I.J. Education and Management Engineering, **2019, 1, 9-19**
Published Online January 2019 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2019.01.02

Available online at http://www.mecs-press.net/ijeme

Aggressive Action Estimation: A Comprehensive Review on Neural Network Based Human Segmentation and Action Recognition

A. F. M. Saifuddin Saif [a], Md. Akib Shahriar Khan [a], Abir Mohammad Hadi [a], Rahul Prashad Karmoker [a], Joy Julian Gomes [a]

[a]*Faculty of Science and Technology, American International University – Bangladesh (AIUB), Dhaka, Bangladesh*

Received: 11 October 2018; Accepted: 17 December 2018; Published: 08 January 2019

Abstract

Human action recognition has been a talked topic since machine vision was coined. With the advent of neural networks and deep learning methods, various architectures were suggested to address the problems within a context. Convolutional neural network has been the primary go-to architecture for image segmentation, flow estimation and action recognition in recent days. As the problem itself is an extended version of various sub-problems, such as frame segmentation, spatial and temporal feature extraction, motion modeling and action classification as a whole, some methods reviewed in this paper addressed sub-problems and some tried to address a single architecture to the action recognition problem. While being a success, convolution neural networks have drawbacks in its pooling methods. CapsNet, on the other hand, uses squashing function to determine the activation. Also it addresses spatiotemporal information with the normalized vector maps while CNN-based methods extracts feature map for spatial and temporal information and later augment them in a fusion layer for combining two separate feature maps. Critical review of papers provided in this work can contribute significantly in addressing human action recognition problem as a whole.

Index Terms: Capsule Network, Neural Network, Image Segmentation, Flow Estimation, Action Recognition.

1. Introduction

Action recognition in videos can have a radical impact on human life. Numerous attempts have been taken to solve the action recognition challenges. Due to huge collaborative efforts in computer vision community,

* Corresponding author:
E-mail address: saif@aiub.edu, akeeebkhan@gmail.com, abir45pro@gmail.com, karmoker.rahul4@gmail.com, joyjuliangomes@gmail.com

simple actions of waving, standing etc. from KTH and Weizmann dataset are now considered as obsolete challenges and the community has moved on to solve more complex actions like sports and human interactions. However, despite having the potential to improve security and surveillance applications, there has not been much improvement in regard to the violent scene and aggressive behavior detection which is a special case of action recognition. Previous works on action recognition heavily relied on the usage of hard-coded techniques such as MoSIFT, Optical Flow and Dense Trajectory. These hard-coded techniques are computationally expensive while offering low performance. In recent times after the success of AlexNet, a wave of works approached the problem from a new viewpoint using convolutional neural networks. Though being incisive in image classification tasks, Convolutional Neural Networks did not fare well immediately against already established methods in action recognition. Different types of fusion techniques using both dense features and CNN improves performance. Two-stream networks and 3D-CNN using motion features such as optical flow and RNN in conjunction with the aforementioned techniques also gave a boost in performance. But these approaches have some severe disadvantages like max-pooling which in most cases suppresses tiny but important features and, susceptible to adversarial attacks. Though these methods work, they do not provide any insight into how the inner mechanism functions. The newly proposed CapsNet architecture can help to bridge the gap as this particular system follows a part-to-whole approach and produces vector outputs, unlike CNN which has scalar outputs. Capsules are particularly good at handling different types of visual stimulus and encoding things like pose (position, size, and orientation), deformation, velocity, albedo, hue, texture etc. that is not possible for CNN. Capsules encapsulate all-important information about the state of the feature they are detecting in vector form.

The rest of the paper is organized as follows. Section 2 discusses the challenges of action recognition and provides a concise view of a broad range of technologies and approaches that are used to solve the problem. In section 3 numerous methods related to action recognition are reviewed. Section 4 elaborates the frameworks used in the method described in section 3. Section 5 provides details on the experimental settings and performance of the methods. In section 6 key findings from the methods are summarized. Section 7 concludes the paper emphasizing the impact of the problem.

2. Core Background Study

Human action recognition is an integral problem in spatiotemporal information extraction, fusion, learning and detection from video streams, both in static and especially in a live feed analysis. Numerous studies have been conducted based on hard-coded feature extraction, pose estimation, frame and dynamics and also as neural network learning problem. The problem in discussion is addressed by sub-problems that include: frame preprocessing (if any), feature extraction (both spatial and temporal), learning the feature sets and classification. Further dividing the problems of feature extraction includes background subtraction for subject(s) isolation and background dynamics for temporal information gathering. For learning the problem is subdivided into two categories: individual action and social action where the individual's actions are considered as a collective action within the context.

Feature extraction is addressed by single-frame detection for the subject identification (E.g.: Human) and multi-frame detection for the flow field estimation that gives subjects' collective movement modeling. Datta et al. (2002) [1] addressed the collective problem with hard-coded feature extractor for a probabilistic model generation which used adaptive background subtraction for human silhouette segmentation and color sum square difference for motion modeling. They used Acceleration Measure Vector and jerk motion modeling which classified action as violent or not. Bagautdonov et al. (2017) [2], addressed the individual person action recognition problem by dense feature representation to obtain feature maps and further refining by inference in a hybrid Markov Random Field. Feature maps were then passed through RNN layer to analyze according to temporal domain. Xiao et al. (2017) [3], used autoencoder to train for both human body segmentation and motion modeling. Then a deep network is used to perform action recognition.

Zhu et al. (2018) [4], also used encoder-decoder networks to address the problems. In Jégou et al. (2018) [5], dense feature maps are also used and fully convolutional DenseNet is used to get the output.

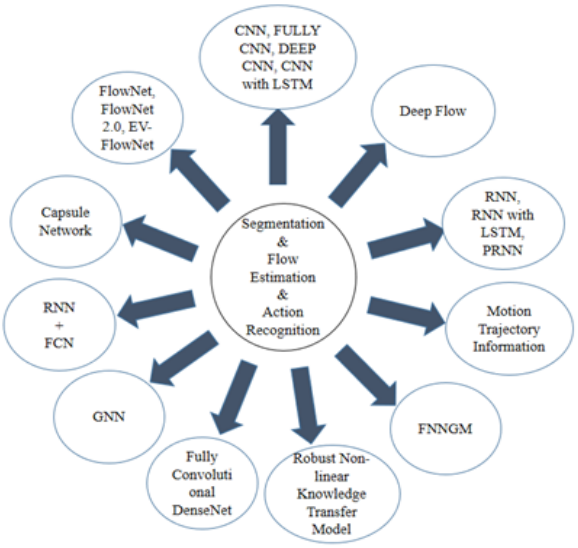

Fig.1. Existing Methods for Segmentation and Action Recognition.

Baccouche et al. (2011) [6] used 3D-CNN to train the feature and motion model and RNN with LSTM were used to perform action recognition. Sun et al. (2017) [7] also used CNN for two-stream network which independently calculates spatial and temporal features. It applied recurrent attention mask to regularize and the modified RNN produced more complex and intricate motion features. Chen et al. (2017) [8] addressed action recognition in extremely low-resolution videos using a semi-coupled ConvNets which share some common filters and some stream-only filters to train and test the data. Dosovitskiy et al. (2015) [9], and Ilg et al. (2015) [10] presented optical flow estimation as a learning problem and uses ConvNets to predict the optical flow. Ilg et al. (2015) [10], further extended the problem to data scheduling showing an interesting insight that scheduling training data improved the performance on most of their tested datasets. Pigou et al. (2018) [11] addressed spatial and temporal information extraction and learning using a strategic temporal pooling using ConvNets with average pooling and bidirectional RNN sequence. Luvizon et al. (2018) [12] used stacking of 2D pose probability heat maps to generate 3D pose estimation and uses Fully Connected Networks to aggregate the pose-based and appearance-based action estimations. Oyedotun et al. (2017) [13] used multilayer CNN and Stock Denoising Autoencoder layers to train and testing. In Wang et al. (2017) [14], authors tried to mitigate the misclassification arose using two-stream networks where they used Spatiotemporal Compact Bilinear Fusion (STCB) that can back-up spatial stream with temporal stream and vice versa. Rahmani et al. (2018) [15], Zhang et al. (2017) [16], Zhang et al. (2017) [17], Li et al. (2018) [18] and Li et al. (2018) [19] used skeleton information to train their system. Rahmani et al. (2018) [15] used dense trajectory calculation and encoded the shape of the trajectory for training and learned through knowledge transfer model while Zhang et al. (2017) [16] and Zhang et al. (2017) [17] utilized LSTM to extract and learn temporal dynamics. Yue-Hei Ng et al. (2015) [20] used stacked LSTM layer with SoftMax for prediction. Sabour et al. (2017) [21] used a completely different approach than all other literature mentioned above.

Recent development in feature detection and extraction has seen a massive leap ahead with the invention of

CapsNet by Sabour et al. (2017) [21]. This method allows tiny features to be detected in low level and provides vector data rather than a scalar. The vector data holds key to improve optical flow estimation. With orientation and magnitude of feature data being embedded in the capsule output, it is possible to calculate the difference in orientation and magnitude between two frames. The proposal is to develop a system to classify the extracted features as aggressive or nonaggressive behavior based on the rate of change of orientation of the interested feature vector.

3. Review Based on Methods

In this section, we first give an overview of different methods for segmentation. Then respectively methods of action recognition were introduced.

3.1 Segmentation:

Jégou et al. (2017) [5] worked on object segmentation based on powerful ResNet architecture but fails to provide insight into the process. Sabour et al. (2017) [21] used dynamic routing instead of pooling and thus building a hierarchical structure among the layers which prevents information loss and enables robust segmentation. SegCaps by LaLonde et al. (2018) [22], is a CapsNet based segmentation method which proposed constrained dynamic routing. Afshar et al. (2018) [23] used a reduced version of CapsNet for medical image segmentation.

3.2 Action Recognition:

Datta et al. (2002) [1] made use of hard-coded features and mathematical conditions for violence detection. MoSIFT Chen et al. (2009) [24] is an extension of SIFT features for videos where SIFT method was used to extract interest point, later action was recognized by creating optical flow pyramid. Also, Nievas et al. (2011) [25] used the Bag-of-Words technique for action recognition with better spatiotemporal information.

FlowNet Dosovitskiy et al. (2015) [9] and DeepFlow by Weinzaepfel et al. (2017) [26], introduced models of CNN to estimate optical flow. FlowNet 2.0 by Ilg et al. (2015) [10], improved the previous model using 2 channels for large and small displacements. Cascade Residual Learning Pang et al. (2017) [27] improved optical flow estimation by utilizing two-stage disparity computations. EV-Flow Zhu et al. (2018) [4] made use of skip connection introduced to estimate optical flow for images from event based camera. All of these techniques rely heavily on motion computation blocks of CNN. CapsNet by Sabour et al. (2017) [21] produced vector activation outputs from the primary capsule layer where each element in the vector represented different state and properties of the input. In this case, CapsNet can produce optical flow estimation without relying on secondary computational unit.

CNN architectures have delivered good results in computer vision problems such as image classification and segmentation. GoogleNet and ResNet models are most popular among CNN architectures because of depth. Though they provide good result, their usage of pooling technique limits the ability of fully utilizing input data. In Table 1, different types of strategy are analyzed considering convolutional neural network for action recognition.

While CNN works on spatial data, RNN can process sequence of inputs. It is composed of neurons connected in a successive sequence where middle layers work as memory cells. Bagautdonov et al. (2017) [2] used RNN extensively. It produced refined detection proposals from scaled CNN outputs which were used by RNN to determine individual and collective action. Though RNN inspired techniques yields convincing results, they are sophisticated and have a high complexity and gradient vanishing problem which makes training a model computationally expensive. On the contrary, LSTM used by Zhang et al. (2017) [16]. Zhang et al. (2017) [17] used an efficient variant of RNN which solves the gradient vanishing problem by introducing exclusive

memory cell and forget gate. Zhang et al. (2017) [16] dynamically adjusted the position of a subject through adaptive LSTM by creating a new virtual camera viewpoint. In Zhang et al. (2017) [17], time-varying geometric shapes extracted from human joints are input to 3 layers LSTM for action recognition.

Xiao et al. (2017) [3] trained autoencoder using deep learning techniques to extract features which are later used by Pattern Recognition Neural Network (PRNN) to detect human action. In Li et al. (2018) [19], authors proposed a new graph based neural network with a fully connected layer for action analysis, where action attending layer produces salient action units. While in Edwards et al. (2016) [29], graph theorem on irregular domain was used in conjunction with convolution and pooling for image processing.

Table 1. Qualitative study of different methods based on Convolutional Neural Network

Method	Strategy	Advantage	Disadvantage
CNN [13, 18, 8, 27, 5], Ren et al. (2015) [27]	Deep Learning	1. Deep learning networks are more optimized and dynamic in nature in recognizing the gestures. 2. Contains more spatial temporal information and is easier to classify 3. Generates high-quality disparities for the inherently ill-posed regions.	1. CNN use max-pooling which in most cases suppresses tiny but important features. 2. robust left-right consistency check module problem 3. Complicated system for image segmentation.
CNN with LSTM [12]	Long Short-Term Memory	1. Single architecture to solve the action recognition and pose estimation problem.	1. Uses max-pooling which throws away all but most active feature. 2. Still computationally expensive to train.
CNN with RNN and LSTM [6, 7, 11]	Recurrent Neural Network with Long Short-Term Memory	1. Two-steps scheme automatically learns spatiotemporal features and uses them to classify the entire sequences 2. Learns complex motion features well.	1. The Accuracy may decrease for crowded cases with different complex human action datasets. 2. Computationally expensive 3. Softmax classifier may overlook and misclassify in a range of action provided.
CNN with FCN [15, 14]	Fully Connected Network	1. Computationally efficient than other available methods 2. Better at classifying unseen video 3. Dynamic spatiotemporal fusion.	1. Adding a new class has a time complexity of training a SVM

4. Review Based on Frameworks

The segmentation problem in image processing and computer vision has been tackled by many algorithms. Before the use of neural network was common, there were algorithms like SIFT (Scale Invariant Feature Transform) and SURF (Speeded-Up Robust Features). The main problem with them is their adaptability and robustness. Since the neural network age began, Convolutional neural networks have been used for segmentation by passing a frame through multiple layers with varied kernel size and strides. But as CNN use pooling strategies, they often leave out the important information that doesn't cut the level of the pooling values. This can be solved using the squashing of the vector fields generated by CapsNet in the final layer which embeds each vector and normalizes the magnitude leaving all the information intact.

Dosovitskiy et al. (2015) [9], Ilg et al. (2015) [10], Zhu et al. (2018) [4], Weinzaepfel et al. (2013) [26], all tried to address optical flow as a learning problem. Dosovitskiy et al. (2015) [9] and Ilg et al. (2015) [10] both used specialized CNN in a two-stream form to extract and augment the spatial and temporal images. But they

are prone to small displacement and sub-pixel motion error. Zhu et al. (2018) [4] used an encoder-decoder network built with CNN and an event-based input feed which is not robust in unseen cases and prone to the pooling error. CapsNet can play a vital role in extracting the motion features and model it through normalized vector field.

Datta et al. (2002) [1] is a good example of hard-coded feature extraction and motion classification where frames are segmented based on color and actions were classified using Acceleration Measure Vectors and Jerk motion. This is not robust to unseen conditions which can be countered by CapsNet as it's a learning method.

Chen et al. (2017) [8], Luvizon et al. (2018) [12], Oyedotun et al. (2017) [13], Li et al. (2018) [17], Pang et al. (2017) [27] used CNN for segmentation of video stream and classify the actions within. Chen et al. (2017) [8] used two-stream CNN, one for spatial and another for temporal information extraction and augment them using fully connected convolutional layer. It's not robust to spatial and temporal information mismatching which was later addressed by Wang et al. (2017) [14]. They used a Spatiotemporal Compact Bilinear Fusion to dynamically extract spatiotemporal information that used average pooling for activation that might have left important information due to the pooling strategy. Luvizon et al. (2018) [12] also used max pooling in several layers of their temporal convolution network. Pang et al. (2017) [27] used CNN based deep learning hence prone to pooling error that can be solved by CapsNet squashing activation.

Rahmani et al. (2018) [15] and Wang et al. (2017) [14] used CNN with Fully Connected Networks (FCN) to segment and classify action from videos. [14] Used CNN to extract spatial and temporal information using STCB and later uses the same method to augment the spatiotemporal features. [15], however, took a different approach using a knowledge transfer model which is built using novel viewpoints of 3D mocap data. Both have high computational complexity and prone to pooling error, moreover, [15] has a lower accuracy of <75%. Using CapsNet will give a single architecture and end to end training with its recursive neuron models also known as capsule which hierarchically extracts high-level information through segmentation and augmentation of the vector fields.

Bagautdonov et al. (2017) [2], Pigou et al. (2018) [11], Sun et al. (2017) [7], Baccouche et al. (2011) [6] used CNN and RNN with LSTM memory cell. Pigou et al. (2018) [11] and Baccouche et al. (2011) [6] used RNN with LSTM to learn and classify actions in a sequential manner where Sun et al. (2017) [7] used a novel Lattice-LSTM which can learn independent memory cell transition. Bagautdonov et al. (2017) [2] deployed one LSTM layer to identify person level dynamics and another one to aggregate scene level information. Bagautdonov et al. (2017) [2], Pigou et al. (2018) [11] and Baccouche et al. (2011) [6] may suffer from the vanishing gradient problem which was address using L-LSTM but with a greater computational expense. This can be overcome using recursive technique of the CapsNet which sequentially and recursively process information from one level to another.

Zhang et al. (2017) [17] deployed a stacked layer of LSTM with RNN models for spatiotemporal information extraction with a tree-based traversal method. LSTM, though solved the vanishing gradient problem, requires a lot of computational power and cannot store large amount of sequences. This can be solved using CapsNet's instantaneous feature extraction and vector modeling in a hierarchical way.

Li et al. (2018) [19] and Edwards et al. (2016) [29] used a graph based approach. Edwards et al. (2016) [29] tried to solve the signal processing techniques for convolutional network on irregular domain problem. These methods are good for single human body feature detection and are not suitable for crowd-facing. Due to the robustness to variation in posture and gesture, CapsNet will be a suitable candidate in this matter.

Xiao et al. (2017) [3] used encoder-decoder model using backpropagation to train the autoencoders which encode the features of the frame and constructs a pattern recognition network. This requires a lot of computational power to encode and decode the models.

Jégou et al. (2017) [5] used a fully convolutional DenseNet to address the problems. Due to iterative feature map concatenation in the dense block, this requires the gradients to be passed through networks of different depth (with different numbers of nonlinearities) forcefully.

5. Review Based on Experimental Results

For segmentation, using the CapsNet in brain tumor MRI images in Afshar et al. (2018) [23], has the highest accuracy of 86.56% with one convolutional layer including 64 feature maps. An added layer of SegCaps in Capsule Network architecture LaLonde et al. (2018) [22] slightly outperformed all other compared approaches with an average accuracy 98.479% for performing segmentation on LUNA 16 lung images. In terms of action recognition, estimation of large displacement in optical flow field, all the versions of FlowNet 2.0 from Ilg et al. (2015) [10] performed better than typical FlowNetS by Dosovitskiy et al. (2015) [9]. Among them, FlowNet2-CSS-ftsd has the highest accuracy rate of 79.64% is very close to the state-of-the-art DeepFlow from Weinzaepfel et al. (2013) [26] accuracy rate of 81.89% while FlowNetS has 55.27%. Zhu et al. (2018) [4] tested their method on Multi-Vehicle Stereo Event Camera dataset (MVSEC). They used qualitative evaluation on the dataset. In Jégou et al. (2017) [5], proposed two-steps sequence labeling scheme achieves an overall accuracy of 94.39% on KTH1 and 92.17% on KTH2. Following Table 2 summarizes the performance of other models:

Table 2. Performance Summarization of Different Methods for Different Datasets

Method	Reference	Dataset	Accuracy(%)
CNN	[26]	Middlebury 2009	*AEE (0.42)
CNN with Convolutional Fusion at Conv3 layer with eLR and HR shared filter	[8]	IXMAS	93.7
CNN	[18]	NTU RGB+D (Cross-View)	82.1
CNN with DispResNet	[27]	Kitti 2015	*AEE (0.68)
CNN with LSTM	[12]	Penn Action	98.6
Robust Non-linear Knowledge Transfer Model	[15]	IXMAS	80.7
Base Normalization-Inception	[14]	UCF-101	94.6
3D-ConvNet with LSTM	[6]	KTH	94.39
Lattice Long Short-Term Memory	[7]	UCF-101	93.6
Temp Convolution with RNN, LSTM	[11]	Montalbano	94.49 (Precision)
JL_d Model	[17]	Berkeley MHAD	100
VA-LSTM	[16]	SBU Kinect	97.2
MRF-GT-Temporal	[2]	Volleyball	89.9 (Collective) 82.4 (Individual)
Action Attending Graphic Neural Network	[19]	Florence 3D	98.6
Graph CNN	[29]	MNIST	94.23
Pattern Recognition Neural Network (PRNN)	[3]	Weizmann	96

*AEE = Average Endpoint Error

6. Observation and Discussion

Human aggressive action recognition involves several sub-problems to be solved in order to come to a decision whether some aggressive action is contained within the frames or not. The critical reviews on the previous works presented in the paper look toward an all-in-one framework, which in most of the cases were not met. The limitations not only include image segmentation, feature extraction or motion modeling but also shed light into the limitations of convolutional neural network as a whole. The review summarizes the following observations:

- Spatial and temporal information extraction based on convolutional segmentation has a drawback in case any one of the information in one frame is faulty, which however is tried to address through a coupling network and modified LSTMs. This has a drawback regarding information loss in the convolutional layer due to pooling strategies.
- Convolutional neural networks work with one-dimensional feature map for each spatial and temporal feature extraction, which can lead to bottleneck due to fusion in the later layers.
- Background removal technique will result in an information gap between spatial and temporal feature if any low impact but significant information is missed which particularly happens in low resolution video streams.
- For combining a solution for all the aforementioned problem, a learning mechanism should be deployed that uses hierarchical feature extraction with embedded motion vector which is addressed by CapsNet[21]. This method embeds spatiotemporal features and generates a normalized vector output with squashing method. It is also very efficient in high-level feature extraction. Information loss is also addressed through the squashing method, which is basically normalization as unit vectors.

Based on the reviews, CapsNet can address several human action related problems with improvements on the top. These include:

- Overlapping feature segmentation and classification based on vector map matching which can be used for Face Recognition in a crowd with member estimation using differential map matching
- Human action for behavioral pattern recognition can be determined with the intensity and direction of the displacement of the vectors
- The action recognition problem can address aggressor-victim recognition as a segmentation solution
- Further, these can be integrated into a human emotion feature estimation problem based on weighting.

6. Conclusion

Violent behavior detection can help to ensure the safety of people and help law enforcement agencies to identify the perpetrator. This review task highlights the importance of the impact of detecting violence and aggressive behavior. In detecting violent scenes and to segment them, methods have to have robust spatiotemporal feature extraction and augmentation technique, focus on hierarchical relations and ability to discriminate inherent ambiguous human actions. Current methods which tackle this interesting problem were thoroughly analyzed to shed light on the weaknesses and strengths. From this review, it is apparent that there is a trend of using Convolutional Neural Networks as it provides discriminative power. But CNN is not without fault as it heavily relies on different pooling strategies resulting in information loss. CNN also lacks the ability to encode motion features inherently and have to depend on specialized motion feature mapping networks. Moreover, they do not perform well in scenarios with overlapping subjects. CapsNet though also lacking the

motion features, have the ability of hierarchical segmentation and input reconstruction as well as performs better on overlapping scenarios with a multitude of orientation variations.

After analyzing existing frameworks, this paper presents a modified and updated framework for segmentation and action recognition, which are expected to handle robustness indicated in observations. Judging from the previous research in computer vision field, it is certain that it will continue to be among the best research field in the future.

References

[1] Datta, A., Shah, M., & Lobo, N. D. V. (2002). Person-on-person violence detection in video data. In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 1, pp. 433-438). IEEE.

[2] Bagautdinov, T. M., Alahi, A., Fleuret, F., Fua, P., & Savarese, S. (2017, July). Social Scene Understanding: End-to-End Multi-Person Action Localization and Collective Activity Recognition. In CVPR (pp. 3425-3434).

[3] Xiao, Q., & Si, Y. (2017, December). Human action recognition using autoencoder. In Computer and Communications (ICCC), 2017 3rd IEEE International Conference on (pp. 1672-1675). IEEE.

[4] Zhu, A. Z., Yuan, L., Chaney, K., & Daniilidis, K. (2018). EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras. arXiv preprint arXiv:1802.06898.

[5] Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017, July). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on (pp. 1175-1183). IEEE.

[6] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011, November). Sequential deep learning for human action recognition. In International Workshop on Human Behavior Understanding (pp. 29-39). Springer, Berlin, Heidelberg.

[7] Sun, L., Jia, K., Chen, K., Yeung, D. Y., Shi, B. E., & Savarese, S. (2017, October). Lattice Long Short-Term Memory for Human Action Recognition. In ICCV (pp. 2166-2175).

[8] Chen, J., Wu, J., Konrad, J., & Ishwar, P. (2017, March). Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on (pp. 139-147). IEEE.

[9] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., ... & Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2758-2766).

[10] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017, July). Flownet 2.0: Evolution of optical flow estimation with deep networks. In IEEE conference on computer vision and pattern recognition (CVPR) (Vol. 2, p. 6).

[11] Pigou, L., Van Den Oord, A., Dieleman, S., Van Herreweghe, M., & Dambre, J. (2018). Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. International Journal of Computer Vision, 126(2-4), 430-439.

[12] Luvizon, D. C., Picard, D., & Tabia, H. (2018, June). 2d/3d pose estimation and action recognition using multitask deep learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 2).

[13] Oyedotun, O. K., & Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition. Neural Computing and Applications, 28(12), 3941-3951.

[14] Wang, Y., Long, M., Wang, J., & Philip, S. Y. (2017, July). Spatiotemporal Pyramid Network for Video Action Recognition. In CVPR (Vol. 6, p. 7).

[15] Rahmani, H., Mian, A., & Shah, M. (2018). Learning a deep model for human action recognition from novel viewpoints. IEEE transactions on pattern analysis and machine intelligence, 40(3), 667-681.

[16] Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2017). View adaptive recurrent neural networks for high performance human action recognition from skeleton data. arXiv, no. Mar.
[17] Zhang, S., Liu, X., & Xiao, J. (2017, March). On geometric features for skeleton-based action recognition using multilayer lstm networks. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 148-157). IEEE.
[18] Li, C., Sun, S., Min, X., Lin, W., Nie, B., & Zhang, X. (2017, July). End-to-end learning of deep convolutional neural network for 3D human action recognition. In 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 609-612). IEEE.
[19] Li, C., Cui, Z., Zheng, W., Xu, C., Ji, R., & Yang, J. (2018). Action-Attending Graphic Neural Network. IEEE Transactions on Image Processing, 27(7), 3657-3670.
[20] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4694-4702).
[21] Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In Advances in Neural Information Processing Systems (pp. 3856-3866).
[22] LaLonde, R., & Bagci, U. (2018). Capsules for Object Segmentation. arXiv preprint arXiv:1804.04241.
[23] Afshar, P., Mohammadi, A., & Plataniotis, K. N. (2018). Brain tumor type classification via capsule networks. arXiv preprint arXiv:1802.10200.
[24] Chen, M. Y., & Hauptmann, A. (2009). Mosift: Recognizing human actions in surveillance videos.
[25] Nievas, E. B., Suarez, O. D., Garc á, G. B., & Sukthankar, R. (2011, August). Violence detection in video using computer vision techniques. In International conference on Computer analysis of images and patterns (pp. 332-339). Springer, Berlin, Heidelberg.
[26] Weinzaepfel, P., Revaud, J., Harchaoui, Z., & Schmid, C. (2013). DeepFlow: Large displacement optical flow with deep matching. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1385-1392).
[27] Pang, J., Sun, W., Ren, J. S., Yang, C., & Yan, Q. (2017, October). Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching. In ICCV Workshops (Vol. 7, No. 8).
[28] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).
[29] Edwards, M., & Xie, X. (2016). Graph based convolutional neural network. arXiv preprint arXiv:1609.08965.

Authors' Profiles

A.F.M. Saifuddin Saif received PhD from Faculty of Information Science and Technology, University Kebangsaan Malaysia (UKM) in 2016. He received M.Sc. in Computer System Engineering (Software System) from University of East London, UK and B.Sc. (Eng.) degree in Computer Science and Engineering from Shahjalal University of Science and Technology, Bangladesh in 2012 and 2008, respectively. Most of his contributions in Computer Vision and Artificial Intelligence Research field were published in ISI Q1 journals. He has published many papers in ISI indexed Journals, Scopus indexed Journals, Book Chapters, Conferences and Proceedings. He served as Technical Committee Members, Reviewers, Guest Speakers, Session Chairs in many Conferences and Workshops. Currently, Dr. A.F.M. Saifuddin Saif is an Assistant Professor at **Faculty of Science and Technology**, American International University – Bangladesh. Before joining the university, he did Post Doctorate at Faculty of Information Science & Technology, University Kebangsaan Malaysia. He spent more than 6 years in IT industry such as Advanced Software Development, Web eMaze etc as IT researcher. His research interests include Image Processing, Computer Vision, Artificial Intelligence, Augmented Reality, 3D

reconstruction and Medical Image Processing.

Md. Akib Shahriar Khan is an undergraduate student enrolled at the Computer Science and Engineering program at the Faculty of Information Science and Technology of the American International University Bangladesh. Currently he is also a Research Assistant in the Department of Computer Science, American International University - Bangladesh. His research interests and passions are mostly based on Computer Vision and Pattern Recognition, Artificial Intelligence, Action Recognition and Motion Analysis, Neural Network and Machine Learning.

Abir Mohammad Hadi is an undergraduate student of Computer Science and Engineering (CSE) program at the Faculty of Science and Information Technology of American International University – Bangladesh (AIUB). He is currently also serving as Research Assistant in the Department of Computer Science. His area of research is machine intelligence, computer vision, image processing, autonomous system, neural network and artificial intelligence.

Rahul Proshad Karmoker is an undergraduate student of Computer Science and Software Engineering under the Department of Science and Information Technology of American International University Bangladesh. His research interests and passions are mostly based on Convolutional Artificial Neural Networks, Object Detection and Tracking included in the field of Computer Vision, Image Processing and Machine Learning.

Joy Julian Gomes is an undergraduate student of Computer Science and Engineering under the Department of Science and Information Technology of American International University Bangladesh. His research interests and passion are mostly based on Computer Vision and Pattern Recognition, Image processing, Artificial Neural Network, Colour Segmentation, Machine Learning, Object Tracking and Detection, Evolutionary Computation.

How to cite this paper: A. F. M. Saifuddin Saif, Md. Akib Shahriar Khan, Abir Mohammad Hadi, Rahul Prashad Karmoker, Joy Julian Gomes,"Aggressive Action Estimation: A Comprehensive Review on Neural Network Based Human Segmentation and Action Recognition", International Journal of Education and Management Engineering(IJEME), Vol.9, No.1, pp.9-19, 2019.DOI: 10.5815/ijeme.2019.01.02