Pattern Recognition 127 (2022) 108621

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Human and action recognition using adaptive energy images

Onur Can Kurban[a,*], Nurullah Calik[b], Tülay Yildirim[a]

[a] Department of Electronics and Communications Engineering, Yildiz Technical University, Istanbul, 34220, Turkey
[b] Department of Biomedical Engineering, Istanbul Medeniyet University, Istanbul, 34700, Turkey

ARTICLE INFO

Article history:
Received 11 December 2019
Revised 18 February 2022
Accepted 2 March 2022
Available online 3 March 2022

Keywords:
Motion recognition
Human recognition
Correlation coefficients
Deep learning
Behavioral biometrics

ABSTRACT

In this paper, we propose a new temporal template approach for action recognition and person identification based on motion sequence information in masked depth video streams obtained from RGB-D data. This new representation creates a membership function that models the change in motion based on the correlation between frames that occur during motion flow. The energy images created with this function emphasize the intervals of motion with more change, while the intervals with less change are suppressed. To understand the distinctive features, the obtained energy images by using the proposed function are given as input to the convolutional neural networks and different handcrafted classifiers. The proposed method was observed on the BodyLogin, NATOPS, and SBU Kinect datasets and compared with the existing temporal templates and recent methods. The results indicate that the proposed method provides both higher performance and better motion representation.

1. Introduction

In recent years, human behaviour analysis has become one of the most popular fields of research. This analysis aims to understand the behaviour of humans at a certain time interval using measurable behavioural and physical motion information. Behavioural analysis is categorized into motions, gestures, actions, interactions, activity or person recognition and authentication depending on the time of the observation. This analysis investigates not only the behaviour of one person but also the interpersonal interactions and the interaction between persons and objects. The analysed behavioural movements can also be divided into subcategories such as upper or lower body, hand or arm movements, and face expressions. The recognition of complex human activities makes the construction of several important applications possible. Automated surveillance systems in public places such as airports and subway stations to detect abnormal activities; real-time monitoring of patients, elderly people, and children; gesture-based human-computer interfaces; and vision-based intelligent environments are examples of these applications [1].

In [2], Bobick and Davis presented the construction of a binary motion-energy image (MEI) that demonstrated where motion has occurred in an image sequence. In this view, the interaction area of the movement can be seen. This method has been widely used to recognize gait and motion. In [3], the MEI method was used for spatiotemporal gait representation called the gait energy image (GEI); human walking properties were characterized for individual recognition by gait. However, this representation loses detailed information and does not involve temporal changes. Next, in [4], Davis generated a scalar-valued image called a motion-history image (MHI), where intensity is a function of recency of motion. In [5], the gait history image (GHI) was developed based on GEI and MHI. This method eliminates the lack of temporal change in GEI. However, in this method, insufficient training cycles lead to problems. In [6], a gait moment image (GMI) was created with GEI. GMI is the gait probability image at every important moment of all gait cycles. These important moments of all gait cycles are averaged as evaluated in the GEI. However, it is not simple for GMI to select these important moments from cycles with different periods. In [7], Chen et al. divided the movement into clusters, obtained dominant-energy images (DEIs), and suggested the frame difference-energy image (FDEI) method by taking the difference between the images obtained by the GEI and DEI. Additionally, Megavannan et al. used an SVM to recognize human actions from depth information of MHI and average depth images [8].

Pose-based action recognition approaches mostly use full-body tracking information and different skeletal joints to extract features for action recognition. For recognizing human interaction, such as relationships between two people where one person moves and the other reacts, body-pose features, including joint features, plane features, and velocity features, are evaluated in [9] using an SVM and multiple instance learning. Zhu et al. in [10] suggest a

* Corresponding author.
E-mail addresses: ockurban@yildiz.edu.tr (O.C. Kurban), nurullah.calik@medeniyet.edu.tr (N. Calik), tulay@yildiz.edu.tr (T. Yildirim).

https://doi.org/10.1016/j.patcog.2022.108621
0031-3203/

skeleton-based action recognition approach that relies on the co-occurrence features of skeleton joints by using a frame that includes the 3D positions of the skeleton joints as input to a deep long short-term memory (LSTM) network. As an example of challenging situations, in [11] the authors presented a method to propose a solution to action recognition with partially occluded silhouette sequences. By adding occlusion to the lower and upper parts of the silhouettes, 2 different groups were created, namely lower occluded silhouettes and upper occluded silhouettes. On average, 90% success has been achieved. In [12], on the other hand, the authors have attempted to detect human joints from the image and to solve the problem of action recognition through their movements. First, independently human joints were detected, then these points were treated as a node system. Thus, they enabled the model to focus on motion rather than recognizing the human body. Data were classified with an LSTM-based recurrent neural network. Thanks to this approach, the movements of people in the same view can even be classified independently of each other. They achieved a score of 95.96% in the UTkinect action 3D dataset.

In [13] Ke et al. proposed transforming a skeleton sequence to three video clips for action recognition. Using a pretrained CNN model, they extract a compact representation of each frame. The temporal information of the entire skeleton sequence and one particular spatial relationship between the joints are concatenated in a single feature vector. The obtained features were implemented in the proposed MTLN network. Another CNN-based study is [14]. The model proposed by Russell et al. performs action recognition by processing spatial and temporal streams. The model uses two images. First CNN structure is fed via a frame produced over the motion vectors at time t. Another CNN structure working in parallel uses dynamic image created by using f1,f2,...,ft. Classification is made by fusioning the result scores obtained. In addition to the KTH dataset, Weizmann, UCF sports and UCF101 datasets were used in the study. Besides, Singh et al. proposed a two-stream deep ConvNet model that includes both an upper layer using RGB frames and a lower layer trained with a single dynamic motion image for recognition of single, multi-person, and human-object interaction in [15]. The results from both layers are fused at the decision level.

In [16], Wang et al. proposed a method to encode spatiotemporal information of skeleton sequences referred to as joint trajectory maps (JTMs), and ConvNets were adopted to exploit the discriminative features for real-time human action recognition. In addition to the front, top and side features of the skeleton sequences used in the JTM method, Li et al. in [17] added a 4th CovNet channel where all of these features are applied together. This new method, referred to as joint distance maps (JDMs), was used for human action and interaction recognition. In this way, the JDM method obtained by the improvement of the JTM achieved more successful results on the NTU RGB+D and small UTD-MHAD datasets. In [18] Simonyan and Zisserman proposed a two-stream CNN architecture that combines RGB and optical flow data for action recognition on the UCF101 and HMDB51 datasets. The fusion of these spatial and temporal features has provided performance increments. In addition, this architecture proposed in [18] was further improved in [19] by Bilen et al. with an idea similar to the previous study. They presented a four stream architecture that combines RGB data, optical flow data and their dynamical streams for action recognition. They noted that using the dynamic image method provided a more detailed representation than the MHI [2] method. One of the recent studies on human action recognition is [20]. The authors created energy images by taking the projections of the motion series images not only in the plane (x) where the image is formed but also in the y and z directions. Later, they contributed to the improvement of improved dense trajectories by extracting the histogram of oriented gradients (HOG), histogram of optical flow (HOF) and motion boundary histogram (MBH) feature vectors from these three images. KTH, UFC101 and HMDB51 datasets were used for benchmarking.

A recent study, the Correlation of Temporal Difference Frame (CTDF) method is proposed by Poonkodi and Vadivu for action recognition in [21]. This method, generates interest points locally according to the correlation between three consecutive frames of silhouette sequences. Spatio-temporal neighborhoods are obtained using the Harris mathematical operator. In [22], Background subtraction is made using the frame difference technique based on the correlation coefficient and silhouette images are obtained. Euclidean distance transform and Shannon entropy are used to extract features from these images. Combinations of these features are given as input to a feed-forward neural network to recognize various human actions. In [23], on the other hand, the 3D CNN model is used for the recognition of human movement through short videos. The proposed architecture provides action classification, taking into account the time-motion features with the spatial location of the activities. While spatial features are extracted with CNN, a temporal relationship is established between frames thanks to the convolutional long short-term memory layer. And finally in [24], Chen et al. presented a skeleton-based human action recognition method that provides spatial relationships of human skeletal joints and long-term temporal information in time. In this approach, the skeleton sequences are transformed into color motion images, and they are applied as input to a shallow CNN model for action classification.

In summary, energy images have become a popular subject of study in recent years, and the development of new methods in this area continues.

1.1. Contribution

Previous studies on motion recognition, especially the recent articles mentioned above, create ideas for new researchers and show us that researchers have developed these ideas and carried them to a more efficient structure. In addition to these improvements, Ijjina et al. propose three specific temporal templates (TT), unlike the traditional MHI and MEI methods, that emphasize various temporal regions (beginning, middle, and ending) of motion for recognizing actions based on motion sequence information in RGB-D video using deep learning [25].Another study using energy images is [26]. In this study, Abdelbaky and Aly propose a human action recognition method by fusing spatial and temporal features learned from principal component analysis network (PCANet) in combination with bag-of-features (BoF) and vector of locally aggregated descriptors (VLAD) encoding schemes. In this study, each video dimensionality is reduced using whitening transformation (WT), and both spatial and temporal features are learned through PCANet. The short-time motion energy images (ST-MEI) are used to calculate the temporal stream. The stream features are fused at the feature level and classified using SVM. Another two-stream structure is presented in [27]. The authors proposed a two-stream convolution neural network model to recognize single-person behavior and two-person interaction behaviors. In the VMHI structure, motion history images (MHI) are obtained from video frames of motions. Then, they are applied to a VGG-16 network. In the FRGB structure, a Faster R-CNN algorithm is used to detect human activities from the RGB frames. M-scores and R-scores obtained from these parts are fused at the decision level. Mishra et al. used a static 5-part function as membership in their study in [28]. Herein, to model the flow of motion well, they aimed to multiply different frames with high values while multiplying the places where the emphasis of motion decreased with low coefficients. They obtained a precision of 0.96 on the HMDB51 dataset by using a CNN they proposed. However, the increase in the number of temporary templates as in Ijjina's study and the use of fixed templates for special inter-

vals as in Mishra's study also bring some problems. Each template focuses on different action areas, which will produce different results for different motion types. As a result, these methods may produce good results for actions that coincide with the highlighted areas, while the representation ability will be reduced for actions outside the highlighted area. If the action area is wider than the area highlighted by any TT, fusion of multiple TTs is required to better represent the action. To overcome these difficulties, we propose a new temporal template representation that is adaptive to the movement, that is, generated according to the changes that occur throughout the movement.

In the proposed method, to determine the changes that occur during a motion sequence, the first frame is taken as a reference, and the correlation coefficients between all frames are calculated. The obtained coefficients are used to construct a motion-adaptiveTT function. Energy images generated by the adaptive TT function will have a better representation of the movements than energy images generated by existing TTs that highlight specific periods. In this way, a practical method is proposed that provides a better representation of actions in energy images and reduces the cost of processing, as it eliminates the need for fusion of multiple TTs.

In addition, RGB-D-based studies are mostly used in the literature for motion recognition [29]. Another contribution in this study is that the existing and our proposed methods were also examined for person identification.

VGG-F, ResNet-18 and handcrafted feature extractors and classifiers were used for performance analysis. The VGG-F and ResNet-18 networks, a type of CNN, are used to evaluate our method and to compare it with existing temporal template representations presented in [25] for motion recognition and person identification. First, the results of the training were compared with previous studies in the literature. Then, the advantage and performance of the adaptive TT we proposed over previous TTs were tested under the same conditions. Finally, the performance between the existing TTs and our method has been examined with handmade feature extractors and classifiers to better understand whether the performance is from the power of CNN networks or the proposed adaptive method. The BodyLogin [30,31], Naval Air Training and Operating Procedures Standardization (NATOPS) [32], and SBU Kinect [9] databases are used for experiments. The results indicate that the proposed adaptive TT provides high performance and shorter process duration than the fusion of three existing templates. The person identification results showed that the images obtained with RGB-D sensors can be used effectively for person identification.

In summary, the advantages of this method can be listed as follows:

1. Adaptable to the completion time of the action.
2. The intervals where changes are greater during movement are identified and highlighted. Conversely, The intervals that do not contain any changes are suppressed.
3. There is no need to determine which static TT is more suitable for an action using empirical methods.

1.2. Remainder of the paper

This paper is organized as follows: In Section 2, first, the NATOPS, SBU and BodyLogin databases and datasets derived from them are described. Then, the existing methods and the details and advantages of the proposed method are explained. Finally, the training models and parameters used for classification and identification are presented. In Section 3, first, the comparisons of the obtained results with the studies in the literature are presented, and then the proposed TT and existing TTs are applied to the same networks, so the performance differences in equal conditions are examined. Finally, all TT performance results in handcrafted classifiers are presented. The discussion and conclusion are presented in Section 4 and Section 5, respectively.

2. Material and methods

2.1. Benchmark datasets

In this study, three databases were used to observe the contribution of the proposed method. Visuals in the NATOPS and SBU Kinect databases were evaluated according to the body and the moving parts of the individuals. With this method, two different energy image datasets were created from each. In addition, one dataset was generated using the BodyLogin database. In this way, five different datasets were obtained. The accuracy performance between existing methods and our proposed method has been examined using these datasets.

Details and properties of the datasets are described in the subsections. The training networks used for comparisons and their parameters are also explained in this section.

2.1.1. NATOPS dataset

The NATOPS dataset consists of 24 aircraft handling gestures. These gestures are arranged according to the NATOPS manual. The movements in this dataset include upper-body and arm motions and hand gestures (thumb up/down, hands opened/closed). Twenty subjects performed 24 upper-body gestures 20 times.

The average length of all samples was 2.34 sec (std. dev=0.62). The collected dataset was recorded in two different formats. The first one is the feature dataset (CSV format). This includes body and hand features as well as segmentation labels. The second one is the video dataset. This video dataset includes stereo camera-recorded images, depth maps, and mask images of gestures [32]. These gestures were recorded using a Microsoft Kinect sensor at 20 FPS with 320×240 resolution, as shown in Figure-1.

In our study, the video dataset was used because our motivation was based on the energy image. Therefore, subject-gesture pairs are created for motion recognition. For evaluation criteria, training and test groups were established at different rates.

In [25], the RGB features were found to be more successful than the depth features. This may be due to the yellow colour vest worn by the subjects that caused the high contrast difference between the arms and the body. Additionally, in the depth template, the arms of the subjects overlap with their bodies. Due to this overlap, the arm actions are well acquired from RGB templates rather than the depth templates. However, while this database was being created, subjects were wearing special clothes. This situation will not always be the same for every subject. This advantage will be lost when subjects wear clothes with the same arm and body colour, in which case the RGB templates will not be able to make a robust evaluation. Thus, in this study, it was preferred to use masked depth templates, which can show the same robustness in all cases, regardless of the colour of clothing worn by the subjects. The acquisition of the masked depth images in the video dataset is shown in Figure-2. Here, the first line contains RGB images. The images in the second line are raw depth images collected with the Kinect camera. These raw images contain a large amount of background noise. To remove this unnecessary information, RGB images and raw depth images are masked. The masked depth dataset created in this way is shown in the last line.

Two different energy images were generated when creating the datasets. In the first type of image, moving limbs and body images are used together. These energy images are shown on left side of Figure-3. This dataset will hereafter be referred to as NATOPS*Full – Body* after this. In the second type of image, only the moving limb (hand-arm) information is used, and the body information is removed from the image, as used in [25] These images are shown

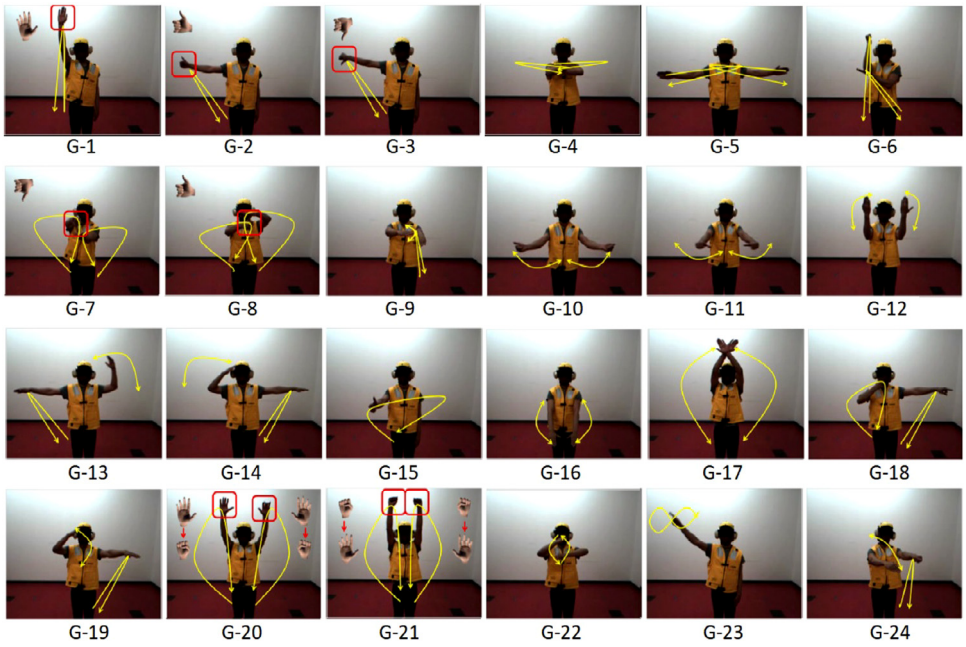


Fig. 1. The patterns of 24 motions in the NATOPS database as shown in [32]. The movements are completed in the direction indicated by arrows.

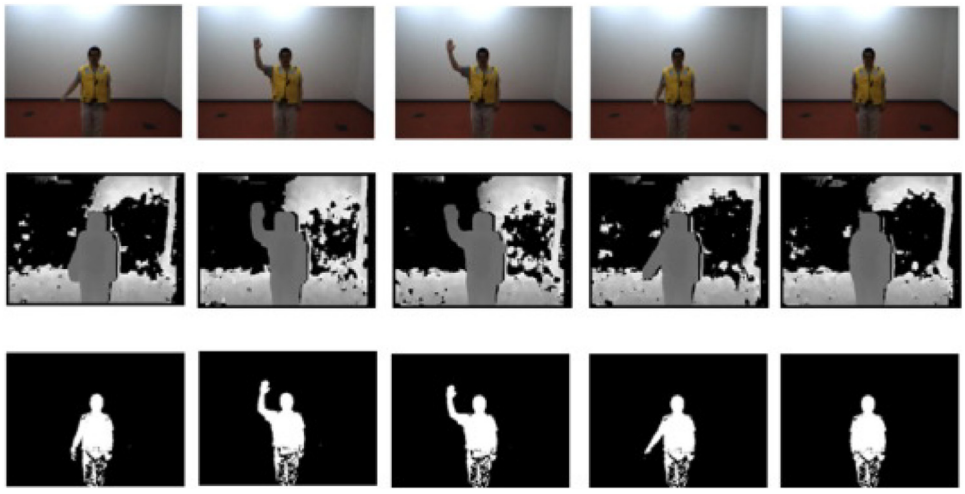


Fig. 2. The first line represents RGB images. The images in the second line are raw depth images collected with the Kinect camera. The last line shows the masked images generated with RGB images and raw depth images.

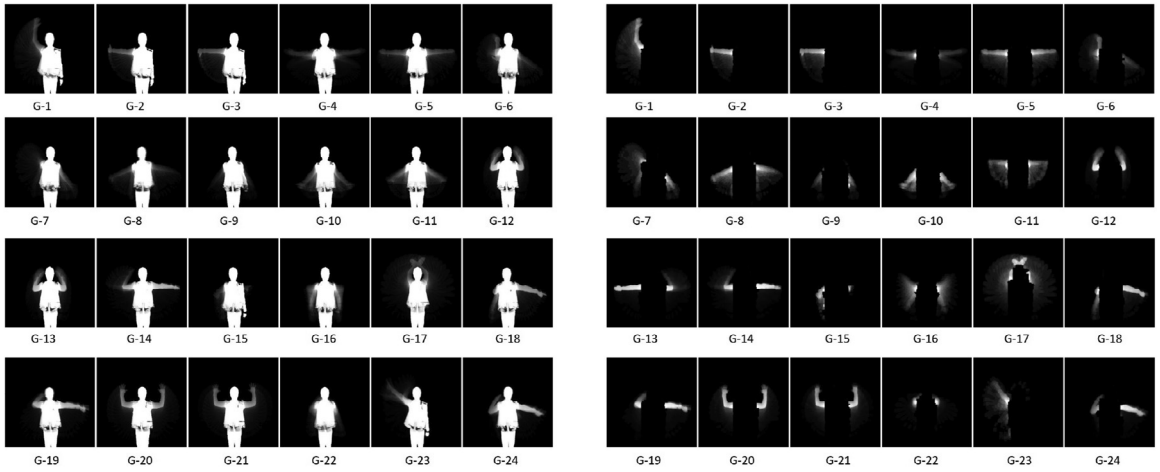


Fig. 3. Two different energy image datasets were created using NATOPS database. This image shows samples of "NATOPS*Full – Body*" (left side) and "NATOPS*Moving – Limbs*" (right side) datasets generated using μ_R in masked depth video streams.

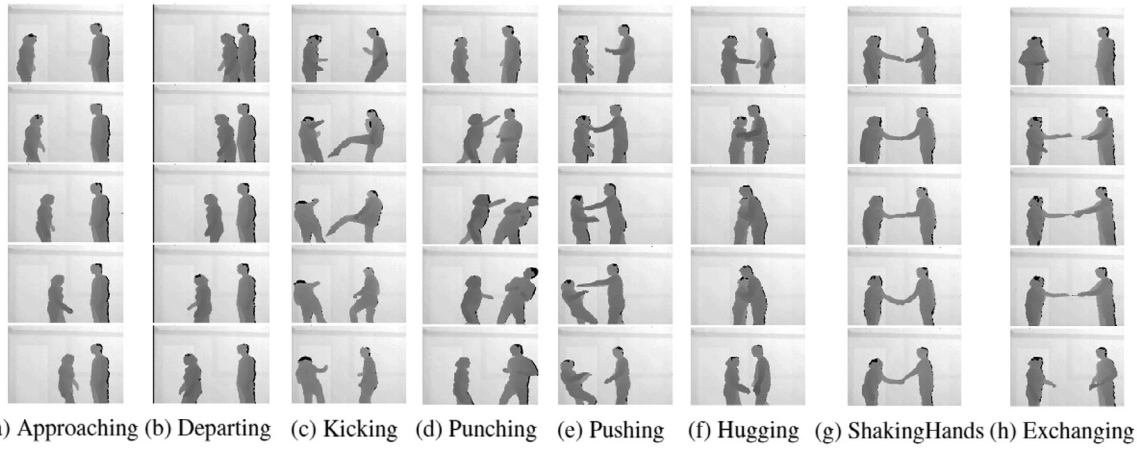


Fig. 4. The patterns of 8 interactions in SBU kinect database as shown in [9]. Some movements are similar to each other like punching-pushing, hugging-approaching and shaking hands-exchanging. This makes the data set more challenging, even if there are fewer movements.

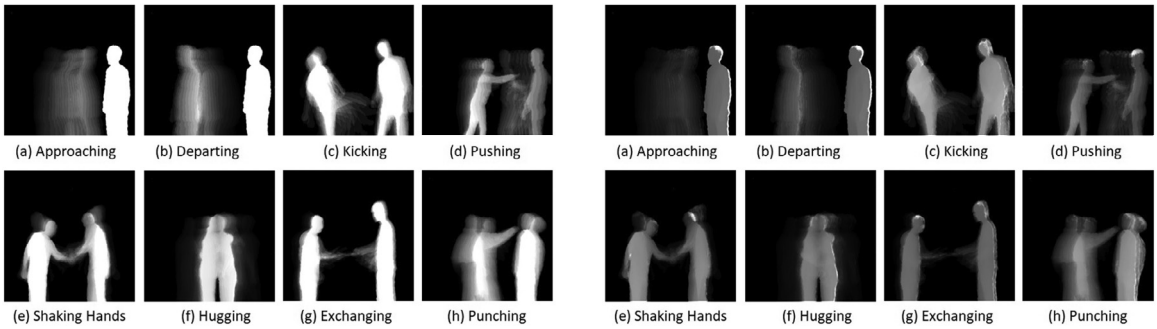


Fig. 5. Two different energy image datasets were created using SBU Kinect database. This image shows samples of "SBU$Full-Body$" (left side) and "SBU$Moving-Limbs$" (right side) datasets generated using μ_R on depth video streams.

on the right side of Fig. 3.This dataset will hereafter be referred to as NATOPS$Moving-Limbs$ after this. In this way, two different energy image datasets were produced using the NATOPS database.

2.1.2. SBU kinect dataset

This dataset consists of eight movements involving the interaction of two persons. These movements are approaching, departing, kicking, punching, pushing, hugging, shaking hands, and exchanging, as described in [9] (the frames are shown in Fig. 4).

Some of these actions, such as exchanging an object and shaking hands, are structurally similar to each other because both subjects extend their arms. This makes the database challenging. In some of these actions (pushing, kicking, approaching, departing, and punching), one subject initiates the movement, and the second subject reacts to it.

For each interaction, RGB and depth video images are acquired with a resolution of $640X480$ pixels with 15 frames per second (FPS). Depending on the type of motion, each action is completed within the range of 20 to 40 frames. The observations in this dataset are obtained using 21 different interaction pairs.

As in the NATOPS dataset, two different energy image datasets were generated using the SBU Kinect database. In the first dataset, a full body image is used. These energy image samples are shown on the left side of Fig. 5. This dataset will hereafter be referred to as SBU$Full-Body$. In the second dataset, only the 'moving limbs of the body' information is used, and the 'stable body' information is removed from the images. Some of these images are shown on the right side of Fig. 5. This dataset will hereafter be referred to as SBU$Moving-Limbs$. As shown in Fig. 5, SBU$Moving-Limbs$ contains a whole-body image as opposed to NATOPS$Moving-Limbs$. The movements in the SBU Kinect dataset are made up of fully active body actions during the interaction. In this way, two different energy image datasets were obtained from the SBU database

2.1.3. BodyLogin dataset

The BodyLogin database consists of different dataset groups recorded at different times. In this study, two of these datasets are combined and used. The first of these is the Silhouettes and Skeletons dataset (BLD-SS) and contains two types of gestures performed by 40 different subjects (27 men and 13 women aged 18–33 years). While the dataset was created, each subject was asked to perform two unique short movements with 20 samples, each nearly 3 seconds long. Both gestures were made to cover both the lower and upper bodies.

BLD-SS S gesture: The subject draws an S shape with two hands. Despite being simple, this gesture is done in a similar way by all users. Although easy for motion recognition, it is more difficult for person identification because of this reason. *BLD-SS User-defined gesture*: Unlike the previous gesture, the subject selects his or her gesture without any instructions. Although potentially complex, this action is unique to most users. This makes the dataset easier for person identification [30].Illustrations of movements are shown in Fig. 6.

The second dataset is the Posture, Build, and Dynamics dataset (BLD-PBD), which includes three types of gestures performed by 36 different subjects (25 men and 11 women aged 18–33 years). While the dataset was created, each subject was asked to perform three unique short movements with 20 samples, each nearly 3 seconds long. These gestures (*left-right, double-handed arch, and balancing*) were planned with different complexity levels and involved motions in both the upper and lower body.

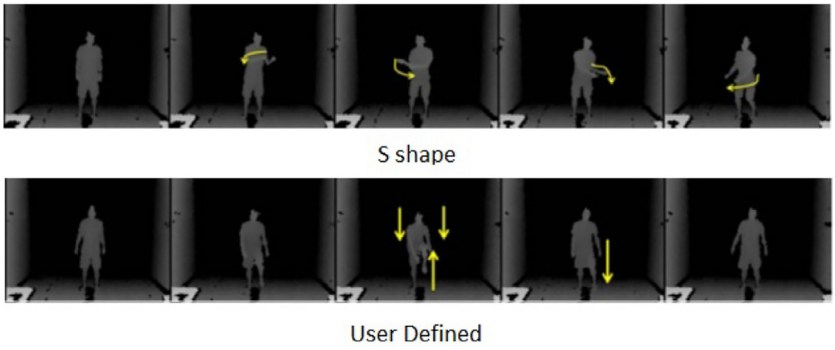


Fig. 6. The patterns of S-shape and User Defined gestures in BodyLogin Silhouettes and Skeletons dataset as shown in [30].

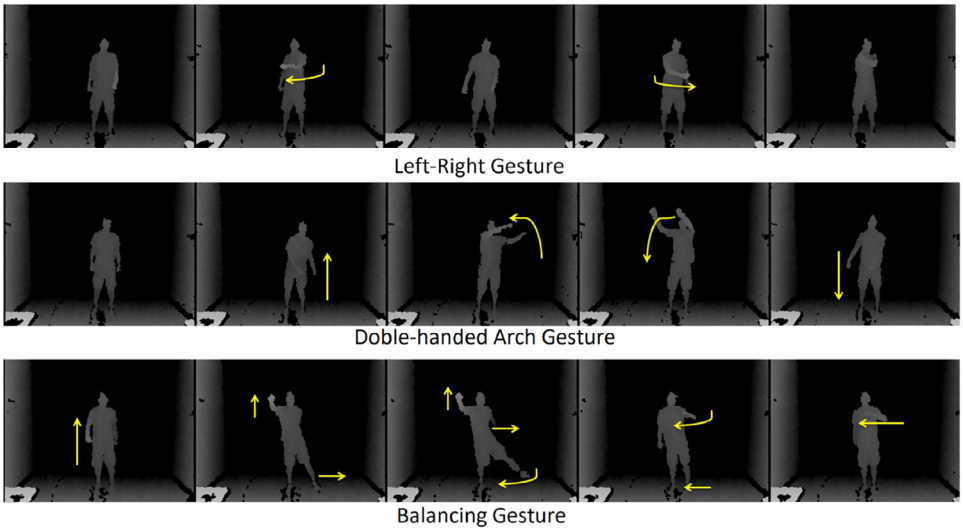


Fig. 7. The patterns of Left-Right, Double-handed Arch and Balancing gestures in BodyLogin Posture, Build, and Dynamics dataset as shown in [31].

BLD-PBD Left-Right gesture: The subject reaches the right shoulder with the left hand and then reaches the left shoulder with the right hand. *BLD-PBD Double-handed arch gesture*: Subject draws an arch shape from left to right with both hands. *BLD-PBD Balancing gesture*: The subject tries to stand in balance and raises the right arm forward while pulling the left arm back while sweeping the left leg forward. Each subject gesture was acquired using a Windows Kinect camera (version 1) at a distance of approximately 2 metres. Using an official Kinect camera SDK, $640X480$ depth images and skeletal joint coordinates were captured at 30 FPS [31]. Illustrations of movements are shown in Fig. 7.

The gestures in these two datasets were combined to form a new dataset containing four movements. As in the previous sections, this dataset was also used to identify subjects with the proposed method. Only the BLD-SS $User-Defined$ gesture dataset is used for motion recognition because it contains 40 different movements.

Unlike previous datasets, only one type of energy image dataset called BodyLogin$Full-Body$ was generated due to the structure of the dataset. These energy images are shown in Fig. 8.

2.2. Existing approaches

The TTs emphasize the frames that occurred during the movement, such as the MEI or MHI methods, and try to reveal the distinctive features of the movement.

The TTs used for this purpose in previous studies are shown in Fig. 9, and the equations of the calculation methods are given in Equations (1–7), where n represents the number of frames in motion, $m(i)$ indicates the i^{th} frame of motion and w_i is the weight value that multiplies the greyscale image pixels. Additionally, the frame pixels in the range of $0 to 255$ are normalized to the $0 to 1$ range after applying the w_i weights. This value has been expressed in previous studies as a fuzzy membership function $\mu(i)$. If $\mu(i)$ parameters can be applied to each frame specifically, this can help emphasize movement locally and use motion information more successfully for recognition and identification. Ijjina and Chalavadi [25] studied three temporal template models by comparing the traditional energy image-generating model to demonstrate this behaviour. These models were investigated as the $\{begin, middle, end\}$ of movements as shown in Figure-9 and given in Equations 4–7, respectively.

$$TT = (\frac{1}{255}) \sum_{i=1}^{n} w_i.m(i) \tag{1}$$

$$= \sum_{i=1}^{n} (\frac{w_i}{255}).m(i) \tag{2}$$

$$= \sum_{i=1}^{n} \mu_i.m(i) \tag{3}$$

In Fig. 9, μ_1 corresponds to the computation of MEI, which is the most basic method for creating an energy image. Since μ_1 has a constant value, the same weight is assigned to motion in all temporal regions.

$$\mu_1(i) = 1, \forall i \in [0\, n] \tag{4}$$

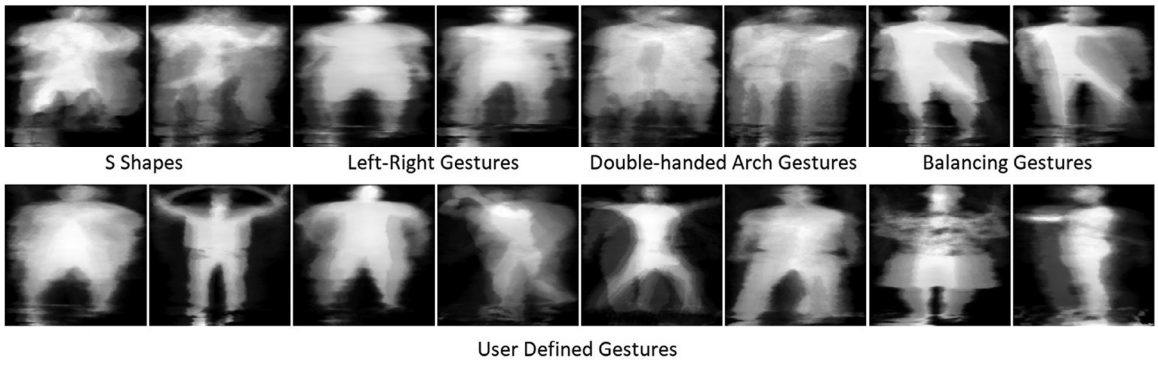


Fig. 8. Showing "BodyLogin$Full-Body$" energy image datasets generated using μ_R on depth frames. The upper line shows two images of the four movements used for person identification. The bottom line shows some of the energy images used for motion recognition.

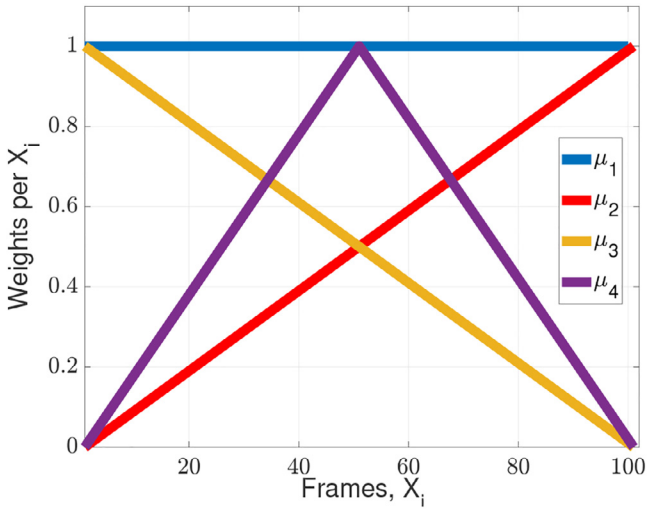


Fig. 9. Graphs of previously used fuzzy membership functions (μ_i). Each μ_i emphasizing a different temporal region.

The μ_2 function computes the traditional MHI. Since μ_2 is a linearly increasing function, in an MHI, the importance assigned to motion information increases linearly with time, i.e., towards the end of action, the motion information has the highest importance.

$$\mu_2(i) = \frac{i}{n}, \forall i \in [0\,n] \tag{5}$$

In case of μ_3, the weights assigned to motion information are in the opposite direction of μ_2, i.e., the weights decrease linearly with time. Therefore, the oldest motion information has the highest importance.

$$\mu_3(i) = 1 - \frac{i}{n}, \forall i \in [0\,n] \tag{6}$$

The last computing function μ_4 assigns weights in a triangular shape. The highest weight is assigned in the middle of the motion, with the weights decreasing towards the beginning and end of the motion.

$$\mu_4(i) = \begin{cases} \frac{2i}{n}, & 0 < i \leq \frac{n}{2} \\ 2 - \frac{2i}{n}, & \frac{n}{2} < i \leq n \end{cases} \tag{7}$$

2.3. Novel data adaptive correlation-based TT

In the current studies, energy images are obtained by applying any equation selected from the template models shown in Fig. 9 to the n images taken during the motion. These images can be used for motion classification using machine learning methods. However, the success of these methods varies depending on the applied motion. That is, the template model to be used varies according to the shape of the movement, the timing of the movement, the speed of the person making the move, and the repetition of motion [25]. Unfortunately, this situation requires experimenting with either all the template methods to obtain the most successful result or analysing in which region the motion contains more intense information.

Table 1
Correlation coefficients and their significance levels[33].

R (Coefficients Value)	Correlation Status
0.00	No Correlation
0.01 – 0.29	Low Level Correlation
0.30 – 0.69	Medium Level Correlation
0.70 – 0.99	High Level Correlation
1	Perfect Correlation

In this study, we propose a motion-energy calculation method that can better represent the changed information of movements without the need for an empirical method. The proposed method ensures that the temporal weights are generated adaptively to the motion and that the energy images are obtained with these temporal weights. This method is based on the calculation of the correlation coefficients between images. Correlation, as is known, shows (more or less) a linear relationship between two variables. Thus, changes between frames occurring during an action can be examined by correlation. If the correlation between the two images is low, we can infer that there is a change in the movement of the individual, and if the correlation is high, there is almost no change in the movement of the individual. To express it better, the 'Pearson Product-Moment Correlation Coefficients' can be examined in 1.

The correlation coefficients of two random variables are a measure of their linear dependence. If each variable has N scalar observations, then the Pearson correlation coefficient is defined as

$$P(A,B) = \frac{1}{N-1}\sum_{i_1}^{N}\left(\frac{\overline{A_i - \mu_A}}{\sigma_A}\right).\left(\frac{B_i - \mu_B}{\sigma_B}\right) \tag{8}$$

μ_A and σ_A are the mean and standard deviation of A, respectively, and μ_B and σ_B are the mean and standard deviation of B, respectively. Alternatively, the correlation coefficient can be defined in terms of the covariance of A and B.

$$P(A,B) = \frac{cov(A,B)}{\sigma_B.\sigma_A} \rightarrow R = \begin{pmatrix} P(A,A) & P(A,B) \\ P(B,A) & P(B,B) \end{pmatrix} \tag{9}$$

The correlation coefficient matrix of two random variables is the matrix of correlation coefficients for each pairwise variable

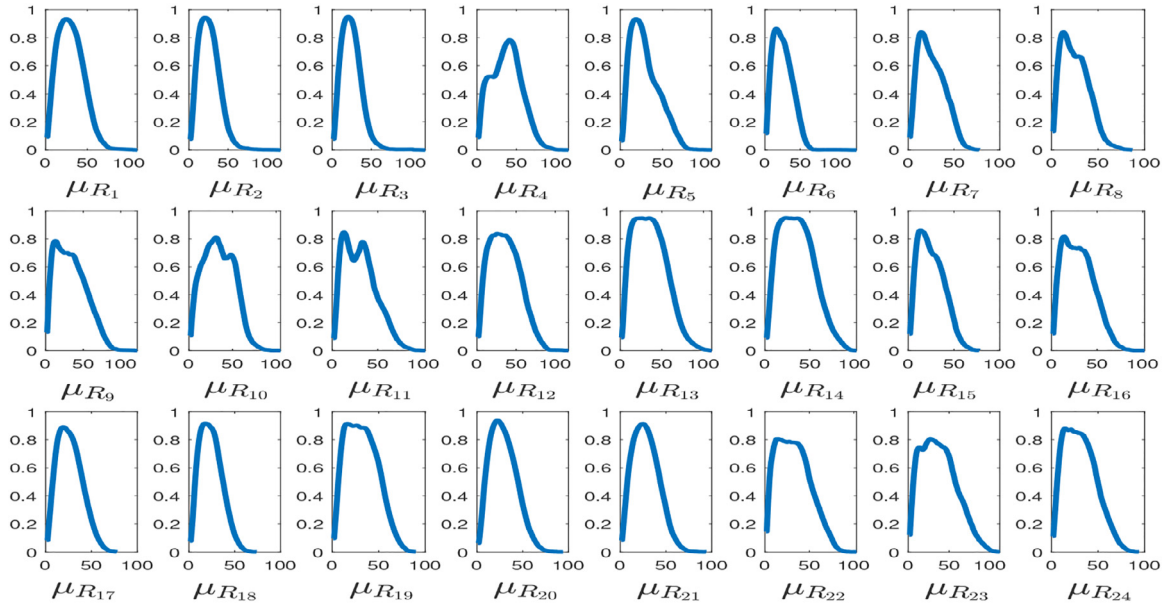


Fig. 10. TT graphs obtained by $(1 - R_N)$ operation on 24 different motions from the NATOPS dataset. Although the movement frames are fixed, some movements end earlier and some movements end later. In some movements, the beginning parts are emphasized, while in others, both the beginning and middle parts are emphasized. Since the proposed method is adaptive to movements, the TTs to be obtained will vary according to the data set and the content of the movements.

combination. Since A and B are always directly correlated to themselves, the diagonal entries are just 1 [34].

If all frames acquired during a movement are correlated with a frame that is selected as a reference frame and the reference frame is chosen as a static pose before movement begins, the correlation value will decrease because of any change in movement. On the other hand, the correlation value will rise again when the subject returns to the former position while completing movement. The resulting graph will provide information about the areas where the motion is changed, the duration of motion, and the rapidity of change.

In this process, $P(A, B)$ in Eq. (9) will have a value that can be expressed as low, medium or high correlation, as indicated in Table 1 since all the pictures belong to the same person during an action. The correlation values generated during motion are normalized within the range of $0 to 1$ and expressed as R_N so that these values can be better represented for using the temporal template.

Here, our method goes through a process to be adaptive. This process is expected to be as short and fast as possible. The correlation coefficients are calculated quickly, but if these values are used as they are, they will highlight similarities and suppress differences. The aim is to suppress the areas where there is little variability in the movement and to highlight the parts where there is considerable variability. To eliminate these antipodes, the function coefficients are used as $(1 - R_N)$. In this way, the function is adapted to the movement.

The suggested method in this work for the generation of energy images used in motion recognition and person identification is defined as in Equations (10) and (11).

$$\mu_{corr}(i) = (1 - R_N), \forall i \in [0\, n] \tag{10}$$

$$EI_{corr} = \sum_{i=1}^{n} \mu_{corr}(i).m(i) \tag{11}$$

In Fig. 10, examples of temporal templates generated by the method given in this study are presented. These weights were obtained from 24 different actions in the NATOPS dataset.

The adaptive TTs generated for each movement can be seen here. Additionally, if all templates are analysed, it is seen that the completion times of the movements, the areas of action, and the intensity of change are different. Even though the graphics are shown on a scale of 0–100 frames, some movements end in the range of 70–80 frames. However, since the video recordings continue after the movement ends, the total number of frames is more than the range of motion. The frame coefficients that continue at the 0 level towards the end in the graphic are the result of this. For example, in Fig. 10, the second and third examples of the first line peaked at approximately the 30th frame and ended at approximately the 60th frame, and the video continued until the 100th frame. In these movements, the middle part of the movement was emphasized, and its beginning and ending parts were suppressed. In addition, when the 6th function of the second line and the 7th function of the last line are examined, it is seen that the movement continues as much as the 100th frame. Here, the beginning and middle parts of the movements were emphasized, and the ending parts of movements were suppressed. When energy images are created with these suggested adaptive TTs, there will be no need to examine empirical methods or areas of change of motion for selecting the appropriate temporal template, as in the previous methods.

Examples of energy images generated by the adaptive TT and existing TTs are shown in Fig. 11. Three samples were added for each dataset. Examples are presented for μ_2, μ_3, μ_4, and μ_R. The effects of the differences in TT functions on the energy image can be better examined and understood with this figure. A more detailed illustration of the differences between the outputs of the proposed method and the existing methods is presented.

2.4. Classifiers

Three classifiers are chosen for testing the proposed method and other TT algorithms. The first and second classifiers are VGG-F and ResNet-18, which are deep learning-based and frequently used in the literature for feature extraction. The third set of classifiers consists of an SVM with different kernel types (RBF, Poly, and Linear) and k-NN with k = 3, 5, 7.

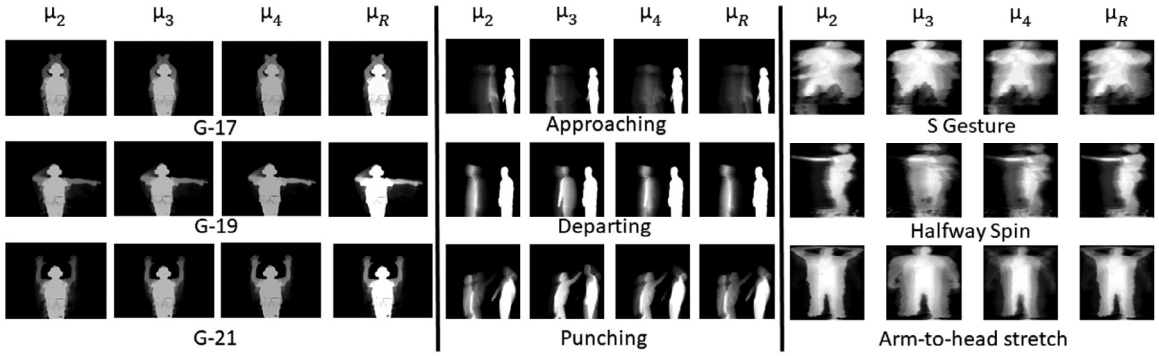


Fig. 11. Samples of energy images generated by the adaptive TT and existing TTs are presented for NATOPS, SBU Kinect and BodyLogin datasets, respectively. The samples are energy images obtained using the u2, u3, u4 and uR functions from left to right. The outputs of the functions are better illustrated by these samples.

2.4.1. VGG-F

The VGG-F network is one of three networks proposed by the Visual Geometry Group [35]. In their study, networks were developed using different parameters, and their performances were observed under various scenarios. The aim of the work is to understand which mechanisms are effective in designing a good CNN network. Based on this study, the VGG-16/19 networks proposed in [36] were developed. In the 2014 ImageNet competition, this study took first and second place in the localization and classification fields.

The VGG-F network architecture has been developed for performing fast and practical experiments. Many studies use the VGG-F network as a back-end for developing their algorithms [37,38]. CNNs are used as an effective classifier for testing the methods proposed in these studies. Usage of pretrained networks is especially popular. In this work, instead of a pretrained network, we use Glorot (or Xavier) random weight initialization proposed in [39]. The VGG-F network was trained using a back-propagation algorithm in batch mode. The batch size and training epoch number were selected as 16 and 50, respectively, for all training networks.

2.4.2. ResNet-18

In ResNet-type networks, there are alternative paths for information flow compared to VGG-type networks. If the weights in a layer negatively affect learning, these layers can be bypassed by shortcut connections. In this way, the model tries to find a suitable solution where it can learn data by minimizing the vanishing gradient problem in deeper networks.

When ResNet was published, it appeared with 5 different models [40]. Of these, ResNet-18 and 50 are among the most noted and used in many studies as backbone netwoks. In this study, *ResNet-18* is used to make a quick comparison for the classification of feature silhouette images.

The ResNet-18 model is created using MATLAB with random weights according to Glorot weight initalization [39]. The reason for not using pretrained model is to demonstrate the capabilities of TT methods more fairly. The input for the ResNet model is given as $224 \times 224 \times 3$. The TT silhouette image is transformed into a 3-depth image by placing a grey-level version on the R, G, B channels. The last neuron layer, called the logit layer, is replaced by an FC layer in which the number of neurons corresponds to the number of classes in the problem. Then, the model is trained and tested on 5 different randomly allocated clusters for 80%, 50% and 20% training ratios for each problem. The average of these 5 results reveals the success of the TT algorithms. Thirty epochs of 48 minibatches were selected for hyperparameter analysis. The cross-entropy is used as the loss function.

A summary of CNN classifier usage is given in Fig. 12. Classifications are made through this systematic approach. Networks are trained separately for PI and MR tasks.

2.4.3. Handcrafted classifiers

In addition, we analysed whether the success achieved depends on the generalization capability of the CNN network or the proposed method. To perform this analysis and observe the effectiveness of the proposed μ_R function, the features of energy images generated in this work were extracted with principal component analysis (PCA), local binary patterns (LBP), and histogram of gradient (HoG) methods and then evaluated with classifiers such as the deep neural network-based multilayer perceptron, k-nearest neighbours (k-NN), and SVM.

3. Experiments and results

In this work, we used the NATOPS Gesture, SBU Kinect Interaction, and BodyLogin datasets to evaluate our proposed method and the existing approach for both motion recognition and person identification. The datasets used contain RGB-D videos and images captured using a Microsoft Kinect depth sensor. Since the depth information captured by the Kinect sensor at each pixel is of low accuracy and can be ignored according to the scope of the study, instead of using the greyscale, we masked and binarized the depth video stream frames. Thus, the temporal templates computed using binarized depth video streams have spatial positions of the subjects, similar to a silhouette.

The first of the membership functions given in Fig. 9, μ_1, was proposed by Bobick and Davis to create an energy image in [2] and led to the development of different methods in subsequent studies. In [25], Ijjina et al. proposed three new membership functions, μ_2, μ_3 and μ_4, which provide an alternative to the μ_1 method with emphasis on temporal regions to create an energy image. The energy images obtained by these functions achieved better results than those obtained with μ_1. In addition, the advantages of μ_2, μ_3 and μ_4 compared to μ_1 are shown in detail. However, the choice of which of these three membership functions can give better results in which dataset or movements is entirely experimental. Therefore, in this study, a method based on changes in movement is proposed instead of the μ_2, μ_3 and μ_4 methods presented in [25]. Using the μ_R function, we employed a method that will eliminate the experimental selection difficulties of these three functions and will better represent the changes in movements in the creation of the energy image. In this context, we consider that it is not necessary to make a comparison between μ_1 and μ_R.

The 2- and 5-fold cross-validation methods were used for evaluation of all the datasets. In 2-fold cross-validation, the data were randomly divided into two equal parts: one group was used for

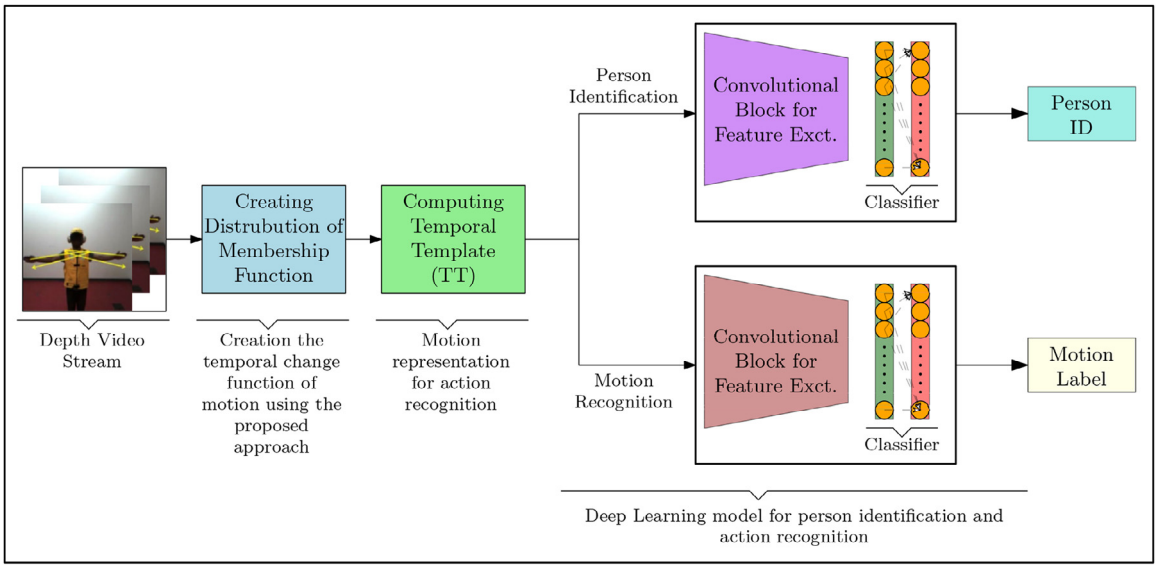


Fig. 12. General framework for CNN classification of energy images generated by TT. VGG-F and ResNet18 are used by creating from scratch. FC layers of CNNs are changed with class number of task.

training, and the other group was used for testing (50%*testing*).). In 5-fold cross-validation, the data were randomly divided into five different groups. In the first evaluation, four groups were used for training, and one group was used for testing (20%*testing*). In the second evaluation, one group was used for training, and four groups were used for testing (80%*testing*). These operations were repeated five times, and averages of all performances were taken. Both easy and challenging situations have been evaluated for recognition and identification of all generated energy images.

This section is organized into three performance evaluation procedures. In the first part, the results obtained with the proposed method are compared with previous studies. In this way, the performance between the proposed method and the studies in the literature was revealed. In the second part, the existing and proposed TT functions were applied to the same networks. In this way, the performance of each TT function under the same conditions was compared, and the contribution of the proposed method was observed. In the third part, different handcrafted methods are used to understand whether the success in the first and second parts comes from the robustness of the proposed method or the power of the CNN classifiers. Thus, the performances of TT functions were compared with different feature extractors and classifiers. By these evaluations, the contribution of the proposed method was examined with a larger ablation.

3.1. Comparison with existing results

The results are compared with other state-of-the-art methods reported in the literature. Since the results presented in previous studies are usually 5-fold cross-validation (20% testing) results, comparisons were made according to these results. Table 2 shows the performance comparisons of the proposed approach with existing approaches that use skeletal and hand signal knowledge and RGB-D video streams. The proposed approach achieved 99.58% accuracy using the VGG-F network and 96.49% accuracy using the ResNet-18 network for motion recognition. Compared to current studies, very high success has been achieved with this method. The reasons for the high performance can be listed as follows: the feature extraction capability of the network used is high, the dataset features used are simpler and the depth data can focus on changes in the silhouette rather than complex RGB-D data, and the adaptive TT function we have proposed better emphasizes the motion changes. The comparisons in Table 2 contain different types of data. Some studies use skeleton and hand features, while others use RGB-D features. We also used the masked depth features in our study. Although different features are used for the current methods and the proposed approach, comparisons can be made between studies, as all data are obtained from the NATOPS dataset.

The performance comparison of the proposed method with previous studies on the SBU Kinect dataset is shown in 3. The results show that the proposed approach achieved 96.56% accuracy and gives better results than previous studies. Even though different types of data are extracted from the same dataset, they are comparable because all of them are taken from the same source. In Table 3, Khaire et al. [44] seem to have achieved slightly better results on the VGG-16 network. However, the method of combining CNN streams proposed in [44] requires training of more than one network and fusion of these networks, as in [25]. This inevitably increases the process costs. The method we propose involves lower process costs and achieves significantly more successful results in comparisons on SBU Kinect data.

Since there is no specific study on motion recognition and person identification with the BodyLogin dataset, no comparison could be made with previous studies. In the 2nd and 3rd comparison titles, it was used for the comparison between existing TTs and our proposed adaptive TT.

The proposed method yielded better results than previous studies. However, these findings raised the following questions. How much better was the adaptive TT's ability to emphasize movement compared to existing TTs, and what would be the result if existing TTs were applied to the same networks? If the answer to these questions can be examined, the success of the adaptive TT will be compared more fairly to other TTs. For this purpose, all templates (previously used $\mu_2 - \mu_4$ and proposed μ_R) were used to generate energy images, and comparisons were made between templates trained on the same CNN networks.

3.2. CNN-based comparison of TTs

While comparing all TT functions, datasets are organized not only for motion recognition but also for person recognition. These datasets have mostly been used for motion recognition in the literature. In our study, person recognition is studied with datasets, and the performance and comparisons of TT functions on person recog-

Table 2
The performance results of proposed and existing approaches on motion recognition using NATOPS dataset (in terms of classification accuracy in %). This table shows the accuracy rates obtained with the methods and properties used in previous studies. Although, different features were used for the existing methods and proposed approach, all data were obtained from NATOPS dataset.

Approaches	Using Features	Accuracy %
Song et al.[41]	Skeletal & hand	75.37
CRF, Song et al. [42]	Skeletal & hand	53.30
HMM, Song et al. [42]	Skeletal & hand	77.67
HCRF, Song et al. [42]	Skeletal & hand	78.00
Couples HCRF, Song et al. [42]	Skeletal & hand	86.00
Linked HCRF, Song et al. [42]	Skeletal & hand	87.00
TTs Fusion Top-1, Ijjina et al. [25]	RGB-D raw video streams	72.58
TTs Fusion Top-2, Ijjina et al. [25]	RGB-D raw video streams	86.58
2D-MRCN, Elbouhaki et al. [43]	RGB-D video streams	89.83
3D-CDCN, Elbouhaki et al. [43]	RGB-D video streams	94.54
MultiD-CNN, Elbouhaki et al. [43]	RGB-D video streams	95.87
ResNet-18, Proposed Approach	**Masked depth video streams**	96.49
VGG-F, Proposed Approach	**Masked depth video streams**	99.58

Table 3
The performance results of proposed and existing approaches on motion recognition using SBU dataset (in terms of classification accuracy in %). In this table, the performance of approaches and features used in previous studies are compared with the our proposed method. Although different features were used for the existing methods and proposed approach, all data were obtained from SBU Kinect dataset.

Approaches	Using Features	Accuracy%
Raw skeleton, Yun et al.[9]	Skeletal Data	49.70
Raw skeleton, Ji et al. [45]	Skeletal Data	79.40
Joint features, Yun et al. [9]	Skeletal Data	80.30
Joint features, Ji et al. [45]	Skeletal Data	86.90
Hierarchical RNN, Du et al. [46]	Skeletal Data	80.35
Cluster analysis of pose, Edwards et al. [47]	Skeletal Data	83.90
Deep LSTM, Zhu et al. [10]	Skeletal Data	86.03
Generative topic model, Huynh-The et al. [48]	Skeletal Data	90.30
STA-LSTM, Song et al. [49]	Skeletal Data	91.51
ST-LSTM + Trust Gate, Liu et al. [50]	Skeletal Data	93.30
Radius-margin bound, Lin et al. [51]	Skeletal Data	93.40
TTs Fusion of {begin and middle}, Ijjina et al. [25]	RGB-D Raw Video Stream	90.98
Multi-Task Learning Network, Ke et al. [13]	Concatenated Skeleton Sequences	93.57
Non-Linear Mapping (ConvNet + ELM classifier), Gharahdaghi et al.[52]	Depth Features	78.00
Non-Linear Mapping (ConvNet + ELM classifier), Gharahdaghi et al.[52]	RGB Features	81.50
Deeply Coupled ConvNet (Inception-v3 + LSTM), Singh et al. [15]	RGB Features	76.60
Deeply Coupled ConvNet (Inception-v3 + LSTM), Singh et al. [15]	Dynamic Motion Image Features	91.40
Deeply Coupled ConvNet (Inception-v3 + LSTM), Singh et al. [15]	RGB + DMI Max Fusion	98.70
Combining CNN streams (VGG-F), Khaire et al.[44]	RGB + Depth + Skeletal Data	96.26
Combining CNN streams (VGG-16), Khaire et al.[44]	RGB + Depth + Skeletal Data	96.67
Proposed Approach with ResNet-18	**Depth video stream**	81.25
Proposed Approach with VGG-F	**Depth video stream**	96.56

Table 4
A comparison of individual success rates of the proposed adaptive TT and existing TT functions on NATOPS *Full – Body* dataset using CNN networks. Motion recognition and person identification results for each CNN network are presented separately in the table. Results in bold show the best average score.

	VGG-F						ResNet-18					
Temporal Template	Motion Recognition			Person Identification			Motion Recognition			Person Identification		
Testing(%)	20%	50%	80%	20%	50%	80%	20%	50%	80%	20%	50%	80%
μ_2	**99.81**	**99.43**	97.97	99.99	100	99.96	96.26	93.95	**89.46**	100	99.96	99.92
μ_3	99.75	99.40	**98.02**	100	100	99.99	96.17	92.97	87.31	99.95	99.98	**99.96**
μ_4	99.59	99.30	97.56	100	99.99	99.88	95.23	91.16	86.01	99.95	99.90	99.71
μ_R	99.58	99.11	97.94	**100**	**100**	**100**	**96.49**	**94.10**	88.30	**100**	**100**	99.95

nition is discussed in a broad ablation study as in motion recognition. While the motions of all subjects are labelled identically for motion recognition, the motions of each subject are labelled individually for person identification. That is, samples of the first motion of all subjects have an identical label for motion recognition; on the other hand, the samples have individual labels according to the subject performing the action for correct identification.

The performance comparison of the proposed template and the existing templates in CNN classifiers for the two generated datasets are given in Table 4 and Table 5.. Existing templates achieved better classification with VGG-F for motion recognition. In the ResNet-18 network, although the average performance is generally lower than VGG-F, the proposed adaptive TT performed better than the other methods. Additionally, for person recognition, the proposed template outperformed existing templates on the NATOPS*Full – Body* dataset in both CNN networks. By considering the previous studies, it is apparent that no work has been conducted on person classification with the NATOPS database; thus, the proposed method could only be compared with existing templates.

When Table 5 is examined, it is seen that μ_2 in the VGG-F network and μ_3 in the ResNet-18 network are more successful for motion recognition on the NATOPS*Moving – Limbs* dataset. While adaptive μ_R has an average success for all TTs in both networks, μ_4 gives the worst result in motion recognition. In addition, it appears

Table 5
A comparison of individual success rates of the proposed adaptive TT and existing TT functions on NATOPS*Moving – Limbs* dataset using CNN networks. Motion recognition and person identification results for each CNN network are presented separately in the table. Results in bold show the best average score.

	VGG-F						ResNet-18					
Temporal Template	Motion Recognition			Person Identification			Motion Recognition			Person Identification		
Testing(%)	20%	50%	80%	20%	50%	80%	20%	50%	80%	20%	50%	80%
μ_2	**99.19**	**98.87**	**96.51**	98.80	98.06	94.84	91.76	88.66	88.70	**97.92**	95.98	91.33
μ_3	99.16	98.58	96.34	**98.96**	**98.13**	**95.40**	**92.24**	**88.73**	**89.11**	97.60	**96.77**	**91.88**
μ_4	97.99	97.06	93.79	97.72	96.16	90.68	83.57	81.15	81.97	93.13	89.83	78.54
μ_R	98.73	97.90	95.14	98.76	97.55	94.02	87.60	83.76	85.38	95.63	93.10	84.65

Table 6
A comparison of individual success rates of the existing TT's functions and proposed adaptive TT on SBU Kinect datasets. The first row of the table contains the results in the reference study. The other lines contain the comparison results made within the scope of this study.

				5-Fold Cross Val. Averg. Score			
Dataset	Using Features	Classifier	Testing Ratio	μ_2	μ_3	μ_4	μ_R
SBU in [25]	RGB-D data	ConvNet with ELM Classifier	20%	75.58	85.06	79.85	-
		VGG-F	20%	96.25	96.41	95.94	**96.56**
SBU *Full Body*	Masked Depth data	ResNet-18	20%	79.38	73.75	70.63	**81.25**
		VGG-F	20%	95.31	94.22	94.06	**95.94**
SBU *Moving Limbs*	Masked Depth data	ResNet-18	20%	71.88	68.75	67.50	**79.38**

that the results are similar for person recognition. Among the datasets generated from the NATOPS database and applied to CNN networks, the best result on the *Full – Body* dataset was obtained from the adaptive μ_R function. On the *Moving – Limbs* dataset, μ_R reached average results, and the best results were obtained from the μ_3 function.

In the NATOPS*Moving – Limbs* dataset, the body silhouette information was removed because people are moving their limbs while they are stationary. In existing TTs, especially in μ_2 and μ_3 functions, the emphasis on and average of the body silhouette in any sample are equal. That is, for a stationary body silhouette in a 100 frame video stream, the functions will obtain the same average. These two functions have a fixed length, independent of the motion change. This will reduce the discrimination feature of body silhouette information between functions.

However, in the μ_R function, the body silhouette will be highlighted in proportion to the number of moving frames. In this case, the silhouette will have values that vary according to the movement. The Moving-Limbs dataset was obtained by removing the body silhouette. When the results are examined in Table 5, μ_2 and μ_3 have obtained the best results, and these values are very close to each other. However, the performance of the μ_4 and μ_R functions decreased. This situation reveals that the μ_R function benefits more from silhouette information due to motion changes. The μ_R performance decreased with the removal of body silhouette information. Therefore, it can be said that μ_R also emphasizes the changes in the silhouette body and μ_R gives better results in silhouette-based depth images such as those in the *Full – Body* dataset.

In some samples in the SBU Kinect database, such as approaching and departing movements, one subject has no or small motions, while the other subject is more active. The positions of the subjects towards each other may change, but since the interactions remain the same, horizontally inverted interactions can also be used to evaluate the temporal template. Thus, the number of samples for motion recognition can be doubled. However, this situation restricts the use of the dataset for person identification. Therefore, the SBU Kinect dataset is only used for motion recognition.

Table 6 shows the comparison of the results obtained with existing membership functions with the proposed method. In Table 3,we present a comparison of the performance of the existing TTs and the adaptive TT using the SBU Kinect database. However, this table contains the results obtained by fusion of TTs. In Table 6, the individual performance of each TT is presented. The existing TT results presented in [25] are given in the first row. In the next lines, the energy images obtained with all TTs were applied to CNN networks with the same parameters. Thus, a comparison was made between adaptive TT and existing TT functions under equal conditions. While the performance of all TTs varied depending on the CNN networks, the performance of the adaptive TT gave the best result in all comparisons. Among the classifiers, the best result was obtained with the VGG-F network.

In Table 7, the comparison of all TTs using 2 different datasets obtained from the SBU Kinect database is presented. Each test group was trained randomly five times with the same training and test sequence for each function, and their average performance was compared. The proposed method generally yields better results, but in the challenging group that was tested with 80% of the data, the best result was obtained with μ_2. Although the fusion of existing functions has proven to be more successful, the usage of more than one function for motion recognition will be of long duration and have process load problems. The proposed approach provides a solution to these problems. The last column of Table 7 consists of *N/A* entries. These entries indicate that the results obtained are not applicable for evaluation. For this reason, values for the last column are not included in the table.

In contrast to the NATOPS*Moving – Limbs* dataset results, the adaptive TT yielded the best results in the SBU*Moving – Limbs* dataset. This is because the body silhouettes mentioned previously are not completely lost in the SBU*Moving – Limbs* dataset. The SBU Kinect database, unlike the NATOPS, contains interaction movements between two persons, so the bodies of the persons are also in motion. Therefore, the body silhouettes have not disappeared completely and affect the accuracy performance of the adaptive TT.

The VGG-F and ResNet-18 classifiers were trained for the BodyLogin dataset as in the previous datasets. Using the generated energy images, 20% testing, 50% testing and 80% testing operations were applied for recognition and identification to evaluate both easy and challenging situations. The training was repeated five times, and the average scores are presented in Table 8.

BodyLogin*Full – Body* dataset was used for person identification. In Table 8, it is seen that the proposed function in both CNN networks achieves the best performance in 20% and 50% testing operations. In the 80% testing operation, the μ_2 function in the VGG-F network and the μ_4 function in the ResNet-18 network achieved the best results.

Table 7
A comparison of individual success rates of the proposed adaptive TT and existing TT functions on SBU$Full-Body$ and SBU$Moving-Limbs$ datasets using CNN networks. Motion recognition results for each dataset and CNN network are presented in the table respectively. Results in bold show the best average score.

	VGG-F						ResNet-18					
Temporal Template	**SBU *Full Body***			**SBU *Moving Limbs***			**SBU *Full Body***			**SBU *Moving Limbs***		
Testing(%)	20%	50%	80%	20%	50%	80%	20%	50%	80%	20%	50%	80%
μ_2	96.25	85.25	**65.31**	95.31	**79.25**	**58.28**	79.38	67.75	49.06	71.88	**66.25**	N/A
μ_3	96.41	85.50	65.16	94.22	70.50	53.44	73.75	65.25	45.78	68.75	55.75	N/A
μ_4	95.94	75.50	60.78	94.06	74.25	55.31	70.63	63.25	49.53	67.50	58.75	N/A
μ_R	**96.56**	**86.00**	64.69	**95.94**	**79.25**	57.66	**81.25**	**69.00**	**54.69**	**79.38**	**66.25**	N/A

Table 8
A comparison of individual success rates of the proposed adaptive TT and existing TT functions on BodyLogin datasets using CNN networks. Motion recognition and person identification results are presented in the table for each CNN network, respectively. Results in bold show the best average score.

	VGG-F						ResNet-18					
Temporal Template	Person Identification BodyLogin *Full body*			Motion Recognition BLD-SS *User-Defined*			Person Identification BodyLogin *Full body*			Motion Recognition BLD-SS *User-Defined*		
Testing(%)	20%	50%	80%	20%	50%	80%	20%	50%	80%	20%	50%	80%
μ_2	99.63	98.75	**94.25**	98.87	**97.29**	85.74	97.38	95.90	85.97	94.25	87.63	**68.22**
μ_3	99.00	98.05	91.63	98.41	96.02	83.52	96.88	94.40	81.19	91.63	85.45	64.67
μ_4	99.25	98.55	94.00	98.72	96.79	85.20	**98.13**	95.66	**86.22**	93.63	86.16	66.27
μ_R	**99.63**	**98.80**	93.81	**99.12**	97.28	**86.13**	**98.13**	**96.05**	85.72	**94.88**	**88.28**	67.99

When the results of the BLD-SS $User-Defined$ dataset used for motion recognition are examined, it is seen that the proposed method is more successful than other functions. The μ_R function achieved the best performances in 20% and 80% testing operations on the VGG-F network and 20% and 50% testing operations on the ResNet-18 network. The μ_2 function achieved the best results in the 50% testing operation using the VGG-F network and 80% testing operation using the ResNet-18 network on the BLD-SS $User-Defined$ dataset.

In addition to these tables, the validation loss curves of the training in the ResNet-18 network are presented in Fig. 13 and Fig. 14. To better observe the differences between the curves, the last parts are enlarged and shown in the subfigures. The iterations of some training runs in the graphs are different. Since the ratio of the number of images per class to batch size is higher in some training runs, the number of iterations is higher. Although the number of classes for a dataset is higher, due to the small number of samples per class and the training runs with the same batch size, different iteration values were emerged.

According to the results obtained from the BodyLogin datasets, the proposed adaptive TT showed higher performance among all TTs trained in two different CNN networks. This proved that the μ_R function is better at representing movements.

Although these results show the success of μ_R, another question is the effect of CNN networks on the performance of the μ_R function. In addition to these results, using handcrafted feature extractors and classifiers, we analysed whether the success achieved depends on the generalization capability of the CNN network or the proposed method.

3.3. Handcrafted-based comparison

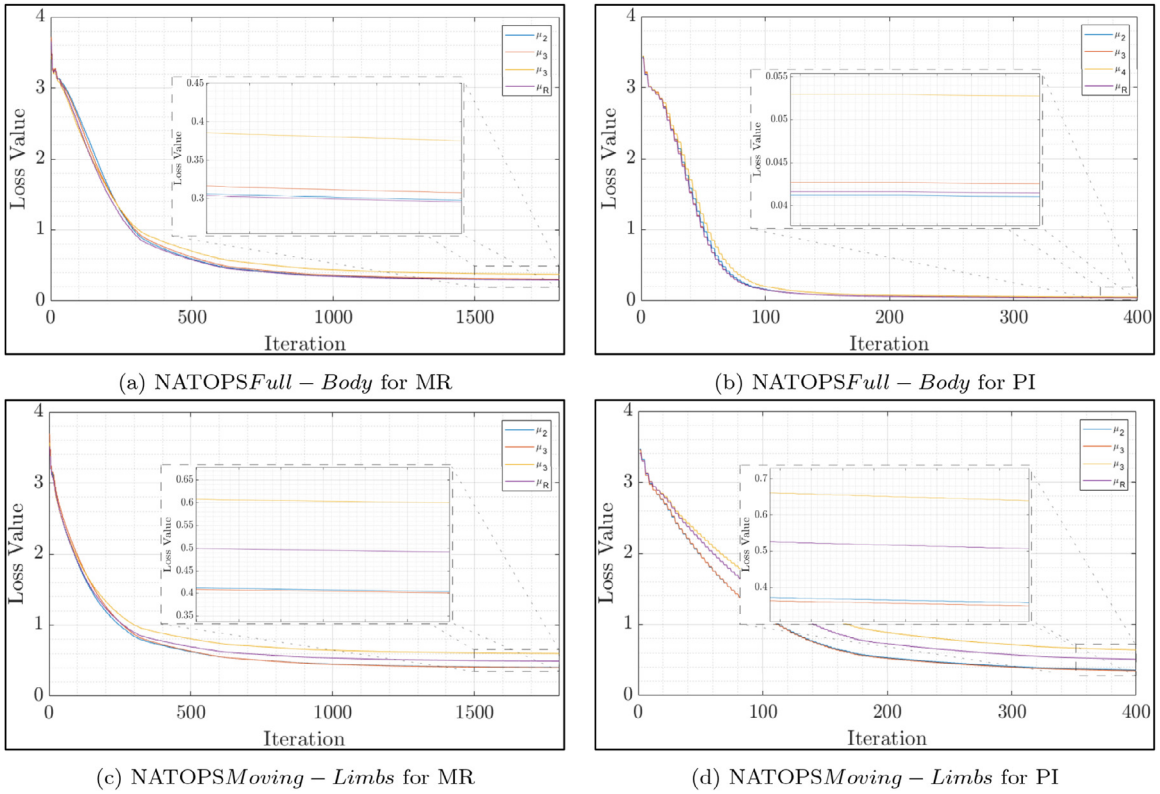
To perform this analysis and observe the effectiveness of the proposed μ_R function, the features of energy images generated in this work were extracted with principal component analysis (PCA), local binary patterns (LBP) [53], and histogram of gradient (HoG) methods [54] and then performed with classifiers such as deep neural network-based multilayer perceptron, k-nearest neighbours (k-NN), and SVM. Analogous to the VGG-F network, these classifiers were trained with 20%, 50%, and 80% testing to evaluate both easy and challenging situations. Similarly, this procedure was repeated five times, and the average performance scores were obtained.

As a classifier, a modified MLP style Deep Neural Network (DNN) [55] model was also used. It is developed by using innovative layers such as batch normalization (BNM) [56] and rectified linear unit (ReLU) within the standard MLP structure. The architecture of the network classifying the LBP and HoG vectors is given as $v_i \rightarrow FC_1 \rightarrow BNM \rightarrow ReLU \rightarrow FC_2 \rightarrow Softmax$. $v_i \in \Re^{1\times1\times N_f}$ is the extracted feature vector where N_f is the length of vector. FC_1 and FC_2 are *fully-connected* layers which have $2 \times N_f$ and N_c neurons. 100 epochs have been selected for the training of the network and the minibatch is defined as 1/5 of the dataset used. Adam optimizer [57] was used in the training phase.

The lengths of the feature vectors obtained in these three methods are quite high. Hence, the vectors were reduced to C-1 using PCA and linear discriminant analysis (LDA). The obtained features were classified with variations of DNN, kernel SVM, and k-NN algorithms and the effects of the proposed algorithm were compared with previous functions.

As with the CNN networks, these classifiers were trained with 20%, 50%, and 80% testing to evaluate both easy and challenging situations. Similarly, this procedure was repeated five times, and the average performance scores were obtained. Showing all experimental results obtained by handcrafted methods in a single table makes it difficult to understand. For this reason, only 20% of the results are included in the tables. Comments on the all results are mentioned here. However, 50% and 80% testing results are presented in the appendix section at the end of the article. The results that are more detailed can be observed from tables in the appendix.

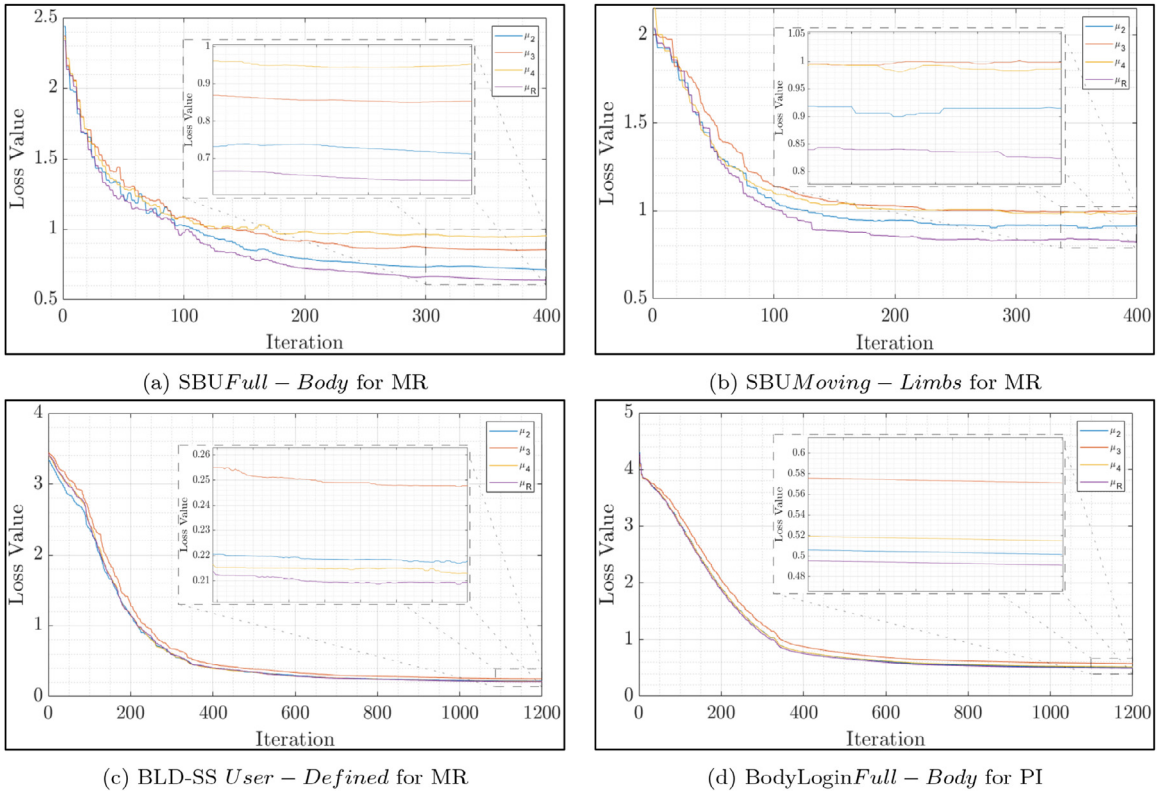
When the motion recognition results are examined in Table 9, μ_2 has a better result for PCA and HOG features in the 20% testing process for NATOPS$Full-Body$ data, and μ_3 has the best result for LBP features in the 20% testing process. However, when the results in Table A.13, which includes all test procedures, are observed, it is seen that the μ_R function performs better than other methods with PCA features. In the same table, the μ_2 function reached the best results in the average of all test processes with HOG features and the μ_3 function with LBP features. While motion recognition results obtained from handcrafted classifiers are similar to ResNet-18 results for PCA features, the results obtained with LBP and HOG



(a) NATOPS$Full-Body$ for MR (b) NATOPS$Full-Body$ for PI

(c) NATOPS$Moving-Limbs$ for MR (d) NATOPS$Moving-Limbs$ for PI

Fig. 13. The Validation-Loss plots of ResNet-18 network. The Validation-Loss changes of the μ_2, μ_3, μ_4 and μ_R functions are compared, respectively. In addition to the Loss curves, the differences in the last iterations are enlarged and shown more clearly in Sub-figures. a) NATOPS*Full – Body* for Motion Recognition. b) NATOPS*Full – Body* for Person Identification. c) NATOPS*Moving – Limbs* for MR. d) NATOPS*Moving – Limbs* for PI.



(a) SBU$Full-Body$ for MR (b) SBU$Moving-Limbs$ for MR

(c) BLD-SS $User-Defined$ for MR (d) BodyLogin$Full-Body$ for PI

Fig. 14. The Validation-Loss plots of ResNet-18 network. The Validation-Loss changes of the μ_2, μ_3, μ_4 and μ_R functions are compared, respectively. In addition to the Loss curves, the differences in the last iterations are enlarged and shown more clearly in Sub-figures. a) SBU*Full – Body* for Motion Recognition. b) SBU*Moving – Limbs* for Motion Recognition. c) BLD-SS *User – Defined* for MR. d) BodyLogin*Full – Body* for PI.

Table 9
The table refers the Motion Recognition performances of membership function on NATOPS*Full – Body* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

NATOPS Motion Recognition	Full Body Dataset - 20% Testing Results							Moving Limbs Dataset - 20% Testing Results						
	DNN	SVM			k-NN			DNN	SVM			k-NN		
		RBF	LIN	POL	3	5	7		RBF	LIN	POL	3	5	7
Features				PCA							PCA			
μ_2	94.59	99.24	98.54	99.13	97.84	97.61	97.04	93.95	94.90	94.60	94.64	95.27	95.21	94.98
μ_3	94.10	99.39	98.34	99.26	97.97	97.29	96.82	93.73	94.83	94.60	94.51	95.10	95.18	94.96
μ_4	92.36	99.00	97.92	98.79	97.33	96.68	96.31	83.52	85.32	84.11	85.20	86.31	86.29	86.09
μ_R	95.55	99.06	98.39	98.79	97.94	97.52	97.02	87.54	88.95	88.28	88.78	90.77	90.33	89.94
Features				LBP							LBP			
μ_2	94.28	95.53	93.73	93.18	96.36	96.27	96.15	92.31	94.14	91.64	92.31	94.81	94.70	94.44
μ_3	94.35	95.59	94.43	92.78	96.50	96.25	95.94	93.00	94.47	91.84	93.23	95.22	95.07	94.89
μ_4	94.19	95.35	93.75	93.49	96.09	95.89	95.72	90.50	91.69	88.92	90.26	93.21	92.96	92.49
μ_R	94.03	95.53	94.27	92.02	96.17	96.1	95.98	90.54	92.21	89.14	90.47	93.47	93.49	92.94
Features				HOG							HOG			
μ_2	98.72	98.68	98.28	98.34	98.81	98.79	98.73	97.56	97.82	97.88	97.78	98.01	97.99	97.91
μ_3	98.67	98.48	98.21	98.50	98.65	98.73	98.63	97.13	97.40	97.28	97.39	97.36	97.51	97.35
μ_4	98.09	98.44	98.27	98.07	98.73	98.65	98.61	93.89	94.57	94.35	94.36	95.01	94.89	94.86
μ_R	98.13	98.14	97.75	98.07	98.23	98.17	98.09	94.88	95.77	95.65	95.48	95.68	95.66	95.46

Table 10
The table refers the Person Identification performances of membership function on NATOPS*Full – Body* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

NATOPS Person Identification	Full Body Dataset - 20% Testing Results							Moving Limbs Dataset - 20% Testing Results						
	DNN	SVM			k-NN			DNN	SVM			k-NN		
		RBF	LIN	POL	3	5	7		RBF	LIN	POL	3	5	7
Features				PCA							PCA			
μ_2	100	100	100	100	100	100	100	99.69	99.53	97.34	96.72	99.53	99.53	99.53
μ_3	100	100	100	100	100	100	100	99.53	99.69	98.91	98.75	99.84	99.84	99.84
μ_4	100	100	100	100	100	100	100	99.69	98.44	97.97	96.72	98.59	98.59	98.59
μ_R	100	100	100	100	100	100	100	100	99.38	99.38	98.91	99.69	99.53	99.69
Features				LBP							LBP			
μ_2	100	100	99.38	99.84	100	100	100	99.22	99.69	98.75	98.91	100	99.84	99.69
μ_3	99.84	99.69	99.69	99.69	99.69	99.69	99.69	99.06	99.22	98.28	98.59	99.22	99.22	99.22
μ_4	99.84	99.84	99.38	99.53	99.84	99.84	99.84	99.69	99.53	98.28	98.28	99.69	99.53	99.69
μ_R	99.53	99.69	98.59	98.75	99.69	99.69	99.69	99.84	99.69	97.34	98.13	100	99.69	99.84
Features				HOG							HOG			
μ_2	100	100	100	100	100	100	100	99.69	99.69	99.53	99.53	99.69	99.69	99.69
μ_3	100	100	99.69	100	100	100	100	99.22	99.06	98.13	98.28	99.22	99.22	99.22
μ_4	99.84	99.84	99.84	99.84	100	100	100	99.22	98.75	98.13	97.50	99.22	99.22	99.22
μ_R	100	100	100	100	100	100	100	99.84	99.84	99.38	99.53	99.84	99.84	99.84

features are similar to VGG-F network results. It can be observed that feature selection methods are very effective on this dataset.

On the right side of Table 9, the results of the NATOPS*Moving – Limbs* dataset are presented. When this part of the table and all test results in Table A.14 are examined, it is seen that the μ_2 and μ_3 functions have the best results and μ_4 has the worst. As explained before, because there is no body silhouette in the *Moving – Limbs* dataset, μ_2 and μ_3 achieved very close performance, and μ_R was negatively affected by this deficiency. While the feature extraction methods had an effect on the performance of the *Moving – Limbs* dataset, they had no effect on the ordering of TT functions. The results and rankings in this dataset are similar to those obtained from CNN networks. This shows that the motion representation of functions is not affected much by the classifiers.

When the results of personal identification using the NATOPS*Full – Body* dataset in Table 10 are examined, it was observed that the features extracted with PCA achieved very good results in all classifiers. In the expanded results in Table A.15, all functions achieved similar performance with the PCA feature in handcrafted classifiers.

In Table 10 and Table A.15, the adaptive TT function achieved lower performance in all handcrafted classifiers using LBP features. However, in the classification results using HoG features, the adaptive TT function has significantly better and robust results.

Although the results of the HOG features in Table 10 seem similar for all functions, when all testing results in Table A.15 are examined, the μ_R function has reached more robust results in challenging situations. The results obtained with HoG and PCA features are similar to the results obtained in CNN networks. The μ_2 function was more successful in the results obtained with LBP features.

According to these results, all TT functions in handcrafted classifiers showed similar performance for person recognition on the NATOPS*Full – Body* dataset. In addition, adaptive TT achieved more robust results in challenging situations. The μ_2 function showed better representation ability in LBP features in person identification. When interpreted in general, it has been observed that the adaptive TT function achieves better results using handcrafted classifiers on the NATOPS*Full – Body* Dataset, although not as much as in CNN networks.

As shown on the left side of Table 10, feature extraction methods have an effect on the results of the NATOPS*Moving – Limbs* dataset on the right side of Table 10. The μ_R and μ_3 functions showed similar performances in the results obtained using the PCA features. In the results obtained using the LBP features, the μ_2

Table 11
The table refers the Motion Recognition performances of membership function on SBU*Full – Body* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

SBU Kinect Motion Recognition	Full Body Datase - 20% Testing Results							Moving Limbs Dataset - 20% Testing Results						
	DNN	SVM			k-NN			DNN	SVM			k-NN		
		RBF	LIN	POL	3	5	7		RBF	LIN	POL	3	5	7
Features				PCA							PCA			
μ_2	73.75	72.50	72.50	66.88	73.75	75.00	74.38	73.75	69.38	68.13	64.38	68.75	69.38	68.75
μ_3	65.63	66.88	63.13	65.00	66.88	68.13	67.50	70.00	71.25	69.38	66.25	68.75	69.38	70.63
μ_4	71.25	73.13	71.88	67.50	71.88	76.25	75.63	69.38	68.13	66.88	71.88	70.63	71.25	72.50
μ_R	77.50	70.63	72.50	67.50	78.13	78.13	77.50	74.38	70.63	71.88	68.75	73.13	74.38	76.25
Features				LBP							LBP			
μ_2	88.13	88.75	85.63	84.38	88.75	88.75	89.38	89.38	86.88	85.63	81.88	89.38	90.00	90.63
μ_3	86.25	84.38	78.13	77.50	87.50	87.50	86.25	85.00	85.00	81.25	80.00	86.88	86.25	86.88
μ_4	86.88	85.63	82.50	83.13	87.50	87.50	88.75	88.13	87.50	83.13	84.38	89.38	88.75	89.38
μ_R	85.63	88.75	85.00	85.63	90.00	90.63	90.00	82.50	88.13	83.75	84.38	88.75	89.38	90.00
Features				HOG							HOG			
μ_2	87.50	78.75	71.88	71.88	79.38	79.38	80.00	85.00	79.38	73.13	76.25	83.13	82.50	82.50
μ_3	80.63	81.88	78.13	78.13	84.38	83.75	84.38	85.00	80.00	73.75	71.25	82.50	82.50	83.13
μ_4	86.25	80.00	77.50	74.38	81.88	81.88	81.88	85.63	78.13	70.63	73.75	80.00	80.63	80.63
μ_R	87.50	81.88	74.38	76.25	84.38	84.38	84.38	88.13	84.38	78.13	80.00	86.25	86.25	86.25

function achieved better performance. In the results obtained using the HOG features, the μ_R function achieved the best success in almost all classifications.

When Table-A.16, which includes the expanded results of this dataset, is examined, using PCA and LBP features, the μ_2 function has obtained better results for classifications. In almost all of the classification results using HoG features, the μ_R function achieved the best performance. It is observed that the performance of all functions varies considerably depending on the classifier and feature extraction methods. However, in overall performance, the adaptive TT function was less affected by this variation. This shows that the proposed μ_R function can represent the motion-energy images better than existing TT functions in classification.

When the results of SBU*Full – Body* dataset in Table 11 were investigated, the results show that the μ_R function is significantly better for motion recognition, as in the performance of CNN networks on this dataset. Although the performance scores are not as high as those of the CNN networks, the μ_R function achieved better results than the others and showed robustness with different classifier and feature extraction methods. The performances of other methods have more variable results according to classifier and feature extraction methods. The results obtained using 80% testing (i.e., 20%*training*) and some 50% testing rate scores in Table A.17 and Table A.18 were not included in the tables because they did not yield comparable results for some classifier and feature extractor pairs. Therefore, these fields are marked as not applicable (*N/A*).

In the right part of Table 11, it is seen that the performances of the μ_2, μ_3 and μ_4 functions are quite variable. Without empirical observations of these methods, it is not possible to reach a definite conclusion regarding which one is the best. This situation will cause an increase in process cost.

On the other hand, when the results of the μ_R function for the SBU*Moving – Limbs* dataset were examined, the first two best results were obtained with the adaptive TT function. This shows that the μ_R function both represents the movement better and provides a lower time cost. Additionally, the results in Table A.18 are similar to those obtained using the CNN networks in Table 7.

Classification results for motion recognition with the BodyLogin dataset are shown in Table 12. The μ_R function achieved better results than the other functions for all feature extractors. When Table A.19, containing the extended test results, is examined, the μ_R function achieved better results than the other functions for PCA and HoG features. The μ_2 function performed better in extended results in LBP features, while the μ_R function achieved the 2nd best success. As in previous datasets, the results obtained here are similar to those obtained with CNN networks in Table 8.

The last comparison was made with the results of BodyLogin person identification on the right side of Table 12. According to the results of person identification, the adaptive TT function, μ_R, achieved the best success in LBP and HOG properties in the 20% testing procedure and the second-best success with PCA features.

In our last extended table, Table A.20, it is seen that the μ_R method generally gives the most successful results for the first- and second-best results in the person identification process with the BodyLogin dataset in all classifier and feature extraction pairs.

When Table A.20 containing the extended test results for person identification is examined, the adaptive TT function generally gave the first or second most successful results in all classifier and feature extractor pairs. When the existing TT functions for person identification on the BodyLogin dataset were examined, the closest performance to the adaptive TT was obtained with μ_4. The μ_2 function usually yielded average results, and the worst result among all functions was obtained with μ_3.

4. Discussion

The proposed method was tested on the NATOPS, SBU Kinect, and BodyLogin datasets. Using these datasets, we had the opportunity to experiment on upper-body and full-body gesture recognition, person identification and two-person interaction. When all the results were examined, it was seen that the μ_R method, which is a membership function based on the change in the correlation coefficient with which we propose to create energy images for person identification and motion recognition processes, obtained better and generalizable results. In the Experiments and Results section, it was questioned whether the results obtained with the proposed method were due to the classification power of the CNN networks or the robustness of the proposed method. Therefore, the proposed function and the current membership functions have been tested on the machine learning structures obtained with different feature extraction methods, different networks and their different parametric combinations. Although different machine learning and feature extraction pairs give lower results compared to CNNs, CNNs and other networks give similar ranking results for the same datasets. It is clear that the feature extraction and clas-

Table 12
The table refers the Motion Recognition performances of membership function on BodyLogin BLD-SS *User – Defined* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

BodyLogin Datasets	Motion Recognition - 20% Testing Results							Person Identification - 20% Testing Results						
	DNN	SVM			k-NN			DNN	SVM			k-NN		
		RBF	LIN	POL	3	5	7		RBF	LIN	POL	3	5	7
Features	PCA							PCA						
μ_2	99.63	99.75	99.00	96.38	99.88	99.88	99.88	98.60	98.50	97.03	96.22	96.66	98.78	98.75
μ_3	99.13	98.75	98.50	98.25	98.87	99.00	98.87	98.22	98.10	96.47	95.63	98.38	98.41	98.41
μ_4	99.62	99.50	98.75	98.25	99.63	99.75	99.75	99.01	99.16	97.88	97.19	99.19	99.25	99.25
μ_R	99.63	99.63	99.13	97.50	100	99.88	99.75	98.63	98.82	97.66	96.88	98.97	98.97	99.10
Features	LBP							LBP						
μ_2	96.25	100	100	100	100	100	100	86.04	99.22	98.63	98.56	99.31	99.41	99.31
μ_3	97.38	100	99.88	99.88	100	100	100	85.63	99.13	98.78	98.75	99.16	99.19	99.22
μ_4	97.38	100	100	100	100	100	100	84.47	99.28	98.75	98.88	99.31	99.34	99.38
μ_R	96.88	100	100	100	100	100	100	84.82	99.28	98.53	98.88	99.41	99.38	99.38
Features	HOG							HOG						
μ_2	100	99.88	99.75	99.63	99.88	99.88	99.88	98.88	99.07	94.58	97.91	99.66	99.66	99.66
μ_3	99.62	99.88	99.63	99.38	100	99.88	99.88	98.75	98.88	95.65	98.28	99.56	99.59	99.59
μ_4	99.75	99.88	99.63	99.63	99.88	99.88	99.88	99.16	99.22	95.48	98.34	99.59	99.56	99.56
μ_R	100	99.88	99.63	99.63	99.88	99.88	99.88	98.94	99.32	95.65	98.35	99.72	99.72	99.72

sification abilities of CNNs are better than those of the others. However, obtaining similar ranking results using the same datasets in all classifiers shows that our proposed adaptive TT function is more robust than the other functions regardless of classification networks.

When the NATOPS dataset is examined, the actions are more prominent at the beginning and middle intervals of the motion. This may be why templates that highlight the beginning and middle areas yield better results than the template we proposed. However, not every dataset will contain actions that prominently include certain intervals, such as NATOPS. In the SBU Kinect and BodyLogin datasets, unlike the NATOPS datasets, actions continue from the beginning to the end of the motion. In such a case, templates that highlight certain intervals will be insufficient. s seen in the results obtained, the proposed method is more effective for motion recognition and person identification in SBU Kinect and BodyLogin datasets compared to previous studies and existing templates. Consequently, it can be said that our approach has the power to represent movements more effectively.

5. Conclusion

Machine learning models have become standard tools in science and industry due to their high performance for high-dimensional problems such as person recognition and motion recognition. For this reason, models and networks are presented ready for application in many areas. In practice, however, the impact of the input on the performance achieved is becoming increasingly important. To demonstrate this, we proposed a new approach to temporal templates used in energy image acquisition and constructed it according to the correlation coefficients of motion sequences. The correlation coefficients of the motion sequences were calculated, and the ranges of variation on the motion were determined quickly and simply with the proposed μ_3 function. In this way, unlike the previous functions, an energy image was created by considering the changes in the whole movement, not specific intervals of it. This approach increased the energy image's ability to represent better changes in motion. Our method has been tested under different machine learning and CNN models using publicly available datasets and has provided robust results.

We would like to emphasize that we are free to use either a trained and carefully tuned machine learning model of our own design or existing pretrained CNN models as a starting point of our method. This can be leveraged because the results we obtained have shown that our method provides better performance than previous functions regardless of classification networks and can better represent the movements. In this way, it is easy to use this function in different networks for researchers who want to use this method in the future.

Considering the existing approaches emphasizing specific intervals, their robustness and generalization capacity seem to be limited. Employing certain intervals will cause loss of discriminative information. However, the proposed method can be used with more robustness for real-time action recognition and person identification with a CNN model that can run in parallel on the graphics processing unit (GPU).

In summary, the contributions of our study are (1) suppressing motionless ending parts by adapting to the completion time of movements, (2) providing functions that can identify and highlight areas of changes in movement, and (3) unlike in previous studies, preventing time loss by avoiding empirical methods to determine which highlighted TT is more appropriate for which action.

In our future studies, our aim is to test the proposed method with larger and more diverse datasets, including both lower-body and upper-body movements and more people, and to compare it with different current methods. In addition, we plan to work on the fusion of energy images obtained with the proposed function with different biometric data.

Declaration of Competing Interest

All authors are fully involved in the study and preparation of the manuscript and that the material within has not been and will not be submitted for publication elsewhere.

There are no other contributors and funding sources that need to be mentioned.

Acknowledgement

The authors would like to acknowledge that this paper is submitted in partial fulfilment of the requirements for Ph.D. degree at Yildiz Technical University.

Appendix A. Extended Tables of Hand Crafted Based Comparison

Table A1
The table refers the Motion Recognition performances of membership function on NATOPS*Full – Body* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

NATOPS Motion Recog. Full Body	DNN			SVM									k-NN								
				RBF			LIN			POL			3			5			7		
	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test
Features									***PCA***												
U2	94.59	92.71	88.02	99.24	98.81	94.42	98.54	97.48	90.7	99.13	98.29	93.04	97.84	96.85	85.02	97.61	95.98	75.63	97.04	94.90	64.6
U3	94.10	91.31	86.84	99.39	98.98	94.89	98.34	97.36	90.67	99.26	98.52	93.56	97.97	96.82	85.31	97.29	95.92	75.44	96.82	94.80	64.67
U4	92.36	91.24	86.05	99.00	98.43	93.92	97.92	96.59	89.69	98.79	97.99	92.08	97.33	96.01	84.13	96.68	94.86	75.14	96.31	93.58	65.11
UR	95.55	93.07	88.59	99.06	98.58	94.51	98.39	97.50	92.26	98.79	98.11	92.85	97.94	97.09	86.32	97.52	96.23	76.27	97.02	95.17	65.30
Features									***LBP***												
U2	94.28	93.36	88.22	95.53	94.58	89.98	93.73	93.03	88.52	93.18	92.50	88.25	96.36	95.52	90.84	96.27	95.28	90.52	96.15	95.19	90.02
U3	94.35	93.85	88.66	95.59	94.89	90.35	94.43	93.70	89.03	92.78	92.67	88.48	96.50	95.59	91.46	96.25	95.47	91.01	95.94	95.30	90.64
U4	94.19	92.94	87.34	95.35	94.36	89.24	93.75	92.82	87.81	93.49	92.28	87.34	96.09	94.94	89.92	95.89	94.95	89.49	95.72	94.60	89.17
UR	94.03	93.20	87.98	95.53	94.41	89.78	94.27	93.25	88.49	92.02	92.55	88.19	96.17	95.35	90.64	96.10	95.14	90.38	95.98	94.95	89.84
Features									***HOG***												
U2	98.72	98.29	95.25	98.68	98.44	96.96	98.28	98.05	96.15	98.34	98.06	95.99	98.81	98.68	97.22	98.79	98.63	97.13	98.73	98.54	97.09
U3	98.67	98.13	94.74	98.48	98.12	96.49	98.21	97.92	95.82	98.50	97.86	95.51	98.65	98.42	96.86	98.73	98.29	96.67	98.63	98.23	96.52
U4	98.09	97.84	94.31	98.44	97.85	96.02	98.27	97.49	95.09	98.07	97.45	94.94	98.73	98.23	96.37	98.65	98.10	96.26	98.61	98.05	96.20
UR	98.13	97.43	94.09	98.14	97.78	96.15	97.75	97.48	95.30	98.07	97.50	95.40	98.23	98.05	96.34	98.17	97.94	96.28	98.09	97.87	96.29

Table A2
The table refers the Motion Recognition performances of membership function on NATOPS*Moving – Limbs* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

NATOPS Motion Recog. Moving Limbs	DNN			SVM									k-NN								
				RBF			LIN			POL			3			5			7		
	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test
Features									***PCA***												
U2	93.95	91.75	83.86	94.90	93.15	86.16	94.60	92.23	83.88	94.64	92.10	83.51	95.27	93.58	86.70	95.21	93.39	86.42	94.98	93.23	86.32
U3	93.73	91.48	83.84	94.83	93.09	86.33	94.60	92.20	84.05	94.51	92.27	83.29	95.10	93.46	86.88	95.18	93.43	86.75	94.96	93.32	86.63
U4	83.52	80.09	71.07	85.32	82.32	75.13	84.11	80.94	72.40	85.20	81.35	72.57	86.31	82.95	75.41	86.29	83.01	75.35	86.09	82.82	75.11
UR	87.54	84.11	75.18	88.95	86.4	78.99	88.28	85.46	76.78	88.78	85.64	76.08	90.77	87.34	79.7	90.33	87.28	79.62	89.94	86.93	79.32
LBP									***LBP***												
U2	92.31	91.49	86.69	94.14	92.74	88.57	91.64	90.73	86.92	92.31	91.39	86.7	94.81	93.83	89.13	94.70	93.55	88.51	94.44	93.12	87.75
U3	93.00	91.79	86.92	94.47	92.93	88.7	91.84	91.01	86.95	93.23	91.73	86.97	95.22	94.17	89.33	95.07	93.81	88.7	94.89	93.61	88.03
U4	90.50	89.08	83.28	91.69	90.47	85.36	88.92	87.86	83.17	90.26	88.75	82.34	93.21	91.62	85.82	92.96	91.53	85.32	92.49	90.91	84.78
UR	90.54	89.11	83.84	92.21	90.84	86	89.14	88.68	84.1	90.47	89.29	83.89	93.47	91.96	86.71	93.49	91.78	86.08	92.94	91.34	85.51
HOG									***HOG***												
U2	97.56	96.77	93.05	97.82	97.41	93.86	97.88	97.16	93.09	97.78	96.89	92.72	98.01	97.62	94.22	97.99	97.57	94.07	97.91	97.44	93.97
U3	97.13	96.22	92.07	97.40	96.75	93.26	97.28	96.45	92.27	97.39	96.18	91.97	97.36	97.03	93.59	97.51	96.95	93.45	97.35	96.91	93.31
U4	93.89	92.25	86.65	94.57	93.67	88.85	94.35	93.01	87.34	94.36	92.75	87	95.01	93.57	88.83	94.89	93.86	88.99	94.86	93.73	88.92
UR	94.88	93.82	89.04	95.77	94.96	90.63	95.65	94.59	89.48	95.48	94.43	88.85	95.68	95.1	90.78	95.66	95.11	90.85	95.46	95.06	90.73

Table A3
The table refers the Person Identification performances of membership function on NATOPS*Full – Body* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

NATOPS Person Ident. Full Body	DNN			SVM RBF			SVM LIN			SVM POL			k-NN 3			k-NN 5			k-NN 7		
	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test
Features									***PCA***												
U2	100	100	100	100	100	99.97	100	100	99.59	100	100	99.53	100	100	100	100	100	100	100	100	100
U3	100	100	100	100	100	100	100	100	99.78	100	100	99.66	100	100	100	100	100	100	100	100	100
U4	100	100	100	100	100	100	100	100	99.91	100	100	99.81	100	100	100	100	100	100	100	100	100
UR	100	100	100	100	100	100	100	100	99.69	100	100	99.72	100	100	100	100	100	100	100	100	100
LBP									***LBP***												
U2	100	99.90	99.44	100	100	98.31	99.38	99.10	92.94	99.84	99.70	92.31	100	100	98.84	100	100	98.84	100	100	98.97
U3	99.84	99.80	98.81	99.69	99.65	97.19	99.69	99.25	88.84	99.69	99.30	87.13	99.69	99.60	98.22	99.69	99.60	98.28	99.69	99.60	98.34
U4	99.84	99.60	99.34	99.84	99.70	97.38	99.38	98.60	89.28	99.53	98.50	86.88	99.84	99.75	98.13	99.84	99.70	98.03	99.84	99.65	98.03
UR	99.53	99.70	97.22	99.69	99.55	91.00	98.59	96.75	78.34	98.75	97.50	75.84	99.69	99.75	92.75	99.69	99.80	92.72	99.69	99.75	92.97
HOG									***HOG***												
U2	100	100	99.72	100	100	91.88	100	99.80	86.37	100	99.85	82.84	100	100	94.5	100	100	94.59	100	100	94.59
U3	100	100	99.47	100	99.90	89	99.69	99.60	86.88	100	99.55	85.81	100	99.95	90.5	100	99.95	90.5	100	99.95	90.50
U4	99.84	99.85	97.91	99.84	99.80	78.88	99.84	99.65	76.25	99.84	99.75	75.84	100	99.85	80.38	100	99.85	80.38	100	99.85	80.38
UR	100	100	99.91	100	100	99.84	100	99.60	99.75	100	99.50	99.69	100	100	99.84	100	100	99.84	100	100	99.84

Table A4
The table refers the Person Identification performances of membership function on NATOPS*Moving – Limbs* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

NATOPS Person Ident. Moving Limbs	DNN			SVM RBF			SVM LIN			SVM POL			k-NN 3			k-NN 5			k-NN 7		
	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test
Features									***PCA***												
U2	99.69	99.70	97.09	99.53	99.50	90.72	97.34	97.95	83.47	96.72	96.90	79.91	99.53	99.60	93.13	99.53	99.60	93.31	99.53	99.60	93.28
U3	99.53	99.25	97.84	99.69	99.20	86.84	98.91	98.00	78.00	98.75	97.20	73.97	99.84	99.25	89.13	99.84	99.25	89.22	99.84	99.25	89.28
U4	99.69	99.25	95.44	98.44	98.30	64.91	97.97	95.70	54.75	96.72	94.00	50.84	98.59	98.70	69.31	98.54	98.75	69.28	98.59	98.70	69.28
UR	100	99.80	98.87	99.38	99.35	80.81	99.38	97.40	73.13	98.91	95.85	68.56	99.69	99.50	83.44	99.53	99.50	83.41	99.69	99.50	83.59
LBP									***LBP***												
U2	99.22	98.95	95.69	99.69	99.20	92.41	98.75	97.00	85.00	98.91	97.75	83.94	100	99.40	94.34	99.84	99.35	94.63	99.69	99.30	94.38
U3	99.06	99.00	96.94	99.22	99.30	92.66	98.28	97.10	83.69	98.59	97.05	81.50	99.22	99.25	94.31	99.22	99.25	94.28	99.22	99.15	94.28
U4	99.69	99.40	96.81	99.53	99.25	94.28	98.28	97.45	83.91	98.28	97.35	83.75	99.69	99.55	95.75	99.53	99.60	95.78	99.69	99.55	95.91
UR	99.84	99.25	96.81	99.69	99.65	89.69	97.34	97.10	77.75	98.13	96.90	74.16	100	99.65	91.81	99.69	99.65	91.78	99.84	99.70	91.88
HOG									***HOG***												
U2	99.69	99.30	94.09	99.69	99.00	85.50	99.53	98.45	85.00	99.53	98.25	83.84	99.69	99.25	86.06	99.69	99.30	86.03	99.69	99.30	86.13
U3	99.22	98.95	97.81	99.06	98.85	97.88	98.13	96.85	97.09	98.28	97.10	96.91	99.22	99.05	98.03	99.22	99.00	98.00	99.22	99.05	98.09
U4	99.22	98.95	95.03	98.75	99.05	95.66	98.13	96.25	94.44	97.50	96.30	93.66	99.22	99.20	95.87	99.22	99.20	95.91	99.22	99.20	95.97
UR	99.84	99.75	98.06	99.84	99.15	98.47	99.38	97.60	98.06	99.53	97.30	97.47	99.84	99.35	98.75	99.84	99.40	98.78	99.84	99.35	98.81

Table A5
The table refers the Motion Recognition performances of membership function on SBU KINECT*Full – Body* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

SBU Motion Recog. Full Body	DNN			SVM									k-NN								
				RBF			LIN			POL			3			5			7		
	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test
Features									***PCA***												
U2	73.75	65.25	N/A	72.50	61.00	N/A	72.50	64.00	N/A	66.88	55.75	N/A	73.75	65.25	N/A	75.00	64.75	N/A	74.38	64.75	N/A
U3	65.63	59.25	N/A	66.88	58.50	N/A	63.13	55.88	N/A	65.00	52.50	N/A	66.88	60.25	N/A	68.13	59.50	N/A	67.50	59.75	N/A
U4	71.25	64.25	N/A	73.13	57.50	N/A	71.88	61.00	N/A	67.50	52.50	N/A	71.88	59.25	N/A	76.25	62.25	N/A	75.63	62.75	N/A
UR	77.50	66.75	N/A	70.63	63.00	N/A	72.50	64.38	N/A	67.50	55.50	N/A	78.13	64.75	N/A	78.13	65.50	N/A	77.50	65.00	N/A
LBP									***LBP***												
U2	88.13	77.25	61.88	88.75	79.50	70.78	85.63	75.00	61.56	84.38	75.75	70.00	88.75	81.25	73.75	88.75	81.00	74.06	89.38	81.75	74.84
U3	86.25	76.25	61.25	84.38	71.00	61.56	78.13	67.25	59.53	77.50	69.00	56.09	87.50	72.50	61.88	87.50	81.00	62.66	86.25	71.75	61.72
U4	86.88	76.75	62.81	85.63	78.00	71.56	82.50	73.50	62.66	83.13	74.00	70.00	87.50	81.25	74.06	87.50	80.75	76.25	88.75	82.25	75.94
UR	85.63	72.25	61.25	88.75	81.00	72.66	85.00	76.25	63.28	85.63	78.50	71.25	90.00	83.75	75.16	90.63	85.00	75.94	90.00	85.00	77.03
HOG									***HOG***												
U2	87.50	81.00	72.50	78.75	62.00	72.97	71.88	61.75	70.94	71.88	61.75	67.34	79.38	62.50	73.59	79.38	63.75	73.44	80.00	63.75	72.97
U3	80.63	73.75	68.44	81.88	48.25	66.72	78.13	47.25	64.69	78.13	45.75	63.28	84.38	48.50	68.44	83.75	48.50	69.53	84.38	48.25	69.53
U4	86.25	80.75	71.88	80.00	70.00	71.56	77.50	64.50	69.53	74.38	65.50	66.09	81.88	70.50	72.34	81.88	70.50	72.50	81.88	70.50	72.19
UR	87.50	77.25	72.34	81.88	63.00	73.59	74.38	61.75	73.59	76.25	60.50	72.66	84.38	64.50	74.69	84.38	63.50	74.53	84.38	63.25	74.38

Table A6
The table refers the Motion Recognition performances of membership function on SBU KINECT*Moving – Limbs* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

SBU Motion Recog. Moving Limbs	DNN			SVM									k-NN								
				RBF			LIN			POL			3			5			7		
	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test
Features									***PCA***												
U2	73.75	N/A	N/A	69.38	N/A	N/A	68.13	N/A	N/A	64.38	N/A	N/A	68.75	N/A	N/A	69.38	N/A	N/A	68.75	N/A	N/A
U3	70.00	N/A	N/A	71.25	N/A	N/A	69.38	N/A	N/A	66.25	N/A	N/A	68.75	N/A	N/A	69.38	N/A	N/A	70.63	N/A	N/A
U4	69.38	N/A	N/A	68.13	N/A	N/A	66.88	N/A	N/A	71.88	N/A	N/A	70.63	N/A	N/A	71.25	N/A	N/A	72.50	N/A	N/A
UR	74.38	N/A	N/A	70.63	N/A	N/A	71.88	N/A	N/A	68.75	N/A	N/A	73.13	N/A	N/A	74.38	N/A	N/A	76.25	N/A	N/A
LBP									***LBP***												
U2	89.38	72.00	N/A	86.88	81.00	69.38	85.63	73.25	N/A	81.88	77.50	67.34	89.38	81.75	70.63	90.00	72.50	N/A	90.63	82.50	70.47
U3	85.00	75.25	N/A	85.00	82.00	71.09	81.25	74.50	N/A	80.00	75.75	68.59	86.88	84.25	71.72	86.25	72.00	N/A	86.88	84.25	72.34
U4	88.13	73.50	N/A	87.50	80.25	70.47	83.13	73.75	N/A	84.38	73.75	67.5	89.38	83.75	70.63	88.75	73.50	N/A	89.38	84.50	70.78
UR	82.50	73.25	N/A	88.13	82.50	69.53	83.75	74.25	N/A	84.38	76.75	67.66	88.75	83.25	70.63	89.38	72.25	N/A	90.00	84.00	70.47
HOG									***HOG***												
U2	85.00	77.00	70.31	79.38	N/A	71.41	73.13	N/A	70.47	76.25	N/A	68.75	83.13	N/A	72.5	82.50	N/A	72.81	82.50	N/A	72.5
U3	85.00	73.75	69.53	80.00	67.75	67.81	73.75	64.25	63.91	71.25	63.50	62.81	82.50	69.00	69.06	82.50	69.00	68.59	83.13	69.00	68.13
U4	85.63	75.75	70.78	78.13	70.00	70.00	70.63	64.50	69.22	73.75	66.00	66.09	80.00	71.75	72.03	80.63	71.00	71.88	80.63	71.50	71.88
UR	88.13	79.50	72.97	84.38	N/A	75.31	78.13	N/A	73.91	80.00	N/A	70.00	86.25	N/A	75.47	86.25	N/A	75.78	86.25	N/A	75.47

Table A7
The table refers the Motion Recognition performances of membership function on BodyLogin BLD-SS *User – Defined* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

BodyLogin Motion Recog. Full Body	DNN			SVM									KNN								
				RBF			LIN			POL			3			5			7		
	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test
Features									***PCA***												
U2	99.63	99.30	94.75	99.75	98.75	84.69	99.00	98.10	89.47	96.38	91.85	61.53	99.88	99.15	87.09	99.88	99.25	87.19	99.88	99.10	87.44
U3	99.13	97.45	91.00	98.75	97.20	71.03	98.50	96.93	82.06	98.25	93.35	49.00	98.87	97.80	75.63	99.00	97.70	75.84	98.87	97.60	75.72
U4	99.62	98.75	94.69	99.50	98.60	88.59	98.75	98.10	87.47	98.25	95.45	65.72	99.63	98.95	90.94	99.75	98.95	90.97	99.75	98.90	91.12
UR	99.63	99.40	95.59	99.63	99.25	85.72	99.13	98.18	89.72	97.50	93.70	61.69	100	99.20	88.38	99.88	99.25	88.5	99.75	99.30	88.28
LBP									***LBP***												
U2	96.25	94.05	81.75	100	99.75	97.94	100	99.20	97.72	100	99.40	97.81	100	99.85	98.06	100	99.85	98.06	100	99.85	98.09
U3	97.38	93.20	81.13	100	99.75	97.44	99.88	99.15	96.47	99.88	99.40	97.31	100	99.75	97.63	100	99.75	97.59	100	99.75	97.66
U4	97.38	93.35	80.91	100	99.85	97.5	100	99.15	96.97	100	99.40	97.34	100	99.90	97.56	100	99.90	97.66	100	99.90	97.44
UR	96.88	92.95	81.75	100	99.70	97.81	100	99.25	97.44	100	99.50	97.78	100	99.75	97.88	100	99.75	97.91	100	99.75	97.84
HOG									***HOG***												
U2	100	99.70	98.31	99.88	99.40	98.66	99.75	97.90	98.47	99.63	97.85	98.25	99.88	99.60	98.75	99.88	99.60	98.84	99.88	99.60	98.94
U3	99.62	99.20	96.72	99.88	98.85	98.06	99.63	97.95	97.63	99.38	97.35	96.78	100	99.10	98.06	99.88	99.05	98.13	99.88	99.05	98.09
U4	99.75	99.85	97.94	99.88	99.40	98.84	99.63	98.40	98.75	99.63	98.10	98.66	99.88	99.50	98.94	99.88	99.50	98.97	99.88	99.50	99.06
UR	100	99.65	98.72	99.88	99.50	98.75	99.63	98.50	98.59	99.63	98.25	98.5	99.88	99.70	98.78	99.88	99.60	98.78	99.88	99.60	98.75

Table A8
The table refers the Person Identification performances of membership function on BodyLogin*Full – Body* dataset using Masked Depth Video Stream features and various classifiers. (The following table refer to colored squares performance as Best 2nd 3rd Worst scores.).

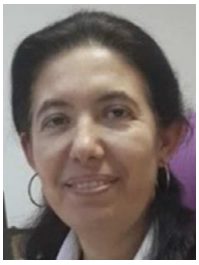
BodyLogin Person Ident. Full Body	DNN			SVM									KNN								
				RBF			LIN			POL			3			5			7		
	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test	20% Test	50% Test	80% Test
Features									***PCA***												
U2	98.60	97.36	83.09	98.50	97.35	67.13	97.03	93.54	58.6	96.22	92.13	56.26	98.66	97.84	69.62	98.78	97.85	69.71	98.75	97.93	69.83
U3	98.22	96.75	82.56	98.10	96.29	68.38	96.47	92.76	63.37	95.63	91.38	61.14	98.38	96.90	70.11	98.41	96.96	70.30	98.41	97.00	70.24
U4	99.01	97.53	87.18	99.16	97.69	70.83	97.88	94.95	60.90	97.19	93.75	58.64	99.19	98.16	73.48	99.25	98.23	73.48	99.25	98.25	73.67
UR	98.63	97.70	87.22	98.82	97.81	73.16	97.66	94.65	64.72	96.88	93.59	62.11	98.97	98.24	75.58	98.97	98.40	75.54	99.10	98.38	75.63
LBP									***LBP***												
U2	86.04	77.26	N\A	99.22	97.89	88.27	98.63	94.46	86.02	98.56	94.91	87.25	99.31	98.18	89.19	99.41	98.23	89.25	99.31	98.23	89.24
U3	85.63	78.40	N\A	99.13	97.43	87.94	98.78	94.40	85.46	98.75	95.19	87.08	99.16	97.79	88.31	99.19	97.81	88.47	99.22	97.83	88.37
U4	84.47	77.46	N\A	99.28	98.01	88.35	98.75	95.20	85.84	98.88	95.84	87.53	99.31	98.31	88.82	99.34	98.35	88.91	99.38	98.33	88.87
UR	84.82	76.56	N\A	99.28	97.84	88.57	98.53	95.10	86.76	98.88	95.39	87.85	99.41	98.26	89.09	99.38	98.25	89.07	99.38	98.24	89.07
HOG									***HOG***												
U2	98.88	98.00	87.83	99.07	97.56	89.25	98.16	94.58	88.51	97.91	93.93	87.71	99.66	99.08	90.52	99.66	99.05	90.63	99.69	99.11	90.75
U3	98.75	97.69	87.47	98.88	97.60	88.4	98.13	95.65	87.6	98.28	94.96	86.37	99.56	98.53	90.26	99.59	98.53	90.36	99.59	98.54	90.38
U4	99.16	98.19	89.9	99.22	97.89	90.94	98.63	95.48	89.76	98.34	94.96	88.94	99.59	98.95	90.7	99.56	98.96	90.69	99.56	98.95	90.73
UR	98.94	97.96	90.21	99.32	97.83	90.68	98.44	95.65	89.57	98.35	94.99	88.82	99.72	98.99	90.75	99.75	98.96	90.79	99.75	98.94	90.89

References

[1] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Computing Surveys (CSUR) 43 (3) (2011) 16.
[2] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans Pattern Anal Mach Intell 23 (3) (2001) 257–267.
[3] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Transactions on Pattern Analysis & Machine Intelligence 28 (2) (2006) 316–322.
[4] J.W. Davis, Hierarchical motion history images for recognizing human motion, in: Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on, IEEE, 2001, pp. 39–46.
[5] J. Liu, N. Zheng, Gait history image: A novel temporal template for gait recognition, in: 2007 IEEE International Conference on Multimedia and Expo, 2007, pp. 663–666.
[6] Q. Ma, S. Wang, D. Nie, J. Qiu, Recognizing humans based on gait moment image, in: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007., volume 2, IEEE, 2007, pp. 606–610.
[7] C. Chen, J. Liang, H. Zhao, H. Hu, J. Tian, Frame difference energy image for gait recognition with incomplete silhouettes, Pattern Recognit Lett 30 (11) (2009) 977–984.
[8] V. Megavannan, B. Agarwal, R.V. Babu, Human action recognition using depth maps, in: Signal Processing and Communications (SPCOM), 2012 International Conference on, IEEE, 2012, pp. 1–5.
[9] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, IEEE, 2012, pp. 28–35.
[10] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, et al., Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, in: AAAI, volume 2, 2016, p. 6.
[11] K. Gościewska, D. Frejlichowski, Action classification for partially occluded silhouettes by means of shape and action descriptors, Applied Sciences 11 (18) (2021) 8633.
[12] X. Cao, et al., Human motion recognition information processing system based on LSTM recurrent neural network algorithm, J Ambient Intell Humaniz Comput (2022) 1–13.
[13] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3288–3297.
[14] N.S. Russel, A. Selvaraj, Fusion of spatial and dynamic cnn streams for action recognition, Multimedia Systems 27 (5) (2021) 969–984.
[15] T. Singh, D.K. Vishwakarma, A deeply coupled convnet for human activity recognition using dynamic and RGB images, Neural Computing and Applications 33 (1) (2021) 469–485.
[16] P. Wang, Z. Li, Y. Hou, W. Li, Action recognition based on joint trajectory maps using convolutional neural networks, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 102–106.
[17] C. Li, Y. Hou, P. Wang, W. Li, Joint distance maps based action recognition with convolutional neural networks, IEEE Signal Process Lett 24 (5) (2017) 624–628.
[18] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[19] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, Action recognition with dynamic image networks, IEEE Trans Pattern Anal Mach Intell 40 (12) (2017) 2799–2813.
[20] J.M. Carmona, J. Climent, Human action recognition by means of subtensor projections and dense trajectories, Pattern Recognit 81 (2018) 443–455.
[21] M. Poonkodi, G. Vadivu, Action recognition using correlation of temporal difference frame (CTDF)-an algorithmic approach, J Ambient Intell Humaniz Comput 12 (7) (2021) 7107–7120.
[22] P. Ramya, R. Rajeswari, Human action recognition using distance transform and entropy based features, Multimed Tools Appl 80 (6) (2021) 8147–8173.
[23] Y.A. Andrade-Ambriz, S. Ledesma, M.-A. Ibarra-Manzano, M.I. Oros-Flores, D.-L. Almanza-Ojeda, Human activity recognition using temporal convolutional neural network architecture, Expert Syst Appl 191 (2022) 116287.
[24] J. Chen, W. Yang, C. Liu, L. Yao, A data augmentation method for skeleton-based action recognition with relative features, Applied Sciences 11 (23) (2021) 11481.
[25] E.P. Ijjina, K.M. Chalavadi, Human action recognition in RGB-D videos using motion sequence information and deep learning, Pattern Recognit 72 (2017) 504–516.
[26] A. Abdelbaky, S. Aly, Two-stream spatiotemporal feature fusion for human action recognition, Vis Comput 37 (7) (2021) 1821–1835.
[27] C. Liu, J. Ying, H. Yang, X. Hu, J. Liu, Improved human action recognition approach based on two-stream convolutional neural network model, Vis Comput 37 (6) (2021) 1327–1341.
[28] S.R. Mishra, T.K. Mishra, G. Sanyal, A. Sarkar, S.C. Satapathy, Real time human action recognition using triggered frame extraction and a typical cnn heuristic, Pattern Recognit Lett 135 (2020) 329–336.
[29] B. Liu, H. Cai, Z. Ju, H. Liu, Rgb-d sensing based human action and interaction analysis: a survey, Pattern Recognit 94 (2019) 1–12.
[30] J. Wu, P. Ishwar, J. Konrad, Silhouettes versus skeletons in gesture-based authentication with kinect, in: 2014 International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2014, pp. 99–106.
[31] J. Wu, P. Ishwar, J. Konrad, The value of posture, build and dynamics in gesture-based user authentication, in: Biometrics (IJCB), 2014 IEEE International Joint Conference on, IEEE, 2014, pp. 1–8.
[32] Y. Song, D. Demirdjian, R. Davis, Tracking body and hands for gesture recognition: Natops aircraft handling signals database, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 500–506.
[33] T. Theoharis, G. Passalis, G. Toderici, I.A. Kakadiaris, Unified 3D face and ear recognition using wavelets on geometry images, Pattern Recognit 41 (3) (2008) 796–804.
[34] Correlation description of two random variables, (https://www.mathworks.com/help/matlab/ref/corrcoef.html), Accessed: 2018-03-22.
[35] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in: British Machine Vision Conference, 2014.
[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
[37] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 86–94.
[38] M. Xie, N. Jean, M. Burke, D. Lobell, S. Ermon, Transfer learning from deep features for remote sensing and poverty mapping, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 3929–3935.
[39] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.
[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[41] Y. Song, D. Demirdjian, R. Davis, Continuous body and hand gesture recognition for natural human-computer interaction, ACM Transactions on Interactive Intelligent Systems (TiiS) 2 (1) (2012) 5.
[42] Y. Song, L.-P. Morency, R. Davis, Multi-view latent variable discriminative models for action recognition, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2120–2127.
[43] A. Elboushaki, R. Hannane, K. Afdel, L. Koutti, Multid-cnn: a multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in rgb-d image sequences, Expert Syst Appl 139 (2020) 112829.
[44] P. Khaire, P. Kumar, J. Imran, Combining cnn streams of rgb-d and skeletal data for human activity recognition, Pattern Recognit Lett 115 (2018) 107–116.
[45] Y. Ji, G. Ye, H. Cheng, Interactive body part contrast mining for human interaction recognition, in: Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on, IEEE, 2014, pp. 1–6.
[46] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.
[47] M. Edwards, X. Xie, Generating local temporal poses from gestures with aligned cluster analysis for human action recognition, UK Computer Vision Student Workshop (BMVW), BMVA Press, 2015. 1–1
[48] T. Huynh-The, O. Banos, B.-V. Le, D.-M. Bui, S. Lee, Y. Yoon, T. Le-Tien, Pam-based flexible generative topic model for 3d interactive activity recognition, in: Advanced Technologies for Communications (ATC), 2015 International Conference on, IEEE, 2015, pp. 117–122.
[49] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: AAAI, volume 1, 2017, pp. 4263–4270.
[50] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 816–833.
[51] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, L. Zhang, A deep structured model with radius–margin bound for 3D human activity recognition, Int J Comput Vis 118 (2) (2016) 256–273.
[52] A. Gharahdaghi, F. Razzazi, A. Amini, A non-linear mapping representing human action recognition under missing modality problem in video data, Measurement 186 (2021) 110123.
[53] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis & Machine Intelligence 24 (7) (2002) 971–987.
[54] R.K. McConnell, Method of and apparatus for pattern recognition, 1986, US Patent 4,567,610.
[55] N. Calik, M.A. Belen, P. Mahouti, Deep learning base modified MLP model for precise scattering parameter prediction of capacitive feed antenna, Int. J. Numer. Modell. Electron. Networks Devices Fields (2019) e2682.
[56] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).
[57] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

Onur Can Kurban received the B.S. degree in electrical and electronics engineering Erciyes University, Kayseri, Turkey in 2009 and the M.S., degree in electronics and communication engineering from Yildiz Technical University (YTU), Istanbul, Turkey, in 2014. Currently, he is in the department of Electronics and Communication Eng. in YTU as a PhD candidate. His research interests are image processing, human-computer interaction, behavioral and soft biometrics and machine learning.

Nurullah Calik received the B.S., M.S. degrees in Electronics and Communications Eng. From Yildiz Technical University (YTU), Istanbul, Turkey, in 2010 and 2013, respectively, and the Ph.D. degree in Electronics and Communications Engineering from the YTU in 2019. He completed post-doctoral studies in Istanbul Technical University Informatics Institute. He is currently working as an assistant professor member at Istanbul Medeniyet University. His research interests are signal and image processing and machine learning.



Tülay Yildirim received the B.S. and M.S. degrees in electronics and communication engineering from Yildiz Technical University (YTU), Istanbul, Turkey, in 1990 and 1992, respectively, and the Ph.D. degree in electrical and electronics engineering from the University of Liverpool, Liverpool, U.K., in 1997. Currently, she is a Full Professor with YTU. Her current research interests include analog and digital integrated circuit design, hardware implementations of neural networks, medical electronics, biometrics, and artificial intelligence.