Proceedings Paper:

Al-Obaidi, S. and Abhayaratne, C. orcid.org/0000-0002-2799-7395 (2019) Privacy protected recognition of activities of daily living in video. In: 3rd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2019). 3rd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2019), 25 Mar 2019, London, UK. Institution of Engineering and Technology . ISBN 9781839530883

https://doi.org/10.1049/cp.2019.0101

Privacy Protected Recognition of Activities of Daily Living in Video

Salah Al-Obaidi and Charith Abhayaratne

Department of Electronic and Electrical Engineering, The University of Sheffield
Sheffield, S1 3JD, United Kingdom
Email: {s.alobaidi, c.abhayaratne}@sheffield.ac.uk

Keywords: Privacy Protection, Video Anonymisation, Temporal Saliency, Activities of Daily Living Recognition.

Abstract

This paper proposes a new method to protect the privacy while retaining the ability to accurately recognise the activities of daily living for video-based monitoring in ambient assisted living applications. The proposed method obfuscates the human appearance by modelling the temporal saliency in the monitoring video sequences. It mimics the functionality of neuromorphic cameras and explores the temporal saliency to generate a mask to anonymise the human appearance. Since the anonymising masks encapsulate the temporal saliency with respect to motion in the sequence, they provide a good basis for further utilisation in activity recognition, which is achieved by representing the HOG features on privacy masks. The proposed method has resulted in excellent anonymising performances compared using the cross correlation measures. In terms of activity recognition, the proposed method has resulted in 5.6% and 5.4% improvements of accuracies over other anonymisation methods for Weizmann and DHA datasets, respectively.

1 Introduction

Video and computer vision have become important tools for monitoring activities of daily living and person wellbeing in ambient assisted living (AAL) [1–4]. Although; these systems and tools have performed well in monitoring, decreasing the concerns on privacy of the elderly users remains a highly important requirement [1, 5]. There have been a few solutions to address the privacy concerns, which can be mainly categorised into two themes: low resolution sensors [6–8] and image processing based [9–11]. Using low resolution sensors adopt a network of extremely low resolution cameras [6, 7] or low resolution colour sensors [8] to capture low-resolution visual images, which have been successfully exploited in the applications of activity recognition [6], behaviour understanding [7] and object localisation [8]. However, these sensors are more sensitive to the changes in the light conditions [7, 8] resulting in less accuracy in activity recognition. The second category of solutions is to adopt the image processing techniques, such as, blocking [9], cartooning [10], Gaussian blurring [11], pixelation [12] and masking with silhouettes, to obfuscate the sensitive information. However, these methods require consideration of the trade-off between the privacy protection and utility of the anonymised sequences for monitoring tasks [13]. Naturally, a high level of privacy protection leads to low level of utility and vice versa. This is one of the major challenges associated with using video camera sensors in the field of AAL.

To address this, in this paper, we propose a new methodology for anonymising video, which can anonymise the persons and content in the video while retaining high utility of anonymised video for activity recognition. The proposed anonymising method captures the motion activity in the scene to replace the corresponding spatial information. This mimics the functionality of emerging neuromorphic (event-based) cameras [14] to capture events by modelling the temporal saliency. The anonymised video maps derived from the temporal saliency modelling are further analysed by extracting histograms of oriented gradients (HOG) features for activity recognition tasks. The proposed method provides useful anonymous information which can be further explored in activities of daily living monitoring applications, such as, activity recognition and activity level recognition, efficiently without doing any further processing, like motion estimation. The main contributions of this work include:

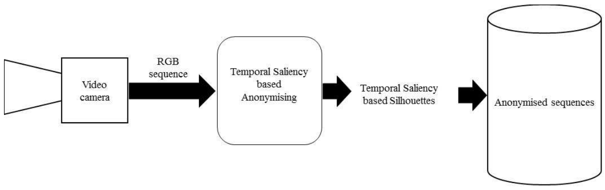
1. A new methodology to achieve a high level of privacy protection using the proposed temporal saliency estimation model; and

2. A new methodology for highly accurate activity recognition using such anonymised video sequences.

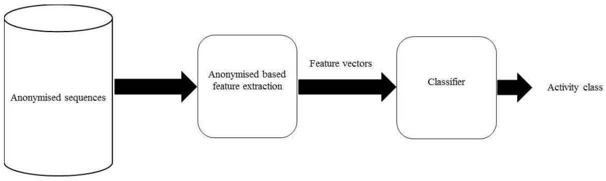
The rest of the paper is organised as follows: Section 2 presents the proposed method for anonymising video sequences followed by activity recognition. The performance evaluation of the proposed system for anonymising efficiency, as well as, activity recognition efficiency is presented in Section 3 followed by concluding remarks in Section 4.

2 The proposed method

This section presents the proposed method to protect the privacy and explore the obtained obfuscated information in activity recognition. Figure 1 depicts the proposed approach to preserve the privacy and recognise the protected information.

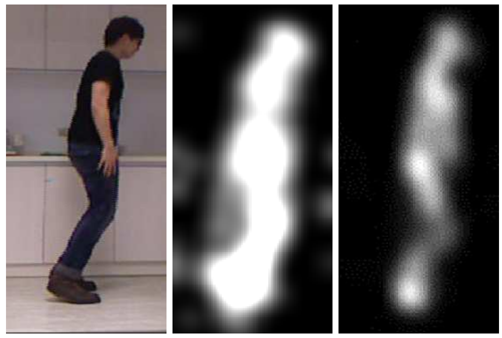


(a) Temporal saliency based anonymising



(b) Utility of the anonymised information

Figure 1. Proposed privacy protection and intelligibility model



(a) (b) (c)

Figure 2. Anonymising modelling of frame #20 from the jumping sequence of the participant #1: (a) original frame, (b) single threshold based anonymising and (c) weighted threshold based anonymising.

2.1 Anonymising mask generation using temporal saliency modelling

The proposed anonymity method extracts the anonymising mask of the human to protect the privacy by modelling the temporal saliency in the video sequence. This method depends on distributing the temporal saliency magnitudes based on the most dynamic parts to anonymise the privacy, which is crucially attributed in activity signature and then activity representation. This enhances the ability of utilising this anonymised information without using extra information or additional algorithms for extracting such information.

The anonymity mask generation consists of the following steps. Let s be a video sequence with F frames. First, for each two successive frames, f_t and f_{t-1}, where $t \in [1, \cdots, F]$, the frame difference D_t is computed to obtain the temporal changes in each pixel intensity in the current frame with respect to the previous frame. It is important to measure the magnitude of the movement and to differentiate the changes that come from the light conditions and global motion problems. These pixel differences are compared with a user-defined threshold, τ, to keep the magnitude of pixels' movement and eliminate unnecessary changes as follows:

$$\tilde{D}_t(x,y) = \begin{cases} D_t(x,y) & \text{if } D_t(x,y) \geq \tau \\ 0 & Otherwise \end{cases}, \tag{1}$$

where $D_t(x,y)$ and $\tilde{D}_t(x,y)$ are the difference maps at location (x,y) before and after comparing with a threshold, respectively. However, it is difficult to determine the perfect value of τ. Therefore, we further vary τ using $\mathcal{N}$ user-defined thresholds to make this filtration more robust and comprehensive.

Second, the difference map based on the user-defined threshold τ_n, e.g. $\tilde{D}_t^{\tau_n}$, where $n \in \{1, 2, \cdots, \mathcal{N}\}$, is then analysed using an overlapped block-based two dimensions fast Fourier transform (2DFFT) based entropy in order to scoring these differences and distribute the scores differently based on the magnitude of the pixel's change. Due to $\tilde{D}_t^{\tau_n}$ has magnitudes that are distributed over the foreground region, a $N \times N$ entropy based operator is adopted to score these magnitudes to model the temporal salience. Accordingly, $\tilde{D}_t^{\tau_n}$ is partitioned into overlapped $N \times N$ blocks and then 2DFFT is applied to each block, separately, to analyse the frequencies of differences in each block. To make sure all the pixels in $\tilde{D}_t^{\tau_n}$ will be processed, we pad zero values around the original borders. This will assist in highlighting the salience regions compared to the background.

Third, the Power Spectral Density (PSD), $P_{bi}^t(u,v) = \frac{1}{N^2}\mathcal{A}_{b_i}^t(u,v)^2$, where $\mathcal{A}$ is the matrix of magnitudes of frequencies in block b_i, where $i = 1, 2, \cdots, B$ and B is the number of blocks. The obtained densities in P_{bi}^t are then normalised to suppress the high variation in b_i and create a homogeneous distribution in this block. In the fourth step, these probability densities are subjected to scoring approach using the local Shannon entropy formula to obtain the temporal salience magnitude in each pixel location. This entropy assigns a score for the pixel that has been located in the centre of b_i. Then, the weighted entropy, $\hat{\mathcal{E}}$, at location (x,y) of all entropies over $\mathcal{N}$ thresholds is computed as

$$\hat{\mathcal{E}}(x,y) = \frac{\sum_{n=1}^{\mathcal{N}} \tau_n \, \mathcal{E}^{\tau_n}(x,y)}{\sum_{n=1}^{\mathcal{N}} \tau_n}, \tag{2}$$

where $\hat{\mathcal{E}}(x,y)$ and $\mathcal{E}^{\tau_n}(x,y)$ are the weighted entropy and the entropy value based on threshold τ_n at location (x,y), respectively. This leads to highlight the silhouette of the foreground object by distributing the saliency magnitudes exploiting the relation between the action and the human body parts. Furthermore, this approach is considered a crucial factor in the distinction between actions since the magnitudes of the entropies will be distributed across the silhouette based on the action class. This entropies map is then normalised to the [0 255] values and smoothed by applying a 2-D Gaussian kernel with $\sigma = 6$ in order to fill the small holes if found and obtain the final temporal saliency map, S_T, leading to the anonymising mask.

Figure 2 shows the result of applying the above procedure on a sample frame from the jumping sequence of DHA dataset. As we can see, our anonymising modelling approach leads to further highlighting the most dynamic parts of the body, Figure 2 (c). These moving parts are represented by increasing the highlighting using the proposed approach of anonymity, which reflects the optimal performance of the proposed model of privacy preservation. If we depend only on a single threshold formula, Figure 2 (b), it is difficult to represent the silhouette of the activity perfectly since the human body parts seem to have the same magnitude. Instead, with the local entropy based obfuscating approach, the variation in movements will be fairly modelled. This method protects the privacy as well as maintaining the most useful information about the action/activity.

The example in Figure 2 shows that the proposed method can be used for two useful purposes: protecting the privacy efficiently by obfuscating the most significant information in the scene and eliminating unnecessarily redundant information. This makes the video-based sensor less intrusive and more acceptable in the real world. Furthermore, modelling the anonymity in the form of saliency maps provides useful information to be explored by the descriptor to extract powerful discriminating features and achieve efficiency in the activity recognition application.

2.2 Analysing the anonymised silhouette for activity recognition

The above privacy anonymising approach is beneficial to design a robust discriminating descriptor. The proposed descriptor focuses on analysing the useful information that has been modelled in the anonymised silhouette. The proposed descriptor relies on exploring the HOG features of anonymised video maps (HOG-A). The description is obtained by calculating the horizontal and vertical gradients of the silhouette region R, where $R \subset S_T$, in each video frame. Then, the magnitude, G_S, and orientation, θ_S, of the gradients are computed as follows:

$$G_S = \sqrt{d_x^2(R) + d_y^2(R)}, \tag{3}$$

$$\theta_S = arctan\left(\frac{d_y(R)}{d_x(R)}\right), \tag{4}$$

where $d_x(R)$ and $d_y(R)$ are the horizontal and vertical gradients of R, respectively. The gradient orientations are then quantized into 9-bins orientation histogram. Let G be a $K \times L$ matrix containing the gradient response. Thus, there are $B_K \times B_L$ blocks from which HOG-A features are extracted. Each block in turn is subdivided into P patches and each patch gives us 9-bins histogram. Thus, when we concatenate them all into one response vector, $\mathbf{v}$, we obtain a $9 \times P \times B_K \times B_L$ dimensions vector. Then, $\mathbf{v}$ is normalised in order to make the description more light-invariant.

However, the discriminating power of the descriptor can not be perfect in accordance to the variation inside the action and similarities among the actions. To address this, we improve the HOG-A descriptor by doing a time down-sampling formula in order to find two groups of vectors in each time: even and odd time-based vectors. Thus, the set of normalised vectors are down-sampled by 2 with respect to time, e.g., $T\downarrow 2$, where T is the length of the sequence. The final feature vector at time instant t, $\tilde{\mathbf{v}}_t$, is computed as

$$\tilde{\mathbf{v}}_t = \left| \sum_{k=0}^{t-1} \hat{\mathbf{v}}_{t-2k} - \sum_{k=0}^{t-1} \hat{\mathbf{v}}_{t-(2k+1)} \right|, \tag{5}$$

where $\hat{\mathbf{v}}$ is the normalised feature vector and $t - k > 0$.

The purpose from Equation (5) is to increase the dissimilarity and the similarity between the activities and inside the activity, respectively, in order to improve the discriminating power of HOG-A descriptor.

3 Performance Evaluation

3.1 Datasets

Depth-included Human Action (DHA) video dataset [15] contains 23 action categories performed by 21 participants (12 males and 9 females). It is recorded using a static Kinect camera in three different scenes with 480×640 resolution. The RGB version of videos is used in the experiments. Weizmann dataset [16] contains 93 low-resolution (144×180, 50 fps) video sequences showing 10 actions achieved by 9 actors.

3.2 Evaluation of anonymising

Several approaches to anonymise the privacy using filtering algorithms are evaluated on our datasets. These methods are Gaussian blurring [17], silhouette, pixelation [18], and binary silhouette. For the pixelation method, we used two different resolutions based on the size of the frame; e.g. 20×16 and 12×12 for DHA and Weizmann , respectively. Mainly, all the filtering algorithms and the proposed anonymity method follow the same scenario to process the video sequences to anonymise the privacy; i.e. 1) the method, 2) HOG extraction and 3) KNN classifier. The most important issue that has been considered to evaluate the efficiency of the algorithms is the trade-off between the privacy level and utility. We adopted an objective method, the magnitude of mean cross correlation, to assess the robustness of the proposed privacy protection approach and the current filtering algorithms w.r.t. the unmodified video frames.

Table 1 depicts the calculated magnitudes of mean cross correlation of the four filtering methods and the proposed method over all the sequences in each dataset. As can be clearly seen, the proposed method outperforms all filtering methods. The proposed method with the binary silhouette got the lowest similarity with w.r.t. the original video frames. This means that our method achieves a high level of privacy preservation which was expected due to its strong anonymised activity abstraction and modelling.

This means that the temporal saliency maps are considered as masks which makes the proposed anonymising method achieves a high level of privacy protection such the binary mask. However, the difference between the binary mask and the saliency map is that the saliency map distributed the magnitudes in the silhouette region based on the dynamic changes acquired at each location. This can be shown in Figure 3;

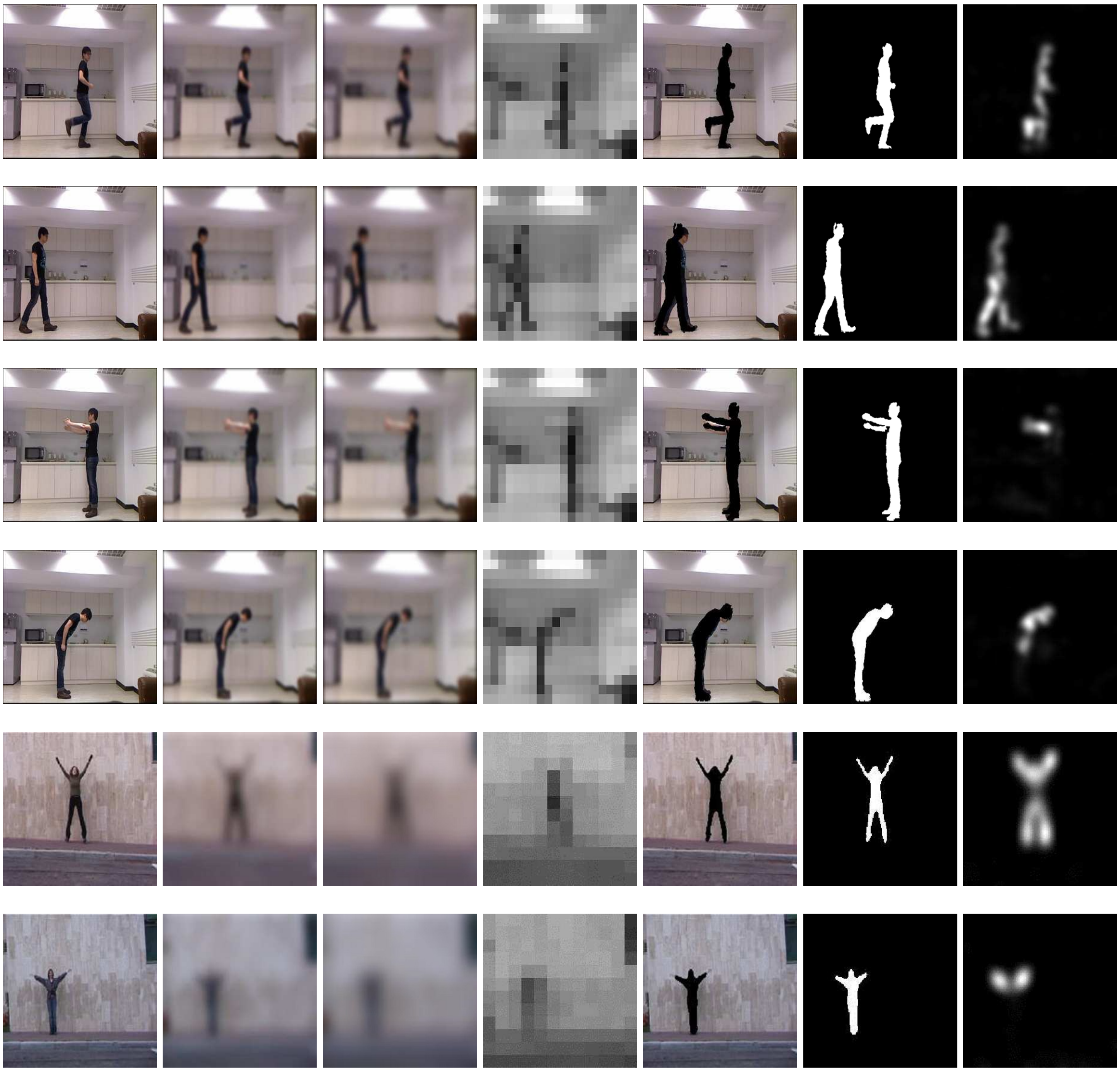


Figure 3. Comparison of globally Gaussian blurred, pixelated, silhouette, binary silhouette and the proposed anonymising method for different actions from both datasets. From top to bottom: run, walk, side clap, bend, jack and two hand waving. From left to right: *column-1* original frame, *column-2* Gaussian blurring with $\sigma = 5$, *column-3* Gaussian blurring with $\sigma = 8$, *column-4* pixelation, *column-5* silhouette, *column-6* binary silhouette and *column-7* proposed anonymising method, respectively.

the columns 6 and 7, respectively. The main difference between these two columns is that the silhouette region in the binary mask contains the same magnitude in the overall locations while these locations highlighted differently in our proposed anonymising method. These variations will be reflected on the human action recognition stage.

Furthermore, we can see that pixelation filter achieved privacy protection level better than Gaussian blurring. On the other hand, this can limit the utility of the obfuscated information compared to blurring because the pixelation leads to distort the useful visual information in order to address the problem of privacy preservation. This reduces the possibility of using this information, e.g. low level utility. Columns 2, 3 and 4 illustrate clearly this trade off.

3.3 Evaluation of activity recognition using the silhouette

Several experiments have been done to evaluate the proposed approach of vision-based privacy preservation. The results of converting the RGB frames into obfuscated maps are evaluated in terms of human action recognition task. We explained previously that the proposed method preserves the privacy and explores the obtained saliency maps to recognise the actions. This means that the proposed method exploits the same approach to achieve two objectives instead of using multiple algorithms to achieve each one. Table 2 explains a comparison between the proposed method and the filtering techniques used to preserve the privacy using the DHA and Weizmann datasets.

We can see that the proposed method achieves the higher level of utility than blurring and pixelation methods. This is due to the efficiency of the proposed method to anonymise the

Method	DHA	Weizmann
Blurring $\sigma = 8$	0.90	0.66
Blurring $\sigma = 5$	0.94	0.78
Pixilation	0.74	0.58
Silhouette	0.68	0.84
Binary Silhouette	0.48	0.75
Proposed Method	**0.02**	**0.27**

Table 1. Anonymising performance using the average cross correlation magnitudes computing on the bounding box.

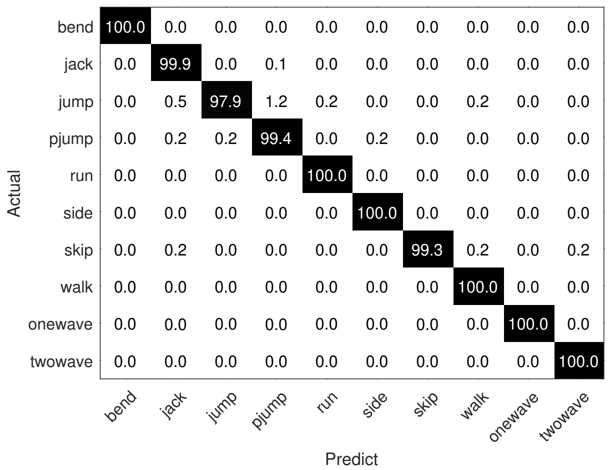


Figure 4. The KNN confusion matrix of Weizmann dataset using HOG-A (Overall accuracy is 99.64 %)

silhouette of the human object by distributing the highlighting based on the most dynamic parts. This was explained in the previous sections and Figures 2 and 3. This distribution of the salience makes the maps intelligible and the descriptor more robust to discriminate between the actions/activities because the values of the description are extracted based on this variation of the magnitudes.

Figures 4 and 5 show the confusion matrices of the proposed method applied on Weizmann and DHA using KNN classifiers, respectively. The results in confusion matrices prove that the proposed anonymising method contributes perfectly to design a robust descriptor, e.g. HOG-A. This descriptor discriminates the actions based on exploring the salience magnitudes to extract powerful discriminating features. The confusion matrices of KNN's classifier achieve 100% of accuracies for a 50% and 60% of the actions in DHA and Weizmann datasets, respectively, despite the similarities among the actions.

4 Conclusions

This paper has proposed a new anonymising method to address the problems of preserving the privacy efficiently while retaining a high level of utility of the anonymised video in video

Method	DHA	Weizmann
Original (not anonymised)	96.69 [19]	95.70 [20]
Blurring $\sigma = 8$	94.51	91.15
Blurring $\sigma = 5$	94.62	94.05
Pixilation	79.35	69.16
Silhouette	93.9	89.36
Binary Silhouette	91.97	93.61
Proposed Method	**99.39**	**99.64**

Table 2. Activity recognition accuracy (%) of the anonymised sequences in DHA and Weizmann datasets.

based recognition of activities of daily living. The proposed method anonymised the useful visual information as a mask corresponding to the temporal saliency in the video frames. This obfuscated information was explored in activity recognition without using extra algorithms to revise or recovery of the original visual data. The proposed method achieved a high level of privacy protection and high rates of accuracy in activity recognition when compared with the other methods. It showed improvements of 5.6% and 5.4% for Weizmann and DHA datasets, respectively, for activity recognition over the other methods used for anonymising the video.

References

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

[2] R. Li, B. Lu, and K. D. McDonald-Maier, "Cognitive assisted living ambient system: A survey," *Digital Communications and Networks*, vol. 1, no. 4, pp. 229–252, 2015.

[3] S. Pal and C. Abhayaratne, "Video-based activity level recognition for assisted living using motion features," in *International Conference on Distributed Smart Cameras*, Seville (Spain), 2015, pp. 62–67.

[4] T.-H. Tsai and K.-L. Zhang, "Implementation of intelligent home appliances based on IoT," in *Asia Pacific Conference on Circuits and Systems (APCCAS)*, Jeju (Korea), 2016, pp. 146–148.

[5] 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 Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[6] 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

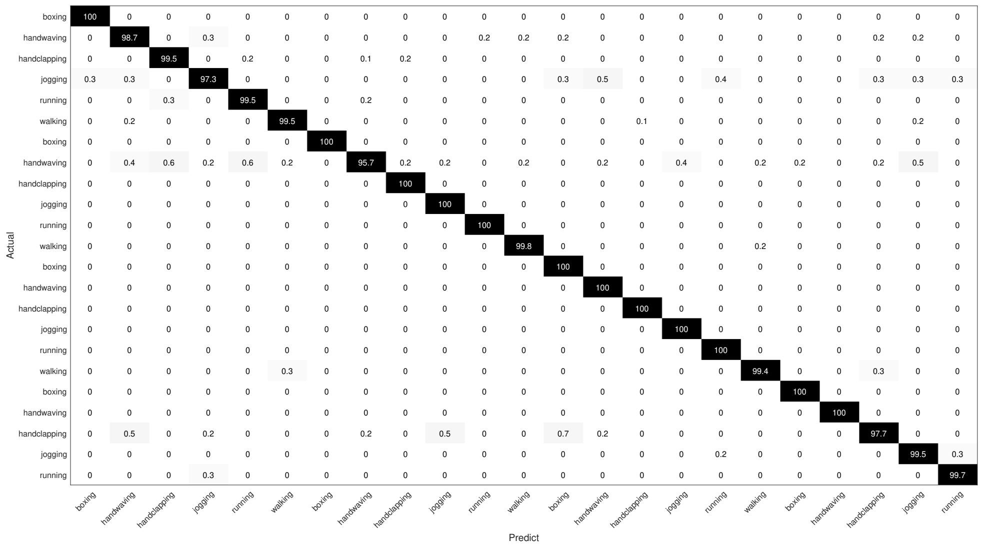


Figure 5. The KNN confusion matrix of DHA dataset using HOG-A (Overall accuracy is 99.39%)

International Conference on Computer Vision and Pattern Recognition Workshops, Boston (USA), 2015, pp. 68–76.

[7] M. Eldib, F. Deboeverie, W. Philips, and H. Aghajan, "Behavior analysis for elderly care using a network of low-resolution visual sensors," *Journal of Electronic Imaging*, vol. 25, no. 4, pp. 041 003–041 003, 2016.

[8] J. Zhao, N. Frumkin, J. Konrad, and P. Ishwar, "Privacy-preserving indoor localization via active scene illumination," in *International Conference on Computer Vision and Pattern Recognition Workshops*, Utah (USA), 2018, pp. 1580–1589.

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

[10] E. T. Hassan, R. Hasan, P. Shaffer, D. J. Crandall, and A. Kapadia, "Cartooning for enhanced privacy in lifelogging and streaming videos." in *International Conference on Computer Vision and Pattern Recognition Workshops*, Hawaii (USA), 2017, pp. 1333–1342.

[11] D. J. Butler, J. Huang, F. Roesner, and M. Cakmak, "The privacy-utility tradeoff for remotely teleoperated robots," in *International Conference on Human-Robot Interaction*, Portland (USA), 2015, pp. 27–34.

[12] B.-J. Han, H. Jeong, and Y.-J. Won, "The privacy protection framework for biometric information in network based CCTV environment," in *International Conference on Open Systems (ICOS)*, Langkawi (Malaysia), 2011, pp. 86–90.

[13] M. Saini, P. K. Atrey, S. Mehrotra, and M. Kankanhalli, "Adaptive transformation for robust privacy protection in video surveillance," *Advances in Multimedia*, vol. 2012, p. 4, 2012.

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

[15] 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 *International Conference on Multimedia*, Nara (Japan), 2012, pp. 1053–1056.

[16] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[17] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *International Workshop on Privacy Enhancing Technologies*, Cavtat (Croatia), 2005, pp. 227–242.

[18] I. Fan, "Image pixelization with differential privacy," in *Annual Conference on Data and Applications Security and Privacy*, Bergamo (Italy), 2018, pp. 148–162.

[19] 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 Transactions on Image Processing*, vol. 26, no. 10, pp. 4648–4660, 2017.

[20] J. Cai, X. Tang, and R. Zhong, "Silhouettes based human action recognition by procrustes analysis and fisher vector encoding," in *International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 10836, Shanghai (China), 2018, p. 1083612.