IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of current version 21/04/2022.

A Review of Methods for Human Action Detection Using Deep Learning with Applications

Michael Gillespie[1], Abhishek Gowda K S[2], Nilesh Ohol[3], Yifan Huang[4],
[1,2,3,4]Department of Electrical and Electronic Engineering, University of Strathclyde, Glasgow, G1 1XQ, UK

Corresponding author: Michael Gillespie (e-mail: Michael.Gillespie.2021@ uni.strath.ac.uk).

This work was carried out as part of MSc course "EE986: Assignment and Professional Studies".

ABSTRACT In recent times the need for systems that can automatically recognize human actions has become more prevalent. Driven by the ever-increasing presence of autonomous vehicles and systems, the demand for smarter surveillance and video analysis, and the prospect of automatic medical diagnoses, to name a few, the interest in human action detection has skyrocketed. Consequently, over the last decade, the application of deep learning to perform human action recognition has also grown. This paper focuses on three human action detection areas of interest, namely, human object interaction detection, image and video captioning and human movement detection. For each area key methodologies used to perform the detection are detailed along with qualitative comparisons of strengths and weaknesses. This paper then concludes with several case studies which detail real world applications of these methodologies in industry.

INDEX TERMS Action detection, artificial intelligence, computer vision, deep learning, human object interaction, image and video captioning, movement detection.

I. INTRODUCTION

As with many computer vision tasks the application of deep learning to human action detection has exploded in recent times, driven by increased computational capacity and the availability of larger and more diverse datasets the capability of systems to recognize and categorize human actions has reached new levels of complexity. Although recent research focuses on deep learning methods human action detection predates the explosion of deep learning with a variety of traditional machine learning methods laying the foundations for what would become a thriving area of research. These traditional models utilized hand crafted feature extraction methods then using feature information such as orientation and spatial relations between features classified actions and interactions using traditional machine learning models [1,2,3,4]. These hand-crafted methods are however limited in the fact that they use relatively few numbers of features to perform the action detection and so tend to only perform well in more specific applications. Deep learning methods, namely convolutional neural networks (CNNs) on the other hand can automatically identify and extract huge amounts feature information from the images and videos provided. This allows them to utilize a much larger amount of visual scene information when attempting to classify human actions. This has led to better overall performance when as well as the capability to classify actions with a higher degree of accuracy on datasets featuring a much larger array of objects, scenes, and actions [5].

Three human action detection tasks are explored in this paper with methods for each being discussed and compared to provide an overview of the current state of the art.

The first area of interest explored is human object interaction (HOI) detection. HOI detection is a branch of human action detection that attempts to create models that are capable of localizing both humans and objects present in an image or video and then, infer whether the two are interacting and if so, what action the human is performing with or on the object. The output from these models is generally an object, human, action triplet i.e., the location and class of the object and human then the verb that describes their interaction.

The second task discussed is image and video captioning. The intention of models created here are to generate captions of natural language which describe the contents of an image or video. Although this task is a more general area of research rather than purely a human action detection process many datasets available exclusively feature humans performing actions or have a large quantity of examples. The output from these models is a short natural language caption that describes the action being taken by the human.

The final area of research presented in the paper is human body movement detection. The work presented in this

section focuses on tracking the movement of the human body in videos or from live streams in order to classify or quantify movements according to the use case.

II. HUMAN OBJECT INTERACTION DETECTION

Traditional methods to perform HOI detection focused on using specifically crafted feature extraction models such as Scale Invariant Feature Extraction (SIFT) and similarly Speeded Up Robust Features (SURF) to identify objects and specific human poses from images. These were then coupled with traditional machine learning models to perform action identification in images and videos [9, 10, 1]. These methods provided a good level of performance but like many computer vision applications, in recent years performance has been dwarfed by the advent of deep learning.

This increase in performance can be attributed to the fact these deep networks are able to extract a wider array of features from the large amount of data they encounter and utilize more information about the scenes than can be achieved using traditional approaches where features in the images are identified using human created models of objects and poses [11].

Much of the work in recent years on modeling HOI can find its origins in the work by Gupta and Malik [12]. In this piece of research, they postulated that the state of the art at the time was too coarse in its approach to performing action detection either classifying a whole image/video based on its content or classifying the image/video then applying methods to add a bounding box around the proposed areas the actions are occurring. As a result, the classification does not form a complete understanding of the relationships between objects and the humans present in the images. They argue that to learn the relationships between human and objects the model first must identify and localize the object and the human within the scene before attempting to identify the action occurring. This approach to the problem has since become the norm with most modern methods initially performing human and object identification and localization before attempting to infer the relationship.

FIGURE 1. **Image with Human and Object Localization Applied for HOI Detection [12]**

Modern approaches to performing HOI detection are defined here in two main categories: single stream or multiple stream methods. Single stream models utilize one single model trained end to end, or several models stacked in series with the output of one flowing into the input of the next [12,13,14,15]. Multiple stream approaches have models which utilize parallel streams where a greater number of features about the scene are analyzed in parallel and the outcome of these different feature analysis streams are then used in combination to perform HOI detection [6,28,29,30,34,35,36,38,39].

A. DATASETS

As with many computer vision tasks as the interest in the topic area has grown so too have the number of complete datasets available which can be used to benchmark different methodologies. These datasets were either created with the intent of being used to perform HOI detection [16, 17, 18] or are modifications of existing datasets that are used for similar tasks such as image and video captioning that has been modified to be applicable to HOI detection [5,6]. Generally, the modified sets such as V-COCO and HICO-DET cover a wider array of actions and are a better option when evaluating the performance as they are modifications of extensive sets which were previously used for image classification or object detection. Many of the datasets created specifically for HOI are much more truncated in the number of objects/actions featured and are often very domain specific.

B. SINGLE STREAM METHODS

The models detailed in this section all utilize similar methodologies which start by first identifying regions of interest within an image before, attempting to identify humans and objects within these areas and finally categorizing the consequent actions. When the methodologies are discussed generally, they are done so from the point of view that they are applied to images. Although there are instances where the methods are applied to videos unless otherwise stated it can be assumed that this is done so on a frame-by-frame basis with no temporal tracking so is considered equivalent.

The R-CNN first introduced by Girshick et al [19] was a novel way to perform accurate object detection and classification in images by utilizing a combination of existing computer vision methods to perform object localization and classification. It has since become a popular way to perform object identification and localization in the HOI space. The two-stage method first performs object detection using the exhaustive selective search method [20] to extract regions of interest likely to contain objects. Then it performs object classification by utilizing a CNN [21] to perform feature extraction. The new embedded features are then passed into a support vector machine to perform classification. The original R-CNN architecture suffered from slow processing times so since its inception improvements have been made with the creation of fast R-CNN, Faster R-CNN, and Mask R-CNN [22]. These new methods improve on the R-CNN by utilizing novel methods such as a CNN to perform region proposals in place of selective search, improving accuracy and computational time

or utilizing transfer learning of image classification CNNs to perform the object classification stage such as VGG-16 CNN architecture trained on the imagenet dataset [23, 24].

First introduced by Gupta and Malik as a baseline model architecture [12], the use of R-CNNs became a very widely accepted approach for performing HOI detection. These methods are relatively simple to implement and do not require a huge amount of training data due to the use of transfer learning of pretrained architectures.

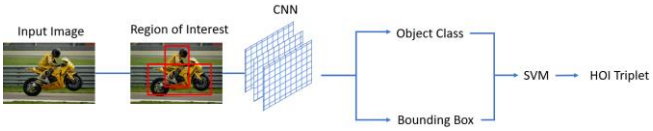

FIGURE 2. **Visualization of Single Stream System Detailed in [12]**

In order to perform HOI detection, the object classification and bounding box output from the R-CNN is usually passed to another machine learning model which has been trained on the dataset in order to perform the final HOI classification. Initially, the use of SVMs proved effective and had the added benefit of creating a simple single stream model [12]. Dogariu et al [13] took an even simpler approach to perform HOI detection utilizing a similar region-based approach in order to identify abandoned luggage in surveillance videos. The model utilizes output from an R-CNN and classifies luggage as abandoned if there is no bounding box overlap between the baggage and human object. This model is simple and reasonably human interpretable. However, the model lacks any understanding of how a person would interact with a piece of baggage and although the reported performance is good it is suspected that in practice it would perform poorly for instance a busy train station could have lots of bounding boxes for humans overlapping with a bag which has been sat unattended for some time.

Utilizing a more complex model in the action classification stage seems like a logical way to improve applicability and extract more detail about the relationships between bounding box location, object classification, and the resultant action. Zhang et al [14] did just this, they extended the architectures seen previously utilizing an R-CNN with VGG-16 backbone to identify and extract human and objects in a scene before passing this human object information to a second "Relation Prediction Module". This module takes the output from the final convolutional layer of a fast R-CNN and then passes that as input into a second CNN which extracts more detailed features of the objects or humans. The output from this CNN is then passed as input, concatenated with class and bounding box location from the R-CNN into a softmax layer which classifies the action between the objects. This more complex approach has the added benefit of being trainable from end to end. It also greatly increases the amount of detail about the scene and the interactions that are extracted by the models so is more likely to perform better on more complex datasets than the methods shown previously.

Although most single stream methods for performing HOI detection are based around some variation of the region-based R-CNN to perform object and human detection there has been considerable research undertaken in recent times in the area of object detection whereby new alternative methods are available. The most significant is the creation of the You Only Look Once (YOLO) algorithm [25]. The YOLO, and subsequent iterations thereof, differ from R-CNNs as the model performs both the bounding box regression and object classification in one feedforward model. Although this algorithm has been applied to human action detection in general [26, 27] very little research is available detailing its application to HOI detection. Zapf et al [15] utilized the YOLO algorithm to perform HOI detection in real time for use in autonomous robotics applications. This paper is a good example of the possible benefits of utilizing YOLO over R-CNN as it is much more suited to real-time applications due to its very fast processing speeds.

The single stream models referenced in this section include some of the first instances of performing HOI detection using deep learning and introduced a variety of methods and approaches to the task that have translated into the state of the art today. Their performance is however partly limited by the fact they only draw on a small amount of information about the scene, generally only considering bounding box locations and object categories and do not consider things like human pose or divide the human body into subsections for different body parts to gain a more nuanced interpretation of the scene.

C. MULTIPLE STREAM METHODS

1) Overall Relative Distance Methods

Several models have been designed in recent years to improve the performance of single stream models by utilizing the relative distance between humans and objects as a measure of their interactivity or as an additional input to inform on what action is being performed.

The work by Chao et al [6] introduces the idea of utilizing the relative distance between the objects and human bounding boxes in the frame to identify firstly, if the object and the person are interacting and secondly to provide extra information to the system about the interaction occurring. The model works by extracting the output from an R-CNN and then calculating the relative distance between the bounding boxes of the human and the object. The object bounding box image is then passed to its own deep neural network, the same is done for the human bounding box image and then finally the spatial relation between the two is passed to a third parallel network. The output of these three networks is then summed in order to provide a classification score which represents the likely hood of the given class. This methodology although novel in its creation is heavily limited by the fact the network is only designed to be able to recognize one action for example someone riding a bike. For

multiple HOI classes multiple binary classifiers would need to be trained in a one vs rest or one vs all architecture. This severely limits the applications in practice. Gao et al [28] utilize a very similar model, however they are able to extend functionality to make the model applicable in multiclass classifications by using similar softmax classifiers at the output of each network for all possible actions. The scores from the object image network and human image network are then added together before being multiplied by the output from the spatial configuration network improving on the work in [6].

The paper by Qi et al [29] offers an alternative approach on this whereby a graph is generated for a scene where each node represents the center of the bounding box for a given human or object and then each edge represents the spatial vector between every human and object present in the scene. This graph is then passed to a neural network which infers what the most likely human object pair is and the most likely interaction. This model is simpler in its design but loses scene information which is utilized in the three stream models [6, 28].

Prest et al [30] extended the idea of using relative distance between objects and humans to assist in action detection to videos, applying the method for object detection and relative spatial regression detailed in [31] and applying it to subsequent frames in a video. This then provides a spatial vector that changes with time which can be used to classify the actions.

2) Human Part Based Methods

Human body part detection has been around for some time [32,33] allowing models to identify and label sections of a person relative to one another. This can be useful in HOI detection as it allows a model to infer with greater detail the 'interactiveness' of objects and people and what action is occurring. For instance, if a hat is detected next to a head the person is probably wearing a hat whereas if it is detected next to a hand, they are probably carrying it.

In recent years some models have taken advantage of this and implemented body part attention in their detection methodology for HOI classification. Wan et al [34] utilized a multi-level approach in their model for HOI detection whereby the object and human bounding boxes as a whole were both considered by the model as well as the positions of human body parts that are likely to be interacting with the object. After human, object and body part detection has been performed the detected object and humans are organized into human object pairs which are each then processed separately. The bounding box locations for the human, detected body parts and object are then passed into two parallel neural networks labelled the holistic module and the zoom-in module. The output from these networks is then concatenated and passed into a fully connected network for action classification. The output from the Holistic network is also passed into a separate fully connected network to determine if the pairs are interacting or not. The overall output from this network then gives a series of human object interactions that are present in the scene. This method was able to achieve state of the art performance by having the ability to recognize subtle differences in HOIs that would previously go unnoticed by earlier networks.

Fang et al [35] created a very similar architecture with the main difference being body part detection uses pairwise correlations between body parts in the model so extracts pairs of body parts that appear to be involved together in the action and use that information to perform HOI detection. This means the model can more easily ignore irrelevant body parts as well as gain an understanding of how body parts would relate to one another when performing certain actions.

These models improve on the research utilizing only relative distsance between bounding boxes and achieved state of the art results in comparison as they are able to identify much more subtle differences in the relationships between humans and object and a finer scale.

3) Skeletal Pose Methods

While the success of human part-based methods is clear similar work has been done in recent years which utilizes the skeletal pose of the humans in the image to improve HOI detection.

Gupta et al [36] take what they describe as a 'no frills' approach to performing HOI detection. In their research, they use transfer learning and a number of pre-trained models to perform each of their feature extraction steps. An R-CNN implementation is used for object and human bounding box regression and classification and then the OpenPose CNN [37] is used to generate a 2D skeletal representation of the human pose. This information is then fed into a multilayer perceptron with the output being the action classification. This model excels in its simplicity allowing for fast implementation by utilizing existing knowledge in the area to a much higher degree than other works. This work appears to be very successful however, it is unclear how they would cope with real world issues such as occlusion for instance if an individual's lower half is blocked it may cause issues in the classification process.

Li et al [38] Explore the use of skeletal pose to evaluate the 'interactiveness' of an object and a human i.e whether the two are indeed interacting. Similar to the works in [6,28,30] Three parallel networks are used to classify the HOI based on the human features, the object features, and the spatial relation between them. It differs however in the fact that there is a parallel network that utilizes the skeletal pose of the detected human to evaluate whether the object and human pair are indeed interacting. This is an effective method for reducing instances where interaction is labeled which may not necessarily be occurring due to the model falsely labeling a human and object as interacting. The most apparent downfall of this approach seems to be the fact that

the interactiveness is processed in parallel to the classification. Most models perform some form of pairwise removal of human object pairs that don't appear to be interacting before the HOI detection is performed. By performing the action in parallel the HOI classification is performed regardless of whether they are indeed interacting meaning additional computation is performed for instances where it is redundant.

Finally, the work by Yan et al [39] takes the analysis of skeletal pose a step further by including the movement of hands. This architecture consists of three parallel models to perform HOI detection. Firstly a YOLO algorithm implementation is used to perform human and object bounding and classification. Secondly, the human skeletal movement was captured using Microsoft Kinect, gaming hardware and software product, and then these images were passed into a CNN to extract features about body movement. Finally, a set of custom-made digital gloves was used to track hand movement, this was then fed into a third CNN. The outputs from these three networks were then fed into a fully connected feed-forward network for performing classification. This paper offers an interesting approach to human action detection as well as detailing interesting new hardware options. However, when examining the dataset used there is very little intra-class variation and large inter-class variations. It appears not enough emphasis was placed on challenging the system. One option would be to create a set similar to the Rochester Daily Activities Dataset [40] as it was designed so that many of the actions appear similar and are harder to distinguish between.

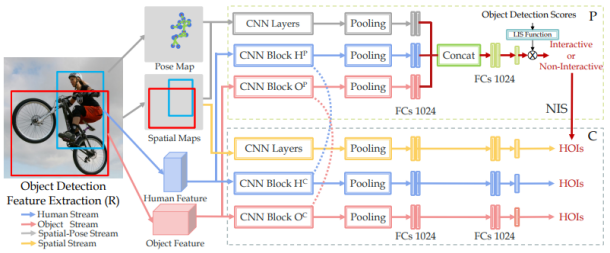

FIGURE 3. **Multiple Stream Example from [38]**

III. IMAGE AND VIDEO CAPTIONING

A. DATASETS

Similarly, to most computer vision challenges, as interest in the field has grown, so has the number of complete datasets available for comparing different techniques. The most commonly used datasets for evaluating the performance of image captioning models are Flickr 8k, Flickr 30k, MS COCO, and PASCAL 2008 images which can be used to evaluate the state-of-the-art models and image annotations in the PASCAL dataset.

B. IMAGE CAPTIONING

1) Template-based

Fixed templates with several empty slots are used to produce captions in template-based techniques. Different objects, attributes, and actions are first identified, and then the empty spots in the most relevant templates are filled in the models described in this section.

The image captioning model introduced in [42] by Farhadi uses a triplet of scene elements to fill the template slots for generating image captions. To assess the similarity between phrases and images, the model employs a scoring mechanism where the image and each sample phrase are mapped in an intermediate space of meaning that consists of different projections i.e., keywords from the space of images and keywords from the sentences that describe images, a score is given to each sentence based on similarity of sentence to the image, sentences with similar meaning then get a high score. The space of meanings is represented by object, scene, and action triplets and each slot in the triplet can take up a value from a set of distinct values. Solving a multi-label Markov random field is required to predict a triplet from a picture. There is an object node that can accept a value from a collection of expected nouns, an action node that has a set of distinct anticipated verbs, and a scene node that can pick each of the predicted nouns. The edges represent the binary connections that exist between the nodes. The model can score a match between an image and a sentence once it predicts triplets for pictures and sentences using the greedy strategy. It generates sentences by looking through a list of possible phrases for one that has a high match score to the image and then substituting the appropriate words that describe the nodes. Although this work is a relatively simple and effective approach to perform image captioning the capability of the system is extremely limited by the fact is can only select from a predetermined pool of captions.

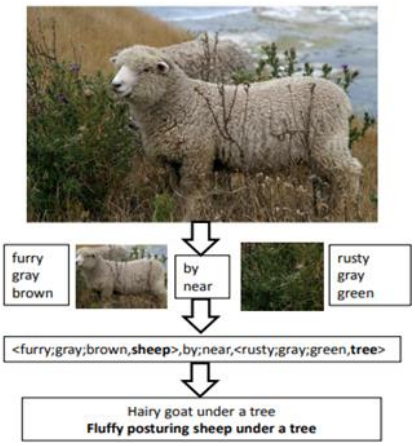

FIGURE 4. **Template Based Captioning [41]**

In order to improve on the work illustrated in [42] Li et al. [41] introduces the idea of extracting relevant phrases and then fusing these together to create a more complex caption. This is performed using a unique surface realization technique based on web-scale n-gram data. The sentences generated by this model seem like lines from a poem. This method consists of two steps: phrase selection (n-grams) and phrase fusion (n-grams). The first phase, phrase selection, gathers candidate phrases that might be used to

generate a description for a certain image. To increase fluency, this stage naturally accommodates ambiguity in image recognition inputs, as well as synonyms and word re-ordering. The image recognition system extracts visual information and encodes it as a set of triples introduced by Farhadi [42]. The second stage, phrase fusion, uses dynamic programming to discover the most suitable group of phrases and builds a new (and more complicated) phrase that characterizes the image allowing for improved caption complexity and robustness to unseen images when compared with [42].

The suggested method by Girish Kulkarni [43] is divided into two steps. The occasionally noisy output of computer vision recognition algorithms is smoothed with statistics acquired from visually descriptive natural language in the first stage, content planning. After deciding on the content for creation, the following step is surface realization, which entails choosing words to express the material. Text statistics are employed once again to select a surface realization that is more comparable to frequent language constructs.

Although template-based methods shown here are relatively simple and effective ways to produce image captions there are limitations to this approach. It does not allow creative writing as templates and phrases are pre-defined meaning it cannot add interesting new words or replace existing content words with better ones which might be necessary to create an articulate sentence and generate captions of variable length.

2) Retrieval based

With a basic retrieval (pre-retrieval) model, retrieval-based techniques explain pictures by extracting a candidate caption set from a pre-constructed image-caption repository. The re-ranking approach selects the final best-matching captions for the input picture from the captions pool. It can create relevant and grammatically acceptable captions, but it has trouble coming up with creative and diverse captions that accurately depict the new photos.

The model introduced by Ordonez [44] takes a query image as input and matches scene structure and overall color of images using two global image descriptors, firstly an image descriptor related to perceptual dimensions and secondly an image descriptor computed by resizing an image into a smaller image. It then finds visually relevant images similar to the input image in a large novel data set, which consists of images from the web with associated captions written by people and filtered such that they are visually relevant. After matching a set of images based on similarity, the system re-ranks the selected images to select the caption of the best image by computing the similarity of image contents such as attributes of objects, people with action, and background scene in the query image and matched images. In another approach, captions can also be composed by using only global image descriptors.

The preceding model provided by Ordonez [44] created captions for images automatically, but the model developed by Hodosh [45] ranks a list of unseen captions for each test image seen. The author presents the PASCAL VOC-2008 and the Flickr 8K dataset, both of which were generated for sentence-based image descriptions. Each image in the collection comprises five captions defining the image. The ranking-based technique returns a ranked list of captions, allowing for large-scale quantitative evaluations and direct comparisons of other methods. Using Kernel Canonical Correlation Analysis, to learn projections so that images and captions connected with it are optimally linked, images and captions related to that image are projected into a common space. When nouns and verbs refer to the same object, the text kernel in the model uses lexical-based similarity to capture words/phrases describing the same concept, and distributional similarity to capture the occurrence of information to extract important information in an image. VGG16, a current state-of-the-art deep learning algorithm, uses the dataset produced to train this model.

Kuznetsova et al proposed a complex tree-based approach that consists of image descriptions that make use of web images with captions. The model makes use of the dataset introduced by Ordonez [44]. The fundamental concept behind this method is to grasp visual content similarity to find expressive phrases, which are key pieces of text in an existing description, and then combine them to generate a new description. It uses image caption generation along with image caption generalization as an optional preprocessing step to perform compression to remove text information that does not directly describe visual content, this process is referred to as pruning. Images that are visually similar to the given query image are retrieved, and four potentially useful types of phrases (noun phrase, verb phrase, preposition phrase based on relative spatial relation and overall scene) are extracted from their corresponding image descriptions, and these retrieved text fragments are used to compose a new image description. Visual resemblance and the grammatical parse of the related textual description drive the extraction of valuable phrases. This extraction method tries to make the most of language regularities in terms of objects, actions, and scenes, allowing for richer textual descriptions. The model accepts phrases as input and aims to create a final sentence by rearranging the selected phrases using sequencing, an objective function, and numerous constraints by encoding using Integer Linear Programming (ILP).

Retrieval-based approaches are also able to accurately caption images although there are some drawbacks to this approach as it generates descriptions for query images by transferring well-formed human-written captions. Even though the resulting outputs are often grammatically acceptable and fluent, confining visual descriptions to pre-existing phrases prevents them from adapting to new scenarios. In some cases, created

descriptions may even be unrelated to the contents of the image.

3) Novel Methods

Image captioning has improved a great deal with the aid of state-of-the-art deep neural networks which can be able to handle complexities quite well. In general, in this approach features in a given image are obtained which is utilized to construct a caption for the image using a language model to generate grammatically correct words, as opposed to earlier approaches. Features are learned automatically from training data in deep machine learning-based techniques, and they can handle a huge and diverse set of images and videos. Convolutional Neural Networks (CNN) are commonly used for feature learning, while a classifier like SoftMax is commonly used for classification. Recurrent Neural Networks (RNN) are commonly used to generate captions after CNN. Image captioning with neural networks is divided into many categories based on the model's architecture, the number of captions created, the learning type, mapping features, and the language model employed, among other factors.

One of the first to suggest a work in multimodal space, which consists of a language Encoder part, a vision part, a multimodal space part, and a language decoder part, was Kiros [50]. This model uses CNN to extract image features and a multimodal space that represents both image and text simultaneously. It relies on high-level image attributes and word representations learned by deep neural networks and multimodal neural language nets, respectively, and does not require any additional templates, structures, or constraints. Because neural language models are unable to process vast amounts of data rapidly, hence work badly with long-term memory.

A reinforcement learning-based image captioning model is introduced by Ren [51] which consists of two networks that jointly compute the next best word at each time step. The policy network serves as local advice for predicting the next word depending on the existing situation and value. The network functions as a global guidance system, evaluating the reward value while considering all conceivable expansions of the present state. By modifying the network in predicting accurate words, the model provides captions that are similar to the ground truth. The model is trained by employing an actor-critic reinforcement learning model. The exact reward value in anticipating the correct word is calculated via visual semantic embedding. It aids in determining the degree of resemblance between pictures and sentences, which may be used to determine the accuracy of a generated phrase.

C. VIDEO CAPTIONING

Video captioning is a verbal explanation of the process of creating video material. It has more information in it than a static image. As a result, video captions require more elements to be extracted, which is more complex than image captioning. The video feature extraction component and the video description generating section are the two parts of the video captioning tasks.

1) Templated-based approach

Template-based methods construct a video description by slot filling appropriate words after choosing an output sentence structure from a collection of predefined sentence structures that is relevant to the video.

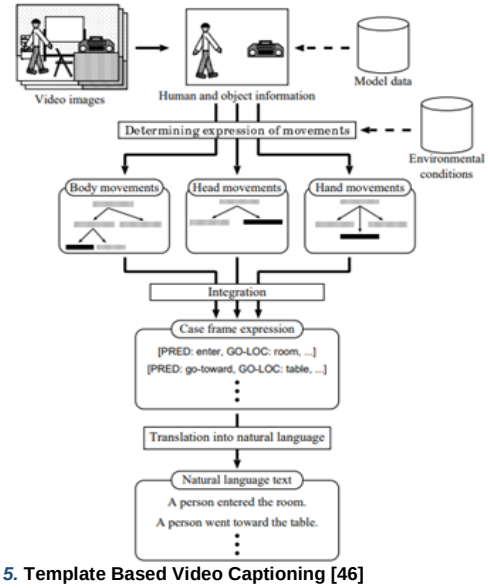

FIGURE 5. **Template Based Video Captioning [46]**

By extracting semantic aspects of human motions and forming connections with the notion of the hierarchy of actions, Kojima et al. [46] present a technique of generating captions that explain the human behavior seen in the video. The body and skin areas of a human are extracted pixel by pixel from each frame of the input video by calculating the difference in colors between input and background images. Perspective transformation is used to determine the positions of the head and hands. In addition, by comparing the edges and color histograms of extracted object areas with those of object models, the object may be recognized. Then, using domain knowledge and the position/posture of the person retrieved from the video, conceptual descriptions of activities are constructed for each body component. A collection of semantic primitives is used to generate and classify hierarchies of activities for each bodily component of the human. The most relevant predicate, object, and so on are chosen by establishing a link between a semantic primitive of action and a feature derived from the video pictures. These syntactic components are combined into a case frame, which is combined into a single frame that depicts complete body motion. The case frame is then transformed into a natural language statement using syntactic rules and a natural dictionary.

Rather than trying to predict a caption based on the most visually likely objects and events, Guadarrama [47] introduces a model that can predict a less specific or more general sentence that is both visually reasonable and

instructive. Semantic hierarchies are utilized to learn from the data to assist an acceptable amount of generalization. Firstly, subject, verb, and object triplets are extracted from natural language descriptions of the videos where each node represents the collection of labels associated with it, and a different semantic hierarchy is built for each member of the triplet (subject, verb, and object). For each leaf of the subject, verb, and object hierarchical trees, a visual model is created that maximizes the semantic similarity to the training data to learn to predict node triplets over the learned hierarchies, a language model from large-scale text corpora is used as a prior on triplets to infer verbs missing in the vocabulary. Lastly, the best possible triplets are used to produce phrases.

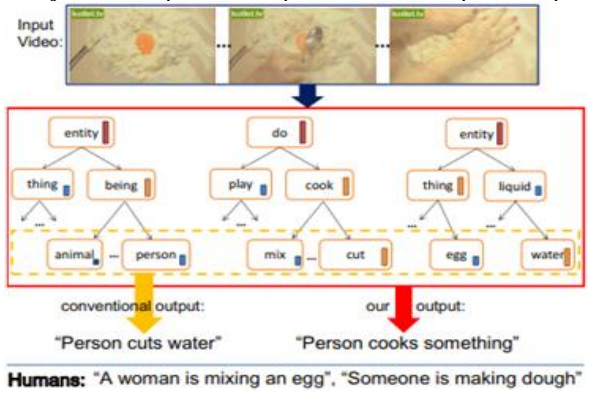

FIGURE 6. **Less Specific Action Captioning [47]**

Although template-based methods may produce full sentences, the descriptions that are generated are quite constrained. Meanwhile, evaluations are typically constrained to a tiny language and a narrow topic. The requisite complexity of rules and templates makes manual template design impractical or too expensive for any suitably rich domain.

2) Sequential classification-based method

Sequential classification-based algorithms provide a video description by repetitively anticipating the next words.

Previous research by Guadarrama et al [47] has reduced the challenge by recognizing a predetermined set of semantic roles, such as subject, verb, and object, as an intermediate representation. This fixed representation is challenging for extensive vocabularies, and it also leads to oversimplified restrictive sentence templates that fail to replicate natural language's complex patterns. The research by Venugopalan et al [48] suggests using a deep learning network to translate video pixels into conversational words. To extract essential characteristics, the model uses long short-term memory (LSTM) neural network to describe sequencing but links it directly to a deep CNN to analyze incoming video frames, bypassing the intermediate representations entirely. This model is similar to the image-to-text methodology; however, it has been modified to work with video sequences. The method uses a spatially invariant convolutional network pre-trained on 1.2M+ pictures with category labels, to perform feature extraction on each frame of the movie. The meaning state and sequence of words are then generated by a recurrent LSTM deep network pre-trained on 100K+ Flickr and COCO images with associated phrase descriptions. Finally, the deep ConvNet creates a robust visual representation of the video's objects, activities, and scenes. The model collects features for each frame, pools them throughout the whole video, and feeds them into the LSTM network at each step. The LSTM outputs one word at a time until it selects the end-of-sentence tag, depending on the video characteristics and preceding word in some cases. These models build a fixed dimensional vector representation by first applying a feature transformation to an image. RNN is used to decode the vector into a phrase or a sentence.

The above method falls short of making greater use of temporal information in the video as it averages features from all frames into a single vector representation of the video which is overcome by using the temporal attention mechanism introduced by Yao [49] which makes use of 3D CNN-RNN encoder-decoder framework. The 3D Conventional Neural Network retrieves more abstract action-related characteristics from the video while preserving and emphasizing key local structures. By raising the attention weights of the associated temporal characteristic, the attention mechanism allows the RNN decoder to selectively focus on a limited selection of frames. If the data has a meaningful temporal structure, the attention mechanism allows the decoder to take advantage of it. This method increases a level of flexibility and generalizability far above what has been seen previously as this model gives an indication that the new state of the art in video captioning is likely to be a single model trained end to end that is capable to take pixel input only and output a natural language caption.

IV. HUMAN MOVEMENT DETECTION

In recent years, human movement detection, as a research topic in the field of computer vision, has been widely explored and studied by scholars all over the world. Compared with the research on human movement in static images, video human movement detection pays more attention to the temporal and spatial changes of the human body in the video image sequence. According to the image frame or image sequence in the video, systems automatically identify human movement through computer processing and analysis of visual information [52]. As a key technology in video understanding, human movement detection is a significant research topic with many applications such as intelligent surveillance, intelligent homes, virtual reality and video retrieval. The diversity of human movement, the variation of video perspective and the complexity of non-rigid motion make human movement detection a great

challenge [53].

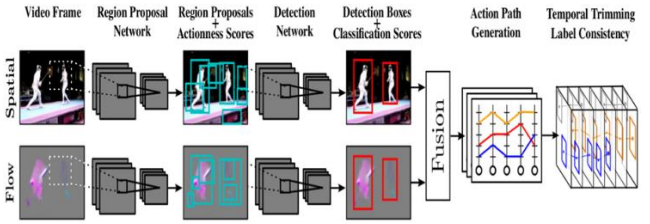

FIGURE 7. **The stream chart of human movement detection [52]**

Human movement features are the representation of key information extracted from video data, and are the key part of movement detection. The classification of human movement is to take the feature vector of human movement as input, train a classifier through the method of machine learning, input the feature vector to be recognized into the classifier and receive the classification result of the movement [55]. Traditional manual features are not universal for problems such as illumination, occlusion, and perspective change in different complex scenes. Therefore, automatic feature learning from data in the way of deep learning is more effective [56]. Different from methods based on traditional manual features, the process framework of the human movement detection method based on deep learning is shown in the figure below.

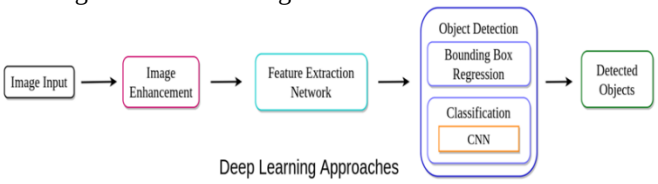

FIGURE 8. **Deep learning approaches on human action detection**

The human movement detection method based on deep learning uses a trainable feature extraction model to automatically learn movement representation from video in an end-to-end manner to complete classification. At present, the Network structure of movement detection methods based on deep learning mainly includes Two-Stream Networks and 3D Convolution Networks. There are also some lesser established methodologies in literature such as restricted Boltzmann machine, Recurrent Neural Networks (RNN), and Independent Subspace Analysis (ISA) that also achieved good performance [57].

A. DATASETS

The performance of human movement detection methods needs to be analyzed and compared under the same dataset. Many of the original datasets used for human movement detection such as KTH and the Weizmann set [58,59] were a good starting point for investigation in the area but lack real world features such as non-static backgrounds, varied camera angles and changes in lighting and occlusions. These sets were effective for validating some of the first work on human movement detection but lacked the complexity to allow models to function well in the real world.

In order to improve on these sets such as UCF-Sports, Hollywood2, HMDB-51 and UCF-101 utilized video from movies, television, and the internet, labelling these videos with the appropriate movement classifications in order to create sets with more diverse camera angles, backgrounds and intra-class variation. The most diverse set to date is the Kinetics data set. Published in 2017 it includes more than 30,000 video clips from 400 categories of human movement, all taken from the YouTube video library, making it the largest human movement detection dataset publicly available [60].

B. CURRENT NETWORKS AND ALGORITHMS

3) Two-Stream Network

The two-stream network structure to perform movement detection was proposed by Simon Yan in 2014, and its basic process is shown in FIGURE 9[61]. A two-stream network structure is divided into two branches: time-stream convolutional neural network and spatial-stream convolutional neural network, and the two branches have the same network structure. The temporal convolutional neural network firstly calculates the optical flow image of two adjacent frames in the video sequence and then obtains temporal information in the form of an optical flow image of multiple stacked frames. In the case of Spatial stream convolution, the spatial features of the RGB are extracted to classify action based on objects and features of the images. The final classification results are then obtained by merging the two networks. This method greatly improves the detection of visual frequency Specific accuracy [62].

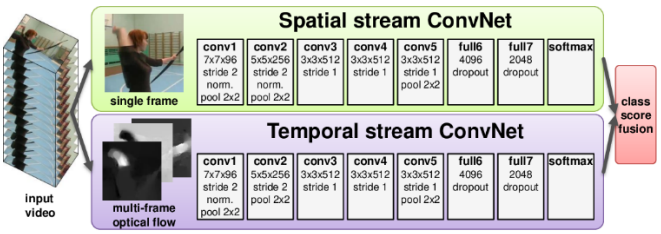

Figure 1: **Two-stream architecture for video classification.**

FIGURE 9. **Two stream networks [61]**

The initial two-stream Network suffers from insufficient modeling ability for long videos, this led to the creation of the Temporal Segment Network (TSN) [63]. The principal difference being it uses a sparse time sampling strategy to segment videos in time domain and then randomly extract fragments. However, with the further development of human motion detection, two problems of two-stream structure emerged. Firstly, the pixel-level correspondence between spatial and temporal features cannot be learned. Secondly, space domain convolutions are only on single RGB frames, and time-domain convolution is only on the adjacent, stacked time-sequence optical flow frames, with a very limited time scale, resulting in the inability to use two very important cues in the video for behavioural recognition, i.e. the inability to observe changes in the corresponding optical flow (temporal cues) at the same

time in the specified appearance location area (spatial cues), and the changing trend of spatial cues over time [64].

Inspired by the outstanding performance of residual networks in the field of image recognition, the ST ResNet (spatio temporal Residual Network) was developed on the basis of the two-stream network. The ST ResNet expands on the time domain and continues the characteristics of network integration by finding a pixel-level correspondence between the time and space domains. In two-stream networks, the time domain network and the space domain network also use residual connection to transfer parameters, and this spatio-temporal coherence is an important clue in the process of processing video data. In addition, the authors adopt a new approach to temporal residuals, using a small asymmetric filter, by converting the spatial filter dimension mapping in the residual path into a temporal filter for temporal convolution. This temporal filter learns the changes in scene and motion. By stacking these filters, deeper spatio-temporal features can be learned. [65].

Fernando et al. [66] effectively encode the timing information of video clips from CNN features based on video frames by using the pooling method of Rank-Pooling to learn all parameters and video features of the model end-to-end in a double-layer optimization mode. Bilen et al. [67] proposed Dynamic Image Networks on this basis, which collects videos into an Image called Dynamic Image (DI) and then learns a Ranking Machine to capture the temporal evolution of the data and use the parameters of the ranker as a representation. Using the idea of multi-stream networks, the author takes static image, dynamic image, optical stream image and dynamic optical stream image as input, proposes a four-stream network model, and combines it with Improved Dense Trajectories, and achieves a good detection effect. The feature of this method is to convert the video into a DI, so that the existing CNN model can be immediately extended to the video after pre-training the still image [68].

4) 3D Convolutional Network

At present, most movement detection methods use 2D convolutional neural network based on images to learn CNN features of a single image, this tends to ignore the connection between consecutive frames, resulting in the loss of video action information. Therefore, using 3D convolutional networks to learn video movement representation has become an important area of research.

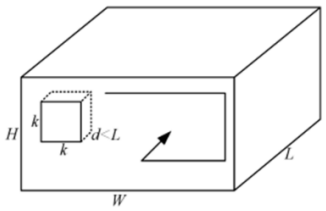

FIGURE 10. **3D Convolution Algorithm [69]**

3D CNNs differ from 2D CNNs in the fact that the data passed into the network is 3 dimensional with the 3rd dimension in this case representing time. Therefore, the kernels used in the convolution layers are also 3 dimensional. This allows the pixel-level temporal changes across frames in a video to be learned by the network. Wang [70] proposed the C3D network to learn spatiotemporal features and found the most appropriate length of sequential convolution kernel for 3D convolution through systematic study, including 8 convolution layers with 3×3×3 convolution kernels, 5 pooling layers, and 2 fully connected layers. They proved that the features extracted by C3D are universal, efficient, and compact. C3D has an accuracy of 82.3% on UCF-101 dataset and 51.6% on HMDB-51 dataset. Subsequently, Tran et al. [71] proposed Res3D network by combining residual network (Res-Net) with C3D network. Res3D network further improves network performance, running twice as fast as C3D, with half the mode size and 3.5% and 3.3% higher accuracy on UCF-101 and HMDBG51 datasets. Mansimov et al.[72] proposed several methods to initialize 3D convolution weights using 2D convolution weights, including averaging, scaling, zero weight and negative weight initialization, in order to avoid zero-training 3D convolution network and to be able to use the knowledge learned from 2D convolution. Experimental results show that the negative weight initialization method achieves the best result among all initialization methods. By demobilizing a 3×3×3 3D convolutional filter into a 1× 3×3 convolutional filter and a 3×1×1 convolutional filter, Qiu et al. [73] designed three library blocks based on a residual network and proposed a pseudo-3D residual network Pseudo-3D ResNet not only significantly reduces the size of the model, but also makes use of 2D convolutional neural networks trained in image datasets

5) Restricted Boltzmann Machine

A restricted Boltzmann machine is a generation network model that can learn probability distribution through input data sets. Neurons in the input/output layer and neurons in the hidden layer are connected by a weight matrix W and bias vector. Neurons in the same layer are independent of each other, and neurons in different layers relate to each other. Taylor et al. [74] used gated restricted Boltzmann machine to learn motion information in videos in an unsupervised way and fine-tune network parameters in combination with convolution, which could effectively extract motion-sensitive features and achieved satisfactory results in KTH dataset and Hollywood2 dataset. Tanaka used such an efficient method for Gaussian-constrained Boltzmann machines to learn differences in human motion features in video, defining a difference subtraction function between two frames, creating a simple temporal and spatial saliency map that highlights motion patterns in space by eliminating motion-related shapes, and background image detection are not relevant, making it easier for shallow RBMs to learn the actions in these saliency maps [75]. The advantage of methods based on restricted Boltzmann

machine is that they can use unlabeled data to carry out unsupervised learning and obtain spatial-temporal feature representation method.

6) Attention Based Methods

Limited by hardware such as GPU and CPU, many movement detection methods based on deep learning cannot directly input the whole video into the network model to extract features the depth feature can be extracted by using the information redundancy between continuous frames to represent the whole video. At present, most of the existing research methods still use the whole image for feature extraction, which cannot distinguish the foreground and background well, and the global motion information and the local human motion information may have the defect of losing the key motion information. Wang [76] captured long-term dependencies directly by calculating the relationship between any two positions between adjacent frames, independent of position distance, and was able to effectively turn off motion images in visual frequencies and model the non-local part of the visual frequency. Zhang et al. [77] proposed the interactive perception space-time pyramid attention network, which takes advantage of the high correlation of local features of adjacent positions in space and constructs multi-scale feature maps into spatial pyramids, adding self-attention to weight each spatial position and focusing more on moving objects themselves. Recently, methods combined with the attention mechanisms have demonstrated the effectiveness of the attention mechanism, with initial success in improving detection performance.

7) Unsupervised Methods

At present, the identification of human movement requires a lot of labeled samples for training, but in practice, because of the huge amount of video data and diverse content available, the cost of labelling such huge datasets is also large and therefore difficult to put into use in industry. Discrim-Net, proposed by Rychener et al. [78], uses unlabeled data to train a generative adversarial network model in an unsupervised pre-training way, and then fine-tuning movement data sets with a smaller labeled set to perform specific tasks. However, due to the difficulty of training generative adversarial network, unsupervised or semi-supervised attempts are still in the preliminary stages, however, future human movement detection methods are expected to develop towards unsupervised or semi-supervised learning [79].

8) Computationally Simple Methods

Current models for human movement detection based on deep learning are often computation complex and slow, meaning they do not meet the requirements of real-time applications in the industry. Therefore, it is expected that the efficiency and timeliness of human movement detection methods will be improved in the future. The efficient convolutional network was proposed by Zolfaghari et al [80]. processes each video quickly while considering the long-term content using a selective sampling strategy, and uses the inter-frame redundancy to quickly achieve high-quality movement classification. Moreover, the whole network model only has fewer layers. At present, the state of the art models are too large to be used in mobile devices, networks with smaller computational demands such as that proposed by Zolfaghari would allow for use in these devices however currently they cannot match the detection performance of some of the more complex models.

V. CASE STUDIES

A. HUMAN OBJECT INTERACTION DETECTION:

1) Spiideo

Spiideo is a Sweden based company that develops and manufactures camera systems for recording, analysing, and streaming sports. The camera resolution could range anywhere from 4K to 12K with horizontal view ranging from 100° to 170°.[94] The camera's software has a feature that auto follows the main object of the game so that the users don't miss the real action. They have a range of camera systems that are flexible for any type of sport, lighting, and weather.[82]

The Auto Follow feature is an AI tracker that positions the camera towards the action. Using the image from the camera the model recognizes the human-object interaction and positions its view. This feature is also known as," Virtual Camera Man".[83] Apart from delivering high-quality footage to the audience, the system's footage is also used by the suitable coach and analyst to level up their game in the next match.

2) Tesla

Tesla is a clean energy company based in Austin, Texas which was founded by a group of engineers in the year 2003. The total revenue generated by the company in 2021 was $ 53.82 Billion.[93]

FIGURE 11. **Display on 15" Touchscreen on a dashboard in one of the Tesla cars[102]**

In 2014, Tesla was able to build its very own autonomous vehicle called, Model S which could autonomously steer, adjust speed and brake on the principle of signal image processing. All the Tesla cars have 8 cameras for 360° view, 12 Sonar (sound navigation and ranging) sensors, 1 RADAR and GPS.[88] Through these sensors,

information is transferred to the Full Self-Driving computer and using the data from the cameras with the help of neural networks the model identifies the objects in the image and converts it into vector space, and predicts an outcome for that event. This allows the car to react depending on the human actions for instance if the car encounters a cross walk it will either stop or drive through the crossing depending on whether a human can be seen to be interacting with the crossing [88].

B. HUMAN MOVEMENT DETECTION:

1) Stonkam

Stonkam, is a high-tech enterprise that aims to research, develop and manufacture vision products that can work as advanced driving assistance. The company was founded in 2003 in Guangzhou, China [95]. As per the company's report, every year 1.25 million lives are lost due to road-related accidents. Among these traffic accidents, one cause is due to the blind spot.[81]

FIGURE 12. **STONKAM 1080P HD Intelligent Real-Time Pedestrian Detection and Alarm Camera [90]**

They have developed a product called STONKAM 1080P HD Intelligent Real-Time Pedestrian Detection and Alarm Camera which can detect pedestrians from any side of the car. The product has a 170° horizontal viewing angle. The product can be used on large vehicles like trucks, buses, construction machines, etc [94]. Apart from pedestrian detection technology the company also focuses on applications like Face Recognition, Fatigue Driver monitoring, Distracted Driver monitoring, Cell phone use monitoring, Forward Collision Warning and many more.[81]

2) HireVue

HireVue is a company based in Utah, U.S. where they help American companies to hire talented candidates. Hiring someone means many rounds of assessments and interviews, to simplify the process the company uses an AI-powered interview system that assesses the candidate's potential [90]. During the video interview, the AI application records your video and assesses your answer, attire, hand gestures, and posture by using the video footage from your webcam. The AI performs human-movement detection to understand your non-verbal cues during the interview [89]. The company has managed to host more than 18 million video interviews for over 700 companies around the globe [96].

C. IMAGE AND VIDEO CAPTIONING:

1) Facebook

Facebook is an American social media tech multinational company with over 2.9 Billion users across the world based in California, United States [100]. The company has developed an image analysis AI which can generate captions for media content posted on their platform [97]. The decision-making happens with the help of a neural network model which potentially could be an RNN (Recurrent Neural Network) model. RNNs are often used for Image Captioning. By captioning the image, the company can understand the trends in society and help them to make profitable decisions. According to Statista in 2021, the company changed its name to Meta [100] and managed to generate $1,17,929 Million [99].

2) YouTube

TABLE I
YOUTUBE'S DISTRIBUTION OF VIDEOS REMOVED FROM ITS PLATFORM WORLDWIDE IN Q4 2021 [87]

Category Removed	Percentage
Spam, Misleading and Scams	9.1%
Child Safety	31.5%
Nudity or Sexual	18.4%
Violent or Graphics	19.9%
Harmful or Dangerous	8.1%
Hateful or Abusive	1.9%
Promotion of Violence and Violent Extremism	2.4%
Others	8.6%

YouTube is a multinational video-sharing social media platform based in California, U.S. It has more than 2 billion active monthly users. 300 hours of video content is uploaded on this platform every minute [88]. Also, according to a YouTube Press report, 1 billion hours of video content is watched every day [88]. The parent company is Google. YouTube cannot stop any of its users from posting but it can remove inappropriate content. According to Statista, the company has removed inappropriate content as mentioned in TABLE I above in the fourth quarter of 2021[87]. It is quite unrealistic for a human to remove these types of content manually, so Google uses a sophisticated Machine Learning algorithm to detect explicit content and delete it from the platform [86]. A video file is a combination of multiple image frames. Each image is passed through a CNN. These images are classified and captioned, and if the video is inappropriate, it will be removed [85]. The company has managed to generate $28.8 Billion in 2021 [84] which makes it the second most visited site in the United States [88].

ACKNOWLEDGMENT

The group would like to thank Dr. Mazheruddin Syed for his mentorship and guidance throughout this project.

REFERENCES

[1] A. Bobick and J. Davis, "An appearance-based representation of action," *IEEE Xplore*, Aug. 01, 1996. https://ieeexplore.ieee.org/document/546039 (accessed Apr. 14, 2022).

[2] W.-L. Lu and J. J. Little, "Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor," *IEEE Xplore*,

Jun. 01, 2006. https://ieeexplore.ieee.org/document/1640361 (accessed Apr. 14, 2022).

[3] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *Proceedings of the 15th international conference on Multimedia - MULTIMEDIA '07*, vol. 15, 2007, doi: 10.1145/1291233.1291311.

[4] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010, doi: 10.1109/tpami.2009.154.

[5] S. Sunaina, R. Kaur, and D. V. Sharma, "A Review of Vision-Based Techniques Applied to Detecting Human-Object Interactions in Still Images," *Journal of Computing Science and Engineering*, vol. 15, no. 1, pp. 18–33, Mar. 2021, doi: 10.5626/jcse.2021.15.1.18.

[6] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to Detect Human-Object Interactions," *IEEE Xplore*, Mar. 01, 2018. https://ieeexplore.ieee.org/abstract/document/8354152 (accessed Feb. 18, 2022).

[7] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap," *IEEE Access*, vol. 8, pp. 218386–218400, 2020, doi: 10.1109/access.2020.3042484.

[8] B. X. B. Yu, Y. Liu, and K. C. C. Chan, "Skeleton-Based Detection of Abnormalities in Human Actions Using Graph Convolutional Networks," *IEEE Xplore*, Sep. 01, 2020. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9253160 (accessed Apr. 15, 2022).

[9] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," *IEEE Xplore*, Jun. 01, 2010. https://ieeexplore.ieee.org/document/5540234 (accessed Apr. 15, 2022).

[10] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," *Procedings of the British Machine Vision Conference 2010*, 2010, doi: 10.5244/c.24.97.

[11] S. Dara and P. Tumma, "Feature Extraction By Using Deep Learning: A Survey," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1795–1801, Mar. 2018, doi: 10.1109/ICECA.2018.8474912.

[12] S. Gupta and J. Malik, "Visual Semantic Role Labeling," *NASA ADS*, May 01, 2015. https://ui.adsabs.harvard.edu/abs/2015arXiv150504474G/abstract (Accessed Apr. 15, 2022).

[13] M. Dogariu, L.-D. Stefan, M. G. Constantin, and B. Ionescu, "Human-Object Interaction: Application to Abandoned Luggage Detection in Video Surveillance Scenarios," *IEEE Xplore*, Jun. 01, 2020. https://ieeexplore.ieee.org/document/9141973 (accessed Apr. 15, 2022).

[14] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual Translation Embedding Network for Visual Relation Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, doi: 10.1109/cvpr.2017.331.

[15] M. P. Zapf, A. Gupta, L. Y. Morales Saiki, and M. Kawanabe, "Data-Driven, 3-D Classification of Person-Object Relationships and Semantic Context Clustering for Robotics and AI Applications," *IEEE Xplore*, Aug. 01, 2018. https://ieeexplore.ieee.org/document/8525654 (accessed Apr. 15, 2022).

[16] D.-T. Le, J. Uijlings, and R. Bernardi, "'TUHOI: The Universal Human Object Interaction Dataset,'" *COLING'14 workshop on Vision and Language (VL'14)*, 2014.

[17] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "PPDM: Parallel Point Detection and Matching for Real-Time Human-Object Interaction Detection," *IEEE Xplore*, Jun. 01, 2020. https://ieeexplore.ieee.org/abstract/document/9156683 (accessed Apr. 15, 2022).

[18] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual Relationship Detection with Language Priors," *arXiv:1608.00187 [cs]*, Jul. 2016, Accessed: Feb. 15, 2022. [Online]. Available: https://arxiv.org/abs/1608.00187

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, doi: 10.1109/cvpr.2014.81.

[20] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective Search for Object Recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, Apr. 2013, doi: 10.1007/s11263-013-0620-5.

[21] Y. Jia *et al.*, "Caffe: Convolutional Architecture for Fast Feature Embedding," *arXiv.org*, 2014. https://arxiv.org/abs/1408.5093

[22] Y. Verma, "R-CNN vs Fast R-CNN vs Faster R-CNN - A Comparative Guide," *Analytics India Magazine*, Sep. 10, 2021. https://analyticsindiamag.com/r-cnn-vs-fast-r-cnn-vs-faster-r-cnn-a-comparative-guide/ (accessed Nov. 26, 2021).

[23] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv.org*, 2014. https://arxiv.org/abs/1409.1556

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Neural Information Processing Systems*, 2015. https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html (accessed Feb. 08, 2022).

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, doi: 10.1109/cvpr.2016.91.

[26] S. Shinde, A. Kothari, and V. Gupta, "YOLO based Human Action Recognition and Localization," *Procedia Computer Science*, vol. 133, pp. 831–838, 2018, doi: 10.1016/j.procs.2018.07.112.

[27] Kajabad, E. and Ivanov, S., 2019. People Detection and Finding Attractive Areas by the use of Movement Detection Analysis and Deep Learning Approach. *Procedia Computer Science*, 156, pp.327-337.

[28] Gao, C., Zou, Y., & Huang, J. (2018). iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. *ArXiv, abs/1808.10437*.

[29] Qi, S., Wang, W., Jia, B., Shen, J., & Zhu, S. C. (2018). Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 401-417).

[30] A. Prest, V. Ferrari and C. Schmid, "Explicit Modeling of Human-Object Interactions in Realistic Videos," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 835-848, April 2013, doi: 10.1109/TPAMI.2012.175.

[31] A. Prest, C. Schmid and V. Ferrari, "Weakly Supervised Learning of Interactions between Humans and Objects," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 601-614, March 2012, doi: 10.1109/TPAMI.2011.158.

[32] S. -C. Hsu, Y. -W. Wang and C. -L. Huang, "Human Object Identification for Human-Robot Interaction by Using Fast R-CNN," 2018 Second IEEE International Conference on Robotic Computing (IRC), 2018, pp. 201-204, doi: 10.1109/IRC.2018.00043.

[33] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 17-24, doi: 10.1109/CVPR.2010.5540235.

[34] Wan, B., Zhou, D., Liu, Y., Li, R., & He, X. (2019). Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9469-9478).

[35] Fang, H. S., Cao, J., Tai, Y. W., & Lu, C. (2018). Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European conference on computer vision*

(ECCV) (pp. 51-67).

[36] Gupta, T., Schwing, A.G., & Hoiem, D. (2019). No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9676-9684.

[37] Cao, Z., Hidalgo, G., Simon, T., Wei, S. and Sheikh, Y., 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1), pp.172-186.

[38] Y. -L. Li et al., "Transferable Interactiveness Knowledge for Human-Object Interaction Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3580-3589, doi: 10.1109/CVPR.2019.00370.

[39] W. Yan, Y. Gao and Q. Liu, "Human-object Interaction Recognition Using Multitask Neural Network," 2019 3rd International Symposium on Autonomous Systems (ISAS), 2019, pp. 323-328, doi: 10.1109/ISASS.2019.8757767.

[40] Messing, R., Pal, C. & Kautz, H., 2009 "Activity recognition using the velocity histories of tracked keypoints" *ICCV 2009*.

[41] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing Simple Image Descriptions using Web-scale N-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, USA. Association for Computational Linguistics.

[42] Farhadi, A. *et al.* (2010). Every Picture Tells a Story: Generating Sentences from Images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg.

[43] G. Kulkarni et al., "Baby talk: Understanding and generating simple image descriptions," CVPR 2011, 2011, pp. 1601-1608, doi: 10.1109/CVPR.2011.5995466.

[44] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2Text: describing images using 1 million captioned photographs. In Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11). Curran Associates Inc., Red Hook, NY, USA, 1143–1151.

[45] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: data, models and evaluation metrics. J. Artif. Int. Res. 47, 1 (May 2013), 853–899.

[46] Kojima, A., Tamura, T. & Fukunaga, K. Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. *International Journal of Computer Vision* **50,** 171–184 (2002).

[47] S. Guadarrama et al., "YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition," 2013 IEEE International Conference on Computer Vision, 2013, pp. 2712-2719, doi: 10.1109/ICCV.2013.337.

[48] Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

[49] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision* (pp. 4507-4515).

[50] Kiros, R., Salakhutdinov, R., & Zemel, R. (2014, June). Multimodal neural language models. In *International conference on machine learning* (pp. 595-603). PMLR.

[51] Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L. J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 290-298).

[52] T. Kirishima, K. Sato and K. Chihara, "Real-time gesture recognition by learning and selective control of visual interest points," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 351-364, March 2005, doi: 10.1109/TPAMI.2005.61.

[53] J.T. Chien and C.C. Wu, "Discriminant Waveletfaces and Nearest Feature Classifiers for Face Recognition", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1644-1649, Dec. 2002.

[54] L. Lam and C.Y. Suen, "Optimal Combinations of Pattern Classifiers", *Pattern Recognition Letters*, vol. 16, no. 9, pp. 945-954, 1995.

[55] K. Woods, W.P. Kegelmeyer and K. Bowyer, "Combination of Multiple Classifiers Using Local Accuracy Estimates", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 405-410, 1997.

[56] X. Gao, S. Lin and T. Y. Wong, "Automatic Feature Learning to Grade Nuclear Cataracts Based on Deep Learning," in *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 11, pp. 2693-2701, Nov. 2015

[57] Q. V. Le, W. Y. Zou, S. Y. Yeung and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," *CVPR 2011*, 2011, pp. 3361-3368.

[58] L. Deng, "A tutorial survey of architectures algorithms and applications for deep learning", *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[59] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi and R. V. Babu, "A taxonomy of deep convolutional neural nets for computer vision", *Frontiers in Robotics and AI*, 2016.

[60] Forsyth and M. Fleck, "Body plans", *IEEE Conference on Computer Vision and pattern Recognition*, 1997

[61] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:*1409.1556 (2014).

[62] Donner, Tobias H., et al. "Population activity in the human dorsal pathway predicts the accuracy of visual motion detection." *Journal of neurophysiology* 98.1 (2007): 345-359.

[63] Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." *European conference on computer vision. Springer, Cham*, 2016.

[64] J. White, T. Kameneva and C. McCarthy, "Deep reinforcement learning for task-based feature learning in prosthetic vision", *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 2809-2812, 2019.

[65] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[66] Fernando, Basura, and Stephen Gould. "Learning end-to-end video classification with rank-pooling." *International Conference on Machine Learning.* PMLR, 2016.

[67] H. Bilen, B. Fernando, E. Gavves and A. Vedaldi, "Action Recognition with Dynamic Image Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2799-2813, 1 Dec. 2018,

[68] Bay, T. Tuytelaars and L. Van Gool, "LNCS 3951 - SURF: Speeded Up Robust Features", *Comput. Vision- ECCV 2006*, pp. 404-417, 2006.

[69] Maturana, Daniel, and Sebastian Scherer. "Voxnet: A 3d convolutional neural network for real-time object recognition." *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE, 2015.

[70] Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." *European conference on computer vision. Springer, Cham*, 2016.

[71] Tran, Du, et al. "Convnet architecture search for spatiotemporal feature learning." *arXiv preprint arXiv*:1708.05038 (2017).

[72] Mansimov, Elman, Nitish Srivastava, and Ruslan Salakhutdinov. "Initialization strategies of spatio-temporal convolutional neural networks." *arXiv preprint arXiv*:1503.07274 (2015).

[73] Qiu, Zhaofan, Ting Yao, and Tao Mei. "Learning spatio-temporal representation with pseudo-3d residual networks." *proceedings of the IEEE International Conference on Computer Vision*. 2017.
[74] Taylor, Graham W., and Geoffrey E. Hinton. "Factored conditional restricted Boltzmann machines for modeling motion style." *Proceedings of the 26th annual international conference on machine learning.* 2009.
[75] Graham W. Taylor and Geoffrey E. Hinton, "Factored conditional restricted Boltzmann Machines for modeling motion style." *the 26th Annual International Conference on Machine Learning (ICML '09). Association for Computing Machinery*,vol 1025–1032.
[76] Wang, Shuo, et al. "Pm2. 5-gnn: A domain knowledge enhanced graph neural network for pm2. 5 forecasting." *Proceedings of the 28th International Conference on Advances in Geographic Information Systems.* 2020.
[77] Zhang, Hong-Bo, et al. "A comprehensive survey of vision-based human action recognition methods." *Sensors* 19.5 (2019): 1005.
[78] Peng Zhang, Yongju Bai and Mu, "Renlong used image processing algorithms [J]", 12 (05): 61-63. *Research and Exploration of Computer Vision Lab*, vol. 12, no. 05, pp. 61-63, 2018.
[79] Singh, Ghanapriya, et al. "A personalized classifier for human motion activities with semi-supervised learning." *IEEE Transactions on Consumer Electronics* 66.4 (2020): 346-355.
[80] Zolfaghari, Mohammadreza, Kamaljeet Singh, and Thomas Brox. "Eco: Efficient convolutional network for online video understanding." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
[81] AI STONKAM CO. LTD [Online]. Available: https://stonkam.com/AI/ai_2.html
[82] Sports Video Camera Systems – Spiideo [Online]. Available: https://www.spiideo.com/products/sports-video-camera/ (accessed Apr. 19, 2022).
[83] Spiideo AI game tracking system - automatically analyse and follow the play -Spiideo [Online]. Available: https://www.spiideo.com/automatic-player-tracking/ (accessed Apr. 19, 2022).
[84] 84 YouTube statistics you can't ignore in 2022 [Online]. Available: https://invideo.io/blog/youtube-statistics/ (accessed Apr. 19, 2022).
[85] YouTube's new AI will block videos that are inappropriate for kids [Online]. Available:https://thenextweb.com/news/youtubes-new-ai-will-block-videos-that-are-inappropriate-for-kids (accessed Apr. 19, 2022).
[86] How Video Classifier Help Youtube? | Techfastly. Available:https://techfastly.com/youtube-content-filters/ (accessed Apr. 19, 2022).
[87] Share of removed YouTube videos by reason 2021 | Statista [Online]. Available: https://www.statista.com/statistics/1132956/share-removed-youtube-videos-worldwide-by-reason/ (accessed Apr. 19, 2022).
[88] Watch Tesla Unveil its Full Self-Driving Computer in Under 5 Minutes – YouTube [Online]. Available: https://www.youtube.com/watch?v=bZxTG7DmB_0 (accessed Apr. 19, 2022).
[89] 5 Tips for AI-Powered Interview Success [Online] Available:https://firsthand.co/blogs/interviewing/tips-for-ai-powered-interview-success (accessed Apr. 19, 2022).
[90] HireVue – Hiring Experience Platform [Online]. Available: https://www.hirevue.com/ (accessed Apr. 19, 2022).
[91] Facebook: Annual Revenue | Statista [Online] Available;https://www.statista.com/statistics/268604/annual-revenue-of-facebook/ (accessed Apr. 19, 2022).
[92] These are Facebook's secret rules for removing posts [Online] Available:https://eu.usatoday.com/story/tech/news/2018/04/24/facebook-discloses-secret-guidelines-policing-content-introduces-appeals/544046002/ (accessed Apr. 19, 2022).
[93] Tesla's turnover 2008-2018 | Statista [Online]. Available: https://www.statista.com/statistics/272120/revenue-of-tesla/#:~:text=Tesla's%20revenue%20grew%20to%20around,increase%20from%20the%20previous%20year. (accessed Apr. 19, 2022).
[94] STONKAM 1080P & 720P HD Vehicle Vision Systems | For safer driving – STONKAM CO., LTD [Online] Available: https://stonkam.com/products/HD-Systems.html?gclid=Cj0KCQjwjN-SBhCkARIsACsrBz4k2g3AZVYvACjMUZ-qF97d9PzddtrVf6GKxxuQHxTrxvOwCE8znmgaApQVEALw_wcB (accessed Apr. 19, 2022).
[95] STONKAM CO., LTD.| Introduction [Online] Available: https://stonkam.com/aboutus/introduction.html (accessed Apr. 19, 2022).
[96] HireVue – LinkedIn | About [Online]. Available: https://www.linkedin.com/company/hirevue/about/ (accessed Apr. 19, 2022).
[97] Facebook and Instagram's AI-generated image captions now offer far more details – TechCrunch [Online]. Available: https://tcrn.ch/3iqRoQZ (accessed Apr. 19, 2022).
[98] Number of monthly active Facebook users worldwide as 4th quarter 2021 [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (accessed Apr. 19, 2022).
[99] Meta's (foremerly Facebook Inc.) annual revenue from 2009 to 2021 [Online]. Available: https://www.statista.com/statistics/268604/annual-revenue-of-facebook/#:~:text=In%202021%2C%20Meta's%20(formerly%20Facebook,of%20income%20is%20digital%20advertising. (accessed Apr. 19, 2022).
[100] Introducing Meta: A Social Technology Company [Online]. Available:https://about.fb.com/news/2021/10/facebook-company-is-now-meta/ (accessed Apr. 19, 2022).
[101] YouTube's Revenue | Google [Online] Available: https:/www.google.com/search?q=youtube+revenue (accessed Apr. 19, 2022).
[102] "Tesla unveils autopilot system, but comes with caution notice," The Indian Express [Online] https://indianexpress.com/article/technology/tech-news-technology/tesla-unveils-autopilot-system-but-comes-with-caution-notice/ (accessed Apr. 19, 2022).