





## Deep temporal motion descriptor (DTMD) for human action recognition

Nudrat NIDA<sup>1</sup> , Muhammad Haroon YOUSAF<sup>1,2,\*</sup> , Aun IRTAZA<sup>3</sup> , Sergio A. VELASTIN<sup>4,5,6</sup> 

<sup>1</sup>Department of Computer Engineering, University of Engineering and Technology-Taxila, Taxila, Pakistan

<sup>2</sup> Swarm Robotics Lab, National Centre of Robotics and Automation, Islamabad, Pakistan

<sup>3</sup>Department of Computer Science, University of Engineering and Technology-Taxila, Taxila, Pakistan

<sup>4</sup>Cortexica Vision Systems Ltd., London, UK

<sup>5</sup>Department of Computer Science, Applied Artificial Intelligence Research Group, University Carlos III de Madrid, Madrid, Spain

<sup>6</sup>School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

Received: 29.07.2019

Accepted/Published Online: 18.10.2019

Final Version: 08.05.2020

**Abstract:** Spatiotemporal features have significant importance in human action recognition, as they provide the actor's shape and motion characteristics specific to each action class. This paper presents a new deep spatiotemporal human action representation, the deep temporal motion descriptor (DTMD), which shares the attributes of holistic and deep learned features. To generate the DTMD descriptor, the actor's silhouettes are gathered into single motion templates by applying motion history images. These motion templates capture the spatiotemporal movements of the actor and compactly represent the human actions using a single 2D template. Then deep convolutional neural networks are used to compute discriminative deep features from motion history templates to produce the DTMD. Later, DTMD is used for learning a model to recognize human actions using a softmax classifier. The advantage of DTMD are that DTMD is automatically learned from videos and contains higher-dimensional discriminative spatiotemporal representations as compared to handcrafted features; DTMD reduces the computational complexity of human activity recognition as all the video frames are compactly represented as a single motion template; and DTMD works effectively for single and multiview action recognition. We conducted experiments on three challenging datasets: MuHAVI-Uncut, iXMAS, and IAVID-1. The experimental findings reveal that DTMD outperforms previous methods and achieves the highest action prediction rate on the MuHAVI-Uncut dataset.

**Key words:** Human activity recognition, deep convolutional neural network, motion history images, deep temporal motion descriptor, computer vision

### 1. Introduction

Human action recognition (HAR) is a significant research area in computer vision for generating context-aware applications for human assistance. Human action recognition is utilized in various fields including video surveillance systems [1], group activity recognition [2], video summarization [3], and smart education [4, 5]. HAR is a challenging problem because of real-world constraints such as viewpoint variation, background clutter, changes in scale, partial occlusion, lighting, appearance, and frame resolution that affect recognition accuracy. Similarly, describing behavioral components, e.g., gait style, gesture, posture, and pose, demand significant effort and tedious learning of a particular action. Moreover, certain actions are challenging to differentiate due to visual similarities between them, like walking and jogging [6]. The way in which an action is performed also

\*Correspondence: haroon.yousaf@uettaxila.edu.pk

makes action recognition difficult because actions are relative in nature. Moreover, due to fine-grained activities, action classes have less intraclass variability that can confuse the model, as well. Another challenge in the action recognition domain is a scarcity of application-specific benchmark datasets to model the progressive movements of human activities.

Recent research in HAR can broadly be categorized into two classes: representation-based and deep network-based techniques [6]. Representation-based techniques use handcrafted features for classifying human actions into different action classes and can be further classified into the following subcategories: spatiotemporal interest point extractors [7, 8], holistic representation [9–11], and motion trajectory extractors [12, 13]. The spatiotemporal interest point extractors consider the spatial interest points along with action-time to recognize certain actions [7]. However, such features generate a varying number of interest points for different video sequences and generate sparse representations. In [9], HOG (histogram of oriented gradients) features were extracted from the MHIs (motion history images) that were classified through an SVM (support vector machine) classifier for action recognition. However, the major limitation of the HOG-based representation appears in action recognition when actors perform a similar action with different poses, e.g. bending left, right, or in forward positions. The primary reason for performance deficit is that HOG-based representations are not rotation-invariant and even a slight change in pose results in different HOG representations. Later, motion trajectory extractors were proposed in [12, 13] to encode spatiotemporal action representations. These extractors capture pixel variations within video sequences to discriminate actions. Among shallow handcrafted action representations, motion trajectories are proven to be one of the strongest descriptors for human action recognition [12]. In [14] HOG features were combined with the LBP (linear binary pattern) descriptor and action recognition was performed with an SVM classifier. However, the technique results in degraded performance due to the limited ability of HOG features to address various action poses. Moreover, the LBP is severely affected by noise and so classification performance significantly falls. On the other hand, silhouette extraction depends on accurate and robust segmentation techniques, which rely on time evolution.

In addition to representation-based HAR techniques, recently deep learning techniques have become popular, as deep learning networks can effectively handle nonlinear boundaries, thus reducing misclassification rates. Recent works on deep action representation (DAR) can be categorized into various groups including static frame learning [15], transformed frames [16], 3D-CNN [17], multistream networks [18, 19], and recurrent networks [20–22]. In static frame learning methods, video frames are used to learn action sequences without capturing the temporal information [15]. Another limitation of the static frame learning technique is the requirement for a fixed number of video frames with consistent resolution. Transformed frame learning techniques can be used to overcome such limitations by incorporating a temporal representation of action sequences [16]. These techniques establish a generative probabilistic model for learning higher-dimensional frame transformations and fuse the motion information from neighboring frames to capture temporal information. For the case of 3D-CNN models for HAR, convolutional operations are performed in a spatial and temporal dimension through 3D cubes that are constructed by assembling multiple simultaneous frames [17]. In [18] a two-stream network was used having two CNN networks, a spatial network and a temporal network, to extract a spatiotemporal action representation. In [19], optical flow was computed from successive video frames to train 2D CNN networks for human action recognition. However, the resulting method performed lower than handcrafted features due to the imprecise spatiotemporal representation of actions and less diversity of action datasets. The temporal information was extracted by transforming frames to a lower resolution to reduce computational complexity [20]. However, the recognition performance of 3D-CNN is equivalent to 2D-CNN operated on spatial video frames, which indicates

no significant performance improvement in motion information of 3D-CNN compared to 2D-CNN for HAR. Moreover, the reported results of 3D-CNN are also lower than for some handcrafted representations. Sun et al. [21] improved an LSTM (long short-term memory) network to learn a model for activity recognition by encoding movement information at each pixel of the video frame. In [22], a long-term recurrent convolutional network (LRCN) was presented for action recognition using LSTM on convolutional feature maps of RGB spatial frames or optical flow. Bilen et. al [23] used an LSTM network to generate a representation in the form of RGB images for action videos and encoded motion information across each pixel using rank pooling. Afterwards, the motion-encoded RGB representations were used for action classification.

Overall, deep learning-based action techniques usually require powerful computational resources and a large amount of data. Action representation from the fusion of handcrafted and learned approaches may achieve better prediction performance in HAR. Therefore, we have explored the fusion of spatiotemporal holistic motion templates and deep convolutional features to encode an actor's motion representation for action recognition. For evaluation, we tested our scheme on the benchmark activity datasets MuHAVi-Uncut [24] and iXMAS. We also evaluated our technique on a newer video dataset, Instructor Activity Video-1 (IAVID-1). Our contributions can be summarized as follows:

- We propose a deep temporal motion descriptor (DTMD) for human action recognition using motion history images and a 2D CNN (convolutional neural network). DTMD utilizes the CNN model to capture spatiotemporal deep representations from MHI using a backpropagation algorithm. The proposed DTMD outperforms recent methods in silhouette-based activity recognition.
- Secondly, DTMD is capable of recognizing human actions in a multiview scenario. Its spatiotemporal representation of actions improves performance under occlusion, scale, temporal variation, and viewpoint variation. This contribution is explored in the experimental section using the MuHAVI-Uncut and IXMAS multiview action recognition datasets.
- DTMD reduces the computational complexity of human action recognition as all the video frames are compactly represented by a single holistic motion template. Then these motion templates are used for DTMD computation.
- To evaluate the performance of the proposed approach, we perform a case study to recognize instructor actions within the lecture room using DTMD. The work illustrates how important it is to develop an autofeedback mechanism within lecture rooms to improve the quality of lecture delivery within academic institutes. However, this goal cannot be achieved until and unless instructor actions are recognized. DTMD successfully recognizes the eight basic actions of the instructor within the classroom.

The rest of the paper is organized as follows. Section 2 introduces the proposed approach for human action recognition. Then, in Section 3, experimental results and findings are discussed. Section 4 concludes the paper.

## 2. Deep temporal motion descriptor (DTMD) for action recognition

For training, we formed a set of videos  $V$  along with respective action labels  $L$  to build a model for action recognition using the DTMD descriptor. The  $DTMD_{v_t}$  descriptor gathers spatiotemporal information to describe the action class, where  $v_t$  are the total training videos (Figure 1). There are four main steps for  $DTMD_{v_t}$  generation: silhouette extraction and refinement, motion information gathering in the spatiotemporal

template, deep spatiotemporal representation of the actor’s motion template, and, finally, recognition of human action classes using a softmax classifier as illustrated in Figure 1. The following subsections describe the methodology in more detail.

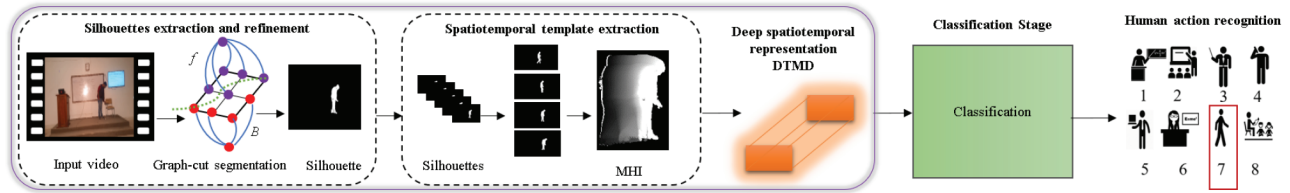


Figure 1. Architecture diagram of the proposed methodology.

### 2.1. Spatiotemporal template computation

The actor silhouettes are processed to form actor motion templates by applying the holistic motion history images (MHIs) technique. MHIs [25] are computed from actor silhouettes  $f$  to produce a spatiotemporal information template. The MHI is computed using Eq. (1), where MHI is a spatiotemporal template of all the actor’s silhouette frames  $f$  in the video, representing the object of interest, in this case an actor at time  $t$  and location  $(x,y)$  [9], as described in Eq.(1):

$$MHI = \begin{cases} \tau & \text{if } f = 1 \\ \max(0, (0, f - 1)) & \text{otherwise.} \end{cases} \quad (1)$$

Here,  $\tau$  is the total number of frames to generate the MHI for each action sequence. The benefit of using MHI is to reduce the spatial and computational complexity of an action, as the entire human action video is represented by a 2D single MHI image. The MHIs are centered and resized into fixed dimensions to eliminate redundant background information and make them compatible for the CNN as input. MHI template intensities are normalized within the maximum range to reduce more computations. Normalization also helps overcome undesirable factors like nonuniform illumination and contrast variations [9].

These motion templates are not invariant to spatial location or viewpoint. Therefore, as the prime motive of the action recognition system is to identify an action irrespective of spatial location, normalization is applied to create correspondences between training and testing action templates [11]. Scale and spatial location constraints are minimized through centering the motion template with respect to the maximum occupied area and wrapping the silhouettes to a predefined  $\tau$  template. In this way, the challenges of viewpoint, spatial location, and scale are reduced through the spatiotemporal action template.

### 2.2. Deep representation of the actor’s movement

The next step is the computation of deep features to represent each action class using CNN models. In CNN networks, the input layer receives the MHIs as input data and passes them to the next convolutional layer (conv layer). As the conv layer performs convolution of input data in the smaller region with weights to generate a feature map, the MHI is rescaled to  $N \times N$  dimensions to make it compatible with the CNN and generating spatiotemporal DTMD description. The intuition of using the actor’s motion templates or MHIs for action representation is derived from the hypothesis that motion templates are not significantly affected by imperfect

segmentation of silhouettes, being less sensitive to noisy artifacts, shadows, and occluded body parts. MHIs are robust to variations in pose due to changes in direction of motion like walking left and walking right; therefore, MHIs are capable of recognizing the same actions in different directions, viewpoints, and scales [9]. AlexNet [26], VGG16 [27], VGG-19 [27], Squeezenet [28], GoogleNet [29], and ResNet [30] have been used here for DTMD extraction. The last layer of the network is a soft-max classification layer, which performs the task of action recognition as shown in Figure 1.

### 3. Experiments

The proposed approach has been evaluated on the RGB IAVID-I, MuHAVI-Uncut, and iXMAS datasets, which are single and multiview datasets captured from static cameras. IXMAS and MuHAVI-Uncut have been chosen for evaluation of DTMD because the proposed approach involves a segmented actor's silhouettes (provided by these standard datasets) with a static background. A set of experiments has been performed to evaluate the performance of DTMD using the following steps:

- Examine the impact of different types of CNN architectures on DTMD for HAR.
- Quantitative analysis.
- Case study: Instructor action recognition in lecture room scenario using DTMD.
- Comparison of the proposed approach with other state-of-the-art methods.

#### 3.1. Datasets

New (IAVID-I) and benchmark (MuHAVI-Uncut, IXMAS) action recognition datasets are used to evaluate DTMD. The main characteristics of the used datasets are listed below.

##### 3.1.1. IAVID-1

We have constructed the IAVID-1 dataset to evaluate the proposed scheme. Twelve actors participated in data recordings in a realistic lecture room environment. There are 100 videos having  $1088 \times 1920$  high-resolution 24-bit RGB videos. Twelve actors performed the 8 instructor actions and approximately 12 instances of each action class are present in the dataset. There are eight actions in this dataset, i.e. interacting or idle, pointing towards the board, pointing towards the screen, using a mobile phone, using a laptop, sitting, walking, and writing on the board. IAVID-1 is a challenging dataset as it contains illumination and contrast variation, moving objects like an electric fan, and multimedia slide transitions.

##### 3.1.2. MuHAVi Uncut

The MuHAVI-Uncut dataset is a multi-action recognition dataset acquired from multiple cameras. It constitutes 17 activities performed by 14 actors at multiple durations. Eight CCTV cameras were mounted at  $45^\circ$  view difference to capture an action sequence. The MuHAVI-Uncut dataset is a large video dataset (2898 videos) and a single actor's silhouettes are segmented.

##### 3.1.3. IXMAS

INRIA Xmas Motion Acquisition Sequence (IXMAS) is a multiview dataset that constitutes twelve human actions performed by twelve actors. DTMD requires an actor's silhouettes to form motion templates since better silhouettes form better MHIs; therefore, MuHAVI-Uncut and iXMAS are the most suitable datasets for

evaluation of DTMD. Another benefit for DTMD's evaluation is that MuHAVI-Uncut and iXMAS allow us to examine the performance of action prediction at multiview settings and all the actions in these datasets are performed by a single actor with a static background.

### 3.2. Implementation details

In this section, we describe the optimal hyperparameters arrived at. The learning rate is set to 0.0001, while the batch size is set to 50 to train CNN models with 200 epochs using DTMD action representation. The batch size is set to 50 because there are 50 MHI training samples to optimize gradient loss and update the weight of pretrained CNNs using a stochastic gradient descent algorithm in each epoch within 12 iterations for HAR. To evaluate the performance of DTMD we have used cross-validation splits, leave one actor out (LOAO), leave one camera out (LOCO), and leave one sequence out (LOSO) validation schemes for HAR.

### 3.3. The impact of the type of CNN architecture on DTMD for HAR

To examine the performance of the proposed technique with respect to different types of CNNs, various serial and directed acyclic graph (DAG)-based CNN networks (i.e. AlexNet [26], VGG16 [27], VGG-19 [27], Squeezenet [28], GoogleNet [29], and ResNet [30]) have been used.

Sequential CNN architecture designs such as Alexnet and VGG follow a hierarchical sequential convolutional layer for computation of DTMD using a high-level representation of the input classes. On the other hand, CNN models based on DAGs, such as Squeezenet, GoogleNet, and ResNet, use multiple CNN parallel layers to capture lower, middle, and higher levels for DTMD extraction.

For simplicity, the DTMDs obtained from AlexNet, VGG16, VGG-19, Squeezenet, GoogleNet, and ResNet are denoted as DTMD11, DTMD22, DTMD38, DTMD44, DTMD68, DTMD99, and DTMD175, respectively, because DTMD11 is computed from the 11 layers of AlexNet, DTMD22 from the 22 layers of Alexnet, DTMD38 from the 38 layers of VGG16, DTMD44 from the 44 layers of VGG19, and so on. The benefits of deep network transfer learning have been exploited by fine-tuning the pretrained models for DTMD generation. The performance of the proposed DTMD descriptor is influenced by the quality of silhouettes, the span of activities, and the type of CNN architecture, as the quality of the actor's silhouettes and span of actions are the essential factors for generating class-specific MHIs. The resultant MHIs generated from short-duration activities of visually similar actions or imprecisely segmented silhouettes sometimes confuse the prediction of action recognition. Moreover, the type of CNN architecture for DTMD generation also affects the performance of action prediction.

It can be observed from Table 1 that DTMDs computed from sequential CNN models for HAR are generally better than DTMDs computed from DAG-based CNN models, as the smaller sequential CNN models hold a smaller number of hyperparameters helping the transfer learning training process, especially with modest amounts of data [28]. On the other hand, DAG-based CNN models used for DTMD generation accumulate lower, middle, and higher levels of representation of MHIs, reducing granularity for precise action prediction. Among sequential CNN models, the DTMD extracted from Alexnet outperformed on IAVID-1, MuHAVi-Uncut, and IXMAS datasets. The reason behind the good performance of the Alexnet model within sequential CNN models is its capability to resolve nonsaturating activation of nodes and apply overlapping pooling, local response normalization, and dropout regularization. After evaluation of the optimal model for DTMD computation, the rest of the experiments will consider DTMD22 extracted with Alexnet.

**Table 1.** Impact of CNN architecture on DTMD performance.

CNN type	DTMD	CNN models	Parameters	Prediction accuracy		
				IAVID	MuHAVi	IXMAS
				2/3 splits	LOAO	LOAO
Serial	DTMD11	Alexnet	60 K	75.09	70.90	63.78
	DTMD22	Alexnet	60 M	78.13	89.66	70.70
	DTMD38	VGG16	110 M	62.50	79.85	68.05
	DTMD44	VGG19	138 M	78.13	83.94	68.98
DAG	DTMD68	SqueezeNet	421,098	50.73	65.32	63.77
	DTMD99	GoogLeNet	4 M	65.65	71.44	65.56
	DTMD175	ResNet50	25M	71.87	75.75	67.39
	DTMD205	ResNet101	44.5M	72.81	78.45	68.97

**Table 2.** DTMD performance on MuHAVi-Uncut, IXMAS, and IAVID-1 using LOAO validation scheme.

Dataset	DTMD dimension	Accuracy
MuHAVi-Uncut		89.66
IXMAS	4096	70.70
IAVID-1		68.19

### 3.4. Quantitative evaluation of DTMD

#### 3.4.1. DTMD performance for person-invariant HAR

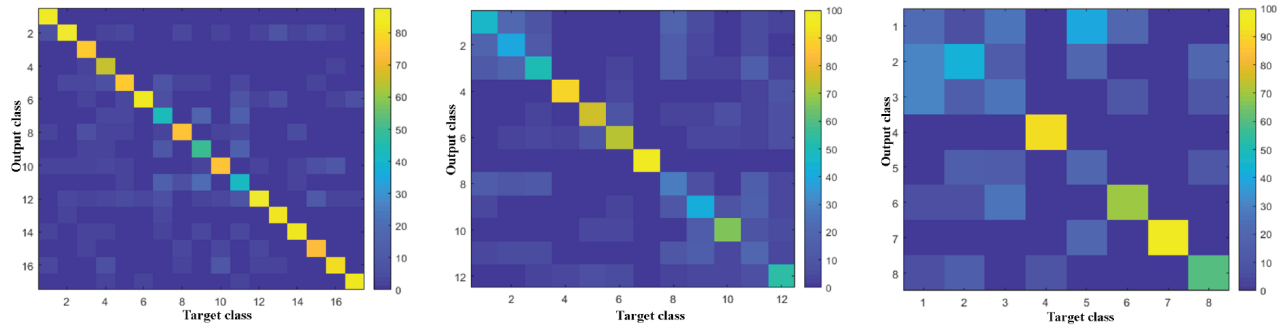
The performance of DTMD is examined using the leave-one-actor-out (LOAO) validation scheme, using training video data of all the actors except one actor and using the remaining actor's video sequences for testing. The entire process is repeated for all the actors and the average accuracy is reported. In MuHAVi-Uncut, IXMAS, and IAVID-1, there are fourteen, ten, and twelve actors participating in the data acquisition process and these actors perform multiple actions several times.

The observations recorded in Table 2 show that DTMD overcomes the high intra-class variation that rises due to multiple actors performing the same action multiple times and accurately recognizes human activities. The MuHAVi-Uncut dataset has seventeen actions performed by fourteen actors multiple times, resulting in an action prediction rate of 89.66%, which is higher than state-of-the-art methods. The IXMAS dataset has twelve activities performed by twelve actors multiple times and the action prediction rate is 70.70%. Similarly, IAVID-1 has eight activities performed by twelve actors multiple times and the action prediction rate is 68.19%. The per-class accuracies for LOAO validation of MuHAVi-Uncut and IXMAS are illustrated in Figure 2.

#### 3.4.2. DTMD performance for view-invariant HAR

The leave-one-camera-out (LOCO) validation scheme is used to estimate the stability of action recognition algorithm using video representations from multiple camera views. In LOCO, action videos from multiple cameras except one camera view are used for training and action videos from the remaining camera are used for testing. The process is repeated for all camera views and average prediction accuracies are reported in Table 3.



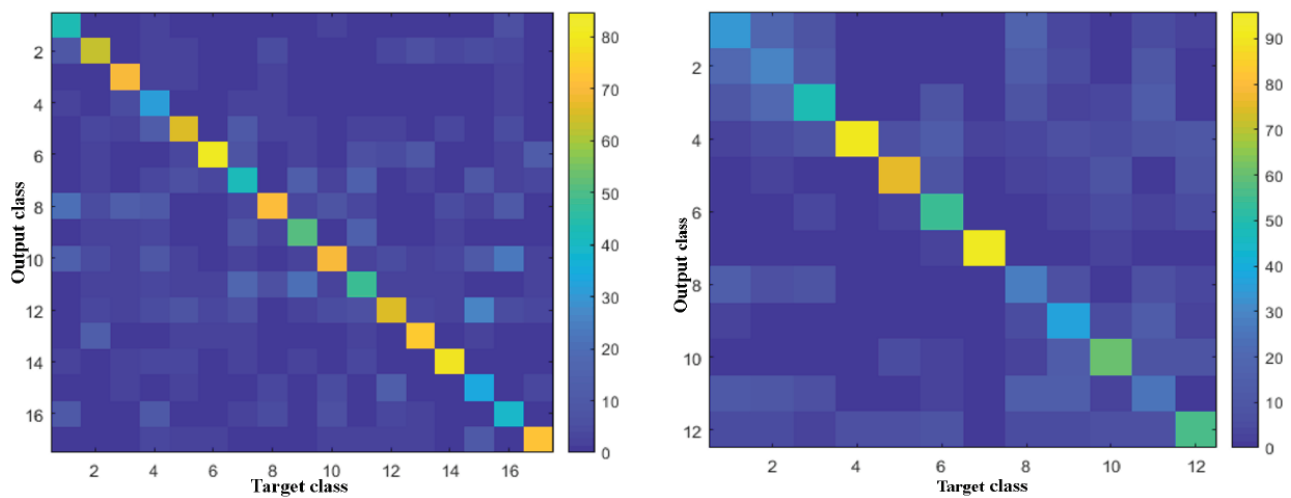


**Figure 2.** Confusion matrices for LOAO validation scheme on MuHAVi-Uncut, IXMAS, and IAVID-1 datasets using DTMD descriptor.

**Table 3.** Performance evaluation of DTMD using LOCO validation scheme on MuHAVi-Uncut and IXMAS datasets.

Dataset	DTMD dimension	Accuracy
MuHAVi-Uncut	4096	52.42%
IXMAS	4096	60.80%

The MuHAVi-Uncut and IXMAS datasets are captured from eight and five camera views. The observations recorded in Table 3 illustrate that DTMD does a reasonable job of dealing with the high intraclass variation that arises due to multiple views of the same action. DTMD is able to recognize human actions in the multiview setting because the deep spatiotemporal representation of action templates strengthens the learning model to predict actions from missing views. The MuHAVi-Uncut dataset’s seventeen actions are predicted by 52.42%, which is higher than other state-of-the-art methods. The IXMAS dataset has twelve activities captured from five camera views that are predicted by 60.8%. The confusion matrices of MuHAVi-Uncut and IXMAS are presented in Figure 3.



**Figure 3.** Confusion matrices for LOCO validation scheme on MuHAVi-Uncut (left) and IXMAS dataset (right) using DTMD descriptor.



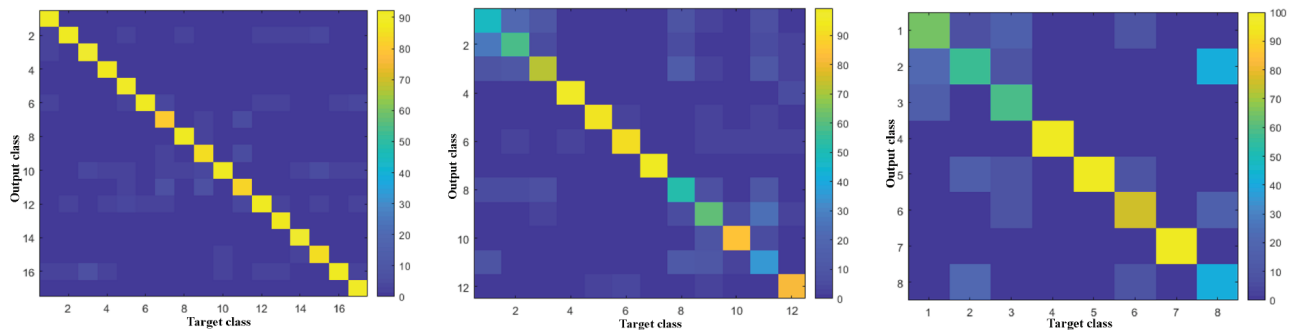
**3.4.3. DTMD performance for large training data**

In the leave one sequence out (LOSO) validation scheme, the DTMD descriptions for all the action sequences (except one) are used to generate the training model. The remaining action video sequence is used as a test sample. To determine the performance of the system, the process is repeated for all possible combinations and average accuracy is calculated, as illustrated in Table 4.

**Table 4.** Performance evaluation of DTMD using LOSO validation scheme on MuHAVi-Uncut, IXMAS, and IAVID-1 datasets.

Dataset	DTMD dimension	Accuracy
MuHAVi-Uncut		97.07%
IXMAS	4096	70.83%
IAVID-1		71.42%

To examine the behavior of DTMD on a large training dataset, we evaluated the LOSO validation scheme on MuHAVi-Uncut, IXMAS, and IAVID-1 datasets. The MuHAVi-Uncut, IXMAS, and IAVID-1 datasets have 2898, 1800, and 100 action video samples for model learning, respectively. Table 4 shows that DTMD is capable of recognizing human activities for a large amount of training data due to higher-dimensional spatiotemporal representation. The MuHAVi-Uncut dataset’s seventeen actions for approximately 3000 videos samples are predicted by 97.07%, which is higher than state-of-the-art methods. The IXMAS dataset has twelve activities for 1800 video samples, which are predicted by 79.83%. The confusion matrices of MuHAVi-Uncut, IXMAS, and IAVID-1 are presented in Figure 4, which portrays the stability of DTMD for HAR on large training data. The per-class accuracies for LOSO validation of MuHAVi-Uncut and IXMAS are illustrated in confusion matrices in Figure 4.

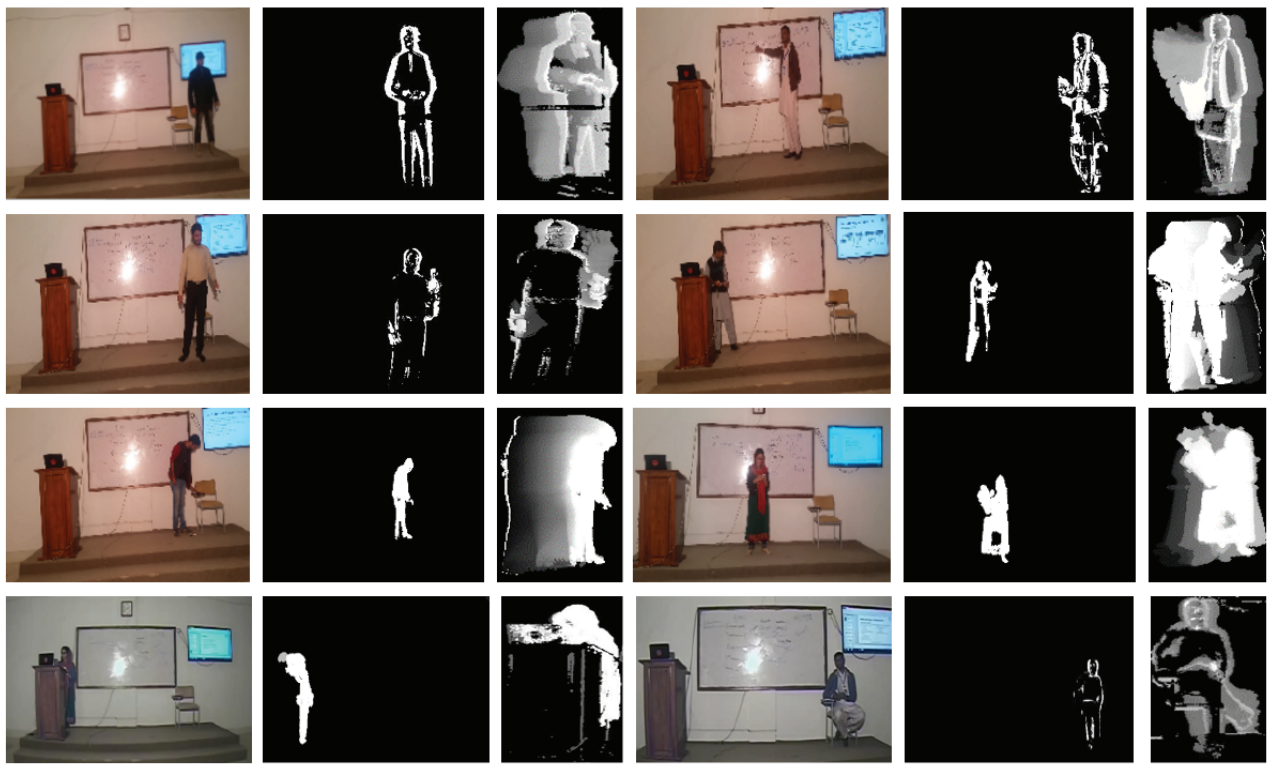


**Figure 4.** The performance of DTMD using LOSO validation scheme on MuHAVi-Uncut, IXMAS, and IAVID-I dataset.

**3.5. Case study: Instructor action recognition in lecture room scenario using DTMD**

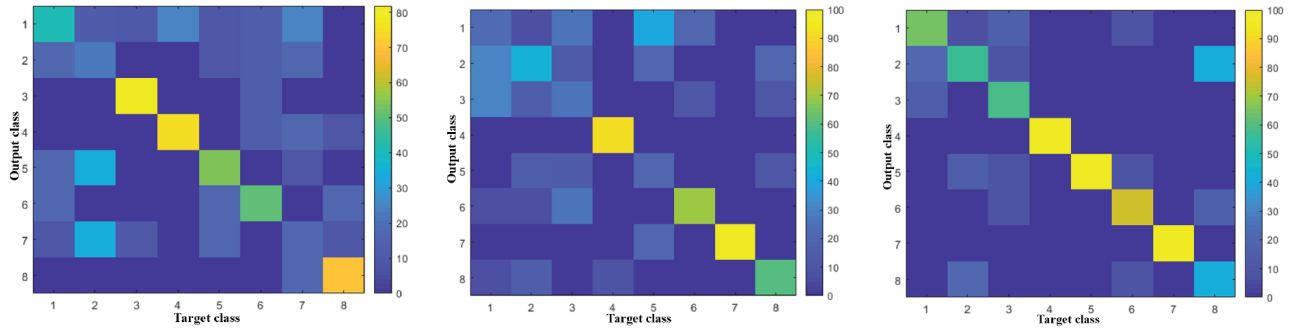
HAR techniques can be applied in applications for societal growth, like in academic institutes, for estimating the effectiveness of lecture delivery within the lecture room. The classroom visual information can be utilized to provide genuine feedback for the instructor to self-evaluate and plan for subsequent lectures. There is a need to utilize modern computer vision technology to evaluate instructor behavior within the lecture room for improving the quality of lecture delivery. Manual self-evaluation is tedious for teachers, as remembering all the

details and shortcomings in lecture delivery is not possible or involves time analyzing video recordings. Peer-evaluation is also useful but time-consuming and may be biased. Therefore, automation using video analysis is useful for teachers and instructors [31, 32]. In this case study, input lecture video streams are captured in an uncontrolled realistic environment and the instructor action recognition methodology begins with foreground extraction and the computing of MHIs for each video. Spatiotemporal characteristics are described using DTMD. Although resulting silhouettes of IAVID-1 obtained from graph-cut segmentation do not completely capture the instructor as the movement may become semistatic during lecture delivery, causing the partial segmentation of the instructor’s silhouettes, as shown in Figure 5, the working hypothesis is that the MHIs generated from these silhouettes are capable of representing the instructor’s action, as this overcomes the issues of undersegmentation of human silhouettes. The hypothesis is later tested in experiments.



**Figure 5.** Instructor’s actions in IAVID-1 within the classroom in columns I and IV, and relevant silhouette representations in columns II and V and resulting MHIs in columns III and VI.

The per-class action recognition accuracy for cross-validation, LOAO, and LOSO validation schemes are shown in Figure 6. From Figure 6, a confusion matrix of 70:30 cross-validation, it can be observed that five of the instructor’s actions, ‘Interacting or idle’, ‘Pointing to board or screen’, ‘Using laptop’, ‘Using phone’, and ‘Walking’, are classified accurately, whereas ‘Pointing to students’, ‘Writing on board’, and ‘Sitting’ are misclassified with ‘Using laptop’, ‘Using phone’, ‘Pointing towards board screen’, and ‘Interacting or idle’, respectively. The reason for misclassification of these actions is the existence of some visual similarities in their poses. Increasing the training sample will improve the recognition capability of DTMD. However, average action recognition accuracy is 78.13% for 70:30 cross-validation, as shown in Table 5.



**Figure 6.** DTMD performance on IAVID-1 dataset for instructor analysis on 70-30 split cross-validation, LOAO, and LOSO validation scheme, respectively.

**Table 5.** Performance of DTMD on IAVID-1 for instructor actions recognition.

Validation schemes	DTMD dimension	Accuracy
Cross-validation		78.13%
LOAO	4096	68.19%
LOSO		71.42%

From the LOAO confusion matrix of IAVID-1 (Figure 6), it can be observed that the class *pointing towards board or screen* has the lowest prediction accuracy using the LOAO scheme. This is because a comprehensive representation of actions from MHIs requires a large ratio of motion variation. Consequently, actions that have a small span of motion are not effectively predicted by the softmax classifier due to visual similarity with other action classes like *sitting*, *using a laptop*, *using the phone*, and *walking*. This issue could be resolved by increasing the training data. Average recognition accuracy is 68.19% for the LOAO validation scheme, as shown in Table 5. The confusion matrix for the LOSO validation scheme on IAVID-1 is given in Figure 6 for actions of *Interacting or idle*, *Pointing towards board or screen*, *Pointing to students*, *Sitting*, *Using laptop*, *Using phone*, *Walking*, and *Writing on board* as 75%, 91.7%, 57.15%, 64.35%, 100%, 100%, 58.3%, and 25%, respectively. It is noted (Figure 6) that the class of *Writing on board or screen* has the minimum recognition accuracy of 25% for the LOSO validation scheme while *Pointing to board*, *Using laptop*, and *Using phone* achieved maximum recognition rates of 91.7%, 100%, and 100%, respectively. Overall, the average recognition accuracy is 71.42% for the LOSO validation scheme, as shown in Table 5.

### 3.6. Comparison of the proposed DTMD descriptor with state-of-the-art HAR techniques

In this section, the proposed approach is compared with relevant state-of-the-art approaches. We have implemented the ones listed in Table 6. Action prediction accuracy of representation-based (handcrafted) methods [9, 14] is lower than that of the proposed scheme due to a low-dimension representation of motion templates. A holistic motion template with handcrafted features was computed in [9, 14], but the fusion of the holistic motion template and deep convolutional neural network enhances the performance of the proposed system due to the higher level action representation, as shown in Table 6. Comparing deep learning methods C3D [17] and LRCN [22] with handcrafted HAR techniques has also shown that the former performed better. However, the performance of deep learning methods is lower than that of the proposed DTMD due to the fact that C3D

performs similarly to action recognition using 2D-CNN at frame level, and LRCN used optical flow to encode action motion information that is sensitive to noise and fails to encode long-term temporal information. Conversely, the DTMD, due to the fusion of a motion template with numerous deep spatiotemporal features, shows better performance.

**Table 6.** Performance DTMD22 on action dataset.

Dataset	Validation scheme	Method	Accuracy
IAVID-I	Cross-validation	Proposed technique	78.12
		Motion template and HOG description [9]	63.5
		Motion template and HOG-LBP descriptor [14]	50
		C3D features and CNN [17]	70
		LRCN [22]	70.05
MuHAVi-Uncut	LOAO	Proposed technique	89.66
		Motion template and HOG description [9]	84.1
		Observable Markov model [11]	83.9
		C3D features and CNN [17]	84.56
		LRCN [22]	86.87
	LOCO	Proposed technique	73.29
		Motion template and HOG description [9]	52.2
		C3D features and CNN [17]	72.98
		LRCN [22]	71.23
	LOSO	Proposed technique	97
		Motion template and HOG description [9]	96.6
		C3D features and CNN [17]	95.78
		LRCN [22]	90.78
IXMAS	LOSO	Proposed technique	70.83
		The sequence of the key poses [10]	85.9
		C3D features and CNN [17]	71.09
		Substructure and boundary modeling [33]	76.5
	LOCO	Proposed technique	71.92
		C3D features and CNN [17]	70.09
		LRCN [22]	69.43
		Spatiotemporal visual words to learn the SVM [34]	57.3

The proposed technique is robust in addressing the challenges of action recognition in multiview settings due to the higher-dimensional representation of actions. The problem of viewpoint dependency for action recognition was addressed in [11] using 3D human body modeling for action recognition. The major drawback of this method is that various fixed calibrated cameras are required for precise training and testing, which is difficult to obtain. The observed results validate DTMD’s view-invariant properties. The experiments on the MuHAVi-Uncut dataset revealed that in contrast to [33, 34], DTMD produces view-invariant human action representation that does not require multiview feature fusion or calibrated cameras; hence, a single viewpoint video is used for testing and the remaining views are used for training in the LOCO validation scheme. Although most practical applications use single cameras, view invariance is important to minimize retraining when the views/cameras are changed.

The performance of DTMD on the iXMAS dataset has been evaluated, as well. However, its prediction accuracy is not outstanding. This is due to the fact that the iXMAS dataset is challenging and has inconsistent angular positions of actors with respect to cameras views. As in IAVID-1, iXMAS action classes are fairly visually similar, e.g. watching watch or stretching head and folding arms. This variation in the actor's pose causes no significant visual differences in MHIs within each view and this is a major reason for lower action prediction accuracy. Nevertheless, still, the proposed technique recognizes actions of iXMAS to some extent, as shown in Table 6. As shown in the same table, the method outperformed other state-of-the-art action recognition approaches in terms of accuracy.

It is notable from Table 6 that DTMD is capable of recognizing actions in single and multiview settings and robust to recognizing actions performed by multiple actors. DTMD improved the baseline results on the MuHAVI-Uncut dataset in the LOCO scheme by 21.09%, the LOAO scheme by approximately 5.56%, and the LOSO scheme by approximately 0.4%, respectively. Thus, the improvement in performance validates the reliability of DTMD for multiple views and multiple actors in action recognition tasks. Moreover, improvements in the recognition rate on MuHAVI-Uncut also indicate the practical bearing of the DTMD descriptor for action recognition tasks in another application domain.

#### 4. Conclusion

This paper has proposed a DTMD descriptor that is a deep spatiotemporal representation of action sequences for human action recognition. The method consists of the actor's silhouettes extraction, followed by the actor's motion encoding in the form of motion history templates then described through a deep feature representation. DTMD is capable of precise recognition of human actions performed by multiple actors in single or multiview settings. In comparison to some handcrafted features and deep learning HAR approaches, DTMD has shown better performance on IAVID-1, as well as on standard action recognition datasets, due to the deep spatiotemporal representation of action templates. In IAVID-1, DTMD successfully recognizes the eight basic actions of the instructor, which indicates the practical application of the DTMD descriptor for action recognition tasks in another domain. In the future, we plan to construct a larger instructor activity dataset to promote research in smart applications for educational institutes.

#### References

- [1] Chua JL, Chang YC, Lim WK. A simple vision-based fall detection technique for indoor video surveillance. *Signal, Image and Video Processing* 2015; 9 (3): 623-633. doi: 10.1007/s11760-013-0493-7
- [2] Ibrahim MS, Muralidharan S, Deng Z, Vahdat A, Mori G. A hierarchical deep temporal model for group activity recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016; Las Vegas, NV, USA; 2016*. pp. 1971-1980.
- [3] Zhang K, Grauman K, Sha F. Retrospective encoders for video summarization. In: *Proceedings of the European Conference on Computer Vision 2018; Munich, Germany; 2018*. pp. 383-399.
- [4] Yousaf MH, Habib HA, Azhar K. Fuzzy classification of instructor's morphological features for autonomous lecture recording system. *Information - An International Interdisciplinary Journal* 2013; 16: 6367-6382. doi: 10.1007/3-540-44967-1\_67
- [5] Yousaf MH, Azhar K, Sial HA. A novel vision based Approach for instructors performance and behavior analysis. In: *2015 International Conference on Communications, Signal Processing, and their Applications 2015; UAE; 2015*. pp. 1-6.

- [6] Zhu F, Shao L, Xie J, Fang Y. From handcrafted to learned representations for human action recognition: a survey. *Image and Vision Computing* 2016; 55: 42-52. doi: 10.1016/j.imavis.2016.06.007
- [7] Laptev I. On space-time interest points. *International Journal of Computer Vision* 2005; 64 (2-3): 107-123.
- [8] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2010*; San Francisco, CA, USA; 2010. pp. 2046-2053.
- [9] Murtaza F, Yousaf MH, Velastin SA. Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description. *IET Computer Vision* 2016; 10 (7): 758-767. doi: 10.1049/iet-cvi.2015.0416
- [10] Chaaaraoui AA, Climent-Pérez P, Flórez-Revuelta F. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters* 2013; 34 (15): 1799-1807. doi: 10.1016/j.patrec.2013.01.021
- [11] Orrite C, Rodriguez M, Herrero E, Rogez G, Velastin SA. Automatic segmentation and recognition of human actions in monocular sequences. In: *22nd International Conference on Pattern Recognition 2014*; Stockholm, Sweden; 2014. pp. 4218-4223.
- [12] Wang H, Schmid C. Action recognition with improved trajectories. In: *IEEE International Conference on Computer Vision 2013*; Sydney, Australia; 2013. pp. 3551-3558.
- [13] Wang Y, Mori G. Human action recognition by semilattent topic models. *IEEE Transactions on Pattern Analysis And Machine Intelligence* 2009; 31 (10): 1762-1774. doi: 10.1109/tpami.2009.43
- [14] Ahad MA, Islam MN, Jahan I. Action recognition based on binary patterns of action-history and histogram of oriented gradient. *Journal on Multimodal User Interfaces* 2016; 10 (4): 335-344. doi: 10.1007/s12193-016-0229-4
- [15] Ning F, Delhomme D, LeCun Y, Piano F, Bottou L et al. Toward automatic phenotyping of developing embryos from videos. *IEEE Transactions on Image Processing* 2005; 14: 1360-1371. doi: 10.1109/tip.2005.852470
- [16] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Computation* 2006; 18 (7): 1527-1554. doi: 10.1109/tip.2005.852470
- [17] Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2012; 35 (1): 221-231.
- [18] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: *Advances in Neural Information Processing Systems 2014*; Montreal, Canada; 2014. pp. 568-576.
- [19] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R et al. Large-scale video classification with convolutional neural networks. In: *IEEE Conference on Computer Vision and Pattern Recognition 2014*; Columbus, OH, USA; 2014. pp. 1725-1732. doi: 10.1109/cvpr.2014.223
- [20] Memisevic R, Hinton G. Unsupervised learning of image transformations. In: *IEEE Conference on Computer Vision and Pattern Recognition 2007*; Minneapolis, MN, USA; 2007. pp. 1-8. doi: 10.1109/cvpr.2007.383036
- [21] Sun L, Jia K, Chen K, Yeung DY, Shi BE et al. Lattice long short-term memory for human action recognition. In: *IEEE International Conference on Computer Vision 2017*; Venice, Italy; 2017. pp. 2147-2156.
- [22] Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S et al. Long-term recurrent convolutional networks for visual recognition and description. In: *IEEE Conference on Computer Vision and Pattern Recognition 2015*; Boston, MA, USA; 2015. pp. 2625-2634.
- [23] Bilen H, Fernando B, Gavves E, Vedaldi A, Gould S. Dynamic image networks for action recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition 2016*; Las Vegas, NV, USA; 2016. pp. 3034-3042.
- [24] Sepulveda J, Velastin SA. F1 score assessment of Gaussian mixture background subtraction algorithms using the MuHAVi dataset. In: *6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15) 2015*; London, UK; 2015. pp. 6-8.
- [25] Bobick AF, Davis JW. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 2001 (3): 257-267. doi: 10.1109/cvpr.1997.609439

- [26] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 2012; Lake Tahoe, NV, USA; 2012. pp. 1097-1105. doi: 10.1145/3065386
- [27] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint, arXiv:1409.1556, 2014. doi: 10.14257/astl.2016.140.36
- [28] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. arXiv preprint, arXiv: 1602.07360, 2016.
- [29] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S et al. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition 2015; Boston, MA, USA; 2015. pp. 1-9. doi: 10.1109/cvpr.2015.7298594
- [30] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016; Las Vegas, NV, USA; 2016. pp. 770-778. doi: 10.1109/cvpr.2016.90
- [31] Raza A, Yousaf MH, Sial HA, Raja G. HMM-based scheme for smart instructor activity recognition in a lecture room environment. SmartCR 2015; 5 (6): 578-590. doi: 10.6029/smartcr.2015.06.008
- [32] Nida N, Yousaf MH, Irtaza A, Velastin SA. Bag of deep features for instructor activity recognition in lecture room. In Proceedings of the Springer International Conference on Multimedia Modeling 2019; Athens, Greece; 2019. pp. 481-492.
- [33] Wang Z, Wang J, Xiao J, Lin KH, Huang T. Substructure and boundary modeling for continuous action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition 2012; Rhode Island, USA; 2012. pp. 1330-1337. IEEE. doi: 10.1109/cvpr.2012.6247818
- [34] Huang CH, Yeh YR, Wang YC. Recognizing actions across cameras by exploring the correlated subspace. In: European Conference on Computer Vision 2012; Berlin, Germany; 2012. pp. 342-351. doi: 10.1007/978-3-642-33863-2\_34