doi: 10.1093/comjnl/bxaa061

Fast Learning Through Deep Multi-Net CNN Model For Violence Recognition In Video Surveillance

Aqib Mumtaz, Allah Bux Sargano and Zulfiqar Habib*

Department of Computer Sciences, COMSATS University Islamabad, Lahore Campus, Lahore 54700, Pakistan

**Corresponding author: drzhabib@cuilahore.edu.pk*

The violence detection is mostly achieved through handcrafted feature descriptors, while some researchers have also employed deep learning-based representation models for violent activity recognition. Deep learning-based models have achieved encouraging results for fight activity recognition on benchmark data sets such as hockey and movies. However, these models have limitations in learning discriminating features for violence activity classification with abrupt camera motion. This research work investigated deep representation models using transfer learning for handling the issue of abrupt camera motion. Consequently, a novel deep multi-net (DMN) architecture based on AlexNet and GoogleNet is proposed for violence detection in videos. AlexNet and GoogleNet are top-ranked pre-trained models for image classification with distinct pre-learnt potential features. The fusion of these models can yield superior performance. The proposed DMN unleashed the integrated potential by concurrently coalescing both networks. The results confirmed that DMN outperformed state-of-the-art methods by learning finest discriminating features and achieved 99.82% and 100% accuracy on hockey and movies data sets, respectively. Moreover, DMN has faster learning capability i.e. 1.33 and 2.28 times faster than AlexNet and GoogleNet, which makes it an effective learning architecture on images and videos.

Keywords: deep CNN; multi-net; fast learning; violence recognition; video surveillance

Received 1 March 2019; Revised 29 January 2020; Accepted 22 April 2020
Handling editor: Suchi Bhandarkar

1. INTRODUCTION

Video surveillance is subjected to monitor surveillance cameras for suspected human behaviors and violent activities. For this purpose, hundreds and thousands of cameras are deployed within cities. Due to drastic increase in security cameras deployment, it is impossible these days to physically monitor all the deployed cameras. Rather, there is a significant requirement for developing automated video surveillance system, to track and monitor violent activities in an automated fashion, eliminating the need of manual surveillance. Automated surveillance system will help alarming the controlling authorities in case of emergencies, to take appropriate actions against the detected violence. Violence detection in videos is a prime step toward developing such systems, to distinguish normal human activities from abnormal/violent actions. Normal human activities are often categorized as routine life human behaviors, such as walking, jogging, running and hand waving [1, 2]. Whereas, violence is an unusual human action, such as fight activity happening between two or more people [3].

Over the past few years, computer vision community has shown keen interest in the task of video surveillance. Modern deep architectures have been proposed for the person re-identification problem by matching pedestrian across the disjoint camera views [4–6]. For human action recognition, researchers have suggested numerous methods to detect normal day human activities through video analysis; see surveys [7, 8]. However, little attention has paid toward human violence detection, until the availability of violent sequences specifically designed for fight activity detection. To achieve violence recognition, the authors have created two video data sets to differentiate violent/fight incidents from the normal events [9]. Before the production of these two data sets, most of the renowned data sets were particularly focused on detection of general human actions. Whereas, these data sets are primarily developed for violent/fight scenes detection, to assist

building precise automated surveillance systems, for indoor and outdoor environments.

In literature, human activity recognition (HAR) is first accomplished through traditional handcrafted feature descriptors, such as histogram of oriented gradient (HOG), scale-invariant feature transform (SIFT), Hessian3D, local binary pattern, etc. Later on, it is achieved through learning-based deep representations, such as convolutional neural networks (CNNs), 3D-CNN, CNN followed by recurrent neural network (RNNs) and spiking neural networks, etc.; see survey [10, 11].

Similar to human action recognition, for human violence detection, most of the existing approaches addressed this problem through handcrafted features descriptors, to differentiate fight sequences from normal frames. For this reason, since the introduction of the violent sequences, numerous approaches were applied on two violent data sets. These techniques constructed handcrafted feature detectors, such as space-time interest points (STIP), motion SIFT (MoSIFT), motion features and motion blobs performed on audio–visual analysis along blood and flame detection [9, 12–15]. Besides that, a few researches inquired deep representations such as 2D-CNN, 3D-CNN and C3D [16–18]. However, there is a scarcity in examining deep representations based models using transfer learning, to identify fight scenes in videos.

Deep representation due to self-reliant learning capacity is generally known as end-to-end learning. It has history starting from CNNs, an end-to-end learning model designed for handwritten digit classifications [19]. The deep representation architecture unleashed its hidden potential to computer vision community, when state-of-the-art AlexNet CNN architecture stood winner of ImageNet large scale visual recognition challenge (ILSVRC), due to its astonishing results for image classifications, trained on 15 million annotated images for 1000 categories [20]. Following AlexNet, in coming years, VGGNet [21], GoogleNet [22] and ResNet [23] were presented as competitive deep CNN (DCNN) models on ImageNet data set. These architectures are winner model of ImageNet ILSVRC, due to their remarkable accuracy for image classification task. However, in order to effectively train a deep network, a very large data set is required to learn most discriminating features. To combat the challenge of huge data requirement, a concept of transfer learning is adopted by many researchers as a hallmark strategy. In transfer learning, a CNN model pre-trained on specific data set, which has already learnt specific features for some specific task, can be transferred to be fine tuned for a new task, even to an entirely different domain [24]. Due to this powerful concept, researchers started using transfer learning for numerous tasks of images/videos classification and action detection. For instance, large-scale video classification for distinct videos [25] and HAR [26] is achieved through transfer learning. Although transfer learning has optimal strategies [27] [28], it has ability for learning generalized representations through fine tuning or as feature extractor machine for large or small size data sets [29]. However, in order to make an effective transfer learning, a pre-trained network has to be fine tuned on target data set to successfully perform the new task in target domain [30].

Furthermore, in deep representation models, 3D-CNN based methods employ both spatial and temporal information for action recognition, as compare to 2D-CNN. Hence, these methods are considered more robust and may produce more accurate results [31]. However, 3D-CNN has more parameters than 2D-CNN models. This makes 3D-CNN computationally complex and relatively hard to train. Moreover, 3D-CNN models are prevented to take advantage of ImageNet pre-training. Whereas, 2D-CNN models have key benefit of models pre-training on large-scale data sets such as ImageNet [32]. The 2D-CNN models due to pre-learnt feature representation on large-scale data set are effectively useful in transfer learning. Besides, the light weight 2D-CNN models are considered to be the suitable candidate for real-time applications in surveillance systems [33].

Recently, 3D-CNN and C3D models with spatio-temporal features have been evaluated for violence recognition. However, these models are unable to produce superseding results [16, 17]. Therefore, this research work revisited the 2D-CNN models using transfer learning as a base line strategy to recognize violent/fight actions on two benchmark videos data sets. Although deep learning approaches are successfully used for human action recognition, however, these techniques in coordination with transfer learning are not much considered for violence discovery.

The contribution of the paper is 2-folds. First, this research work has examined deep representations models using transfer learning for handling the issue of abrupt camera motion. For this purpose, AlexNet [20] and GoogleNet [22] are fine tuned on hockey and movies data sets. Second, a novel deep multi-net (DMN) architecture has been proposed with the fusion of two pre-trained models i.e. AlexNet and GoogleNet. This architecture has fast learning capability and significantly reduced training time on benchmark data sets. The results confirm that DMN architecture has produced state-of-the-art performance. The DMN has learning ability up to 2.28 times faster than AlexNet and GoogleNet achieving equivalent accuracies. The rest of the paper is organized as follows: related work, methodology, data sets, experiments, results and conclusion are presented in Sections 2, 3, 4, 5, 6 and 7, respectively.

2. RELATED WORK

Violence recognition in videos started with initial proposal of adopting the methodology using blood and flame detection, capturing the degrees of motion and exploiting audio–visual correlation for recognizing sounds features, for violent scenes detection [14]. Then, an audio analysis is performed to detect gun shots, explosions and car-breaking activity using hierarchal approach, modeling statistical characteristics of audio events, based upon hidden Markov models and Gaussian mixture models [15].

Giannakopoulos *et al.* [34] proposed violence detection based on audio features from the time and frequency domain, using support vector machine (SVM) classifier. Clarin *et al.* [35] proposed a system to classify extreme actions by identifying skin and blood, featuring through self-organizing map and motion intensity of pixels for violence classification of scenes. Zajdel *et al.* [36] presented CASSANDRA system detecting motion features in videos and scream like cues in audio exploiting the complimentary nature of video and audio, using dynamic Bayesian network for discovering aggressive human behavior in public areas. Gong *et al.* [37] presented a method for classifying low-level integrated visual and auditory features using semi-supervised classifier in coordination with high-level audio features detection, to identify potential violent contents. Chen *et al.* [38] used spatio-temporal video cubes and local binary motion descriptor for recognizing aggressive behaviors in video. Lin and Wang [39] proposed an exploited weekly supervised method classifying audio violence combined with co-training of motion, blood and explosion video classifier to detect violent shot scenes in movies. Moreover, Giannakopoulos *et al.* [40] proposed an approach to identify violent action in movies by analyzing audio–visual features information using statistics, average motion and motion orientation variance features, followed by *k*-nearest neighbors (KNN) classifier to figure out whether given action scene is violent. Chen *et al.* [41] presented an approach based upon motion detection of faces and blood presence. Hassner *et al.* proposed solution for the problem of monitoring and detecting crowded events for outbreaks of violence. The approach used statistical method of flow vector for short frame video sequences followed by a linear SVM to detect violent or non-violent event in video surveillance [42].

Bermejo *et al.* exhibited unique contribution in the domain of human violence detection based upon generic action recognition approaches using MoSIFT feature descriptor, demonstrating an encouraging 90% accuracy results. This research work revealed two potential data sets designed specifically for the violent/fight scenes detection based on real-time videos 'hockey data set' and 'movies data set' [9]. Since then, these data sets are established benchmark source of motivation for research community to adhere challenges in violence recognition. Thereafter, kernel density estimation was exploited to obtain feature selection on MoSIFT descriptor with sparse coding. This approach achieved 94.3% accuracy on hockey data set [43]. Motion binary patterns (MBPs) were originally adopted for human action recognition [44]; afterward, MBPs were also used for violence action detection for performing motion intensity variation estimation using local changes in pixel intensity.

Recently, an author measured fuzzy region pattern, emerges in image frames due to sudden abrupt motion (fight) patterns, to discriminate fight and non-fight images sequences, reporting 98.9% accuracy on movies data set [12]. Motion blobs, another form of motion features, are used to discriminate fight and non-fight video frames by extracting features from motion blob after thresholding the absolute difference of consecutive frame sequences for fight activity classification in movies producing 97.8% accuracy [13].

More on, a modified version of SIFT features attempt encoding appearance and Lagrangian-based features motion models, in a bag-of-words framework, classifying through SVM [45]. Similarly, a specialized Lagrangian technique, introducing a novel feature Lagrangian direction fields based on spatio-temporal model is presented, to perform automated violent video scenes detection [46]. A substantial derivative-based novel video descriptor is proposed, which is formed on a concept from fluid mechanics. This approach estimates the consecutive and local field from the optical flow, by using bag-of-words procedure for each motion pattern and finally forming the final descriptor by concatenating the resulting histograms [47]. A spatio-temporal feature based on interest points, detected in spatial and time domain using optical flow information, is used to learn an SVM binary classifier, which describes violence in videos of crowded and uncrowded scenes [48].

Recently, 3D ConvNets model is inquired on hockey data set [17]. Deep representation-based 3D-CNN architecture called C3D has 3D convolutional and 3D pooling layers for effectively learning temporal information from adjacent video frames [49]. Learning spatio-temporal features through CNN in 3D fashion was originally proposed for learning action recognition [50], used in airport video surveillance [51].

More recently, Serrano *et al.* evaluated C3D architecture [49] on hockey and movies data sets to combat the challenge of fight detection producing adequate results [16]. Apart from that, the author proposed 2D-CNN model using Hough forest features. This system revealed exceptional results on hockey and movies data sets with 94.6% and 99% accuracies, respectively, exceeding all previous techniques of handcrafted features detectors and deep representation models [16]. Moreover, a deep representation-based GoogleNet model using transfer learning is potentially inquired to identify violent human actions in hockey and movies data sets achieving benchmark results [52]. Besides that, to detect long-term temporal structure in violent sequences, a pertained model on UCF101 data set for action recognition is used as classifier for fight scenes on violent interaction detection (VID), a data set composed of hockey, movies, HMDB51 and UCF10. However, author is inclined toward training a deep learning model on action data set UCF101, to be used as pre-trained model for fine tuning on VID, exploring new modality acceleration field to capture the motion features in target videos [18].

Furthermore, modern deep representation architectures are proposed to solve person re-identification problem in surveillance videos. A deep feature embedding approach is proposed to combat the challenge of cross-view visual variation, matching pedestrian across the disjoint camera views, through deep learning feature space [4]. Two CNN-based networks generating relevant descriptors for pedestrian matching,

followed by spatially recurrent pooling via a four-directional RNN, solve what and where to match problem in video analysis [5]. A deep Siamese attention network is proposed to jointly learn spatio-temporal video representations and similarity metrics. The model contains an attention mechanism embedded in spatial gated recurrent units for where and when to look for relevant features [6]. Another, deep attention-based spatially recursive model operates on two CNN streams, to learn deep local features. These learned features are fed into spatial LSTMs, with soft attention mechanism embedded into LSTM units, to recognize the visual objects with subtle appearances differences [53]. An end-to-end trainable deep 3D convolutional architecture is presented to generate global video-level features over the entire video span, for person re-identification in video surveillance [54]. Deep representation-based methods demonstrate more power to learn deep features from images and videos sequences, as compare to handcrafted feature detectors.

However, deep representation architectures are resource hungry and demanding for computational efficiency. They require huge computational power and enormous amount of domain specific data to construct learned representations. The preparation of huge labeled domain-specific data set is itself a challenging, laborious and expensive task. This shortcoming leads to a major bottleneck in training deep learning model from scratch for target domain. To combat this challenge, an approach of transfer learning is an optimal strategy, in which a network pre-trained on a huge data set is taken as a source model, to create a target model by retraining on the small data set for target domain [24]. This strategy eliminates the need of producing huge labeled domain-specific data set along with training model from scratch. Taking this into account, winner models of ILSVRC, such as AlexNet [20] VGGNet [21], GoogleNet [22] and ResNet [23] trained on 15 million annotated images for 1000 categories, are fortunately publicly available as pre-trained open source networks. Researchers have successfully used these source networks, as deep representation models incorporating transfer learning strategy, to address large-scale video classification problem [25]. Moreover, this scheme is adopted in HAR domain to determine human actions in videos sequences, in coordination with SVM and KNN classifier [26]. Hence, open source models can be used as pre-trained networks employing transfer learning, to be retrained on target domain-specific small data sets, such as for the task of violent behaviors detection.

3. METHODOLOGY

This section describes the details about proposed methodology. At first DCNN transfer learning approach is used. Next, a DMN CNN model is proposed to address the task of violence detection in videos.

3.1. DCNN transfer learning

In the domain of machine learning, learning-based algorithms learn deep feature through iterative optimization procedure. Feature learning is very influential toward making deep learning models to learn complex underlying data representation, especially learning complex patterns for image recognition, as compare to handcrafted feature descriptors. The learnt features representation achieved through learning a specific problem can be re-utilized to solve a new task, even to an entirely new domain, a concept known as transfer learning. This approach has been successfully used for object classification and categorization in computer vision field [55].

Since DCNN model is primarily data driven for feature learning, for which it requires large labeled data set. However, preparing large volume of annotated data set is complex, demanding and time-consuming task, which requires enormous amount of development effort. Whereas, providing insufficient amount of training data would not leverage CNN model to learn complex deep features, instead network suffer from significant overfitting problem. To counter overfitting issue for small data set, utilizing modern deep learning network architectures, the approach of transfer learning becomes very handy. In transfer learning, existing network architecture with already learnt features as source task network is employed to build new target task network architecture, to be trained on limited data set [56]. Figure 1 shows general representation of source task network with convolutional blocks followed by dense fully connected subsequent layers, pre-trained on source data set ImageNet with 1000 output classes. The source task network is transformed to create a target task network through transfer learning, to be trained on hockey and movies data set, with two output classes for violent/fight and non-fight activities.

In this paper, AlexNet [20] and GoogleNet [22] are selected as a pre-trained source networks with pre-learnt features from ImageNet data set on 15 million annotated images for 1000 classes. AlexNet is first developed DCNN architecture for image recognition and possesses rapid down sampling capacity of the intermediate representations through strided convolutions and max-pooling layers [57]. The AlexNet architecture is comprised of five convolutional blocks followed by three fully connected layers with 60 million trainable parameters, making it an effective model for image classification [20]. GoogleNet codenamed Inception is a 22-layer deep network architecture with repeated inception modules. Although the network is 22 layers deep, it has 12 times fewer parameters than AlexNet making it an efficient deep neural network architecture for computer vision applications [22].

Based upon these characteristics, AlexNet and GoogleNet are selected for transfer learning experiments as source networks. These source networks are pre-trained on ImageNet data set and has 1000 classes. While, targeted hockey and movies data sets have only two classes i.e. fight and non-fight. To achieve transfer learning, the fully connected classification

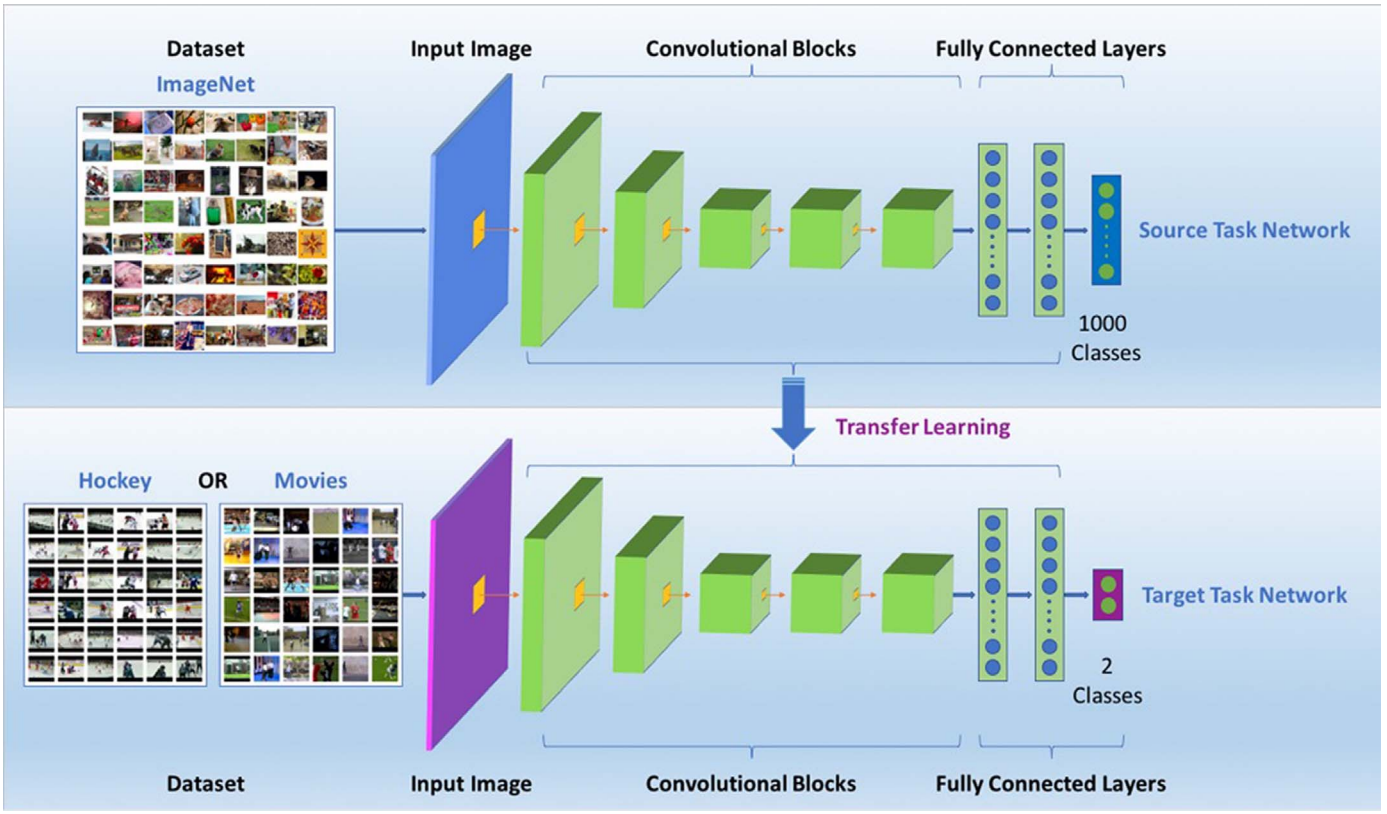

FIGURE 1. Transfer learning concept with DCNN model for image classification task, a source task network architecture pre-trained on source data set is fine-tuned on target data set to achieve target task network.

layers of the source network for 1000 classes are replaced with two classes of targeted data sets. The networks are further fine-tuned for optimization [30]. In practice during network retraining, if targeted data set is small, then transfer learning layers are left frozen to avoid overfitting, due to large number of training parameters. Optionally, if targeted data set has reasonable or large size then instead of freezing transfer learning layers, one can choose to back-propagate errors to previous layers for network fine tuning [29]. Since, hockey and movies data sets have reasonable size (images/frames in thousands), so transfer learning layers are not frozen during fine tuning. Instead, an error is back propagated to previous layers to adjust weights of network, which is producing dominating accuracies, for aggressive behaviors detection in videos.

3.2. DMN CNN model

AlexNet and GoogleNet are state-of-the-art deep representation architectures and winner models of ILSVRC. However, each network possesses distinct pre-learnt features, making them unique composite representational model for image classification. Consequently, each model has discrete potential to recognize target activity classes. This research work has key contribution in detection of aggressive human behaviors in videos by developing novel DMN CNN architecture. The DMN-proposed architecture leverages the integrated potential of AlexNet and GoogleNet by concurrently coalescing both networks. The AlexNet has 60 million trainable parameters densely packed in eight-layer network, whereas GoogleNet has 12 times fewer parameters distributed over 22-layer deep network. In proposed scheme, both networks with diverse set of distinct pre-learnt features are integrated together to build a fast learning system, yielding high-end accuracy for images/videos classification.

To achieve transfer learning in DMN model through k-fold cross validation, a systematic implementation is accomplished. The AlexNet and GoogleNet are pre-trained networks on ImageNet. In order to build DMN architecture, both networks are fine-tuned on target data set to form a multi-net composite model. Image sequences are obtained for 10-fold cross validations scheme from videos. These video frames are passed to DMN model through images resizing data pipeline distinctly connected to AlexNet and GoogleNet networks for fine tuning. The images resizing data pipeline maintains the streams of scaled images ready for fine tuning, independently compliant the input requirements of both networks simultaneously. Moreover, each fold possesses a stream of image sequences through individual images resizing data pipeline,

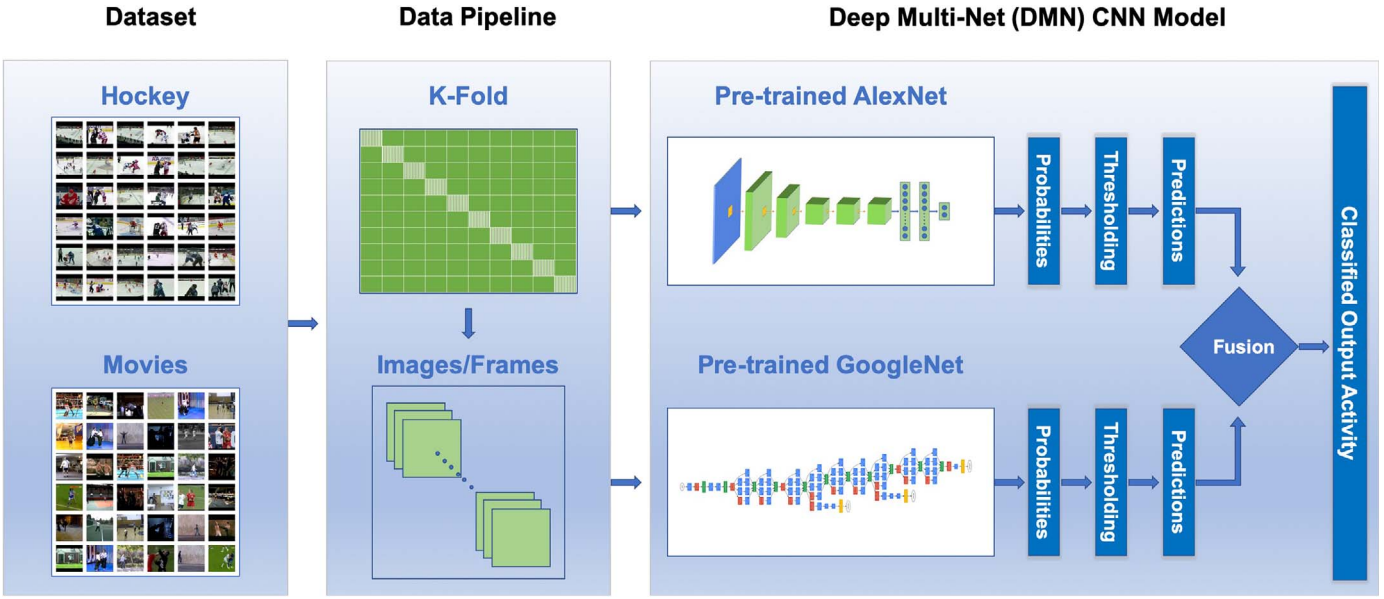

FIGURE 2. Overview of proposed system. Left two sections indicate the images/frames of hockey and movies data sets prepared for DMN architecture through images resizing data pipeline using k-fold scheme. Right section represents DMN CNN model comprised of AlexNet and GoogleNet pre-trained networks.

during 10-fold cross validation. The DMN consume images input stream for fine tuning AlexNet and GoogleNet concurrently throughout training process. To assure concurrent data integrity for both networks, the labels of image input stream are cross-validated through both networks during each training epoch. Hence, DMN fine-tuned networks exhibit complimentary strength in activity classification. Finally, during testing, the output probabilities are computed on 10-fold for both networks. These probabilities possess the output predictions of the networks after thresholding at default value 0.5. The distinct output predictions/labels of both networks are fused together for an overall prediction. The fusion of the computed predictions/labels of both models prevails the true prediction computed by distinct model to demonstrate overall system classification accuracy. The DMN architecture is examined on both hockey and movies data sets independently, with different training epochs, as discussed in Section 5 and 6.

The DMN system revealed superior accuracy as compare to all previous techniques, even surpassing AlexNet and GoogleNet accuracies, with reduced learning time. The results illustrate that proposed model has advantage over AlexNet and GoogleNet with minimized training time. DMN not only outperforms with best accuracy results but also it is more efficient with fast learning capabilities up to 2.28 times faster than AlexNet and GoogleNet, while achieving similar competing accuracies. Figure 2 shows an overview of proposed DMN system comprised of AlexNet and GoogleNet, with images/frames systematic acquisition from data sets using 10-fold cross validation.

4. DATA SETS

The experiments are conducted on two benchmark violence action detection hockey and movies data sets [9].

The hockey data set is first of its nature, specifically developed for violence detection systems. It has 1000 video clips of 360×288 resolution each. Data set has further two categories for fight and non-fight, with each category enclosing 500 clips. The data set is obtained from hockey games of National Hockey League with real-life violent events.

The movies data set is also exclusively designed for fight activity detection. It is comprised of 200 video clips for both fight and non-fight actions. Fight scenes are extracted from diverse actions movie clips, whereas non-fight scenes are obtained from publicly available action detection data sets. Unlike hockey data set, this data set is comprised of wide range of diverse scenes collection, recorded at different resolutions under different circumstances, with an average resolution of 360×250 pixels per frame.

In comparison, the hockey data set is challenging due to abrupt camera motion in recording for non-fight scenes, happening due to real-time hockey players movement tracking during game. Movies data set has views complexities due to diverse collection of scenes, captured in different environmental conditions, exhibiting variations in views background with different illuminations and occlusions. The challenging characteristics of both hockey and movies data sets are making them best suitable source to develop human violence recognition model for automated video surveillance system. See Fig. 3

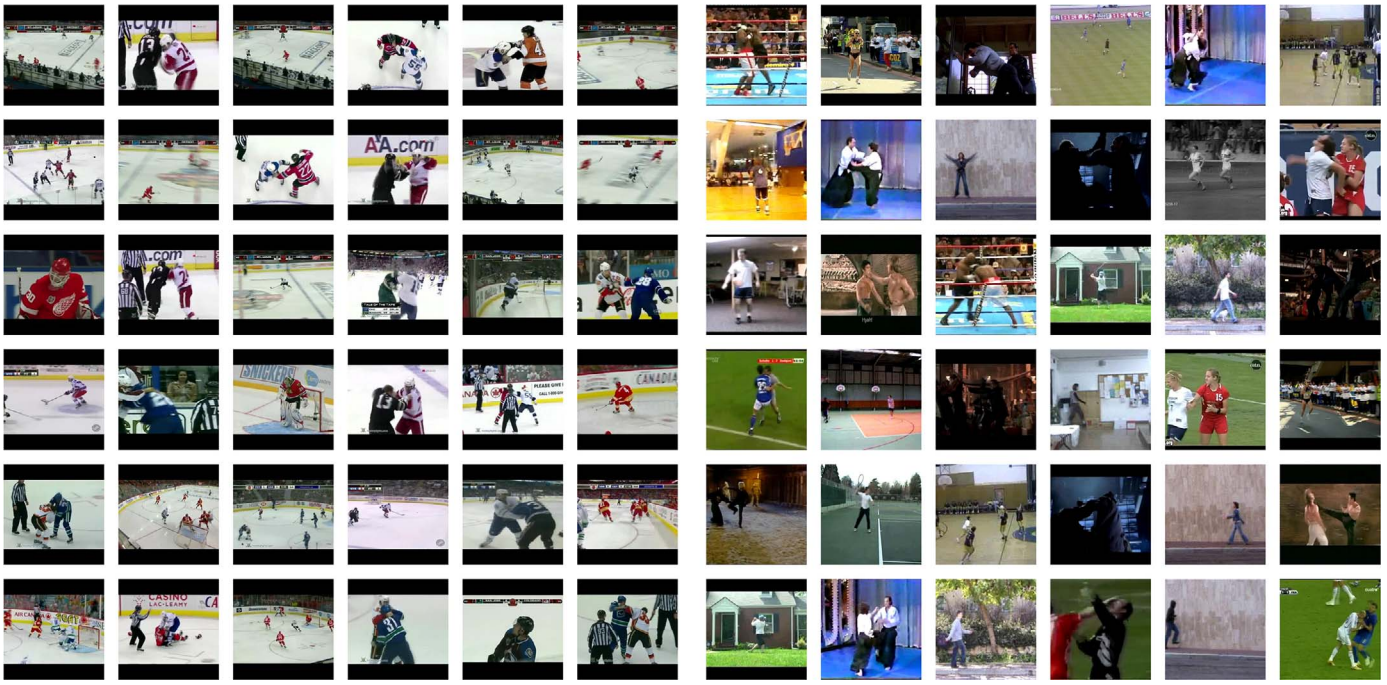

FIGURE 3. Hockey and movies data sets samples. Left image is a collection of fight and non-fight frames from hockey data set. Right image is a collection of fight and non-fight sequences from movies data set.

showing fight and non-fight frames collection for hockey and movies data sets.

5. EXPERIMENTS

This section describes details of experiments carried out using DCNN transfer learning and proposed method of DMN as discussed in Section 3.

5.1. Experimental setup

The AlexNet and GoogleNet models learn spatial features from images as DCNN network. These networks can learn features from videos, by converting annotated video clips into labeled images sequences. In experiments, as pre-training step both videos data sets are transformed into images/frames to efficiently train deep models.

The hockey data set has 1000 video clips. These clips has generated 20 557 and 20 499 annotated images for fight and non-fight, respectively, producing 41 056 total frames for both activities. The converted images represent adjacent video frames. Similarly, movies data set has 200 video clips. These clips produced 4791 and 5050 labeled images for fight and non-fight activities, respectively, producing 9841 frames for both activities.

5.2. Parameters

To achieve superior accuracy results, optimal training parameters are discovered through experimental work. Several experiments are conducted on both data sets, by examining different mini batch sizes, to choose best mini batch size for networks training. Based on examination, using mini batch size range between 4, 8, 16, 32, 64, 128 and 256, it is observed that lowest mini batch size is although effective to achieve high accuracy quickly in a smaller number of training epochs, however, training time required by each training cycle is exponentially high. This causes network to take much longer to learn deep features on whole data set. Alternatively, highest mini batch size is although taking very short span of training time per cycle, however, network ends up with very slow learning throughput. This causes greater number of epochs required to achieve desired accuracy, eventually exponentially increasing training time on whole data set. Based on these observations, an optimal mini batch size 64 is selected for networks training, producing best results.

AlexNet and GoogleNet networks are trained using stochastic gradient descent (SGD) with momentum optimizer. Each iteration of SGD is set to mini batch size 64, with a momentum of 0.9. Network learning rate is set to 0.0001 as a constant learning rate throughout training. During each iteration, network fine tuning is performed by back propagating error to previous layers throughout the network. Experiments are conducted on NVIDIA 1080ti GPU, Intel Xeon X5670 @ 2.93 GHz (two processors) with 32GB RAM.

For training DCNN network, although train/test split is recognized approach. However, to perform comprehensive analysis on target data sets by assuring networks are not overfitting during training, 10-fold cross validation scheme is adopted. It

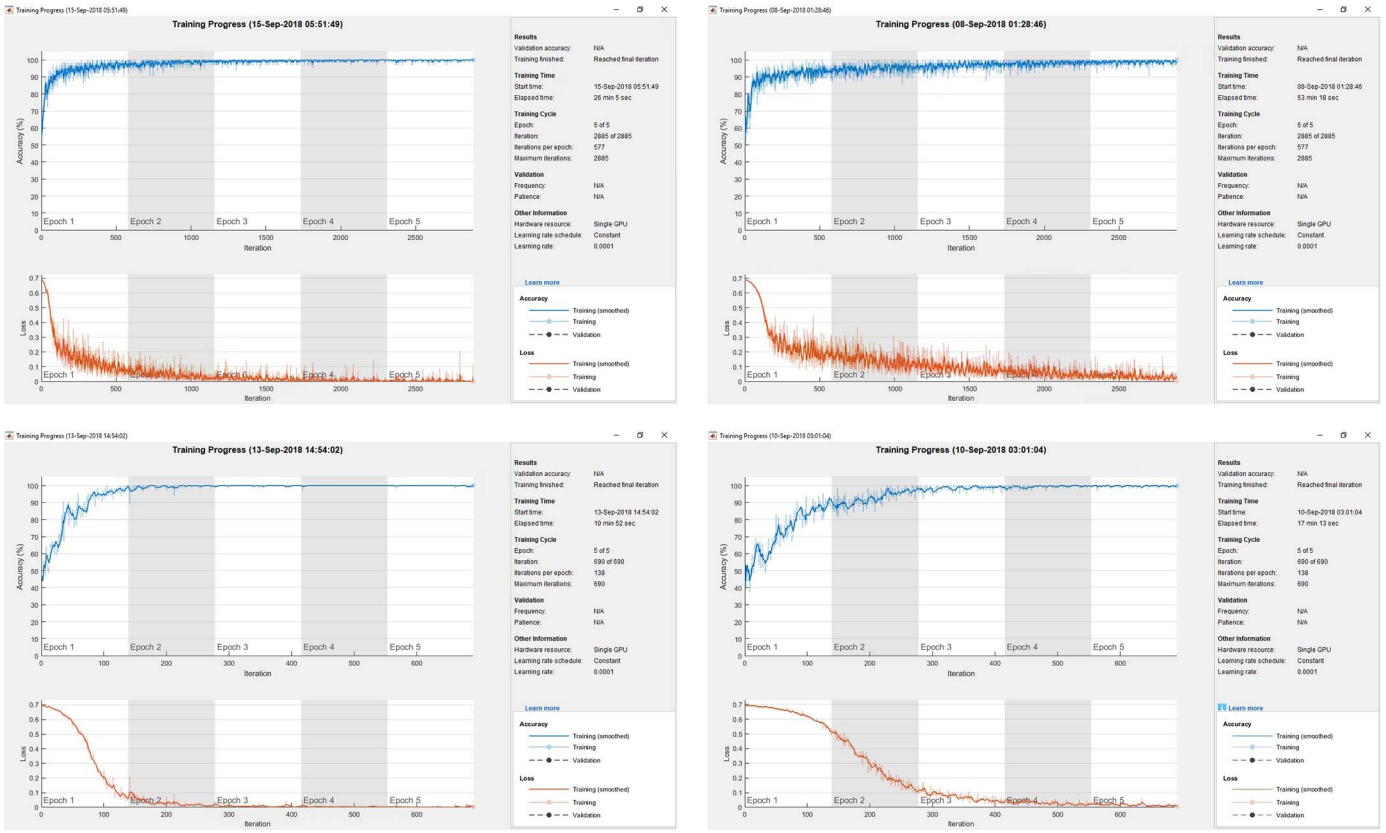

FIGURE 4. Hockey and movies data sets training progress samples of 1-fold for five epochs. Top row shows 1-fold training progress on hockey data set for AlexNet and GoogleNet from left to right, respectively. Bottom row shows 1-fold training progress on movies data set for AlexNet and GoogleNet from left to right, respectively.

is complicated to implement image data pipeline for 10-fold cross-validation scheme, to accommodate two distinct DCNN networks (AlexNet and GoogleNet) simultaneously in DMN system. Moreover, images in each data pipeline are scaled to 227×227 and 224×224 to develop a persistent required input stream for AlexNet and GoogleNet correspondingly. Lastly, networks are trained using original images/frames. Data frames are not augmented at any stage to create multiple augmented images for training, instead original images sequences are provided to networks for learning discrete features for violence classification.

5.3. Training progress

To achieve 10-fold cross validation training strategy, the AlexNet and GoogleNet are trained for each fold distinctly. Networks are trained in sequential fashion for each fold trained for n epochs, on both data sets. To perform a comprehensive analysis, both networks are trained for different number of n epochs, where n varies from 1, 5 and 10 for AlexNet and GoogleNet in all experiments. However, DMN architecture is solely examined for n epochs, where n varies from 1, 2 and 3, producing finest results.

Figure 4 shows training progress samples of 1-fold for five epochs on hockey and movies data sets, for both networks. Progress indicators shows drastic increase in accuracies by reducing loss in very initial training cycles. Moreover, networks achieved highest accuracy rapidly in subsequent epochs.

Figure 5 shows accuracies samples depicting each fold accuracy for five epochs on hockey and movies data sets, for both networks. Therefore, 10 distinct accuracies are reported for all 10-folds. Bar chart that shows hockey data set with abrupt camera motions in video frames is yielding arbitrary highest accuracies for different folds. However, movies data set with diverse collection of scenes have more consistent accuracies peaks for each fold.

6. RESULTS

This section describes results of the experiments carried out using DCNN transfer learning and proposed method of DMN architecture as suggested in Section 3. First DCNN and pro-

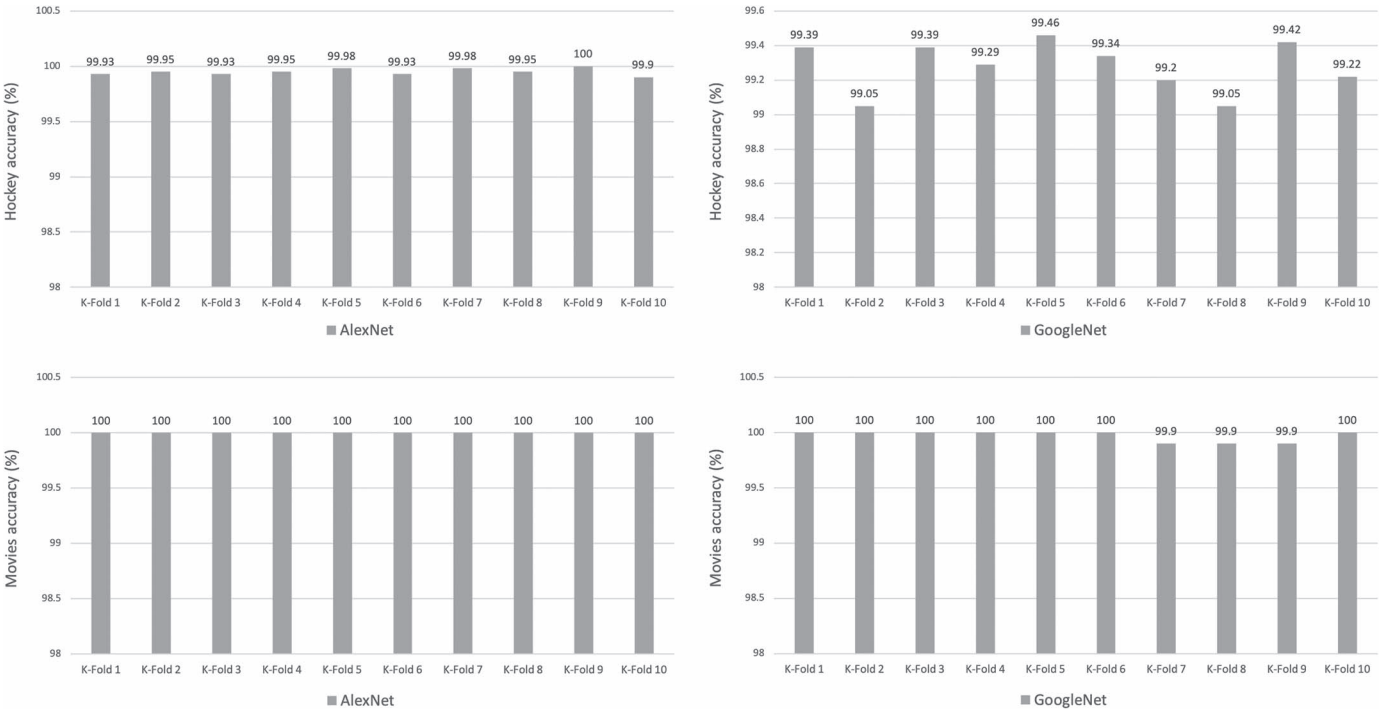

FIGURE 5. Hockey and movies data sets accuracies samples for each fold for five epochs, with 10-fold cross validation scheme. Top row shows accuracies for each fold of hockey data set, for AlexNet and GoogleNet from left to right, respectively. Bottom row shows accuracies for each fold of movies data set, for AlexNet and GoogleNet from left to right, respectively.

posed DMN system accuracy results on both data sets are discussed, then proposed DMN architecture with fast learning capabilities, producing superior accuracy as compare to DCNN models (AlexNet and GoogleNet) is examined.

6.1. DCNN and DMN results

The DCNN transfer learning and proposed method of DMN architecture is evaluated against two benchmark violent activity recognition data sets i.e. the hockey and movies data sets [9]. To perform a comprehensive comparative analysis, a wide range of reported algorithms are taken from literature with benchmark accuracies. These algorithms belong to both domains of handcrafted feature descriptors and deep representation-based models, as discussed in Section 2.

In handcrafted feature descriptor domain, Bermejo *et al.* [9] proposed approach using STIP (HOG), STIP (HOF) and MoSIFT features accomplished 90% as benchmark accuracy with the introduction of hockey and movies data sets for violence detection. Utilizing these data sets, Deniz *et al.* [12] suggested method incorporating SVM and Adaboost achieved 98.9% accuracy on movies data set. Following that, the violent flows (ViF) and LMP-based model are inquired [13]. Then, three contemporary classifiers, SVM, Adaboost and random forests, are reported [12, 13], to assess aggressive human behaviors in videos.

Furthermore, using deep representations approaches, Ding *et al.* [17] employed 3D-CNN model with train/test data split scheme. In recent times, Serrano *et al.* evaluated performances for C3D producing adequate results. The author further proposed finest approach for human violence detection by incorporating 2D-CNN with Hough forest features. This approach elevated hockey and movies data sets accuracies to $94.6 \pm 0.6\%$ and $99 \pm 0.5\%$ respectively, setting the accuracy bar to the next level [16].

Thereby, the DCNN transfer learning and proposed approach of DMN is extensively evaluated for human violence detection. Table 1 results illustrate a comprehensive comparison between DCNN transfer learning models (AlexNet and GoogleNet) and proposed strategy of DMN architecture against established techniques. Results are formulated as mean of accuracy for each experiment, using 10-fold cross-validation scheme as discussed in training progress part of Section 5.

The results for hockey data set show that our approaches, AlexNet and GoogleNet with 99.99% and 99.84% from deep 2D-CNN transfer learning models and proposed method of DMN with 99.82%, depicted superior accuracies as compare to all existing contemporary algorithms. Moreover, AlexNet has slightly higher accuracy as compare to GoogleNet and DMN in this case. Similarly, results for movies data set show that our approaches, AlexNet and GoogleNet from deep 2D-

TABLE 1. Comparison of classification accuracies results on hockey and movies data sets.

Author/method/year	Features/classifiers	Testing scheme	Data sets accuracy (%)	
			Hockey [9]	Movies [9]
Bermejo *et al.* [9]	STIP (HOG) + HIK	5-Fold CV	91.7± -%	49.0± -%
	STIP (HOF) + HIK		88.6± -%	59.0± -%
	HIK MoSIFT + HIK		90.9± -%	89.5± -%
Deniz *et al.* [12]	SVM	10-Fold CV	90.1±0%	85.4±9.3%
	Adaboost		90.1±0%	98.9±0.2%
	Random forests			
Ding *et al.* [17]	3D-CNN	Train/test split	91± -%	
ViF (2015) [13]	SVM	10-Fold CV	82.3±0.2%	96.7±0.3%
	Adaboost		82.2±0.4%	92.8±0.4%
	Random forests		82.4±0.6%	88.9±1.2%
LMP (2015) [13]	SVM	10-Fold CV	75.9±0.3%	84.4±0.8%
	Adaboost		76.5±0.9%	81.5±2.1%
	Random forests		77.7±0.6%	92±0.1%
Serrano *et al.* [13]	SVM	10-Fold CV	72.5±0.5%	87.2±0.7%
	Adaboost		71.7±0.3%	81.7±0.2%
	Random forests		82.4±0.6%	97.8±0.4%
Serrano *et al.* [16]	C3D [49]	10-Fold CV	87.4±1.2%	93.6±0.8%
	2D-CNN		87.8±0.3%	93.1±0.3%
	2D-CNN + HOG Forest		94.6±0.6%	99±0.5%
AlexNet	DCNN	1 Epoch	97.94%	98.42%
		5 Epoch	99.95%	**100%**
		10 Epoch	**99.99%**	**100%**
GoogleNet	DCNN	1 Epoch	94.06%	87.82%
		5 Epoch	99.28%	99.97%
		10 Epoch	**99.84%**	**100%**
Proposed	DMN	1 Epoch	98.32%	99.58%
		2 Epoch	99.57%	99.98%
		3 Epoch	**99.82%**	**100%**

CNN transfer learning models and proposed method of DMN, achieved 100% accuracy superseding each existing benchmark algorithm.

Conclusively, the DCNN and DMN proposed approaches outperforms state-of-the-art accuracies. Results confirm highest accuracies on both data sets i.e. the hockey and movies. The transfer learning strategy used in DCNN, which is essence of DMN, has specifically improved hockey data set accuracy by learning generalize deep representations for abrupt camera motion sequences, as compared to benchmark algorithms. Similarly, DCNN and DMN are also able to distinguish violent action in a wide variety of movie clips with dynamic scenes, on movies data set.

6.2. DMN fast learning evaluation

DMN has fast learning capabilities to learn deep features with reduced training time producing superior accuracy, as compare to the most high-tech models of computer vision field. The accuracy and learning time are strategically measured for DCNN models (AlexNet and GoogleNet) and proposed

TABLE 2. Comparison of accuracies results and learning time (minutes) on hockey and movies data sets.

Methodology		Data sets accuracy (%)		Learning time (minutes)	
		Hockey	Movies	Hockey	Movies
AlexNet	1 Epoch	97.94%	98.42%	50.53 m	20.57 m
	5 Epoch	99.95%	**100%**	262.15 m	108.85 m
	10 Epoch	**99.99%**	**100%**	527.48 m	217.65 m
GoogleNet	1 Epoch	94.06%	87.82%	107.78 m	34.08 m
	5 Epoch	99.28%	99.97%	538.88 m	172.38 m
	10 Epoch	**99.84%**	**100%**	1100.3 m	345.38 m
Proposed	1 Epoch	98.32%	99.58%	157.15 m	54.57 m
	2 Epoch	99.57%	99.98%	316.67 m	111.13 m
	3 Epoch	**99.82%**	**100%**	483.37 m	164.13 m

DMN system for cross examination. Thereby, in addition to mean accuracy, the learning time (training time) in minutes is computed for each model for n epochs, where n varies from 1, 5 and 10 for AlexNet and GoogleNet and 1, 2 and 3 for DMN, on both data sets experiments.

There is an obvious increase in accuracy value with the increase in n training epochs for DCNN models. However, results show that AlexNet is able to learn more comprehensive classification features at fifth epoch as compare to GoogleNet, which learn at 10th epoch. For that reason, AlexNet is producing 99.95% and 100% accuracies at fifth epoch as compare to GoogleNet 99.84% and 100% accuracies at 10th epoch for hockey and movies data sets, respectively, which means that certain deep networks can learn certain discriminating features faster as compare to others with less training epochs. The discovery of fast feature learning capacity of models led this research to inquire and develop DMN system to learn most comprehensive differentiating features with minimal training effort, as suggested in Section 3.

This section evaluates the following key characteristics of proposed architecture, which make DMN an effective fast learning system exhibiting high-end accuracies:

1. Proposed model is yielding superior accuracy results as compare to other networks (AlexNet and GoogleNet), when trained at similar number of n epochs.
2. Proposed model is learning comprehensive discriminating features faster in very few n epochs, as compare to contemporary models (AlexNet and GoogleNet).

For DMN fast learning evaluation, Table 2 reports mean accuracy and learning time (in minutes) for DCNN (AlexNet and GoogleNet) and DMN models, on both hockey and movies data sets, for n epochs, where n varies from 1, 5 and 10 epochs for DCNN and 1, 2 and 3 for DMN.

The hockey data set results show that proposed DMN approach achieved higher accuracy 98.32%, in comparison

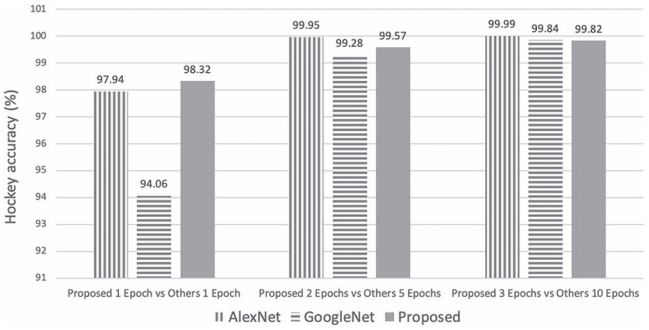

FIGURE 6. Hockey data set accuracies for DCNN (AlexNetand GoogleNet) and proposed DMN model. Proposed model is producing higher accuracy at epoch 1, whereas learning features faster producing similar accuracies at epoch 2 and 3, which others learns at epoch 5 and 10, respectively. (Higher value is better.)

with AlexNet and GoogleNet achieved accuracies 97.94% and 94.06%, respectively, when trained at similar number of n epochs i.e., n set to 1. Moreover, proposed approach learns features faster achieving similar accuracy 99.57% at epoch 2, as compare to AlexNet and GoogleNet accuracies 99.95% and 99.28% correspondingly, achieved at epoch 5. Similarly, proposed approach learns features even faster achieving similar accuracy 99.82% at just epoch 3, as compare to AlexNet, GoogleNet accuracies 99.99%, 99.84% respectively, achieved at epoch 10. See Fig. 6 for graphical representation details.

Similarly, the movies data set results describe that proposed strategy produced higher accuracy 99.58%, as compare to AlexNet and GoogleNet achieved accuracies 98.42% and 87.82%, respectively, when trained for equal number of n epochs i.e. n set to 1 for all networks. Besides, proposed approach learns features faster achieving similar accuracy 99.98% at epoch 2, as compare to AlexNet and GoogleNet accuracies 100% and 99.97% correspondingly, achieved at

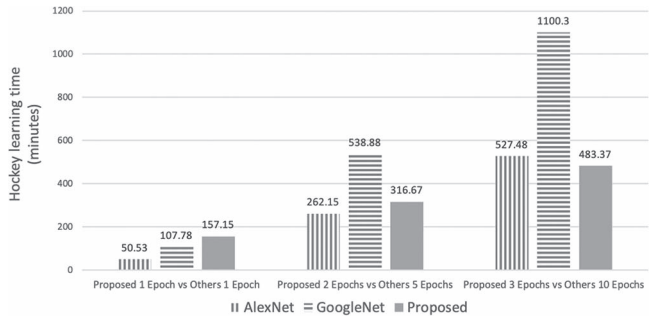

FIGURE 7. Hockey data set computed learning time for DCNN (AlexNet and GoogleNet) and proposed DMN model. Proposed model has although higher learning time when trained at similar n epochs i.e. n equals to 1, however it is learning features faster with drastically reduced training time at epoch 2 and 3, which other learns at epoch 5 and 10, respectively. (Lower value is better.)

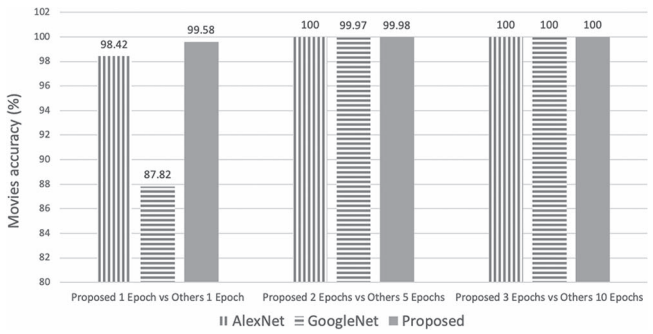

FIGURE 8. Movies data set accuracies for DCNN (AlexNet and GoogleNet) and proposed DMN model. Proposed model is producing higher accuracy at epoch 1, whereas learning features faster producing similar accuracies at epoch 2 and 3, which others learns at epoch 5 and 10, respectively. (Higher value is better.)

epoch 5. Likewise, proposed approach learns features even faster achieving an equal accuracy 100% at just epoch 3, as compare to AlexNet and GoogleNet accuracies 100%, achieved at epoch 10. See Fig. 8 for graphical representation details.

The learning time for hockey data set show that training time required to train proposed model is although higher when trained at similar number of n epochs i.e. n equals to 1, however model training time drastically reduces by learning features faster with higher number of training epochs. For epoch 1, proposed approach learning time is 157.15 m; however, AlexNet and GoogleNet have 50.53 m and 107.78 m, respectively. For epoch 2, proposed approach learning time is 316.67 m, close to AlexNet 262.15m at fifth epoch, but smaller than GoogleNet 538.88m at fifth epoch. Moreover, for epoch 3, proposed approach learning time 483.37m is smaller than AlexNet 527.48m at 10th epoch; however, it is significantly smaller than GoogleNet 1100.3 m at 10th epoch, simultaneously achieving similar accuracy results. See Fig. 7 for graphical representation details.

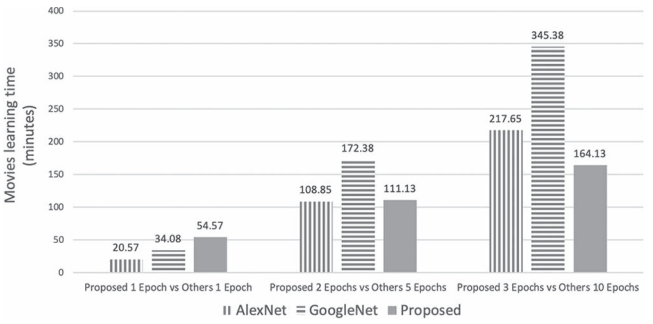

FIGURE 9. Movies data set computed learning time for DCNN (AlexNet and GoogleNet) and proposed DMN model. Proposed model has although higher learning time when trained at similar n epochs i.e. n equals to 1, however it is learning features faster with drastically reduced training time at epoch 2 and 3, which other learns at epoch 5 and 10, respectively. (Lower value is better.)

Likewise, the learning time for movies data set confirms similar findings about learning behavior of proposed model when trained for equal number of n epochs, and drastic reduced training time by learning features faster with higher number of training epochs. For epoch 1, proposed approach learning time is 54.57 m; however, AlexNet and GoogleNet have 20.57 m and 34.08 m, respectively. For epoch 2, proposed approach learning time is 111.13 m, close to AlexNet 108.85 m at fifth epoch, but smaller than GoogleNet 172.38 m at fifth epoch. Moreover, for epoch 3, proposed approach learning time 164.13 m is smaller than AlexNet 217.65 m at 10th epoch; however, it is significantly smaller than GoogleNet 345.38 m at 10th epoch, at the same time achieving similar accuracy results. See Fig. 9 for graphical representation details.

Finally, DMN fast learning capacity is evaluated assessing how many times proposed technique is faster than high-tech AlexNet and GoogleNet models.

Performance evaluation results for hockey data set show that proposed technique at epoch 2 is although 0.17 times slower than AlexNet trained at fifth epoch, however, it is 1.7 times faster than GoogleNet trained at fifth epoch, producing similar accuracies. Moreover, proposed approach at epoch 3 is 1.09 times faster than AlexNet trained at 10th epoch, besides that it is 2.28 times even faster than GoogleNet trained at 10th epoch, producing similar accuracies. See Fig. 10 for graphical representation details.

Likewise, performance evaluation results for movies data set illustrate that proposed approach at epoch 2 is although just 0.02 times slower than AlexNet trained at 5th epoch; however, it is 1.55 times faster than GoogleNet trained at 5th epoch, generating similar accuracies. Furthermore, proposed approach at epoch 3 is 1.33 times faster than AlexNet trained at 10th epoch, beyond that it is 2.1 times even faster than GoogleNet trained at 10th epoch, generating similar accuracies. See Fig. 10 for graphical representation details.

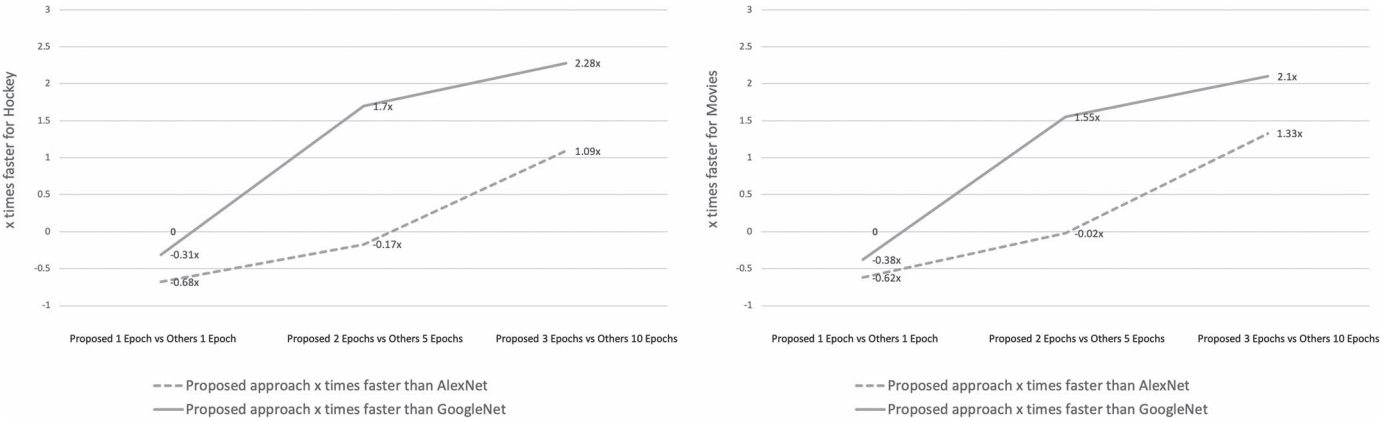

FIGURE 10. DMN fast learning capacity evaluation assessing x-times faster than AlexNet and GoogleNet. Left image show performance evaluation for hockey data set and right image show performance evaluation for movies data set.

The results confirm that proposed DMN CNN architecture has huge capacity of learning most discriminating features much faster, in few epochs with minimal training effort. DMN achieved similar accuracy results at just second and third epoch, which were achieved by AlexNet and GoogleNet at 5th and 10th epoch. Proposed approach is up to 1.33 times faster than AlexNet and 2.28 times faster than GoogleNet in learning features, simultaneously outperforming existing accuracies on both hockey and movies data sets, for human violence recognition. The DMN has potential to learn features even faster, by saving enormous amount of training time for large-scale images/videos recognition. The system can rapidly learn deep generalized discriminating features even beyond the domain of human violence recognition in video surveillance.

7. CONCLUSION

Human violence detection has attracted computer vision community during the past few years, to detect the aggressive human behaviors such as fight scene recognition for development of automated video surveillance systems. Historically, violence recognition is usually achieved through handcrafted feature descriptors. Yet, some researchers also proposed deep representation-based models to detect potential violent contents. Although, deep representation-based transfer learning techniques have been successfully used for human action detection such as walking, jogging, running and hand waving. However, there is scarcity in using transfer learning-based deep representation models for human violence detection. The hockey and movies data sets are first of their nature specifically designed for violent/fight scenes identification, as compare to other available data sets designed for general human action detection.

In this research, the learned representation-based DCNN models incorporating transfer learning is firstly in-depth examined, then a novel DMN CNN architecture is proposed to identify violent/fight scenes in images sequences. For deep representation-based models, training from scratch ends up with network overfitting issues. Therefore, transfer learning is adopted as an alternative strategy. AlexNet and GoogleNet, very deep network architectures, are adopted as source task networks, pre-trained on ImageNet data set with 15 million annotated images for 1000 categories. A source task network is utilized in transfer learning to develop a target task network, fine tuned on target data sets. AlexNet and GoogleNet are fine tuned as DCNN models using suggested optimal parameters on hockey and movies data sets. DMN-proposed architecture has profound advantages over state-of-the-art AlexNet and GoogleNet networks. DMN exhibits the integrated potential of AlexNet and GoogleNet by concurrently coalescing both networks, to build an effective fast learning system yielding high-end accuracy results for violence detection.

The DCNN transfer learning networks and proposed DMN model are evaluated using 10-fold cross validation scheme, on both data sets independently i.e. the hockey and movies. The DCNN networks are fine-tuned, by developing images resizing data pipeline as input stream being fed to fine tuning networks, for each distinct fold. Following that in DMN system, pre-validated images resizing data pipeline is distinctly connected to AlexNet and GoogleNet networks for fine tuning, for each distinct fold and for each distinct network, through a systematic implementation.

The results show that AlexNet and GoogleNet with 99.99% and 99.84% from DCNN and proposed DMN with 99.82% achieved superior accuracies for hockey data set as compare to all existing top-ranked techniques. Similarly, AlexNet and GoogleNet from DCNN and proposed DMN achieved 100% superseding accuracies for movies data set in comparison with existing benchmark algorithms.

Finally, the DMN-proposed architecture accuracy and performance is evaluated against most high-tech (AlexNet and GoogleNet) models in computer vision field. The proposed DMN architecture is not only yielding superior accuracy results at similar training epochs. Additionally, it has fast learning

capabilities, to learn comprehensive discriminating features faster in very few epochs with reduced training time, as compare to contemporary (AlexNet and GoogleNet) models. The hockey and movies data sets results show that proposed DMN approach achieved highest accuracy as compare to AlexNet and GoogleNet, when trained at similar number of *n* epochs. Moreover, proposed approach learns features faster achieving similar accuracy at just 2nd and third epoch, as compare to AlexNet and GoogleNet accuracies achieved at 5th and 10th epoch. Furthermore, the proposed approach is up to 1.33 times faster than AlexNet and 2.28 times faster than GoogleNet in features learning, achieving superior accuracy results on both hockey and movies data sets. The proposed DMN architecture is examined for human violence detection; however, the system has potential to learn discriminating features even faster, for large-scale images/videos classification, beyond the domain of human violence recognition in video surveillance.

Funding

European Union's Horizon 2020 Research and Innovation Program (778035).

REFERENCES

[1] Schuldt, C., Laptev, I. and Caputo, B. (2004) Recognizing Human Actions: A Local AVM approach. In *Proc. Int. Conf. Pattern Recognition*. IEEE.

[2] Gorelick, L., Blank, M., Shechtman, E., Irani, M. and Basri, R. (2007) Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29, 2247–2253.

[3] Marszałek, M., Laptev, I. and Schmid, C. (2009) Actions in Context. In *2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*. IEEE.

[4] Wu, L., Wang, Y., Gao, J. and Li, X. (2018) Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognit.*, 73, 275–288.

[5] Wu, L., Wang, Y., Li, X. and Gao, J. (2018) What-and-where to match: deep spatially multiplicative integration networks for person re-identification. *Pattern Recognit.*, 76, 727–738.

[6] Wu, L., Wang, Y., Gao, J. and Li, X. (2019) Where-and-when to look: deep Siamese attention networks for video-based person re-identification. *IEEE Trans. Multimedia*, 21, 1412–1424.

[7] Poppe, R. (2010) A survey on vision-based human action recognition. *Image Vision Comput.*, 28, 976–990.

[8] Ke, S.-R., Thuc, H., Lee, Y.-J., Hwang, J.-N., Yoo, J.-H. and Choi, K.-H. (2013) A review on video-based human activity recognition. *Computers*, 2, 88–131.

[9] Nievas, E.B., Suarez, O.D., García, G.B. and Sukthankar, R. (2011) Violence Detection in Video Using Computer Vision Techniques. In *CAIP'11 Proc. 14th Int. Conf. Computer Analysis of Images and Patterns—Volume Part II*, 2, 332–339. Springer.

[10] Sargano, A., Angelov, P. and Habib, Z. (2017) A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Appl. Sci.*, 7, 110.

[11] Wu, D., Sharma, N. and Blumenstein, M. (2017) Recent Advances in Video-Based Human Action Recognition using Deep Learning: A Review. In *2017 Int. Joint Conf. Neural Networks (IJCNN)*, pp. 2865–2872. IEEE.

[12] Deniz, O., Serrano, I., Bueno, G. and Kim, T.-K.T.-K. (2014) Fast Violence Detection in Video. In *2014 Int. Conf. Computer Vision Theory and Applications (VISAPP)*, pp. 478–485. IEEE.

[13] Gracia, I.S., Suarez, O.D., Garcia, G.B. and Kim, T.K. (2015) Fast fight detection. *PLoS One*, 10, e0120448.

[14] Nam, J., Alghoniemy, M. and Tewfik, A.H. (1998) Audio-Visual Content-Based Violent Scene Characterization. In *1998 Int. Conf. on Image Processing, 1998. ICIP 98*, pp. 353–357. IEEE.

[15] Cheng, W.-H., Chu, W.-T. and Wu, J.-L. (2003) Semantic Context Detection Based on Hierarchical Audio Models. In *Proc. 5th ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, pp. 109–115. ACM.

[16] Serrano, I., Deniz, O., Espinosa-Aranda, J. and Bueno, G. (2018) Fight recognition in video using Hough forests and 2D convolutional neural network. *IEEE Trans. Image Process.*, 27, 4787–4797.

[17] Ding, C., Fan, S., Zhu, M., Feng, W. and Jia, B. (2014) Violence Detection in Video by Using 3D Convolutional Neural Networks. In *Int. Symposium on Visual Computing*, pp. 551–558. Springer, Cham.

[18] Zhou, P., Ding, Q., Luo, H., Hou, X., Jin, B. and Maass, P. (2017) Violent Interaction Detection in Video Based on Deep Learning. In *Journal of Physics: Conf. Series 12044*. IOP Publishing.

[19] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document. *Proc. IEEE*, 86, 2278–2323.

[20] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. In *Advances In Neural Information Processing Systems*, pp. 1097–1105. NIPS.

[21] Simonyan, K. and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. 1–14. *arXiv*:1409.1556. cs.CV.

[22] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) Going Deeper with Convolutions. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pp. 1–9. cv-foundation.

[23] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. In *2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

[24] Zeiler, M.D. and Fergus, R. (2014) Visualizing and Understanding Convolutional Networks, *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8689, pp. 818–833.

[25] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014) Large-Scale Video Classification with Convolutional Neural Networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1725–1732.

[26] Sargano, A.B., Wang, X., Angelov, P. and Habib, Z. (2017) Human Action Recognition Using Transfer Learning with Deep Representations. In *2017 Int. Joint Conf. Neural Networks (IJCNN)*, pp. 463–469.

[27] Razavian, A.S., Azizpour, H., Sullivan, J. and Carlsson, S. (2014) CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops*, pp. 806–813.

[28] Garcia-Gasulla, D., Vilalta, A., Parés, F., Ayguadé, E., Labarta, J., Cortés, U. and Suzumura, T. (2018) An Out-of-the-Box Full-Network Embedding for Convolutional Neural Networks. In *Proc. 9th IEEE Int. Conf. Big Knowledge, ICBK 2018*, pp. 168–175.

[29] Yosinski, J., Clune, J., Bengio, Y. and Lipson, H. (2014) How Transferable are Features in Deep Neural Networks? In *Advances in Neural Information Processing Systems*, pp. 3320–3328.

[30] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2014) DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Int. Conf. Machine Learning*, pp. 647–655.

[31] Tran, D., Wang, H., Torresani, L., Ray, J., Lecun, Y. and Paluri, M. (2018) A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459. CVF.

[32] Kong, Y. and Fu, Y. (2018) Human action recognition and prediction: a survey (in press), *arXiv*:1806.11230, 13.

[33] Hoang, V.-D., Hoang, D.-H. and Hieu, C.-L. (2018) Action Recognition Based on Sequential 2D-CNN for Surveillance Systems. In *IECON 2018-44th Annual Conf. IEEE Industrial Electronics Society*, pp. 3225–3230. IEEE.

[34] Giannakopoulos, T., Kosmopoulos, D., Aristidou, A. and Theodoridis, S. (2006) Violence Content Classification Using Audio Features. In *Hellenic Conf. Artificial Intelligence*, pp. 502–507. Springer.

[35] Clarin, C., Dionisio, J., Echavez, M. and Naval, P. (2005) DOVE: detection of movie violence using motion intensity analysis on skin and blood. *PCSC*, 6, 150–156.

[36] Zajdel, W., Krijnders, J.D., Andringa, T. and Gavrila, D.M. (2007) CASSANDRA: Audio-Video Sensor Fusion for Aggression Detection. In *2007 IEEE Conf. Advanced Video and Signal Based Surveillance, AVSS 2007 Proc.*, pp. 200–205. IEEE.

[37] Gong, Y., Wang, W., Jiang, S., Huang, Q. and Gao, W. (2008) Detecting Violent Scenes in Movies by Auditory and Visual Cues. In *Pacific-Rim Conference on Multimedia*, pp. 317–326. Springer.

[38] Chen, D., Wactlar, H. and Chen, M.-y., Gao, C., Bharucha, A., and Hauptmann, A. (2008) Recognition of Aggressive Human Behavior Using Binary Local Motion Descriptors. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual Int. Conf. IEEE*, pp. 5238–5241. IEEE.

[39] Lin, J. and Wang, W. (2009) Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training. In *Pacific-Rim Conference on Multimedia*, pp. 930–935. Springer.

[40] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S. and Theodoridis, S. (2010) Audio-Visual Fusion for Detecting Violent Scenes in Videos. In *Hellenic Conf. Artificial Intelligence*, pp. 91–100. Springer.

[41] Chen, L.-H., Su, C.-W. and Hsu, H.-W. (2011) Violent scene detection in movies. *Int. J. Pattern Recognit. Artif. Intell.*, 25, 1161–1172.

[42] Hassner, T., Itcher, Y. and Kliper-Gross, O. (2012) Violent Flows: Real-Time Detection of Violent Crowd Behavior. In *2012 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–6. IEEE.

[43] Xu, L., Gong, C., Yang, J., Wu, Q. and Yao, L. (2014) Violent Video Detection Based on MoSIFT Feature and Sparse Coding. In *ICASSP, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 3538–3542.

[44] Baumann, F., Lao, J., Ehlers, A., Rosenhahn, B. and Yang, M.Y. (2014) Motion Binary Patterns for Action Recognition. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops; The Seventh IEEE Workshop on Embedded Computer Vision (ECVW2011)*, pp. 385–392.

[45] Senst, T., Eiselein, V. and Sikora, T. (2015) A Local Feature based on Lagrangian Measures for Violent Video Classification. In *Int. Conf. Imaging for Crime Prevention and Detection*. IET.

[46] Senst, T., Eiselein, V., Kuhn, A. and Sikora, T. (2017) Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation. *IEEE Trans. Inf. Forensics Secur.*, 12, 2945–2956.

[47] Mohammadi, S., Kiani, H., Perina, A. and Murino, V. (2015) Violence Detection in Crowded Scenes Using Substantial Derivative. In *AVSS 2015 - 12th IEEE Int. Conf. Advanced Video and Signal Based Surveillance*. IEEE.

[48] Ben Mabrouk, A. and Zagrouba, E. (2017) Spatio-temporal feature using optical flow based distribution for violence detection. *Pattern Recognit. Lett.*, 92, 62–67.

[49] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proc. IEEE Int. Conf. Computer Vision*. CVF.

[50] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C. and Baskurt, A. (2011) Sequential Deep Learning for Human Action Recognition. *International Workshop on Human Behavior Understanding*, Vol. 7065, pp. 29–39. Springer.

[51] Ji, S., Yang, M. and Yu, K. (2013) 3D Convolutional Neural Networks for Human Action Recognition. *Pami*, 35,221–231.

[52] Mumtaz, A., Sargano, A.B. and Habib, Z. (2018) Violence Detection in Surveillance Videos with Deep Network using Transfer Learning. In *2nd European Conf. Electrical Engineering & Computer Science, Bern, Switzerland*. IEEE Computer Society Press, USA. http://www.eecs-conf.org/.

[53] Wu, L., Wang, Y., Li, X. and Gao, J. (2019) Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE Trans. Cybern.*, 49, 1791–1802.

[54] Wu, L., Wang, Y., Shao, L. and Wang, M. (2019) 3-D PersonVLAD: Learning deep global representations for video-based person reidentification. *IEEE Trans. Neural Networks Learn. Syst.*, 30, 3347–3359.

[55] Aytar, Y. (2014) Transfer learning for object category detection. PhD Thesis, University of Oxford.

[56] Su, Y.-C., Chiu, T.-H., Yeh, C.-Y., Huang, H.-F. and Hsu, W.H. (2014) Transfer learning for video recognition with scarce training data for deep convolutional neural network (in press). CoRR, 1409.4127, 1–12. Citeseer.

[57] Grm, K.*et al.* (2017) Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biom.*, 7, 81–89.

Aqib Mumtaz received MS from COMSATS University Islamabad (CUI) with distinction. He served for US based companies for 15+ years, delivering diversified high-tech research projects in the field of mobile computing, web systems, health, fitness and embedded systems. Presently, he is associated with COMSATS University Islamabad, Lahore campus, Pakistan, doing PhD in Department of Computer Science and serving as a Project Manager managing Data Science Team building AI/ML based solutions in the cyber-security domain. Mumtaz's research interests are human action recognition, violence detection, computer vision, image analysis, machine learning, deep learning and artificial intelligence.

Allah Bux Sargano earned his PhD degree in computer science from Lancaster University, United Kingdom and MS degree from COMSATS University Islamabad, Pakistan. Currently he is working as Assistant Professor at the Department of Computer Science, COM-SATS University Islamabad (CUI), Lahore campus, Pakistan. Sagano's research interests broadly span the areas of human activity recognition from videos using deep learning techniques, currency recognition, machine learning, and computer vision. He has published several articles in international journals and conferences.

Zulfiqar Habib earned his PhD degree in Computer Science in 2004 from Kagoshima University Japan followed by the award of Postdoctoral fellowship of two years by Japan Society for the Promotion of Science (JSPS). Dr. Habib has served as the Chairman of Department of Computer Science, COMSATS University Islamabad (CUI) and currently holding the position of Professor. He is also working as the coordinating principal investigator and country representative for European Union's Horizon 2020, MSCA-RISE-2017. Dr. Habibs teaching and research interests include Computer Graphics, Computer Vision, Machine Learning, and Robotics. Dr. Habib has achieved various awards in education and research including two Research Productivity Awards at national level, best researcher award from CUI, and two graduate merit fellowships by Japan. Since 2009, he has been invited to give keynote lectures or tutorials in numerous national and International conferences, and as the guest researcher in the universities of Germany, Turkey and UK.