Violence Detection in Surveillance Videos with Deep Network using Transfer Learning

Aqib Mumtaz
Department of Computer Sciences
COMSATS University Islamabad,
Lahore, Pakistan
fa16-rcs-022@cuilahore.edu.pk
aqib.mumtaz@gmail.com

Allah Bux Sargano
Department of Computer Sciences
COMSATS University Islamabad,
Lahore, Pakistan
allahbux@cuilahore.edu.pk

Zulfiqar Habib
Department of Computer Sciences
COMSATS University Islamabad,
Lahore, Pakistan
drzhabib@cuilahore.edu.pk

***Abstract*—Violent action recognition has significant importance in developing automated video surveillance systems. Over last few years, violence detection such as fight activity recognition is mostly achieved through hand-crafted features detectors. Some researchers also inquired learning based representation models. These approaches achieved high accuracies on Hockey and Movies benchmark datasets specifically designed for detection of violent sequences. However, these techniques have limitations in learning discriminating features for videos with abrupt camera motion of Hockey dataset. Deep representation based approaches have been successfully used in image recognition and human action detection tasks. This paper proposed deep representation based model using concept of transfer learning for violent scenes detection to identify aggressive human behaviors. The result reports that proposed approach is outperforming state-of-the-art accuracies by learning most discriminating features achieving 99.28% and 99.97% accuracies on Hockey and Movies datasets respectively, by learning finest features for the task of violent action recognition in videos.**

Keywords—Violence Detection, Fight Recognition, Surveillance Videos, Deep CNN, GoogleNet, Transfer Learning

I. INTRODUCTION

In video surveillance, to critically assure public safety hundreds and thousands of surveillance cameras are deployed within cities, but it is almost impossible now a day to manually monitor all cameras to keep an eye on violent activities. Rather, there is a significant requirement for developing automated video surveillance systems to automatically track and monitor such activities. Thereby, in case of emergencies alarming the controlling authorities to take appropriate measures against detected violence. Violence recognition is a key step towards developing automated security surveillance systems, to distinguish normal human activities from abnormal/violent actions. Normal human activities are often categorized as routine life interactive behaviors, such as walking, jogging, running, hand waving [1] [2]. However, violence is subjected to unusual furious actions, such as fight activity happening between two or more people [3].

In last few years, the task of human action recognition has received much attention of the researcher community, to detect normal day human activities through video analysis, see surveys [4] [5]. However, little attention has paid to the problem of human violent action detection, until the availability of violent sequences for fight activities. The authors has created two datasets specifically for fight activities detections, to distinguish violent/fight incidents from normal events [6]. Before the availability of this dataset, most of the available datasets were particularly concerned about general human activities. Conversely, this dataset is first of its nature focused on violent scenes detection, to build precise surveillance systems, for monitoring indoor and outdoor environments.

Historically, human activity recognition is achieved through traditional hand-crafted feature representation approaches such as Histogram of Oriented Gradient (HOG), Scale-Invariant Feature Transform (SIFT), Hessian3D and Local Binary Pattern (LBP) etc. More on, there is growing tendency to solve this problem by adopting learning based deep representation techniques, such as Convolutional Neural Networks (CNN), 3D-CNN for spatio-temporal analysis, CNN followed by Recurrent Neural Network (RNN) and Spiking Neural Networks (SNN) etc. see survey [7] [8].

For violence detection, most of the existing approaches relay on hand-defined features descriptors, to distinguish fight sequences from normal ones, a scheme often used in human action recognition domain. Thereby, since the introduction of the violent/fight specific two datasets, most of the techniques are dependent on formulating hand-crafted feature representations for violence identification, such as Space-Time Interest Points (STIP), Motion SIFT (MoSIFT), Motion features, Motion blobs performed on audio-visual analysis along blood and flame detection [6], [9]–[12]. However, A few researches are conducted using deep learning techniques such as 2D-CNN, 3D-CNN, C3D [13]–[15]. Besides that, there is a scarcity in using deep representations models based transfer learning approach, to solve violent/fight detection problem, in violent action recognition domain.

Deep learning based approaches generally called end-to-end learning. It has history of deep representation based Convolutional Neural Network (CNN) model, starting from hand-written digit classifications. Then, over the years evolved as state-of-the-art deep CNN architectures such as AlexNet, GoogleNet, ResNet. These architectures are winner model of ImageNet Large Scale Visual Recognition Challenge (ILSVRC), due to their remarkable accuracy for image classification task. These networks are trained on 15 million annotated images for 1000 categories [16]–[19]. However, in order to successfully train a deep learning network, a very large dataset is required to learn generalized features. To combat the challenge of huge data requirement, the concept of transfer learning is adopted by many researchers as a promising strategy. In transfer learning, a CNN model pre-trained on specific dataset, 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 [20]. Due to this powerful concept, researchers started using transfer learning for numerous tasks

978-1-7281-1929-8/18/$31.00 ©2018 IEEE
DOI 10.1109/EECS.2018.00109

of images classifications, as well as action recognition [21] [22]. Transfer learning has optimal strategies for fine-tuning network, in which a pre-trained network has to be fine-tuned on target dataset to successfully perform the new task in new domain [23].

Despite to the fact that deep learning techniques are successfully used for human action recognition, however these techniques in coordination with transfer learning concept have not been considered by researchers for violence detection. This research work proposed transfer learning based deep CNN model to detect violent/fight activates in videos sequences. GoogleNet [18] is selected due to deep network architecture with 12 times fewer parameters than AlexNet as a pre-trained model. It is fine-tuned on Hockey and Movies datasets using transfer learning to create a deep representation classifier for violent scenes identification. Results show that proposed approach is out performing on both datasets as compare to all competitive state-of-the-art published approaches from hand-crafted and deep learning domains. Following paper is organized in Related Work, Methodology, Datasets, Experiments, Results, and Conclusion in section II, III, IV, V, VI, and VII respectively.

II. Related Work

Initial proposals adopted the methodology of violence recognition using blood and flame detection, capturing the degrees of motion, recognizing sounds features by exploiting audio-visual correlation, skin and blood patterns exposure and discovering scream like cues in audio exploiting audio-video correlation for violent scenes detection [11], [24]–[26]. Then, audio features are used to detect gunshots, explosions and car-breaking activities, using Hidden Markov models (HMM) and Gaussian mixture models [12]. Audio characteristics from time and frequency domain are classified using Support Vector Machine (SVM) [27].

Furthermore, Chen et al. used spatio-temporal video cubes and local binary motion detectors [28]. Lin and Wang exploited weekly-supervised audio classifier co-trained with video features of motion, blood and explosion [29]. Giannakopoulos et al. performed audio-visual features analysis using statistics, average motion followed by K-Nearest Neighbors (KNN) classifier [30]. And, Chen et al. suggested detection of faces and blood presence [31] for determining potential violent contents in videos.

Bermejo et al. exhibited encouraging results with 90% accuracy using MoSIFT feature descriptor revealing two potential datasets "Hockey dataset" and "Movies dataset", specifically designed for violence detection job [6]. Following that, Kernel Density Estimation (KDE) was exploited to obtain feature selection on MoSIFT descriptor with sparse coding reporting accuracy 94.3% on Hockey dataset determining aggressive human behaviors [32]. Another approach describes fuzzy region emerges in image frames due to abrupt violent motion patterns, reporting 98.9% accuracy on Movies dataset [9]. Motion blob, another form of motion features is used to discriminate fight and non-fight video frames, by extracting basic features of blobs (perimeter, area etc.) yielding 97.8% accuracy [10].

Recently, 3D ConvNets based model with the prior knowledge is investigated on Hockey dataset [14]. 3D Convolutional Neural Network architecture C3D [33] is experimented on Hockey and Movie sequences [13]. More recently, a 2D-CNN model using Hough Forest features is proposed. This system is revealing finest accuracy results 94.6%, 99% on Hockey and Movies datasets respectively, as compare to all previous techniques of hand-defined features detectors and deep representation models [13].

In short, a significant number of previous algorithms perform audio-visual cue analysis by recognizing audio cues for violent activities or by examining blood and flame visual color patterns using hand-defined features. A few deep learning based approaches also incorporated CNN and 3D-CNN architectures. Moreover, a 2D-CNN model, by taking advantage of network deep representations, in combination with hand-defined Hough Forest features, is constructing finest classifier to discriminate violent human behaviors. However, deep learning models have certain limitations. They require huge computational power with enormous amount of domain specific data. Developing huge amount of labeled dataset is laborious and time-consuming task. This shortcoming is leading to a major bottleneck in training deep learning model from scratch for target domain. To swiftly combat this challenge, an approach of transfer learning becomes useful. In which, a source network pre-trained on a huge dataset is re-trained on the target domain specific dataset [20]. This scheme eliminates the need of producing huge dataset as well as training model from scratch. In this regards, winner models of ImageNet Large Scale Visual Recognition Challenge (ILSVRC), such as AlexNet [17] VGGNet [34], GoogleNet [18] and ResNet [19] trained on 15 million annotated images for 1000 categories, are fortunately publically available as open source pre-trained models. These models can be used as pre-trained networks employing transfer learning to develop domain specific target networks, such as for the task of violent human behaviors detection.

III. Methodology

In machine learning domain, learning based representation techniques achieve feature learning through iterative optimization procedure. Feature learning is very appealing due to learning complex underlying data representation, especially for complex task of image recognition, as compare to hand-crafted feature descriptors. The learnt features acquired through learning a specific problem, can be re-utilized for solving another problem in a new task, a concept known as transfer learning. This approach has been successfully used in object classification and categorization domain [35].

The CNN deep model is originally data driven, it requires large labeled dataset for training. Annotated dataset preparation is complex and demanding task. On the other hand, providing insufficient amount of data would not leverage CNN model to learn optimal deep features instead network suffer from significant overfitting issue. To solve the problem of overfitting for small dataset, utilizing modern deep learning network architectures, the approach of transfer learning comes into play. In which, existing network architecture with pre-trained learned features as source task network is employed to build new target task network architecture for limited dataset [36]. Fig. 1. shows general representation of source task network, with convolutional blocks followed by dense fully connected subsequent layers, pre-trained on ImageNet with 1000 output classes. The source task network is utilized for transfer learning to create a target task network, to be trained on Hockey and Movies dataset with 2 output classes for violent/fight and non-fight activities.

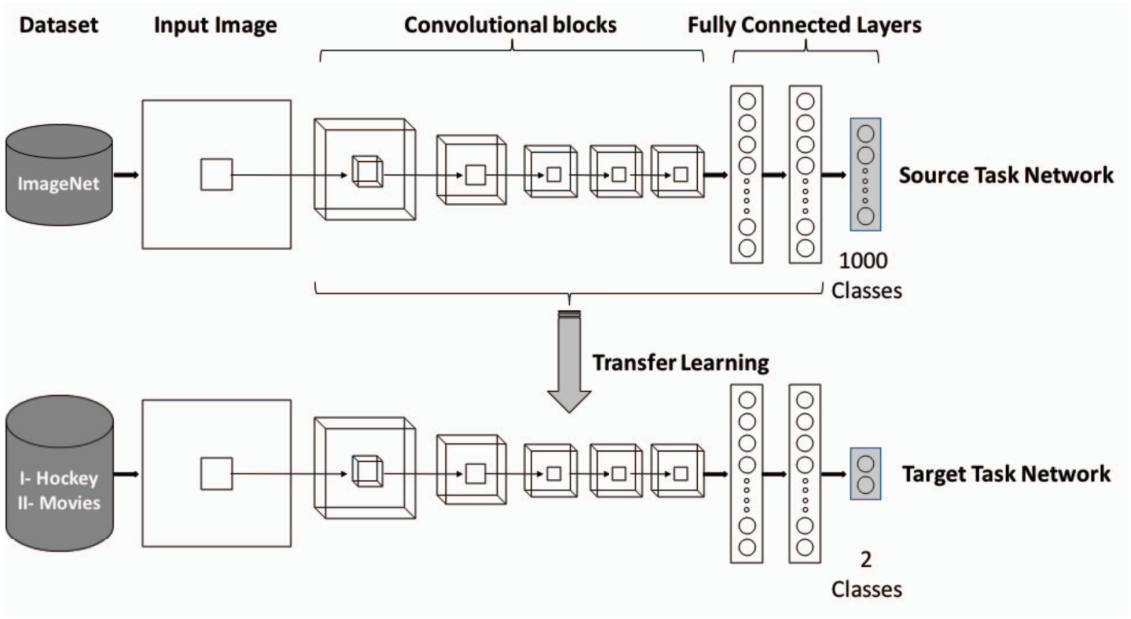

Fig. 1. Transfer learning concept with source task network architecture transformed into target task network architecture. A source task network pre-trained on source dataset is fine-tuned on target dataset to be a target task network

In this paper, GoogleNet [18] is selected as a pre-trained source network architecture with learned features from ImageNet dataset on 15 million annotated images for 1000 categories. GoogleNet codenamed Inception is a 22 layers deep network, composed of repeated inception modules. Although network is 22 layers deep but it has 12 times fewer parameters than AlexNet to make it an efficient deep neural network architecture for computer vision. Based on these characteristics GoogleNet is selected for transfer learning experiments, by removing last dense fully connected classification layer classifying 1000 ImageNet classes and replacing it with 2 classes of Hockey and Movies dataset to discriminate violent/fight actions from non-fight.

Fig. 1. shows overall scheme of transfer learning with source task network architecture transformed into target network architecture, classifying 2 output classes of target datasets.

IV. Datasets

The experiments are conducted on two benchmark violence activity detection Hockey and Movies datasets [6].

The Hockey dataset is first of its kind specifically designed for fight activity recognitions. It has 1000 video clips of 360x288 resolution. Dataset is further sub-divided into two categories, fight and non-fight, with each category containing 500 clips. The dataset is obtained from National Hockey League (NHL) of hockey games with real-life violent events.

The Movies dataset is also made particularly for fight activity detection. It has 200 video clips for both fight and non-fight activities. Fight scenes are extracted from different actions movie clips. Whereas, non-fight scenes are extracted from publically available action recognition datasets. Unlike Hockey dataset, this dataset has collection of wide range of diverse scenes recorded at different resolutions on different occasions, with an average resolution of 360x250 pixels.

The Hockey dataset is challenging due to abrupt camera motion in recording non-fight scenes of real-time hockey games. Movies dataset has views complexities due to diverse collection of scenes, exhibiting variations in background with different illumination conditions and occlusions. The challenging characteristics of both Hockey and Movies datasets is making them best suitable source for the task of violence recognition. See Fig. 2. for dataset details.

V. Experiments

This section describes the experimental details of proposed methodology of learned features representation model using transfer learning as discussed in section III.

A. Experimental setup

The GoogleNet model learn spatial features from images as 2D deep CNN network. This model can be trained on videos dataset by converting annotated video clips into labeled images sequences. As pre-training step both videos datasets are converted into corresponding frames to efficiently train deep model.

Hockey dataset with 1000 video clips generated 41056 annotated images for both activities of fight and non-fight. Where each image represents adjacent video frames. Similarly, Movies dataset with 200 video clips is converted into 9841 annotated images of adjacent frames for fight and non-fight actions.

B. Parameters

Network is fine-tuned on both datasets using batch size 64, constant learning rate 0.0001, with momentum set to 0.9. Based on experiments 5 epochs training scheme is adopted to find optimal results on both datasets. Due to significant size of target datasets, network is fine-tuned by back propagating errors throughout the network to previous layers. During training GoogleNet is provided with resized images pipeline of 224x224 pixels for each fold. Experiments are conducted on NVIDIA 1080 ti GPU.

Fig. 2. Hockey and Movies datasets samples. First row, left two images are fight and right two images are non-fight frames from Hockey dataset. Second row, left two images are fight and right two images are non-fight frames from Movies dataset.

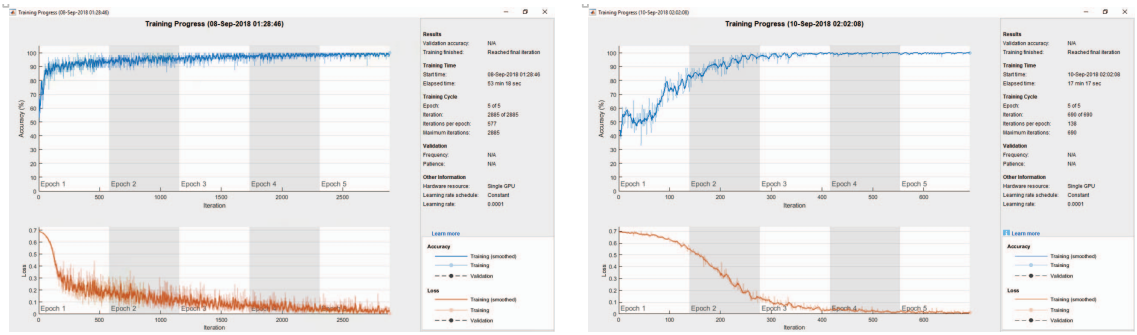

Fig. 3. Hockey and Movies datasets training progress samples for 5 epochs of 1st fold. Left image shows 1st fold training progress of Hockey dataset, Right image reports 1st fold training progress of Movies dataset.

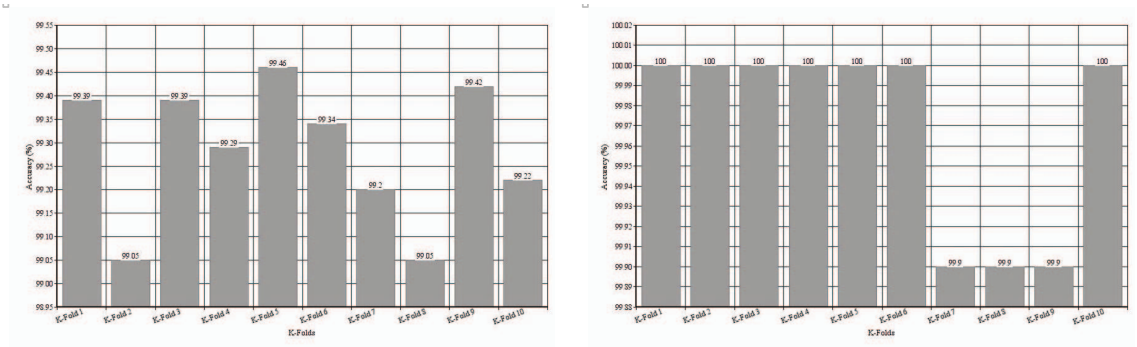

Fig. 4. Hockey and Movies datasets accuracies for 10-Folds. Left image shows accuracies for each fold of Hockey dataset, Right image reports accuracies for each fold of Movies dataset.

C. *Training progress*

The distinct GoogleNet models are trained for each fold separately. Models are trained in sequential order with each fold trained for 5 epochs, on both datasets.

Fig. 3. shows training progress for 5 epochs of 1st fold on Hockey and Movies datasets. Accuracy indicators shows drastic increase in accuracy by reducing loss in very 1st and 2nd training epoch. Model approaches to highest accuracy quickly in 4th and 5th epoch. So, training is stopped at this stage to avoid network overfitting problem.

Fig. 4. show accuracy progress achieved at the end of each fold. 10 distinct accuracy values are reported for all 10-folds. Hockey dataset with abrupt camera motions in video frames is producing arbitrary highest accuracy values for different folds. However, Movies dataset is achieving more consistent highest accuracy rate for each fold.

TABLE I. COMPARISON OF CLASSIFICATION ACCURACIES RESULTS ON HOCKEY AND MOVIES DATASETS

Year	Author/Method	Features/Classifiers	Testing Scheme	Datasets Accuracy (%)	
				Hockey	*Movies*
2013	Bermejo et al. [6]	STIP (HOG) + HIK	5-Fold CV	91.7± -%	49.0± -%
		STIP (HOF) + HIK		88.6± -%	59.0± -%
		MoSIFT + HIK		90.9± -%	89.5± -%
2014	Deniz et al. [9]	SVM	10-Fold CV	90.1±0%	85.4±9.3%
		Adaboost		90.1±0%	98.9±0.2%
2014	Ding et al. [14]	3D-CNN	Train/Test split	91± - %	-
2015	ViF [10]	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%
2015	LMP [10]	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%
2015	Serrano [10]	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%
2018	Serrano et al. [13]	C3D [33]	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%
-	**Proposed Method**	Deep CNN using Transfer learning	10-Fold CV	**99.28%**	**99.97%**

VI. RESULTS

The proposed method is evaluated against two benchmark violent activity recognition datasets. i.e., the Hockey and Movies datasets [6]. To perform a comprehensive comparison a wide range of existing algorithms are taken from literature with benchmark accuracies results, from both domains of hand-defined features detectors and deep representations based models, as discussed in section II.

In hand-defined domain; Bermejo et al. proposed method achieved 90% as benchmark accuracy with the introduction of Hockey and Movies datasets [6]. Following that, Deniz et al. suggested technique using SVM and Adaboost reported 98.9% accuracy on Movies dataset [9]. Later on, The Violent Flows (ViF), LMP methods using SVM, Adaboost and Random Forests classifiers are reported [10].

Moreover, in deep learning domain; Ding et al. implemented 3D-CNN model with train/test split scheme [14]. In recent times, Serrano et al. evaluated C3D and 2D-CNN models. Author further proposed finest approach incorporating 2D-CNN with Hough Forest features. This approach elevated accuracies to 94.6±0.6%, 99±0.5% for Hockey and Movies datasets correspondingly, setting the accuracy bar to the next level [13].

Thereby, the proposed approach of transfer learning using GoogleNet deep model with already learnt features is compared, to assess the performance of suggested methodology against established techniques. See Table I, results are formulated as mean of accuracy with 10-fold cross validation scheme as described in training progress part of section V.

Finally, the proposed approach outperforms state-of-the-art accuracies on both datasets. Results show highest accuracies 99.28% and 99.97% on Hockey and Movies datasets respectively. The proposed strategy specifically improved Hockey dataset accuracy by learning generalize deep features for abrupt camera motion sequences, as compared to benchmark techniques. Similarly, model is able to distinguish violent action in a wide variety of movie clips with dynamic scenes, on Movies dataset.

Foremostly, the proposed model achieved superior accuracy results in just 5 training epochs on both datasets, eventually reducing the effort required to train a model on target dataset.

VII. CONCLUSION

Violent action detection such as fight scene recognition has attracted computer vision researchers during last few years, because detection of aggressive human behaviors is preliminary requirement to develop automated video surveillance systems. Historically, violent action recognition tasks are usually achieved through hand-crafted feature detectors. However, some approaches also proposed deep learning based models to detect aggressive human behaviors. Although, deep representation based transfer learning approaches have been used for human action recognition such as walking, jogging, running, hand waving. However, there is scarcity in using transfer learning based deep model for violent sequences detection. To train model, Hockey and Movies datasets are first of their kind specifically designed for violent/fight action recognition, as compare to other available human action detection datasets.

In this research, the learned representation based deep CNN model is proposed to identify aggressive behaviors in videos. Since training a deep network from scratch ends up facing network overfitting issues. Therefore, an alternative transfer learning training strategy is adopted. GoogleNet; a very deep network architecture is adopted as a source task network, pre-trained on ImageNet dataset with 15 million annotated images for 1000 categories. By incorporating the concept of transfer learning, source network is utilized to develop target task network. Which is then fine-tuned on Hockey and Movies datasets with discussed optimal parameters. The proposed model is trained using 10-fold cross validation scheme, by developing dataset images pipeline with images resizing, as input to fine-tuning network for each distinct fold. Distinct models are trained for 5 epochs, for each fold, producing mean accuracy results across 10-folds, on both datasets.

Results show that, proposed model is outperforming top ranked approaches by learning finest features on challenging datasets. Model achieved 99.28% and 99.97% accuracies on Hockey and Movies datasets respectively, in just 5 training epochs. Proposed method is able to learn most discriminating features for abrupt camera motions and dynamic scenes sequences, for the task of violent action detection in videos.

ACKNOWLEDGMENT

This research is supported by the PDE-GIR project which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 778035.

REFERENCES

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proceedings - International Conference on Pattern Recognition*, 2004.

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

[3] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, 2009.

[4] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[5] S.-R. Ke, H. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A Review on Video-Based Human Activity Recognition," *Computers*, 2013.

[6] E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques," *CAIP'11 Proc. 14th Int. Conf. Comput. Anal. images patterns - Vol. Part II*, 2011.

[7] A. Sargano, P. Angelov, and Z. Habib, "A Comprehensive Review on Handcrafted and Learning-Based Action Representation Approaches for Human Activity Recognition," *Appl. Sci.*, vol. 7, no. 1, p. 110, 2017.

[8] D. Wu, N. Sharma, and M. Blumenstein, "Recent Advances in Video-Based Human Action Recognition using Deep Learning: A Review," pp. 2865–2872, 2017.

[9] O. Deniz, I. Serrano, G. Bueno, and T.-K. T. Kim, "Fast violence detection in video," in *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, 2014, vol. 2, pp. 478–485.

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

[11] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, 1998, vol. 1, pp. 353–357.

[12] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic context detection based on hierarchical audio models," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, 2003, pp. 109–115.

[13] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," *IEEE Trans. Image Process.*, 2018.

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

[15] P. Zhou, Q. Ding, H. Luo, X. Hou, B. Jin, and P. Maass, "Violent Interaction Detection in Video Based on Deep Learning," in *Journal of Physics: Conference Series*, 2017, vol. 844, no. 1, p. 12044.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "lecun- at el - Gradient-based learning applied to document," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.

[18] C. Szegedy *et al.*, "Going Deeper with Convolutions," pp. 1–9, 2014.

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

[20] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8689 LNCS, no. PART 1, pp. 818–833, 2014.

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," *Comput. Vis. Pattern Recognit. (CVPR), 2014 IEEE Conf.*, pp. 1725–1732, 2014.

[22] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," *2017 Int. Jt. Conf. Neural Networks*, pp. 463–469, 2017.

[23] J. Donahue *et al.*, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," 2013.

[24] C. Clarin, J. Dionisio, M. Echavez, and P. Naval, "DOVE: Detection of movie violence using motion intensity analysis on skin and blood," *PCSC*, vol. 6, pp. 150–156, 2005.

[25] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila, "CASSANDRA: Audio-video sensor fusion for aggression detection," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007 Proceedings*, 2007, pp. 200–205.

[26] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao, "Detecting violent scenes in movies by auditory and visual cues," in *Pacific-Rim Conference on Multimedia*, 2008, pp. 317–326.

[27] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," in *Hellenic Conference on Artificial Intelligence*, 2006, pp. 502–507.

[28] D. Chen, H. Wactlar, M. Chen, C. Gao, A. Bharucha, and A. Hauptmann, "Recognition of aggressive human behavior using binary local motion descriptors," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 5238–5241.

[29] J. Lin and W. Wang, "Weakly-supervised violence detection in movies with audio and video based co-training," in *Pacific-Rim Conference on Multimedia*, 2009, pp. 930–935.

[30] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos," in *Hellenic Conference on Artificial Intelligence*, 2010, vol. 6040, no. December, pp. 91–100.

[31] L.-H. Chen, C.-W. Su, and H.-W. Hsu, "Violent scene detection in movies," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 08, pp. 1161–1172, 2011.

[32] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014, pp. 3538–3542.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[34] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," pp. 1–14, 2014.

[35] Y. Aytar, "Transfer learning for object category detection," University of Oxford, 2014.

[36] Y.-C. Su, T.-H. Chiu, C.-Y. Yeh, H.-F. Huang, and W. H. Hsu, "Transfer Learning for Video Recognition with Scarce Training Data for Deep Convolutional Neural Network," *arXiv Prepr. arXiv1409.4127*, 2014.