Accepted Manuscript

Human action recognition in RGB-D videosusing motion sequence information and deep learning

Earnest Paul Ijjina, C.Krishna Mohan

PII: S0031-3203(17)30284-4
DOI: 10.1016/j.patcog.2017.07.013
Reference: PR 6216

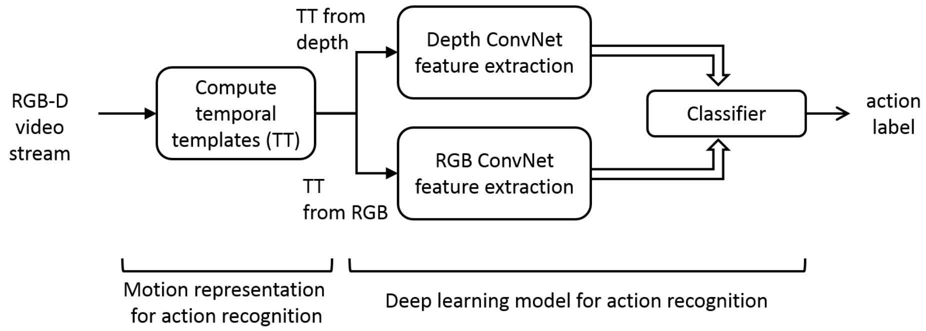
To appear in: *Pattern Recognition*

Received date: 15 December 2016
Revised date: 5 July 2017
Accepted date: 7 July 2017

Please cite this article as: Earnest Paul Ijjina, C.Krishna Mohan, Human action recognition in RGB-D videosusing motion sequence information and deep learning, *Pattern Recognition* (2017), doi: 10.1016/j.patcog.2017.07.013

Highlights

- An approach to recognize human actions in RGB-D videos using motion sequence information and deep learning is proposed.
- Proposed a new representation of motion information for human action recognition that emphasizes motion in various temporal regions.
- The use of motion information in RGB and depth video streams.
- Analysis using t-SNE visualization of ConvNet features to show the discriminative characteristics of the proposed representation.


RGB-D
video
stream
Compute
temporal
templates (TT)
TT from
depth
TT
from RGB
Depth ConvNet
feature extraction
RGB ConvNet
feature extraction
Classifier
action
label
Motion representation
for action recognition
Deep learning model for action recognition

Human action recognition in RGB-D videos using motion sequence information and deep learning

Earnest Paul Ijjina[a,*], Krishna Mohan C[a]

[a] *Visual Learning and Intelligence Group (VIGIL)*
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Telangana, INDIA-502285

Abstract

In this paper, we propose an approach for recognizing human action based on motion sequence information in RGB-D video using deep learning. A new representation that gives emphasis to the key poses associated with each action is presented. The features obtained from motion in RGB and depth video streams are given as input to the convolutional neural network to learn the discriminative features. The efficacy of the proposed approach is demonstrated on MIVIA action, NATOPS gesture, SBU Kinect interaction, and Weizmann datasets.

Keywords: Multi-modal action recognition, Deep learning, Motion information, Extreme Learning Machines

1. Introduction

The field of human behavior analysis aims to understand the subjects behavior over time using motion information. This analysis is categorized into motion, gesture, action, event or activity recognition depending on the duration of the observation. It can be further classified into a single person behavior, inter personal interaction, interaction with an object, group, and crowd behavior analysis based on the number of subjects and objects involved in the motion.

*Corresponding author
Email address: cs12p1002@iith.ac.in (Earnest Paul Ijjina)

The sub-categories of single person behavior into full body action, upper (or) lower body action, hand gesture, and facial expression differ in the region of interest used for recognition. Over decades, computer vision algorithms relied on only visual information to recognize these broad range of human behavior. With the availability of Kinect, a low cost RGB-D camera by Microsoft for its XBox gaming platform, there is rapid growth in the use of RGB-D videos for computer vision research [1].

A review of existing single/multiple-view and multi-person RGB-D datasets for human action recognition was conducted by Jing Zhang et al. in [2], summarizing the environmental conditions used for data acquisition, the characteristics of actions, recommended evaluation protocol, and state-of-the-art results for each dataset. In [3], Michael Firman et al. reviewed RGB-D datasets for various visual recognition tasks like object tracking, pose estimation, and action recognition. New modalities like infra-red vision and internal measurements units (IMU) sensor information are also becoming popular for surveillance videos, smart-homes, activities of daily living (ADL) monitoring, and fall detection. Among the existing RGB-D datasets, NTU RGB+D [4] is one of the largest database with 60 actions performed by 40 subjects that also includes infrared visual information. Fusion approaches on multiple modalities were used to recognize human actions using RGB-D video and wearable internal sensors [5]. In [6], spatio-temporal interest points (STIPs) and motion history images (MHIs) features extracted from RGB-D information along with fusion schemes are utilized to design a human daily activity recognition model for home environment. A descriptor for action representation using depth information is proposed in [7], to capture the structural relation of spatio-temporal points in action volumes for human action recognition. A deep architecture of comparative coding descriptor (DA-CCD) is used to learn high-level representation of depth information in [8], for human action recognition. The visual data from different views is mapped to a discriminative common feature space to learn a cross-view action recognition model [9]. The projection matrices necessary to map the data are simultaneously learnt for optimum discrimination. A pyramid part-wise bag of words (PPBoW) representation capturing the visual characteristics associated with actions is utilized in a multi-task learning model to discover and utilize the correlation between multiple views and body parts for multi-view human action recognition [10]. A review of multi-view learning approaches for exploring the consistency and complementary information across different views using co-training, multiple kernel learning, and subspace learning is discussed in [11].

A majority of pose based action recognition approaches use tracking information of various skeletal joints to compute features for action recognition. Features based on joint distance and joint motion are evaluated in [12], to recognize human interaction using support vector machine and multiple instance learning. The action recognition approach that relies on the co-occurrence of joints was proposed by Wentao Zhu et al. in [13] by using the 3D location of skeletal joints as input to an LSTM classifier, that is regularized using dropout. A local view-invariant skeletal descriptor, skeletal quads is proposed in [14]. A Gaussian mixture model (GMM) learnt on the training data is used to encode the quad as a Fisher vector which is inturn used by the support vector machine (SVM) for classification. To capture the joint shape motion ques in a depth image, HON4D, a descriptor for activity recognition using depth videos is proposed in [15] using SVM for classification. Models with inhomogeneous symmetric bias are trained with examples from an action domain in [16] and [17] for correcting the estimated human-pose. A descriptor to capture depth and spatial information from the segmentation mask of subjects pose, computed from depth information was proposed in [18]. The temporal ordering of these poses is used to learn subsequences of codewords for each activity and a boosted ensemble of discriminative subsequences is used for action recognition. In [19], a hierarchical recurrent network fusing the pose information from five parts of the skeletal structure is proposed to recognize actions from the temporally accumulated output. To recognize actions in RGB videos, action-bank features extracted from visual information are used to train discriminative dictionaries using 'label consistent K-SVD' algorithm in [20]. Human trajectories are modeled as heat sources to recognize group activities from the similarity of heat-maps [21]. Techniques for human detection, object detection, and tracking are combined to recognize human-human and human-object interactions in [22]. The gray-level, gradient and optical-flow information of RGB videos are given as input to a 3D convolutional neural network [23], to recognize human actions. The temporal evolution of pose associated with an action is modeled by a hidden Markov model (HMM) [24] to recognize human actions. In [25], Hu moments extracted from the depth motion history image and average depth image are used for action recognition using support vector machine.

The existing approaches either utilize engineered features (like HOG/HOF) or learn the discriminative features from input data using techniques like deep learning and dictionary learning. The approaches using hand-crafted features can exploit the domain knowledge but have limited generalization capability. On the other hand, feature learning models can generalize across various tasks but cannot utilize the domain knowledge of a system. To overcome these limitations, we propose a new temporal template representation to capture the motion in an entire video (i.e., not engineered for a particular task) while utilizing the domain knowledge to highlight motion in certain temporal frames. In addition, a convolutional neural network (deep learning model) is used to learn the local features from this temporal template representation for action recognition. Thus, by exploiting an input representation (that preserves the motion information in observations) and a discriminative feature learning model based on deep learning, we aim to design a classification framework with better generalization capability.

In this work, we present a new representation of motion information for human action recognition that emphasizes motion information in various temporal regions in contrast to the traditional motion history image that assigns higher weight to motion in the last frames. This motion representation computed from RGB and depth video streams is given as input to a convolutional neural network (CNN) for recognizing the human actions. The motion information computed from both modalities is used for action recognition, to overcome the limitations of individual modalities, namely, the need for high color contrast between the subject and background to capture accurate motion information in RGB video and the lack of sufficient discriminatory motion information for overlapping entities in depth video. Also, classification evidences using action representations highlighting motion in different regions is combined to exploit their complementary information for action recognition. The reminder of this paper is organized as follows: Section 2 describes the proposed human action recognition approach, the action representation, and the deep learning model used for action detection. Section 3 covers the experimental setup, results, and analysis of the proposed approach for MIVIA action, NATOPS gesture, SBU Kinect interaction, and Weizmann datasets. Finally, Section 4 gives concluding remarks and the future work.

2. Proposed approach

In this work, we present a new representation of motion information for human action recognition that emphasizes motion in different temporal regions to achieve better discrimination among actions. This representation of videos is given as input to a convolutional neural network (CNN) [26] model to extract ConvNet features. A classifier trained to recognize the human actions from these ConvNet features is used for action recognition. The block diagram of the proposed architecture is shown in Fig. 1. The following sub-sections explain each of these components in detail.

2.1. Motion representation for action recognition

In this work, we use temporal templates for action recognition due to their ability to capture the whole motion sequence in a single image. The temporal templates like the traditional motion history image (MHI) and motion energy image (MEI) are computed as the weighted sum of motion information in a

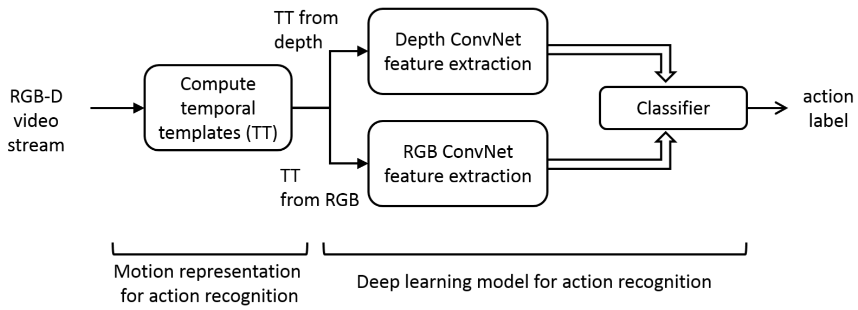


Figure 1: Block diagram of the proposed multimodal action recognition approach

video, where frame difference is used to compute motion between frames. The generalized formulation for computing temporal templates (TT) is given in Eq. 1, where n represents the number of frames in the observation, $m(i)$ denotes the motion in i^{th} frame of the observation and w_i represents the assigned weight (gray scale) value varying between 0 to 255. By replacing the fraction $\frac{w_i}{255}$ that varies from 0 to 1, with a fuzzy membership function $\mu(i)$, whose membership value also varies from 0 to 1 in Eq. 2, we get Eq. 3. It can be observed that w_i determines the significance assigned to the motion information in the i^{th} frame, $m(i)$, in the computed temporal template. This enhancement to the computation of temporal templates gives emphasis to motion information in a temporal region with the selection of the fuzzy membership function $\mu(i)$. To demonstrate this behavior considering three temporal regions, namely, {*begin, middle, end*} of observation, Fig. 2 shows four membership functions μ_1 to μ_4 whose corresponding equations are given in Eq. 4 to Eq. 7, respectively.

From this plots, it can be observed that μ_1 corresponds to the computation of motion energy image (MEI) and μ_2 computes the traditional motion history image (MHI). Since μ_1 is a constant function, an MEI assigns same weight to motion in all temporal regions. As μ_2 is a linearly increasing function, the significance assigned to motion information increases linearly with time in an MHI i.e., recent motion information has the highest significance. In case of μ_3, the weight assigned to motion information decreases linearly with time i.e., oldest motion information has highest significance. The last membership function μ_4 assigns higher weight to motion in the middle of the observation. Thus, the functions μ_2, μ_3, and μ_4 emphasize motion in the beginning, middle, and ending (i.e, different temporal regions) of the observation, respectively. In this work, we explore the representations computed from these functions for human action recognition. The next sub-section describes the use of convolutional neural networks for recognizing human actions from this representation.

$$\begin{aligned} TT &= \left(\tfrac{1}{255}\right) \textstyle\sum_{i=2}^{n} w_i \,.\, m(i) && (1)\\ &= \textstyle\sum_{i=2}^{n} \left(\tfrac{w_i}{255}\right) \,.\, m(i) && (2)\\ &= \textstyle\sum_{i=2}^{n} \mu(i) \,.\, m(i) && (3) \end{aligned}$$

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

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

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

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

2.2. Action recognition using deep learning

The previous section described the procedure to compute the motion representation of video observations that is considered in this section for

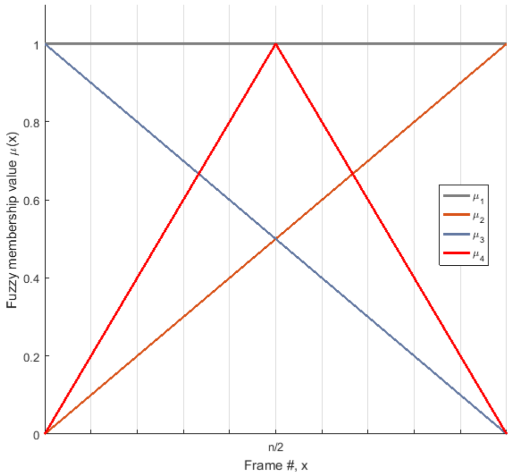


Figure 2: Plot of distribution of membership functions μ_1 to μ_4.

recognizing human actions. As the feature representation can entangle and hide more or less the different explanatory factors of variation behind the data, we use a convolutional neural network to learn the discriminative feature representation for human action recognition. This temporal template representation of videos with distinct local patterns for each action is given as input to a convolutional neural network (CNN) [26] to learn robust (ConvNet) features [27] [28] associated with each action, that are in turn used for action recognition. In this work, we use a 5C-2S-5C-2S CNN architecture for ConvNet feature extraction, where 5C represents a convolution layer with 5×5 kernels and 2S denotes a max-pooling sub-sampling layer using 2×2 kernels. To overcome the limitation of individual modalities in RGB-D videos, motion information computed from both the streams is processed separately to compute the ConvNet features. The ConvNet features computed from RGB and depth information are used to recognize human actions. Due to the better generalization capability of extreme learning machines (ELM) [29], ELM classifiers are used for action recognition. The next section discusses the experimental study of the proposed approach on MIVIA action, NATOPS gesture, SBU Kinect interaction, and Weizmann datasets.

3. Experimental study

The proposed approach was evaluated on MIVIA action [30], NATOPS gesture [31], SBU Kinect interaction [12], and Weizmann [32] datasets that contain RGB-D videos captured using Microsoft Kinect depth sensor. In this work, we compute temporal templates for each observation in these datasets from the RGB and binarized depth video streams. Due to the low accuracy of depth information captured by Kinect sensor, we binarize the depth video stream rather than using the gray scale value indicative of the depth, at each pixel location. The binarization of the depth video stream uses a threshold to binarize all the frames in a depth video. As a result, the binarized depth images have the spatial location of the subjects, similar to a silhouette. The experimental setup, results, and analysis for these datasets are given in the following sub-sections.

3.1. MIVIA action dataset

The MIVIA actions dataset [33] [30] consists of RGB-D video of 7 actions namely : *opening a jar*, *drinking*, *sleeping*, *random motion*, *stopping*, *interacting with a table*, and *sitting* performed by 14 subjects. Due to the absence of motion in RGB-D video for actions like *sleeping* and *sitting*, we consider binarized depth information as motion information in computing depth temporal templates. The leave-one-subject-out (LOSO) evaluation protocol is used to evaluate the performance of the proposed approach on this dataset. The optimum number of filters in CNN and the number of hidden nodes in ELM are empirically determined. The performance of the proposed approach for various membership functions is given in Table 1. The temporal templates computed using μ_4 has better performance than the other temporal templates. The consideration of binarized depth information as motion information in the computation of depth temporal templates is the possible cause for better performance of depth ConvNet features over RGB ConvNet features for recognizing some of the actions with small motion. The performance improves when both depth and RGB ConvNet features are considered for action recognition, which could be due to the complementary information captured by these modalities. The performance of the proposed approach, considering fusion across models trained on different temporal templates is given in Table 2. It can be observed that best performance of 93.37% can be achieved when temporal templates emphasizing motion in the beginning (μ_3), and the middle (μ_4) temporal regions are considered in fusion. The confusion matrix corresponding to this combination is given in Fig. 3. The performance comparison of the proposed approach with existing methods is given in Table 3. It can be observed that the proposed approach using temporal template features achieves better performance than the existing approaches. As the proposed approach uses raw video for action recognition (by computing temporal templates), a parallelized GPU implementation of the proposed approach could be used for real-time action recognition. The next sub-section covers the experimental study of the proposed approach on NATOPS gesture dataset.

Table 1: Classification performance (in %) of various membership functions for ConvNet features extracted from depth, RGB and RGB-D information on MIVIA action dataset. (Here, info. represents information)

Membership	ConvNet features + ELM classifier		
function	Depth info.	RGB info.	RGB-D info.
μ_1	87.85	42.10	88.95
μ_2	84.53	43.65	88.95
μ_3	87.29	39.78	90.06
μ_4	90.16	40.3	**90.61**

Table 2: Classification performance (in %) of groups of membership functions, combined using *Avg*-fusion rule on MIVIA action dataset using RGB-D ConvNet features

Fusion of membership functions (*with emphasis*)			Accuracy
begin	*middle*	*end*	
μ_3	μ_4		**93.37**
	μ_4	μ_2	91.71
μ_3		μ_2	90.61
μ_3	μ_4	μ_2	91.71

Table 3: Performance of existing and proposed approaches using RGB-D information on MIVIA actions dataset (in terms of classification accuracy in %)

Approach	RGB-D info.	Accuracy
Reject mechanism [33]	$\surd$	79.8
HaCK [34]	$\surd$	80.1
BoW [30]	$\surd$	84.1
Deep Learning [35]	$\surd$	84.7
Edit distance [36]	$\surd$	85.2
Proposed approach	$\surd$	**93.37**

3.2. NATOPS dataset

This dataset consists of 24 aircraft handling signals from the Naval Air Training and Operating Procedures Standardization (NATOPS) manual for the US naval aircraft. The motion involved in performing these gestures in given in Fig. 4. It can be observed that some of these gestures involve the movement of arms before the body and some of them require the changes in hand sign (thumb-up, thumb-down, open hand and closed hand) for gesture recognition. These gestures were captured using a Kinect sensor at 20 FPS with a resolution of 320×240. The location of skeletal joints in the upper

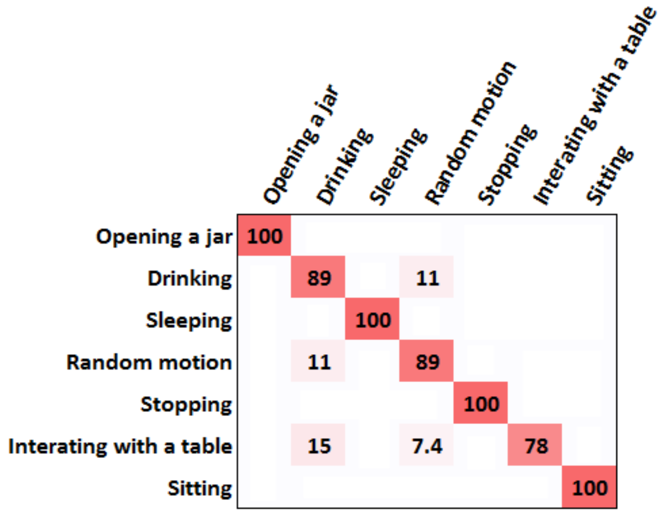


Figure 3: Confusion matrix of the proposed approach for MIVIA action dataset using fusion (Avg-rule) of evidences across membership functions$\{\mu_3, \mu_4\}$.

body along with the hand sign are available with the dataset. These 24 upper body gestures were performed by 20 subjects for 20 times, resulting in 400 observations for each (subject, gesture) pair. The evaluation criteria of using the observations corresponding to the first five subjects for testing and the last 10 subjects for training, as suggested in [31] is followed.

Similar to the previous dataset, the binarized depth and RGB information is used in the computation of temporal templates, that are in-turn used in depth and RGB ConvNet feature extraction. The temporal templates are down-sampled to 64×48 before feature extraction and the ELM classifiers with 10000 hidden nodes are considered for classification. The performance of the four membership functions for depth, RGB, and RGB-D (depth+RGB) ConvNet features is given in Table 4. It can be observed that RGB features are more effective when compared to depth features for these observations, which could be due to the yellow color vest worn by the subjects. Due to the high color contrast between the subjects arms and body, the arm movements

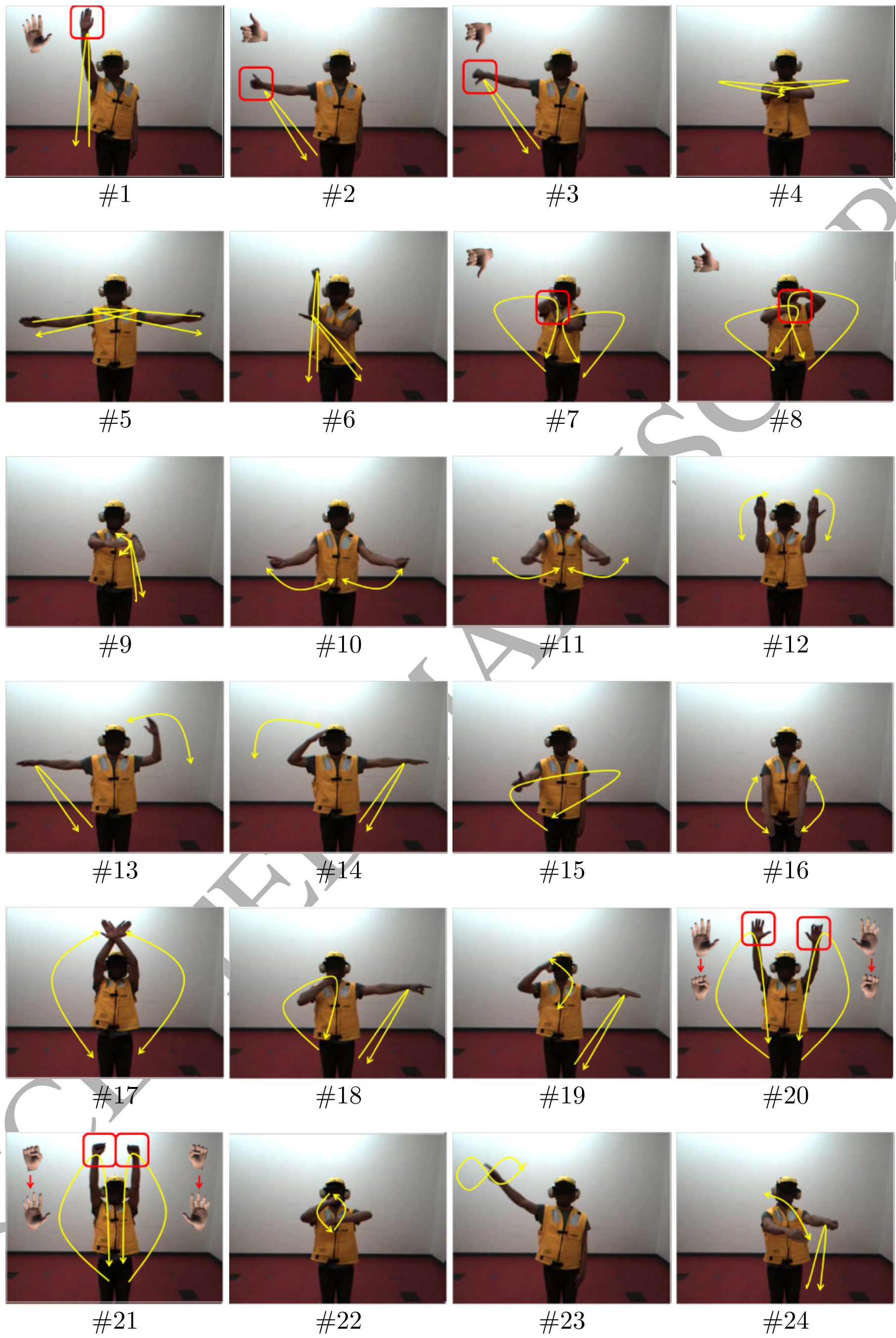


Figure 4: Movements involved in performing the 24 gestures of NATOPS dataset. (Fig. 9 in [31]) (Best viewed in color)

are well captured in RGB templates that could not be captured in the (binarized) depth temporal templates due to their overlap with the body. The fusion of evidences for temporal templates μ_2, μ_3, and μ_4 that emphasize motion in different temporal region, using RGB-D ConvNet features is given in Table 5. An accuracy of 72.58% is achieved by combining evidences of temporal templates generated by μ_2, μ_3, and μ_4 using *Avg*-rule, whose confusion matrix is given in Fig. 5. The unavailability of hand signal information to discriminate gestures (G2, G3) and (G20, G21) is the possible reason behind the high confusion between these gestures in the confusion matrix. This study also considers Top-n analysis, to identify how close this approach is in recognizing the correct class label. The Top-n analysis reports an observation as correctly classified if the actual class label is in the Top-n (determined from the confidence value associated with each label) predicted labels. The performance comparison of the proposed approach with existing approaches is given in Table 6. As the current experimental setup does not utilize the explicit hand signal information for recognizing the human actions, we normalize the results of the proposed approach using Top-2 analysis for comparison with existing approaches. The table suggests that the performance of the proposed approach is comparable with existing approaches that use both skeletal and hand signal information. As skeletal and hand signal information are extracted from RGB-D data, a parallel (GPU) implementation of the proposed approach using raw video could be faster than the existing approaches. The next sub-section covers the experimental study of the proposed approach on SBU interaction dataset.

3.3. SBU Kinect interaction dataset

This dataset consists of 8 types of two-person interactions, namely, *approaching, departing, pushing, kicking, punching, exchanging objects, hugging*, and *shaking hands*, whose typical key frames are shown in Fig. 6. This is a challenging database due to similarity in motion for some actions. For instance, *exchanging object* and *shaking hands* involves extending the arms

Table 4: Performance of various membership functions (in terms of classification accuracy in %) for depth, RGB and RGB-D (depth+RGB) ConvNet features on NATOPS dataset

Membership function	ConvNet features + ELM classifier		
	Depth info.	RGB info.	RGB-D info.
μ_1	35.21	60.88	59.29
μ_2	44.33	64.08	61.58
μ_3	44.21	61.92	62.92
μ_4	44.33	66.83	**68.83**

Table 5: Performance of fusion of membership functions, combined using *Avg* fusion rule on NATOPS (in terms of classification accuracy in %) for test data using RGB-D ConvNet features

Fusion of membership functions (*with emphasis*)			Performance of Top-n				
begin	*middle*	*end*	Top-1	Top-2	Top-3	Top-4	Top-5
μ_3	μ_4		71.58	83.71	87.83	90.29	92.42
	μ_4	μ_2	72.88	85.50	90.04	92.25	93.92
μ_3		μ_2	68.42	82.63	87.42	91.04	92.96
μ_3	μ_4	μ_2	72.58	**86.58**	91.08	93.29	94.75

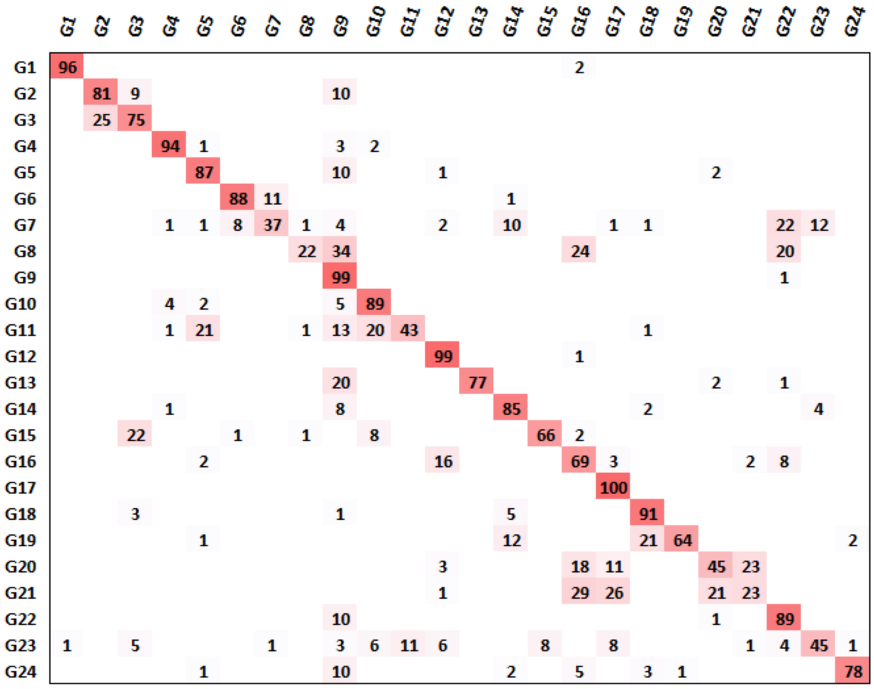


Figure 5: Confusion matrix of the proposed approach for NATOPS dataset.

Table 6: Performance of existing and proposed approach on NATOPS (in terms of classification accuracy in %) dataset

Approach	RGB-D raw video streams	Features: Skeletal	Features: Hand	Accuracy
Yale Song *et al.* [37]		√	√	75.37
CRF, Yale Song *et al.* [38]		√	√	53.30
HMM, Yale Song *et al.* [38]		√	√	77.67
HCRF, Yale Song *et al.* [38]		√	√	78.00
Couples HCRF, Yale Song *et al.* [38]		√	√	86.00
Linked HCRF, Yale Song *et al.* [38]		√	√	87.00
Proposed approach	√			72.58
Proposed approach (Top-2)	√			**86.58**

by both subjects. Some of these interactions (*approaching, departing, pushing, kicking* and *punching*) involve initiation of the action by one subject and the second subject responds to the action. As there are two subjects in each interaction, observations are captured when the left subject initiates the action as well as when the right subject initiates the action. For each observation, RGB and depth video streams at 15 frames per second (FPS) with a resolution of 640×480 pixels is provided. The observations in this dataset are captured using 21 different subject pairs. The 5-fold cross validation is used for evaluating this dataset [31]. The observations are divided into 5 groups and 4 groups are used for training and the remaining group is used for testing. This process in repeated for 5 times, changing the group considered for testing in each repetition.

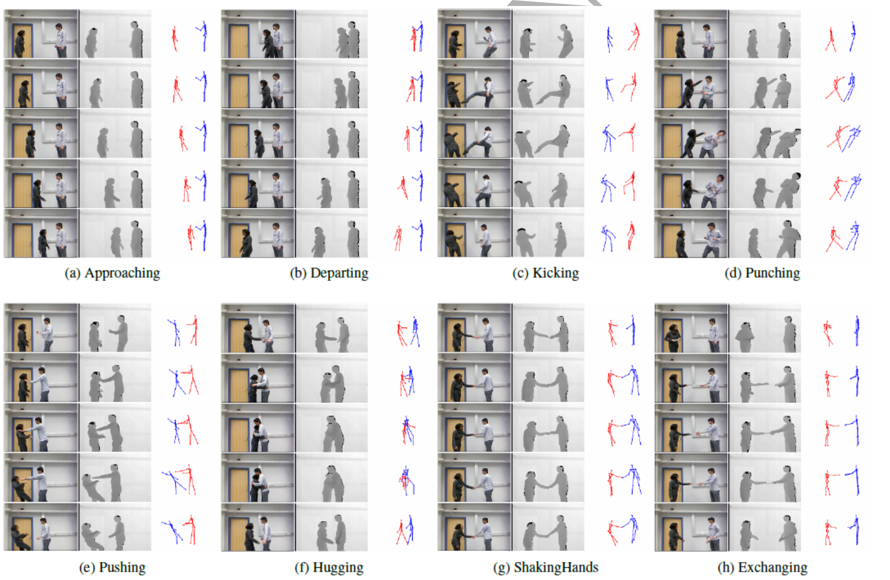


Figure 6: Typical key frames for various actions is SBU Kinect interaction dataset. (Fig. 1 in [12])

The (binarized) depth and RGB streams of the observations are used in the computation of depth and RGB temporal templates. As there is no or small movement of one subject during actions like *approaching* and *departing*, the computation of depth temporal templates is modified to utilize the individual frames as the motion information instead of frame difference. As the interactions remain the same even when the movements of the subjects are exchanged, horizontally flipped temporal templates are also considered in the evaluation thereby doubling the number of observations available for recognition. During testing, the maximum of the confidences corresponding to the original and the horizontally flipped data is used to determine the class label for an observation. The 640×480 pixel temporal templates computed from depth and RGB data are down-sampled to obtain a 40×32 representation, that is given as input to the feature extraction module. The CNN classifier in the ConvNet feature extraction module is trained using back-propagation algorithm in batch mode with a batch size of 16, for 50 epochs. The optimum batch size and number of training epochs is determined empirically. The generated ConvNet features are used to train ELM classifiers for action detection.

The performance of various membership functions used in the computation of temporal templates is given given in Table 7. From the average performance over the 5-splits using depth, RGB, and RGB-D features, it can be observed that better performance is obtained using RGB-D information i.e., the combination of depth and RGB ConvNet features. This may be due to the availability of complementary motion information in RGB and depth motion representations. Among the four fuzzy membership functions defined to compute temporal templates, μ_2 and μ_3 emphasize motion in the last and first frames of the observation, respectively. The experimental results obtained by combining evidences across models emphasizing different temporal regions is given in Table 8. An accuracy of 90.98% is achieved by combining evidences across temporal templates of μ_3 and μ_4 using Avg fusion-rule, whose confusion matrix is shown in Fig. 7. The performance comparison of the proposed approach against existing approaches is shown in Table 9. It can be observed that the proposed approach achieves better performance when compared to majority of the existing approaches. The performance of the action recognition

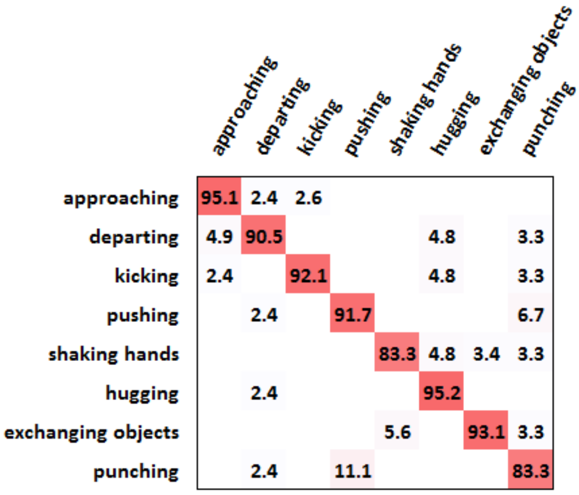


Figure 7: Confusion matrix of the proposed approach for 5-fold cross validation on SBU Kinect interaction dataset.

model in [39] is better than the proposed approach due to the use of pre-trained 3D CNN for optimization. Also, the methods in [40] and [41] achieve better performance because skeletal features with Long shot-term memory (LSTM) (that is efficient for recognizing time series data) are used to recognize human actions.

Overall, our proposed approach performs better than all other existing approaches on MIVIA action and NATOPS gesture datasets. On SBU Kinect interaction dataset, our approach performs better than majority of the existing approaches due to the above-mentioned reasons. As the skeletal information is computed from RGB-D video, a (GPU based) parallel implementation of the proposed approach using raw RGB-D data will be faster than the existing approaches. The next sub-section covers the experimental study of the proposed approach on Weizmann action dataset.

Table 7: Performance (in terms of classification accuracy in %) of various membership functions for 5-fold cross validation on SBU Kinect interaction dataset

Membership function	Set	ConvNet features with ELM classifier		
		Depth info.	RGB info.	RGB-D info.
μ_1	1	62.26	71.70	60.38
	2	88.24	72.55	82.35
	3	80.00	67.27	74.55
	4	62.75	78.43	66.67
	5	72.31	67.69	73.85
	Average	73.11	71.52	71.56
μ_2	1	56.60	69.81	64.15
	2	86.27	82.35	82.35
	3	69.09	60.00	74.55
	4	86.27	76.47	78.43
	5	69.23	73.85	78.46
	Average	73.49	72.49	75.58
μ_3	1	64.15	69.81	69.81
	2	92.16	74.51	92.16
	3	87.27	74.55	85.45
	4	84.31	80.39	90.20
	5	83.08	78.46	87.69
	Average	82.19	75.54	**85.06**
μ_4	1	71.70	60.38	66.04
	2	88.24	70.59	88.24
	3	81.82	63.64	80.00
	4	82.35	68.63	80.39
	5	78.46	70.77	84.62
	Average	80.51	66.80	79.85

Table 8: Performance of fusion of membership functions using RGB-D information, for 5-fold cross validation on SBU Kinect interaction dataset (in terms of classification accuracy in %)

Fusion of membership functions (*with emphasis*)			Accuracy for 5-fold cross validation					
begin	*middle*	*end*	Set-1	Set-2	Set-3	Set-4	Set-5	Total
μ_3	μ_4		76.60	97.83	96.36	93.62	90.00	**90.98**
	μ_4	μ_2	72.34	95.65	92.73	82.98	78.33	84.71
μ_3		μ_2	78.72	97.83	96.36	89.36	91.67	90.20
μ_3	μ_4	μ_2	78.72	97.83	98.18	89.36	88.33	90.20

Table 9: Performance (in terms of classification accuracy in %) of existing and proposed approach using 5-fold cross-validation on SBU Kinect interaction dataset

Approach	RGB-D raw video stream	Skeletal features/data	Accuracy
Raw skeleton [12]		$\surd$	49.7
Joint features [12]		$\surd$	80.3
Raw skeleton [42]		$\surd$	79.4
Joint features [42]		$\surd$	86.9
Hierarchical RNN [19]		$\surd$	80.35
Cluster analysis of pose [43]		$\surd$	83.9
Deep LSTM [13]		$\surd$	86.03
Generative topic model [44]		$\surd$	90.3
STA-LSTM [40]		$\surd$	91.51
ST-LSTM + Trust Gate [41]		$\surd$	93.3
Radius-margin bound [39]	$\surd$		93.4
Proposed approach	$\surd$		**90.98**

3.4. Weizmann action dataset

The Weizmann action dataset [32] consists of RGB video of 9 actions namely : *bend*, *jack*, *jump*, *pjump*, *run*, *side*, *walk*, *wave1*, and *wave2* performed by 9 subjects. Due to the unavailability of depth information, the foreground information obtained from background subtraction is used to compute foreground temporal template, that is used in the proposed approach in place of the depth temporal template. The proposed approach is evaluated on this dataset using leave-one-sequence-out (LOSO) test strategy. The performance of the proposed approach for various membership functions is given in Table 10. The temporal templates computed using μ_4 has better performance than the other temporal templates. The performance improves when both foreground and RGB ConvNet features are considered for action recognition,

Table 10: Classification performance (in %) of various membership functions for ConvNet features extracted from foreground, RGB, and both information on Weizmann action dataset. (Here, info. represents information)

Membership function	ConvNet features + ELM classifier		
	Foreground info.	RGB info.	Both info.
μ_1	88.89	92.59	93.83
μ_2	91.36	92.59	93.83
μ_3	90.12	91.35	92.59
μ_4	93.83	95.06	**96.30**

which could be due to the complementary information captured by these templates. The performance of the proposed approach, considering fusion across models trained on different temporal templates is given in Table 11. It can be observed that best performance of 100% is achieved when temporal templates emphasizing motion in the middle (μ_4), and the end (μ_2) temporal regions are considered in combining the evidence. The performance comparison of the proposed approach with existing methods is given in Table 12. The proposed approach achieved an accuracy of 100% which is also the state-of-the-art performance on this dataset. The next sub-section includes comments on the proposed approach and the experimental study conducted on these datasets.

3.5. Comments and discussion

As discussed in the previous sections, the proposed approach was evaluated on MIVIA action, NATOPS gesture, SBU Kinect interaction, and Weizmann datasets. These experiments on two-person interaction and upper-body gesture recognition suggest the ability to extend this approach to other problem domains. The extraction of ConvNet features from the temporal template representation of actions could be the primary reason behind the adaptability of the proposed model. Some of the potential factors contributing to the high performance of the proposed approach are: 1) the design of new

Table 11: Classification performance (in %) of groups of membership functions, combined using *Avg*-fusion rule on Weizmann action dataset using foreground and RGB ConvNet features

Fusion of membership functions (*with emphasis*)			Accuracy
begin	*middle*	*end*	
μ_3	μ_4		97.53
	μ_4	μ_2	**100.0**
μ_3		μ_2	95.06
μ_3	μ_4	μ_2	98.76

Table 12: Performance of existing and proposed approaches on Weizmann action dataset

Approach	Accuracy (%)
S. Ali *et al.* [45]	92.6
Boiman and Irani *et al.* [46]	97.5
Kellokumpu *et al.* [47]	98.7
Blank *et al.* [32]	99.6
Yang Wang *et al.* [48]	100.0
Proposed approach	**100.0**

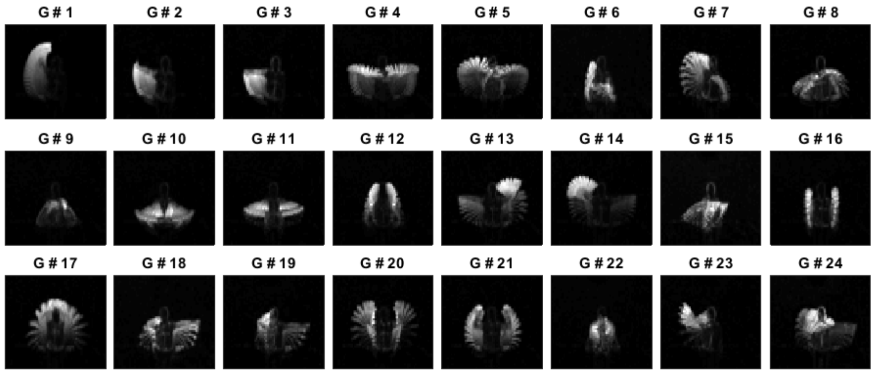


Figure 8: The typical temporal templates computed for the 24 gestures in NATOPS dataset using RGB vidoe with μ_4

action representation to emphasize motion in different temporal regions, 2) the use of binarized depth frames as motion information in the computation of depth temporal templates to recognize interactions involving static subjects, and 3) combining evidences across models with complementary characteristics (i.e., μ_2, μ_3, and μ_4 highlighting motion in the ending, beginning, and middle of observations, respectively). The typical temporal templates generated for NATOPS gestures is shown in Fig 8. It can be observed that this temporal template representation contains necessary discriminative information for action recognition. The generalization capability of deep learning architectures and the hardware implementations of CNN and ConvNet feature extraction facilitates the possibility for designing a real-time implementation of the proposed approach.

We analyze the effectiveness of the proposed ConvNet features and ELM classifier used in this work, with other classifiers and features extracted by pre-trained CNN models. We consider NATOPS dataset for this study due to its large number of classes and observations among the datasets evaluated in this work. The performance of the proposed ConvNet features with extreme learning machine (ELM), neural network (NN), and support vector machine (SVM) classifiers is given in Table 13. From the table, it can be observed that

ELM performs better than NN and SVM classifiers, because of its better generalization capability [49]. Also, from the last column of Table 13, it can be observed that the proposed ConvNet features are more discriminative than AlexNet features, that are obtained from a pre-trained CNN. This could be due to the training of AlexNet on natural color images whereas the proposed CNN is trained on the corresponding temporal templates which are gray scale images. For unconstrained videos, some of the existing deep learning action recognition models are more efficient than the proposed approach using temporal template representation which is sensitive to the angle of view. But, when observations are captured at the same angle of view (like in the cases of human computer interaction using Kinect), the proposed approach might outperform existing approaches in terms of both speed and accuracy. This could be due to the use of temporal templates as input to the deep learning model instead of raw video data.

Table 13: Performance of ConvNet features extracted from various temporal templates (TT) with ELM, NN and SVM classifiers on NATOPS dataset.

Temporal template	ConvNet features			AlexNet
computed using	ELM	NN	SVM	features
μ_1	59.29	52.08	56.42	55.54
μ_2	61.58	58.96	59.88	55.50
μ_3	62.92	53.46	60.33	55.96
μ_4	68.83	55.25	61.54	62.92

In MIVIA action and SBU Kinect interaction datasets, temporal templates generated using depth information are more effective than the once generated using RGB information. This may be due to the use of binarized depth information as motion information in the computation of depth temporal templates when compared to the use of frame difference in RGB temporal templates. As a result, depth templates will be able to recognize actions like *approaching*, *departing*, *Sleeping* and *Sitting* with a static subject. An illustration of typical templates generated for *approaching* and *departing* actions using binarized depth and frame difference for motion in the computation of temporal templates is shown in Fig 9. In NATOPS gesture dataset, RGB temporal templates outperformed depth templates, which may be due to the presence of gestures with arm movement in front of the body. Even when depth templates are unable to capture these movements, this information is captured in RGB temporal templates due to the high color contrast difference between subject's arms and the yellow color vest. As a result, RGB temporal templates have better discriminative information than the depth temporal templates in this dataset. The performance of the proposed approach is comparable with existing approaches, that use skeletal joints and hand signal information. The key contribution of this work is in redefining the computation of temporal templates using fuzzy membership functions that in-turn supports complex weight assignment through the distribution of membership function. It also provides the flexibility to learn the distribution of membership function as a curve fitting problem to optimize the performance for a set of actions. The t-SNE [50] visualization of the proposed action representation for SBU Kinect interaction dataset is shown in Fig. 10. The observations from this visualization are: i) the clusters in the visualization of the proposed representation indicate their ability to capture the necessary discriminative information for action recognition and ii) the well separated clusters in the visualization of ConvNet features indicate the robustness of deep learning features used for discrimination. Thus, by utilizing this representation for action recognition using a convolutional neural network, we propose a robust human action recognition. By exploiting the parallelism involved in the computation of this representation and the CNN using a GPU environment, this approach can be used for real-time action recognition.

From the comparative studies in Tables 3, 6, and 9, it can be observed that the existing approaches use either MOCAP information or other hand engineered features computed from visual information. As visual information is sensitive to noise, their robustness and generalization capability is limited. The use of temporal templates (capturing the motion history information) for

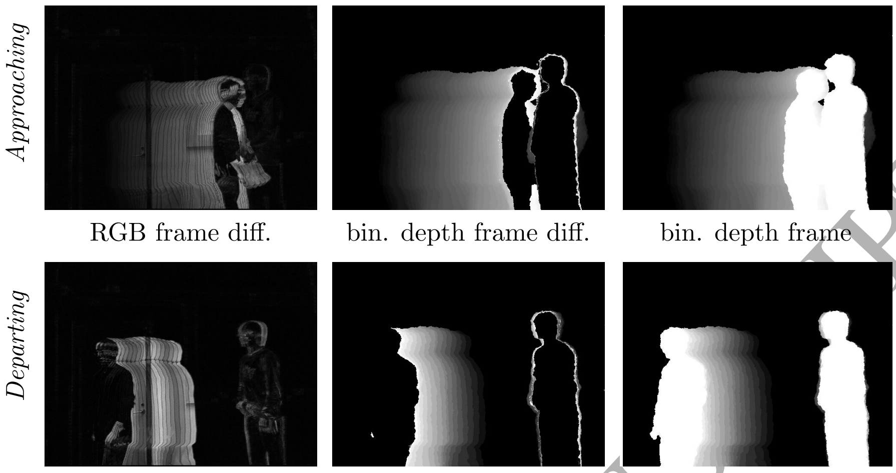
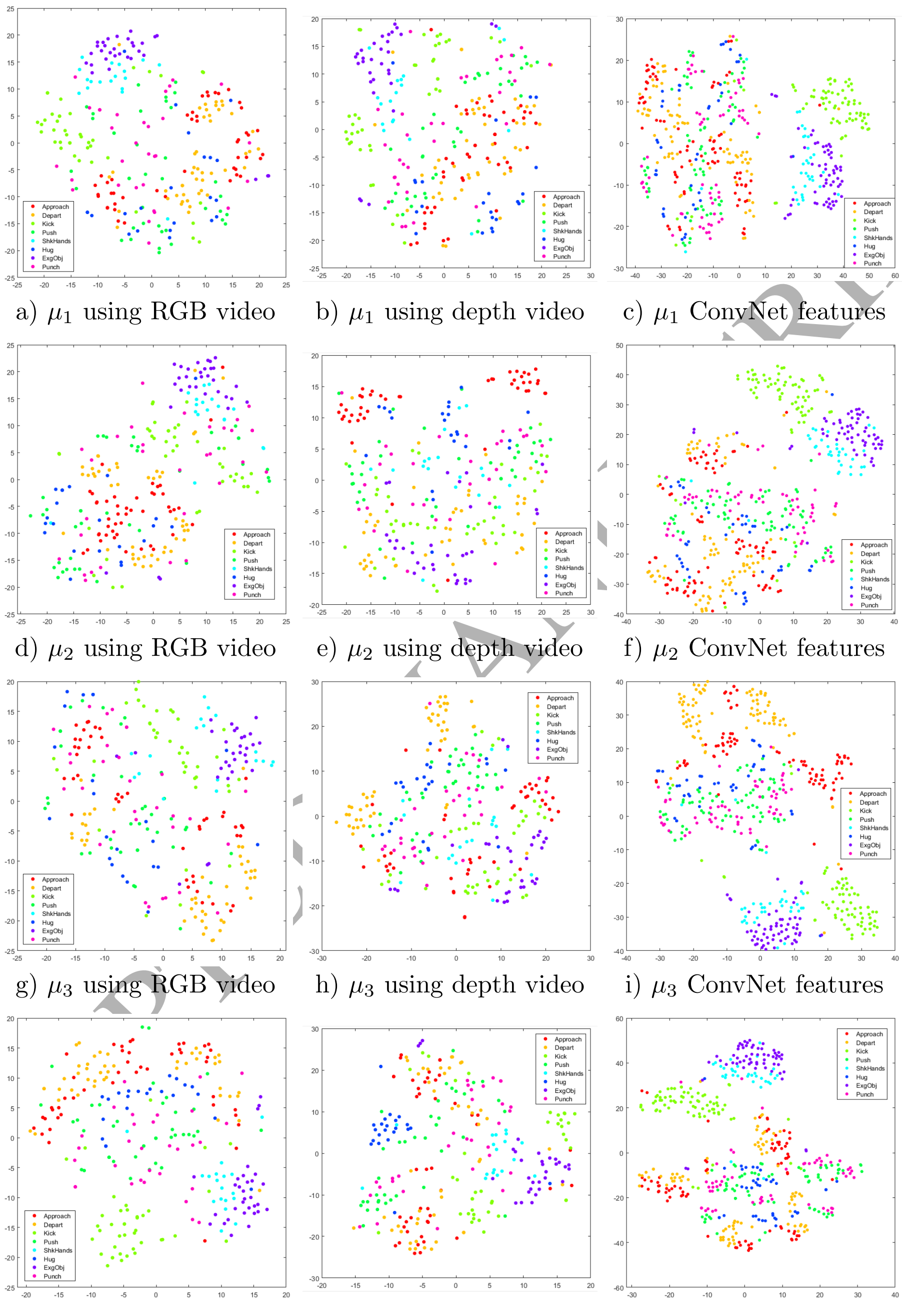


Figure 9: Temporal templates computed with μ_2 for *Approaching* and *Departing* actions, using RGB frame difference, binarized depth frame difference and binarized depth frame as motion information. (Here, diff. represents difference and bin.represents binary)

input representation reduces the loss of discriminative information available from raw visual data. The membership functions are used to emphasize motion information in certain temporal regions instead of using the entire motion sequence information (spatio-temporal volume) as input representation. The robustness of the proposed approach is further improved by considering ConvNet features extracted from temporal templates generated from different modalities.

The proposed approach is independent of the environment (indoor/outdoor) in which the observations are captured and can even be used with other modalities like infrared and thermal video. This work can be extended to recognize human actions from videos captured in real world by incorporating techniques like background subtraction for eliminating noise due to complex background and using subjects bounding-box obtained through tracking as region-of-interest to recognize actions performed by multiple subjects. To recognize actions in a streaming video (without action boundaries), this approach can be extended to process short fixed-length videos obtained by



a) μ_1 using RGB video b) μ_1 using depth video c) μ_1 ConvNet features

d) μ_2 using RGB video e) μ_2 using depth video f) μ_2 ConvNet features

g) μ_3 using RGB video h) μ_3 using depth video i) μ_3 ConvNet features

j) μ_4 using RGB video k) μ_4 using depth video l) μ_4 ConvNet features

Figure 10: t-SNE visualization of the proposed action representation generated for μ_1 to μ_4 using motion in RGB and depth video of SBU Kinect interaction dataset. (Best viewed in color)

running a sliding window on the video stream. As a result, in addition to the recognizing the action, the temporal duration/occurrence of the action with also be identified. To manage the variation in speed of execution of the action, multiple window sizes can be used (i.e., a shorter window for faster execution and longer window for slower execution), to obtain similar temporal template representation. Non-continuous actions like cooking activities can be recognized by identifying the temporal occurrence of the primary actions and using other approaches like fusion strategy in [51] or temporal fusion in [52] that can handle the temporal discrepancies on these actions in recognizing the activity. Even though temporal templates are affected by the capturing conditions (like distance from the subject, angle of view) and appearance of the subject, the use of ConvNet features extracted from temporal templates computed from depth information makes this a robust approach. Similar to observations in NATOPS dataset, there are areas in real world environment where this approach is applicable to discriminate the actions. The next section concludes this work.

4. Conclusions and future work

In this work, new representation for action recognition capable of emphasizing motion in different temporal regions is presented. A multi-modal action recognition approach, utilizing ConvNet features extracted from this new representation computed from RGB and depth information is proposed. The use of multi-modal information with noise tolerance of ConvNet features, gives the robustness and adaptability of this approach to other recognition tasks. Fusion of evidences across models suggests that optimum performance can be achieved by combining evidences across models emphasizing different temporal regions. The proposed approach would be faster than the existing approaches due to the simple arithmetic in computing the new representation (that can be parallelized) and the parallel implementations of ConvNet features extraction. In future, this work can be extended to other modalities like infrared images and other types of human behavior like hand gestures and group activity.

References

References

[1] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft kinect sensor: A review., IEEE Transactions on Cybernetics 43 (5) (2013) 1318–1334.

[2] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, C. Tang, Rgb-d-based action recognition datasets: A survey, CoRR abs/1601.05511.

[3] M. Firman, RGBD datasets: Past, present and future, CoRR abs/1604.00999.

[4] A. Shahroudy, J. Liu, T. Ng, G. Wang, Ntu rgb+d: A large scale dataset for 3d human activity analysis, CoRR abs/1604.02808.

[5] C. Chen, R. Jafari, N. Kehtarnavaz, Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: IEEE International Conference on Image Processing (ICIP), 2015, pp. 168–172. doi:10.1109/ICIP.2015.7350781.

[6] B. Ni, G. Wang, P. Moulin, Rgbd-hudaact: A color-depth video database for human daily activity recognition., in: A. Fossati, J. Gall, H. Grabner, X. Ren, K. Konolige (Eds.), Consumer Depth Cameras for Computer Vision, Advances in Computer Vision and Pattern Recognition, Springer, 2013, pp. 193–208.

[7] Z. Cheng, L. Qin, Y. Ye, Q. Huang, Q. Tian, Human daily action analysis with multi-view and color-depth data., in: A. Fusiello, V. Murino, R. Cucchiara (Eds.), ECCV Workshops (2), Vol. 7584 of Lecture Notes in Computer Science, Springer, 2012, pp. 52–61.

[8] Y. Liu, L. Qin, Z. Cheng, Y. Zhang, W. Zhang, Q. Huang, Da-ccd: A novel action representation by deep architecture of local depth feature, in: IEEE International Conference on Image Processing (ICIP), 2014, pp. 833–837.

[9] W. Sui, X. Wu, Y. Feng, Y. Jia, Heterogeneous discriminant analysis for cross-view action recognition, Neurocomputing 191 (2016) 286 – 295. doi:http://dx.doi.org/10.1016/j.neucom.2016.01.051.

[10] A.-A. Liu, N. Xu, Y.-T. Su, H. Lin, T. Hao, Z.-X. Yang, Single/multi-view human action recognition via regularized multi-task learning, Neurocomputing 151, Part 2 (2015) 544 – 553.

[11] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, CoRR abs/1304.5634.
URL http://arxiv.org/abs/1304.5634

[12] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 28–35.

[13] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks, CoRR abs/1603.07772.

[14] G. Evangelidis, G. Singh, R. Horaud, Skeletal quads: Human action recognition using joint quadruples, in: International Conference on Pattern Recognition (ICPR), 2014, pp. 4513–4518.

[15] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences., in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 716–723.

[16] K. D. Wei Shen, X. Bai, T. Leyvand, B. Guo, Z. Tu, Exemplar-based human action pose correction and tagging, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1784–1791.

[17] S. Wei, D. Ke, B. Xiang, L. Tommer, G. Baining, T. Zhuowen, Exemplar-based human action pose correction, IEEE Trans. Cybernetics 44 (7) (2014) 1053–1066.

[18] R. Gupta, A. Y.-S. Chia, D. Rajan, Human activities recognition using depth images, in: ACM International Conference on Multimedia, MM '13, ACM, New York, NY, USA, 2013, pp. 283–292.

[19] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1110–1118.

[20] Z. Jiang, Z. Lin, L. S. Davis, Label consistent k-svd: Learning a discriminative dictionary for recognition., IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11) (2013) 2651–2664.

[21] W. Lin, H. Chu, J. Wu, B. Sheng, Z. Chen, A heat-map-based algorithm for recognizing group activities in videos., IEEE Transactions on Circuits and Systems for Video Technology 23 (11) (2013) 1980–1992.

[22] A. Prest, V. Ferrari, C. Schmid, Explicit modeling of human-object interactions in realistic videos, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (4) (2013) 835–848.

[23] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition., IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1) (2013) 221–231.

[24] L. Xia, C.-C. Chen, J. K. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D), Rhode Island, USA, 2012, pp. 20–27.

[25] V. Megavannan, B. Agarwal, R. Venkatesh Babu, Human action recognition using depth maps, in: International Conference on Signal Processing and Communications (SPCOM), 2012, pp. 1–5.

[26] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324. doi:10.1109/5.726791.

[27] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1798–1828. doi:10.1109/TPAMI.2013.50.

[28] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (Eds.), Advances in Neural Information Processing Systems (NIPS) 27, Curran Associates, Inc., 2014, pp. 3320–3328.

[29] Extreme learning machine: Theory and applications, Neurocomputing 70 (13) (2006) 489 – 501, neural Networks Selected Papers from the 7th Brazilian Symposium on Neural Networks (SBRN '04).

[30] P. Foggia, G. Percannella, A. Saggese, M. Vento, Recognizing human actions by a bag of visual words, in: IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 2910–2915. doi:10.1109/SMC.2013.496.

[31] Y. Song, D. Demirdjian, R. Davis, Tracking body and hands for gesture recognition: Natops aircraft handling signals database., in: IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), 2011, pp. 500–506.

[32] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: IEEE International Conference on Computer Vision (ICCV), Vol. 2, 2005, pp. 1395–1402. doi:10.1109/ICCV.2005.28.

[33] V. Carletti, P. Foggia, G. Percannella, A. Saggese, M. Vento, Recognition of Human Actions from RGB-D Videos Using a Reject Option, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 436–445. doi:10.1007/978-3-642-41190-8_47.

[34] L. Brun, G. Percannella, A. Saggese, M. Vento, Hack: A system for the recognition of human actions by kernels of visual strings, in: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2014, pp. 142–147. doi:10.1109/AVSS.2014.6918658.

[35] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, Exploiting the deep learning paradigm for recognizing human actions, in: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2014, pp. 93–98. doi:10.1109/AVSS.2014.6918650.

[36] L. Brun, P. Foggia, A. Saggese, M. Vento, Recognition of human actions using edit distance on aclet strings, in: International Conference on Computer Vision Theory and Applications (VISAPP), 2015, pp. 97–103.

[37] Y. Song, D. Demirdjian, R. Davis, Continuous body and hand gesture recognition for natural human-computer interaction, ACM Trans. Interact. Intell. Syst. 2 (1) (2012) 5:1–5:28.

[38] Y. Song, L.-P. Morency, R. Davis, Multi-view latent variable discriminative models for action recognition., in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2120–2127.

[39] 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, CoRR abs/1512.01642.

[40] 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, CoRR abs/1611.06067.

[41] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition, Springer International Publishing, Cham, 2016, pp. 816–833. doi:10.1007/978-3-319-46487-9_50.

[42] Y. Ji, G. Ye, H. Cheng, Interactive body part contrast mining for human interaction recognition, in: IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2014, pp. 1–6.

[43] M. Edwards, X. Xie, Generating local temporal poses from gestures with aligned cluster analysis for human action recognition, in: G. K. L. Tam (Ed.), UK Computer Vision Student Workshop (BMVW), BMVA Press, 2015, pp. 1.1–1.12.

[44] 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: International Conference on Advanced Technologies for Communications (ATC), 2015, pp. 117–122. `doi:10.1109/ATC.2015.7388302`.

[45] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, in: IEEE International Conference on Computer Vision (ICCV), 2007, pp. 1–8. `doi:10.1109/ICCV.2007.4409046`.

[46] O. Boiman, M. Irani, Similarity by composition, in: B. Schlkopf, J. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, 2006, pp. 177–184.

[47] P. Crook, V. Kellokumpu, G. Zhao, M. Pietikainen, Human activity recognition using a dynamic texture based method, in: British Machine Vision Conference (BMVC), BMVA Press, 2008, pp. 88.1–88.10.

[48] Y. Wang, G. Mori, Human action recognition by semilatent topic models, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 31 (10) (2009) 1762–1774.

[49] G. B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42 (2) (2012) 513–529. `doi:10.1109/TSMCB.2011.2168604`.

[50] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (Nov) (2008) 2579–2605.

[51] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1725–1732. doi:10.1109/CVPR.2014.223.

[52] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1933–1941. doi:10.1109/CVPR.2016.213.

Earnest Paul Ijjina received his Masters in Computer Science and Engineering from Indian Institute of Technology Kharagpur in 2007 and is currently pursuing his Ph.D in Computer Science and Engineering at Indian Institute of Technology Hyderabad. Previously, he was an IT manager for four years at Morgan Stanley and an Assistant Professor for one year. His research interests include computer vision, video content analysis and deep learning.

Dr. C. Krishan Mohan received the Ph.D. Degree in Computer Science and Engineering from Indian Institute of Technology, Madras India in 2007. He received the Master of Technology in System Analysis and Computer Applications from National Institute of Technology, Surathkal India in 2000. He is currently Associate Professor with the Department of Computer Science and Engineering, Indian Institute of Technology, Hyderabad India. His research interests include video content analysis, pattern recognition, and neural networks.