Artificial Intelligence Review (2024) 57:178
https://doi.org/10.1007/s10462-024-10730-5

Multimodal vision-based human action recognition using deep learning: a review

Fatemeh Shafizadegan[1] Ahmad R. Naghsh-Nilchi[1] Elham Shabaninia[2]

Published online: 19 June 2024

Abstract

Vision-based Human Action Recognition (HAR) is a hot topic in computer vision. Recently, deep-based HAR has shown promising results. HAR using a single data modality is a common approach; however, the fusion of different data sources essentially conveys complementary information and improves the results. This paper comprehensively reviews deep-based HAR methods using multiple visual data modalities. The main contribution of this paper is categorizing existing methods into four levels, which provides an in-depth and comparable analysis of approaches in various aspects. So, at the first level, proposed methods are categorized based on the employed modalities. At the second level, methods categorized in the first level are classified based on the employment of complete modalities or working with missing modalities at the test time. At the third level, complete and missing modality branches are categorized based on existing approaches. Finally, similar frameworks in the third category are grouped together. In addition, a comprehensive comparison is provided for publicly available benchmark datasets, which helps to compare and choose suitable datasets for a task or to develop new datasets. This paper also compares the performance of state-of-the-art methods on benchmark datasets. The review concludes by highlighting several future directions.

Keywords Deep learning · Human action recognition · Multimodality · Visual modality

✉ Ahmad R. Naghsh-Nilchi
nilchi@eng.ui.ac.ir

Fatemeh Shafizadegan
fatemeh.shafizadegan.1990@eng.ui.ac.ir

Elham Shabaninia
e.shabaninia@kgut.ac.ir

[1] Department of Artificial Intelligence, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

[2] Department of Applied Mathematics, Faculty of Sciences and Modern Technologies, Graduate University of Advanced Technology, Kerman, Iran

1 Introduction

Human action recognition is a hot topic in computer vision with a wide range of applications such as video security and surveillance systems (Kwak and Song 2013), human-computer interaction (Choi et al. 2008), smart homes (Amiri et al. 2014), robotics (Akkaladevi and Heindl 2015; Van Amsterdam et al. 2022), content-based video retrieval (Jones and Shao 2013), entertainment (Shotton et al. 2011), sports events (Soomro and Zamir 2015), sign language recognition (Azar and Seyedarabi 2020), etc. A generic framework for action recognition typically consists of some components, including data acquisition, data preprocessing, feature extraction, temporal modeling, classification, and evaluation.

Vision-based human action recognition is the task of assigning a label representing the recognized action of a human (performing a simple gesture/action or an inter-object interaction) or a group of people (performing human-human interactions or group activities) in a trimmed (segmented) or untrimmed (continues) clip of video or in a still image. The inter-class and intra-class variations of different actions (due to different viewpoints, appearances, lighting conditions, occlusion, cluttered background, and varying speed) make human activity recognition a challenging task.

Fortunately, the previous decade began with surprising news from the computer vision community by announcing the success of deep neural networks (DNNs) and especially convolutional neural networks (CNNs) in challenging computer vision tasks such as image classification (He et al. 2016), object detection (Girshick et al. 2014), object recognition (Liang and Hu 2015), image segmentation (Bi et al. 2018), and action recognition (Simonyan and Zisserman 2014a). These approaches and the development of computing technologies contributed to expanding research in the computer vision field. Although, this new approach started with a short delay for human action recognition due to difficulties in using DNNs for video analysis. Today, almost all state-of-the-art methods for HAR are based on deep learning approaches.

Distinct modalities offer complementary information for robust action recognition and provide compensatory information in the case of missing modalities. Activity recognition with multiple visual modalities is regarded as a new promising approach that offers an in-depth understanding of different actions (Yadav et al. 2021; Sun et al. 2022; Majumder and Kehtarnavaz 2020). However, action recognition with multimodal data is challenging due to the heterogeneity of different data sources, large amounts of data, different fusion strategies, and the need for transferring knowledge.

Although different review papers exist in the literature for deep-based HAR, limited papers reviewed multimodal vision-based HAR using deep approaches. As there exist lots of papers in HAR, this review paper tried to compare and classify the existing approaches from different points of view. This review focuses on multimodal visual approaches and categorizes them into four levels (see Fig.1). This review helps readers better comprehend the HAR approaches as well as provides a means for comparing the frameworks in various aspects. Besides, vision-based HAR benchmark datasets are studied, and the best results on popular and newest ones are reported as well.

Our four-level categorization is based on the mentioned components in a generic framework for HAR except data acquisition. In the first level of proposed categorization, methods are classified based on used modalities, i.e., RGB & depth, RGB & skeleton, depth & skeleton, RGB & depth & skeleton, and infrared & other visual modalities.

There are two different approaches for using multimodal data in the second level of our categorization. The first approach considers more than one modality to make use of different

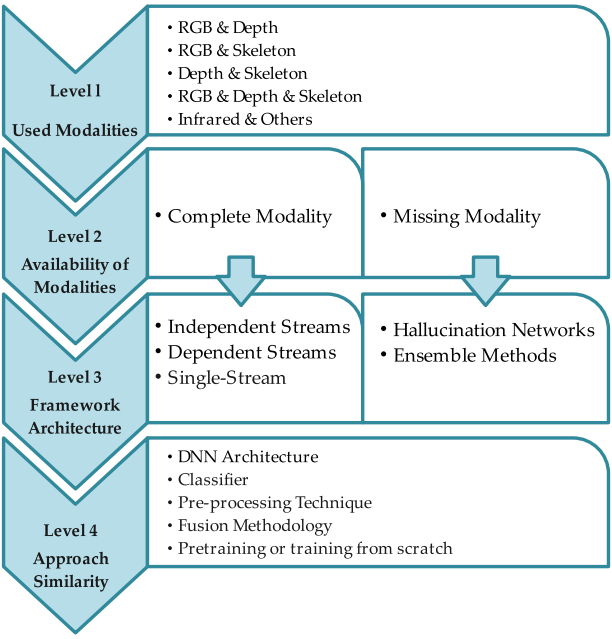


Fig. 1 Four-level classification of deep-based HAR approaches using multiple visual modalities

and complementary information of distinct modalities. Methods in this approach usually work with multiple streams of data (dependent or independent) and try to fuse features from different modalities. Here, all modalities are present both in training and test time. However, real-life applications usually miss one or more modalities during test time due to the cost, noise, sensor failure, privacy, etc. The second approach targets missing modality scenarios by using multiple modalities during training to compensate for missed modality in the test time through knowledge transfer or co-learning. So, methods in the second level are categorized into two branches: complete modality and missing modality.

At the third level, methods in the second level are grouped based on the framework architecture. Complete modality approaches are grouped into independent streams, dependent streams, and single-stream. Missing modality approaches are grouped into hallucination networks and ensemble methods. Finally, in the fourth level, similar frameworks in the third level are grouped according to the DNN architecture, classifier, fusion, and preprocessing methodology.

All methods are also summarized in Appendix A based on our four-level categorization for ease of access in network architecture comparison and gaining more insights into framework design.

Besides, this paper reviews almost all available and related benchmark datasets, along with categorizing and comparing them with each other. Datasets are grouped based on providing trimmed or untrimmed videos, the number of viewpoints (as single-view or multi-view), and visual data modalities (as RGB+depth, RGB+skeleton, depth+skeleton, and RGB+depth+skeleton). Furthermore, the results of state-of-the-art methods are reported on popular and the newest benchmark datasets. The main contributions of this review in supervised deep-based HAR techniques using multiple visual data modalities are three folds as follows:

1. A new categorization is proposed for the first time that classifies methods into four levels. Used modalities, using complete or missing modalities, architecture (based on the number and the dependency of network streams and learning methodology), and framework similarities.
2. Novel categorization and comparison are proposed on available multimodal vision-based benchmark datasets of HAR.
3. Different methods are discussed along with challenges, open issues, and new trends to provide insightful guidance on future directions for research.

Also, methods with the best results on popular and the newest benchmark datasets are highlighted.

The rest of the paper is organized as follows. In Sect. 2, a brief review of relevant surveys is presented. Methodology is stated in Sect. 3. A generic framework for HAR is presented in Sect. 4. Section 5 provides a brief review of unimodal vision-based HAR using deep learning. In Sect. 6, various multimodal vision-based HAR methods are studied and analyzed in detail. In Sect. 7, multimodal visual HAR datasets are categorized. We discuss the studied methods and some future research directions in Sects. 8 and 9, respectively. Finally, the paper concludes in Sect. 10.

2 Relevant surveys

As mentioned before, considerable research has been devoted to human activity recognition during the last decades. Besides, many surveys were published based on different characteristics of deep-based HAR methods (see Table 1). Some focused on both traditional and deep-based approaches (Yuanyuan et al. 2021; Pareek and Thakkar 2021; Khan and Ghani 2021; Rangasamy et al. 2020; Jegham et al. 2020b; Zhang et al. 2019; Dhiman and Vishwakarma 2019; Estevam et al. 2021; Özyer et al. 2021).

While others concentrated only on deep-based methodologies (Shabaninia 2022; Ulhaq et al. 2022; Ahmad et al. 2021; Islam et al. 2022; Zhu et al. 2020; Yao et al. 2019; Sreenu and Durai 2019). A group of reviews studied a specific data modality such as visual or sensor-based methods (Chen et al. 2021; Nguyen et al. 2021; Hussain et al. 2020; Dang et al. 2020; Beddiar et al. 2020; Al-Faris et al. 2020; Wang et al. 2019, 2018). Some others dealt with multiple data modalities in HAR (Yadav et al. 2021; Majumder and Kehtarnavaz 2021; Sun et al. 2022; Majumder and Kehtarnavaz 2020; Li et al. 2020; Roitberg et al. 2019; Liu et al. 2019). Also, a branch of surveys focused on HAR applications (Prati et al. 2019; Mar et al. 2019). Further, some surveys reviewed benchmark multimodal visual or RGB-D datasets in HAR (Singh and Vishwakarma 2019b, a; Zhang et al. 2016; Cai et al. 2017).

Although numerous surveys concentrated on deep-based HAR, limited works focus on combining different visual data modalities. Some papers focused on a few combinations of visual data modalities. For example, in (Majumder and Kehtarnavaz 2021; Roitberg et al. 2019), only the fusion of RGB & depth modalities is investigated. The fusion of depth & skeleton is also studied in (Liu et al. 2019). As outlined above, Sun et al (Sun et al. 2022) investigated combinations of visual data modalities; however, limited papers were surveyed and categorized. Compared to (Sun et al. 2022), this review categorizes more methods into four levels, which provides an in-depth analysis of issues.

Table 1 Recent surveys on HAR

Category	References	Main focus
Conventional and deep-based	Yuanyuan et al. (2021)	Categorizing deep approaches into Two-stream, 3D CNN, and LSTM groups.
	Pareek and Thakkar (2021)	Reviewing several extracted features in action presentation, action analysis-based HAR, applications, and challenges in HAR.
	Khan and Ghani (2021)	Categorizing deep approaches into CNN, RNN, and hybrid groups.
	Rangasamy et al. (2020)	Studying methods in sport video analysis and grouping deep-based architectures into CNN, 3D CNN, RNN, and LSTM.
	Jegham et al. (2020b)	Reviewing methods for challenges and issues in HAR. Besides, approaches are grouped into template-based, generative, and discriminative models.
	Zhang et al. (2019)	Overviewing methods in action feature representation based on deep learning.
	Dhiman and Vishwakarma (2019)	Investigating methods in abnormal human action recognition.
	Estevam et al. (2021)	Studying zero-shot video-based action recognition methods.
Deep-based	Shabaninia (2022)	Reviewing vision Transformer and temporal modeling.
	Ulhaq et al. (2022)	Reviewing vision Transformer.
	Islam et al. (2022)	Surveying CNNs based on input device.
	Ahmad et al. (2021)	Studying GCNs.
	Zhu et al. (2020)	Categorizing deep-based approaches into CNNs, two-stream, and 3D CNNs.
	Yao et al. (2019)	Studying CNN-based approaches.
	Sreenu and Durai (2019)	Reviewing application of HAR in video surveillance for crowd analysis.
Specific unimodal data modality	Chen et al. (2021)	Surveying deep approaches in sensor-based methods.
	Nguyen et al. (2021)	Reviewing deep-based approaches and power requirements in mobile and wearable sensors.
	Hussain et al. (2020)	Categorizing methods into wearable, object-tagged, and device-free approaches with a specific focus on device-free methods, especially RFID. Further, device-free studies are grouped into action, motion, and interaction.
	Dang et al. (2020)	Analyzing vision-based and sensor-based methods and their corresponding procedure of data collection, preprocessing, feature engineering, and training.
	Beddiar et al. (2020)	Reviewing vision-based approaches according to feature extraction process, recognition stage, source of input data, and machine learning supervision level. Besides, body organs are compared during actions.
	Al-Faris et al. (2020)	Categorizing vision-based approaches using deep learning into generative and discriminative models.
	Wang et al. (2019)	Analyzing ten approaches using conventional and deep strategies on visual modality (Kinect-based).
	Wang et al. (2018)	Classifying deep-based segmented and continuous motion recognition approaches into RGB, depth, skeleton, and hybrid.

Table 1 continued

Category	References	Main focus
Multimodal data modalities	Yadav et al. (2021)	Grouping deep and conventional approaches into vision-based (including RGB camera, RGB-D sensor, and 3D point cloud), wearables (including inertial sensor, footwear, and egocentric camera), and multimodal (including RGB & inertial sensor, depth & inertial sensor, and multiple wearable sensors) categories.
	Majumder and Kehtarnavaz (2021)	Categorizing deep and conventional approaches into RGB, depth, and RGB & depth.
	Sun et al. (2022)	Categorizing deep approaches into visual modality (including RGB, 3D skeleton, depth, infrared, point cloud, and event stream) and non-visual modality (including audio, acceleration, radar, and WiFi). Further, fusion and co-learning-based approaches with visual modalities and visual & non-visual modalities are considered.
	Majumder and Kehtarnavaz (2020)	Grouping traditional and deep approaches into the fusion of RGB & inertial, depth & inertial, and RGB & depth & inertial.
	Li et al. (2020)	Classifying methods in multi-user or group activity recognition into vision-based, sensor-based, radiofrequency, and hybrid groups.
	Roitberg et al. (2019)	Grouping fusion strategies in multimodal (RGB & depth) gesture recognition on C3D architecture into late fusion, information exchange, and linking information.
	Liu et al. (2019)	Reviewing deep and conventional methods with depth, skeleton, and hybrid features.

3 Methodology

The guidelines followed in this paper are taken from (Harris et al. 2014; Wright et al. 2007). The performed literature review process comprises four steps: formulating research objectives, selecting eligibility criteria, identifying a search strategy, and conducting data extraction (Adewopo et al. 2022).

3.1 Research objectives

This study aims to address the following research questions:

- **RQ1:** What are the main deep-based HAR techniques that use multimodal visual data modalities?
- **RQ2:** What are the primary datasets and metrics used in this scope?
- **RQ3:** What are the best results and future directions in this field of study?

3.2 Eligibility criteria

This review includes the papers related to action recognition. It includes topics in deep-based HAR, multimodal action recognition, activity recognition using visual data modalities, multimodal vision-based HAR, gesture recognition, and group activity recognition published in journals and conferences between 2016 and 2023. The initial works commenced in 2016 and have since been ongoing. Only papers published in the English language were used. The publications needed to meet the following characteristics in order to be included:

- Action recognition or gesture recognition tasks,
- Deep-based approaches,
- Multiple visual data modalities,
- Trimmed (segmented) datasets,
- Within 2016–2023 (7 years).

The following exclusions were implemented:

- Employing untrimmed datasets.
- Does not provide clear findings and analysis of results.
- Written in other languages excluding English.

3.3 Information sources

The selection of papers for this review was conducted through a comprehensive search of electronic databases that specifically included articles published in the English language. The databases, including IEEE Xplore, Wiley Online Library, Springer Link, Science Direct, ACM, and arXiv were utilized as the primary source for identifying relevant articles on action recognition tasks. These databases encompass a wide range of full-text journals and conference papers.

3.4 Search strategy

The following keywords are combined with conjunctions “AND” and disjunctions “OR” in our search. The most common terms used for our search were:

- Action recognition.
- Activity recognition.
- Gesture recognition.
- Recognizing action.
- Motion recognition.

Above terms were combined with:

- Deep network.
- Multimodal.

- Visual data.
- Vision-based modalities.

The abstracts, titles, keywords, and employed datasets from selected articles were reviewed to assess their relevance according to the inclusion and exclusion criteria. Articles that did not meet the eligibility criteria or were not pertinent to the research questions were excluded from the study.

3.5 Data extraction

A full text reading of selected articles was conducted to retrieve the relevant data that enabled to answer the research questions, categorize the studies and identify the future research prospects. The following data are extracted from the selected studies:

(1) Document title, authors' names, publication year and journal/conference name,
(2) Used modalities,
(3) Fusion techniques,
(4) Framework architecture,
(5) Datasets and corresponding results.

Appendix A summarizes the extracted data from selected studies.

4 A generic framework for HAR

A generic unimodal or multimodal HAR system typically consists of several components, as illustrated in Fig. 2. These components include data acquisition, data preprocessing, spatial feature extraction and temporal modeling, classification, and evaluation.

Initially, the system needs to capture data that contains human actions. This data can be obtained from either an egocentric or third-person view. Egocentric videos are captured from a first-person perspective, where the camera is mounted on the head or body of the person recording the video. On the other hand, third-person videos are captured from a third-person perspective, where the camera is positioned outside of the person being recorded.

Preprocessing techniques are utilized to enhance the quality of inputs before they are fed into next stage. These techniques can be customized and combined to meet the requirements of a specific computer vision task. Common data preprocessing techniques include resizing

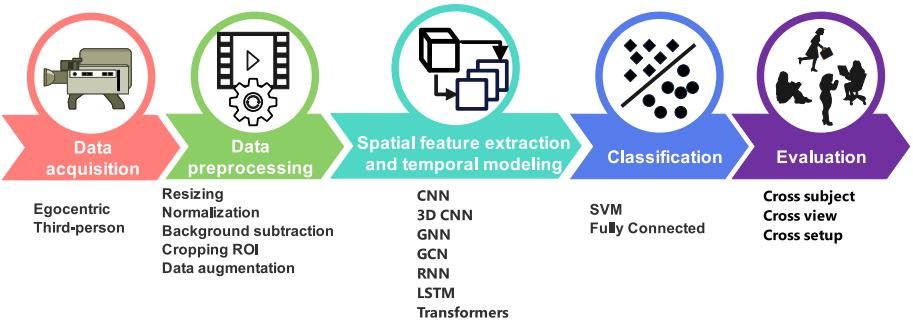


Fig. 2 A generic framework for HAR

(changing input dimensions), normalization (scaling pixel values), cropping regions of interest (eliminating irrelevant parts of the input), background subtraction (removing distracting objects in the background), and data augmentation (creating new data by applying random transformations to existing data).

The next stage involves spatial feature extraction and temporal modeling, which is the most crucial phase in the system. Various techniques are utilized for this purpose and DNNs are the focus of this paper as they are commonly used and provide state-of-the-art results.

CNNs are mainly used to extract local features that are meaningful and shared throughout the data. They operate on fixed-size vectors with a fixed number of computational steps (Alom et al. 2019). 3D CNNs have been developed using 3D convolutions (Ji et al. 2012). Compared to CNNs, 3D CNNs can better extract dependencies between adjacent frames (Ji et al. 2012). However, their performance requires a lot of computational resources during the training stage, and they are rigid in capturing action sequences with fine-grained visual patterns (Köpüklü et al. 2022). Graphical neural networks (GNNs) extend deep learning techniques to non-Euclidean or graph data (Ahmad et al. 2021). CNNs have also been extended to non-Euclidean data via graphical convolutional networks (GCNs). Compared to CNNs, GCNs provide unordered and variable-sized structures.

Recurrent neural networks (RNNs) (Alahi et al. 2016), Long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) and Transformers (Vaswani et al. 2017) are applied for temporal sequence modeling in video analysis. However, RNNs suffer from vanishing gradient problems. LSTM overcame this issue, but the number of learnable parameters in LSTM is high (Salehinejad et al. 2017). Transformers are capable of modeling long-range dependencies between input elements of a sequence, and they support parallel sequence processing, unlike RNNs and CNNs. Additionally, their design requires minimum biases (Khan et al. 2022).

After extracting features, the actions are labeled in the classification stage. Fully connected networks or SVMs are used as classifiers. Different protocols are used to evaluate the HAR system. These include cross-subject (people with different appearances, cultures, genders, and ages), cross-view (different views like front or side), and cross-setup (different distances, heights, etc.).

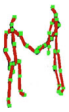
The design of an action recognition system is a complex task that requires careful consideration of various factors in order to achieve high accuracy.

5 Unimodal vision-based HAR

RGB-D sensors usually provide RGB (a three-channel data encoding color information), depth (containing information relating to the distance of surfaces from a viewpoint), and skeleton (encoding 3D location of joints). Each modality possesses its own specification in HAR. Table 2 lists the pros and cons of using different visual modalities.

RGB-based HAR is very common in computer vision tasks. However, depth information is lost in traditional RGB images. With the emergence of low-cost depth sensors such as Microsoft Kinect (Tölgyessy et al. 2021), Intel RealSense (Keselman et al. 2017), and ASUS Xtion (Gonzalez-Jorge et al. 2013), the use of depth data has increased a lot for HAR. Depth sensors use different technologies, such as structured light (measuring the distortion of a pattern projected on an object caused by the object's surface) or time of flight (ToF) (measuring the phase delay of reflected infrared light) (Zanuttigh et al. 2016). Due to the sensitivity of IR cameras to sunlight, cameras that rely on structured light technology are not

Table 2 Different visual modalities used in HAR with an example (Hands Shaking) from NTU RGB+D dataset (Shahroudy et al. 2016)

Modality	Example	Pros	Cons
RGB		• Providing appearance information • Ease of access • Applicable in indoor and outdoor environments	• Sensitive to viewpoint • Sensitive to illumination • Sensitive to background
Depth		• Providing 3D structures • Invariant to illumination • Invariant to appearance • Usually applicable in in-door	• Noisy data • Lack of color and texture
Skeleton		• Providing higl-level 3D information of joints • Invariant to background • Invariant to motion speed and scale • Low dimensional space	• Lack of appearance • Human-object interaction • Noisy data
Infrared		• Applicable in dark environments	• Sensitive to illumination • Lack of color and texture

usable outdoors. However, they are employed in HAR tasks, which do not require very high-depth resolution (Kazmi et al. 2014).

Deep-based activity recognition with visual data modalities is classified into unimodal and multimodal approaches (see Fig. 3). Most of unimodal approaches use a single stream in their framework, while some with multi-stream frameworks utilize several streams from different shapes of one modality (Özyer et al. 2021; Fu et al. 2020). These unimodal multi-stream approaches are categorized into multi-resolution (Karpathy et al. 2014), multi-rate (Feichtenhofer et al. 2019), and multi-feature (Simonyan and Zisserman 2014a) structures. The multi-resolution structure employs different resolutions of one data modality in each stream (Fig. 4a). In the multi-rate structure, different frame rates of one input are fed to the multiple streams of the network (Fig. 4b). Lastly, the multi-feature structure uses various extracted features (such as optical flow) from one data modality in the streams (Fig. 4c). Although all these approaches use multiple data streams, they are not considered multi-modal, which is the focus of this review.

Although many studies concentrate on unimodal approaches, the interest in multimodal ones is also increasing. This paper focuses on HAR methods using multiple visual data modalities.

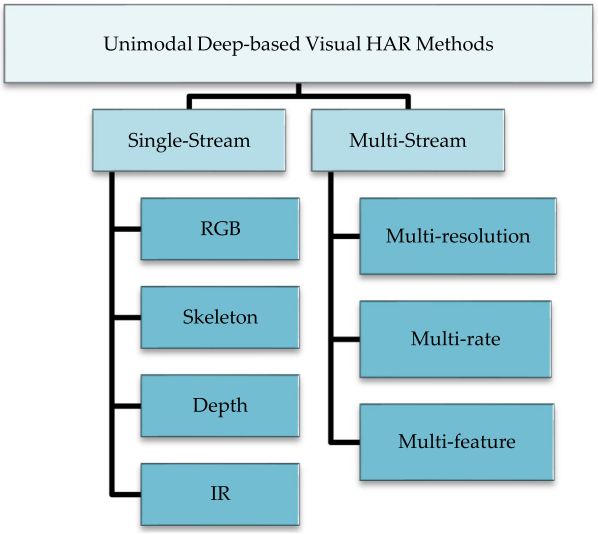


Fig. 3 Classification of unimodal visual HAR methods using deep approaches

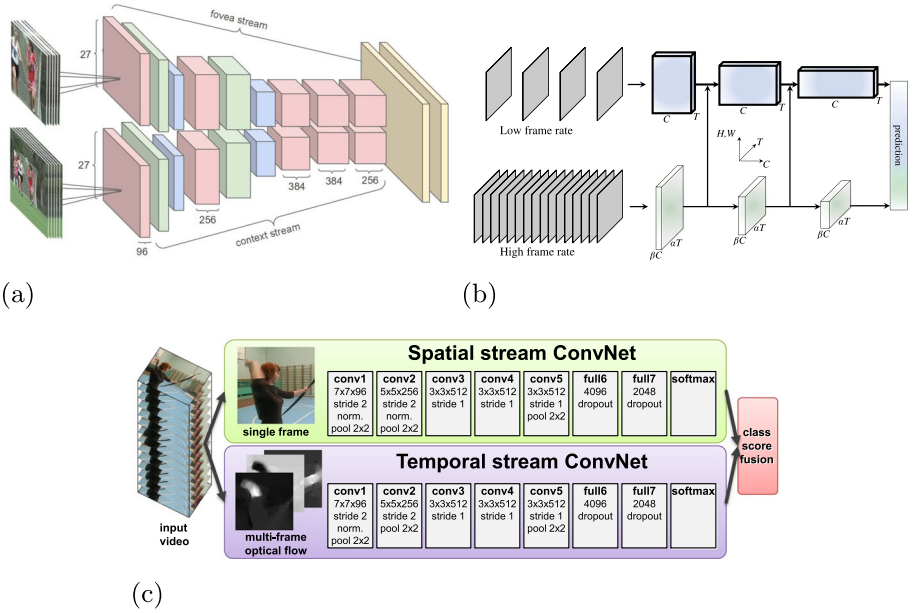


Fig. 4 Examples of unimodal multi-stream DNNs. **a** Multi-resolution (Karpathy et al. 2014), **b** Multi-rate (Feichtenhofer et al. 2019), and **c** Multi-feature (Simonyan and Zisserman 2014a) structures

6 Multimodal vision-based HAR

Research on multimodal vision-based activity recognition shows that using multiple modalities can achieve higher accuracy than unimodal approaches because they benefit from different and complementary sources of information (Sun et al. 2022; Roitberg et al. 2019; Liu et al. 2019).

Action recognition with multimodal data is challenging due to several factors. Firstly, the heterogeneity of different data sources poses a challenge, as modalities can have different data types, formats, and noise levels. Handling such heterogeneous data requires careful consideration. Secondly, collecting large amounts of multimodal action datasets takes time and effort. Compared to single-modal datasets, large-scale multimodal datasets are relatively scarce and small. This limitation restricts the availability of training data for model training. Thirdly, information fusion from different modalities effectively is a complex task. Various modalities may have varying levels of importance for different actions, and finding the proper fusion strategy is crucial. Moreover, simultaneously processing multiple modalities increases the computational complexity of action recognition frameworks. This can pose challenges for real-time performance, especially when dealing with high-dimensional data. Additionally, different modalities may require specific feature extraction techniques for effective fusion. Therefore, adapting feature extraction methods to suit different modalities is necessary. Lastly, aligning different modalities in time is crucial for accurate action recognition. However, each distinct modality may have a different sampling rate or temporal resolution, making it difficult to synchronize them. Temporal misalignment can lead to inaccurate recognition results.

This paper reviews studies that employ two or more data modalities as inputs. For example, in (Qin et al. 2020; Luvizon et al. 2018; Liu and Yuan 2018; Zolfaghari et al. 2017), depth or pose are created from RGB and used for action classification. Others (Davoodikakhki and Yin 2020) use pose only in the preprocessing stage. As mentioned before, approaches that construct optical flow from RGB (Liu et al. 2018) or similar features from a modality lie in the multi-feature subgroup of the multi-stream unimodal category, as shown in Fig. 3. These papers are beyond the scope of this paper.

This paper proposes a novel four-level categorization of multimodal approaches (see Fig. 1). First, multimodal vision-based HAR methods are categorized into five major categories: RGB & depth, RGB & skeleton, depth & skeleton, RGB & depth & skeleton, and infrared & other visual modalities. Since few works employ infrared along with other visual modalities, the existing papers are analyzed only in a single subsection.

Second, methods of five major categories are grouped into complete and missing modalities. As previously stated, all data modalities available in the training phase may not be available during the test due to different reasons. For the case of complete modalities, three framework architectures are popular in the third level of categorization based on the number of used streams and their dependencies: independent streams, dependent streams, and single-stream. Two different approaches in the literature try to solve the missing modality problem in the third level: Hallucination networks and Ensemble methods (see Fig. 1). Lastly, methods of the third level with similar approaches in DNN architecture, classifier, fusion, and preprocessing techniques are placed in the fourth level.

As Baltrusaitis et al. (Baltrušaitis et al. 2018) categorized multimodal machine learning challenges, multimodal HAR approaches lie in fusion, translation, and co-learning groups. Fusion joins information from all modalities in training and test time. In contrast, co-learning could handle missing modalities or assist learning in modality with fewer samples.

In translation, a modality is derived from another modality before the training process, and then all modalities are used in the training and test stage. That is why the framework architecture of fusion and translation HAR methods are analogous.

Fusion is used in multimodal learning algorithms to benefit from various predictive powers (Baltrušaitis et al. 2018). There are different taxonomies for the fusion approaches in the literature. Fusion approaches in (Ramachandram and Taylor 2017) are categorized into early, late, and intermediate (as Fig. 5 shows), while authors in (Jain et al. 2005) grouped them into feature-level, score-level, and decision-level (as illustrated in Fig. 6).

Early or data-level fusion involves the integration of multiple raw or preprocessed data modalities into a feature vector before using it in the learning stage. Late or decision-level fusion refers to collecting decisions from multiple classifiers, each trained on distinct modalities. Various rules like maximum or average scores can be used in the late fusion. Intermediate deep fusion is adopted where a shared representation layer is constructed by merging units from multiple paths coming into this layer (Ramachandram and Taylor 2017).

In the other taxonomy, feature-level fusion relates to merging different feature vectors extracted by different algorithms at any layer before fully connected, softmax, and classification layers. The weighted average is used for homogeneous feature vectors and concatenation of non-homogeneous ones (Jain et al. 2005). Decision-level fusion is accomplished after the network prediction. Also, score-level fusion can be performed after or between fully connected and softmax layers (Lai and Yanushkevich 2018).

In independent streams, distinct streams are usually considered for feature extraction of each modality, while the streams do not have any connection or intrusion during extracting features. Then, extracted features are late fused via different fusion strategies in the feature-level (often concatenated), score-level, and decision-level.

Two or more modalities participate in extracting features of each stream in frameworks with dependent streams, i.e., mid-level features of a stream are used in the feature extraction of other streams via an intermediate fusion strategy. These frameworks fuse ultimate features via feature-level, score-level, and decision-level fusion strategies.

Frameworks with only one mainstream lie in the third group. Using an early-level fusion strategy, various data modalities are fed to the network as an N-D entity (for example, 4D for RGB & depth). Other studies make use of a modality as auxiliary data in weighting the features of the primary modality before the classifier layer via late fusion strategies (feature-level fusion).

Further, frameworks that employ one modality as auxiliary data or attention for other primary modalities could possess dependent streams or single-stream. In attention-based

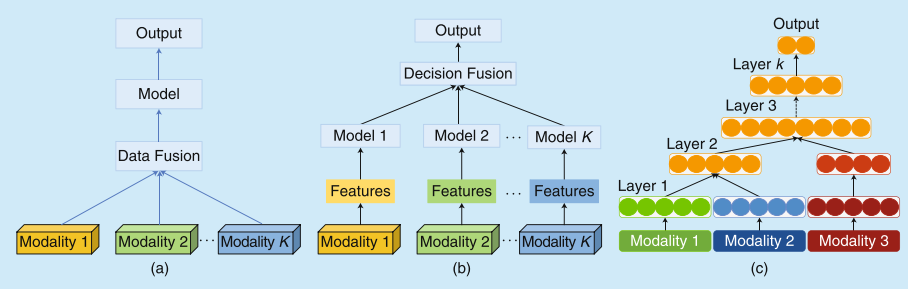


Fig. 5 Various fusion models (Ramachandram and Taylor 2017). (a) Early or data-level fusion, (b) late or decision-level fusion, and (c) intermediate fusion

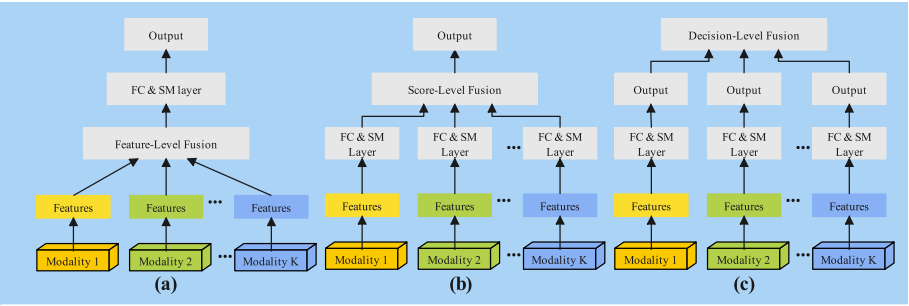


Fig. 6 Various fusion models according to (Jain et al. 2005). (a) Feature-level fusion, (b) score-level fusion, and (c) decision-level fusion

approaches with dependent streams, auxiliary data is trained in a separate stream and used as attention for other modalities during the training phase. The features of all streams are employed in the classification process. Nevertheless, features of auxiliary data in single-stream attention-based approaches are employed only in the training of other modalities and do not participate directly in the classification stage. The difference between attention-based approaches with dependent streams and single-stream is shown in Figs. 7 and 8.

Real-life applications deal with missing modality at test time. Some approaches use multiple modalities to profit from diverse information for missed modalities. Methods of this category mainly employ co-learning approaches. In co-learning, "the knowledge transfers between modalities, their representations, and their predictive models" (Baltrušaitis et al. 2018). Co-learning investigates the ways of transferring knowledge learned from one modality to the model trained on different modalities. It is an appropriate approach in missing or noisy modality issues or when one modality has limited labeled data or samples (Rahate et al. 2022). Therefore, it is a promising approach in HAR working with missing modalities.

Knowledge distillation and transfer learning are co-learning-based techniques usually used to handle missing modalities (Rahate et al. 2022). Transfer learning aims to leverage knowledge from the source domain to the target domain (Zhuang et al. 2020; Tan et al. 2018), while in knowledge distillation, the generalization of a complex model (teacher) is

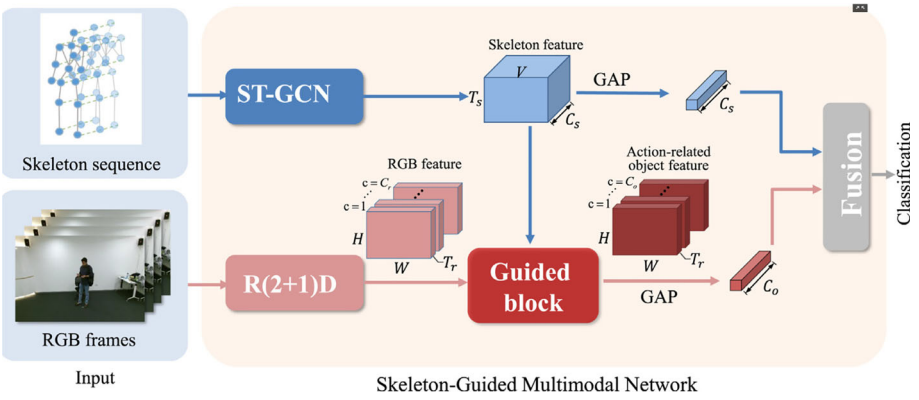


Fig. 7 Attention-based approach with dependent streams (Li et al. 2020)

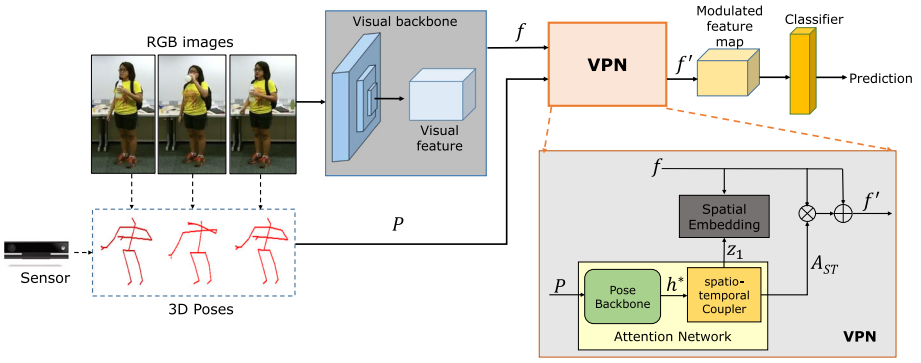


Fig. 8 Attention-based approach with single-stream architecture (Das et al. 2020)

transferred to a more simple model (student) (Wang and Yoon 2021; Gou et al. 2021). Thereby, both techniques train a prototype network, which can generate a new, so-called hallucination network for missed modality (Rahate et al. 2022).

The hallucination network learns privileged information for missing (or noisy) modalities at test time via co-learning approaches. This is primarily proposed by (Hoffman et al. 2016) presented a convolutional hallucination architecture for training an RGB object detection model, which includes depth information at training time. The hallucination network is trained to mimic mid-level features of missed modality and learns a new representation of available ones, as shown in Fig. 9a. At test time, images are processed mutually through the available modalities and hallucination network to increase the detection performance, as shown in Fig. 9b.

Since hallucination networks use the whole data of all modalities indiscriminately and require a pre-trained network, some works suggest ensemble approaches, which do not need a pre-training step or extra networks at test time (Garcia et al. 2019).

At the fourth and last level, methods of the third level are categorized firstly based on DNN architecture, including CNNs, 3D CNNs, RNNs, GNNs, and Transformers. However, multiple DNN architectures are also used in some approaches. Next, methods in each category are grouped based on fusion methodologies, classifiers, and preprocessing techniques. In addition, a group of methods concentrates on extracting specific features such as spatio-temporal, common-specific, or pose features. This level captures the similarities and differences between methods by considering general factors. With the introduction of new approaches over time, it is possible to have a more precise categorization for each of the general factors mentioned here.

In each of the following subsections, a combination of modalities is described (first level of our categorization). Complete and missing modalities are explained for each combination in two subsubsections (second level). Complete modality approaches contain independent streams, dependent streams, and single-stream, while missing modality approaches contain hallucination networks and ensemble methods (third level). Similar frameworks are explained in each paragraph, along with similarities specified at the beginning (fourth level).

6.1 RGB and depth

The main branch of multimodal visual action recognition methods focuses on RGB and depth. As previously mentioned, RGB data represents the appearance information of the

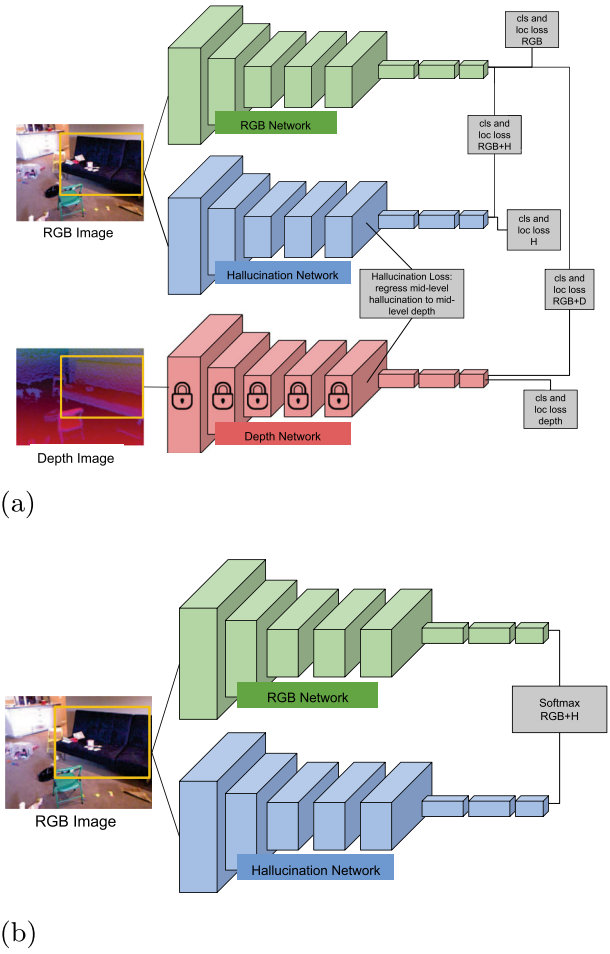


Fig. 9 Hallucination network (Hoffman et al. 2016) in **a** training and **b** test stage

scene and objects, while depth data encodes body silhouette, 3D shape, and scene structure. The depth data is robust against illumination, color, and texture variations. These complementary aspects of RGB and depth encourage many studies to make use of these two modalities. Results from (Das et al. 2017) show that extracting skeleton data is improved via the fusion of a depth-based approach using Kinect and an RGB-based framework using CNNs.

6.1.1 Complete modality

6.1.1.1 Independent streams Most multimodal methods focus on using multiple distinct streams to learn features separately. They either fuse each streams' recognition score or concatenate the extracted features before the classifier layer. In this architecture, streams do not have any connection or do not intrude into other streams while extracting features of other modalities. However, they cannot learn from mid-level complementary information of different heterogeneous modalities.

In Twinanda et al. (2016), Ijjina and Chalavadi (2017), Asadi-Aghbolaghi et al. (2017), Mukherjee et al. (2020) and Sun et al. (2023), multi-stream CNNs are used for extracting features. Extracted features are then concatenated before the classifier. Twinanda et al. (2016) propose a four-stream network pre-trained on AlexNet (Krizhevsky et al. 2017) with RGB, depth, and their motions as inputs for surgical recognition tasks. In Ijjina and Chalavadi (2017), the represented framework for action recognition emphasizes motion in different temporal regions using key poses. It is shown that using multimodal information with the noise tolerance property of convnet features can improve the results. Asadi-Aghbolaghi et al. (2017) consider the fusion of hand-crafted features and deep strategies for RGB-D-based action recognition. For this purpose, dense multimodal trajectories (MMDT) and multimodal 2DCNN approaches are proposed via RGB, depth, scene flow, and optical flow modalities. A two-stream CNN network pre-trained on VGG16 (Simonyan and Zisserman 2014b), or Resnet-101 (He et al. 2016a), is proposed in (Mukherjee et al. 2020) using dynamic images as network inputs. Dynamic images, first introduced in (Bilen et al. 2016), are based on rank pooling; and summarize motion and action information of a video in a single image. Dynamic images are made separately from RGB and depth videos and are fed into the network. Extracted features are concatenated and passed through a fully-connected layer to predict the action class. Sun et al (2023) introduce the multi-level feature fusion of a two-stream CNN.

In Singh et al. (2020), Rajput et al. (2020), Imran and Kumar (2016), a four-stream CNN network is suggested. It takes RGB and depth maps from three different views (i.e., top, front, and side) as inputs. The scores of different streams are late fused at the end of the network to classify the action. Singh et al. (2020) use dynamic images constructed from RGB videos and depth motion map (DMM) and fed to the pre-trained VGG-F model (Zhou et al. 2017). Depth map motion computes the difference of consecutive frames projected on XY, YZ, and XZ planes (corresponding to front, side, and top). A weighted product model is used to classify the action. In Rajput et al. (2020), a motion history image (MHI) is constructed from RGB video and three DMMs as inputs of a pre-trained Mobilenet network. Scores are late fused using product rules of posterior probability generated in each stream. Imran et al. (2016) propose a framework similar to (Rajput et al. 2020) with a pre-trained VGG16 network structure.

3D CNNs are offered in (Zhu et al. 2016; Li et al. 2016b; Miao et al. 2017; Duan et al. 2016; Zhang et al. 2018a; Bini et al. 2022; Chen et al. 2022). To learn gestures from the whole video, Zhu et al. (2016) use pyramid input and pyramid fusion with multiscale contextual information using 3D CNNs. Li et al. (2016b) employ a pre-trained C3D (Tran et al. 2015) network for RGB and depth to recognize gestures. Extracted features are concatenated or averaged. Finally, the gestures are classified with a linear SVM. In (Miao et al. 2017), different features (RGB, flow, and depth) are extracted by the ResC3D network (Miao et al. 2017) and fuse with canonical correlation analysis. The final recognition results are derived with a linear SVM classifier. Chen et al. (2022) propose a local attention- and dual attention-based multimodal 3D Convolutional Network. The I3D model base extracts features of RGB data and uses the I3D model with dual spatio-temporal attention to extract depth data features. The extracted features are multiplied element-wise as the final classification result. Bini et al. (2022) concentrate on gesture recognition in real-time and over embedded platforms with limited resources. They suggest a four-stream network of 2D CNNs and 3D CNNs (MobileNet and ResNeXt as backbone) using RGB, depth, optical flow, and MHI. Finally, the fusion of streams are performed at the decision level. Duan et al. (2016) suggest a four-stream network for continuous gesture recognition. This approach uses a two-stream convolutional consensus voting network (2SCVN) to model short and

long-term video sequences. Furthermore, a two-stream 3D depth-saliency ConvNet (3DDSN) is used to learn subtle motions and remove background clutter. 3D CNNs suffer from computational inefficiency since they comprise numerous parameters. Zhang et al. (2018a) proposed a series of 3D lightweight structures for action recognition based on RGB-D data to resolve this deficiency. The proposed lightweight 3D CNNs have significantly fewer parameters with lower computation costs, and it results in desirable recognition performance compared to conventional 3D CNNs (Zhang et al. 2018a).

A branch of studies (Chai et al. 2016; Zhu et al. 2017; Zhang et al. 2017a) employs RNNs and LSTM in their frameworks. In Chai et al. (2016), two streams recurrent neural network (2S-RNN) is presented using RGB-D data which model the contextual information of temporal sequences. Frameworks in (Zhu et al. 2017; Zhang et al. 2017a; Elboushaki et al. 2020) extract spatio-temporal features which are more robust to background clutter. In Zhu et al. (2017), a two-stream network based on 3D CNN, convolutional LSTM (ConvLSTM), spatial pyramid pooling, and a FC layer is proposed to enhance better long-term spatio-temporal learning. 3D CNNs extract spatio-temporal features of RGB-D data. In Zhang et al. (2017a), a three-stream network based on 3D CNN, ConvLSTM, 2D CNN, temporal pooling, and a FC layer with softmax is used to extract spatio-temporal features of RGB, depth, and optical flow. In Elboushaki et al. (2020), a deep-based framework called MultiD-CNN is proposed for gesture recognition which learns spatio-temporal features from RGB-D videos. This method incorporates spatial and temporal information through two different recognition models: a 3D color-depth convolutional network (3D-CDCN) and a 2D motion representation convolutional network (2D- MRCN). 3D-CDCN adds the temporal dimension and makes use of 3D ResNets and ConvLSTM to learn spatio-temporal features simultaneously. On the other hand, 2D- MRCN accumulates the motion across the video sequences into a motion representation and uses 2D ResNets to learn high-level gesture representation. Obaid et al. (2020) utilize CNNs and RNNs in hand gesture recognition to extract temporal features. Dhiman et al. (2020) suggest motion and shape temporal dynamics (STD) as action cues. They propose a framework with RGB dynamic images in motion stream and depth silhouette in STD stream for recognizing action from an unknown view.

Various research (Shahroudy et al. 2017; Qin et al. 2018; Tang et al. 2018; Qin et al. 2019) concentrate on extracting common-specific RGB-D features. The combination of the shared and specific components in input features can be complex and highly nonlinear (Shahroudy et al. 2017). To disentangle specific features from common ones, Shahroudy et al. (2017) stack layers of nonlinear autoencoder-based component factorization to form a deep shared-specific analysis network. Even though RGB and depth images are inherently different in appearance, there is a certain high-level consistency between them (Qin et al. 2018). Qin et al. (2018) propose a novel two-stream model to extract common-specific features via the constraint of similarity at the high level. In (Tang et al. 2018), a method based on multi-stream deep neural networks is proposed for egocentric action recognition. This method exploits the complementary aspects of RGB and depth by learning the non-linear structure of heterogeneous representations. It tries to preserve the distinctive property for each modality and simultaneously explore their sharable information in a unified architecture. In addition, it deploys a Cauchy estimator (Mizera and Müller 2002) to maximize the correlations of the sharable components and impose the orthogonality constraints on the individual components to guarantee their high independencies. Qin et al. (2019) employ a novel end-to-end trainable framework called TSN-3DCSF (two-stream network with 3D common-specific features) that uses 3D CNNs to extract common-specific features.

In Ren et al. (2021), Wang et al. (2018c), Ren et al. (2021b), Wang et al. (2020) segmented or bidirectional rank pooling is proposed in the frameworks. Dynamic images are created from the RGB-D sequence as inputs to the ConvNets to extract spatio-temporal information (Ren et al. 2021). Then, a segmented cooperative ConvNet is utilized to learn the complementary features of RGB-D modalities. Wang et al. (2018c) use two separate cooperative convolutional networks (c-ConvNet), which extract information from dynamic images constructed from both visual RGB (VDIs) and depth (DDIs) modalities. By applying bidirectional rank pooling, VDIs, and DDIs are represented by two dynamic images, i.e., forward (f) and backward (b), namely VDIf & VDIb and DDIf & DDIb, respectively. The c-ConvNet consists of one feature extraction network and two branches, one for ranking loss and another for softmax loss. Ren et al. (2021b) employ the segmented bidirectional rank pooling to acquire spatio-temporal information, as shown in Fig. 10. Moreover, the multimodality hierarchical fusion scheme gets the most out of the complementarity of different modalities. The multimodality hierarchical scheme includes VDIs-f, VDIs-b, DDIs-f, DDIs-b, and optical flow fields (flow X and flow Y) trained on ConvNets. In Wang et al. (2020), the network is built upon weighted dynamic images, bidirectional rank pooling, CNNs, and 3D ConvLSTM to extract complementary information from the depth and RGB video sequences. Canonical correlation analysis is adopted for feature-level fusion, and a linear SVM is used for classification.

Several methods use RGB-predicted pose in action recognition (Al-Faris et al. 2020; Wu et al. 2021). In Al-Faris et al. (2020), a framework is proposed for hierarchical region-adaptive multi-time resolution DMM (RAMDMM) and multi-time resolution RGB action recognition system. The proposed method introduces a feature representation technique for RGB-D data that enables multi-view and multi-temporal action recognition. Original and synthesized viewpoints employed for multi-view HAR. To be invariant to variations in an action's speed, it also used temporal motion information by integrating it into the depth sequences. Appearance information in terms of multi-temporal RGB data is used to help focus on the underlying appearance information (that would otherwise be lost with depth data alone). It helps to provide sensitivity to interactions with small objects. Wu et al. (2021) utilize 3D CNNs with multimodal inputs to enhance spatio-temporal features. This approach proposes two different video presentations: depth residual dynamic image sequence

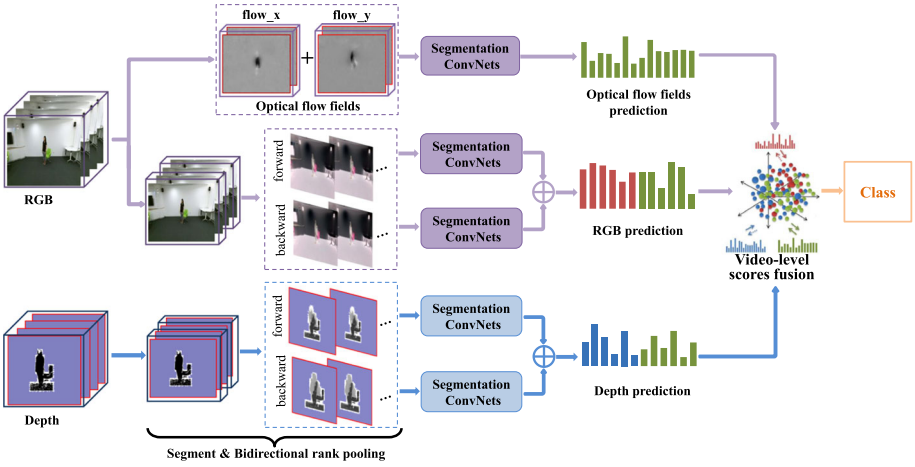


Fig. 10 An example of independent streams (Ren et al. 2021b)

(DRDIS), which reflects spatial motion changes of an action over time, and pose estimation map sequence (PEMS), which is constructed by pose (skeleton) estimation from an RGB video. DRDIS is robust under lighting conditions, texture, and color changes. PEMS eliminates the background clutter.

6.1.1.2 Dependent streams In some other methods, there are dependent streams, i.e., each stream uses extracted features of other streams and shares its features with others during feature extraction. The following approaches employ dependent streams in their architectures.

In Li et al. (2021), Transformers are used in the inter-frame and modality mutual attention for egocentric action recognition (see Fig. 11). Transformers are more robust in modeling long-term sequences than RNNs or LSTM (Khan et al. 2022). Li et al. (2021) apply position encoding to frames in order to emphasize the frame orders. A two-stream network with Transformer encoders is applied to inputs. Then, extracted features are passed to the mutual-attentional fusion block to exchange cross-modality information. Zhao et al. (2021) use a capsule network, Kalman filter, and Transformers to resolve CNN's over-sensitivity to rotation and scaling. Body parts are extracted using a capsule network, and their attribute is determined via a Kalman filter.

In Li et al. (2019), a spatio-temporal attention mechanism is proposed to select the most representative regions and frames in a video. Different features (RGB, flow, and depth) are extracted by the ResC3D network (Miao et al. 2017) and fuse with canonical correlation analysis. The final recognition results are derived with a linear SVM classifier.

Li et al. (2023) propose a hierarchical gesture prototype framework to handle two problems in gesture recognition, including redundancy in the gesture-relevant features of different modalities and exploiting the complementarity of modalities. The framework highlights gesture-relevant features such as poses and motions using a sample-level prototype and a modal-level prototype. The sample-level gesture prototype uses a memory bank to extract the essential features of a specific gesture class with different phenotypes. Then,

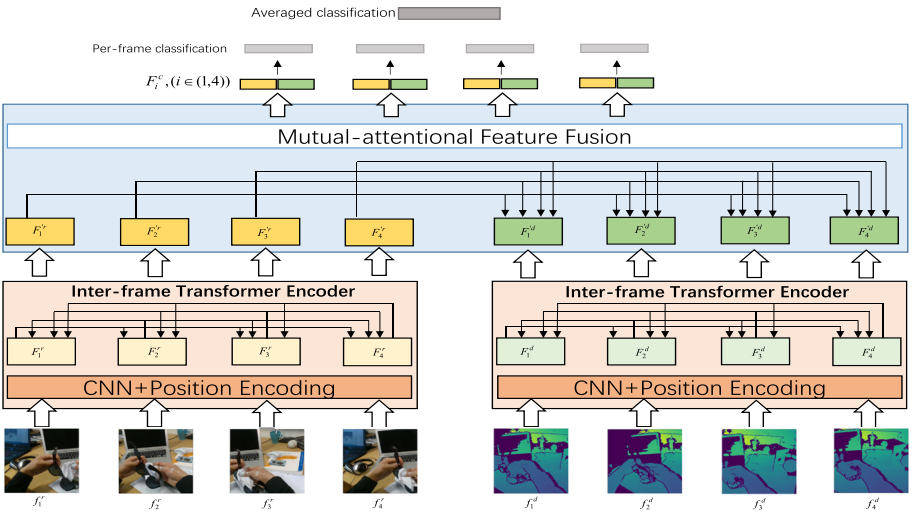


Fig. 11 An example of dependent streams using Transformers (Li et al. 2021)

the modal-level prototype is obtained via a GAN-based subnetwork, in which the modal-invariant features are extracted and pulled together.

New fusion approaches are proposed in a group of studies (Cheng et al. 2021; Zhou et al. 2021; Tian et al. 2020; Wang et al. 2019a; Hampiholi et al. 2023; Lee et al. 2023; Cheng et al. 2022). In Cheng et al. (2021), a cross-modality compensation block (CMCB) is developed to learn the cross-modality complementary features from RGB and depth modalities. The CMCB first gathers features from the two isolated information flows, then sends and intensifies them to the RGB-D paths using the convolution layers. CMCB is incorporated into two typical network architectures of ResNet and VGG to improve action recognition performance. In Zhou et al. (2021), Adaptive Cross-modal Weighting (ACmW) approach is employed to extract complementary features from RGB-D data. ACmW scheme intuitively evaluates the relationship between the complementary features from different streams and fuses them in the spatial and temporal dimensions. In Zhou et al. (2021), a pair of CNNs is used to exploit the features of RGB and depth images then the extracted features at different levels are evaluated. Wang et al. (2019a) address multi-view and missing view problems in action recognition. The adversarial generative network deploys to generate one view conditioning on the other view, fully exploring the latent connections in intra-view and cross-view. Hampiholi et al. (2023) introduce the Convolutional Transformer Fusion Blocks (CTFBs) in multimodal gesture recognition using RGB and depth modalities. A CTFB consists of a Convolutional Self-Attention (ECSA) mechanism, a fusion operation, and an MLP module. 3D convolution layers are used in ECSAs to capture local key spatio-temporal features from each modality. Then, output feature maps of each modality from two ECSA modules are fused using elementwise addition operation. An MLP is used in the final classification. In Lee et al. (2023), multimodal data are fused with recurrent units. Authors have proposed the Modality Mixer (MMixer) network containing a key component recurrent unit called Multimodal Contextualization Unit (MCU). MCU extracts complementary information across modalities and temporal information of the action. Cheng et al. (2022) suggest a multimodal interactive network (MMINet) for RGB-D-based action recognition using two proposed modules in the two-stream CNNs. The first module is called the spatial-temporal information aggregation module (STIAM) to extract richer spatial-temporal features with limited extra memory and computational cost. A cross-modality interactive module (CMIM) is the second one proposed to fuse the multimodal complementary information fully. The final recognition is based on the score fusion of two stream outputs.

6.1.1.3 Single-stream Another architecture contains a single stream in which different modalities are considered an N-dimensional (N-D) entity, or one modality is used as an attention guide for other modalities.

In Adhikari et al. (2017), Pigou et al. (2018), Wang et al. (2017a), Zhou et al. (2021a), RGB and depth are combined as a four-channel data or a 4D entity. Adhikari et al. (2017) use CNNs via RGB-D input. This approach uses human postures to detect fall events or non-fall events. Pigou et al. (2018) propose a novel end-to-end trainable network using temporal convolutions and bidirectional recurrence. In this approach, RNNs respond to high-level spatial features and do not need to consider the temporal aspect in the lower layers of the network. In addition, RNNs estimate the beginning and end frames of gestures. In Wang et al. (2017a), extracting scene flow from RGB-D videos for action recognition is considered. Scene flow to action map (SFAM) is presented to summarize RGB-D videos. In Zhou et al. (2021a), a regional attention with architecture-rebuilt 3D network

(RAAR3DNet) is proposed for gesture recognition. Fixed inception modules are replaced with the automatically rebuilt structure via neural architecture search (NAS) to get the different representations of features in the early, middle, and late stages of the network. In addition, a stackable regional attention module called dynamic-static attention is designed to emphasize the hand/arm regions and the motion information.

Another class of methods uses an auxiliary modality as an attention for other modalities. In Jegham et al. (2020a), a depth-based spatial attention network is suggested, which focuses on the driver's silhouette and motion in a scene. Each time, a new weighted RGB frame is fed to the network with the relevant depth frame as attention. Soft spatial attention enhanced the CNN's recognition by selectively highlighting relevant frame regions.

6.1.2 Missing modality

Useful information from multimodal data in the multimodal action recognition can be performed to include a variety of sensory modalities. However, it is often the case that not all of the modalities are accessible in real-life scenarios due to restrictions such as noise or missing modalities (Fig. 12). The main challenge in missing modalities is training the model in a form that can be utilized at the test time.

6.1.2.1 Hallucination Network Some studies utilize distillation knowledge or transfer learning to possess complementary information of all modalities in the missing scenario. In these approaches, hallucination networks are considered in the context of learning with privileged information to address the challenge of considering a missing (or noisy) modality at test time.

The teacher-student framework is introduced to distill knowledge. It deals with missing or noisy data samples (Rahate et al. 2022). The student model is typically faster than the teacher model (Rahate et al. 2022). In Garcia et al. (2018, 2019), the teacher-student framework is used for missed modality in HAR. In Garcia et al. (2018), a hallucination (student) network is trained to simulate the depth stream. The authors similarly use RGB and depth frames as inputs for training, but only RGB at the test time. In this approach, a technique based on inter-stream connection is implemented to enhance the learning process of the hallucination network, as shown in Fig. 13. Also, a general loss function is designed in (Lopez-Paz et al. 2015) that unifies distillation and privileged information learning theories. The proposed scheme in (Garcia et al. 2018) is revisited in (Garcia et al. 2019) where the hallucination network is trained via discriminative adversarial learning. This method does not need to balance the different losses used in other methods (Garcia et al. 2018, 2019).

Fig. 12 Missing modality at test time [images are from NTU RGB+D Dataset (Shahroudy et al. 2016)]

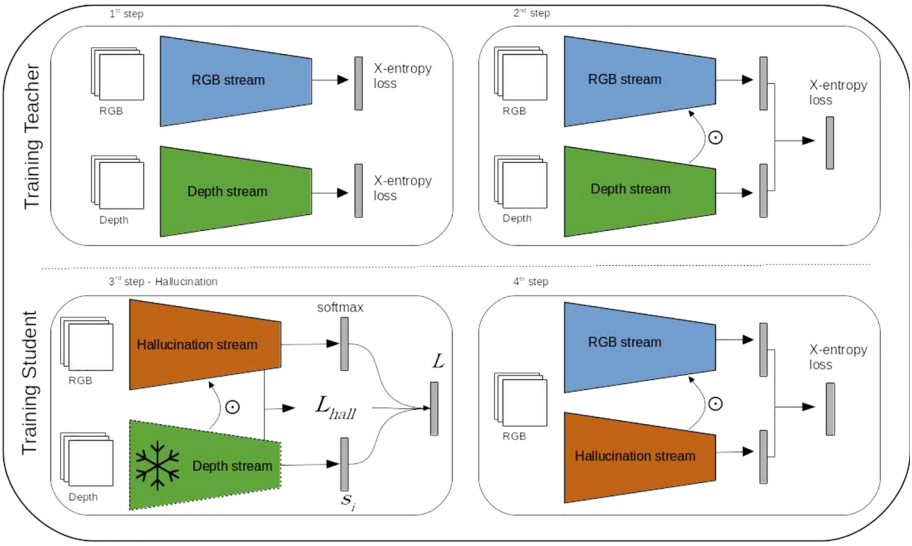


Fig. 13 Hallucination Network proposed in (Garcia et al. 2018)

Park et al. (2023) have proposed Cross-Modal Alignment and Translation (CMAT) in action recognition. The framework first aligns representations of multiple modalities from the same video sample through contrastive learning by R(2+1)D-18 architecture (Tran et al. 2018). Then, CMAT learns to translate representations of one modality into that of another modality using CNNs. This allows the representations of the missing modalities to be generated from the remaining modalities during the testing.

6.1.2.2 Ensemble methods An ensemble method is proposed in (Garcia et al. 2021), where the complementary information of multiple modalities is leveraged to the benefit of the ensemble and each network (without the need for a pre-training step or another network at test time, as shown in Fig. 14). The introduced distillation multiple choice-learning framework is trained from scratch, and modalities are strengthened.

6.2 RGB and skeleton

Skeletal data, used as high-level information, is robust against different views, backgrounds, and motion speeds (Shabaninia et al. 2019). However, sparse information of 3D joints in skeletal data is insufficient to model human actions, especially human-object interactions

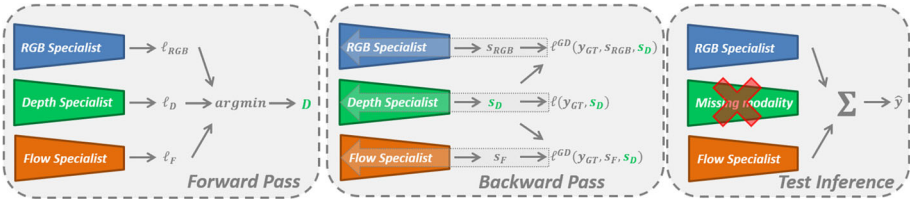


Fig. 14 Ensemble methods proposed in (Garcia et al. 2021)

fully. Some papers suggest considering skeletal information with RGB data to gain complementary features of both modalities in an accurate recognition framework.

6.2.1 Complete modality

6.2.1.1 Independent streams Some studies (Jang et al. 2020; Das et al. 2019c; Tomas and Biswas 2017) employ CNNs in their architectures. Jang et al. (2020) propose a four-stream adaptive CNN (FSA-CNN) framework, robust to spatio-temporal variations. The activation function adapts without using multiple activation layers. Streams consist of raw data, short-term temporal differential, long-term temporal differential, and spatial differential sequences of actions. Using 2D skeletal data (created from RGB sequences) and 3D skeletons (captured by Kinect sensors) improves the accuracy of action recognition. Although 2D and 3D skeletons appear similar, their sources differ, and extra information is obtained in the 3D case. Das et al. (2019c) propose a framework with an action-pair memory module to disambiguate similar actions. Moreover, a two-level fusion mechanism employs various information from three modalities, including RGB, 3D skeletons, and 2D skeletons & RGB. Tomas et al. (2017) employ appearance and motion information from RGB and skeletal joints, respectively, to capture subtle motions. Motion representations are learned via CNN from MHIs created from RGB images. Besides, stacked autoencoders (SAE) attain discriminative movements of human skeletal joints by taking the distance of joints from the mean joint at each frame.

A variety of methods exploits spatio-temporal features using both CNNs and RNNs. Debnath et al. (2021) offer a two-stream attention-based framework that learns the 3D position of body joint relationships during action sequences. 3D poses are inputs of two streams; one learns spatial features, and the other learns temporal features. Then, a multi-head attention mechanism fuses pose streams. Along with the 3D pose streams, an RGB stream extracts appearance information using an Inception-ResNet-V2 pre-trained model, multi-head attention block, and bidirectional LSTM. These two streams are fused by the concatenation method and global average pooling. An FC layer for final classification. Verma et al. (2020) propose a two-stream framework using MHI and motion energy image (MEI) as RGB descriptors. Skeleton modality is used after developing intensity images in three views: top, side, and front. Feature-level fusion is applied in each stream. Afterward, scores of two-stream are fused via the weighted product rule. The multimodal network is trained only once using the cyclic learning rate concept. Liu et al. (2018) offer a multi-modality multi-task RNN for online action detection. The framework contains classification and regression subnetworks for temporal modeling. After extracting dynamic features, classification, and regression, subnetworks share identical structures with different weights among different modalities. Zhao et al. (2017) suggest 3D CNNs for processing RGB videos and RNNs to extract features from 3D skeleton data with SVM as a classifier.

Cai et al. (2021) utilize a two-stream GCN using a human pose skeleton and joint-centered light-weight information, namely, JOLO-GCN, as inputs. Each joint local motion is captured as the pivotal joint-centered visual information via joint-aligned optical flow patches (JFP). The proposed scheme is accurate while keeping low computational and memory overheads. Duan et al. (2022) propose PoseC3D using a 3D heatmap stack for skeletal data representation instead of a graph sequence. The suggested framework can handle multiple-person scenarios in HAR.

6.2.1.2 Dependent streams Several methods employ skeletal data as an attention guide for the RGB stream while training them separately. However, high computational cost limits the utilization of these two modalities. A novel pose-driven attention mechanism on 3D ConvNets is suggested in (Das et al. 2019b) to point out activities of daily living (ADL) recognition challenges. Time-series representation of pose dynamics extracts human activities' spatial and temporal saliency. Song et al. (2018) offer an end-to-end trainable three-stream skeletal attention-based framework from RGB and optical flow videos. The framework is based on a ConvNet with LSTM. Visual features around critical joints are extracted automatically using a skeleton-indexed transform layer, and via a part-aggregated pooling, the visual features from different body parts and actors are uniformly regulated. Baradel et al. (2018; 2017b; a) propose a two-stream LSTM framework from articulated pose and RGB. A specific joint ordering is processed with the pose stream. The RGB stream, which gives essential cues on hand motion and objects, is handled by a spatio-temporal soft-attention mechanism conditioned on features from the pose network. Lastly, a temporal attention scheme learns to fuse features over time. Liu et al. (2019b) propose an attention-based two-stream framework that uses a 3D skeleton sequence and just a single middle frame from an RGB video as network inputs. Spatial features are extracted from the RGB stream using self-attention and skeleton-attention modules. In parallel, temporal features are obtained from the skeleton sequence by a BI-LSTM network. In the case of processing a single image instead of a video, this network contains lightweight architecture with less computational cost. Li et al. (2020) focus on action problems like "throw up hat," whose related objects are extremely far from actors. The framework is pre-trained on R (2+1)D (Tran et al. 2018) using the Kinetics dataset (Kay et al. 2017). RGB and skeleton data fuse at the feature level via a skeleton-guided multimodal network (SGM-Net) in the proposed framework. Skeleton features guide the attention on the object related to the action and reduce the interference of unnecessary background information, e.g., for the action "throw up hat," the object information of arms and hat are enhanced. In the guided block, two schemes of correlation operation are explored, including feature learning correlation (FLC) and compact bilinear correlation (CBC). Weiyao et al. (2021) propose bilinear pooling and attention network (BPAN) to fuse multimodal data to get the deep semantic relationship between multimodal features. Two-stream adaptive graph convolution network (2S-AGCN) and R(2+1)D are used in feature extraction from RGB and skeleton, respectively. Bruce et al. (2022) have proposed a model-based multimodal network (MMNet) for HAR in RGB-D videos using a model-based multimodal data fusion mechanism. This method borrows the extracted attention feature from the skeleton modality using GCNs and contributes to the CNN-based stream of RGB modality's performance to improve ultimate performance.

Joze et al. (2020) offer a multimodal transfer module (MMTM) to fuse knowledge from multiple modalities in CNN. MMTM uses squeeze operations to make a global feature descriptor. Both tensors map into a joint representation using concatenation and a fully connected layer. Excitation signals are produced based on the joint representation to gate the channel-wise features in each modality. This module can be placed at different positions of networks and fused modality features in convolution layers with different spatial dimensions. It is also used in unimodal stream networks with minimum changes in their architectures to initialize the network with pre-trained weights.

In Ahn et al. (2023), spatio-temporal cross attention transformer (STAR++) is suggested using deformable 3D (D3D) token selection and interval attention (IAttn) modules, which

creates tokens with spatio-temporal cross-attention. They employ 3D CNNs and Transformers in the proposed architecture.

6.2.1.3 Single-stream Various papers introduce skeleton as attention for RGB stream; however, skeletal data is not trained separately. In Mahasseni and Todorovic (2016), using 3D skeleton sequences as attention, large-scale video action recognition improves by providing an auxiliary modality in training data to complete poorly or missing features of human actions. The framework consists of LSTM and deep CNN for recognition. LSTM is regularized based on the output of another LSTM and 3D human-skeletal data. For regularization, the standard backpropagation through time (BPTT) is rectified to address problems with gradient descent in constraint optimization. Das et al. (2019a) offer a 3D convolution network with soft RNN attention. Articulated poses specify the best body part for modeling the action class. The framework consists of three branches to extract features from human body parts: left hand, right hand, and entire body. RNN attention subnetwork allocates different levels of importance to the body parts. Das et al. (2020; 2021) propose pose-driven attention strategies called video-pose network (VPN and VPN++) for recognition of ADL with the ability to distinguish between similar activities with fine-grained details. VPN requires both RGB and 3D poses to classify actions. In contrast, VPN++ requires only RGB images to predict action labels. Further, VPN++ provides high speed and high resiliency to noisy poses. RGB video is fed to the network with the corresponding 3D poses at training time, obtained from Kinect sensors or images using pose estimation methods. Features of inputs are extracted via two distinct videos and pose backbones. The video backbone consists of 3D CNNs to extract spatio-temporal features, and the pose backbone contains a spatio-temporal GCN.

Do et al. (2022) propose a Multimodal Transformer (MMT) to use RGB and skeleton data of only eight input frames. Using the transformer-based structure, MMT can capture the correlation between non-local joints in skeleton data modality. The output of the local patch encoder is concatenated to linear projection of skeletons. The Transformer is trained with the global patch encoder, and the final classification is performed by MLPs.

6.2.2 Missing modality

6.2.2.1 Hallucination network In Xu et al. (2021), a two-stage teacher-student framework is suggested for multi-view and missing modality action recognition. The teacher network leverages multi-view pose and appearance information during training, while the student network uses only RGB sequences at test time. 3D CNN is used for teacher and student frameworks. A cross-modality aggregated transfer (CAT) network transfers multi-view cross-modality aggregated features from the teacher network to the student network. Further, a viewpoint-aware attention (VAA) module taking discriminative information across different views is designed to fuse multi-view features. Then, a multi-view feature strengthening (MFS) network and the VAA module boost the global view-invariance features of the student network. Both CAT and MFS are also trained in an online distillation procedure via jointly training the teacher and the student network. Another teacher-student framework is employed in (Thoker and Gall 2019) for cross-modal action recognition, which nearly achieves the accuracy of a student network trained with full supervision. STGCNs and CNNs are employed as student and teacher architectures, respectively. The student network is trained on sequences of 3D human poses based on a teacher network trained on RGB sequences as supervision. Both RGB videos and human pose sequences

train the student network. The student network employs unlabeled data that does not participate in the training of the teacher network. The knowledge of the trained teacher network for the source modality is transferred to a small ensemble of student networks for the target modality.

6.2.2.2 Ensemble methods Song et al. (2020) suggest a modality compensation network (MCN) to leverage complementary information. The framework consists of CNN and LSTM. RGB and optical flow are source modalities, and skeletal data is the auxiliary modality. The main goal is to compensate for source modality features via auxiliary ones implicitly. A modality adaptation block links source modality to auxiliary modality to compensate for the loss of skeletal data at test time and even at the training.

6.3 Depth and skeleton

Since depth maps are invariant to intra-class variations (such as the appearance of objects), foreground/background segmentation obtains appropriate results using depth data (Camplani and Salgado 2014). HAR using depth maps is poor in noisy data and human-object interactions. Although the computation of skeletal data demands less hardware complexity, the skeleton alone is insufficient to distinguish actions that involve human-object interactions. The fusion of depth and skeleton provides partly discriminative features.

6.3.1 Complete modality

6.3.1.1 Independent streams A class of methods (Rani et al. 2021; Wang et al. 2017b; Rahmani and Bennamoun 2017; De Smedt et al. 2017) use CNNs in their frameworks. Rain et al. (2021) propose three descriptors, including difference depth MHI (D2MHI) descriptor, spherical joint descriptor (SJD), and kinematic joint descriptor (KJD). Difference depth motion map (D2MM) and modified MHI (M2HI) are fused early to form the D2MHI descriptor. In order to be less sensitive to the joint movements, SJD is presented to make the model more robust for actions with similar movements. The motivation of KJD is to model the spatial and temporal changes in actions. Descriptors are fed to three CNNs, and obtained scores from the softmax layer are late fused to get the final action label. Wang et al. (2017b) apply the bidirectional rank pooling method to three hierarchical spatial levels of depth maps guided by skeletons, i.e., body, part, and joint. Each level contains some components and possesses a specific number of joint locations. A spatially structured dynamic depth image (S2DDI) preserves the coordination and synchronization of body parts during the action. It is suggested to learn spatio-temporal and structural information at all levels. This framework carried on with three weights-shared ConvNets and score fusion for classification. Rahmani et al. (2017) use a CNN-based framework to model human-object interactions and intra-class variations under different viewpoints. First, the relative geometry between every body part and others is assessed to transfer the depth map of body parts to a shared view-invariant space. Afterward, the view-invariant body parts of depth and skeletal modalities are combined with learning body part movements during actions. Then, the FC, temporal pooling, and softmax layers recognize the action class. Smedt et al. (2017) use CNNs feeding with keyframes of depth for action recognition.

Lai et al. (2018) suggest a combination of CNNs and RNNs via depth and skeleton for hand gesture recognition. Various fusion techniques are analyzed for improving performance, including feature-level fusion and score-level fusion. In Mahmud et al. (2023)

quantized depth images are employed as an alternative input modality to raw depth images to create sharp relative contrasts between key parts of the hand. The architecture comprises multimodal-fusion CRNNs (Convolutional Recurrent Neural Networks).

Others (Liu et al. 2016; Zhao et al. 2019) employ 3D CNNs in their frameworks. Liu et al (2016) proposed a framework including a 3D-based deep CNN (3D2CNN) to learn the depth and skeleton features (called JointVector) along with fusion of SVM decisions for classification. In Zhao et al. (2019), a fusion-based action recognition framework is proposed, consisting of three parts: 3D CNN, human skeleton manifold representation, and classifier fusion.

6.3.1.2 Dependent streams Authors in (Mahmud et al. 2021; Kamel et al. 2018) use CNNs and score fusion techniques in the frameworks. Mahmud et al. (2021) suggest dynamic hand gesture recognition using depth quantized images and skeleton joints. The fusion of CNN and LSTM is used in this framework to extract depth features, while skeleton features were extracted via LSTM following distinct MLPs. Besides, depth and skeleton data are concatenated and fed to another MLP. Scores of MLPs are fused in prediction. Kamel et al. (2018) focus on a deep CNN framework that is fed with three descriptors, called depth motion image (DMI), moving joint descriptor (MJD), and fusion of DMI with MJD. DMI represents the body changes of depth maps in an image, and MJD shows body joint position changes and directions around a fixed point.

6.3.2 Missing modality

6.3.2.1 Ensemble methods Shi et al. (2017) suggest a privileged information-based RNN (PRNN). The privileged information (PI) is only available throughout training but not during test time. This model considers skeleton joints as PI in three-phase training processes, including pre-training, learning, and refining. The suggested network is end-to-end trainable and learns CNN and RNN parameters cooperatively. The final network enhances latent PI iteratively in an EM process.

6.4 RGB and depth and skeleton

Some research uses all three modalities of RGB & Depth & Skeleton. To compare 18 baseline/state-of-the-art frameworks using RGB-D and posed data, Garcia-Hernando et al. (2018) analyze approaches in the egocentric hand actions interacting with 3D objects. Different aspects, such as appearance, pose, and both are assessed. The results demonstrate the impact of the hand pose as a guide in action recognition with RGB-D data.

6.4.1 Complete modality

6.4.1.1 Independent streams Khaire et al. (2018) use MHI and depth dynamic motion in the top, side, and front view (3-DDM) as RGB and depth descriptors. Their framework offers a new strategy to construct a skeleton image from skeleton joint coordinates. Afterward, five CNNs are trained with constructed descriptors separately. Scores of streams are fused by applying a weighted product model to predict the final action class.

A class of studies (Singh and Vishwakarma 2021; Cardenas and Chavez 2020; Khaire et al. 2018a; Elmadany et al. 2018; Cardenas and Chavez 2018) employ CNNs and SVM in their framework. Singh et al. (2021) present a modality fusion mechanism of RGB, depth,

and skeleton called deep bottleneck multimodal feature fusion (D-BMFF) framework. 3D joints are converted into a single RGB skeleton MHI (RGB-SklMHI). Every ten RGB and depth frames are fed to a DNN to extract spatial features, and a single SklMHI image for each activity demonstrates temporal features. Extracted features of different modalities from three distinct streams are fused via multiset discriminant correlation analysis (M-DCA). Then, action is recognized using a linear multiclass SVM. Cardenas et al. (2020) propose a dynamic hand gesture recognition framework that fuses spatio-temporal features obtained from RGB-D and skeleton joints. Hand poses are detected using skeletons in RGB and depth images. Besides, a descriptor called histogram of cumulative magnitudes (HCM) represents the topology of hand and body to discern similar poses used to extract features from depth. Two distinct CNN streams are trained from RGB and depth modalities. Obtained features from CNNs and HCM are integrated and passed to an SVM for classification. In addition, a method is presented to extract a constant number of keyframes to decrease the computational processing time. Khaire et al. (2018a) improve activity recognition based on a five-stream CNN network. Skeleton images, MHI, and three DMMs from the side, top, and front are fed to the network. Three approaches are presented for fusion to improve the overall accuracy of recognition. The fusion of trained CNN on skeleton images as the fifth CNN stream achieves the best result compared to the other two approaches. Elmadany et al. (2018) suggest two fusion methods, called biset globality locality preserving canonical correlation analysis (BGLPCCA) and multiset globality locality preserving canonical correlation analysis (MGLPCCA) for learning common subspace from two sets and more than two sets, respectively. These methods represent global and local data features with a low-dimensional common subspace. Besides, a bag of angles (BoA) is proposed as a descriptor for skeleton and HPDMM-CNN for depth. Finally, a framework is used for action recognition using proposed fusion methods and descriptors. The multimodal information recorded by a Kinect sensor (RGB-D and skeleton) is also exploited (Cardenas and Chavez 2018). Various rank pooling and skeleton optical spectra methods are tested to generate dynamic images summarizing an action sequence into single flow images. Dynamic images are categorized into five groups: a dynamic color group (DC), a dynamic depth group (DD), and three dynamic skeleton groups (DXY, DYZ, DXZ). Different dynamic images with the main postures for each group are generated to model different action postures. Then, a pre-trained flow-CNN extracting spatio-temporal features is applied with a max-mean aggregation.

Romaissa et al. (2021) suggest a four-step framework for action recognition, including creating dynamic image sets from RGB, depth, and skeleton joints, feature extraction, feature fusion, and classification using LSTM. After constructing three different dynamic image sets, features are extracted from image sets via pre-trained CNN-based models using transfer learning. Canonical correlation analysis fuse extracted features. Eventually, a bidirectional LSTM is trained to recognize action labels.

6.4.1.2 Dependent streams Hu et al. (2018) propose a method to learn modality-temporal mutual information from tensors, called the deep bilinear framework. The bilinear block contains modality pooling and temporal layers, learning the time-varying dynamics and multimodal information. The deep bilinear model is established via accumulating bilinear blocks and other layers to extract video modality-temporal features. Further, a novel descriptor, called modality-temporal cube, characterizing actions from a general schema, is proposed as deep bilinear learning input.

Gan et al. (2023) propose a focal channel knowledge distillation for action recognition to transfer channel semantic correlations and distributions of teacher modalities to the RGB student. The 3D CNN backbone networks extract spatio-temporal features, and an average pooling operation is performed on the teacher features to generate the channel attention map. The channels with large weights are considered as the focal channels. The correlation matrices of these focal channels are measured by inner product, where high relevance represents the homology of channel semantics, and low relevance represents diversity. By minimizing the MSE distance of the focal channel correlation matrices, the student can learn sufficient intrinsic relationships and diversity properties of key semantics. In addition, the teacher's crucial semantic distribution knowledge is transferred to the student by minimizing the weighted sum of KL divergence of channel distribution differences, thus focusing on the salient region of channel features.

Li et al. (2022) suggest a GCN for first-person hand action recognition. They define geometric relationships between two neighboring bones in a hand skeleton as the third-order node feature. Pretrained networks are employed to extract features from RGB images and depth maps. Customized LSTM units fuse RGB-D features. Finally, the skeleton modality and the RGB-D modality are combined by late fusion of classification scores.

6.4.1.3 Single-stream Wu et al. (2016) describe a deep hierarchical dynamic neural network for multimodal gesture recognition. This framework consists of a Gaussian-Bernouilli deep belief network (DBN) to extract dynamic skeletal features and a 3DCNN to represent features from RGB and depth images, as shown in Fig. 15. Further, intermediate and late fusion methods are used to fuse RGB and depth with skeleton data. Emission probability learning of HMM is applied to estimate the gesture class.

6.4.2 Missing modality

6.4.2.1 Ensemble methods Luo et al. (2018) suggest a framework for action detection and classification with only limited training data and partially observed modalities. Their method, called graph distillation, merges rich information from the large-scale multimodal dataset in the source domain and increases learning performance in the target domain with rare training data and modalities, as shown in Fig. 16. A graph distillation layer is suggested to distill knowledge between multiple modalities and attach it to available models. A trained

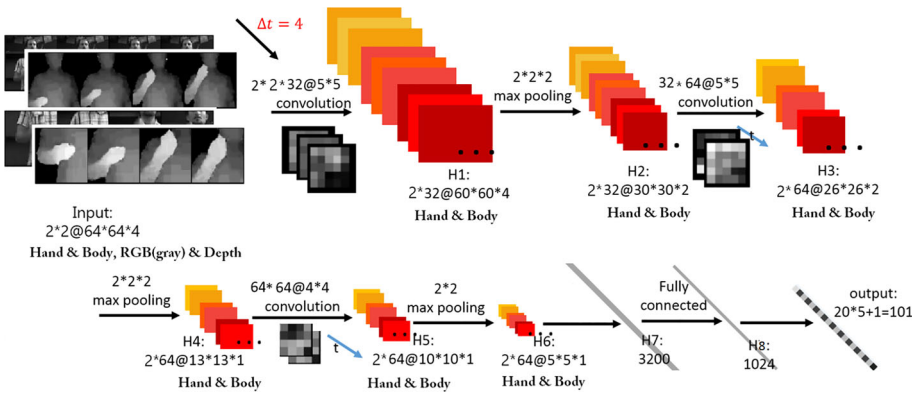


Fig. 15 Single-stream architecture with an N-D entity as input (Wu et al. 2016)

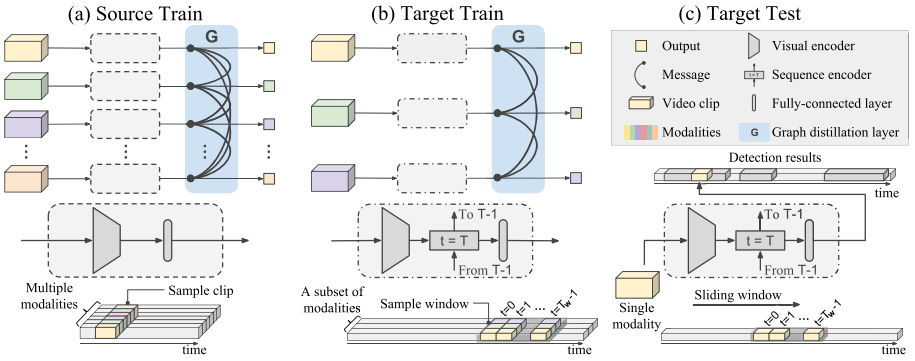


Fig. 16 Graph distillation network in (Luo et al. 2018)

model in the action classification domain is used as a pre-trained model in the action detection task. A short video clip is encoded into a feature vector via a visual encoder for an action recognition task. This step is followed by a sequence to construct the final feature vectors for action detection. The feature vector is fed into task-specific linear and softmax layers to obtain the probability distribution for each clip.

In Li et al. (2023), a deep fusion network (DFN) is proposed to fuse features of different modalities even in missing modality cases. DFN comprised MLP and CNNs.

6.5 Infrared and other visual modalities

Infrared data is another modality used for HAR in a dark environment. It is favored over RGB, as illumination conditions are less affected. However, limited studies concentrated on fusing infrared with other visual modalities. Because the infrared information is more restricted than other modalities and does not comprise complementary information of other visual modalities, this section introduces HAR algorithms using the combination of infrared with other vision-based modalities.

6.5.1 Complete modality

6.5.1.1 Independent streams There are 2D or 3D CNN-based methods for infrared with other modalities. Molchanov et al. (2016) train separate 3D CNNs along with an RNN for RGB, optical flow, depth, IR, and IR disparity modalities. Class conditional probability vectors from all modalities are averaged and fused to detect and classify hand gestures. Boissiere et al. (2020) use infrared and skeletal data in HAR. A pre-trained CNN extracts features from skeleton, which crops the region around the subjects. A pre-trained 3D CNN is designed to extract visual features from infrared videos. Extracted feature vectors are fused and exploited jointly. The main focus in (Rückert et al. 2021) is using RGB, depth, and infrared in HAR to acquire and transfer manual assembly workspaces into a digital environment. A framework based on CNNs and RNNs is suggested to differentiate assembly operations that construct a complex assembly process.

6.5.2 Missing modality

6.5.2.1 Hallucination network The scenario of full-modal learning from partial modalities often arises in practice. For example, RGB surveillance cameras have restrictions according to privacy concerns. In such cases, cross-modal data hallucination is a practical solution (Pahde et al. 2019). Wang et al. (2018a) propose partial-modal generative adversarial networks (PM- GANs) to learn a full-modal model of partial modalities to perform tasks related to data hallucination, as shown in Fig. 17. The complete model is attained via a generated model instead of the missing data channel. In this regard, GANs have shown favorable results for cross-modal sample generation (Pahde et al. 2019). GANs are deep generative models mainly applied for unsupervised tasks that demonstrate significant advances in image generation, image-to-image translation, and facial attribute manipulation (Wang et al. 2021; Pan et al. 2019). Woo et al. (2023) suggested an autoencoder for reconstructing missed modalities. They have used CNNs and Transformers in the proposed architecture.

7 Datasets

With growing attention to deep-based action recognition, there is a need for large datasets describing actions under various conditions, such as different subject appearances (with diverse ages, gender, height, and cultural background), views (first-person or second-person, ego-centric, etc.), illuminations, and environments (Liu et al. 2019). Many different datasets are established in this field for analyzing the efficiency of HAR algorithms. Some datasets contain ADL (Das et al. 2019b; Jang et al. 2020; Liu et al. 2019); others provide data in specific application domains like therapeutic (Negin et al. 2013), sport (Moencks et al. 2019), gaming (Li et al. 2010), (Bloom et al. 2015), human-robot interactions (Jang et al. 2020), and health (Liu et al. 2019). While there are many unimodal datasets in the community, we only concentrate on multimodal visual datasets in this section due to the scope of the paper.

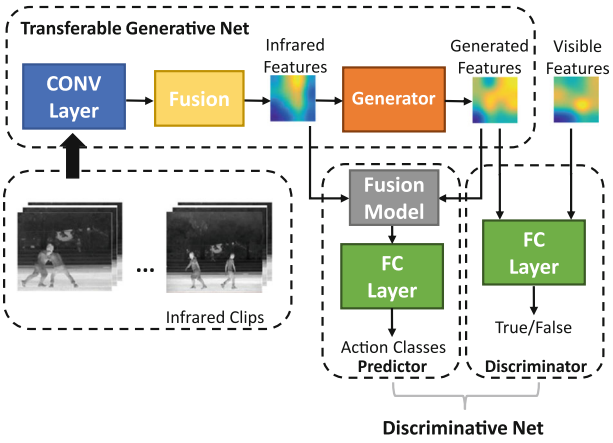


Fig. 17 Full-modal modeling of partial modalities (Wang et al. 2018a)

Multimodal visual datasets offer more than one modality for each action, which is helpful in multimodal HAR algorithms. These datasets usually provide RGB, depth, and skeleton. Nevertheless, IR is also provided in some limited datasets.

We comprehensively review available multimodal vision-based gesture, action, and activity datasets. All these datasets are considered action datasets at the rest of the paper. A three-level categorization is proposed on the multimodal HAR benchmark datasets, as shown in Fig. 18. In the first level, datasets are grouped to provide trimmed (segmented) or untrimmed (continuous) videos. While only one action is performed in a trimmed video, untrimmed ones have more than one action in each video. Next, the number of viewpoints for actions is the criterion categorizing datasets in the second level, including single-view and multi-view. Multi-view datasets (often created by more than one camera) provide different views for action in the scene, while others are single-view (frequently front-view). In multi-view datasets, videos are captured by two different methods. First, several cameras are mounted at different positions and angles, and action is captured synchronously using these cameras. Second, the same action is repeated from different viewpoints with only a single camera. In the last level, datasets are grouped based on the provided data modalities. As described above, multimodal visual HAR datasets usually provide two or three modalities, including RGB and depth (RGB+D), RGB and skeleton (RGB+S), depth and skeleton (D+S), and RGB, depth, and skeleton (RGB+D+S) that are respectively demonstrated with red, pink, green, and blue colors in this paper as shown in Figs. 18, 19, 20, 21, 22. Since few datasets provided infrared data, IR is not considered in the categorizing datasets in the third level.

Therefore, available multimodal vision-based HAR datasets are grouped into four main categories, including trimmed/single-view, trimmed/multi-view, untrimmed/single-view, and untrimmed/multi-view (see Table 3). So datasets are presented accordingly in four diagrams, Figs. 19, 20, 21, 22. Diagrams are provided to compare datasets in each group better. This will help the community choose suitable datasets for their tasks or produce new ones to eliminate available restrictions.

Figures 19, 20, 21, 22 show the publishing year and the average number of videos per class for different datasets. Further, the circle size for each dataset is related to the number of action classes in that dataset, and the color of circles shows the modalities provided in the dataset (identical with Fig. 18). The vertical axis (the average number of videos per class) is plotted on the logarithmic scale to better show differences between different datasets. Therefore, the vertical axis shows the intra-class variations caused by different subjects, views, and environments, whereas the size of circles shows the inter-class variations. Hence, a circle on the upper side of the graph means more intra-class diversity in the dataset,

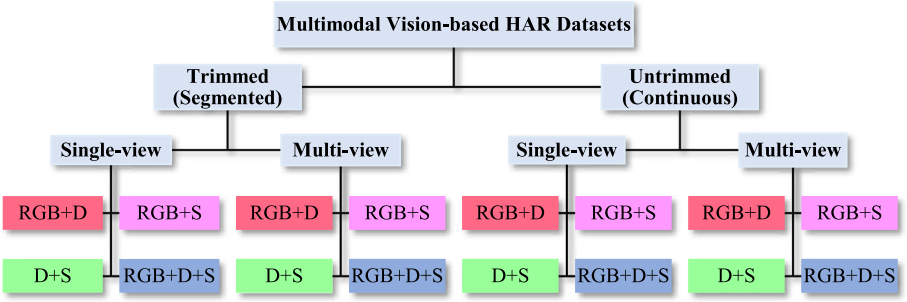


Fig. 18 Proposed taxonomy of multimodal vision-based HAR datasets

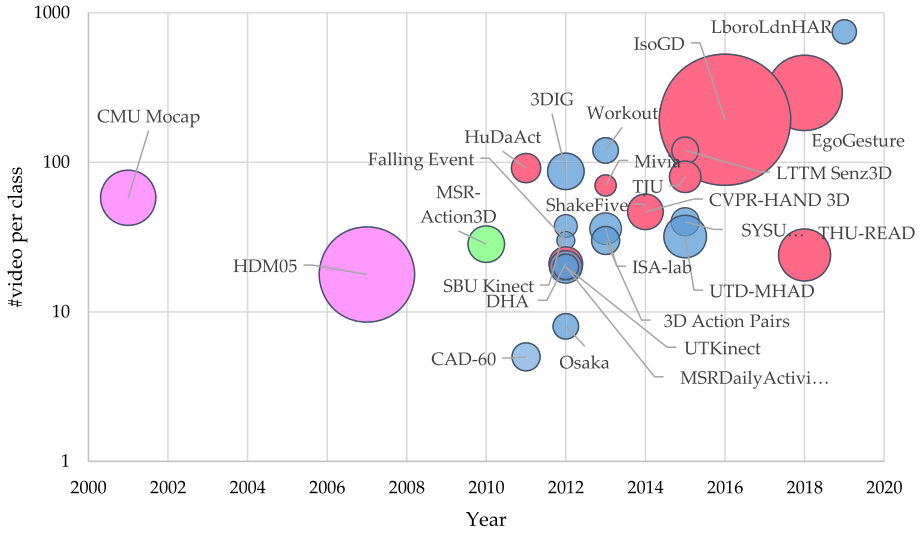


Fig. 19 Trimmed/single-view multimodal vision-based HAR datasets. Circle size is related to the number of action classes compared to other datasets, and circle color demonstrates provided modalities in the dataset; red: RGB+D, pink: RGB+S, green: D+S, and blue: RGB+D+S

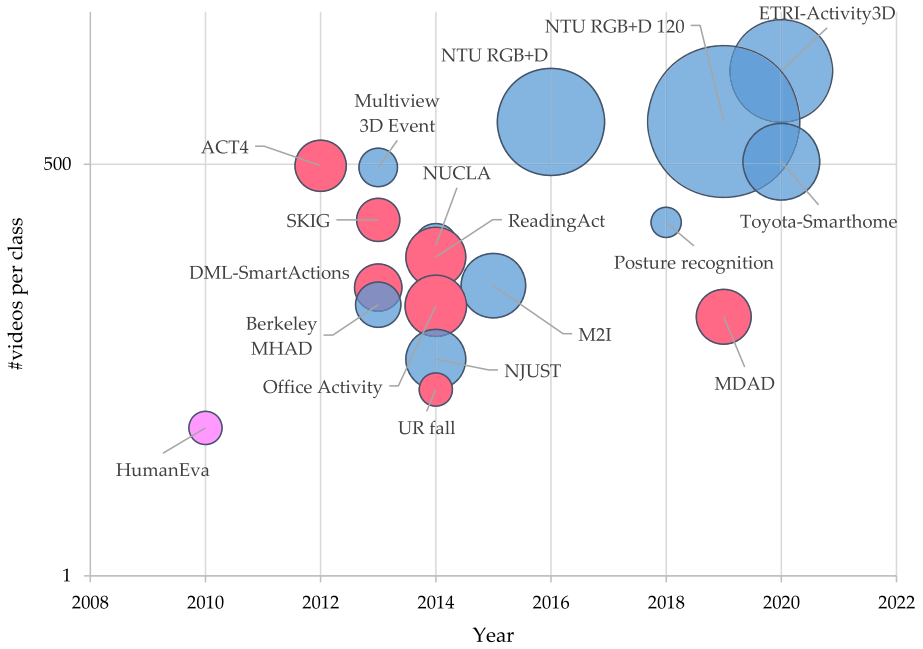


Fig. 20 Trimmed/multi-view multimodal vision-based HAR datasets. Circle size is related to the number of action classes compared to other datasets, and circle color demonstrates provided modalities in the dataset; red: RGB+D, pink: RGB+S, and blue: RGB+D+S

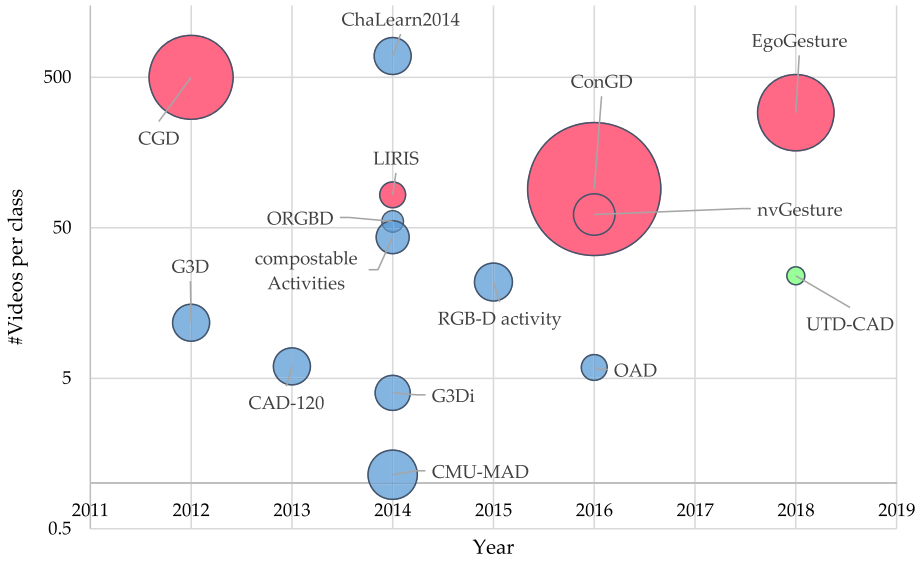


Fig. 21 Untrimmed single-view multimodal vision-based HAR datasets. Circle size is related to the number of action classes compared to other datasets, and circle color demonstrates provided modalities in the dataset; red: RGB+D, green: D+S, and blue: RGB+D+S

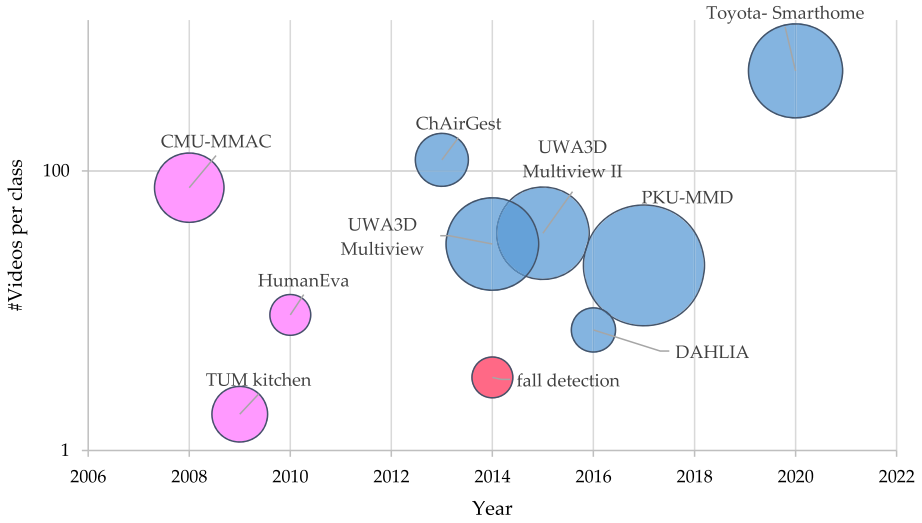


Fig. 22 Untrimmed/multi-view multimodal vision-based HAR datasets. Circle size is related to the number of action classes compared to other datasets, and circle color demonstrates provided modalities in the dataset; red: RGB+D, pink: RGB+S, and blue: RGB+D+S

containing more average number of videos per class. Bigger circles reveal more class activities in the dataset, which causes more inter-class diversity. Therefore, the most appropriate datasets for HAR are bigger ones on the upper side of the diagrams.

To compare benchmark HAR datasets in a meaningful way, the average number of videos per class is considered instead of the number of videos in a dataset. For example,

Table 3 Multimodal vision-based action datasets

Category	Name	
Trimmed/single-view	IsoGD (Wan et al. 2016)	LboroLdnHAR (Moencks et al. 2019)
	THU-READ (Tang et al. 2018)	TJU (Liu et al. 2015)
	SYSU-3DHOI (Hu et al. 2015)	UTD-MHAD (Chen et al. 2015)
	ISA-lab (Munaro et al. 2013b, 2013)	3D Action Pairs (Oreifej and Liu 2013)
	UTKinect (Xia et al. 2012)	MSRDailyActivity3D (Wang et al. 2012)
	Osaka (Mansur et al. 2012)	CAD-60 (Sung et al. 2011)
	DHA (Lin et al. 2012)	SBU Kinect (Yun et al. 2012)
	ShakeFive (Van Gemeren et al. 2014)	Mivia (Carletti et al. 2013)
	Workout (Negin et al. 2013)	3DIG (Sadeghipour et al. 2012)
	HuDaAct (Ni et al. 2011)	Falling Event (Zhang and Tian 2012)
	MSR-Action3D (Li et al. 2010)	HDM05 (Müller et al. 2007)
	CMU-Mocap (Carnegie mellon university 2001)	EgoGesture (Zhang et al. 2018)
	CVPR-HAND 3D (Ohn-Bar and Trivedi 2014)	LTTM Senz3D (Memo and Zanuttigh 2018)
Trimmed/multi-view	ETRI-Activity3D (Jang et al. 2020)	Toyota-Smarthome (Das et al. 2019b)
	NTU RGB+D 120 (Liu et al. 2019)	NTU RGB+D (Shahroudy et al. 2016)
	Posture recognition (Elforaici et al. 2018)	MDAD (Jegham et al. 2019)
	Multiview 3D Event (Wei et al. 2013)	NUCLA (Wang et al. 2014a)
	ReadingAct (Chen et al. 2014)	M^2I (Xu et al. 2015)
	$ACT4^2$ (Cheng et al. 2012)	SKIG (Liu and Shao 2013)
	DML-SmartActions (Amiri et al. 2013)	Berkeley MHAD (Ofli et al. 2013)
	Office Activity (Wang et al. 2014b)	UR fall (Kwolek and Kepski 2014)
	NJUST (Song et al. 2014)	HumanEva (Sigal et al. 2010)
Untrimmed/single-view	UTD-CAD(Dawar and Kehtarnavaz 2018)	OAD (Li et al. 2016a)
	ConGD (Wan et al. 2016)	RGB-D activity (Wu et al. 2015)
	ChaLearn2014 (Escalera et al. 2015)	ChaLearn2014 (Escalera et al. 2015)
	LIRIS (Wolf et al. 2014)	ORGBD (Yu et al. 2015)
	composable Activities (Lillo et al. 2014)	CAD-120 (Koppula et al. 2013)
	G3Di (Bloom et al. 2015)	G3D (Bloom et al. 2012)
	CMU-MAD (Huang et al. 2014)	CGD (Guyon et al. 2013)
	nvGesture (Molchanov et al. 2016)	
Untrimmed/multi-view	Toyota-Smarthome (Dai et al. 2022)	PKU-MMD (Liu et al. 2017a)
	UWA3D Multiview II (Rahmani et al. 2016)	UWA3D Multiview (Rahmani et al. 2014)
	ChAirGest (Ruffieux et al. 2013)	CMU-MMAC (Spriggs et al. 2009)
	HumanEva (Sigal et al. 2010)	TUM kitchen (Tenorth et al. 2009)
	fall detection (Adhikari et al. 2017)	DAHLIA (Vaquette et al. 2017)

ChaLearn2014, which contains 13858 videos, provides more intra-class diversity than ConGD, with 22535 videos. Because ConGD has more action classes, the average number of videos in each class decreased with restricted intra-class variations.

According to Fig. 19, among trimmed/ single-view datasets, LboroLdnHAR (Moencks et al. 2019) provides prominently more videos per class; however, it contains few classes. IsoGD (Wan et al. 2016) provides fewer videos per class compared to LboroLdnHAR (Moencks et al. 2019), and EgoGesture (Zhang et al. 2018) also contains the most number of action classes compared to datasets in this group.

As depicted in Fig. 20, ETRI-Activity3D (Jang et al. 2020) comprises more videos per class than others, and it ranked third based on the number of action classes. Although the intra-class diversity is almost the same in NTU RGB+D 120 (Liu et al. 2019) and NTU RGB+D (Shahroudy et al. 2016), NTU RGB+D 120 provides more action classes among trimmed/ multi-view datasets.

As shown in Fig. 21, ChaLearn2014 (Escalera et al. 2015) has the most videos per class, and ConGD (Wan et al. 2016) provides more action classes in the untrimmed/ single-view group. In untrimmed/ multi-view datasets, Toyota-Smarthome (Dai et al. 2022) ranked first

Table 4 Methods with the best accuracy on common and the newest multimodal vision-based human action and gesture datasets

Dataset	Method	Architecture	Modality	Accuracy	# Papers
NTU RGB+D	Dudan et al. (2022)	Independent	RGB & S	97.0(cs)	36
	Das et al. (2021)	Single	RGB & S	96.6(cs) 99.1(cv)	
IsoGD	Elboushaki et al. (2020)	Independent	RGB & D	72.53	14
UTD-MHAD	Romaissa et al. (2021)	Independent	RGB & D & S	98.88	13
MSRDaily Activity-3D	Das et al. (2019c)	Independent	RGB & S	97.81	11
NTU RGB+D 120	Duan et al. (2022)	Independent	RGB & S	95.3(cs)	10
	Das et al. (2021)	Single	RGB & S	90.7(cs) 92.5(cset)	
SBU Kinect	Elmadany et al. (2018)	Independent	RGB & D	100	10
MSR-Action3D	Wang et al. (2017b)	Independent	D & S	100	9
CAD-60	Singh et al. (2021)	Independent	RGB & D & S	98.66	9
NUCLA	Zhang et al. (2018a)	Independent	RGB & D	95.5(cv)	8
SYSU 3D HOI	Wang et al. (2017b)	Independent	D & S	95.42	7
UWA3D-Multiview-II	Garcia et al. (2021)	Ensemble	RGB & D	89.75	5
SKIG	Elboushaki et al. (2020)	Independent	RGB & D	99.72	4
PKU-MMD	Ren et al. (2021)	Independent	RGB & D	92.1(cs) 93.2(cv)	3
Toyota-Smarthome	Das et al. (2019b)	Independent	RGB & S	75.3(cs) 61(cv1) 68.2 (cv2)	3
MDAD	Jegham et al. (2020a)	Single	RGB & D	69.79(sv) 65.63(fv)	1
ETRI-Activity3D	Jang et al. (2020)	Independent	RGB & S	93.7(cs)	1

D Depth, *S* Skeleton, *CS* Cross Subject, *CV* Cross View, *CSET* Cross Setup, *SV* Side View, *FV* Front View

for the number of videos per class and second for the number of action classes. PKU-MMD (Liu et al. 2017a) provides the most action classes.

The most common datasets in HAR along with the newest ones are listed in Table 4. This table also demonstrates the method with the best accuracy for each dataset, employed modalities, framework architecture, and the number of studies reviewed in this paper that use the mentioned dataset as benchmarks in their experiments.

We have studied 66 publicly available multimodal visual HAR datasets, in which 44 of them contain trimmed videos and 25 datasets involve untrimmed videos. Three datasets (HumanEva, Toyota-Smarthome, and Egogesture) offer both trimmed and untrimmed videos. Among trimmed datasets, 26 offer actions from single-view, and 18 datasets provide multi-view action videos. Among untrimmed datasets, 15 datasets contain action videos in single-view, and 10 offer multi-view videos. In total, there are 40 single-view and 26 multi-view datasets. The study shows that most multimodal action datasets contain single-view trimmed videos with three different modalities: RGB, depth, and skeleton.

8 Discussion

According to the studied papers, the combination of RGB and depth is more frequent in HAR algorithms as they provide complementary information about the appearance and 3D structure of the scene. Combining RGB and skeleton ranked second. In contrast, the combination of infrared with other modalities is employed less than other modalities since infrared data is appropriate in restricted applications.

The majority of studies have concentrated on HAR scenarios with complete modality, while HAR dealing with missed modalities is a new approach. Independent streams architecture is a more frequent framework in the complete modality group, employing score fusion or feature fusion techniques. As shown in Table 4, the state-of-the-art methods mainly use independent streams architecture. That is maybe because there is no need to handle heterogeneous data from different modalities in the independent streams architecture. However, independent streams cannot learn from mid-level complementary information of different modalities. Although dependent streams architecture can gain from mid-level complementary information of multiple modalities, it should handle heterogeneous data from different modalities. Additionally, new fusion strategies can also be used with dependent streams (Cheng et al. 2021; Zhou et al. 2021; Tian et al. 2020; Wang et al. 2019a; Joze et al. 2020).

Single-stream architecture does not need to fuse data or scores in the framework; however, the data from different modalities should be handled before feeding to the network. Skeletal data is mostly used as attention combined with RGB in single-stream architectures, and RGB and depth form a 4D input in single-stream architecture.

A Hallucination network provides new representations of modalities, and it is popular in the case of missing modalities. This idea is used in different methods, such as teacher-student frameworks and GANs. However, a modest number of studies have investigated multimodal visual HAR with missing modalities.

According to Table 4, the most commonly used dataset in multimodal vision-based HAR is NTU RGB+D (Shahroudy et al. 2016). State-of-the-art methods on this dataset obtain accuracies above 97% for the cross subject and cross view using RGB and skeleton modalities. In contrast, the newest datasets, MDAD (Jegham et al. 2019) and ETRI-Activity3D (Jang et al. 2020), are less frequent, and lower accuracies are obtained using them, which means more accurate methods are required to learn the diversity of actions. Further, the combination of RGB & depth is more frequent in the state-of-the-art methods.

Almost all of the most accurate methods in Table 4 employ RGB in their framework. Only the MSR-Action3D (Li et al. 2010) and SYSU 3D HOI (Hu et al. 2015) datasets obtain reasonable results without using the RGB modality.

9 Future directions

The following directions are pointed out for future research.

9.1 Transformers

Transformers first applied in NLP are recently entered into computer vision tasks such as HAR. As Transformers capture long-term dependencies and are capable of parallel processing, they are interested in video action recognition (Girdhar et al. 2019; Gavrilyuk et al. 2020; Chen and Mo 2023; Yang et al. 2022). Also, lightweight Transformers are used in specific applications (EK et al. 2022). Nowadays, Transformers are used in multimodal deep-based HAR (Li et al. 2021). Since Perceiver (Jaegle et al. 2021b) and Perceiver IO (Jaegle et al. 2021a) can be applied in several domains (Han et al. 2022), these two Transformer models are capable of being used in multimodal vision-based HAR. However, there are remaining opportunities to effectively model similar, long-term, and complex activities using Transformers and multimodal visual data. Besides, Transformers seem to be powerful in predicting future actions.

9.2 Large language models

Some convolutional networks have demonstrated promising results in HAR using single modality (Wang et al. 2023), While these can be utilized via multiple modalities as well. The rapid development and high capabilities of large language models (LLMs) present remarkable potential for the future. Initially developed for NLP tasks, LLMs have expanded their application to various vision tasks, such as image captioning (Zhu et al. 2023), visual question answering (Salaberria et al. 2023), OCR (Ye et al. 2023), image generation, and style transfer (Fu et al. 2022). Pretrained LLMs are employed as knowledge engine to generate text descriptions for body movements of actions. Training scheme is performed by using text encoder along with visual encoder (Xiang et al. 2023). Consequently, there is a growing interest in exploring the use of LLMs for action classification tasks with multiple data modalities.

9.3 Missing modality

Real-life applications usually deal with partial modalities due to different restrictions, such as noise or failure of sensors. Co-learning approaches try to transfer or distill knowledge from auxiliary modalities and assist in learning the model from them. Methods that employ hallucination networks in their frameworks (such as student-teacher or GANs) benefit from complementary information of all modalities, which are suggested in (Garcia et al. 2018, 2019; Xu et al. 2021; Thoker and Gall 2019; Wang et al. 2018a). Besides, Transformers are used in the teacher-student framework for NLP applications (Mukherjee and Awadallah 2020; Mirzadeh et al. 2020), and GANs (Jiang et al. 2021), which can be used in HAR. Other methods employ autoencoder to produce missed data modalities (Woo et al. 2023). New and accurate methods of transferring knowledge or knowledge distillation can improve recognition accuracy.

9.4 Few-shot and zero-shot learning

Collecting adequate data for all action classes is a big challenge. Few-shot (or one-shot) and zero-shot learning approaches solve this problem. Few-shot learning is the problem of making predictions based on a limited number of samples. However, zero-shot learning tries to predict without any training samples. Co-learning-based approaches like transfer learning, knowledge distillation, and GANs are tools to hallucinate diverse and discriminative features from a few data samples. In (Wang et al. 2021), GANs are used with significant advances in generating new data from a modality with fewer samples by using other modalities with rich samples. Others employ language models in the context of few-shot learning (Brown et al. 2020; Alayrac et al. 2022). Although there are methods for zero-shot action recognition using RGB data (Estevam et al. 2021), one-shot learning with the fusion of vision-based and sensor-based modalities (Memmesheimer et al. 2021), or Recurrent Transformers (Schatz et al. 2020) to synthesize human actions from novel views, there is no considerable attempt in deep-based HAR with multiple visual data modalities using few-shot or zero-shot learning.

9.5 Fusion methods

Section 6 reviews general and commonly used fusion methods. Nevertheless, some studies have introduced new fusion approaches (Cheng et al. 2021; Zhou et al. 2021; Joze et al. 2020; Singh and Vishwakarma 2021; Elmadany et al. 2018); it is anticipated to propose novel fusion methodologies , particularly in the context of Transformers (Hampiholi et al. 2023) to benefit from mid-level and heterogeneous information effectively.

9.6 Unsupervised, semi-supervised, and self-supervised learning

The entire dataset is used in supervised learning tasks while there is a lack or absence of labels in real scenarios. Semi-supervised and unsupervised learning try to solve tasks with limited or no data labels. Since deep approaches need huge datasets and dataset labeling is a labor-intensive processing task, semi-supervised (Singh et al. 2021) and unsupervised (Lin et al. 2022) learning are of outstanding importance. Multimodal data can be employed as additional information in semi-supervised and unsupervised learning as well (Patwary et al. 2022). Further, a model can be trained to learn one part of input from another part in self-supervised learning (Guo et al. 2022). It seems that self-supervised learning is capable of predicting tasks.

9.7 Datasets

Datasets with diverse and enormous samples have a significant role in developing deep-based algorithms. As mentioned in Sect. 7, many datasets are collected for video-based HAR with multiple visual modalities like ETRI-Activity3D (Jang et al. 2020), Toyota-Smarthome (Das et al. 2019b), and NTU RGB+D 120 (Liu et al. 2019). Due to the diverse intra-class and inter-class variations of actions in the video, the lack of large datasets with multiple modalities still remains, especially for particular applications [such as surgery (Twinanda et al. 2016) or industrial assembly process (Rückert et al. 2021)], complex actions (Li et al. 2010), prediction of actions in the actual scenarios (Dai et al. 2022), uncontrolled (Sung et al. 2011), cluttered and crowded environments (You and Jiang 2019) that need more investigation. Huge datasets for data-hungry DNNs are still in demand for HAR.

10 Conclusion

HAR is an important task in computer vision, which has attracted researchers' interest. DNNs make HAR algorithms more accurate. This paper presents a comprehensive review of deep-based HAR methods using multiple visual modalities. Methods are reviewed based on a novel four-level categorization, which considers framework modalities, modality availability, framework architecture, and framework similarities. This four-level categorization facilitates researchers to comprehend and compare methods in detail. Common properties between methods and their differences are highlighted. It is indicated that new approaches are evidently required to achieve higher accuracies in HAR.

Further, available benchmark HAR datasets providing multiple vision-based modalities are categorized into four groups based on providing trimmed or untrimmed videos, single-view or multi-view, and data modalities. Datasets in each group are compared graphically by plotting their characteristics.

Besides, the pros and cons of different architectures in the proposed four-level categorization are discussed. Then, the most accurate methods on more popular as well as the newest datasets are listed and commented on. Finally, some potential research directions are discussed.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A: Four-level categorization of deep-based HAR methods using multiple visual data modalities

All methods are summarized here based on the proposed four-level classification for ease of access. First they are categorized based on framework modalities which are determined in each subsection: RGB & Depth, RGB & Skeleton, Depth & Skeleton, RGB & Depth & Skeleton, and IR & others. Framework architecture is considered in the third level and conveyed the availabiity of data which is considered in the second level of categorization. Framework architecture is determined at the left most of each table. Similar methods have the same table color to identify fourth-level in the proposed categorization.

A.1 RGB and depth

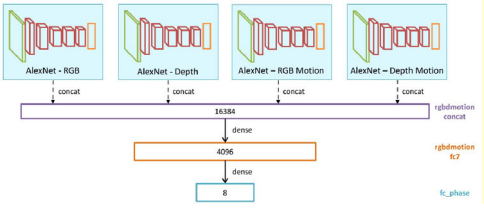
	Paper	Architecture	Pre-train	Classifier	Fusion	Dataset	Accuracy
Independent Streams	Action recognition based on a mixture of RGB and depth based skeleton (Das et al 2017)	RGB-D for skeleton estimation – skeleton for action	VGG-f on ImageNet	SVM	Weighted sum	CAD-60 MSRDailyActivity3D	95.58 84.3
	Multi-stream deep architecture for surgical phase recognition on multi-view RGBD videos (Twinanda et al 2016)	Four-stream CNN, RGB, depth, RGB motion, depth motion	AlexNet	SVM	Late Concatenate	VerCArm24 VerCT13	93.7 93.9
	Human action recognition in RGB-D videos using motion sequence information and deep learning (Ijjina and Chalavadi 2017)	Two-stream CNN	-	ELM	Membership function	SBU Kinect NATOPS MIVIA	90.98 86.58 93.37
	Action recognition from RGB-D data: Comparison and fusion of spatio-temporal handcrafted features and deep strategies (Asadi-Aghbolaghi et al 2017)	Four-stream CNN i.e., RGB, depth, optical flow, and scene flow	on UCF-101	Fully connected (FC)	features and score averaging	MSRDaily MontalbanoII	70.08 96.16
	Human activity recognition in RGB-D videos by dynamic images (Mukherjee et al 2020)	Two-stream CNN	VGG16 or ResNext-101	FC	late concatenate feature vectors of streams	MSR-Action Proposed Dataset	96.17 54.94
	Combining CNN streams of dynamic image and depth data for action recognition (Singh et al 2020)	Four-stream CNN	VGG-F	FC	late fusion of scores	MSRDaily UTD MHAD CAD-60	96.38 95.74 94.80
	Privacy-preserving human action recognition as a remote cloud service using RGB-D sensors and deep CNN (Rajput et al 2020)	Four-stream CNN	MobileNet	FC	late fusion of scores	UTD MHAD	89.30
	Human action recognition using RGB-D sensor and deep convolutional neural networks (Imran and Kumar 2016)	Four-stream CNN	VGG16 On ImageNet	FC	Score fusion: product and avg	UTD MHAD	91.2
	Large-scale isolated gesture recognition using pyramidal 3d convolutional networks (Zhu et al 2016)	Two-stream CNN	C3D on UCF-101	FC	late class score fusion	IsoGD	50.93

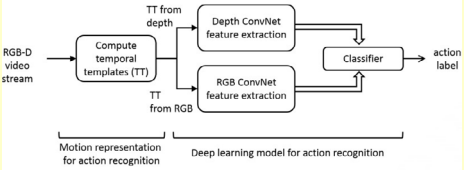
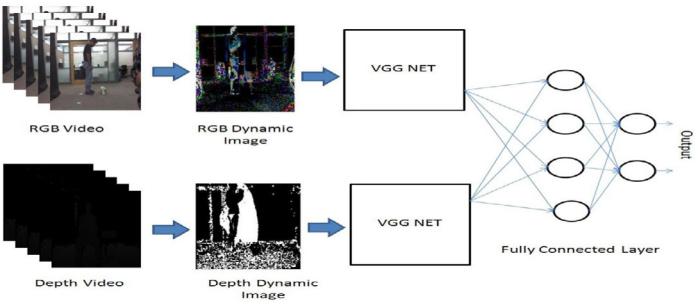
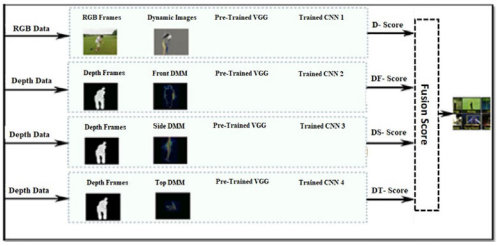
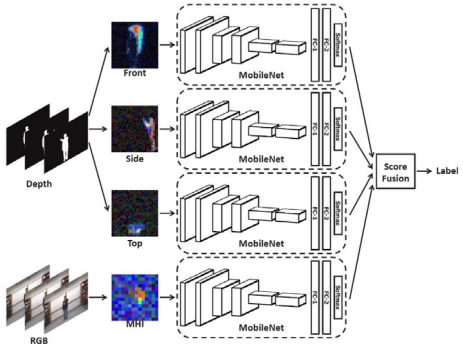
Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model (Li et al 2016b)	Two-stream 3DCNN	C3D on Sports-1M	SVM	Feature-level, average or integrated	IsoGD	49.2
Multimodal gesture recognition based on the resc3d network (Miao et al 2017)	Three-stream 3DCNN	-	SVM	Feature-level: canonical correlation analysis (CCA)	IsoGD	67.71
A spatio-temporal attention-based ResC3D model for large-scale gesture recognition (Li et al 2019)	Three-stream 3DCNN, ResC3D	VGG-16	SVM	Feature-level: canonical correlation analysis (CCA)	IsoGD SKIG	68.14 100
Multimodality fusion based on consensus-voting and 3d convolution for isolated gesture recognition (Duan et al 2016)	Four-stream 3DCNN	on imageNet	SVM	Late: average or integrated	IsoGD RGBD-HuDaAct	67.19 96.74
RGB-D based action recognition with light-weight 3D convolutional networks (Zhang et al 2018a)	Two-stream 3DCNN	on NTU for N-UCLA	FC	Score fusion (avg), merging scores	N-UCLA NTU RGB+D	95.5-cv 93.2-cs 97.6-cv
Two streams recurrent neural networks for large-scale continuous gesture recognition (Chai et al 2016)	Two-stream RNN	-	LSTM	late, concatenate	ConGD	28.6
Multimodal gesture recognition using 3-D convolution and convolutional LSTM (Zhu et al 2017)	Two-stream CNN and RNN	-	FC	late fusion-averaging	IsoGD SKIG	51.02 98.89
Learning spatio-temporal features using 3dcnn and convolutional lstm for gesture recognition (Zhang et al 2017a)	Three-stream 3DCNN and RNN	-	SVM	Feature-level (concate), score fusion (avg)	IsoGD SKIG	62.14 99.53
MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences (Elboushaki et al 2020)	Four-stream: RGB, depth, iMHI, iDMM	-	FC	Late: concatenate feature- (linear, max, product)	IsoGD SKIG NATOPS SBU Kinect	72.53 99.72 95.87 97.51
Hand Gesture Recognition in Video Sequences Using Deep Convolutional and Recurrent Neural Networks (Obaid et al 2020)	Two-stream CNN and RNN	-	FC	Feature-level	VIVA	93

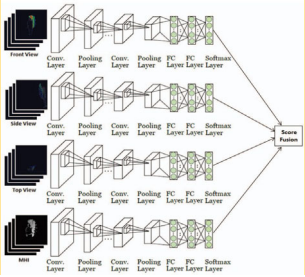
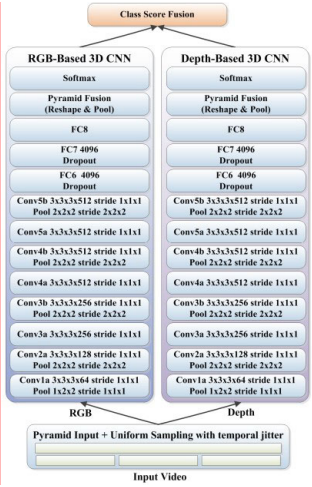
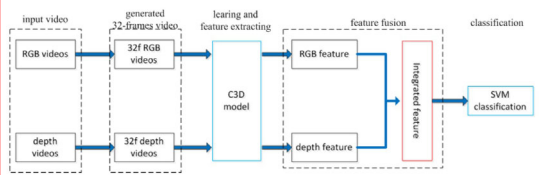
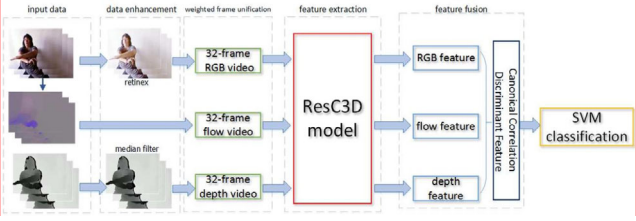
View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics (Dhiman and Vishwakarma 2020)	Two-stream CNN	HPM and InceptionV3	FC	Score fusion: late(max, product, average)	N-UCLA UWA3D...II NTU RGB+D	85.38 86.18 79.4 cs 84.1 cv
Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos (Shahroudy et al 2017)	Two-stream	-	SVM	Feature-level (concate), score fusion	NTU RGB+D MSRDaily action-pairs HuDaAct RGBD action	74.86-cs 97.5 100.0 99.0 95.5-s1 93.7-s2 83.8-s3
Joint Deep Learning for RGB-D Action Recognition (Qin et al 2018)	Two-stream CNN	VGG16	FC	late fusion-merged (features)	CAD-60 MSR daily activity	93.3 81.2
Multi-stream deep neural networks for rgb-d egocentric action recognition (Tang et al 2018)	Three-stream CNN: RGB, optical flow, depth Hand as a cue	VGG16 on ImageNet	FC	MDNN- their method	THU-READ WCVS GUN-71	83.64-cs 71.83-cs 34.04-cs
Two-Stream Network with 3D Common-Specific Framework for RGB-D Action Recognition (Qin et al 2019)	Two-stream 3DCNN	BNInception	FC	Feature-level	CAD-60 SYSU 3D HOI NTU RGB+D	88.2 87.9 79.5
Segment spatial-temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition (Ren et al 2021a)	Two-stream CNN, dynamic images	ResNet on ImageNet	FC	Feature-level	NTU-RGB+D NTU...120 SYSU 3D HOI PKU-MMD	89.4-cs,91.2-cv 86.9-cs 87.7-cv 84.2-s1 82.9-s2 92.1-cs 93.2-cv
Cooperative training of deep aggregation networks for RGB-D action recognition (Wang et al 2018c)	Two-stream c-ConvNet - input of each stream is 4D	VGG16 on ImageNet	FC	product score	ChaLearn LAP IsoGD NTU RGB+D SYSU 3D HOI	44.8 86.42(cs) 89.08(cv)
Multimodality learning for human action recognition (Ren et al 2021b)	Five-stream segmentation ConvNet – RGB(f,b), depth(f,b), optical flow	Inception, InceptionV3, ResNet101	MLP	Score fusion - average	NTU RGB+D SYSU 3D HOI UWA3D Multi-view II	89.70cs,92.97cv 86.71,87.08 84.5
A hybrid network for large-scale action recognition from rgb and depth modalities (Wang et al 2020)	Four-stream CNN and 3D ConvLSTM	ResNet-50 on ImageNet	SVM	Feature-level: canonical correlation analysis (CCA)	IsoGD NTU RGB+D M^2I	68.13 89.51cs,91.68cv 100-sv-fv 93.8-sv-fv 90.6-fv-sv

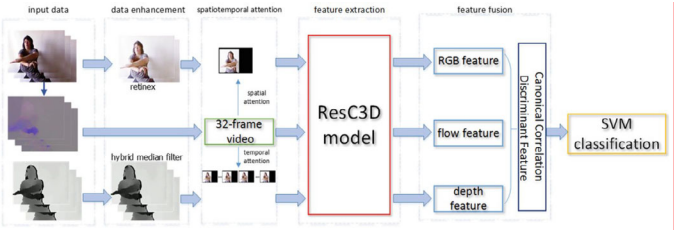
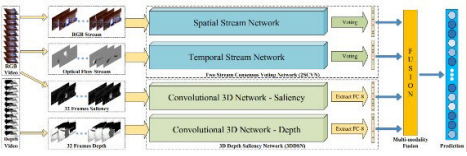
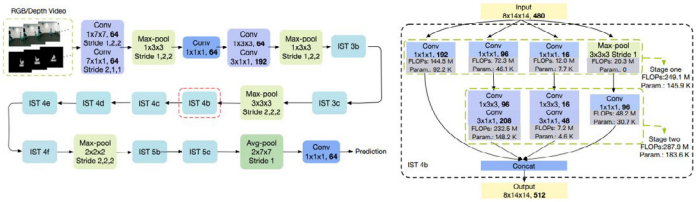
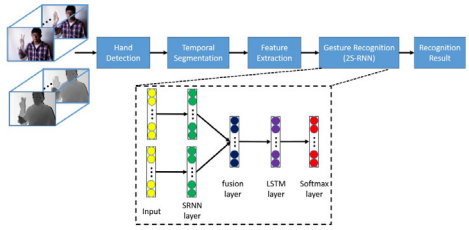
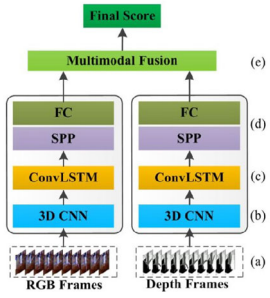
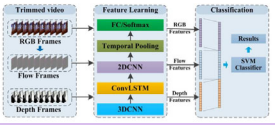
	Multi-view region-adaptive multi-temporal DMM and RGB action recognition (Al-Faris et al 2020b)	Multi-stream 3DCNN	AlexNet, C3D	SVM	Average score fusion	MSR 3DAction N-UCLA MSR3D daily activity	96.5 cs 97.2cs,86.2cv 87.5
	spatio-temporal Multimodal Learning with 3D CNNs for Video Action Recognition (Wu et al 2021)	Two-stream 3DCNN	Res3D on kinetics	FC	late(max, multi-ply,average) and early: concatenate features	SBU Kinect UTD-MHAD NTU RGB+D NTU RGB+D 120	99.88 93.57 91.13cs,94.31cv 81.27cs,83.36cset
Dependent Streams	Trear: Transformer-based rgb-d egocentric action recognition (Li et al 2021)	Two-stream Transformers	ResNet	FC	mutual-attentional-in the training	THU-READ FPHA WCVS	88.33 97.04 71.50
	"Reading Pictures Instead of Looking": RGB-D Image-Based Action Recognition via Capsule Network and Kalman Filter (Zhao et al 2021)	Transformers	-	Transforn	DenseFusion	OAD G3D ChaLearnGes-tur	83.3 86.5 72.2
	Cross-Modality Compensation Convolutional Neural Networks for RGB-D Action Recognition (Cheng et al 2021)	Two-stream CNN	ResNet, VGG,	-	cross-modality compen-sation block	NTU RGB.120 THU-READ PKU-MMD	79.1cs,81.1cv 77.2cs,92.3cg 90.4cs,91.4cv
	Adaptive cross-fusion learning for multimodal gesture Recognition (Zhou et al 2021b)	Two-stream	C3D and ResNet-50	1x1x1 convo-lution layer	Adaptive Cross-modal Weighting (ACmW)	IsoGD NVGesture	59.97 83.96
	Gesture recognition based on multilevel multimodal feature fusion (Tian et al 2020)	Two-stream	-	LSTM	Multi-level and mul-timodal fusion	Self-built gesture	96.4
	Generative Multi-View Human Action Recognition (Wang et al 2019a)	Two-stream	ResNet-101 on ImageNet	-	Feature-level, decision-level (View Correlation Discovery Network)	MHAD UWA3D Multiview DHA	98.94 76.28 88.72
Single Stream	Activity recognition for indoor fall detection using convolutional neural network (Adhikari et al 2017)	One-stream CNN	-	FC	-	Their dataset	74
	Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video (Pigou et al 2018)	One-stream Temporal CNN model	-	RNN(best results) and FC	-	Montalbano Gesture Recognition	90.6

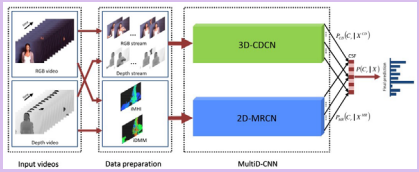
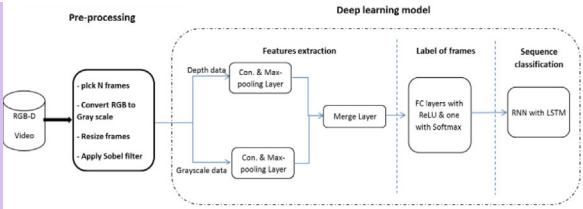
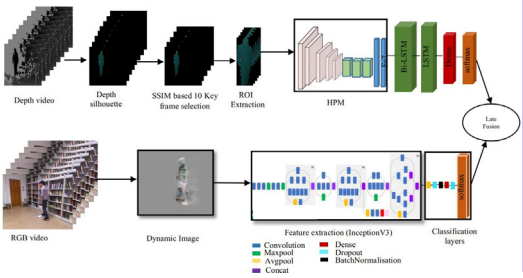
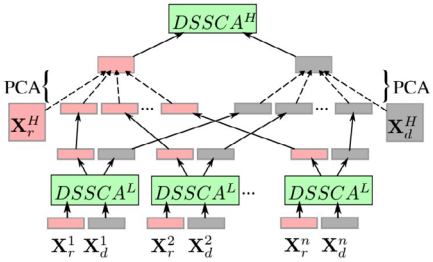
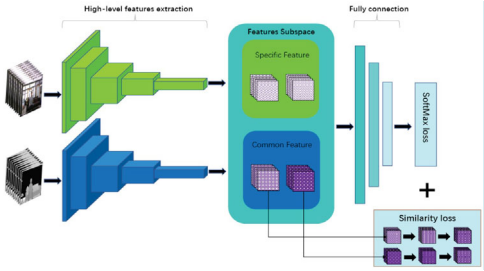
	Scene Flow to Action Map: A New Representation for RGB-D based Action Recognition with Convolutional Neural Networks (Wang et al 2017a)	One-stream CNN- Sceneflow action map as dynamic images (4D input)	AlexNet on ImageNet	-	max-score, average-score, multiply-score	IsoGD M^2I	36.27 89.4sv 91.2fv 87.6 (sv → fv) 76.5 (fv → sv)
	Regional Attention with Architecture-Rebuilt 3D Network for RGB-D Gesture Recognition (Zhou et al 2021a)	4D input , attention based	I3D as backbone	Conv1x1x	-	ChalearnIsoGD NvGesture	66.62 88.59
	Soft Spatial Attention-based Multimodal Driver Action Recognition Using Deep Learning (Jegham et al 2020a)	CNN, depth-based spatial attention network	VGG16	FC	No- for multi-view : late fusion, concatenate features	MDAD	69.79sv 65.63fv
Hallucination Network	Modality Distillation with Multiple Stream Networks for Action Recognition (Garcia et al 2018)	hallucination	Modified ResNet50 on ImageNet	FC	Co-learning	NTU RGB+D UWA3D....II N-UCLA	79.73cs,82.11cv 79.66 88.87
	Learning with privileged information via adversarial discriminative modality distillation (Garcia et al 2019)	GANs	Modified ResNet50 on ImageNet	FC	(average logits)-co-learning	NTU RGB+D N-UCLA	73.11cs,81.50cv 91.64
Ensemble	Distillation multiple choice learning for multimodal action recognition (Garcia et al 2021)	R(2+1)D-18 architecture, DMCL	-	FC	Co-learning	UWA3D...II NTU RGB+D 120	89.75 89.74-cs

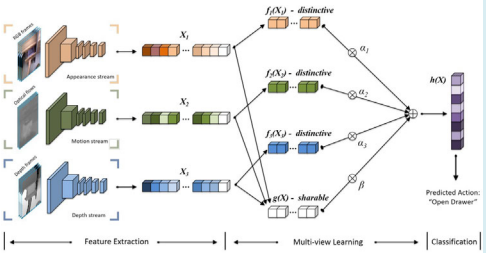
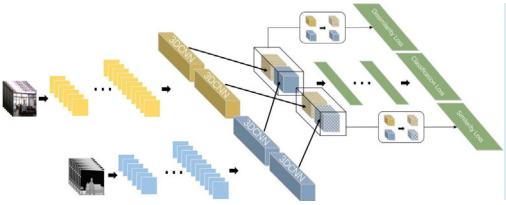
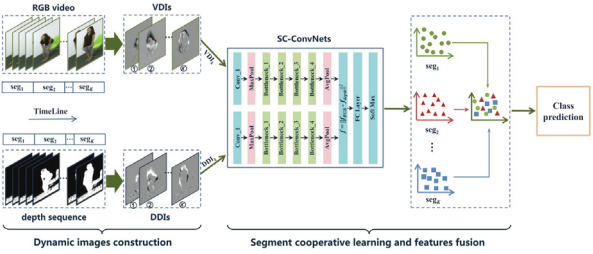
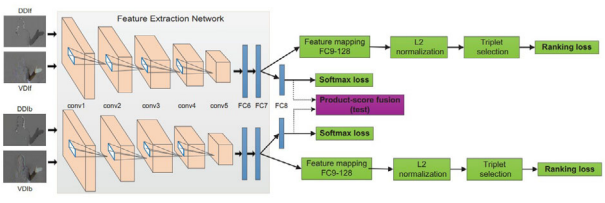
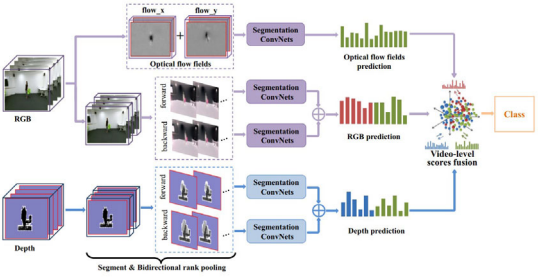
	Paper	Architecture
Independent Streams	Action recognition based on a mixture of RGB and depth based skeleton (Das et al 2017)	None
	Multi-stream deep architecture for surgical phase recognition on multi-view RGBD videos (Twinanda et al 2016)	

Human action recognition in RGB-D videos using motion sequence information and deep learning (Ijjina and Chalavadi 2017)		
Action recognition from RGB-D data: Comparison and fusion of spatio-temporal handcrafted features and deep strategies (Asadi-Aghbolaghi et al 2017)	None	
Human activity recognition in RGB-D videos by dynamic images (Mukherjee et al 2020)		
Combining CNN streams of dynamic image and depth data for action recognition. (Singh et al 2020)		
Privacy-preserving human action recognition as a remote cloud service using RGB-D sensors and deep CNN (Rajput et al 2020)		

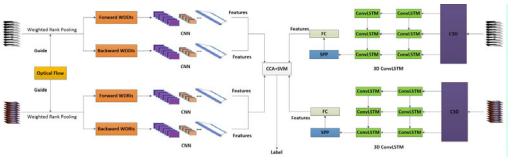
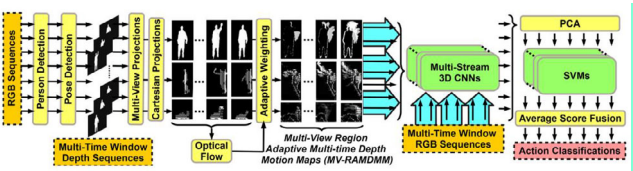
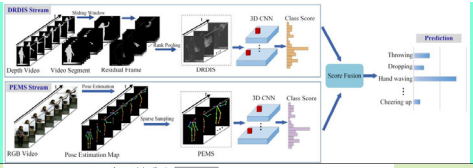
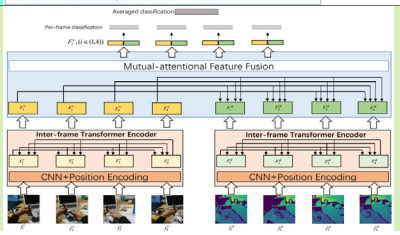
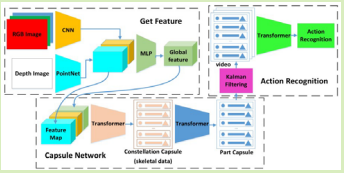
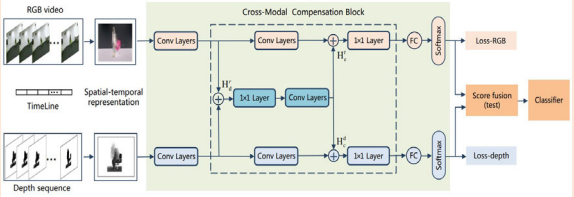
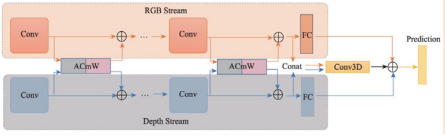
Human action recognition using RGB-D sensor and deep convolutional neural networks (Imran and Kumar 2016)		
Large-scale isolated gesture recognition using pyramidal 3d convolutional networks (Zhu et al 2016)		
Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model (Li et al 2016b)		
Multimodal gesture recognition based on the resc3d network (Miao et al 2017)		

A spatio-temporal attention-based ResC3D model for large-scale gesture recognition (Li et al 2019)	
Multimodality fusion based on consensus-voting and 3d convolution for isolated gesture recognition (Duan et al 2016)	
RGB-D based action recognition with light-weight 3D convolutional networks (Zhang et al 2018a)	
Two streams recurrent neural networks for large-scale continuous gesture recognition (Chai et al 2016)	
Multimodal gesture recognition using 3-D convolution and convolutional LSTM (Zhu et al 2017)	
Learning spatio-temporal features using 3dcnn and convolutional lstm for gesture recognition (Zhang et al 2017a)	

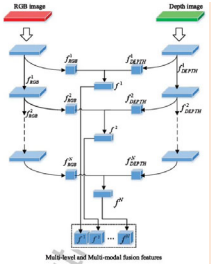
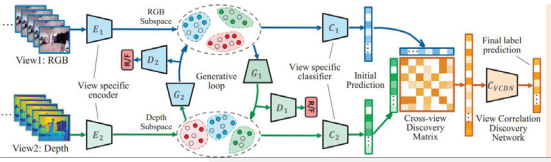
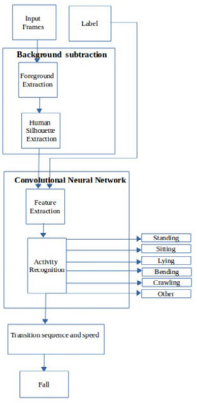
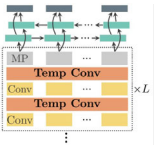
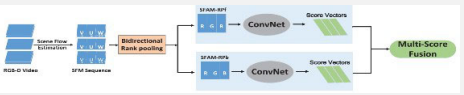
MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences (Elboushaki et al 2020)	
Hand Gesture Recognition in Video Sequences Using Deep Convolutional and Recurrent Neural Networks (Obaid et al 2020)	
View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics (Dhiman and Vishwakarma 2020)	
Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos (Shahroudy et al 2017)	
Joint Deep Learning for RGB-D Action Recognition (Qin et al 2018)	

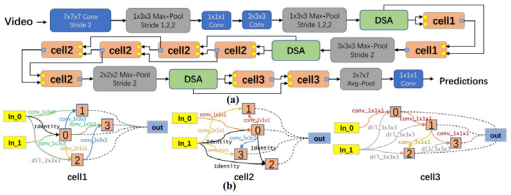
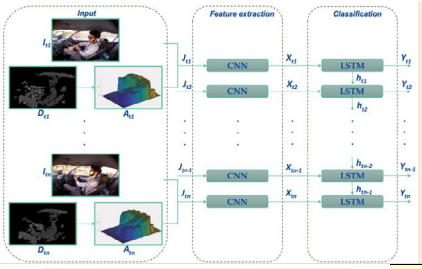
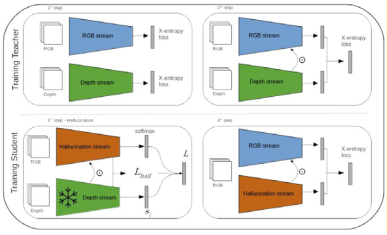
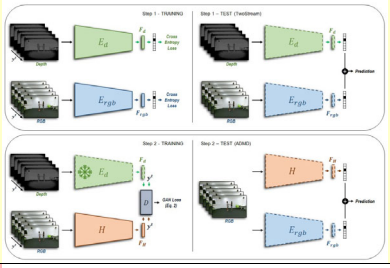
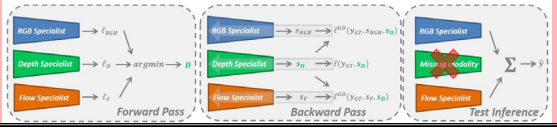
Multi-stream deep neural networks for rgb-d egocentric action recognition (Tang et al 2018)	
Two-Stream Network with 3D Common-Specific Framework for RGB-D Action Recognition (Qin et al 2019)	
Segment spatial-temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition (Ren et al 2021a)	
Cooperative training of deep aggregation networks for RGB-D action recognition (Wang et al 2018c)	
CMultimodality learning for human action recognition (Ren et al 2021b)	

Dependent Streams

A hybrid network for large-scale action recognition from rgb and depth modalities (Wang et al 2020)	
Multi-view region-adaptive multi-temporal DMM and RGB action recognition (Al-Faris et al 2020b)	
spatio-temporal Multimodal Learning with 3D CNNs for Video Action Recognition (Wu et al 2021)	
Trear: Transformer-based rgb-d egocentric action recognition (Li et al 2021)	
Reading Pictures Instead of Looking": RGB-D Image-Based Action Recognition via Capsule Network and Kalman Filter (Zhao et al 2021)	
Cross-Modality Compensation Convolutional Neural Networks for RGB-D Action Recognition (Cheng et al 2021)	
Adaptive cross-fusion learning for multimodal gesture Recognition (Zhou et al 2021b)	

Single-Stream

Gesture recognition based on multilevel multimodal feature fusion (Tian et al 2020)		
Generative Multi-View Human Action Recognition (Wang et al 2019a)		
Activity recognition for indoor fall detection using convolutional neural network (Adhikari et al 2017)		
Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video (Pigou et al 2018)		
Scene Flow to Action Map: A New Representation for RGB-D based Action Recognition with Convolutional Neural Networks (Wang et al 2017a)		

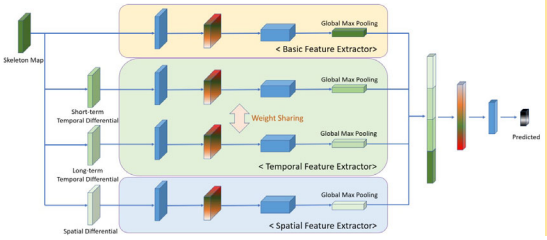
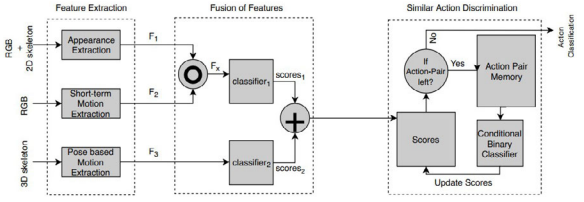
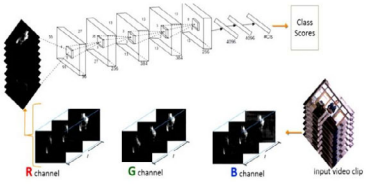
	Regional Attention with Architecture-Rebuilt 3D Network for RGB-D Gesture Recognition (Zhou et al 2021a)	
Hallucination Network	Soft Spatial Attention-based Multimodal Driver Action Recognition Using Deep Learning (Jegham et al 2020a)	
	Modality Distillation with Multiple Stream Networks for Action Recognition (Garcia et al 2018)	
	Learning with privileged information via adversarial discriminative modality distillation (Garcia et al 2019)	
Ensemble	Distillation multiple choice learning for multimodal action recognition (Garcia et al 2021)	

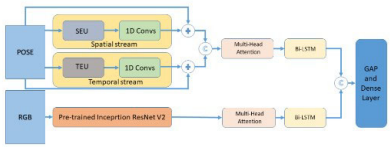
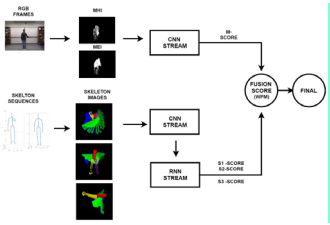
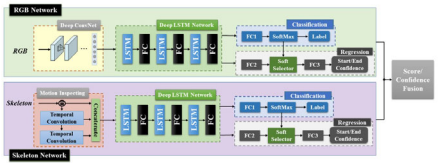
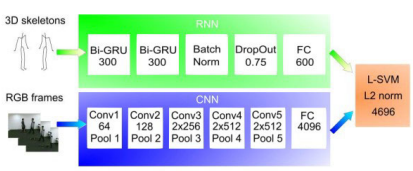
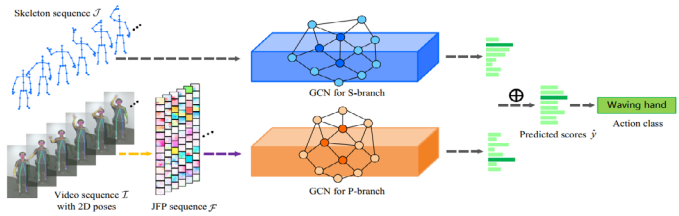
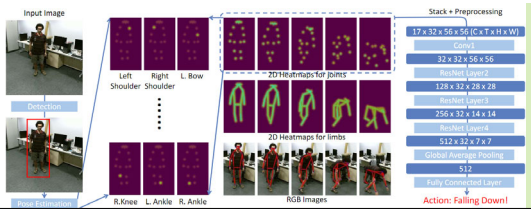
A.2 RGB and skeleton

	Paper	Architecture	Pre-train	Classifier	Fusion	Dataset	Accuracy
Independent Streams	ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly (Jang et al 2020)	Four-stream adaptive CNN (FSA-CNN)	-	-	Feature-level, concatenate	NTU RGB+D ETRI-Activity3D	91.5 (cs) 93.7 (cs)
	A new hybrid architecture for human activity recognition from rgb-d videos (Das et al 2019c)	Three-stream	ResNet-152 pre-trained on ImageNet and I3D	SVM	Score fusion and early fusion	CAD-60 CAD-120 MSRDaily... NTU RGB+D	98.52 94.40 97.81 92.20
	Human activity recognition using combined deep architectures (Tomas and Biswas 2017)	Three-stream	Pre-trained CNN	FC	Score fusion average	MSRDaily... MSR-Action3D	91.3 74.6
	Attention-Driven Body Pose Encoding for Human Activity Recognition (Debnath et al 2021)	Two-stream attentionbased, CNN and RNN	Inception ResNet-v2	FC	Early fusion and late fusion (concatenate)	MSRDaily... NTU RGB+D SBU Kinect	92.5 87.7 (cs) 96.5
	Deep learning-based multimodal approach using RGB and skeleton sequences for human activity recognition (Verma et al 2020)	Two-stream, CNN and RNN	Pre-trained but not mentioned on what	FC	Score level-WPM and feature-level	UTD-MHAD CAD60 NTU RGB+D120	96.5 93.2,91.9 76.7cs 77.9cset
	Multi-modality multi-task recurrent neural network for online action detection (Liu et al 2018b)	Two-stream, CNN and RNN	On UCF-101	FC	Score confidence fusion	G3D OAD	Online detection parameters
	Two-stream RNN/CNN for action recognition in 3D videos (Zhao et al 2017)	Two-stream, CNN and RNN	On Sports-1M	Linear SVM	Feature-level and decision fusion	NTU RGB+D	83.74cs 93.65cv
	JOLO-GCN: mining joint-centered light-weight information for skeleton-based action recognition (Cai et al 2021)	Two-stream GCN-based network	-	-	Score fusion: linear-blending	NTU RGB+D NTU...120 Kinetics-Skeleton	93.8cs,98.1cv 87.6cs,89.7cse 38.3top1 62.3top5
	Revisiting skeleton-based action recognition (Duan et al 2022)	Two-stream 3D CNN	-	FC layer	RGBPose slowfast	NTU RGB+D NTU RGB+D 120	97.0 (cs) 95.3 (cs)
	Toyota smarthome: Real-world activities of daily living (Das et al 2019b)	Two-stream spatial and temporal guided	I3D on Kinetics-400	FC	Feature-level: average pooling	NTU RGB+D N-UCLA ToyotaSmarthome	92.2cs,94.6cv 92.4V31,2 75.3cs 61cv1 68.2cv2

	Skeleton-Indexed Deep Multi-Modal Feature Learning for High Performance Human Action Recognition (Song et al 2018)	Three-stream attention-based RGB, optical flow and skeleton	-	FC	Late, average and early fusion	NTU RGB+D MSRDailyActivity3D	91.4cs,97.15cv 90.63
Dependent Streams	Human activity recognition with pose-driven attention to rgb (Baradel et al 2018)	Two-stream pose+RGB	On NTU for other datasets	FC	Logit level-late	NTU RGB+D SBU Kinect MSR Daily Action	84.8cs,90.6cv 94.1 90.0
	Pose-conditioned spatio-temporal attention for human action recognition (Baradel et al 2017b)	Two-stream attention based. LSTM + CNN	On ILSVRC 2012	FC	Logit level late fusion	NTU RGB+D SBU Kinect MSRDailyActivity3D	84.8cs,90.6cv 94.1 90.0
	Human action recognition: Pose-based attention draws focus to hands (Baradel et al 2017a)	Two-stream attention based	On ILSVRC 2012	FC	Score fusion sum	NTU RGB+D	82.5-cs 88.6-cv
	Action recognition based on 3d skeleton and rgb frame fusion (Liu et al 2019b)	Two-stream attention based	VA-pre+ LSTM,Xception Two-Attention	FC	Feature level concatenate	NTU RGB+D SYSU 3D HOI	85.4cs,91.6cv 80.9set1 82.6set2
	SGM-Net: Skeleton-guided multimodal network for action recognition (Li et al 2020a)	Two-stream skeleton guided-attention-based	R(1+2)D on Kinetics	FC	Score level	NTU RGB+D Sub-JHMDB	89.1cs,95.9cv 84.3
	Fusion of Skeleton and RGB Features for RGB-D Human Action Recognition (Weiyao et al 2021)	Two-stream R(2+1)D and 2S-AGCN	ResNet-18 on kinetics	FC	Bilinear Pooling and Attention Network (BPAN)	NTU RGB+D NTU...120 UTD-MHAD	94.85cs,97.4cv 86.6cs,88.1cse 95.07
	MMTM: multimodal transfer module for CNN fusion (Joze et al 2020)	Two dependent streams	I3D on ImageNet and Kinetics	FC	MMTN for intermediate fusion and late fusion-concatenate	NTU RGB+D	90.11
Single Stream	Regularizing long short term memory with 3D human-skeleton sequences for action recognition (Mahasseni and Todorovic 2016)	Two-stream: attention-based by skeleton: LSTM +CNN	GoogLeNet on ImageNet	FC	No fusion. Skeleton is as a parameter in similarity	Sports-1M HMDB-51 UCF101	75.9 55.3 86.9
	Where to focus on for human action recognition? (Das et al 2019a)	Two-stream attention-based RGB+ 2D skeleton and 3D skeleton, I3D	I3D on ImageNet and Kinetics	FC	Feature level (concatenate and sum) Network (BPAN)	NTU RGB+D N-UCLA	93.0cs,95.4cv 93.1v31,2
	Vpn: Learning video-pose embedding for activities of daily living (Das et al 2020)	Two-stream attention based	I3D on ImageNet and Kinetics	FC	Feature level VPN	NTU RGB+D NTU...120 ToyotaSmarthome N-UCLA	95.5cs,98.0cv 86.3cs,87.8cset 60.8cs,43.8cv1 53.5cv2 93.5v31,2

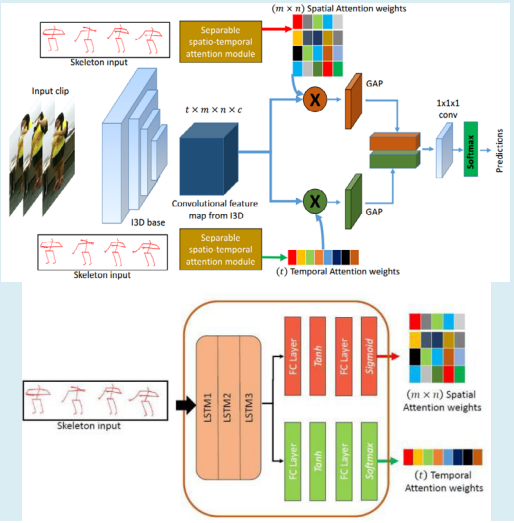
	VPN++: Rethinking Video-Pose embeddings for understanding Activities of Daily Living (Das et al 2021)	Two-stream-attention based	I3D on ImageNet and Kinetics-400	FC	Score-level and feature level	ToyotaSmarthom NTU RGB+D NTU...120 N-UCLA	71.0cs,58.1cv2 96.6cs,99.1cv 90.7cs,92.5cset 93.5v31,2
Hallucination Network	Cross-modality online distillation for multi-view action recognition (Xu et al 2021)	Two-stream attention-based teacher and student	DensePose-RCNN on DensePoseC-OCO	FC	Cross-modality aggregated transfer network	IXMAS N-UCLA	98.53 85.21
	Cross-modal knowledge distillation for action recognition (Thoker and Gall 2019)	ST-GCN	-	FC	Score fusion: avg or max	NTU RGB+D	79.50
Ensemble	Modality compensation network: Cross-modal adaptation for action recognition (Song et al 2020)	Two-stream with auxiliary data	BN-Inception	FC	Score fusion and MCN in feature level	NTU RGB+D MSRDailyActivity3D	90.0cs,96.3cv 91.9

	Paper	Architecture
Independent Streams	ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly (Jang et al 2020)	
	A new hybrid architecture for human activity recognition from rgb-d videos (Das et al 2019c)	
	Human activity recognition using combined deep architectures (Tomas and Biswas 2017)	

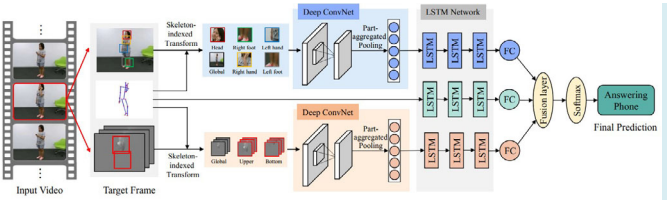
Attention-Driven Body Pose Encoding for Human Activity Recognition (Debnath et al 2021)		
Deep learning-based multi-modal approach using RGB and skeleton sequences for human activity recognition (Verma et al 2020)		
Multi-modality multi-task recurrent neural network for online action detection (Liu et al 2018b)		
Two-stream RNN/CNN for action recognition in 3D videos (Zhao et al 2017)		
JOLO-GCN: mining joint-centered light-weight information for skeleton-based action recognition (Cai et al 2021)		
Revisiting skeleton-based action recognition (Duan et al 2022)		

Dependent Streams

Toyota smarthome: Real-world activities of daily living (Das et al 2019b)



TSkeleton-Indexed Deep Multi-Modal Feature Learning for High Performance Human Action Recognition (Song et al 2018)



Human activity recognition with pose-driven attention to rgb (Baradel et al 2018)

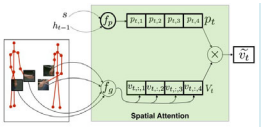
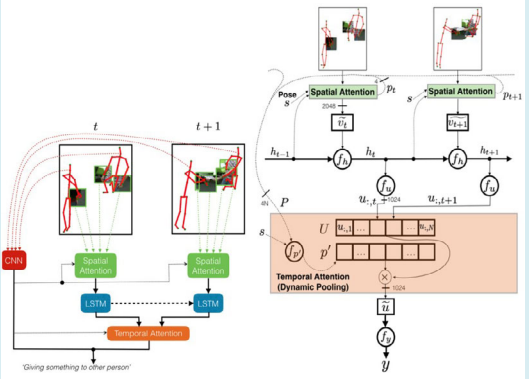
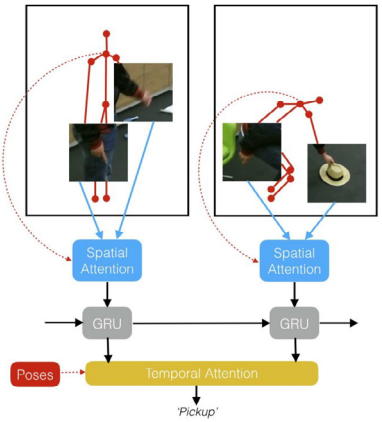
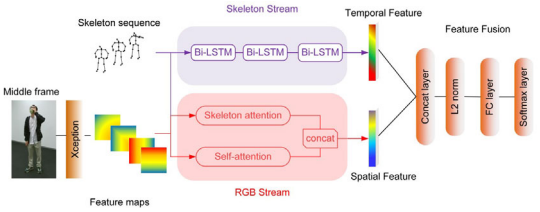
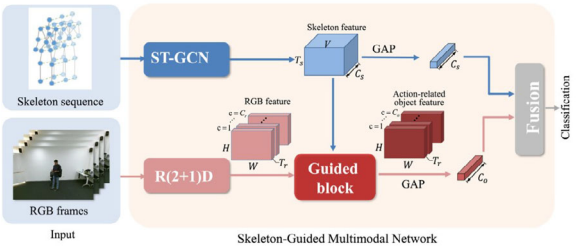
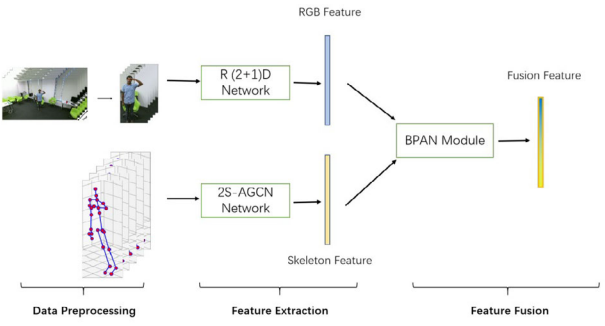
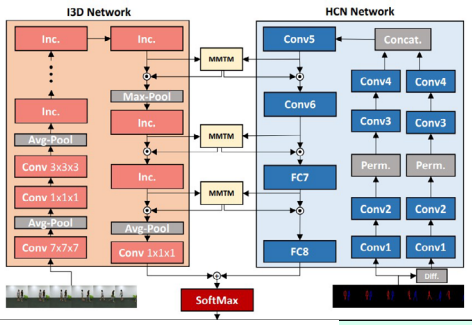
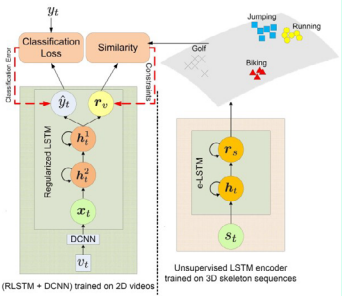
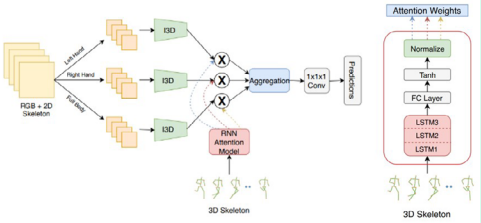
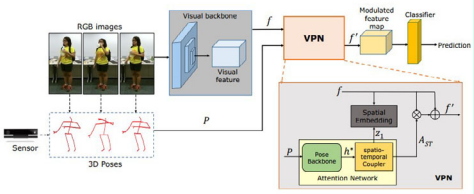
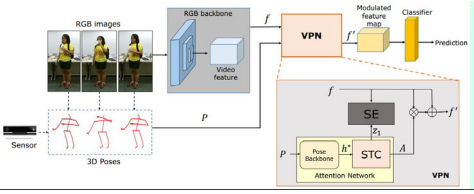


Figure 1: The spatial attention mechanism

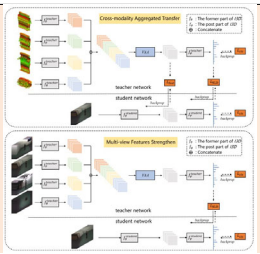
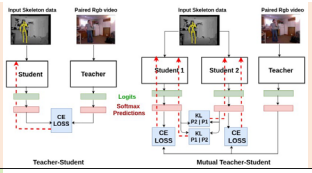
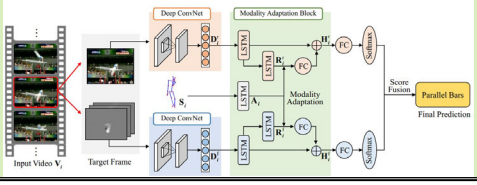
Pose-conditioned spatio-temporal attention for human action recognition (Baradel et al 2017b)



Human action recognition: Pose-based attention draws focus to hands (Baradel et al 2017a)	
Action recognition based on 3d skeleton and rgb frame fusion (Liu et al 2019b)	
SGM-Net: Skeleton-guided multimodal network for action recognition (Li et al 2020a)	
Fusion of Skeleton and RGB Features for RGB-D Human Action Recognition (Weiyao et al 2021)	

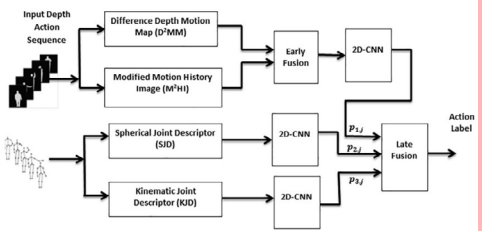
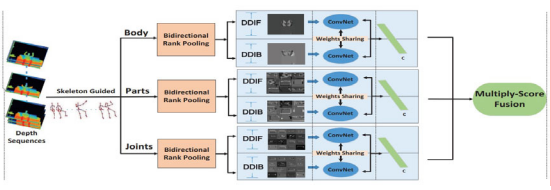
	MMTM: multi-modal transfer module for CNN fusion (Joze et al 2020)	
Single-Streams	Regularizing long short term memory with 3D human-skeleton sequences for action recognition (Mahasseni and Todorovic 2016)	
	Where to focus on for human action recognition? (Das et al 2019a)	
	Vpn: Learning video-pose embedding for activities of daily living (Das et al 2020)	
	VPN++: Rethinking Video-Pose embeddings for understanding Activities of Daily Living (Das et al 2021)	

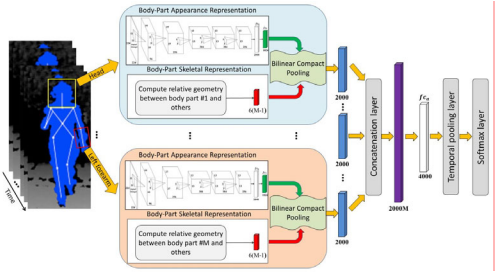
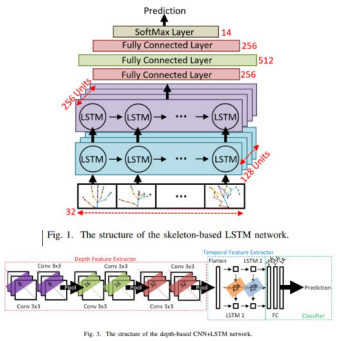
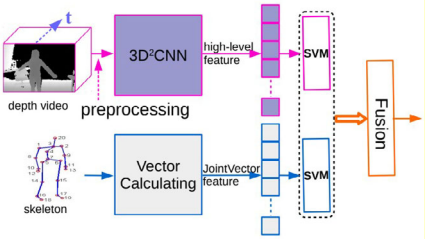
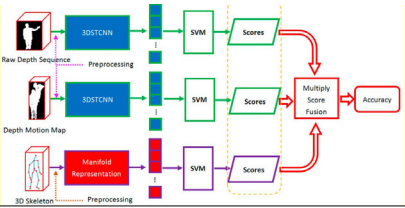
A.3 Depth and skeleton

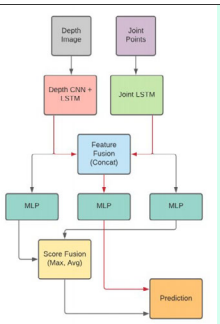
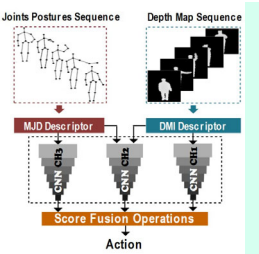
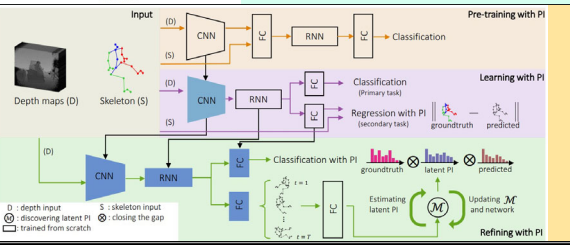
Hallucination Network	Cross-modality online distillation for multi-view action recognition (Xu et al 2021)	
Hallucination Network	Cross-modal knowledge distillation for action recognition (Thoker and Gall 2019)	
Ensemble Method	CModality compensation network: Cross-modal adaptation for action recognition (Song et al 2020)	

	Paper	Architecture	Pre-train	Classifier	Fusion	Dataset	Accuracy
Independent Streams	Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition (Rani et al 2021)	Three-stream 2D-CNN	-	FC	Early fusion and late fusion: Score fusion – average, summation, product and maximum	MSR-Action3D UTD-MHAD	93.4457 88.4725
Independent Streams	Structured images for RGB-D action recognition (Wang et al 2017b)	Six-stream CNN Weight- sharing pairwise, skeleton guided	AlexNet	FC	Concatenate, Score fusion (max, average, multiply)	MSRAct... G3D MSRDaily.... SYSU UTD-MHAD	100 96.05 97.50 95.42 89.04
Independent Streams	Learning Action Recognition Model From Depth and Skeleton Videos (Rahmani and Bennamoun 2017)	Two-stream CNN	-	FC	Concatenate features	NTU RGB+D UWA3DMultivie II	75.2cs,83.1cv 84.2
Independent Streams	3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track (De Smedt et al 2017)	CNN	VGG11 on ImageNet	SVM	-	Shrec'17 track (their dataset)	81.90

	CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition (Lai and Yanushkevich 2018)	Two-stream CNN + LSTM	-	FC or SVM	Feature-level (concat), score-level (Max, Avg)	dynamic hand gesture 14/28 (DHG-14/28)	85.46
	3D-based Deep Convolutional Neural Network for action recognition with depth sequences (Liu et al 2016b)	Two-stream 3DCNN	Model on MSR-Action3D	SVM	Decision level fusion	UTKinect... MSR-Action3D	96 84.07(cs)
	3d behavior recognition based on multimodal deep space-time learning (Zhao et al 2019)	Three-stream 3D CNN: depth, depth motion, skeleton	-	SVM	Score fusion: multiply	UTD-MHAD UTKinect... MSR-Action3D	95.34 97.29 94.15
Dependent Streams	A deep-learning-based multimodal depth-aware dynamic hand gesture recognition system (Mahmud et al 2021)	Two-stream CNN & CNN+LSTM	-	FC	Feature fusion (concatenate)-score fusion (Max, Avg)	SHREC-DHG-14/28	89.61
Dependent Streams	Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures (Kamel et al 2018)	Three-stream CNN	-	FC	Feature-level (concat), score-level (Max, product)	MSR... UTD-MAHD MAD	94.51 cs 88.14 cs 91.86 cross validation
Ensemble	Learning and refining of privileged information-based RNNs for action recognition from depth sequences (Shi and Kim 2017)	3 dependent CNN and RNN stage skeleton guided which is not available in testing	VGG-VeryDepp	Expectati maximization	Attention-based	MSR-Act... SBU Kinect CAD-60 Blanket (their dataset)	94.9 89.2 87.6 53.5

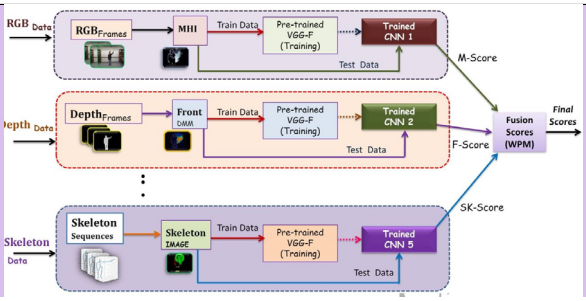
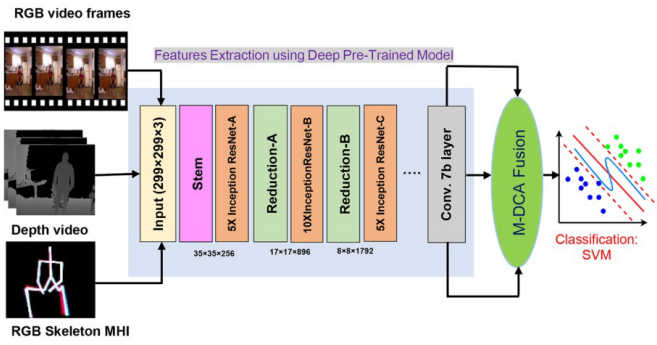
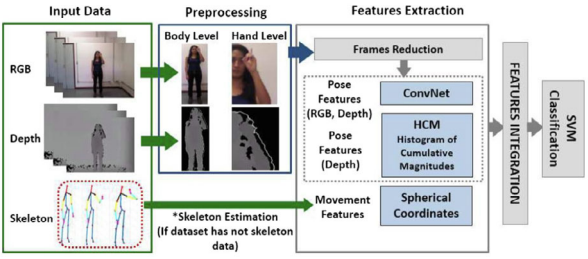
	Paper	Architecture
Independent Streams	Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition (Rani et al 2021)	
Independent Streams	Structured images for RGB-D action recognition (Wang et al 2017b)	

Learning Action Recognition Model From Depth and Skeleton Videos (Rahmani and Bennamoun 2017)		
3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track (De Smedt et al 2017)	None	
CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition (Lai and Yanushkevich 2018)		
3D-based Deep Convolutional Neural Network for action recognition with depth sequences (Liu et al 2016b)		
3d behavior recognition based on multi-modal deep space-time learning (Zhao et al 2019)		

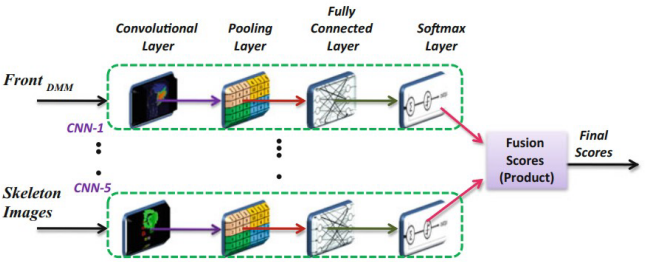
Dependent Streams	A deep-learning-based multimodal depth-aware dynamic hand gesture recognition system (Mahmud et al 2021)	
	Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures (Kamel et al 2018)	
Ensemble Methods	Learning and refining of privileged information-based RNNs for action recognition from depth sequences (Shi and Kim 2017)	

A.4 RGB and depth and skeleton

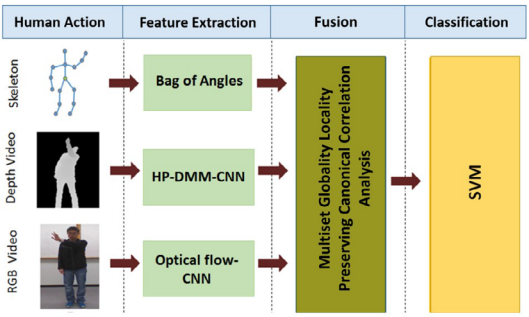
	Paper	Architecture	Pre-train	Classifier	Fusion	Dataset	Accuracy
Independent Streams	First-person hand action benchmark with rgb-d videos and 3d hand pose annotations (Garcia-Hernando et al 2018)	18 baseline architectures				Their dataset	
	Combining CNN streams of RGB-D and skeletal data for human activity recognition (Khaire et al 2018b)	Five-stream CNN	VGG-F VGG-16	FC	Score fusion: Weighted Product Model	CAD-60 SBU Kinect UTD-MHAD	93.06,90.0 96.67 95.11
	A deep multi-modal network based on bottleneck layer features fusion for action recognition (Singh and Vishwakarma 2021)	Three-stream CNN	VGG 19, Inception-v1, ResNet, Inception ResNet-v2	SVM	M-DCA	UT-Kinect CAD-60 Florence3D SBU Kinect	99.0 98.66 97.20 97.75
	Multimodal hand gesture recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes (Cardenas and Chavez 2020)	Two-stream CNN + other descriptors	Vgg-f	SVM	Feature level: Max, mean, concatenate Or max, max-mean, max-min	UTD–MHAD IsoGD UFOP–LIBRAS	94.81 67.36 64.33
	Human activity recognition by fusion of rgb, depth, and skeletal data (Khaire et al 2018a)	Four-stream CNN	VGG-16	SVM	Score fusion : product	UTD-MHAD	95.38
	Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis (Elmadany et al 2018)	Two-stream CNN + an other stream	VGG-f on ImageNet and UCF-101	SVM	Multiset Globality Locality Preserving Canonical Correlation Analysis	MSR-Action3D UTD-MHAD MAD KARD SBU Kinect	96.92 96.63 100 100
	Multimodal human action recognition based on a fusion of dynamic images using cnn descriptors (Cardenas and Chavez 2018)	Five-stream CNN: color dynamic, depth dynamic, DXY, DVZ, DXZ	use a pre-trained flow-CNN architecture	SVM	Feature-level : max-mean aggregation	UTD-MHAD	94.57
	Vision-Based Multi-Modal Framework for Action Recognition (Romaissa et al 2021)	Three-stream CNN	ResNet-50, AlexNet	LSTM	Feature level: Canonical Correlation Analysis (CCA)	UTD-MHAD NTU RGB+D	98.88 75.50
Dependent	Deep bilinear learning for rgb-d action recognition (Hu et al 2018)	fivestream CNN	-	FC	Bilinear learning	NTU RGB+D SYSU 3D HOI	85.4cs,90.7cv 84.8setting1 88.9setting2
Single	Deep dynamic neural networks for multimodal gesture segmentation and recognition (Wu et al 2016)	3DCNN+ DBN 2D entity	-	FC	Late (weighted sum of scores) and intermediate (concatenate features) fusion	ChaLearn Lap	86.04
Ensemble	Graph distillation for action detection with privileged modalities (Luo et al 2018)	Deep network + graph	ResNet-18 with NTU and action recognition task for action detection	FC	One modality in testing	NTU RGB+D PKU-MMD	As source As target

	Paper	Architecture
	First-person hand action benchmark with rgb-d videos and 3d hand pose annotations (Garcia-Hernando et al 2018)	None
Independent Streams	Combining CNN streams of RGB-D and skeletal data for human activity recognition (Khaire et al 2018b)	
	A deep multimodal network based on bottleneck layer features fusion for action recognition (Singh and Vishwakarma 2021)	
	Multimodal hand gesture recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes (Cardenas and Chavez 2020)	

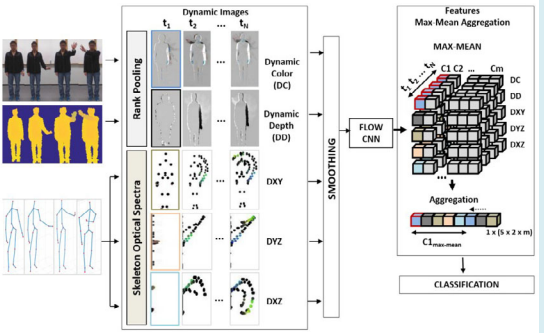
Human activity recognition by fusion of rgb, depth, and skeletal data (Khaire et al 2018a)



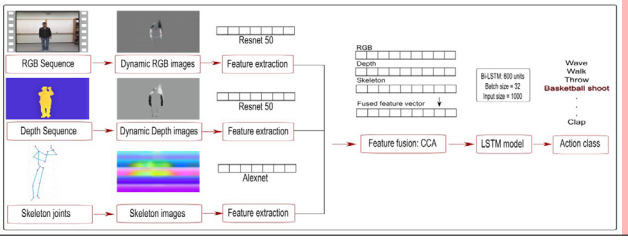
Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis (Elmadany et al 2018)

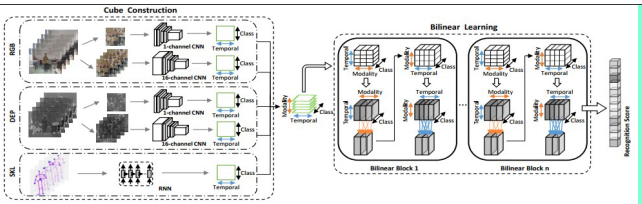
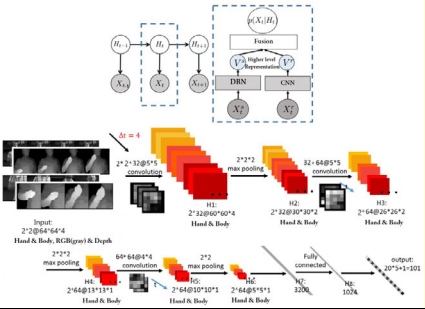
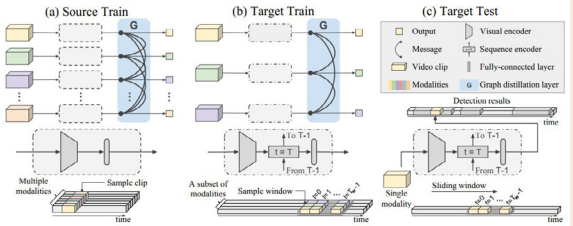


Multimodal human action recognition based on a fusion of dynamic images using cnn descriptors (Cardenas and Chavez 2018)



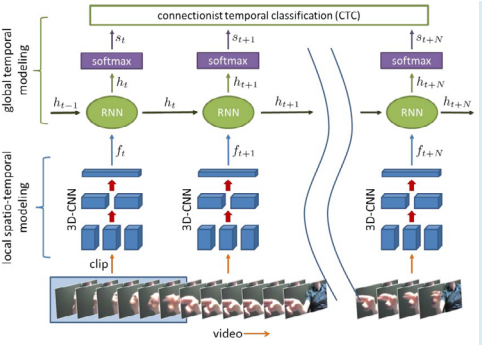
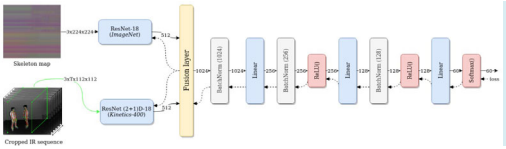
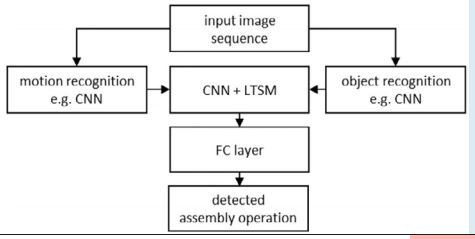
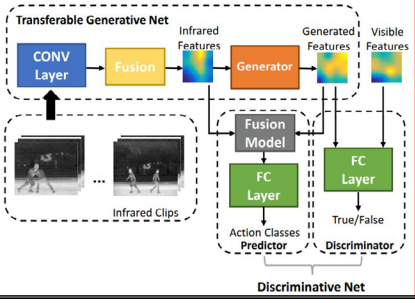
Vision-Based Multi-Modal Framework for Action Recognition (Romaissa et al 2021)



Dependent Streams	Deep bilinear learning for rgb-d action recognition (Hu et al 2018)	
Single-Streams	Deep dynamic neural networks for multimodal gesture segmentation and recognition (Wu et al 2016)	
Ensemble Methods	Graph distillation for action detection with privileged modalities (Luo et al 2018)	

A.5 Infrared and others

	Paper	Architecture	Pre-train	Classifier	Fusion	Dataset	Accuracy
Independent Streams	Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network (Molchanov et al 2016)	Five-stream CNN and RNN: depth, optical flow, RGB, IR, IR_disparity	VGG-Net, C3D on Sports-1M	SVM	average class conditional probebility-late	SKIG ChaLearn2014	98.6 98.2
	Infrared and 3d skeleton feature fusion for rgb-d action recognition (De Boissiere and Noumeir 2020)	IR + skeleton: 3D and 2D CNN	ResNet-18 on ImageNet and ResNet (2+1)D-18 on Kinetics-400	FC	Feature-level (concatenate)	NTU RGB+D	91.8-cs 94.9-cv
	Classification of assembly operations using machine learning algorithms based on visual sensor data (Rückert et al 2021)	CNN + LSTM : IR, RGB, depth	-	FC	Feature-level	Their dataset	54.4
Hallucination	Pm-gans: Discriminative representation learning for action recognition using partial-modalities (Wang et al 2018a)	IR + RGB: CNN Generative and discriminative (GANs)	-	SVM	Feature-level (concatenate)	Their dataset	82.78

	Paper	Architecture
Independent Streams	Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network (Molchanov et al 2016)	
	Infrared and 3d skeleton feature fusion for rgb-d action recognition (De Boissiere and Noumeir 2020)	
	Classification of assembly operations using machine learning algorithms based on visual sensor data (Rückert et al 2021)	
Hallucination Network	Pm-gans: Discriminative representation learning for action recognition using partial-modalities (Wang et al 2018a)	

References

Adewopo V, Elsayed N, ElSayed Z, etal (2022) Review on action recognition for accident detection in smart city transportation systems. arXiv preprint arXiv:2208.09588https://doi.org/10.48550/arXiv.2208.09588

Adhikari K, Bouchachia H, Nait-Charif H (2017) Activity recognition for indoor fall detection using convolutional neural network. In: 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), IEEE, pp 81–84, https://doi.org/10.23919/mva.2017.7986795

Ahmad T, Jin L, Zhang X et al (2021) Graph convolutional neural network for human action recognition: a comprehensive survey. IEEE Trans Artif Intell 2(2):128–145. https://doi.org/10.1109/tai.2021.3076974

Ahn D, Kim S, Ko BC (2023) Star++: rethinking spatio-temporal cross attention transformer for video action recognition. Appl Intell. https://doi.org/10.1007/s10489-023-04978-7

Akkaladevi SC, Heindl C (2015) Action recognition for human robot interaction in industrial applications. In: 2015 IEEE International Conference on Computer Graphics Vision and Information Security (CGVIS). IEEE, pp 94–99, https://doi.org/10.1109/cgvis.2015.7449900

Al-Faris M, Chiverton J, Ndzi D et al (2020) A review on computer vision-based methods for human action recognition. J Imag 6(6):46. https://doi.org/10.3390/jimaging6060046

Al-Faris M, Chiverton JP, Yang Y et al (2020) Multi-view region-adaptive multi-temporal dmm and rgb action recognition. Pattern Anal Appl 23:1587–1602. https://doi.org/10.1007/s10044-020-00886-5

Alahi A, Goel K, Ramanathan V, etal (2016) Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 961–971, https://doi.org/10.1109/cvpr.2016.110

Alayrac JB, Donahue J, Luc P, etal, (2022) Flamingo: a visual language model for few-shot learning. Adv Neural Inf Process Syst 35:23716–23736

Alom MZ, Taha TM, Yakopcic C et al (2019) A state-of-the-art survey on deep learning theory and architectures. Electronics 8(3):292. https://doi.org/10.3390/electronics8030292

Amiri SM, Pourazad MT, Nasiopoulos P, etal (2013) Non-intrusive human activity monitoring in a smart home environment. In: 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013), IEEE, pp 606–610, https://doi.org/10.1109/healthcom.2013.6720748

Amiri SM, Pourazad M, Nasiopoulos P et al (2014) Improved human action recognition in a smart home environment setting. IRBM 35(6):321–328. https://doi.org/10.1016/j.irbm.2014.10.005

Asadi-Aghbolaghi M, Bertiche H, Roig V, etal (2017) Action recognition from rgb-d data: Comparison and fusion of spatio-temporal handcrafted features and deep strategies. In: Proceedings of the IEEE International conference on computer vision workshops, pp 3179–3188, https://doi.org/10.1109/iccvw.2017.376

Azar SG, Seyedarabi H (2020) Trajectory-based recognition of dynamic persian sign language using hidden markov model. Comput Speech Lang 61(101):053. https://doi.org/10.1016/j.csl.2019.101053

Baltrušaitis T, Ahuja C, Morency LP (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443. https://doi.org/10.1109/tpami.2018.2798607

Baradel F, Wolf C, Mille J (2017a) Human action recognition: Pose-based attention draws focus to hands. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 604–613, https://doi.org/10.1109/iccvw.2017.77

Baradel F, Wolf C, Mille J (2017b) Pose-conditioned spatio-temporal attention for human action recognition. arXiv preprint arXiv:1703.10106https://doi.org/10.48550/arXiv.1703.10106

Baradel F, Wolf C, Mille J (2018) Human activity recognition with pose-driven attention to rgb. In: BMVC 2018-29th British Machine Vision Conference, pp 1–14

Beddiar DR, Nini B, Sabokrou M et al (2020) Vision-based human activity recognition: a survey. Multimedia Tools Appl 79(3509–30):555. https://doi.org/10.1007/s11042-020-09004-3

Bi L, Feng D, Kim J (2018) Dual-path adversarial learning for fully convolutional network (fcn)-based medical image segmentation. Vis Comput 34:1043–1052. https://doi.org/10.1007/s00371-018-1519-5

Bilen H, Fernando B, Gavves E, etal (2016) Dynamic image networks for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3034–3042, https://doi.org/10.1109/cvpr.2016.331

Bini S, Greco A, Saggese A, etal (2022) Benchmarking deep neural networks for gesture recognition on embedded devices. In: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, pp 1285–1290, https://doi.org/10.1109/RO-MAN53752.2022.9900705

Bloom V, Makris D, Argyriou V (2012) G3d: A gaming action dataset and real time action recognition evaluation framework. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 7–12, https://doi.org/10.1109/cvprw.2012.6239175

Bloom V, Argyriou V, Makris D (2015) G3di: A gaming interaction dataset with a real time detection and evaluation framework. In: Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13, Springer, pp 698–712, https://doi.org/10.1007/978-3-319-16178-5_49

Brown T, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neural Inf Proces Syst 33:1877–1901

Bruce X, Liu Y, Zhang X et al (2022) Mmnet: a model-based multimodal network for human action recognition in rgb-d videos. IEEE Trans Pattern Anal Mach Intell 45(3):3522–3538. https://doi.org/10.1109/TPAMI.2022.3177813

Cai J, Jiang N, Han X, etal (2021) Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2735–2744, https://doi.org/10.1109/wacv48630.2021.00278

Cai Z, Han J, Liu L et al (2017) Rgb-d datasets using microsoft kinect or similar sensors: a survey. Multimedia Tools Appl 76:4313–4355. https://doi.org/10.1007/s11042-016-3374-6
Camplani M, Salgado L (2014) Background foreground segmentation with rgb-d kinect data: an efficient combination of classifiers. J Vis Commun Image Represent 25(1):122–136. https://doi.org/10.1016/j.jvcir.2013.03.009
Cardenas EE, Chavez GC (2018) Multimodal human action recognition based on a fusion of dynamic images using cnn descriptors. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, pp 95–102, https://doi.org/10.1109/sibgrapi.2018.00019
Cardenas EJE, Chavez GC (2020) Multimodal hand gesture recognition combining temporal and pose information based on cnn descriptors and histogram of cumulative magnitudes. J Vis Commun Image Represent 71(102):772. https://doi.org/10.1016/j.jvcir.2020.102772
Carletti V, Foggia P, Percannella G, etal (2013) Recognition of human actions from rgb-d videos using a reject option. In: New Trends in Image Analysis and Processing–ICIAP 2013: ICIAP 2013 International Workshops, Naples, Italy, September 9-13, 2013. Proceedings 17, Springer, pp 436–445, https://doi.org/10.1007/978-3-642-41190-8_47
Carnegie mellon university (2001) cmu graphics lab—motion capture library. http://mocap.cs.cmu.edu/. Accessed 13 Nov 2021
Chai X, Liu Z, Yin F, etal (2016) Two streams recurrent neural networks for large-scale continuous gesture recognition. In: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, pp 31–36, https://doi.org/10.1109/icpr.2016.7899603
Chen C, Jafari R, Kehtarnavaz N (2015) Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, pp 168–172, https://doi.org/10.1109/icip.2015.7350781
Chen H, Li Y, Fang H et al (2022) Multi-scale attention 3d convolutional network for multimodal gesture recognition. Sensors 22(6):2405. https://doi.org/10.3390/s22062405
Chen K, Zhang D, Yao L et al (2021) Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities. ACM Computing Surveys (CSUR) 54(4):1–40. https://doi.org/10.1145/3447744
Chen L, Wei H, Ferryman J (2014) Readingact rgb-d action dataset and human action recognition from local features. Pattern Recogn Lett 50:159–169. https://doi.org/10.1016/j.patrec.2013.09.004
Chen T, Mo L (2023) Swin-fusion: swin-transformer with feature fusion for human action recognition. Neural Process Lett. https://doi.org/10.1007/s11063-023-11367-1
Cheng J, Ren Z, Zhang Q et al (2021) Cross-modality compensation convolutional neural networks for rgb-d action recognition. IEEE Trans Circuits Syst Video Technol 32(3):1498–1509. https://doi.org/10.1109/tcsvt.2021.3076165
Cheng Q, Liu Z, Ren Z, etal, (2022) Spatial-temporal information aggregation and cross-modality interactive learning for rgb-d-based human action recognition. IEEE Access 10:104190–104201. https://doi.org/10.1109/ACCESS.2022.3201227
Cheng Z, Qin L, Ye Y, etal (2012) Human daily action analysis with multi-view and color-depth data. In: Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7-13, 2012, Proceedings, Part II 12, Springer, pp 52–61, https://doi.org/10.1007/978-3-642-33868-7_6
Choi J, Yi C, Han T et al (2008) A view-based real-time human action recognition system as an interface for human computer interaction. Lect Notes Comput Sci 4820:112–120. https://doi.org/10.1007/978-3-540-78566-8_10
Dai R, Das S, Sharma S et al (2022) Toyota smarthome untrimmed: real-world untrimmed videos for activity detection. IEEE Trans Pattern Anal Mach Intell 45(2):2533–2550. https://doi.org/10.1109/tpami.2022.3169976
Dang LM, Min K, Wang H et al (2020) Sensor-based and vision-based human activity recognition: a comprehensive survey. Pattern Recogn 108(107):561. https://doi.org/10.1016/j.patcog.2020.107561
Das S, Koperski M, Bremond F, etal (2017) Action recognition based on a mixture of rgb and depth based skeleton. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp 1–6, https://doi.org/10.1109/AVSS.2017.8078548
Das S, Chaudhary A, Bremond F, etal (2019a) Where to focus on for human action recognition? In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 71–80, https://doi.org/10.1109/wacv.2019.00015
Das S, Dai R, Koperski M, etal (2019b) Toyota smarthome: Real-world activities of daily living. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 833–842, https://doi.org/10.1109/iccv.2019.00092
Das S, Thonnat M, Sakhalkar K, etal (2019c) A new hybrid architecture for human activity recognition from rgb-d videos. In: MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki,

Greece, January 8–11, 2019, Proceedings, Part II 25, Springer, pp 493–505, https://doi.org/10.1007/978-3-030-05716-9_40
Das S, Sharma S, Dai R, etal (2020) Vpn: Learning video-pose embedding for activities of daily living. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer, pp 72–90, https://doi.org/10.1007/978-3-030-58545-7_5
Das S, Dai R, Yang D et al (2021) Vpn++: rethinking video-pose embeddings for understanding activities of daily living. IEEE Trans Pattern Anal Mach Intell 44(12):9703–9717. https://doi.org/10.1109/tpami.2021.3127885
Davoodikakhki M, Yin K (2020) Hierarchical action classification with network pruning. In: Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part I 15. p 291–305, https://doi.org/10.1007/978-3-030-64556-4_23
Dawar N, Kehtarnavaz N (2018) Real-time continuous detection and recognition of subject-specific smart tv gestures via fusion of depth and inertial sensing. IEEE Access 6:7019–7028. https://doi.org/10.1109/access.2017.2788558
DeBoissiere AM, Noumeir R (2020) Infrared and 3d skeleton feature fusion for rgb-d action recognition. IEEE Access 8:168297–168308. https://doi.org/10.1109/access.2020.3023599
DeSmedt Q, Wannous H, Vandeborre JP, etal (2017) 3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track. In: Proceedings of the Workshop on 3D Object Retrieval, pp 33–38, https://doi.org/10.2312/3dor.20171049
Debnath B, O'Brient M, Kumar S, etal (2021) Attention-driven body pose encoding for human activity recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, pp 5897–5904, https://doi.org/10.1109/icpr48806.2021.9412487
Dhiman C, Vishwakarma DK (2019) A review of state-of-the-art techniques for abnormal human activity recognition. Eng Appl Artif Intell 77:21–45. https://doi.org/10.1016/j.engappai.2018.08.014
Dhiman C, Vishwakarma DK (2020) View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Trans Image Process 29:3835–3844. https://doi.org/10.1109/tip.2020.2965299
Do J, Kim M (2022) Multi-modal transformer for indoor human action recognition. In: 2022 22nd International Conference on Control, Automation and Systems (ICCAS), IEEE, pp 1155–1160, https://doi.org/10.23919/ICCAS55662.2022.10003914
Duan H, Zhao Y, Chen K, etal (2022) Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2969–2978, https://doi.org/10.1109/cvpr52688.2022.00298
Duan J, Zhou S, Wan J, etal (2016) Multi-modality fusion based on consensus-voting and 3d convolution for isolated gesture recognition. arXiv preprint arXiv:1611.06689https://doi.org/10.48550/arXiv.1611.06689
EK S, Portet F, Lalanda P (2022) Lightweight transformers for human activity recognition on mobile devices. arXiv preprint arXiv:2209.11750https://doi.org/10.48550/arXiv.2209.11750
Elboushaki A, Hannane R, Afdel K et al (2020) Multid-cnn: a multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in rgb-d image sequences. Expert Syst Appl 139(112):829. https://doi.org/10.1016/j.eswa.2019.112829
Elforaici MEA, Chaaraoui I, Bouachir W, etal (2018) Posture recognition using an rgb-d camera: exploring 3d body modeling and deep learning approaches. In: 2018 IEEE Life Sciences Conference (LSC), IEEE, pp 69–72, https://doi.org/10.1109/lsc.2018.8572079
Elmadany NED, He Y, Guan L (2018) Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis. IEEE Trans Image Process 27(11):5275–5287. https://doi.org/10.1109/tip.2018.2855438
Escalera S, Baró X, Gonzalez J, etal (2015) Chalearn looking at people challenge 2014: Dataset and results. In: Computer Vision-ECCV 2014 Workshop, PT I, Springer, Cham, pp 459–473, https://doi.org/10.1007/978-3-319-16178-5_32
Estevam V, Pedrini H, Menotti D (2021) Zero-shot action recognition in videos: a survey. Neurocomputing 439:159–175. https://doi.org/10.1016/j.neucom.2021.01.036
Feichtenhofer C, Fan H, Malik J, etal (2019) Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6202–6211, https://doi.org/10.1109/iccv.2019.00630
Fu B, Damer N, Kirchbuchner F, etal, (2020) Sensing technology for human activity recognition: a comprehensive survey. IEEE Access 8:83791–83820. https://doi.org/10.1109/access.2020.2991891
Fu TJ, Wang XE, Wang WY (2022) Language-driven artistic style transfer. In: European Conference on Computer Vision, Springer, pp 717–734, https://doi.org/10.1007/978-3-031-20059-5_41

Gan L, Cao R, Li N et al (2023) Focal channel knowledge distillation for multi-modality action recognition. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3298647

Garcia NC, Morerio P, Murino V (2018) Modality distillation with multiple stream networks for action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 103–118, https://doi.org/10.1007/978-3-030-01237-3_7

Garcia NC, Morerio P, Murino V (2019) Learning with privileged information via adversarial discriminative modality distillation. IEEE Trans Pattern Anal Mach Intell 42(10):2581–2593. https://doi.org/10.1109/tpami.2019.2929038

Garcia NC, Bargal SA, Ablavsky V, etal (2021) Distillation multiple choice learning for multimodal action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2755–2764, https://doi.org/10.1109/wacv48630.2021.00280

Garcia-Hernando G, Yuan S, Baek S, etal (2018) First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 409–419, https://doi.org/10.1109/cvpr.2018.00050

Gavrilyuk K, Sanford R, Javan M, etal (2020) Actor-transformers for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 839–848, https://doi.org/10.1109/cvpr42600.2020.00092

Girdhar R, Carreira J, Doersch C, etal (2019) Video action transformer network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 244–253, https://doi.org/10.1109/cvpr.2019.00033

Girshick R, Donahue J, Darrell T, etal (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEEConference on Computer Vision and Pattern Recognition, pp 580–587, https://doi.org/10.1109/cvpr.2014.81

Gonzalez-Jorge H, Riveiro B, Vazquez-Fernandez E et al (2013) Metrological evaluation of microsoft kinect and asus xtion sensors. Measurement 46(6):1800–1806. https://doi.org/10.1016/j.measurement.2013.01.011

Gou J, Yu B, Maybank SJ et al (2021) Knowledge distillation: a survey. Int J Comput Vis 129:1789–1819. https://doi.org/10.1007/s11263-021-01453-z

Guo T, Liu H, Chen Z, etal (2022) Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 762–770, https://doi.org/10.1609/aaai.v36i1.19957

Guyon I, Athitsos V, Jangyodsuk P, etal (2013) Results and analysis of the chalearn gesture challenge 2012. In: Advances in Depth Image Analysis and Applications: International Workshop, WDIA 2012, Tsukuba, Japan, November 11, 2012, Revised Selected and Invited Papers, Springer, pp 186–204, https://doi.org/10.1007/978-3-642-40303-3_19

Hampiholi B, Jarvers C, Mader W, etal, (2023) Convolutional transformer fusion blocks for multi-modal gesture recognition. IEEE Access 11:34094–34103. https://doi.org/10.1109/ACCESS.2023.3263812

Han K, Wang Y, Chen H et al (2022) A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 45(1):87–110. https://doi.org/10.1109/TPAMI.2022.3152247

Harris JD, Quatman CE, Manring MM et al (2014) How to write a systematic review. Am J Sports Med 42(11):2761–2768. https://doi.org/10.1177/0363546513497567

He K, Zhang X, Ren S, et al (2016a) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778, https://doi.org/10.1109/cvpr.2016.90

He L, Li J, Plaza A et al (2016) Discriminative low-rank gabor filtering for spectral spatial hyperspectral image classification. IEEE Trans Geosci Remote Sens 55(3):1381–1395. https://doi.org/10.1109/tgrs.2016.2623742

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffman J, Gupta S, Darrell T (2016) Learning with side information through modality hallucination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 826–834, https://doi.org/10.1109/cvpr.2016.96

Hu JF, Zheng WS, Lai J, etal (2015) Jointly learning heterogeneous features for rgb-d activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5344–5352, https://doi.org/10.1109/cvpr.2015.7299172

Hu JF, Zheng WS, Pan J, etal (2018) Deep bilinear learning for rgb-d action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 335–351, https://doi.org/10.1007/978-3-030-01234-2_21

Huang D, Yao S, Wang Y, etal (2014) Sequential max-margin event detectors. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13, Springer, pp 410–424, https://doi.org/10.1007/978-3-319-10578-9_27

Hussain Z, Sheng QZ, Zhang WE (2020) A review and categorization of techniques on device-free human activity recognition. J Netw Comput Appl 167(102):738. https://doi.org/10.1016/j.jnca.2020.102738

Ijjina EP, Chalavadi KM (2017) Human action recognition in rgb-d videos using motion sequence information and deep learning. Pattern Recogn 72:504–516. https://doi.org/10.1016/j.patcog.2017.07.013

Imran J, Kumar P (2016) Human action recognition using rgb-d sensor and deep convolutional neural networks. In: 2016 International Conference on Advances in Computing Communications and Informatics (ICACCI), IEEE, pp 144–148, https://doi.org/10.1109/icacci.2016.7732038

Islam MM, Nooruddin S, Karray F, etal, (2022) Human activity recognition using tools of convolutional neural networks: a state of the art review, data sets, challenges, and future prospects. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2022.106060

Jaegle A, Borgeaud S, Alayrac JB, etal (2021a) Perceiver io: a general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795https://doi.org/10.48550/arXiv.2107.14795

Jaegle A, Gimeno F, Brock A, etal (2021b) Perceiver: general perception with iterative attention. In: International Conference on Machine Learning, PMLR, pp 4651–4664

Jain A, Nandakumar K, Ross A (2005) Score normalization in multimodal biometric systems. Pattern Recogn 38(12):2270–2285. https://doi.org/10.1016/j.patcog.2005.01.012

Jang J, Kim D, Park C, etal (2020) Etri-activity3d: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 10,990–10,997, https://doi.org/10.1109/IROS45743.2020.9341160

Jegham I, BenKhalifa A, Alouani I, etal (2019) Mdad: A multimodal and multiview in-vehicle driver action dataset. In: Computer Analysis of Images and Patterns: 18th International Conference, CAIP 2019, Salerno, Italy, September 3–5, 2019, Proceedings, Part I 18, Springer, pp 518–529, https://doi.org/10.1007/978-3-030-29888-3_42

Jegham I, Khalifa AB, Alouani I et al (2020a) Soft spatial attention-based multimodal driver action recognition using deep learning. IEEE Sens J 21(2):1918–1925. https://doi.org/10.1109/jsen.2020.3019258

Jegham I, Khalifa AB, Alouani I et al (2020b) Vision-based human action recognition: an overview and real world challenges. Forensic Sci Int: Digit Invest 32(200):901. https://doi.org/10.1016/j.fsidi.2019.200901

Ji S, Xu W, Yang M et al (2012) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231. https://doi.org/10.1109/TPAMI.2012.59

Jiang Y, Chang S, Wang Z (2021) Transgan: two pure transformers can make one strong gan, and that can scale up. Adv Neural Inf Process Syst 34:14745–14758

Jones S, Shao L (2013) Content-based retrieval of human actions from realistic video databases. Inform Sci 236:56–65. https://doi.org/10.1016/j.ins.2013.02.018

Joze HRV, Shaban A, Iuzzolino ML, etal (2020) Mmtm: Multimodal transfer module for cnn fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13289–13299, https://doi.org/10.1109/cvpr42600.2020.01330

Kamel A, Sheng B, Yang P et al (2018) Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Trans Syst Man Cybern: Syst 49(9):1806–1819. https://doi.org/10.1109/tsmc.2018.2850149

Karpathy A, Toderici G, Shetty S, etal (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 1725–1732, https://doi.org/10.1109/cvpr.2014.223

Kay W, Carreira J, Simonyan K, etal (2017) The kinetics human action video dataset. arXiv preprint arXiv:1705.06950https://doi.org/10.48550/arXiv.1705.06950

Kazmi W, Foix S, Alenyà G et al (2014) Indoor and outdoor depth imaging of leaves with time-of-flight and stereo vision sensors: analysis and comparison. ISPRS J Photogramm Remote Sens 88:128–146. https://doi.org/10.1016/j.isprsjprs.2013.11.012

Keselman L, IselinWoodfill J, Grunnet-Jepsen A, etal (2017) Intel realsense stereoscopic depth cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 1–10, https://doi.org/10.1109/cvprw.2017.167

Khaire P, Imran J, Kumar P (2018a) Human activity recognition by fusion of rgb, depth, and skeletal data. In: Proceedings of 2nd International Conference on Computer Vision & Image Processing: CVIP 2017, Volume 1, Springer, pp 409–421, https://doi.org/10.1007/978-981-10-7895-8_32

Khaire P, Kumar P, Imran J (2018) Combining cnn streams of rgb-d and skeletal data for human activity recognition. Pattern Recogn Lett 115:107–116. https://doi.org/10.1016/j.patrec.2018.04.035

Khan NS, Ghani MS (2021) A survey of deep learning based models for human activity recognition. Wireless Pers Commun 120(2):1593–1635. https://doi.org/10.1007/s11277-021-08525-w
Khan S, Naseer M, Hayat M et al (2022) Transformers in vision: a survey. ACM Computing Surveys (CSUR) 54(10s):1–41. https://doi.org/10.1145/3505244
Koppula HS, Gupta R, Saxena A (2013) Learning human activities and object affordances from rgb-d videos. Int J Robot Res 32(8):951–970. https://doi.org/10.1177/0278364913478446
Köpüklü O, Hörmann S, Herzog F et al (2022) Dissected 3d cnns: temporal skip connections for efficient online video processing. Comput Vis Image Underst 215(103):318. https://doi.org/10.1016/j.cviu.2021.103318
Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90. https://doi.org/10.1145/3065386
Kwak NJ, Song TS (2013) Human action classification and unusual action recognition algorithm for intelligent surveillance system. IT Converg Secur 2012:797–804. https://doi.org/10.1007/978-94-007-5860-5_95
Kwolek B, Kepski M (2014) Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput Methods Programs Biomed 117(3):489–501. https://doi.org/10.1016/j.cmpb.2014.09.005
Lai K, Yanushkevich SN (2018) Cnn+ rnn depth and skeleton based dynamic hand gesture recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, pp 3451–3456, https://doi.org/10.1109/icpr.2018.8545718
Lee S, Woo S, Park Y, et al (2023) Modality mixer for multi-modal action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 3298–3307
Li C, Hou Y, Li W et al (2023) Dfn: a deep fusion network for flexible single and multi-modal action recognition. Available at SSRN 4341613. https://doi.org/10.2139/ssrn.4341613
Li J, Xie X, Pan Q et al (2020) Sgm-net: skeleton-guided multimodal network for action recognition. Pattern Recogn 104(107):356. https://doi.org/10.1016/j.patcog.2020.107356
Li Q, Gravina R, Li Y et al (2020) Multi-user activity recognition: challenges and opportunities. Inf Fusion 63:121–135. https://doi.org/10.1016/j.inffus.2020.06.004
Li R, Wang H (2022) Graph convolutional networks and lstm for first-person multimodal hand action recognition. Mach Vis Appl 33(6):84. https://doi.org/10.1007/s00138-022-01328-4
Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 9–14, https://doi.org/10.1109/cvprw.2010.5543273
Li X, Hou Y, Wang P et al (2021) Trear: transformer-based rgb-d egocentric action recognition. IEEE Trans Cognit Develop Syst 14(1):246–252. https://doi.org/10.1109/TCDS.2020.3048883
Li Y, Lan C, Xing J, etal (2016a) Online human action detection using joint classification-regression recurrent neural networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, Springer, pp 203–220, https://doi.org/10.1007/978-3-319-46478-7_13
Li Y, Miao Q, Tian K, etal (2016b) Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In: 2016 23rd International Conference on Pattern Recognition(ICPR). IEEE, pp 25–30, https://doi.org/10.1016/j.patrec.2017.12.003
Li Y, Miao Q, Qi X et al (2019) A spatiotemporal attention-based resc3d model for large-scale gesture recognition. Mach Vis Appl 30:875–888. https://doi.org/10.1007/s00138-018-0996-x
Li Y, Qi T, Ma Z et al (2023) Seeking a hierarchical prototype for multimodal gesture recognition. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2023.3295811
Liang M, Hu X (2015) Recurrent convolutional neural network for object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3367–3375, https://doi.org/10.1109/cvpr.2015.7298958
Lillo I, Soto A, CarlosNiebles J (2014) Discriminative hierarchical modeling of spatio-temporally composable human activities. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 812–819, https://doi.org/10.1109/cvpr.2014.109
Lin W, Liu X, Zhuang Y et al (2022) Unsupervised video-based action recognition with imagining motion and perceiving appearance. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/TCSVT.2022.3221280
Lin YC, Hu MC, Cheng WH, etal (2012) Human action recognition and retrieval using sole depth information. In: Proceedings of the 20th ACM International Conference on Multimedia, pp 1053–1056
Liu AA, Nie WZ, Su YT et al (2015) Coupled hidden conditional random fields for rgb-d human action recognition. Signal Process 112:74–82. https://doi.org/10.1016/j.sigpro.2014.08.038

Liu B, Cai H, Ju Z et al (2019) Rgb-d sensing based human action and interaction analysis: a survey. Pattern Recogn 94:1–12. https://doi.org/10.1016/j.patcog.2019.05.020
Liu C, Hu Y, Li Y, etal (2017a) Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475https://doi.org/10.48550/arXiv.1703.07475
Liu G, Tian G, Li J et al (2018) Human action recognition using a distributed rgb-depth camera network. IEEE Sens J 18(18):7570–7576. https://doi.org/10.1109/jsen.2018.2859268
Liu G, Qian J, Wen F, etal (2019b) Action recognition based on 3d skeleton and rgb frame fusion. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 258–264, https://doi.org/10.1109/iros40897.2019.8967570
Liu J, Shahroudy A, Xu D, etal (2016a) Spatio-temporal lstm with trust gates for 3d human action recognition. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, pp 816–833, https://doi.org/10.1007/978-3-319-46487-9_50
Liu J, Wang G, Hu P, etal (2017b) Global context-aware attention lstm networks for 3d action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1647–1656, https://doi.org/10.1109/cvpr.2017.391
Liu J, Li Y, Song S et al (2018) Multi-modality multi-task recurrent neural network for online action detection. IEEE Trans Circuits Syst Video Technol 29(9):2667–2682. https://doi.org/10.1109/tcsvt.2018.2799968
Liu J, Shahroudy A, Perez M et al (2019) Ntu rgb+ d 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans Pattern Anal Mach Intell 42(10):2684–2701. https://doi.org/10.1109/tpami.2019.2916873
Liu L, Shao L (2013) Learning discriminative representations from rgb-d video data. In: Twenty-third International Joint Conference on Artificial Intelligence
Liu M, Yuan J (2018) Recognizing human actions as the evolution of pose estimation maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1159–1168, https://doi.org/10.1109/cvpr.2018.00127
Liu Z, Zhang C, Tian Y (2016) 3d-based deep convolutional neural network for action recognition with depth sequences. Image Vis Comput 55:93–100. https://doi.org/10.1016/j.imavis.2016.04.004
Lopez-Paz D, Bottou L, Schölkopf B, etal (2015) Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643https://doi.org/10.48550/arXiv.1511.03643
Luo Z, Hsieh JT, Jiang L, etal (2018) Graph distillation for action detection with privileged modalities. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 166–183, https://doi.org/10.1007/978-3-030-01264-9_11
Luvizon DC, Picard D, Tabia H (2018) 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5137–5146, https://doi.org/10.1109/cvpr.2018.00539
Mahasseni B, Todorovic S (2016) Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3054–3062, https://doi.org/10.1109/cvpr.2016.333
Mahmud H, Morshed MM, Hasan M, etal (2021) A deep learning-based multimodal depth-aware dynamic hand gesture recognition system. arXiv preprint arXiv:2107.02543https://doi.org/10.48550/arXiv.2107.02543
Mahmud H, Morshed MM, Hasan MK (2023) Quantized depth image and skeleton-based multimodal dynamic hand gesture recognition. Vis Computer. https://doi.org/10.1007/s00371-022-02762-1
Majumder S, Kehtarnavaz N (2020) Vision and inertial sensing fusion for human action recognition: a review. IEEE Sens J 21(3):2454–2467. https://doi.org/10.1109/jsen.2020.3022326
Majumder S, Kehtarnavaz N (2021) A review of real-time human action recognition involving vision sensing. Real-Time Image Process Deep Learn 11736:53–64. https://doi.org/10.1117/12.2585680
Mansur A, Makihara Y, Yagi Y (2012) Inverse dynamics for action recognition. IEEE Trans Cybern 43(4):1226–1236. https://doi.org/10.1109/tsmcb.2012.2226879
Mar A, Antar AD, Shahid O (2019) Vision-based action understanding for assistive healthcare: a short review. 2019 CVPR Workshops pp 1–11
Memmesheimer R, Theisen N, Paulus D (2021) Sl-dml: Signal level deep metric learning for multimodal one-shot action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, pp 4573–4580, https://doi.org/10.1109/icpr48806.2021.9413336
Memo A, Zanuttigh P (2018) Head-mounted gesture controlled interface for human-computer interaction. Multimedia Tools Appl 77:27–53. https://doi.org/10.1007/s11042-016-4223-3

Miao Q, Li Y, Ouyang W, etal (2017) Multimodal gesture recognition based on the resc3d network. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 3047–3055, https://doi.org/10.1109/iccvw.2017.360

Mirzadeh SI, Farajtabar M, Li A, etal (2020) Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 5191–5198, https://doi.org/10.1609/aaai.v34i04.5963

Mizera I, Müller CH (2002) Breakdown points of cauchy regression-scale estimators. Statist Probab Lett 57 (1):79–89. https://doi.org/10.1016/s0167-7152(02)00057-3

Moencks M, DeSilva V, Roche J, etal (2019) Adaptive feature processing for robust human activity recognition on a novel multi-modal dataset. arXiv preprint arXiv:1901.02858https://doi.org/10.48550/arXiv.1901.02858

Molchanov P, Yang X, Gupta S, etal (2016) Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4207–4215, https://doi.org/10.1109/cvpr.2016.456

Mukherjee S, Awadallah A (2020) Xtremedistil: multi-stage distillation for massive multilingual models. arXiv preprint arXiv:2004.05686https://doi.org/10.48550/arXiv.2004.05686

Mukherjee S, Anvitha L, Lahari TM (2020) Human activity recognition in rgb-d videos by dynamic images. Multimedia Tools Appl 79(27–28):19787–19801. https://doi.org/10.1007/s11042-020-08747-3

Müller M, Röder T, Clausen M, etal (2007) Mocap database hdm05. Institut für Informatik II, Universität Bonn 2(7)

Munaro M, Ballin G, Michieletto S et al (2013) 3d flow estimation for human action recognition from colored point clouds. Biol Inspir Cognit Arch 5:42–51. https://doi.org/10.1016/j.bica.2013.05.008

Munaro M, Michieletto S, Menegatti E (2013b) An evaluation of 3d motion flow and 3d pose estimation for human action recognition. In: RSS Workshops: RGB-D: Advanced Reasoning with Depth Cameras

Negin F, Özdemir F, Akgül CB, etal (2013) A decision forest based feature selection framework for action recognition from rgb-depth cameras. In: Image Analysis and Recognition: 10th International Conference, ICIAR 2013, Póvoa do Varzim, Portugal, June 26-28, 2013. Proceedings 10, Springer, pp 648–657, https://doi.org/10.1109/siu.2013.6531398

Nguyen B, Coelho Y, Bastos T et al (2021) Trends in human activity recognition with focus on machine learning and power requirements. Mach Learn Appl 5(100):072. https://doi.org/10.1016/j.mlwa.2021.100072

Ni B, Wang G, Moulin P (2011) Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, pp 1147–1153, https://doi.org/10.1109/iccvw.2011.6130379

Obaid F, Babadi A, Yoosofan A (2020) Hand gesture recognition in video sequences using deep convolutional and recurrent neural networks. Appl Comput Syst 25(1):57–61. https://doi.org/10.2478/acss-2020-0007

Ofli F, Chaudhry R, Kurillo G, etal (2013) Berkeley mhad: A comprehensive multimodal human action database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV), IEEE, pp 53–60, https://doi.org/10.1109/wacv.2013.6474999

Ohn-Bar E, Trivedi MM (2014) Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations. IEEE Trans Intell Transp Syst 15(6):2368–2377. https://doi.org/10.1109/tits.2014.2337331

Oreifej O, Liu Z (2013) Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 716–723, https://doi.org/10.1109/cvpr.2013.98

Özyer T, Ak DS, Alhajj R (2021) Human action recognition approaches with video datasets—a survey. Knowl-Based Syst 222(106):995. https://doi.org/10.1016/j.knosys.2021.106995

Pahde F, Ostapenko O, Hnichen PJ, etal (2019) Self-paced adversarial training for multimodal few-shot learning. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 218–226, https://doi.org/10.1109/wacv.2019.00029

Pan Z, Yu W, Yi X et al (2019) Recent progress on generative adversarial networks (gans): a survey. IEEE Access 7:36322–36333. https://doi.org/10.1109/access.2019.2905015

Pareek P, Thakkar A (2021) A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artif Intell Rev 54:2259–2322. https://doi.org/10.1007/s10462-020-09904-8

Park Y, Woo S, Lee S et al (2023) Cross-modal alignment and translation for missing modality action recognition. Available at SSRN 4292746. https://doi.org/10.2139/ssrn.4292746

Patwary MJ, Cao W, Wang XZ et al (2022) Fuzziness based semi-supervised multimodal learning for patient's activity recognition using rgbdt videos. Appl Soft Comput 120(108):655. https://doi.org/10.1016/j.asoc.2022.108655

Pigou L, Van Den Oord A, Dieleman S et al (2018) Beyond temporal pooling: recurrence and temporal convolutions for gesture recognition in video. Int J Comput Vis 126:430–439. https://doi.org/10.1007/s11263-016-0957-7

Prati A, Shan C, Kik W (2019) Sensors, vision and networks: from video surveillance to activity recognition and health monitoring. J Ambient Intell Smart Environ 11(1):5–22. https://doi.org/10.3233/AIS-180510

Qin X, Ge Y, Zhan L, etal (2018) Joint deep learning for rgb-d action recognition. In: 2018 IEEE Visual Communications and Image Processing (VCIP), IEEE, pp 1–6, https://doi.org/10.1109/vcip.2018.8698624

Qin X, Ge Y, Feng J, etal (2019) Two-stream network with 3d common-specific framework for rgb-d action recognition. In: 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE, pp 731–738, https://doi.org/10.1109/smartworld-uic-atc-scalcom-iop-sci.2019.00159

Qin X, Ge Y, Feng J et al (2020) Dtmmn: deep transfer multi-metric network for rgb-d action recognition. Neurocomputing 406:127–134. https://doi.org/10.1016/j.neucom.2020.04.034

Rahate A, Walambe R, Ramanna S et al (2022) Multimodal co-learning: challenges, applications with datasets, recent advances and future directions. Inf Fusion 81:203–239. https://doi.org/10.1016/j.inffus.2021.12.003

Rahmani H, Bennamoun M (2017) Learning action recognition model from depth and skeleton videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp 5832–5841, https://doi.org/10.1109/iccv.2017.621

Rahmani H, Mahmood A, QHuynh D, etal (2014) Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13, Springer, pp 742–757, https://doi.org/10.1007/978-3-319-10605-2_48

Rahmani H, Mahmood A, Huynh D et al (2016) Histogram of oriented principal components for cross-view action recognition. IEEE Trans Pattern Anal Mach Intell 38(12):2430–2443. https://doi.org/10.1109/tpami.2016.2533389

Rajput AS, Raman B, Imran J (2020) Privacy-preserving human action recognition as a remote cloud service using rgb-d sensors and deep cnn. Expert Syst Appl 152(113):349. https://doi.org/10.1016/j.eswa.2020.113349

Ramachandram D, Taylor GW (2017) Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag 34(6):96–108. https://doi.org/10.1109/msp.2017.2738401

Rangasamy K, Asari MA, Rahmad NA et al (2020) Deep learning in sport video analysis: a review. TELKOMNIKA Telecommun Comput Electron Control 18(4):1926–1933

Rani SS, Naidu GA, Shree VU (2021) Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition. Mater Today: Proc 37:3164–3173. https://doi.org/10.1016/j.matpr.2020.09.052

Ren Z, Zhang Q, Cheng J et al (2021) Segment spatial-temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition. Neurocomputing 433:142–153. https://doi.org/10.1016/j.neucom.2020.12.020

Ren Z, Zhang Q, Gao X, etal, (2021b) Multi-modality learning for human action recognition. Multimedia Tools Appl 80:16185–16203. https://doi.org/10.1007/s11042-019-08576-z

Roitberg A, Pollert T, Haurilet M, et al (2019) Analysis of deep fusion strategies for multi-modal gesture recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, https://doi.org/10.1109/cvprw.2019.00029

Romaissa BD, Mourad O, Brahim N (2021) Vision-based multi-modal framework for action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, pp 5859–5866, https://doi.org/10.1109/icpr48806.2021.9412863

Rückert P, Papenberg B, Tracht K (2021) Classification of assembly operations using machine learning algorithms based on visual sensor data. Procedia CIRP 97:110–116. https://doi.org/10.1016/j.procir.2020.05.211

Ruffieux S, Lalanne D, Mugellini E (2013) Chairgest: a challenge for multimodal mid-air gesture recognition for close hci. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp 483–488, https://doi.org/10.1145/2522848.2532590

Sadeghipour A, Morency LP, Kopp S (2012) Gesture-based object recognition using histograms of guiding strokes. In: Proceedings of the British Machine Vision Conference, https://doi.org/10.5244/c.26.44

Salaberria A, Azkune G, de Lacalle OL et al (2023) Image captioning for effective use of language models in knowledge-based visual question answering. Expert Syst Appl 212(118):669. https://doi.org/10.1016/j.eswa.2022.118669

Salehinejad H, Sankar S, Barfett J, etal (2017) Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078https://doi.org/10.48550/arXiv.1801.01078

Schatz KM, Quintanilla E, Vyas S, etal (2020) A recurrent transformer network for novel view action synthesis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, Springer, pp 410–426, https://doi.org/10.1007/978-3-030-58583-9_25

Shabaninia E, Naghsh-Nilchi AR, Kasaei S (2019) A weighting scheme for mining key skeletal joints for human action recognition. Multimedia Tools Appl 78:31319–31345. https://doi.org/10.1007/s11042-019-7740-z

Shabaninia E, Nezamabadi-pour H, Shafizadegan F (2022) Transformers in action recognition: a review on temporal modeling. arXiv preprint arXiv:2302.01921https://doi.org/10.48550/arXiv.2302.01921

Shahroudy A, Liu J, Ng TT, etal (2016) Ntu rgb+ d: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1010–1019, https://doi.org/10.1109/cvpr.2016.115

Shahroudy A, Ng TT, Gong Y et al (2017) Deep multimodal feature analysis for action recognition in rgb+ d videos. IEEE Trans Pattern Anal Mach Intell 40(5):1045–1058. https://doi.org/10.1109/tpami.2017.2691321

Shi Z, Kim TK (2017) Learning and refining of privileged information-based rnns for action recognition from depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3461–3470, https://doi.org/10.1109/cvpr.2017.498

Shotton J, Fitzgibbon A, Cook M, et al (2011) Real-time human pose recognition in parts from single depth images. In: CVPR 2011. IEEE, pp 1926–1933, https://doi.org/10.1109/cvpr.2011.5995316

Sigal L, Balan AO, Black MJ (2010) Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int J Comput Vis 87(1–2):4. https://doi.org/10.1007/s11263-009-0273-6

Simonyan K, Zisserman A (2014a) Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems 27

Simonyan K, Zisserman A (2014b) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556https://doi.org/10.48550/arXiv.1409.1556

Singh A, Chakraborty O, Varshney A, etal (2021) Semi-supervised action recognition with temporal contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10,389–10,399, https://doi.org/10.1109/cvpr46437.2021.01025

Singh R, Khurana R, Aks K et al (2020) Combining cnn streams of dynamic image and depth data for action recognition. Multimedia Syst 26:313–322. https://doi.org/10.1007/s00530-019-00645-5

Singh T, Vishwakarma DK (2019a) Human activity recognition in video benchmarks: a survey. Adv Sign Process Commun: Select Proc ICSC 2018:247–259. https://doi.org/10.1007/978-981-13-2553-3_24

Singh T, Vishwakarma DK (2019b) Video benchmarks of human action datasets: a review. Artif Intell Rev 52:1107–1154. https://doi.org/10.1007/s10462-018-9651-1

Singh T, Vishwakarma DK (2021) A deep multimodal network based on bottleneck layer features fusion for action recognition. Multimedia Tools Appl 80(24):33505–33525. https://doi.org/10.1007/s11042-021-11415-9

Song S, Lan C, Xing J, etal (2018) Skeleton-indexed deep multi-modal feature learning for high performance human action recognition. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6, https://doi.org/10.1109/icme.2018.8486486

Song S, Liu J, Li Y et al (2020) Modality compensation network: cross-modal adaptation for action recognition. IEEE Trans Image Process 29:3957–3969. https://doi.org/10.1109/tip.2020.2967577

Song Y, Tang J, Liu F et al (2014) Body surface context: a new robust feature for action recognition from depth videos. IEEE Trans Circuits Syst Video Technol 24(6):952–964. https://doi.org/10.1109/tcsvt.2014.2302558

Soomro K, Zamir AR (2015) Action recognition in realistic sports videos. In: Computer Vision in Sports. p 181–208, https://doi.org/10.1007/978-3-319-09396-3_9

Spriggs EH, DeLaTorre F, Hebert M (2009) Temporal segmentation and activity classification from first-person sensing. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 17–24, https://doi.org/10.1109/cvprw.2009.5204354

Sreenu G, Durai S (2019) Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data 6(1):1–27. https://doi.org/10.1186/s40537-019-0212-5

Sun Y, Weng Y, Luo B et al (2023) Gesture recognition algorithm based on multi-scale feature fusion in rgb-d images. IET Image Proc 17(4):1280–1290. https://doi.org/10.1049/ipr2.12712
Sun Z, Ke Q, Rahmani H et al (2022) Human action recognition from various data modalities: a review. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/tpami.2022.3183112
Sung J, Ponce C, Selman B, etal (2011) Human activity detection from rgbd images. In: Workshops at the twenty-fifth AAAI Conference on Artificial Intelligence
Tan C, Sun F, Kong T, etal (2018) A survey on deep transfer learning. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27, pp 270–279, https://doi.org/10.1007/978-3-030-01424-7_27
Tang Y, Wang Z, Lu J et al (2018) Multi-stream deep neural networks for rgb-d egocentric action recognition. IEEE Trans Circuits Syst Video Technol 29(10):3001–3015. https://doi.org/10.1109/tcsvt.2018.2875441
Tenorth M, Bandouch J, Beetz M (2009) The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, IEEE, pp 1089–1096, https://doi.org/10.1109/iccvw.2009.5457583
Thoker FM, Gall J (2019) Cross-modal knowledge distillation for action recognition. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, pp 6–10, https://doi.org/10.1109/icip.2019.8802909
Tian J, Cheng W, Sun Y et al (2020) Gesture recognition based on multilevel multimodal feature fusion. J Intell Fuzzy Syst 38(3):2539–2550. https://doi.org/10.3233/jifs-179541
Tölgyessy M, Dekan M, Chovanec L et al (2021) Evaluation of the azure kinect and its comparison to kinect v1 and kinect v2. Sensors 21(2):413. https://doi.org/10.3390/s21020413
Tomas A, Biswas K (2017) Human activity recognition using combined deep architectures. In: 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), IEEE, pp 41–45, https://doi.org/10.1109/siprocess.2017.8124502
Tran D, Bourdev L, Fergus R, etal (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4489–4497, https://doi.org/10.1109/iccv.2015.510
Tran D, Wang H, Torresani L, etal (2018) A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 6450–6459, https://doi.org/10.1109/cvpr.2018.00675
Twinanda AP, Winata P, Gangi A, etal (2016) Multi-stream deep architecture for surgical phase recognition on multi-view rgbd videos. In: Proc. M2CAI Workshop MICCAI, pp 1–8
Ulhaq A, Akhtar N, Pogrebna G, etal (2022) Vision transformers for action recognition: A survey. arXiv preprint arXiv:2209.05700https://doi.org/10.48550/arXiv.2209.05700
Van Amsterdam B, Funke I, Edwards E et al (2022) Gesture recognition in robotic surgery with multimodal attention. IEEE Trans Med Imaging 41(7):1677–1687. https://doi.org/10.1109/TMI.2022.3147640
VanGemeren C, Tan RT, Poppe R, etal (2014) Dyadic interaction detection from pose and flow. In: Human Behavior Understanding: 5th International Workshop, HBU 2014, Zurich, Switzerland, September 12, 2014. Proceedings 5, Springer, pp 101–115, https://doi.org/10.1007/978-3-319-11839-0_9
Vaquette G, Orcesi A, Lucat L, etal (2017) The daily home life activity dataset: a high semantic activity dataset for online recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, pp 497–504, https://doi.org/10.1109/fg.2017.67
Vaswani A, Shazeer N, Parmar N, etal (2017) Attention is all you need. Advances in Neural Information Processing Systems 30
Verma P, Sah A, Srivastava R (2020) Deep learning-based multi-modal approach using rgb and skeleton sequences for human activity recognition. Multimedia Syst 26(6):671–685. https://doi.org/10.1007/s00530-020-00677-2
Wan J, Zhao Y, Zhou S, etal (2016) Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 56–64, https://doi.org/10.1109/cvprw.2016.100
Wang H, Song Z, Li W et al (2020) A hybrid network for large-scale action recognition from rgb and depth modalities. Sensors 20(11):3305. https://doi.org/10.3390/s20113305
Wang J, Liu Z, Wu Y, etal (2012) Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1290–1297, https://doi.org/10.1109/cvpr.2012.6247813
Wang J, Nie X, Xia Y, etal (2014a) Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2649–2656, https://doi.org/10.1109/cvpr.2014.339

Wang K, Wang X, Lin L, etal (2014b) 3d human activity recognition with reconfigurable convolutional neural networks. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp 97–106, https://doi.org/10.1145/2647868.2654912
Wang L, Yoon KJ (2021) Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/tpami.2021.3055564
Wang L, Gao C, Yang L, etal (2018a) Pm-gans: Discriminative representation learning for action recognition using partial-modalities. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 384–401, https://doi.org/10.1007/978-3-030-01231-1_24
Wang L, Ding Z, Tao Z, etal (2019a) Generative multi-view human action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6212–6221, https://doi.org/10.1109/iccv.2019.00631
Wang L, Huynh DQ, Koniusz P (2019) A comparative review of recent kinect-based action recognition algorithms. IEEE Trans Image Process 29:15–28. https://doi.org/10.1109/tip.2019.2925285
Wang P, Li W, Gao Z, etal (2017a) Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 595–604, https://doi.org/10.1109/cvpr.2017.52
Wang P, Wang S, Gao Z, etal (2017b) Structured images for rgb-d action recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 1005–1014, https://doi.org/10.1109/iccvw.2017.123
Wang P, Li W, Ogunbona P et al (2018) Rgb-d-based human motion recognition with deep learning: a survey. Comput Vis Image Underst 171:118–139. https://doi.org/10.1016/j.cviu.2018.04.007
Wang P, Li W, Wan J, etal, (2018c) Cooperative training of deep aggregation networks for rgb-d action recognition. Proc AAAI Conf Artif Intell. https://doi.org/10.1609/aaai.v32i1.12228
Wang Z, She Q, Ward TE (2021) Generative adversarial networks in computer vision: a survey and taxonomy. ACM Computing Surveys (CSUR) 54(2):1–38. https://doi.org/10.1145/3439723
Wang Z, Dong W, Zhang B et al (2023) Gsoanet: Group second-order aggregation network for video action recognition. Neural Process Lett. https://doi.org/10.1007/s11063-023-11270-9
Wei P, Zhao Y, Zheng N, etal (2013) Modeling 4d human-object interactions for event and object recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3272–3279, https://doi.org/10.1109/iccv.2013.406
Weiyao X, Muqing W, Min Z, etal, (2021) Fusion of skeleton and rgb features for rgb-d human action recognition. IEEE Sensors J 21(17):19157–19164. https://doi.org/10.1109/jsen.2021.3089705
Wolf C, Lombardi E, Mille J et al (2014) Evaluation of video activity localizations integrating quality and quantity measurements. Comput Vis Image Underst 127:14–30. https://doi.org/10.1016/j.cviu.2014.06.014
Woo S, Lee S, Park Y, etal (2023) Towards good practices for missing modality robust action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 2776–2784, https://doi.org/10.1609/aaai.v37i3.25378
Wright RW, Brand RA, Dunn W et al (2007) How to write a systematic review. Clin Orthop Relat Res 1976–2007(455):23–29. https://doi.org/10.1097/BLO.0b013e31802c9098
Wu C, Zhang J, Savarese S, etal (2015) Watch-n-patch: Unsupervised understanding of actions and relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4362–4370, https://doi.org/10.1109/cvpr.2015.7299065
Wu D, Pigou L, Kindermans PJ et al (2016) Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE Trans Pattern Anal Mach Intell 38(8):1583–1597. https://doi.org/10.1109/tpami.2016.2537340
Wu H, Ma X, Li Y (2021) Spatiotemporal multimodal learning with 3d cnns for video action recognition. IEEE Trans Circuits Syst Video Technol 32(3):1250–1261. https://doi.org/10.1109/tcsvt.2021.3077512
Wu Z, Pan S, Chen F et al (2020) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24. https://doi.org/10.1109/tnnls.2020.2978386
Xia L, Chen CC, Aggarwal JK (2012) View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops, IEEE, pp 20–27, https://doi.org/10.1109/cvprw.2012.6239233
Xiang W, Li C, Zhou Y, etal (2023) Generative action description prompts for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10,276–10,285
Xu C, Wu X, Li Y et al (2021) Cross-modality online distillation for multi-view action recognition. Neurocomputing 456:384–393. https://doi.org/10.1016/j.neucom.2021.05.077

Xu N, Liu A, Nie W, etal (2015) Multi-modal & multi-view & interactive benchmark dataset for human action recognition. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp 1195–1198, https://doi.org/10.1145/2733373.2806315

Yadav SK, Tiwari K, Pandey HM et al (2021) A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. Chall Fut Direct 223 (106):970. https://doi.org/10.1016/j.knosys.2021.106970

Yang J, Dong X, Liu L, etal (2022) Recurring the transformer for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14,063–14,073

Yao G, Lei T, Zhong J (2019) A review of convolutional-neural-network-based action recognition. Pattern Recogn Lett 118:14–22. https://doi.org/10.1016/j.patrec.2018.05.018

Ye J, Hu A, Xu H, etal (2023) mplug-docowl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499https://doi.org/10.48550/arXiv.2307.02499

You Q, Jiang H (2019) Action4d: Online action recognition in the crowd and clutter. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11,857–11,866, https://doi.org/10.1109/cvpr.2019.01213

Yu G, Liu Z, Yuan J (2015) Discriminative orderlet mining for real-time recognition of human-object interaction. In: Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12, Springer, pp 50–65, https://doi.org/10.1007/978-3-319-16814-2_4

Yuanyuan S, Yunan L, Xiaolong F et al (2021) Review of dynamic gesture recognition. Virt Real Intell Hardw 3(3):183–206. https://doi.org/10.1016/j.vrih.2021.05.001

Yun K, Honorio J, Chattopadhyay D, etal (2012) Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 28–35, https://doi.org/10.1109/cvprw.2012.6239234

Zanuttigh P, Marin G, DalMutto C, etal, (2016) Time-of-flight and structured light depth cameras. Technol Appl. https://doi.org/10.1007/978-3-319-30973-6

Zhang C, Tian Y (2012) Rgb-d camera-based daily living activity recognition. J Comput Vis Image Process 2 (4):12

Zhang H, Li Y, Wang P, etal (2018a) Rgb-d based action recognition with light-weight 3d convolutional networks. arXiv preprint arXiv:1811.09908https://doi.org/10.48550/arXiv.1811.09908

Zhang HB, Zhang YX, Zhong B et al (2019) A comprehensive survey of vision-based human action recognition methods. Sensors 19(5):1005. https://doi.org/10.3390/s19051005

Zhang J, Li W, Ogunbona PO et al (2016) Rgb-d-based action recognition datasets: a survey. Pattern Recogn 60:86–105. https://doi.org/10.1016/j.patcog.2016.05.019

Zhang L, Zhu G, Shen P, etal (2017a) Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 3120–3128, https://doi.org/10.1109/iccvw.2017.369

Zhang S, Liu X, Xiao J (2017b) On geometric features for skeleton-based action recognition using multilayer lstm networks. In: IEEE winter conference on applications of computer vision (CACV), pp 148–157, https://doi.org/10.1109/wacv.2017.24

Zhang Y, Cao C, Cheng J et al (2018) Egogesture: a new dataset and benchmark for egocentric hand gesture recognition. IEEE Trans Multimedia 20(5):1038–1050. https://doi.org/10.1109/tmm.2018.2808769

Zhao B, Wang Y, Su K et al (2021) reading pictures instead of looking: Rgb-d image-based action recognition via capsule network and kalman filter. Sensors 21(6):2217. https://doi.org/10.3390/s21062217

Zhao C, Chen M, Zhao J et al (2019) 3d behavior recognition based on multi-modal deep space-time learning. Appl Sci 9(4):716. https://doi.org/10.3390/app9040716

Zhao R, Ali H, Vander Smagt P (2017) Two-stream rnn/cnn for action recognition in 3d videos. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 4260–4267, https://doi.org/10.1109/iros.2017.8206288

Zhou B, Li Y, Wan J (2021a) Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 3563–3571, https://doi.org/10.1609/aaai.v35i4.16471

Zhou B, Wan J, Liang Y et al (2021) Adaptive cross-fusion learning for multi-modal gesture recognition. Virt Real Intell Hardw 3(3):235–247. https://doi.org/10.1016/j.vrih.2021.05.003

Zhou W, Newsam S, Li C et al (2017) Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. Remote Sensing 9(5):489. https://doi.org/10.3390/rs9050489

Zhu D, Chen J, Shen X, etal (2023) Minigpt-4: enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592https://doi.org/10.48550/arXiv.2304.10592

Zhu G, Zhang L, Mei L, etal (2016) Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp 19–24, https://doi.org/10.1109/icpr.2016.7899601

Zhu G, Zhang L, Shen P et al (2017) Multimodal gesture recognition using 3-d convolution and convolutional lstm. IEEE Access 5:4517–4524. https://doi.org/10.1109/access.2017.2684186

Zhu Y, Li X, Liu C, etal (2020) A comprehensive study of deep video action recognition. arXiv preprint arXiv:2012.06567https://doi.org/10.48550/arXiv.2012.06567

Zhuang F, Qi Z, Duan K et al (2020) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76. https://doi.org/10.1109/jproc.2020.3004555

Zolfaghari M, Oliveira GL, Sedaghat N, etal (2017) Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2904–2913, https://doi.org/10.1109/iccv.2017.316