International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 34 العدد
April 2024 ابريل
المجلة الدولية للعلوم والتقنية
ISTJ
تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

Human Action Detection Using A hybrid Architecture of CNN and Transformer

www.doi.org/10.62341/bsmh2119

Bassma .A. Awad Abdlrazg, Sumaia Masoud, Mnal .M. Ali

University of Omar Al-Mokhtar - Faculty of Science
Department of Mathematics
Somya.masoud2023@gmail.com

Abstract:
This work presents a Deep learning and Vision Transformer hybrid sequence model for the classification and identification of Human Motion Actions. The deep learning model works by extracting Spatial-temporal features from the features of every video, and then we use a CNN model that takes these inputs as spatial features map from videos and outputs them as a sequence of features. These sequences will be temporally fed into the Vision Transformer (ViT) which classifies the videos used into 7 different classes: Jump, Walk, Wave1, wave2, Bend, Jack, and powerful jump. The model was trained and tested on the Weismann dataset and the results showed that such a model was accurately capable of identifying the human actions.

Keywords: Deep Learning, Vision Transformer, Human Motion Action Detection, Spatial features, CNN.

International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 34 العدد
April 2024 ابريل
المجلة الدولية للعلوم والتقنية
ISTJ
تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

التعرف على الفعل البشري باستخدام البنية الهجينة لشبكة CNN والمحول

www.doi.org/10.62341/bsmh2119

أ. بسمه إبراهيم عوض، أ. سمية مسعود، أ. منال محمد علي

قسم الرياضيات /كلية العلوم / جامعة عمر المختار
Somya.masoud2023@gmail.com

الملخص:

يقدم هذا العمل نموذج تسلسل هجيني للتعلم العميق ومحول الرؤية لتصنيف وتحديد أفعال حركة الإنسان. يعمل نموذج التعلم العميق عن طريق استخراج الميزات المكانية والزمنية من ميزات كل فيديو، ثم نستخدم نموذجCNN الذي يأخذ هذه المدخلات كخريطة للميزات المكانية من مقاطع الفيديو و يخرجها كسلسلة من الميزات. سيتم تغذية هذه التسلسلات مؤقتا في محول الرؤية الذي يصنف مقاطع الفيديو إلى 7 فئات مختلفة وهى القفز و المشي وإشارة بيد واحدة و إشارة بيدين و إنحناء و قفز مع دوران و قفزة قويه. تم تدريب النموذج و اختباره على مجموعة بيانات وايزمان و أظهرت النتائج إن هذا النموذج قادر بدقة على تحديد الأفعال البشرية.

الكلمات المفتاحية : التعلم العميق، محولات الرؤية، كشف حركة الإنسان، الميزات المكانية، الشبكات العصبية التلافيفية.

Introduction

Human action recognition, the task of automatically recognizing and understanding human actions from video data, has garnered significant attention in recent years. With the increasing availability of video data and the remarkable progress in deep learning techniques, the field has witnessed a paradigm shift in the way action recognition is approached (Kuehue H et al, 2011). In particular, the integration of deep learning models and Transformers has emerged as a promising approach, offering improved accuracy and capturing long-range dependencies in action sequences.

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

Deep learning has revolutionized the field of computer vision by enabling the development of highly effective models capable of learning intricate representations directly from raw data (Jhuang H et al, 2013). Convolutional Neural Networks (CNNs) have played a pivotal role in action recognition, exploiting their ability to extract spatial and temporal features from video frames and sequences. However, traditional CNN architectures struggle to capture long-range dependencies and context information, which are crucial for accurately recognizing and understanding complex human actions (Zhang W et al,2013).

On the other hand, Transformers, originally introduced for natural language processing tasks, have demonstrated remarkable success in modeling sequential data by capturing global dependencies. Transformers rely on self-attention mechanisms, allowing them to focus on different parts of the input sequence and model complex interactions between elements. By adapting Transformers to action recognition, researchers have been able to effectively address the limitations of traditional CNN-based approaches and achieve state-of-the-art performance on various benchmark datasets (Lillo L, 2014).

This research aims to provide a comprehensive study of human action recognition using deep learning and Vision Transformers. We investigate the recent advancements in deep learning models and Transformers, and explore their fusion techniques to leverage the strengths of both approaches. By combining spatial and temporal information extraction from CNNs with the global context modeling capabilities of Transformers, we aim to enhance the accuracy and robustness of action recognition systems (Bilen H et al, 2016).

Furthermore, this research explores various architectures and configurations of deep learning and Transformer models specifically tailored for human action recognition. We delve into different strategies for incorporating temporal information, such as 3D convolutions and optical flow, and examine the impact of pre-training on large-scale video datasets. Additionally, we investigate the effectiveness of different attention mechanisms and architectural

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

variations in CNN and Transformers to optimize their performance in the context of action recognition.

This paper has been divided into five main sections: Section 1 provides a general background on Human action detection and convolution neural networks. Section 2 discusses the method proposed in this work: the hybrid combination of CNN and ViT, and the dataset used in this work. Section 3 provides an overview of the results achieved in this study. Then, in section 4, a comparison of the obtained results versus the related works results is introduced. Finally, in section 5, conclusions are presented.

Related Work

Face detection and counting people have been active research areas in the field of computer vision, and several studies have been conducted in this domain. Below is a summary of some of the important works related to face detection and counting people. Face detection is the process of locating and identifying human faces in digital images or videos. One of the most popular face detection algorithms is the Viola-Jones algorithm (Viola P & Jones M, 2001). It uses a cascade of classifiers based on Haar-like features and AdaBoost to detect faces in images. The algorithm has been widely used in various applications, including security, surveillance, and human-computer interaction. While traditional methods such as Viola-Jones algorithms have been successful in detecting faces under controlled conditions, they may struggle with complex scenarios such as occlusion and varying lighting conditions. Recent advancements in deep learning have led to improvements in face detection performance, particularly with the development of convolution neural networks (CNNs). A notable method is the Single Shot MultiBox Detector SSD (Liu Y et al, 2016). The SSD is a convolution neural network (CNN) that can detect faces in 3 real-time with high accuracy. Other deep learning-based methods, such as Faster RCNN and YOLO, have also been proposed for face detection (Yao L et al,2016 & Chou K P et al,2018). The RetinaFace algorithm uses a single-stage CNN model to detect faces while achieving significant accuracy (Deng J et al, 2019). Similarly, the

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

CenterFace algorithm uses a lightweight backbone network to detect faces in images and video streams (Xu L et al, 2020). Counting people is another challenging task in computer vision, especially in crowded or complex scenes. One approach is to use object detection algorithms, such as the Viola-Jones algorithm, to detect faces in images and then count them. Ittahir et al. proposed a system that presents a basic yet practical approach for people counting in still images using skin color face detection. The proposed method provides reasonable accuracy given its simplicity, demonstrating its potential as an initial solution for applications where approximate people counts are sufficient (Bilen H et al, 2016). Recently, deep learning-based methods have shown superior performance in crowd density estimation, where CNNs are used to estimate the density of people from input images. One notable study in this area is the work by (Chen et al, 2021) where they proposed a method for counting people in crowded scenes using a combination of scaleinvariant feature transform (SIFT) and support vector regression (SVR) **.** The SIFT algorithm is used to extract local features from the image, and the SVR is used to estimate the number of people based on the extracted features. The method was tested on several datasets and showed promising results. Based on deep learning (Zhang R et al, 2018) they suggested a method for counting people. They used a convolution neural network (CNN) to extract features from the image and then used a fully connected layer to estimate the number of people. It was also suggested a method for crowd counting using a deep neural network that is trained to estimate the density of people in an image (Zhang R et al, 2018).

The method uses a multi-column CNN architecture that can handle different scales of people in the image. The method was tested on several datasets and showed superior performance compared to state-of-the-art methods. A recent study by Chen et al. proposed a method for counting people in videos using a spatiotemporal attention mechanism. The method uses CNN to extract spatial features and a recurrent neural network (RNN) to capture temporal dependencies in the video sequence (Yau G et al, 2019). The spatiotemporal attention mechanism is used to focus on the most

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

relevant regions of the image and video frames, leading to improved accuracy. The main difference between using Viola-Jones face detection methods and deep learning is the type of algorithms used. Viola-Jones face detection methods use supervised learning algorithms such as CART and LBP, while deep learning uses unsupervised learning algorithms such as convolution neural networks. Viola-Jones face detection methods are more accurate for face detection and classification, while deep learning is more accurate for object recognition and image segmentation (Zhou H et al, 2009) . In conclusion, face detection and counting people are important research areas in computer vision with several applications. While deep learning-based methods have shown promising results in recent years, the Viola-Jones algorithms are popular face detection algorithms with high accuracy. Therefore, for this study, Viola-Jones algorithms are used for face detection and people counting in still images.

METHODS

CNN-ViT Hybrid Model

When training video classifiers, one of the challenges is determining how to input videos into a network. Since videos consist of a sequence of frames a method could be to extract these frames and organize them into a tensor. However varying frame counts, across videos may make it difficult to batch them together without using padding. An alternative approach involves saving video frames at intervals until reaching a maximum frame count. In this scenario we will follow these steps;

1. Capture the frames from a video.
2. Extract frames from the videos until reaching the frame count.
3. If a video has frames, than the maximum allowed we will pad it with zeros as needed.

The frames are then fed into a pre-trained CNN which will extract feature maps from every video and output those spatial features as sequences (Figure 1).

International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 34 العدد
April 2024 ابريل
المجلة الدولية للعلوم والتقنية
ISTJ
تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

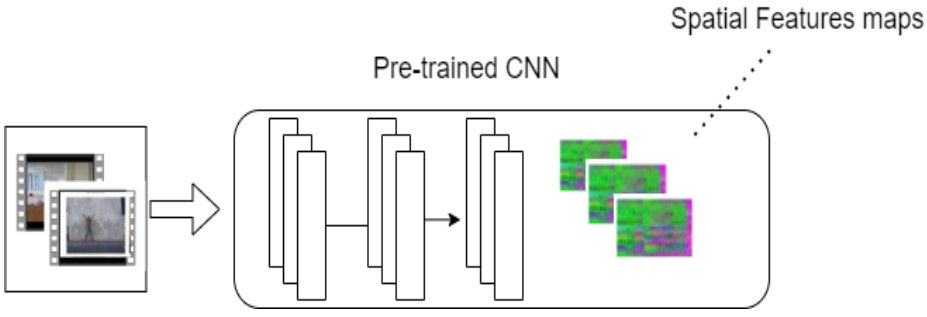

Figure1: Capturing the spatial features using a pre-trained CNN.

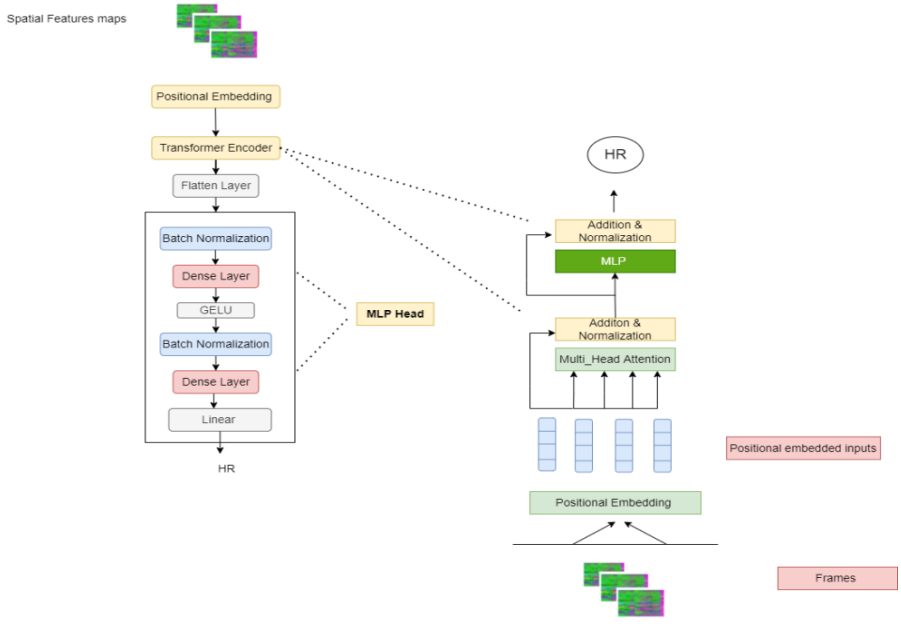

Figure2: The CNN-Transformer hybrid model architecture

The sequences of frames will be then fed into the Transformer as seen in figure 2. FOne of the fundamental challenges in adapting Transformer models for video processing lies in their inherent order-

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

agnostic nature. Unlike language, where the order of words is crucial for understanding, traditional Transformer architectures do not explicitly capture the temporal structure of video sequences, which are inherently ordered by definition. This necessitates the introduction of a mechanism that allows the model to infer and utilize the relative and absolute positions of frames within a video.

This work addresses this challenge by incorporating positional encoding as a crucial step within the video processing pipeline. This technique leverages an Embedding layer to map the position of each frame within the video sequence to a corresponding positional embedding, a low-dimensional vector representation that encapsulates the positional information. These positional embeddings are then added element-wise to the precomputed Convolution Neural Network (CNN) feature maps extracted from each individual frame. This simple yet effective strategy injects the necessary positional information into the model's internal representation, allowing it to exploit the temporal ordering of frames and perform tasks that rely on this structure, such as video classification or action recognition.

By explicitly incorporating positional information through encoding, this work empowers Transformer models to effectively process and understand the inherent temporal dynamics within video sequences, paving the way for their wider application in various video-related tasks.

Dataset

The Weismann data set which is illustrated in figure 3 comprises 90 low-resolution (180x144, Deinterlace50 fps) video sequences. These sequences depict nine individuals performing ten distinct natural actions, including running, walking, skipping, jumping jacks, various jumping variations, a sideways gallop, two-handed waving, one-handed waving, and bending.

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

Figure 3: The Weizmann Dataset

Results

The CNN-ViT model was trained and tested on 70 videos of 7 different types: ‘bend', 'jack', 'jump', 'p jump', 'walk', 'wave1', 'wave2'. The aim of this work is to train deep model to identify the type of action shown in a video by extracting some of its frames.

It was important to partition the dataset into training and testing phases for the models. All divisions were carried out with the purpose of splitting the videos’ classes/categories in an even way, as much as possible. Thus, avoiding data leakage and imbalance between the training and test sets. For the evaluation of the model, the following metrics were used: accuracy, precision, and F1-score. This training phase was repeated for each hyper parameter combination created in the subsequent optimization phase. The goal of hyper parameter optimization was to improve model efficiency and reduce classification errors. The dataset was split into 70% for training and 30% for testing. Note that the Tensor Flow 2.5 framework was used to create the training and testing pipelines for the base ViT model. This was achieved in a Python 3.7 virtual environment. The Adam optimization method has been proved to perform better than its competitors among the ones that are currently accessible. As a result, the Adam optimization technique was used

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

to train the model, with a gradient decay value of 0.9. The regularization factor was set to 0.0001, while the initial learning rate was set to 0.001. Due to memory constraints, the model was ultimately trained for 150 epochs with a mini batch size of 64.

Figure 4 and 5 displays the training accuracy and loss of the best model performance. It can be seen that the model's lowest error occurred at epoch 100, at which point learning ceased due to the implementation of the Early Stopping strategy during training in order to prevent over fitting.

The model was tested on 30% of the data as previously mentioned, and to show the feasibility of the STP and LSA, we also trained and tested a regular CNN –ViT. Table 1 shows the testing results of the CNN -ViT. It is noticed that adding CNN improved the performance of the ViT in recognizing human actions.

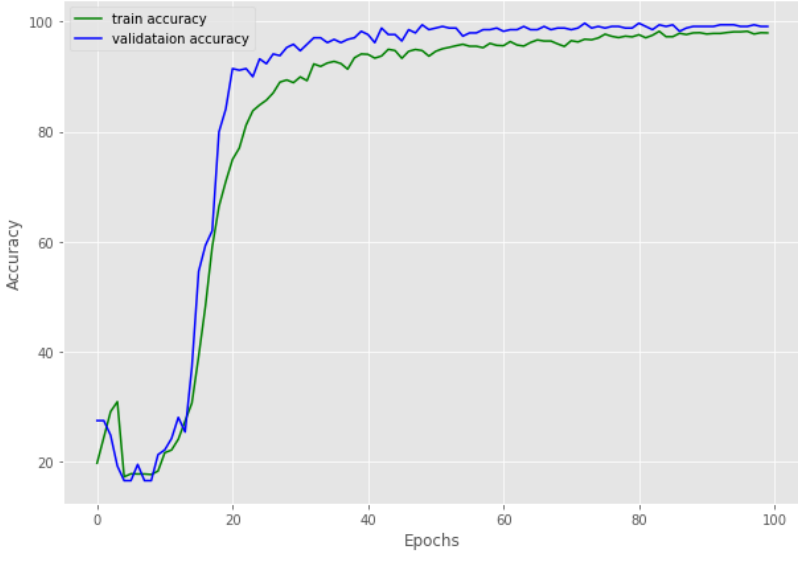

Figure 4: The accuracy versus number of epochs variation of the model during training

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

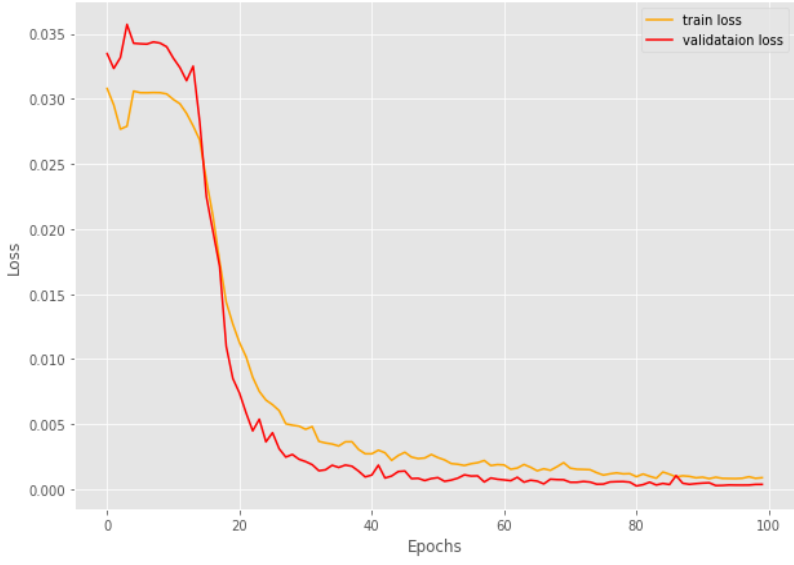

Figure 5: The loss versus number of epochs variation of the model during training

Table 1. Results evaluation

	Train	Test
Nb. of images	3000	1500
Accuracy	100%	97.1%

Figure 4 and 5 show the training learning curve and the loss variation of the model during training.
It is seen that the model loss was gradually decreasing with respect to the increase of number of epochs, which eventually grated significant generalization capability during testing.

Discussion

The performance analysis of the proposed approach on the Wiezmann dataset is presented in Table 2. The proposed method significantly improves the classification results for the action verbs in the Wiezmann dataset. Due to the similarities between certain

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

actions, such as boxing and hand clapping, there is a possibility of misclassification. However, the proposed method effectively classifies actions like jogging, walking, running, and hand waving in the Wiezamann dataset with an impressive accuracy rate of 91%. Moreover, the proposed method successfully recognizes and accurately classifies similar actions, such as boxing, jogging, walking, and running, which share significant resemblances in shape and motion.

Table 2. Results comparison

	Method	**Accuracy**
Zhang et al. []	SIFT	96.6%
Yao et al. []	Fusion + Pool	-
Xhou et al. []	Nearest neighbour + Gaussian Mixture Model, Nearest Mean Classifier	95%
Ours	CNN + ViT	97.1%

Conclusion

Overall, this research explores various architectures and configurations of deep learning and Transformer models specifically tailored for human action recognition. We delve into different strategies for incorporating temporal information, such as 3D convolutions and optical flow, and examine the impact of pre-training on large-scale video datasets. Additionally, we investigate the effectiveness of different attention mechanisms and architectural variations in Transformers to optimize their performance in the context of action recognition.

Ultimately, the findings of this research will contribute to the growing body of knowledge in the field of human action recognition. By elucidating the benefits and limitations of deep learning and Transformers, we aim to facilitate the development of

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

more accurate and robust action recognition systems, enabling applications in areas such as video surveillance, human-computer interaction, and healthcare monitoring.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., and Gould, S. (2016). "Dynamic image networks for action recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3034-3042.

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. (2021). "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities", ACM Computing Surveys (CSUR), 54(4), 1-40.

Chou, K.P. et al. (2018). "Robust Feature-Based Automated Multi-View Human Action Recognition System ", in IEEE Access, vol.6, pp.15283-15296. https://doi.org/10.1109/ACCESS . 2018.2809552.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., and Zafeiriou, S. (2019). "Retinaface: Single-stage dense face localisation in the wild", arXiv preprint arXiv:1905.00641.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., and Black, M.J. (2013). "Towards understanding action recognition," in Proceedings of the IEEE International Conference on Computer Vision, pp. 3192-3199.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). "HMDB: a large video database for human motion recognition ", International Conference on Computer Vision, IEEE, pp. 2556–2563.

Lillo, I., Soto, A., and Niebles, J.C. (2014). "Discriminative hierarchical modeling of spatio-temporally composable

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

human activities", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 812-819.

Viola, P., and Jones, M.(2001). "Rapid object detection using a boosted cascade of sample features", Accepted conference on computer vision and patter.

Wiezmann Dataset. From: Actions as Space-Time Shapes (weizmann.ac.il)

Xu, Y., Yan, W., Yang, G., Luo, J., Li, T., and He, J. (2020). "CenterFace: joint face detection and alignment using face as point". https: ://doi.org/10.1155/2020/7845384.

Yao, G., Lei, T., and Zhong, J. (2019). "A review of convolutional - neural-network-based action recognition", Pattern Recognition Letters,118, 14–22.

Yao, L., Liu, Y. & Huang, S. (2016). "Spatio-temporal information for human action recognition", J Image Video Proc. 2016, 39. https://doi.org/10.1186/s13640-016-0145-2.

Zhang, C., Li, H.,Wang, X., and Yang X. (2015). "Cross-scene crowd counting via deep convolutional neural networks", In: Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Zhang, C., Tian, Y., Guo, X., & Liu, J. (2018). "Daal: Deep activationbased attribute learning for action recognition in depth videos", Computer Vision and Image Understanding, 167, 37–49. https://doi.org/10.1016/j.cviu.2017.11.008

Zhang, W., Zhu, M., and Derpanis, K.G. (2013). "From actemes to action: A strongly-supervised representation for detailed action understanding," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2248-2255.

Zhou, H., Wang, L., and Suter, D. (2009). "Human action recognition by feature-reduced Gaussian process

تم استلام الورقة بتاريخ: 19/ 2/2024م وتم نشرها على الموقع بتاريخ: 21/ 3/2024م

classification", Pattern Recognition Letters, 30(12), 1059-1066.