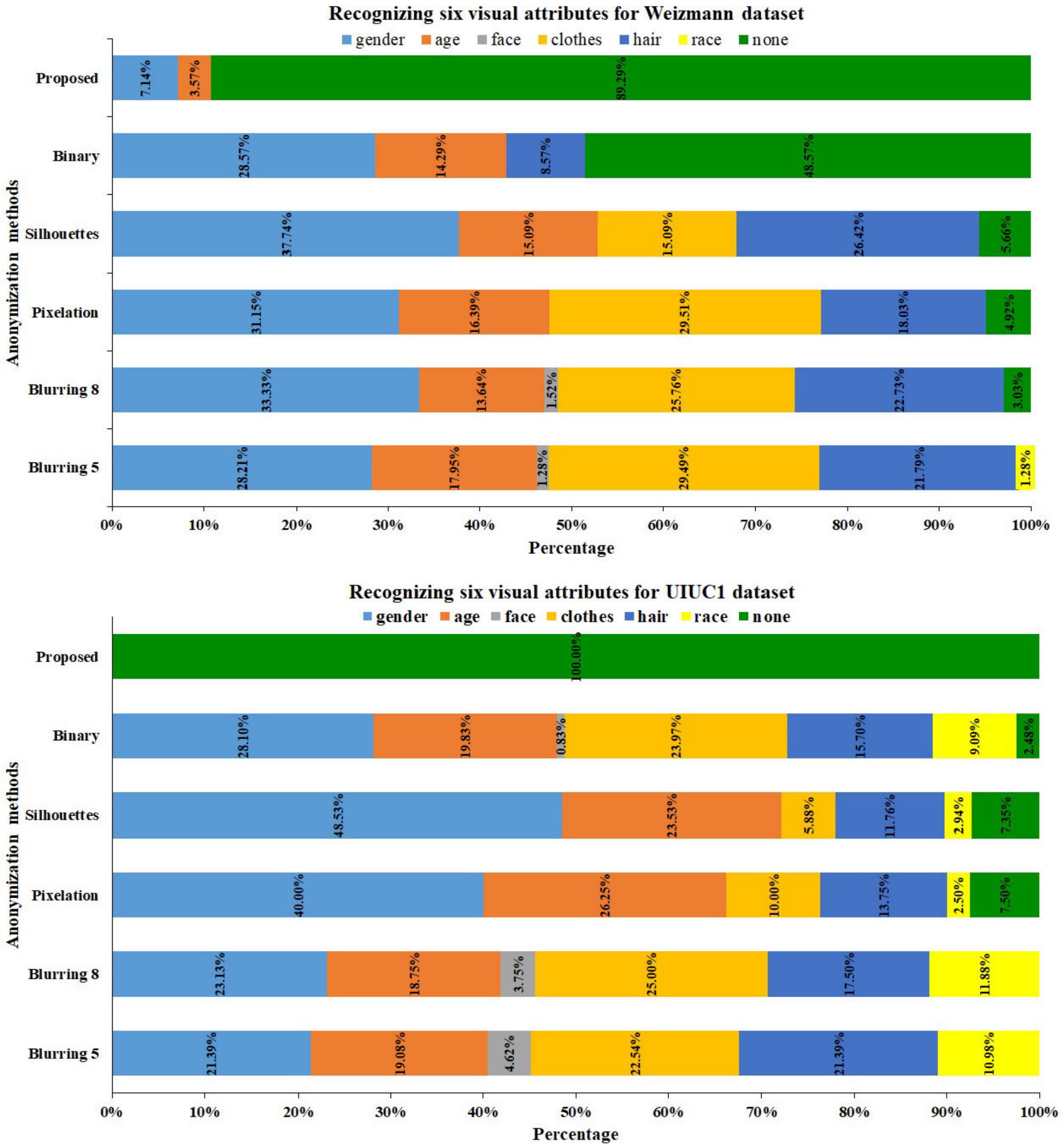


FIGURE 10. Percentages of appearance attribute recognisability for different visual anonymization methods for Weizmann and UIUC1 datasets.

The number of video sequences that have been used in the evaluation depends on the size of the dataset and the number of action labels in each dataset. Thus, the number of the video sequence is distributed among different anonymization models evaluated. Six models were evaluated for DHA, Wizemann and UIUC1 datasets while four models were evaluated for KTH, as Silhouette and Binary masks were not available for the actions in the KHA dataset. With three questions per sequence, the number of responses collected for each dataset is as follows: DHA=720, UIUC1=720, KTH=504 and Weizmann=504 witha total of 2448 responses. The rest of this sub-section shows an analysis of the survey responses to all three questions.

1) ANALYSIS OF RESPONSES TO QUESTION 1

The responses include an anonymity score for each visually anonymized video sequence. We define the Mean Anonymity Score (MAS) for a given method for a given dataset by taking the mean score of all the responses received for the given dataset using the given method. MAS for the six methods, four datasets and the average MAS for all datasets per method

TABLE 4. The questions and their corresponding responses.

#	Question	Possible Responses
1	How well the person is visually anonymized in the video sequence?	A discrete score in the range from 0 (not anonymized) to 5 (perfectly anonymized).
2	Can you recognise the following features from the anonymized sequence ?	tick one or more choices from the following: gender, age, face, clothes, hair, race, and none.
3	Which of the following activities are visible in the anonymized video sequence?	[walking, running, jumping, standing, waving, $\cdots$, I don't know]

TABLE 5. The Mean Anonymity Score (MAS) considering all the participants' response for all datasets and anonymizing methods.

Anony-mizing Method	Datasets				Average for all datasets
	DHA	KTH	Weizmann	UIUC1	
Blurring ($\sigma = 5$)	1.9	3.14	2.28	1.83	2.42
Blurring ($\sigma = 8$)	2.1	3.74	2.71	2.1	2.85
Pixelation	2.07	2.79	2.93	3.23	2.98
Silhouette	3.17	NA	3.04	3.55	3.3
Binary	3.4	NA	4.07	3.65	3.86
Proposed	4.97	4.79	4.86	4.8	4.82

are shown in TABLE 5. A MAS of 0 corresponds to the least anonymity and a MAS of 5 corresponds to the highest anonymity. According to these results, the proposed method has achieved the highest MAS compared to all other methods for all datasets.

2) ANALYSIS OF RESPONSES TO QUESTION 2

The second question aims to collect more details about the appearance attributes recognizable in the anonymized sequences. The question 2 specifically enquires the participants about recognizability of six attributes, *i.e.*, gender, apparent age, facial features, clothes, hair and race of the humans in the test sequences. We have also included the option "none" to indicate if any of the above attributes is not identified. FIGURE 9 and FIGURE 10 summarize the responses presented in stack bars as percentages for each visual anonymization method for different datasets. The proposed anonymization method has recorded between $89\% - 100\%$ of non-recognizable attributes (as shown in green in FIGURE 9), which is the highest compared to the existing anonymization methods. This high level of anonymization proves that our proposed temporal visual salience modeling achieves better anonymity compared to the existing spatial (frame based) approaches for visual anonymization. This result also matches with the highest MAS score reported in Question 1.

TABLE 6. The action recognition accuracy rates considering all participant responses for all datasets and anonymizing methods.

Anony-mizing Method	Datasets				Average for all datasets
	DHA	KTH	Weizmann	UIUC1	
Blurring ($\sigma = 5$)	0.87	0.71	0.98	0.88	0.85
Blurring ($\sigma = 8$)	0.8	0.67	0.86	0.75	0.76
Pixelation	0.7	0.79	0.71	0.83	0.77
Silhouette	0.93	NA	0.8	0.90	0.84
Binary	0.4	NA	0.75	0.83	0.76
Proposed	0.8	0.69	0.82	0.78	0.80

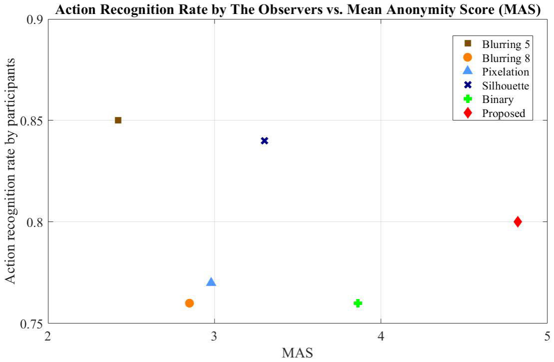
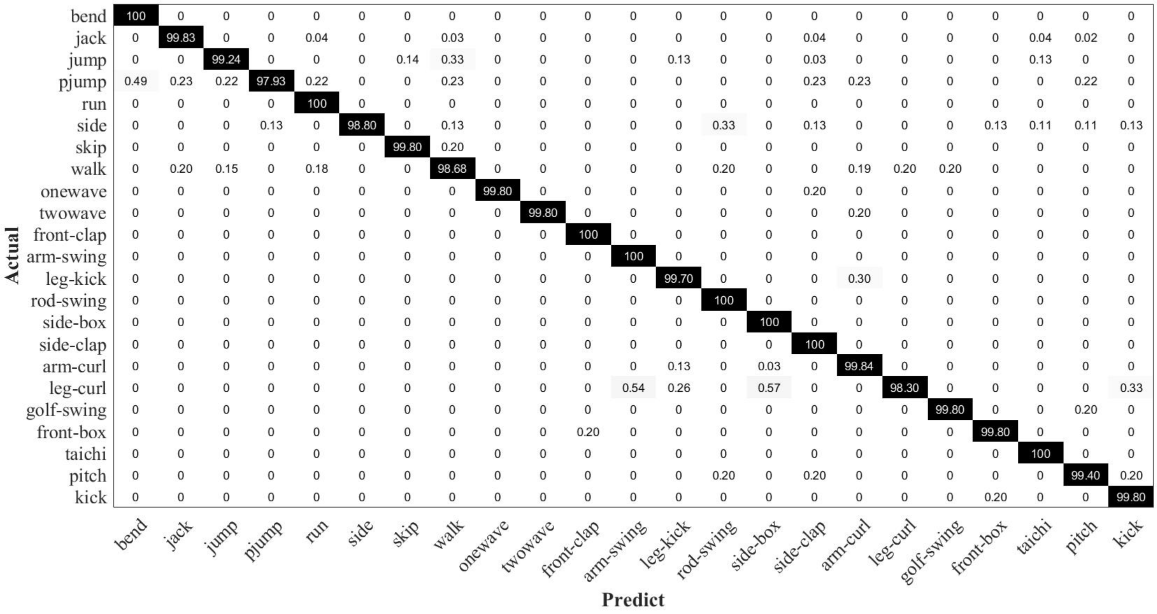


FIGURE 11. Action recognition rate by the participants (the utility of the visually anonymized sequence) vs. the Mean Anonymity Score (MAS) for the existing visual anonymization methods and the proposed method.

3) ANALYSIS OF RESPONSES TO QUESTION 3

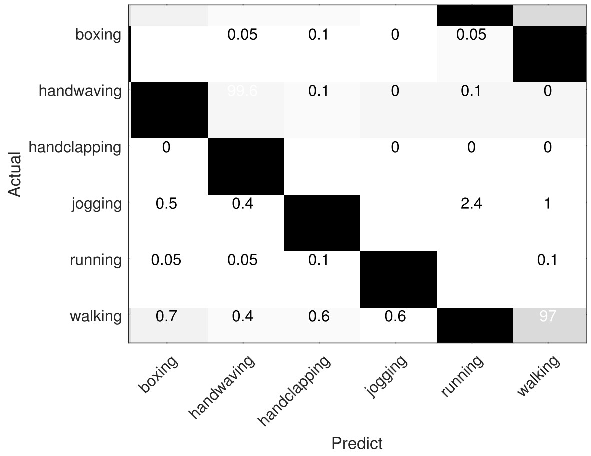
This question evaluates the utility of the anonymized sequence in action recognition as perceived by the participants in the survey. The number of accurately recognized actions in video sequences normalised with respect to the total responses given by the participants for a given anonymizing method for all four datasets are shown in TABLE 6. It is evident from the table that the action recognition rates by participants for some anonymization methods are better than that for the sequences that use the proposed anonymization method. On one hand, for instance, blurring model with $\sigma = 5$ seems to achieve better accuracy rates from the viewpoint of the participants. On the other hand, this means that the quality of the visual anonymity is low, so that it has not distorted the perception of motion present in the action. It can be seen that for some methods there can be a trade-off between the anonymity and the utility of the anonymized sequence. For this reason, we evaluate the anonymization methods using the joint performance in anonymization and utility as shown in FIGURE 11.

It is clear from FIGURE 11 that the proposed temporal visual salience-based anonymity maps achieve the highest level of visual anonymity, outperforming the existing methods. It is also evident that the higher the anonynmity the lower the utility as can be seen for blurring based methods.



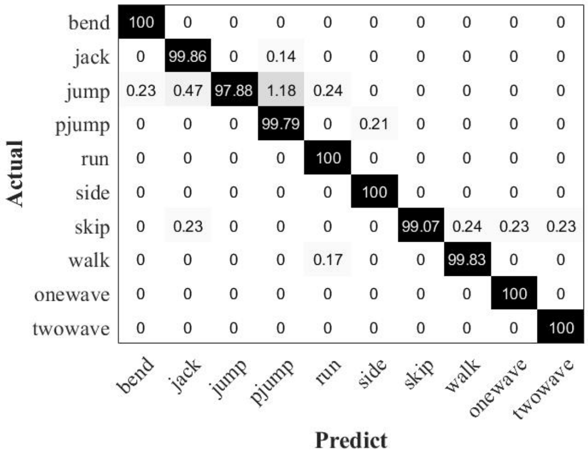
	bend	jack	jump	pjump	run	side	skip	walk	onewave	twowave	front-clap	arm-swing	leg-kick	rod-swing	side-box	side-clap	arm-curl	leg-curl	golf-swing	front-box	taichi	pitch	kick
bend	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
jack	0	99.83	0	0	0.04	0	0	0.03	0	0	0	0	0	0	0	0.04	0	0	0	0	0.04	0.02	0
jump	0	0	99.24	0	0	0	0.14	0.33	0	0	0	0	0.13	0	0	0.03	0	0	0	0	0.13	0	0
pjump	0.49	0.23	0.22	97.93	0.22	0	0	0.23	0	0	0	0	0	0	0	0.23	0.23	0	0	0	0	0.22	0
run	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
side	0	0	0	0.13	0	98.80	0	0.13	0	0	0	0	0	0.33	0	0.13	0	0	0	0.13	0.11	0.11	0.13
skip	0	0	0	0	0	0	99.80	0.20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
walk	0	0.20	0.15	0	0.18	0	0	98.68	0	0	0	0	0	0.20	0	0	0.19	0.20	0.20	0	0	0	0
onewave	0	0	0	0	0	0	0	0	99.80	0	0	0	0	0	0	0.20	0	0	0	0	0	0	0
twowave	0	0	0	0	0	0	0	0	0	99.80	0	0	0	0	0	0	0.20	0	0	0	0	0	0
front-clap	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
arm-swing	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
leg-kick	0	0	0	0	0	0	0	0	0	0	0	0	99.70	0	0	0	0.30	0	0	0	0	0	0
rod-swing	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
side-box	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
side-clap	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
arm-curl	0	0	0	0	0	0	0	0	0	0	0	0	0.13	0	0.03	0	99.84	0	0	0	0	0	0
leg-curl	0	0	0	0	0	0	0	0	0	0	0	0.54	0.26	0	0.57	0	0	98.30	0	0	0	0	0.33
golf-swing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99.80	0	0	0.20	0
front-box	0	0	0	0	0	0	0	0	0	0	0.20	0	0	0	0	0	0	0	0	99.80	0	0	0
taichi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
pitch	0	0	0	0	0	0	0	0	0	0	0	0	0	0.20	0	0.20	0	0	0	0	0	99.40	0.20
kick	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.20	0	0	99.80

FIGURE 12. The confusion matrix of the DHA dataset using the proposed method+KNN (Overall accuracy: 99.59%).



	boxing	handwaving	handclapping	jogging	running	walking
boxing		0.05	0.1	0	0.05	
handwaving		99.6	0.1	0	0.1	0
handclapping	0			0	0	0
jogging	0.5	0.4			2.4	1
running	0.05	0.05	0.1			0.1
walking	0.7	0.4	0.6	0.6		97

FIGURE 13. The confusion matrix of the KTH dataset using the proposed method+KNN (Overall accuracy: 99.06%).



	bend	jack	jump	pjump	run	side	skip	walk	onewave	twowave
bend	100	0	0	0	0	0	0	0	0	0
jack	0	99.86	0	0.14	0	0	0	0	0	0
jump	0.23	0.47	97.88	1.18	0.24	0	0	0	0	0
pjump	0	0	0	99.79	0	0.21	0	0	0	0
run	0	0	0	0	100	0	0	0	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0.23	0	0	0	0	99.07	0.24	0.23	0.23
walk	0	0	0	0	0.17	0	0	99.83	0	0
onewave	0	0	0	0	0	0	0	0	100	0
twowave	0	0	0	0	0	0	0	0	0	100

FIGURE 14. The confusion matrix of the Weizmann dataset using the proposed method+KNN (Overall accuracy: 99.66%).

In conclusion, the proposed approach results in excellent visual anonymity while transforming the original colour pixels into the temporal visual salience leading to an action-related informative domain, which can provide a good indication of the actions in sequences as perceived by the human participants in the survey. In the following section, we demonstrate the utility of our approach in machine-based HAR.

D. HUMAN ACTION RECOGNITION USING THE PROPOSED HOG-S *FEATURES IN THE VISUALLY ANONYMIZED DOMAIN*

Experiments reported in this section perform an objective evaluation of the visually anonymized sequences from the proposed anonymization method from the machine perception. The anonymized sequences are analysed using the *HOG-S* features for HAR as proposed in Section III-B. We report its performance in five datasets using both KNN and SVM classifiers with five fold cross-validation and compare with the existing methods. The number of PCA components that were used for KNN and SVM classifiers to get the HAR accuracy rates reported in this section are shown in TABLE 7. It must be noted that these numbers are much lass than the original feature length, which is 23040.Note that for HMDB51 dataset, we evaluated only using KNN classifier.

TABLE 8 shows the HAR performance of the proposed *HOG-S* features in the temporal visual salience-based visually anonymized domain and compares with the existing methods. The existing methods that are based on deep learning techniques are marked with the suffix (DL) in TABLE 8.

TABLE 7. The number of PCA components used for five datasets.

Dataset	PCA components	
	KNN	SVM
Weizmann	400	900
KTH	700	1300
DHA	300	1400
UIUC1	500	1700
UCF Sports	400	400
HMDB51	900	–

TABLE 8. Human Action Recognition accuracy (%) of the proposed method and comparison with the existing work for all datasets. Existing methods that use deep learning are marked as (DL).

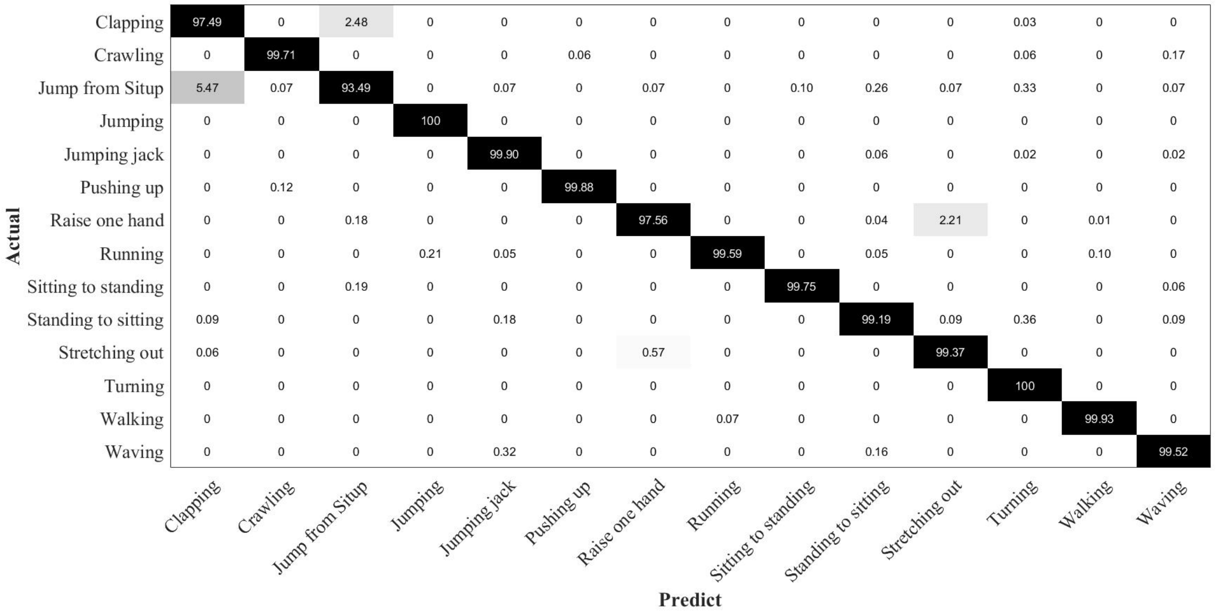
Dataset	Method	Accuracy (%)
DHA	Gao et al. (2015) [58]	95
	Liu et al. (2017) [59]	95.45
	Liu et al. (2018) [60]	95.44
	Zhang et al. (2017) [61]	96.69
	Proposed+SVM	**97.98**
	Proposed+KNN	**99.59**
	Proposed+PCA+SVM	**99.18**
	Proposed+PCA+KNN	**99.73**
KTH	Liu et al. (2013) [62]	94.8
	Shi et al. (2017) [63] (DL)	96.8
	Zhang *et al.*(2020) [64]	97.4
	Yadav et al. (2016) [65]	98.20
	Avola *et al.*(2019) [66] (DL)	98.33
	Proposed+SVM	**98.53**
	Proposed+KNN	**99.06**
	Proposed+PCA+SVM	**99.87**
	Proposed+PCA+KNN	**99.94**
Weizmann	Osayamwen and Tapamo (2019) [67] (DL)	96.40
	Wu and Shao (2013) [68]	97.98
	Zeng et al. (2018) [69]	98.77
	Xu et al. (2017) [70]	99.1
	Angelini et al. (2018) [44]	100
	Parisi et al. (2017) [71] (DL)	100
	Proposed+SVM	**99.46**
	Proposed+KNN	**99.66**
	Proposed+PCA+SVM	**99.81**
	Proposed+PCA+KNN	**99.74**
UIUC1	Wang et al. (2013) [72]	98.4
	Zhang et al. (2015) [73]	98.87
	Shan et al. (2015) [74]	98.9
	Proposed+SVM	**99.06**
	Proposed+KNN	**99.15**
	Proposed+PCA+SVM	**99.73**
	Proposed+PCA+KNN	**99.72**
UCF Sports	Wang et al. (2017) [75](DL)	93.6
	Ghodrati et al.(2017) [76] (DL)	95.7
	Siddiqi et al. (2019) [77]	96.22
	Tu *et al.*(2018) [78] (DL)	97.50
	Dai *et al.*(2020) [79](DL)	98.6
	Proposed+SVM	**98.15**
	Proposed+KNN	**99.71**
	Proposed+PCA+SVM	**99.23**
	Proposed+PCA+KNN	**99.89**
HMDB51	Girdhar, et al. (2017) [80]	66.9
	Ma *et al.*(2019) [81](DL)	69.0
	Song *et al.* [82](DL)	72.7
	Carreira and Zisserman (2017) [83]	80.9
	Choutas et al. (2018) [84]	80.9
	Zheng *et al.*(2020) [85] (DL)	81.1
	Wang et al. (2019) [86] (DL)	82.48
	Proposed+KNN	**99.03**
	Proposed+PCA+KNN	**99.19**

TABLE 9. Accuracy rates (%) for machine-based Human Accurate Recognition in visually anonymized domain for various visually anonymizing methods.

Anonymizing Method	DHA	Weizmann	UIUC1	Average for all datasets
Blurring ($\sigma = 5$)	94.62	94.05	98.94	95.87
Blurring ($\sigma = 8$)	94.51	91.15	99.18	94.95
Pixelation	79.35	69.16	93.23	80.58
Silhouette	93.9	89.36	98.99	94.08
Binary	91.97	93.61	98.19	94.59
Proposed	99.39	99.64	99.15	99. 39

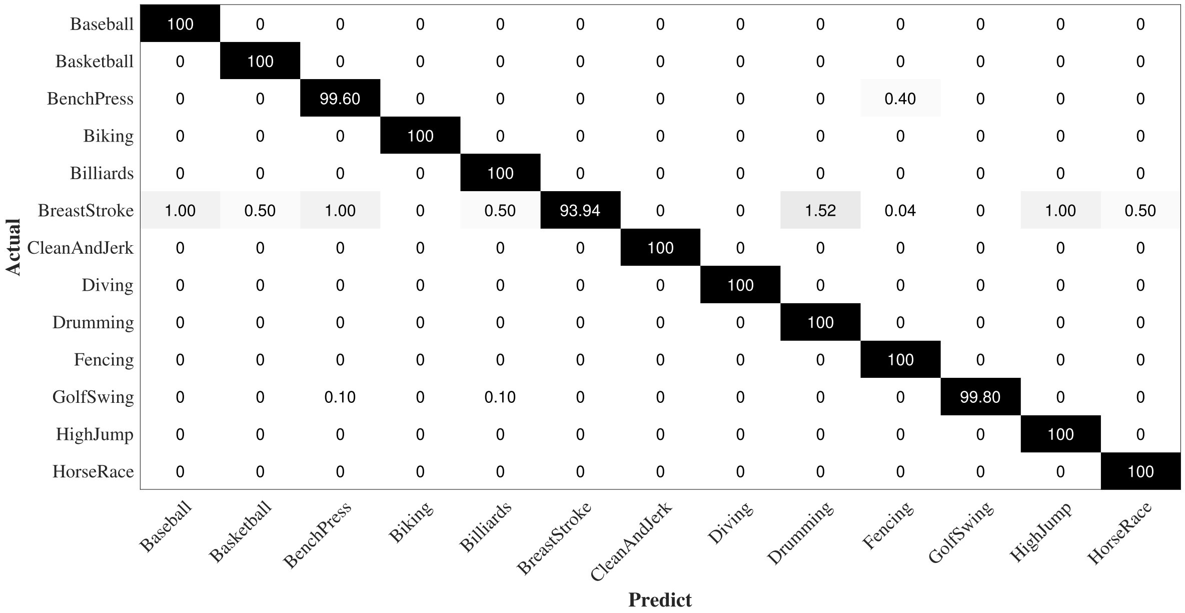
The accuracy percentages for the proposed method are shown in bold font in the table under each dataset. We also show the results with and without the PCA. Overall, the proposed method has resulted in the best performance for all but one datasets outperforming both feature-based and deep learning based methods. Only for the Wizemann dataset, the proposed method is the second best with just 0.19% lower than the best method. Without using the PCA, the KNN has shown better performance compared to that of SVM. Both classifiers have shown improved performance when the PCA is used prior to classification to reduce the dimensionality of the feature space. However, the SVM classifier has benefited the most by using the PCA. FIGURE 12 - FIGURE 17 show the corresponding confusion matrices for the proposed method for the five datasets, respectively.

Though the DHA dataset includes several actions with high similarity, our proposed method discriminates them accurately and outperforms the existing methods to achieve approximately 3% improvement, as can be seen in TABLE 8. The confusion matrix using the KNN classifier in FIGURE 12 shows that 8 out of 23; *i.e.*, 34%, of actions have been fully recognised by proposed modeling method. Similarly, for the KTH dataset, the proposed method shows an improvement of 1.6% compared to the existing methods. For the Weizmann dataset, as shown in FIGURE 14, the proposed method has recognised 50% of actions with 100% accuracy. For the UIUC1 dataset, the proposed method has shown around 0.8% improvement compared to the existing methods. It can be seen in FIGURE 15 that the proposed method recognises the jumping action with 100% accuracy in spite of the similarity between this action and other actions in the dataset. In addition, 79% of action classes have been recognised with more than 99% accuracy. For the UCF Sports dataset, the proposed method has shown improvements around 3.7% compared to the existing feature-based methods, and 1.3% improvement compared to the deep learning based methods. It has also got 77% of action classes having perfect recognition while the rest having accuracy rates higher than 99.6%. For HMDB51 dataset, which is regarded as a complex dataset, our method has outperformed the existing methods, which are mainly deep learning-based, by 16.71%. The confusion matrix in FIGURE 17 shows that 21 out of 51 action classes, *i.e.*, 41%, of classes have achieved 100% accuracy rates using the proposed method. All these datasets contain complex actions with high similarity, yet the proposed method has



	Clapping	Crawling	Jump from Situp	Jumping	Jumping jack	Pushing up	Raise one hand	Running	Sitting to standing	Standing to sitting	Stretching out	Turning	Walking	Waving
Clapping	97.49	0	2.48	0	0	0	0	0	0	0	0	0.03	0	0
Crawling	0	99.71	0	0	0	0.06	0	0	0	0	0	0.06	0	0.17
Jump from Situp	5.47	0.07	93.49	0	0.07	0	0.07	0	0.10	0.26	0.07	0.33	0	0.07
Jumping	0	0	0	100	0	0	0	0	0	0	0	0	0	0
Jumping jack	0	0	0	0	99.90	0	0	0	0	0.06	0	0.02	0	0.02
Pushing up	0	0.12	0	0	0	99.88	0	0	0	0	0	0	0	0
Raise one hand	0	0	0.18	0	0	0	97.56	0	0	0.04	2.21	0	0.01	0
Running	0	0	0	0.21	0.05	0	0	99.59	0	0.05	0	0	0.10	0
Sitting to standing	0	0	0.19	0	0	0	0	0	99.75	0	0	0	0	0.06
Standing to sitting	0.09	0	0	0	0.18	0	0	0	0	99.19	0.09	0.36	0	0.09
Stretching out	0.06	0	0	0	0	0	0.57	0	0	0	99.37	0	0	0
Turning	0	0	0	0	0	0	0	0	0	0	0	100	0	0
Walking	0	0	0	0	0	0	0	0.07	0	0	0	0	99.93	0
Waving	0	0	0	0	0.32	0	0	0	0	0.16	0	0	0	99.52

FIGURE 15. **The confusion matrix of the UIUC1 dataset using the proposed method+KNN (Overall accuracy: 99.15%).**



	Baseball	Basketball	BenchPress	Biking	Billiards	BreastStroke	CleanAndJerk	Diving	Drumming	Fencing	GolfSwing	HighJump	HorseRace
Baseball	100	0	0	0	0	0	0	0	0	0	0	0	0
Basketball	0	100	0	0	0	0	0	0	0	0	0	0	0
BenchPress	0	0	99.60	0	0	0	0	0	0	0.40	0	0	0
Biking	0	0	0	100	0	0	0	0	0	0	0	0	0
Billiards	0	0	0	0	100	0	0	0	0	0	0	0	0
BreastStroke	1.00	0.50	1.00	0	0.50	93.94	0	0	1.52	0.04	0	1.00	0.50
CleanAndJerk	0	0	0	0	0	0	100	0	0	0	0	0	0
Diving	0	0	0	0	0	0	0	100	0	0	0	0	0
Drumming	0	0	0	0	0	0	0	0	100	0	0	0	0
Fencing	0	0	0	0	0	0	0	0	0	100	0	0	0
GolfSwing	0	0	0.10	0	0.10	0	0	0	0	0	99.80	0	0
HighJump	0	0	0	0	0	0	0	0	0	0	0	100	0
HorseRace	0	0	0	0	0	0	0	0	0	0	0	0	100

FIGURE 16. **The confusion matrix of the UCF Sports dataset using the proposed method+KNN (Overall accuracy: 99.71%).**

resulted in excellent recognition rates. The accurate discrimination between the actions in all these dataset proves the superiority of the proposed approach of exploiting the temporal visual salience modeling for visual anonymization followed by learning *HOG-S* features.

Finally, we revisit the utility of the visually anonymized sequence as perceived by the participants in terms of the action recognition rate vs. the level of visual anonymity (measured by MAS) shown in FIGURE 11. Here we evaluate the utility of the visually anonymized streams in terms of machine-based HAR as shown in TABLE 9. The utility in terms of machine-based HAR with respect to the visual anonymizing methods is summarized in FIGURE 18. It is evident that the proposed anonymization method combined with the proposed *HOG-S* based HAR provides the best accuracy rate for HAR as well as the highest MAS resulting in the best joint anonymizing and HAR methodology. It can be also highlighted that the machine-based utility is much higher than the human-perceived utility for the proposed anonymity silhouettes. This confirms the efficiency

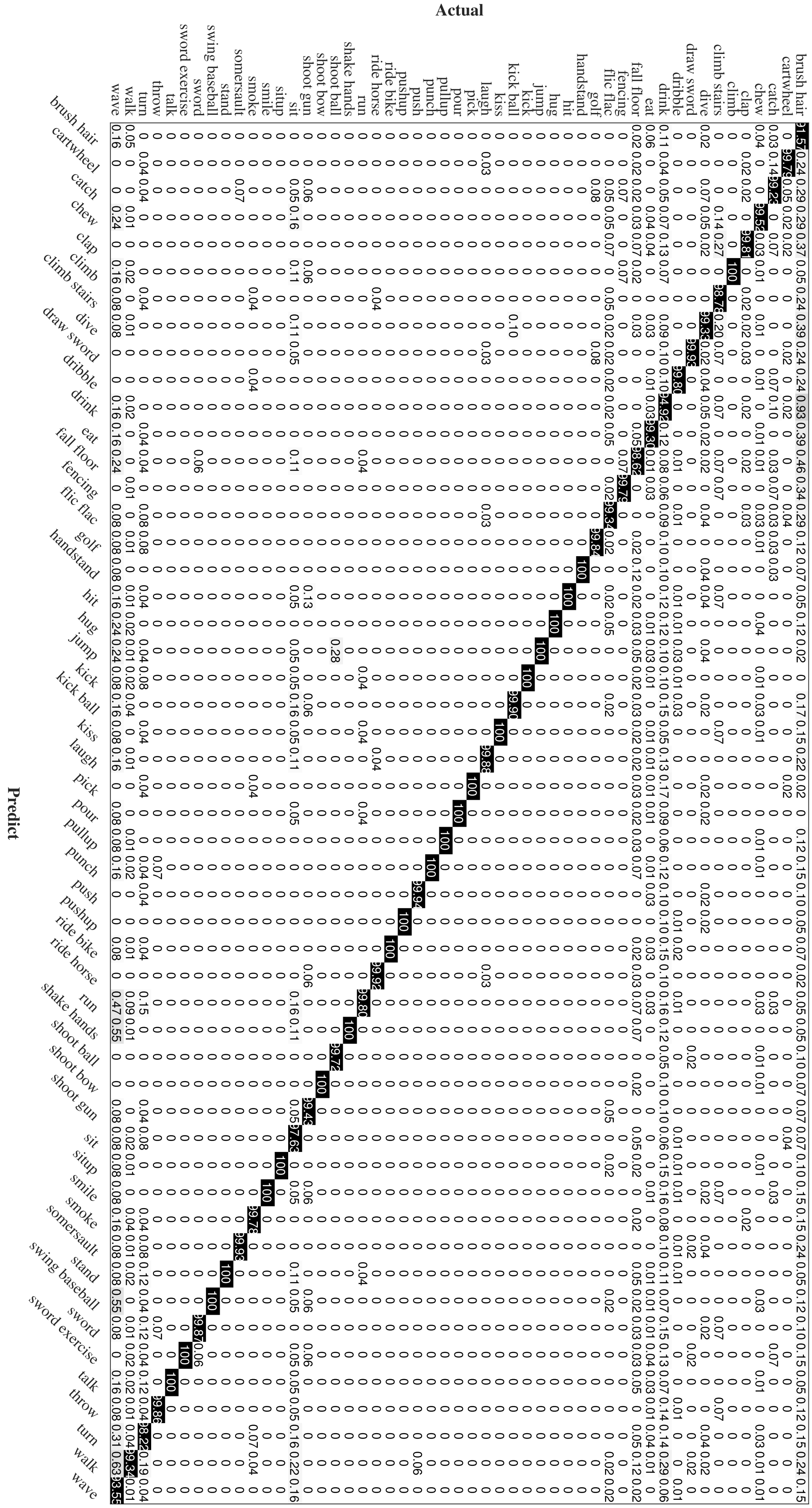


FIGURE 17. The confusion matrix of the HMDB51 dataset using the proposed method + KNN (Overall accuracy: 99.03%).

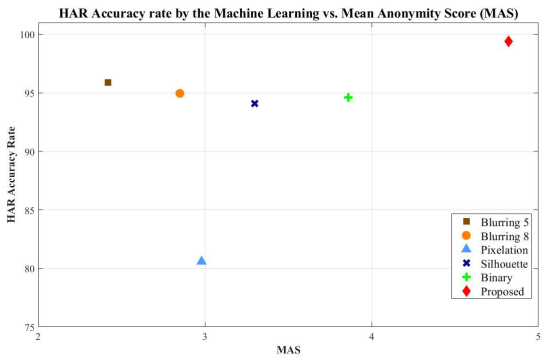


FIGURE 18. HAR accuracy rate (the machine utility of the visually anonymized sequence) vs. the Mean Anonymity Score (MAS) for the existing visual anonymization methods and the proposed method.

of modeling temporal visual saliency to obtain a saliency driven silhouette for anonymization and the ability of the proposed *HOG-S* features to learn the important features in such anonymity maps.

V. CONCLUSIONS

In this paper, we have have presented a methodology for visually anonymizing video clips by modeling the temporal visual salience while retaining the computer-based utility of human action recognition. The novel temporal salience model proposed in this paper encapsulates the intensity of the motion dynamics of the action into the anoymization maps. This is followed by extracting the newly proposed *HOG-S* features for human action recognition in the visually anonymized domain. The proposed visually anonymization method has achieved the highest MAS compared to the existing methods for visually anonymizing. The human observer surveys conducted have confirmed that none of the six appearance attributes were recognizable for all sequences tested for KTH and UIUC1 datasets anonymized using the proposed method. Similarly, for DHA and Weizmann datasets, around 97% and 89% sequences were not able to recognize any of the attributes. The proposed method's high MAS has been justified by these results. In terms of the utility of the anonymized clips, our proposed anonymization method coupled with the proposed *HOG-S* feature learning approach has achieved the best machine perceived human action recognition accuracy rates, compared to those of existing anonymizing methods. The proposed HAR method has also exceeded the performance of human-perceived action recognition from videos anonymized using the proposed temporal salience-based anonymization method. Overall, when considered the proposed work as a holistic human action recognition method, *i.e.*, the temporal salience modeling followed by the *HOG-S* feature extraction, it has resulted in the best human action recognition accuracy rates for datasets DHA, KTH, UIUC1, UCF Sports and HMDB51 with improvements of 3%, 1.6%, 0.8%, 1.3% and 16.7%, respectively outperforming both feature-based and deep learning based existing approaches. It also has shown the second best accuracy rate for the Weizmann dataset, with just 0.19% less than the best method. This superior performance is the result of the way in which the actions are modelled using the proposed temporal salience modeling leading to generation of the silhouette that captures the dynamics of the motion present in the action at a specific time. This work provides a very useful tool for human action recognition in vision-based assisted living.

ACKNOWLEDGMENT

Salah Al-Obaidi and Hiba Al-Khafaji would like to thank the University of Babylon and the Ministry of Higher Education and Scientific Research (MOHESR) in Iraq for their Ph.D. studentships.

REFERENCES

[1] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Activity recognition using a combination of category components and local models for video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1128–1139, Aug. 2008.

[2] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan, and A. A. Abbasi, "Human action recognition using fusion of multiview and deep features: An application to video surveillance," *Multimedia Tools Appl.*, pp. 1–27, Mar. 2020.

[3] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2702–2706.

[4] S. Tanberk, Z. H. Kilimci, D. B. Tükel, M. Uysal, and S. Akyokus, "A hybrid deep model using deep learning and dense optical flow approaches for human activity recognition," *IEEE Access*, vol. 8, pp. 19799–19809, 2020.

[5] E. A. Mosabbeb, K. Raahemifar, and M. Fathy, "Multi-view human activity recognition in distributed camera sensor networks," *Sensors*, vol. 13, no. 7, pp. 8750–8770, Jul. 2013.

[6] F. Cardinaux, D. Bhowmik, C. Abhayaratne, and M. S. Hawley, "Video based technology for ambient assisted living: A review of the literature," *J. Ambient Intell. Smart Environ.*, vol. 3, no. 3, pp. 253–269, 2011.

[7] S. Pal and C. Abhayaratne, "Video-based activity level recognition for assisted living using motion features," in *Proc. 9th Int. Conf. Distrib. Smart Camera (ICDSC)*, 2015, pp. 62–67.

[8] I. Fatima, M. Fahim, Y.-K. Lee, and S. Lee, "A unified framework for activity recognition-based behavior analysis and action prediction in smart homes," *Sensors*, vol. 13, no. 2, pp. 2682–2699, Feb. 2013.

[9] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, T. Dhaene, and W. De Neve, "Indoor human activity recognition using high-dimensional sensors and deep neural networks," *Neural Comput. Appl.*, vol. 32, pp. 12295–12309, Aug. 2019.

[10] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 81–94, Mar. 2016.

[11] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1325–1334.

[12] I. Jegham, A. Ben Khalifa, I. Alouani, and M. A. Mahjoub, "Vision-based human action recognition: An overview and real world challenges," *Forensic Sci. Int., Digit. Invest.*, vol. 32, Mar. 2020, Art. no. 200901.

[13] R. Li, B. Lu, and K. D. McDonald-Maier, "Cognitive assisted living ambient system: A survey," *Digit. Commun. Netw.*, vol. 1, no. 4, pp. 229–252, Nov. 2015.

[14] T.-H. Tsai and K.-L. Zhang, "Implementation of intelligent home appliances based on IoT," in *Proc. IEEE Asia Pacific Conf. Circuits Syst. (APCCAS)*, Jeju, South Korea, Oct. 2016, pp. 146–148.

[15] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb. 2019.

[16] H. Habibzadeh, K. Dinesh, O. Rajabi Shishvan, A. Boggio-Dandry, G. Sharma, and T. Soyata, "A survey of healthcare Internet of Things (HIoT): A clinical perspective," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 53–71, Jan. 2020.

[17] S. Al-Obaidi and C. Abhayaratne, "Privacy protected recognition of activities of daily living in video," in *Proc. 3rd IET Int. Conf. Technol. Act. Assist. Living (TechAAL)*, 2019, pp. 1–6.

[18] B. Myagmar, J. Li, and S. Kimura, "Heterogeneous daily living activity learning through domain invariant feature subspace," *IEEE Trans. Big Data*, early access, Mar. 2, 2020.

[19] J. R. Padilla-López, A. A. Chaaraoui, F. Gu, and F. Flórez-Revuelta, "Visual privacy by context: Proposal and evaluation of a level-based visualisation scheme," *Sensors*, vol. 15, no. 6, pp. 12959–12982, 2015.

[20] D. J. Butler, J. Huang, F. Roesner, and M. Cakmak, "The privacy-utility tradeoff for remotely teleoperated robots," in *Proc. 10th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Portland, OR, USA, Mar. 2015, pp. 27–34.

[21] B.-J. Han, H. Jeong, and Y.-J. Won, "The privacy protection framework for biometric information in network based CCTV environment," in *Proc. IEEE Conf. Open Syst.*, Sep. 2011, pp. 86–90.

[22] A. Edgcomb and F. Vahid, "Privacy perception and fall detection accuracy for in-home video assistive monitoring with privacy enhancements," *ACM SIGHIT Rec.*, vol. 2, no. 2, pp. 6–15, Sep. 2012.

[23] E. T. Hassan, R. Hasan, P. Shaffer, D. Crandall, and A. Kapadia, "Cartooning for enhanced privacy in lifelogging and streaming videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1333–1342.

[24] M. Saini, P. K. Atrey, S. Mehrotra, and M. Kankanhalli, "Adaptive transformation for robust privacy protection in video surveillance," *Adv. Multimedia*, vol. 2012, pp. 1–14, May 2012.

[25] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3910–3921, Sep. 2014.

[26] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[27] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 681–688.

[28] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2552–2564, Aug. 2015.

[29] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, vol. 6, no. 7, p. 9.

[30] M. Oakes, D. Bhowmik, and C. Abhayaratne, "Global motion compensated visual attention-based video watermarking," *J. Electron. Imag.*, vol. 25, no. 6, 2016, Art. no. 061624.

[31] C. Posch, R. Benosman, and R. Etienne-Cummings, "Giving machines humanlike eyes," *IEEE Spectr.*, vol. 52, no. 12, pp. 44–49, Dec. 2015.

[32] S. Al-Obaidi and C. Abhayaratne, "Temporal salience based human action recognition," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2017–2021.

[33] J. Dai, J. Wu, B. Saghafi, J. Konrad, and P. Ishwar, "Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 68–76.

[34] M. Eldib, F. Deboeverie, W. Philips, and H. Aghajan, "Behavior analysis for elderly care using a network of low-resolution visual sensors," *J. Electron. Imag.*, vol. 25, no. 4, 2016, Art. no. 041003.

[35] J. Chen, "Privacy-preserving smart-room visual analytics," Ph.D. dissertation, Boston Univ., Boston, MA, USA, 2019.

[36] J. Zhao, N. Frumkin, J. Konrad, and P. Ishwar, "Privacy-preserving indoor localization via active scene illumination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1580–1589.

[37] J. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe, "Video classification with densely extracted HOG/HOF/MBH features: An evaluation of the accuracy/computational efficiency trade-off," *Int. J. Multimedia Inf. Retr.*, vol. 4, no. 1, pp. 33–44, Mar. 2015.

[38] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *J. Healthcare Eng.*, vol. 2017, pp. 1–31, Jul. 2017.

[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[40] F. Liu, X. Xu, S. Qiu, C. Qing, and D. Tao, "Simple to complex transfer learning for action recognition," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 949–960, Feb. 2016.

[41] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description," *IET Comput. Vis.*, vol. 10, no. 7, pp. 758–767, Oct. 2016.

[42] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–275.

[43] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 635–648.

[44] F. Angelini, Z. Fu, S. A. Velastin, J. A. Chambers, and S. M. Naqvi, "3D-hog embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4219–4223.

[45] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.

[46] F. W. M. Stentiford, "Visual attention: Low-level and high-level viewpoints," *Proc. SPIE*, vol. 8436, Apr. 2012, Art. no. 84360L.

[47] W. Li, J. Qiu, and X. Li, "Visual saliency detection based on gradient contrast and color complexity," in *Proc. 7th Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2015, p. 42.

[48] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, B. Sierra, I. Rodriguez, and E. Jauregi, "Video activity recognition: State-of-the-art," *Sensors*, vol. 19, no. 14, p. 3160, Jul. 2019.

[49] R. D. Singh and N. Aggarwal, "Novel research in the field of shot boundary detection—A survey," in *Advances in Intelligent Informatics* (Advances in Intelligent Systems and Computing), vol. 320, E. S. El-Alfy, S. Thampi, H. Takagi, S. Piramuthu, and T. Hanne, Eds. Cham, Switzerland: Springer, 2015.

[50] S. Gupta, S. Dhiman, A. Makkar, D. Khanna, and A. Arora, "A review of video shot boundary detection (SBD) technique," *J. Image Process. Pattern Recognit. Prog.*, vol. 5, no. 2, pp. 59–70, 2018.

[51] P. Browne, A. F. Smeaton, N. Murphy, N. O'Connor, S. Marlow, and C. Berrut, "Evaluating and combining digital video shot boundary detection algorithms," in *Proc. 4th Irish Mach. Vis. Inf. Process. Conf.*, 2000, pp. 1–8.

[52] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[53] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, Aug. 2004, pp. 32–36.

[54] Y.-C. Lin, M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen, "Human action recognition and retrieval using sole depth information," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, Nara, Japan, 2012, pp. 1053–1056.

[55] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 548–561.

[56] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Advances in Computer Vision and Pattern Recognition*. Cham, Switzerland: Springer, 2014, pp. 181–208.

[57] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[58] Z. Gao, H. Zhang, G. P. Xu, and Y. B. Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition," *Neurocomputing*, vol. 151, pp. 554–564, Mar. 2015.

[59] H. Liu, Q. He, and M. Liu, "Human action recognition using adaptive hierarchical depth motion maps and Gabor filter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1432–1436.

[60] M. Liu, H. Liu, and C. Chen, "3D action recognition using multiscale energy-based global ternary image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1824–1838, Aug. 2018.

[61] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, and L. Shao, "Action recognition using 3D histograms of texture and a multi-class boosting classifier," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4648–4660, Oct. 2017.

[62] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1860–1870, Dec. 2013.

[63] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1510–1520, Jul. 2017.

[64] C.-Y. Zhang, Y.-Y. Xiao, J.-C. Lin, C. L. P. Chen, W. Liu, and Y.-H. Tong, "3-D deconvolutional networks for the unsupervised representation learning of human motions," *IEEE Trans. Cybern.*, early access, Mar. 9, 2020.

[65] G. K. Yadav, P. Shukla, and A. Sethfi, "Action recognition using interest points capturing differential motion information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1881–1885.

[66] D. Avola, M. Cascio, L. Cinque, G. L. Foresti, C. Massaroni, and E. Rodola, "2-D skeleton-based action recognition via two-branch stacked LSTM-RNNs," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2481–2496, Oct. 2020.

[67] F. Osayamwen and J.-R. Tapamo, "Deep learning class discrimination based on prior probability for human activity recognition," *IEEE Access*, vol. 7, pp. 14747–14756, 2019.

[68] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 236–243, Feb. 2013.

[69] S. Zeng, G. Lu, and P. Yan, "Enhancing human action recognition via structural average curves analysis," *Signal, Image Video Process.*, vol. 12, pp. 1551–1558, May 2018.

[70] K. Xu, X. Jiang, and T. Sun, "Two-stream dictionary learning architecture for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 567–576, Mar. 2017.

[71] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Lifelong learning of human actions with deep neural network self-organization," *Neural Netw.*, vol. 96, pp. 137–149, Dec. 2017.

[72] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[73] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Robust relative attributes for human action recognition," *Pattern Anal. Appl.*, vol. 18, no. 1, pp. 157–171, Feb. 2015.

[74] Y. Shan, Z. Zhang, P. Yang, and K. Huang, "Adaptive slice representation for human action classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 10, pp. 1624–1636, Oct. 2015.

[75] T. Wang, Y. Chen, M. Zhang, J. Chen, and H. Snoussi, "Internal transfer learning for improving performance in human action recognition for small datasets," *IEEE Access*, vol. 5, pp. 17627–17633, 2017.

[76] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, "DeepProposals: Hunting objects and actions by cascading deep convolutional layers," *Int. J. Comput. Vis.*, vol. 124, no. 2, pp. 115–131, Sep. 2017.

[77] M. H. Siddiqi, M. Alruwaili, A. Ali, S. Alanazi, and F. Zeshan, "Human activity recognition using Gaussian mixture hidden conditional random fields," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–14, Aug. 2019.

[78] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream CNN: Learning representations based on human-related regions for action recognition," *Pattern Recognit.*, vol. 79, pp. 32–43, Jul. 2018.

[79] C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based LSTM networks," *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105820.

[80] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 971–980.

[81] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib, "TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition," *Signal Process., Image Commun.*, vol. 71, pp. 76–87, Feb. 2019.

[82] X. Song, C. Lan, W. Zeng, J. Xing, X. Sun, and J. Yang, "Temporal–spatial mapping for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 748–759, Mar. 2020.

[83] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[84] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose MoTion representation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7024–7033.

[85] Z. Zheng, G. An, D. Wu, and Q. Ruan, "Global and local knowledge-aware attention network for action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 30, 2020.

[86] L. Wang, P. Koniusz, and D. Huynh, "Hallucinating IDT descriptors and I3D optical flow features for action recognition with CNNs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8698–8708.

SALAH AL-OBAIDI received the B.E. and M.Sc. degrees in computer science from the University of Babylon, Iraq, in 2000 and 2004, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering, The University of Sheffield, U.K. Prior to that, he was a Lecturer with the Department of Computer Science, University of Babylon. His research interests include image and video processing, signal processing, video saliency, surveillance, human action recognition, assisted living, and computer vision. He was a recipient of a Ph.D. Scholarship from the University of Babylon and the Ministry of Higher Education and Scientific Research (MOHESR) in Iraq.

HIBA AL-KHAFAJI received the B.E. and M.Sc. degrees in computer science from the College of Science, University of Babylon, Iraq, in 1996 and 2004, respectively. She is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering, The University of Sheffield, U.K. Her research interests include signal and image processing, security, and computer vision. She was a recipient of a Ph.D. Scholarship from the University of Babylon and the Ministry of Higher Education and Scientific Research (MOHESR) in Iraq.

CHARITH ABHAYARATNE (Member, IEEE) received the B.E. degree in electrical and electronic engineering from The University of Adelaide, Australia, in 1998, and the Ph.D. degree in electronic and electrical engineering from the University of Bath, U.K., in 2002. He is currently a Lecturer with the Department of Electronic and Electrical Engineering, The University of Sheffield, U.K. He has published more than 80 peer-reviewed articles in leading journals, conferences, and book editions. His research interests include multidimensional signal processing, image and video compression, visual content understanding, multimedia content security, and forensics. He was a recipient of the European Research Consortium for Informatics and Mathematics (ERCIM) Postdoctoral Fellowship from 2002 to 2004 to carry out research at the Center of Mathematics and Computer Science (CWI), The Netherlands, and the National Research Institute for Computer Science and Control (INRIA), France. He also serves as an Associate Editor for IEEE Transactions on Image Processing, IEEE Access, and *Journal of Information Security and Applications (JISA)* (Elsevier).