Machine Vision and Applications (2021) 32:121
https://doi.org/10.1007/s00138-021-01249-8

ORIGINAL PAPER

Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition

Said Yacine Boulahia[1] · Abdenour Amamra[1] · Mohamed Ridha Madi[1] · Said Daikh[1]

Received: 8 January 2021 / Revised: 5 July 2021 / Accepted: 9 September 2021 / Published online: 30 September 2021

Abstract

Multimodal action recognition techniques combine several image modalities (RGB, Depth, Skeleton, and InfraRed) for a more robust recognition. According to the fusion level in the action recognition pipeline, we can distinguish three families of approaches: *early* fusion, where the raw modalities are combined ahead of feature extraction; *intermediate* fusion, the features, respective to each modality, are concatenated before classification; and *late* fusion, where the modality-wise classification results are combined. After reviewing the literature, we identified the principal defects of each category, which we try to address by first investigating more deeply the early-stage fusion that has been poorly explored in the literature. Second, intermediate fusion protocols operate on the feature map, irrespective of the particularity of human action, we propose a new scheme where we optimally combine modality-wise features. Third, as most of the late fusion solutions use handcrafted rules, prone to human bias, and far from real-world peculiarities, we adopt a neural learning strategy to extract significant features from data rather than assuming that artificial rules are correct. We validated our findings on two challenging datasets. Our obtained results were as good or better than their literature counterparts.

Keywords Action recognition · Early fusion · Intermediate fusion · Late fusion · Deep learning

1 Introduction

Human action recognition plays a fundamental role in designing intelligent video solutions that can be useful for assistive living applications, health monitoring, smart homes, intrusion detection, robotics, etc. [1–3,10,11,23,26,29,33].

A large number of action recognition approaches have been proposed so far. They mostly rely on the data provided by different types of sensors, which can be sequences of RGB images, skeletons, depth, and infrared images. Each modality has its advantages for identifying action in some circumstances, yet each suffers from some inherent weaknesses. For example, RGB images are good at recognizing actions through the evolution of the color appearance of a person in a video. Still, they do not provide any clues on the 3D structure of the scene and they are sensitive to illumination perturbation. Therefore, the use of a single sensor/modality in an action recognition setup does not ensure optimal robustness, since most sensors have a limited field of view, are unable to overcome occlusions, and are sensitive to illumination change [13,20].

With the advent of the Microsoft Kinect sensor [37], which can simultaneously stream RGB and depth images as well as 3D skeletal data, several research works have been carried out to design robust approaches for action recognition by combining Kinect's modalities. Indeed, they aim to leverage each modality's strengths since they are interdependent and complementary.

Early fusion consists of integrating the separate raw data modalities into a unified representation before proceeding through the learning/feature extraction process. Despite the potential interest of this strategy, there have been few works in the literature that adopted it.

Feature-level fusion, also known as intermediate fusion, refers to the transformation of raw inputs into a higher-level representation by mapping them through a stack of layers. After unifying the feature representation we obtain multimodal feature maps that we can later use for recognition. Nevertheless, feature-level merging has the adverse

✉ Said Yacine Boulahia
boulahia.yacinesaid@gmail.com;
saidyacine.boulahia@emp.mdn.dz

[1] Ecole Militaire Polytechnique, BP 17, Bordj el Bahri 16111, Algiers, Algeria

effect of diluting the single modalities' strengths together [17,31,36,45].

Late fusion consists of extracting decisions from single-modality architectures, then applying a deep fusion to compute the final decision. Nevertheless, the existing approaches [5,12,20,25] have mainly applied rule-based algorithms to compute the final decision. Therefore, the evaluation of the impact of late fusion with a learning procedure should be more sound as it relies on the data without assuming any priors.

Considering the weaknesses of the above-mentioned strategies, we propose alternative approaches for the mentioned fusion schemes (early, intermediate, and late). Also, we discuss the impact of each approach for the type of fusion considered, and validate the most suitable choice for general multimodal action recognition.

The remainder of the paper is organized as follows. In Sect. 2, we discuss previous works on multimodal visual action recognition. In Sect. 3, we present the new approach for early fusion. In Sect. 4, we describe the proposed approach for intermediate fusion. In Sect. 5, we present our late fusion approach. Experimental results and discussions are given in Sect. 6. The paper is concluded with Sect. 7, where we summarize our contributions and give some future directions.

2 Related works

An imaging modality refers to a type of data that a given sensor can deliver for a specific application domain. In the human action recognition domain, it is possible to collect data from several modalities with different degrees of precision depending on the sensor. In particular, these modalities consist of RGB images, skeletons, and more recently depth maps and infrared images. Instead of processing the modalities separately, several action recognition works have explored some multimodal fusion strategies to improve recognition performance.

One of the criteria for dividing the existing multimodal fusion approaches is the learning paradigm on which they are based. The first category includes multimodal fusion approaches based on manually designing the representation of the action by selecting a set of distinctive features. For instance, Zhao et al. [44] proposed an approach to extract keypoints from RGB images and depth maps in a cluttered background and partial occlusions environment. The resulting feature vector is obtained by combining the partial feature vectors based on which a linear SVM selects the class.

Recent multimodal fusion approaches rely on a deep learning paradigm for action classification. The latter are based either on the fusion at the feature level (intermediate fusion) [17,27,31,35,43,46], or at the decision level (late fusion) of the single classifiers' results [20,25].

Regarding the intermediate fusion, Zhao et al. [43] proposed a system that combines recurrent neural networks for skeletal data, and 3D convolutions for RGB videos. The authors integrated the time dimension in the convolution stack, hence allowing the extraction of spatio-temporal patterns from the video. The resulting feature vectors from the first fully connected layer of each model are then concatenated and passed through a linear SVM for classification. Another work based on intermediate fusion was proposed by Liu et al. [27] who introduced a new multimodal feature fusion strategy within the LSTM unit [16]. The aim is to exploit more efficiently the multimodal features extracted for each skeleton joint and to improve the learning capability of their neural network dubbed ST-LSTM. The latter is a Spatio-Temporal LSTM that simultaneously models the temporal and spatial dependencies of the skeletal joints. The authors described their new skeleton LSTM unit as the *geometric unit* and the RGB image LSTM unit as the *visual unit*. They merged the two types of features inside the ST-LSTM unit, rather than concatenating them at the input level.

More recently, Su et al. [39] proposed a novel multimodal fusion module (MSAF) that learns to focus on the most informative features across different modalities. Their module integrated attention components that split each modality into channel-wise equal features blocks and create a joint representation. This later is then used to generate soft attention for each channel across the features blocks. Das et al. [7] proposed another intermediate-based approach that aimed at focusing on the most significant spatio-temporal aspects. In particular, their recognition pipeline relied on spatial embedding and an attention network. On the one hand, the spatial embedding projects the 3D poses and RGB cues in a common semantic space. On the other hand, the attention branch provided joint spatio-temporal attention weights across the whole video. A similar attention-based approach was introduced by Md Mofijul et al. [21,22], called Multi-GAT, which hierarchically learns complementary multimodal features. Their approach consists of a multimodal mixture-of-experts model to disentangle and extract salient modality-specific features that enable feature interactions.

As to the late fusion scheme, the data for each modality undergoes a separate processing pipeline, where each classifier returns a decision for the associated modality. For instance, Khaire et al. [25] proposed a multimodal approach to combine several visual cues. Particularly, they build a five-stream CNN to handle RGB, depth, and skeletal data. From RGB images, the authors proposed to create Motion History Images (MHI), which are later used to train a one-dimensional CNN. Also, they rotate the 3D point clouds of the depth maps to create three Depth Motion Maps (front, side, and top). The last network was obtained using skeleton

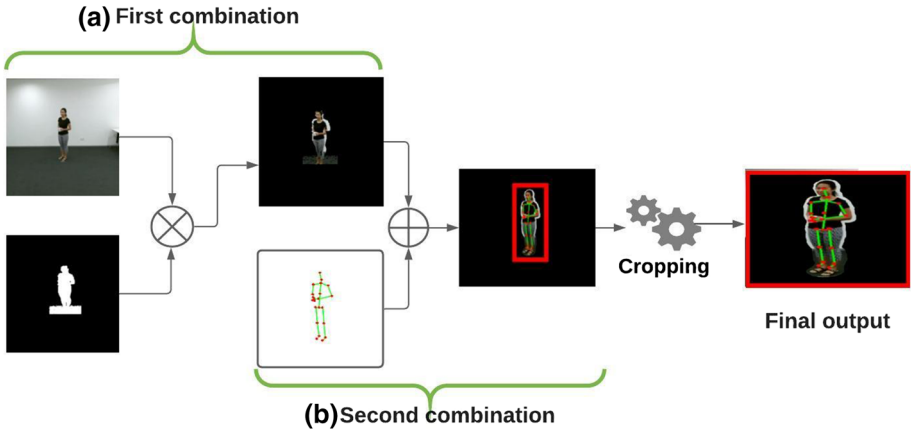

Fig. 1 Alignment and raw fusion of RGB images, depth maps and skeletal sequences

data converted into 2D images. To combine the probability scores generated by the five classifiers, the authors proposed a Weighted Product as an unsupervised decision analysis method that involves merging the scores in order to find the most appropriate class for a given action. Joze et al. [24] presented a simple neural network module for leveraging the knowledge from multiple modalities in convolutional neural networks. Their proposed module, named Multimodal Transfer Module (MMTM), can be inserted into different levels of any late fusion backbone architecture.

Besides, RGB images, depth maps, and skeletal data, other modalities could be used for the sake of action recognition. For instance, Pham et al. [32] proposed a multi-sensing modality framework for human action recognition by combining skeleton and acceleration data. Their approach mainly relies on dilated causal convolution components for learning the feature representation which is fed then into two fully connected layers for the prediction. Memmesheimer et al. [30] introduced a novel approach that combines four modalities including skeleton sequences, inertial and motion capturing measurements as well as Wi-Fi fingerprints. Their approach transforms these individual signals and represents them as an image. The resulting images are then classified using an EfficientNet [40] architecture. Another work conducted by De Boissiere et al. [9] suggested combining infrared and skeleton data. Data from these two modalities are fed into two distinct CNN-based feature extractors and then fused to constitute the final representation. Authors claim that due to the significant difference between these two modalities, their combination significantly increases recognition performance.

After reviewing the literature, we noticed that the approaches based on intermediate fusion are effective in representing the shared and specific modality features. However, the large size and diversity of the feature vectors impinge on real-time recognition. We also found that late-fusion approaches are straightforward, as they do not require an understanding of the individual architectures. However, they are inefficient when the number of modalities increases. Besides, they are based on a weak fusion strategy since they rely on the final decisions of the classifiers, which themselves are modality-wise. In light of the literature, we will address the weaknesses of each fusion scheme before disclosing the most robust configuration that delivers the best performance/speed trade-off.

3 Proposed early fusion approach

Early fusion is applied before entering a recognition network. It transforms raw data into a more succinct intermediate form. To the best of the authors' knowledge, there has not been any action recognition work that investigated in-depth an early fusion strategy. We propose a modality combination approach as an early fusion scheme. We combine RGB images, depth maps, and skeletal sequences since this data can be delivered effortlessly in real-time from RGBD sensors. For instance, by combining the three modalities, we leverage the color and texture information of RGB images, we facilitate background removal with depth maps as well as joint hierarchy computation in skeleton data. The latter can also provide us with the 3D spatial location of the action performer (subject) at no processing cost.

We begin with applying the depth mask to RGB images. Most recent datasets are collected using depth cameras like the Kinect and thus provide depth maps as input data. These maps are timed on the sequence of RGB images. However, these RGB images have larger height and width dimensions than the depth map. Thus, when depth maps are used, we first resize them to reach the RGB image size, and then we apply the mask produced on the RGB image. The result is an RGB image cropped along the subject(s) acting. An illustration of the final result can be seen in the block (a) within Fig. 1 (also see Fig. 7).

Second, we project the skeletal sequences onto the resulting image of the first combination. The aim of such a projection is twofold. First, it benefits from the shape of the human skeleton and allows us to guide the result of the first

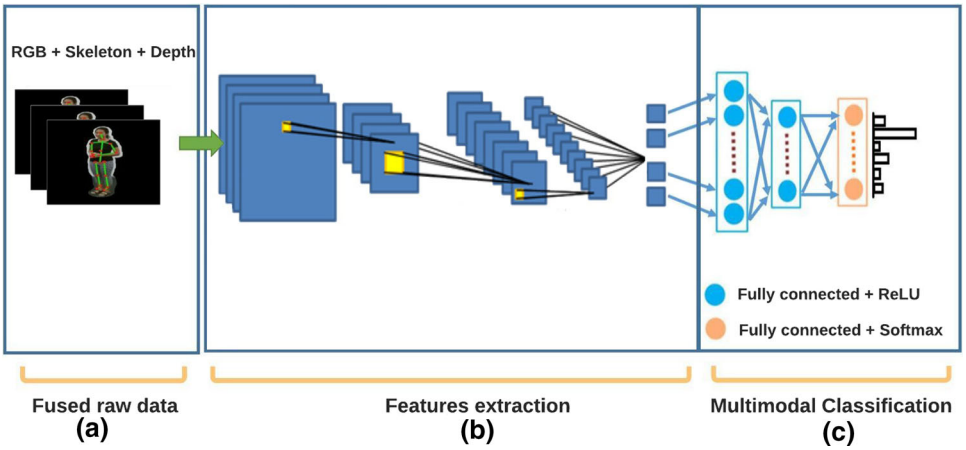

Fig. 2 Proposed early fusion action recognition architecture

combination using 2D skeletal coordinates. For instance, at this level, the subject can be cropped easily in the image. The result of this second combination is shown in block (b) within Fig. 1. The final output of this process is a three-channel (RGB) image with on the one hand a focus on the person acting and on the other hand an overlay of the skeletal data on the color image.

After preprocessing the input data for the two combinations mentioned above, we move now to the part dealing with the model used for the features representation and classification of the action. In fact, our study is intended to be independent of the type of network used insofar. As in this first proposal, the focus is on the combination of data taking place upstream of the network. Also, the choice of network is up to the designer of the final approach, and it is, therefore, possible to substitute the one used as new models are proposed in the literature. Consequently, the different components constituting the recognition process are shown schematically in Fig. 2 while the effective choice would be explained during the evaluations presented in Sect. 6, where different models are used.

In particular, in this figure, the sequence of images produced passes through the layers constituting the recognition model, which must imperatively be a model developed for the representation and classification of RGB images. As such, a model initially developed for skeletal type data cannot be used. Moreover, the models existing today in the literature for modeling RGB images are of two types. Either they model a single image or frame at a time, as is the case with the VGG network [38], Resnet [15] and DenseNet [18], or they make it possible to provide a single representation for a whole sequence of frames. In the first case, an additional step is to be conducted which consists of adding a mechanism to count the score of each frame and to decide based on a majority vote the nature of the predicted action class. In the second case, the models have been explicitly developed for image sequences and therefore the representation produced can be used directly for classification, as is the case with the I3D [4] model for example.

4 Intermediate fusion approach

Intermediate fusion is a type of fusion that takes place inside the recognition model. This kind of fusion combines the features that distinguish each type of data to produce a new representation that is more expressive than the separate representations from which it arose. For example, the fusion of features extracted from RGB images and those extracted from skeletal sequences, allows us to take advantage of the strengths of both representations simultaneously. This could achieve good recognition results compared to using a single representation individually.

As shown in Sect. 2, several approaches proposed to conduct an intermediate fusion [17,24,27,31,43,45]. Nevertheless, we noted that these works do not make a qualitative selection of the distinctive features of human action among the generated feature-set. On the contrary, they combine all of the features thus produced. However, despite being diverse, a lot among these features are of a weak expressive capacity. This leads to confusion between similar classes caused by "noisy" features which can drown out the "good" ones.

Indeed, the overall feature-set does contain some valuable features. But it also contains features of little interest, such as those which do not vary along all the classes, which do not have a characteristic pattern for each class, or which are simply always zero. Therefore, in this second proposition, our objective is to show the procedure used to conduct an intermediate fusion that incorporates a systematic feature selection step. We assume that it would be preferable to apply a selection on these features qualitatively to obtain a more detailed description of the action and therefore results in a better recognition.

An important point to note is that this selection takes place according to a back and forth process, explained later in this

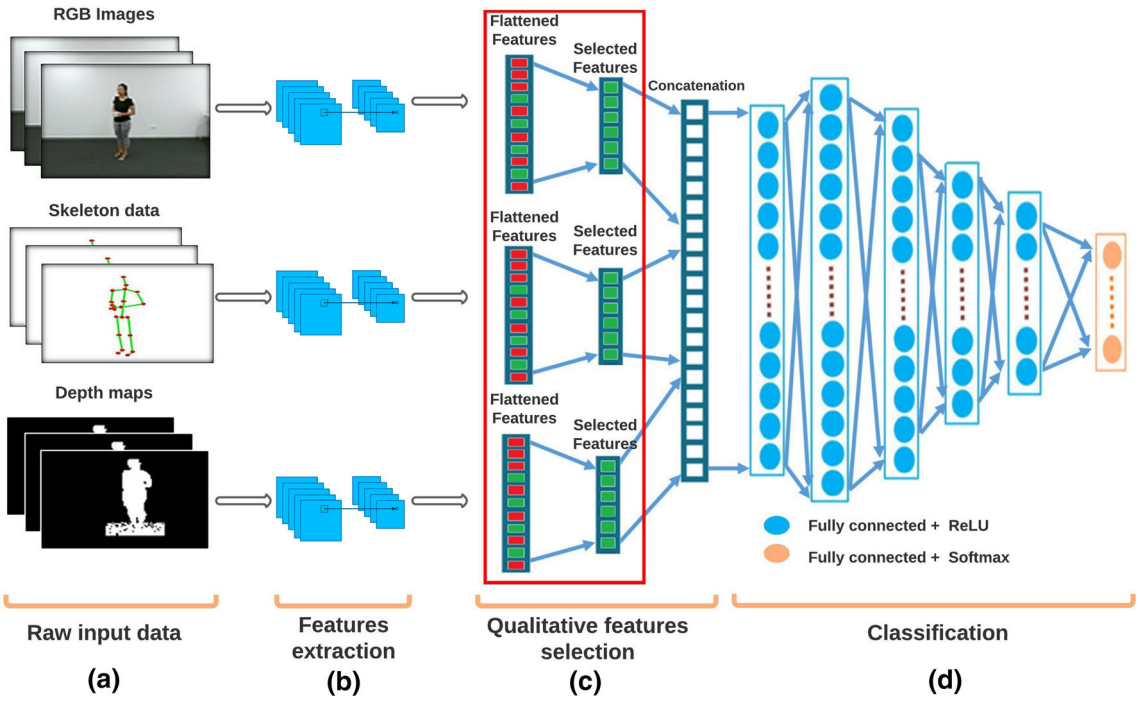

Fig. 3 Proposed intermediate fusion action recognition architecture

section, between the learning and validation set, and that in this sense, once the most valuable features are determined, the final result of the approach is obtained on the test set. In other words, the data constituting the test set does not participate in determining the best features, only the training and validation data are used.

A second important point to note is that our procedure is generic and is presented independently of the model used in practice for each modality. Thus, in this section, we will explain this procedure while intentionally keeping the number of modalities considered, the feature extraction models used for each one and the final classification model as generic. The actual choice would be explained during the evaluations presented in Sect. 6 where results are compared when these different parameters are varied.

In particular, our procedure starts with the extraction of the feature vectors from each considered modality separately. In Fig. 3, the proposed architecture is illustrated for $K = 3$ modalities, namely RGB images, depth maps, and skeletal sequences. For each modality, a specific deep and pre-trained feature extractor is used. Given the action sequence length, the dimension of each such feature vector, respective to each modality, is most often relatively large. For instance, the use of the ST-GCN [41] extraction module provides on average 96, 000 features for each skeleton-based action sample.

We then compute the mean vector $\text{Mean}_{k,c,j}$ of feature j, along the M samples that constitute the validation dataset, for each class c in each modality k separately according to the following formula:

$$\text{Mean}_{k,c,j} = \frac{1}{M}\sum_{i=1}^{M} f_{i,j} \quad 1 \le j \le J_k, \quad 1 \le c \le C, \quad 1 \le k \le K \tag{1}$$

With J_k the total number of features extracted for a given modality k, C the total number of classes, and K the total number of modalities that are considered. The aim is to extract the most distinctive features for each class of actions in each modality separately.

After the computation of the mean, we obtain $K \times C$ vectors (K times the number of classes C). Each characterizes a class in a given modality. We then calculate the variance along the vectors of the same modality as shown in the following formula:

$$\text{Var}_{k,j} = \frac{1}{C}\sum_{c=1}^{C}(\text{Mean}_{k,c,j} - \text{Mean}_{k,c})^2 \quad 1 \le j \le J_k, \quad 1 \le k \le K \tag{2}$$

This operation allows us to deduce the indices of the most significant features that amount to a better distinction

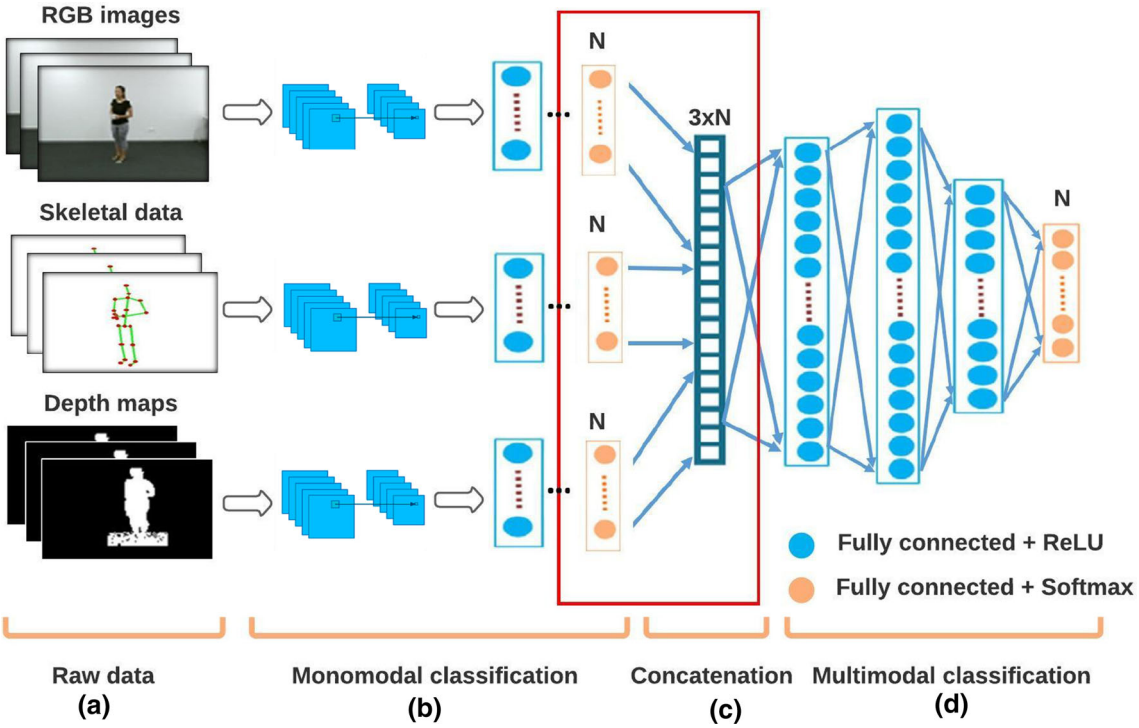

Fig. 4 Proposed architecture for late fusion recognition

between the classes for the same modality. Fig. 3 block (c) illustrates this qualitative selection. In particular, we varied, on the validation set, the variance threshold from the maximum value found in the vector Var_k for the k-th modality to the minimum value in this same vector. The variation curve is parameterized with the number of features that ranges from 0, no feature is retained, up to the total number of features, all features maintained or no selection has been made. Finally, the chosen variance threshold coincides with the peak of the curve. We then keep only the features that have a greater variance value than the determined variance threshold. We will denote by S the total number of features after concatenation.

The last (d) part of Fig. 3 illustrates the classification module. We opted for a structure that contains a total of six fully connected layers. The first five have an activation function *ReLU*. From left to right, the size of the first layer is equal to the number of descriptors resulting from the concatenation, i.e., S. The size of the second layer is equal to $2 \times S$ and is greater than that of the first. This increase makes it possible to avoid the *Bottleneck* phenomenon which causes the network to compress the representations of the features and consequently, to reduce the precision of the classification. The size of the third, fourth and fifth layers is gradually reduced from S to $S/3$. The sixth (last) layer is a *softmax* layer with a size equal to the number of classes and produces a probability distribution. We applied the *dropout* function on the penultimate layer to avoid overfitting the data.

5 End-to-end late fusion approach

Late fusion is a merging strategy that occurs outside of the monomodal classification models. It combines the decisions of each classifier to produce new decisions that are more precise and reliable. In particular, existing works dealing with late fusion do not apply a deep fusion of scores based on neural networks. Indeed, these works simply recover scores from the softmax layers of the monomodal networks, then apply man-designed rules for the fusion of the scores.

Therefore, we design an end-to-end late fusion network where the score of merging is computed by a deep neural network. In particular, for K considered modalities, we used pre-trained architectures to generate score vectors from each modality individually. Each such architecture performs both features extraction and classification and provides a vector of the potential membership scores to each of the considered classes. After being pre-processed, these vectors are used as inputs to our network for training. Such an operation allows us to learn more consistent joint decisions than conventional merging rules.

Our classification network contains only fully connected layers. As we did in the intermediate fusion, at this stage, we do not need any convolutional layers. The results of the K pre-trained architectures are vectors of scores, which have already undergone several convolutions. Figure 4 illustrates the overall architecture of our late fusion approach for $K = 3$ modalities.

In particular, block (a) contains the data flow of the three input modalities. Block (b) holds a scoring step. We notice that the size of the outputs is the same for all three modalities. As illustrated, and for N considered classes, each architecture would produce a score vector of length N. Block (c) concatenates the three decision vectors.

Finally, block (d) is the classification module comprised of four fully connected layers. The size of the first layer is equal to K times the number of classes N. The second layer equals $(4 \times N)$ which is greater than the size of the first layer to again avoid the Bottleneck phenomenon. The size of the third layer is $2 \times N$, which is associated with a dropout. The fourth (last) layer is a softmax function of a size equal to the number of classes.

6 Experimental results

In this section, we evaluate the proposed architectures. In particular, we begin with a description of the two datasets, evaluation protocols, and evaluation metrics. Afterward, we present the experimental results obtained with each architecture, i.e., the early, intermediate, and late fusion. In these experiments, we used a workstation equipped with an 8GB memory graphics card, an 8-th generation i7 processor, and 16GB RAM. We implemented our deep neural networks with the Python programming language and the TensorFlow framework on Ubuntu 18.2LTS operating system.

6.1 Datasets

Our experiments were evaluated on the NTU RGB-D [34] and the SBU Interaction [42] datasets. These datasets are often used for evaluation by most recent action recognition approaches. Also, they are complementary in the sense that the first (NTU RGB-D) contains actions performed by a single subject, while the second (SBU Interaction) contains actions of human interaction between two subjects.

The NTU RGB-D dataset was first introduced by Shahroudy et al. [34] in 2016. It is the largest and best-known multimodal dataset that contains RGB images, 3D skeletal sequences, and depth maps. The sequences are captured by three Microsoft Kinect cameras placed at the same height but oriented at different horizontal angles ($-45°$, $0°$, $45°$). This dataset comprises 56,880 action samples collected from 40 different ages, sex, and size. It includes 60 different classes, covering 40 daily actions, 9 health-related actions, and 11 mutual actions.

On this dataset, we adopt two standard evaluation protocols proposed by its authors [34]. It consists of the Cross-View (CV) protocol based on the angle of view of a given camera and the Cross-Subject (CS) protocol based on the subjects performing the actions.

In particular, according to the Cross-View protocol, samples collected from cameras 2 and 3, corresponding to the frontal and side views of the actions, are used for training. On the other hand, the samples of camera 1 (corresponding to 45-degree left and right views of the actions) are used for testing. The resulting training and test sets comprise 37,920 and 18,960 instances, respectively.

As far as the Cross-Subject protocol is concerned, the 40 subjects constituting the dataset are split into training and testing groups. Each group consists of 20 subjects. The IDs of training subjects are: 1, 2, 4, 5, 8, 9, 13, 14, 15,16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38; while the rest of the subjects are left for the test. Given that, the training and test sets have 40,320 and 16,560 instances, respectively.

On the other hand, the SBU Interaction dataset [42] comprises 21 subsets, each containing sequences from a different pair of subjects. It includes a total of 282 manually segmented sequences. It contains 8 classes of interactions between two persons including actions of approaching, moving away, pushing, kicking, punching, swapping objects, hugging, and shaking hands. Seven subjects participated in the performance of these actions. It is important to mention that like NTU RGB-D, this dataset is also comprised of RGB, depth, and 3D skeletons.

On this dataset, we followed the evaluation protocol first suggested in [42]. As this dataset is split into 5 groups, each containing 4 or 5 sets of two actors, every set of two performers would appear exclusively either in training or in test data. The assessment is made by a 5-permutation cross-validation, that is, 4 groups are used for training and 1 group for the test. The overall performance is obtained by calculating the average of the results obtained for the 5 permutations.

6.2 Evaluation metrics

We based our assessment on two criteria, the first of which was accuracy. The latter evaluates classification performance. By definition, accuracy refers to the portion of correct (positives and negatives) predictions made by the model as follows:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{3}$$

The second criterion is prediction time, which is the time needed, for the model, to classify a single frame. As the latter is very short for a single frame, we compute it for a batch of 100 frames than we average.

6.3 Support deep models

As mentioned during the presentation of the different suggested strategies, our approach is independent of the choice

of models used in practice. However, in order to obtain quantitative results, we selected some pre-trained models found in the literature. The idea is not to find out which of the models is the most efficient for multimodal recognition of actions. We seek rather measure which of the three proposed fusion strategies is the best by evaluating them under the same conditions, in particular in terms of the model supporting each of the approaches.

Therefore, for modeling RGB images and depth maps, we will use four models including two classic models and two more recent ones, namely VGG16 [38], DenseNet [18], I3D [4] and ResNeXt [14], respectively. As for skeletal data, we will use three recent models namely ST-GCN [41], MS-G3D [28] and Shift-GCN [6]. Bellow, we provide a brief description of each one of them.

VGG16 Architecture is a convolutional neural network model proposed by [38]. It was one of the famous models submitted to ILSVRC-2014. It improved existing models by replacing large kernel-sized filters with multiple 3×3 kernel-sized filters one after another. The VGG16 architecture is a feature extractor that operates on a single image. Therefore, when applied to action sequences, a voting method should then be used to get the final output class.

DenseNet Architecture is a network architecture where each layer is directly connected to every other layer in a feed-forward fashion [18]. For each layer, the feature maps of all preceding layers are treated as separate inputs, whereas its feature maps are passed on as inputs to all subsequent layers. As for the VGG16 architecture, the DenseNet is applied to each frame. A voting method is then needed to get the final class.

I3D Architecture is an Inception-type architecture. It operates on a whole sequence. The input size for it is selected as 224×224 and 64 frame length is used conforming with the I3D study [4].

ResNeXt Architecture it is essentially ResNet with group convolutions [14]. Same as the I3D, this architecture provides a single feature vector for the whole sequence. For using this architecture, the input size is selected as 112×112, and a 64 frame length is adopted.

ST-GCN Architecture is a novel model of dynamic skeletons called Spatial-Temporal Graph Convolutional Networks (ST-GCN) [41]. It moves beyond the limitations of previous methods by automatically learning both the spatial and temporal patterns from the skeleton graph structure.

MS-G3D Architecture is a graph convolutional-based architecture that mainly relies on multi-scale graph convolutions and a new unified spatial-temporal graph convolutional operator [28].

Shift-GCN Architecture is another graph convolutional network that is composed of novel shift graph operations and lightweight point-wise convolutions, where the shift graph operations provide flexible receptive fields for both spatial and temporal graphs [6].

Table 1 Results of the ablation study according to the **early fusion** strategy along the Cross-Subject (CS) and Cross-View (CV) protocol on NTU RGB-D dataset

Support model	RGB	Skel.	Depth	CV (%)	CS (%)
VGG16	X	X		**77.60**	**79.45**
		X	X	75.22	78.94
	X		X	75.90	78.19
DenseNet	X	X		**81.03**	**86.11**
		X	X	79.14	83.10
	X		X	80.55	85.42
I3D	X	X		81.44	**86.80**
		X	X	79.66	85.03
	X		X	**81.70**	86.49
ResNeXt	X	X		**82.29**	**86.03**
		X	X	79.37	83.11
	X		X	80.01	85.55

Best achieved performances along each configuration are set in bold

6.4 Ablation studies on NTU RGB-D

In this section, we will analyze two main steps of our multimodal recognition proposals. It concerns mainly the set of considered modalities and the impact of the feature extractor architectures. The latter are used to process on the one hand the image data and the depth map and on the other hand to deal with the skeleton data. In our experiments, we selected four basic models to process the RGB and depth maps and three most recent skeleton-based models. In total, four setups are followed for the early strategy, and twelve others are tested for each of the two remaining strategies, namely the intermediate and late fusion. Notice that the recognition score obtained with the simultaneous use of the three modalities is provided in Sect. 6.6 along with previous works' achievements.

First of all, regarding early recognition, we note from Table 1 that the models obtain different results, for the two adopted protocols, depending on the considered combination of modalities. Nevertheless, we note that the best combination is the one involving RGB image data and skeletal data (except for the I3D model which obtained a slightly higher score with the RGB image and depth map combination according to the Cross-Subject protocol). This may be due to the additional information typology reported in the RGB images by the integration of the skeletal data and consequently allowing the models to focus on certain parts of these images.

Table 2 Results of the ablation study according to the **intermediate fusion** strategy along the Cross-Subject (CS) and Cross-View (CV) protocol on NTU RGB-D dataset

Support model	RGB	Skel.	Depth	CS (%)	CV (%)
VGG16	X	X		89.46	91.33
+		X	X	87.00	88.10
ST-GCN	X		X	86.13	88.50
VGG16	X	X		**91.66**	**92.15**
+		X	X	85.30	86.66
MS-G3D	X		X	88.41	89.65
VGG16	X	X		89.88	91.35
+		X	X	86.10	88.44
Shift-GCN	X		X	86.36	89.95
DenseNet	X	X		87.03	90.44
+		X	X	84.74	87.34
ST-GCN	X		X	87.15	88.90
DenseNet	X	X		88.32	**92.80**
+		X	X	87.93	90.61
MS-G3D	X		X	87.03	88.85
DenseNet	X	X		**88.55**	90.87
+		X	X	84.81	85.97
Shift-GCN	X		X	87.46	88.21
I3D	X	X		90.90	92.50
+		X	X	88.98	90.73
ST-GCN	X		X	89.12	91.74
I3D	X	X		91.28	93.80
+		X	X	89.62	91.31
MS-G3D	X		X	91.01	92.77
I3D	X	X		**93.49**	**96.20**
+		X	X	90.27	93.41
Shift-GCN	X		X	91.11	95.99
ResNeXt	X	X		91.02	93.85
+		X	X	90.11	91.99
ST-GCN	X		X	90.67	91.30
ResNeXt	X	X		**92.11**	94.18
+		X	X	88.55	91.70
MS-G3D	X		X	90.01	93.58
ResNeXt	X	X		91.53	**94.25**
+		X	X	89.63	91.50
Shift-GCN	X		X	90.34	92.60

Best achieved performances along each configuration are set in bold

Table 3 Results of the ablation study according to the **late fusion** strategy along the Cross-Subject (CS) and Cross-View (CV) protocol on NTU RGB-D dataset

Support models	RGB	Skel.	Depth	CS (%)	CV (%)
VGG16	X	X		87.35	88.57
+		X	X	86.00	86.88
ST-GCN	X		X	84.48	85.52
VGG16	X	X		**89.01**	**90.88**
+		X	X	85.52	86.43
MS-G3D	X		X	87.11	88.20
VGG16	X	X		87.14	90.75
+		X	X	86.47	87.70
Shift-GCN	X		X	87.84	90.38
DenseNet	X	X		**90.12**	**92.38**
+		X	X	87.40	88.59
ST-GCN	X		X	89.67	89.10
DenseNet	X	X		88.11	91.02
+		X	X	86.82	88.13
MS-G3D	X		X	88.40	90.59
DenseNet	X	X		87.65	88.59
+		X	X	83.39	85.33
Shift-GCN	X		X	87.22	88.11
I3D	X	X		88.10	91.19
+		X	X	86.62	87.18
ST-GCN	X		X	88.57	90.03
I3D	X	X		**90.70**	91.50
+		X	X	86.63	90.26
MS-G3D	X		X	88.40	90.19
I3D	X	X		90.16	**94.90**
+		X	X	88.73	91.62
Shift-GCN	X		X	89.41	92.11
ResNeXt	X	X		90.55	**94.01**
+		X	X	89.20	91.86
ST-GCN	X		X	90.10	91.00
ResNeXt	X	X		**92.31**	93.21
+		X	X	87.46	90.55
MS-G3D	X		X	88.58	92.32
ResNeXt	X	X		91.00	92.82
+		X	X	89.53	90.70
Shift-GCN	X		X	89.19	91.81

Best achieved performances along each configuration are set in bold

We also note that the results are significantly better with the Cross-View protocol than with the Cross-Subject protocol. This can be explained by the fact that the subjects seen in tests according to the second protocol are not the same as those on which the model was trained, leading to a strong variability, especially when RGB images are used.

Moreover, we note that the models with the best scores are the global models, i.e., those which extract the features for the whole sequence, namely the I3D model and the ResNeXt model. In particular, the I3D model obtains a value of **86.80%** according to the Cross-View protocol while the ResNeXt model reached a score of **82.29%** according to the more difficult Cross-Subject protocol. Compared to local models, such as VGG16 for which a voting method is required, global models better report correlation information between different frames in the produced representation.

Regarding the results of the proposed intermediate strategy reported in Table 2, we note that overall these scores are significantly higher than what has been obtained so far by the early approach. Indeed, most of the modality combinations allow reaching scores close to or above 90% according to the two protocols on the NTU-RGBD dataset. The main objective of this experiment is to quantitatively estimate the impact of the use of different models with the possible modality combinations.

We also note that the combinations of models dedicated to the image and those dedicated to the skeletal data present significantly different results with nevertheless two common points. The best performing combination of modalities is that of RGB image data and skeletal data, as already noted for the early strategy. Moreover, we note that the scores according to the Cross-View protocol are better than those obtained according to the Cross-Subject protocol.

More precisely, this second experimentation allowed to determine among the proposed model combinations the most efficient one for this strategy, namely the combination of the I3D model and the Shift-GCN model. This combination achieved a score of **93.49%** and **96.20%** according to the CS and CV protocols, respectively. These scores are very significant and show that the use of models specific to the considered modalities and for which the descriptors have been cleaned up allows exceeding the recognition scores obtained with the early approach, where the modalities are mixed.

As for the late fusion strategy, we notice that the scores reported in Table 3 are to be situated between on the one hand the scores of the early strategy and on the other hand those obtained according to the intermediate fusion strategy. This observation is valid for all combinations of modalities as well as those of the different combinations of models. However, we note that for this third strategy, several combinations of models are efficient, like the combination of the ResNeXt and MS-G3D models and the ResNeXt and ST-GCN models. Nevertheless, even for this fusion strategy, the combination of the I3D and Shift-GCN models is the most efficient with a score of **94.90%** according to the Cross-View protocol and the use of RGB images and skeletal data as entrance modalities.

In order to provide a qualitative overview of the input data issued from different modalities combinations, we illustrate

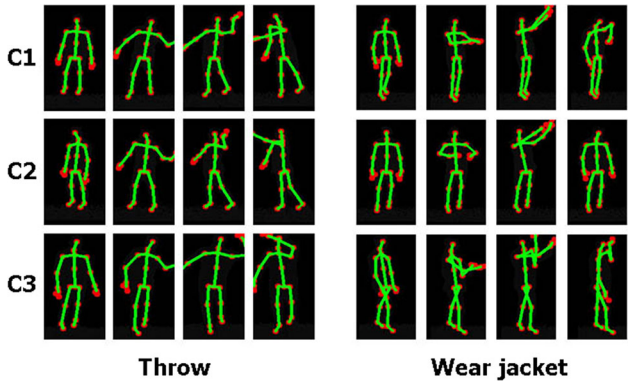

Fig. 5 Illustrations of two actions, Throw and Wear jacket, issued from skeleton and depth frames combinations. C1, C2 and C3 sands for camera 1, 2 and 3, respectively

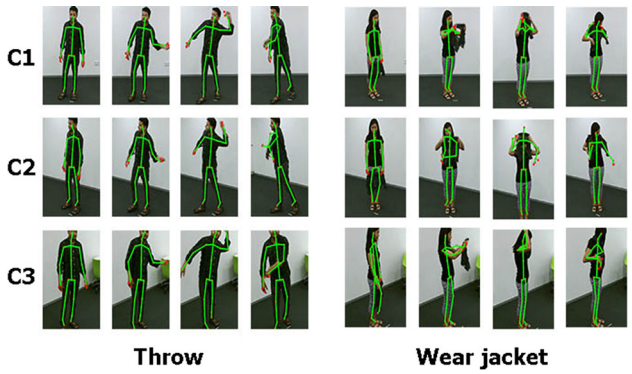

Fig. 6 Illustrations of two actions, Throw and Wear jacket, issued from skeleton and RGB frames combinations. C1, C2 and C3 sands for camera 1, 2 and 3, respectively

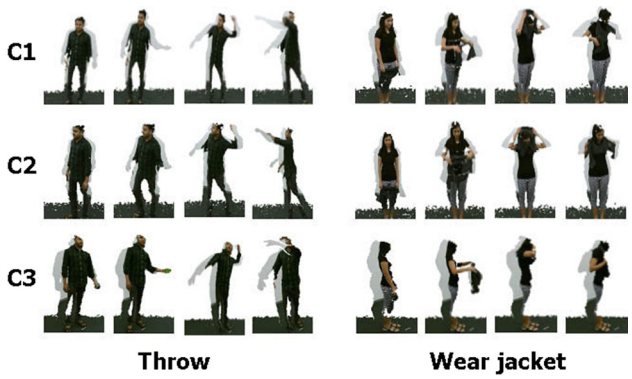

Fig. 7 Illustrations of two actions, Throw and Wear jacket, issued from RGB and depth frames combinations. C1, C2 and C3 sands for camera 1, 2 and 3, respectively

Table 4 Results of the ablation study with the **three multimodal strategies** according to the cross-validation protocol on SBU Interaction dataset

Fusion level	Support model	RGB	Skel.	Depth	Accur. (%)
Early	VGG16	X	X		**87.71**
			X	X	84.23
		X		X	85.90
	DenseNet	X	X		88.11
			X	X	85.25
		X		X	**88.36**
	I3D	X	X		**88.89**
			X	X	86.26
		X		X	87.90
	ResNeXt	X	X		**88.70**
			X	X	84.50
		X		X	87.72
Inter.	VGG16	X	X		94.86
	+		X	X	90.13
	ST-GCN	X		X	93.49
	VGG16	X	X		**95.10**
	+		X	X	92.60
	MS-G3D	X		X	95.84
	VGG16	X	X		94.80
	+		X	X	91.63
	Shift-GCN	X		X	93.50
	DenseNet	X	X		94.00
	+		X	X	91.12
	ST-GCN	X		X	94.20
	DenseNet	X	X		95.82
	+		X	X	91.30
	MS-G3D	X		X	93.92
	DenseNet	X	X		**95.91**
	+		X	X	93.45
	Shift-GCN	X		X	92.38
	I3D	X	X		96.05
	+		X	X	94.70
	ST-GCN	X		X	95.40
	I3D	X	X		96.35
	+		X	X	94.90
	MS-G3D	X		X	96.10
	I3D	X	X		**97.03**
	+		X	X	95.72
	Shift-GCN	X		X	95.48
	ResNeXt	X	X		94.22
	+		X	X	89.82
	ST-GCN	X		X	93.40

Table 4 continued

Fusion level	Support model	RGB	Skel.	Depth	Accur. (%)
Inter.	ResNeXt	X	X		**95.84**
	+		X	X	94.44
	MS-G3D	X		X	94.30
	ResNeXt	X	X		95.00
	+		X	X	93.83
	Shift-GCN	X		X	95.15
Late	VGG16	X	X		92.02
	+		X	X	88.81
	ST-GCN	X		X	91.00
	VGG16	X	X		**92.77**
	+		X	X	91.42
	MS-G3D	X		X	92.23
	VGG16	X	X		91.14
	+		X	X	88.71
	Shift-GCN	X		X	92.47
	DenseNet	X	X		93.70
	+		X	X	89.91
	ST-GCN	X		X	92.18
	DenseNet	X	X		92.77
	+		X	X	90.12
	MS-G3D	X		X	91.45
	DenseNet	X	X		**94.54**
	+		X	X	92.90
	Shift-GCN	X		X	92.40
	I3D	X	X		95.32
	+		X	X	92.10
	ST-GCN	X		X	92.90
	I3D	X	X		94.33
	+		X	X	93.17
	MS-G3D	X		X	94.81
	I3D	X	X		**95.53**
	+		X	X	92.81
	Shift-GCN	X		X	92.10
	ResNeXt	X	X		92.31
	+		X	X	88.56
	ST-GCN	X		X	91.56
	ResNeXt	X	X		**93.81**
	+		X	X	90.39
	MS-G3D	X		X	93.45
	ResNeXt	X	X		93.17
	+		X	X	90.50
	Shift-GCN	X		X	92.28

Best achieved performances along each configuration are set in bold

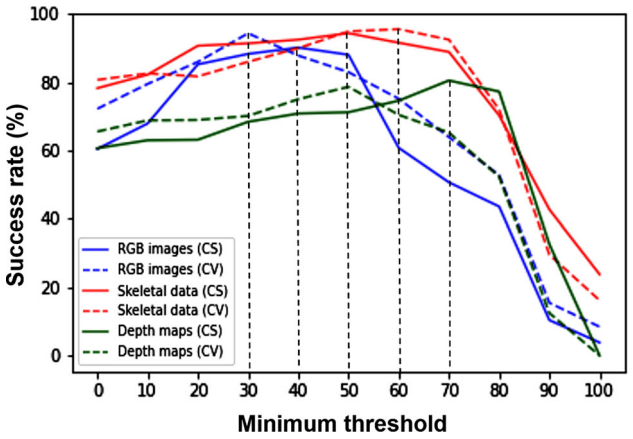

Fig. 8 Performance evolution on NTU RGB-D dataset of models based on RGB images (blue), skeletal data (orange) and depth maps (green) according to variations in minimum thresholds. CS and CV stands for Cross-Subject and Cross-View validation protocols, respectively (color figure online)

in Figs. 5, 6 and 7 two actions, namely Throw and Wear a jacket, using the combinations modalities two by two. It is noted that these kinds of samples are only used according to the early strategy-based recognition models, but they still make it possible to appreciate the difficulty that can be encountered during multimodal recognition with variations such as the viewing angle, occlusions by objects, noisy skeletons, etc.

6.5 Ablation studies on SBU interaction

Similar to the first dataset, we conducted on the SBU interaction dataset another series of ablation experiments where we investigated the impact of both the support model and the combination of considered modalities. It should be noted that this second experiment is complementary to the previous one, in the sense that the dataset comprises a much smaller number of action classes. It is therefore useful to assess the performance of our approaches in early, intermediate, and late fusion when the number of classes is small. Also, the type of actions differs from that of the previous dataset since that all sequences include two performing subjects instead of a single one. Achieved results according to the previously defined cross-validation protocols are presented in Table 4 for the three considered fusion levels namely early, intermediate, and late.

We can globally note that the results on this second dataset are better when compared to those obtained on the first dataset according to each of the three fusion strategies, respectively. This can be mainly due to a better quality of the data and the great distinction between the classes of actions that constitute this dataset. The reduced number of instances per class seems not to impact too much the performances since we relied on pre-trained support architectures.

In particular, we note that the best performing approach is once again the one based on an intermediate fusion which obtains a score of **97.03%** using the I3D and Shift-GCN architectures for the RGB images and the skeletal data, respectively.

These results also allow us to see that among the main architectures intended for modeling actions within RGB images or depth maps, the I3D and ResNeXt architectures are those that allow reaching the best scores about the three fusion strategies. On the other hand, the Shift-GCN architecture is generally the most efficient, except for some cases such as when combined with ResNeXt for the intermediate strategy or with VGG16 according to the late fusion strategy.

We also note that the best performing combination of modalities is the one including RGB images and skeletal sequences due to the very distinct and complementary nature of each of them. Nevertheless, we note on this second dataset that, according to an early fusion strategy, the use of the DenseNet architecture with images and depth maps as input data allows to have a slightly better score compared to the other combinations. This result shows the importance of depth maps, but does not invalidate the overall finding of the superiority of the combinations of RGB images and skeletal data.

6.6 Comparison with state-of-the-arts

After the ablation studies conducted on the two datasets, NTU-RGBD and SBU Interaction dataset, we propose in this section to present the final results of our three fusion approaches obtained with the use of the three modalities. In this same section, we propose a comparison with the best-performing approaches in the literature.

To this end, we report in Table 5 our experimental results. In particular, we provide the recognition accuracies and prediction latency in milliseconds for the most recent best-performing multimodal approaches, according to the two standard validation protocols, namely Cross-Subject (CS) and Cross-View (CV). For each, we specify the modalities that are merged (cross-mark) as well as the fusion level (early, intermediate, or late). We then report the results of the three fusion strategies. Based on the ablation study results, the retained support models are those with the highest scores, namely the I3D and the Shift-GCN.

From Table 5, we first notice that the proposed fusion recognition approaches are significantly better than the monomodal architectures which support them. Furthermore, we notice that the intermediate approach is indeed the best of the three fusion strategies when all three modalities are used. In particular, according to the Cross-Subject protocol, the resulting architecture obtains a score of **95.94%** while it

Table 5 Results of different proposed fusion approaches along with previous state-of-the-art ones on the NTU RGB-D dataset

Approach	Year	Fusion level	RGB	Skeleton	Depth	CS Accur. (%)	CV Accur. (%)	Letency (ms)
Chained Multi-stream [46]	2017	Late	X	X		76.90	–	0.653
ST-LSTM [27]	2017	Intermediate	X	X		73.20	80.60	–
Chained Multi-stream [46]	2017	Intermediate	X	X		–	80.80	0.610
Two-Stream LRCN [43]	2017	Intermediate	X	X		83.74	93.65	0.561
Deep Bilinear [17]	2018	Late	X	X	X	85.40	90.70	0.413
Two-Stream LRCN [43]	2017	Late	X	X		82.05	86.68	0.585
MFAS [31]	2019	Intermediate	X	X		90.04	90.04	–
MMTM [24]	2019	Intermediate	X	X		90.11	–	–
MSAF [39]	2020	Intermediate	X	X		92.24	–	–
VPN [7]	2020	Intermediate	X	X		95.50	98.00	–
Hierarchical [8]	2020	Early	X	X		95.66	**98.79**	–
I3D [4]	2017	–	X			85.81	87.12	0.269
Shift-GCN [6]	2020	–		X		90.70	96.50	**0.214**
I3D [4]	2017	–			X	82.44	85.02	0.237
Our		Early	X	X	X	86.74	87.62	0.218
Our		**Intermediate**	**X**	**X**	**X**	**95.94**	98.11	0.378
Our		Late	X	X	X	94.30	97.61	0.408

Best achieved performances along each configuration are set in bold CS and CV stands for Cross-Subject and Cross-View validation protocols, respectively

reaches a score of **98.11%** according to the more advantageous Cross-View protocol. Moreover, by analyzing the time required for the intermediate approach, we notice that this improvement of the recognition performances, compared to the support models, is obtained at a small extra cost.

As far as existing approaches are concerned, we first notice that despite its simplicity, our early approach achieved a decent performance. Moreover, our late fusion approach outperformed all previous approaches to the same fusion level. The obtained results confirm that relying on a deep neural network for fusing individual decisions is far better than using handcrafted rules. Last, our intermediate fusion approach was better than all previous fusion approaches (according to all levels). As to prediction latency, our intermediate-based approach requires an average of **0.378 ms** for processing 100 frames which is far below the time required by the fastest approach, i.e., **0.413 ms**.

By analyzing more precisely the detailed scores per class for our best intermediate fusion-based approach, we were able to identify the reasons or the type of action that deteriorate our overall performance. In fact, there are mainly three major causes, which are more pronounced under the Cross-Subject protocol than under the Cross-View protocol.

First, for some classes, the definition of the action to be performed is very confusing which leads to having instances of a very different nature and which may even resemble some other action classes. This is mainly the case for class A22 (cheer up) where each subject interprets it in a different way and where the model predicts some of these instances as belonging to class A7 (throw) or A23 (hand waving). In fact, this misclassification is normal from a model point of view, since the subjects who performed the misclassified instances misinterpreted the class cheer up and performed it strongly as some subjects would perform those other classes with which they were confused.

A similar finding is made for some instances of two other classes namely A26 (hopping, i.e., one foot jumping) and A27 (jumping up) which are confused with each other, essentially the instances of class A26 classified as A27 actions. This is due to the fact that some subjects jump up while the camera hides their second leg, and thus the A27 class becomes more encompassing and includes training instances strongly similar to those in the A26 class.

Finally, the last confusion occurs between some two-subjects-based classes (called mutual) with single-subject classes. Indeed, for some instances of mutual actions, the data of the second person are not available or are noisy. In the absence of the second person, these instances are easily confused with single-person classes where the action to be done is very similar. This is the case for example of class A54 (point finger at the other person) which is confused with A31 (pointing to something with a finger).

To dive into our best fusion method, intermediate fusion, we show in Fig. 8 the evolution of the score curve on the validation set of the individual architectures according to the threshold of selected features. An important aspect to point

Table 6 Results of different fusion approaches proposed on the SBU Interaction dataset

Approach	Year	Fusion level	RGB	Skeleton	Depth	Accur.(%)	Latency (ms)
Two CNN [19]	2016	Late	X		X	85.06	0.536
ST-LSTM [27]	2017	Late	X	X		88.6	–
Two CNN [19]	2016	Intermediate	X		X	90.98	0.511
Combining CNN streams [25]	2018	Late	X	X	X	92.45	0.427
ST-LSTM [27]	2017	Intermediate	X	X		93.30	–
Combining CNN streams [25]	2018	Intermediate	X	X	X	96.67	0.403
MSAF [39]	2020	Intermediate	X	X		98.14*	0.351
VPN [7]	2020	Intermediate	X	X		98.77*	0.296
I3D [4]	2017	–	X			88.93	0.265
Shift-GCN [6]	2020	–		X		93.88	**0.203**
I3D [4]	2017	–			X	86.22	0.257
Our		Early	X	X	X	89.57	0.215
Our		**Intermediate**	**X**	**X**	**X**	**99.05**	0.370
Our		Late	X	X	X	95.61	0.386

*Stands for our own evaluation of the approach on this dataset
Best achieved performances along each configuration are set in bold

out is that these detailed results are obtained on the validation dataset, by changing at each time the minimum threshold to be reached by each feature-based variance (see Eq. 2).

From the curves in Fig. 8, we notice that the evolution of the recognition score according to both Cross-Subject and Cross-View protocols of each model follows almost the same pattern. It passes through three stages when the threshold increases gradually. In the first stage (left), we notice an increase in the performance of the three models when the minimum threshold increases (starting from 0), and the number of selected features decreases. This means that the features were too numerous, mostly irrelevant and impinged on classification performance.

This first stage ends at the performance of each model peaks (middle). For instance, the model based on RGB data achieved a success rate of **90.10%** on the validation set according to the Cross-Subject protocol, which corresponds to a threshold of **30** and to **925** selected features. However, for the depth map-based model, the features number corresponding to the best score, using the same Cross-Subject protocol and the same I3D features extractor, is around 530 with a variance of 70. This number is significantly lower than that selected for the RGB images and can be attributed to the richer nature of the RGB images compared to the depth data, which are much more sparse.

More globally, we note a strong disparity between the number of features to be retained according to each of the modalities considered, but also according to the basic model adopted and even according to the validation protocol chosen. Thus, the features to be selected are in fact impossible to deduce beforehand and that in this sense the proposed procedure is essential to determine the most accurate set of features.

In the last step (right), we notice a progressive performance decrease in each model with the increase in the minimum thresholds and the decrease in the number of selected features. This means that the recognition model does not have sufficient relevant features to distinguish between the classes as their number drastically decreases.

Based on these results, we determined the threshold value and consequently the features to select for the model we presented above. For instance, according to the Cross-Subject protocol, we took **30** as a minimum threshold for RGB images, **50** for skeletal data and **70** for depth maps when using the NTU RGB-D dataset. We then determined the indices of the extracted features exceeding that threshold for each modality, respectively.

Regarding the final performances of our approaches on the SBU Interaction dataset [42], we summarize in Table 6 the results achieved as well as those obtained by the previous approaches of the state of the art. We report both recognition accuracies and prediction latency in milliseconds. We also specify the type of fusion adopted by each approach and the modalities used.

We first notice that recognition accuracies, which were **89.57%**, **99.05%** and **95.61%** achieved by the early, intermediate and late fusion, respectively, exceeded the three scores obtained by the monomodal approaches, that were **88.93 %** for RGB images, **93.88 %** for skeletal sequences and **86.22 %** for depth maps. Hence the interest of multimodal fusion when compared with mono-modal recognition. Additionally,

Table 7 Recognition score evolution on the SBU Interaction dataset depending on the minimum variance threshold

	Minimum threshold	100	90	80	70	60	50	40	30	20	10
RGB	Reco. score (%)	–	–	22.38	36.73	50.12	69.28	80.65	**87.26**	83.48	80.77
Skeleton	Reco. score (%)	–	32.98	48.71	70.16	81.41	88.24	88.22	93.06	**95.62**	91.11
Depth	Reco. score (%)	–	–	28.70	35.09	44.47	57.26	70.14	**88.75**	86.56	81.93

Best achieved performances along each configuration are set in bold

we recorded only a small increase in prediction time, principally due to the multimodal fusion operation.

We also find that the intermediate fusion approach outperforms the early and late fusion approaches with an improvement of more than **10%** and **4%**, respectively. Besides, intermediate fusion corresponds to a reasonable prediction time with a value of **0.370 ms**. This suggests that even for a reduced set of action classes, the appropriate selection of features is the most important aspect to consider for the design of fusion strategies.

We notice, also, that our early fusion achieved good results compared to more complex approaches. On the other hand, the proposed late fusion approach outperformed the state-of-the-art late-based counterparts by a margin of **3%** while reducing the prediction time from **0.427** to **0.386 ms** for a set of 100 frames. This represents an enhancement and confirms the interest in learning-based decision-making as opposed to rule design.

Last, our intermediate fusion approach outperforms all previous fusion approaches. This suggests that our feature-selection-based fusion was very efficient. The prediction time was the least among all previous approaches, thus qualifying it for real-time application.

For an in-depth view of the intermediate fusion-based approach, we provide, in Table 7, the recognition rates obtained by increasing the minimum thresholds of accepted variance between the feature vectors of each modality. From these results, we notice that the selection phase is crucial for each modality. One can see that for a small variation of features number, the achieved performance vary significantly, as is the case of RGB modality when the threshold fluctuates from **40** to **30**. This suggests that the features should be carefully selected.

This allows us to conclude that relying on complex deep architectures is interesting as we could extract many features. Nevertheless, the extraction process should be with a feature selection procedure since the features are not equally valuable.

7 Conclusion

We proposed and compared three fusion strategies to improve multimodal action recognition. These strategies differ according to their level of intervention, in the action recognition pipeline, which can be either early, intermediate, or late.

Following an early fusion strategy, we proposed a two-step recognition approach. We first applied the depth mask to the corresponding RGB images to take advantage of the rich and powerful visual texture of RGB images. We then projected the skeletal sequences onto the resulting image to guide to focus on the subject.

Regarding the intermediate fusion, we proposed a deep learning approach for the fusion of features based on a qualitative feature selection method in order to choose the most distinctive ones. The main idea of the proposed selection method is to identify the most diverse features among the classes for each modality.

Finally, we proposed an end-to-end trainable late fusion approach based on a deep neural network. To this end, we relied on three pre-trained architectures to generate score vectors from each modality individually. After pre-processing, the feature vectors are delivered to our network for training.

By following standard experimental protocols, we conducted extensive ablation studies and reported the results obtained by each of the three proposed strategies on two challenging datasets. With a Cross-Subject score of **95.94%** and a Cross-View score of **98.11%** on the NTU RGB-D dataset and a score of **99.05%** on the SBU Interaction dataset, we found that the fusion at the intermediate level has the strongest impact in the multimodal recognition process more than the early and late fusion. Moreover, the intermediate fusion strategy gave the reasonable latency on both datasets, which was, respectively, **0.378 ms** and **0.370 ms** on average.

Our future work will be directed toward improving the search strategies for the best features of the intermediate approach by defining other criteria in addition to variance, and better multimodal fusion strategies.

Declarations

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

References

1. Bouderbal, I., Amamra, A., Benatia, M.A.: How would image down-sampling and compression impact object detection in the context of self-driving vehicles? In: CSA, pp. 25–37 (2020)
2. Boulahia, S.Y., Anquetil, E., Multon, F., Kulpa, R.: Dynamic hand gesture recognition based on 3d pattern assembled trajectories. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6. IEEE (2017)
3. Boulahia, S.Y., Anquetil, E., Multon, F., Kulpa, R.: Cudi3d: curvilinear displacement based approach for online 3d action detection. Comput. Vis. Image Understanding **174**, 57–69 (2018)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
5. Chen, C., Jafari, R., Kehtarnavaz, N.: Fusion of depth, skeleton, and inertial data for human action recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2712–2716. IEEE (2016)
6. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 183–192 (2020)
7. Das, S., Sharma, S., Dai, R., Bremond, F., Thonnat, M.: Vpn: Learning video-pose embedding for activities of daily living. In: European Conference on Computer Vision, pp. 72–90. Springer (2020)
8. Davoodikakhki, M., Yin, K.: Hierarchical action classification with network pruning. In: International Symposium on Visual Computing, pp. 291–305. Springer (2020)
9. De Boissiere, A.M., Noumeir, R.: Infrared and 3d skeleton feature fusion for rgb-d action recognition. IEEE Access **8**, 168297–168308 (2020)
10. Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A., Beghdadi, A.: A combined multiple action recognition and summarization for surveillance video sequences. Appl. Intell. **51**(2), 690–712 (2021)
11. Fan, Y., Weng, S., Zhang, Y., Shi, B., Zhang, Y.: Context-aware cross-attention for skeleton-based human action recognition. IEEE Access **8**, 15280–15290 (2020)
12. Franco, A., Magnani, A., Maio, D.: A multimodal approach for human activity recognition based on skeleton and rgb data. Pattern Recogn. Lett. **131**, 293–299 (2020)
13. Gravina, R., Alinia, P., Ghasemzadeh, H., Fortino, G.: Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges. Inf. Fusion **35**, 68–80 (2017)
14. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6546–6555 (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
17. Hu, J.F., Zheng, W.S., Pan, J., Lai, J., Zhang, J.: Deep bilinear learning for rgb-d action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 335–351 (2018)
18. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
19. Ijjina, E.P., Chalavadi, K.M.: Human action recognition in rgb-d videos using motion sequence information and deep learning. Pattern Recogn. **72**, 504–516 (2017)
20. Imran, J., Raman, B.: Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition. J. Ambient Intell. Hum. Comput. **11**, 1–20 (2019)
21. Islam, M.M., Iqbal, T.: Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. Preprint arXiv:2008.01148 (2020)
22. Islam, M.M., Iqbal, T.: Multi-gat: a graphical attention-based hierarchical multimodal representation learning approach for human activity recognition. IEEE Robot. Autom. Lett. **6**(2), 1729–1736 (2021)
23. Jegham, I., Khalifa, A.B., Alouani, I., Mahjoub, M.A.: Vision-based human action recognition: an overview and real world challenges. For. Sci. Int. Digital Investig. **32**, 200901 (2020)
24. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: Mmtm: Multimodal transfer module for cnn fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13289–13299 (2020)
25. Khaire, P., Kumar, P., Imran, J.: Combining cnn streams of rgb-d and skeletal data for human activity recognition. Pattern Recogn. Lett. **115**, 107–116 (2018)
26. Lin, W., Sun, M.T., Poovandran, R., Zhang, Z.: Human activity recognition for video surveillance. In: IEEE International Symposium on Circuits and Systems, pp. 2737–2740 (2008)
27. Liu, J., Shahroudy, A., Xu, D., Kot, A.C., Wang, G.: Skeleton-based action recognition using spatio-temporal lstm network with trust gates. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 3007–3021 (2017)
28. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 143–152 (2020)
29. Lockhart, J.W., Pulickal, T., Weiss, G.M.: Applications of mobile activity recognition. In: Proceedings of the ACM Conference on Ubiquitous Computing, pp. 1054–1058 (2012)
30. Memmesheimer, R., Theisen, N., Paulus, D.: Gimme signals: discriminative signal encoding for multimodal activity recognition. Preprint arXiv:2003.06156 (2020)
31. Pérez-Rúa, J.M., Vielzeuf, V., Pateux, S., Baccouche, M., Jurie, F.: Mfas: Multimodal fusion architecture search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6966–6975 (2019)
32. Pham, C., Nguyen, L., Nguyen, A., Nguyen, N., Nguyen, V.T.: Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks. Multimedia Tools and Applications pp. 1–22 (2021)
33. Rodríguez-Moreno, I., Martínez-Otzeta, J.M., Goienetxea, I., Rodriguez-Rodriguez, I., Sierra, B.: Shedding light on people action recognition in social robotics by means of common spatial patterns. Sensors **20**(8), 2436 (2020)
34. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Conference on Computer Vision and Pattern Recognition, pp. 1010–1019. IEEE (2016)
35. Shahroudy, A., Ng, T.T., Gong, Y., Wang, G.: Deep multimodal feature analysis for action recognition in rgb+ d videos. IEEE Trans. Pattern Anal. Mach. Intell. **40**(5), 1045–1058 (2017)
36. Shahroudy, A., Wang, G., Ng, T.T.: Multi-modal feature fusion for action recognition in rgb-d sequences. In: 6th International Symposium on Communications, Control and Signal Processing (ISCCSP), pp. 1–4. IEEE (2014)
37. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recogni-

tion in parts from single depth images. In: CVPR, pp. 1297–1304. IEEE (2011)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint arXiv:1409.1556 (2014)
39. Su, L., Hu, C., Li, G., Cao, D.: Msaf: Multimodal split attention fusion. Preprint arXiv:2012.07175 (2020)
40. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
41. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
42. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 28–35. IEEE (2012)
43. Zhao, R., Ali, H., Van der Smagt, P.: Two-stream rnn/cnn for action recognition in 3d videos. In: RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4260–4267. IEEE (2017)
44. Zhao, Y., Liu, Z., Yang, L., Cheng, H.: Combing rgb and depth map features for human activity recognition. In: Proceedings of The Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–4. IEEE (2012)
45. Zhu, Y., Chen, W., Guo, G.: Fusing multiple features for depth-based action recognition. ACM Trans. Intell. Syst. Technol. (TIST) **6**(2), 1–20 (2015)
46. Zolfaghari, M., Oliveira, G.L., Sedaghat, N., Brox, T.: 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 (2017)

Said Yacine Boulahia received an engineering degree in computer science from Ecole Militaire Polytechnique, Algeria, in 2014, a Master's degree in computer science from Rennes 1 University, France, in 2015, and a Ph.D. in computer science from INSA Institute, France in 2018. He joined Artificial Intelligence and Virtual Reality laboratory in Ecole Militaire Polytechnique, Algeria, as a permanent researcher in late 2018. His current research interests include action recognition, medical image processing, computer vision, machine learning, and deep learning.

Abdenour Amamra received an engineering degree in computer science from Ecole Militaire Polytechnique, Algeria, in 2011, and a Ph.D. in computer science from Cranfield University, UK, in 2013. In 2015 he joined Artificial Intelligence and Virtual Reality laboratory in Ecole Militaire Polytechnique, Algeria, as a permanent researcher. In 2019, he obtained his habilitation in computer science. His current research interests include visual recognition, object tracking for virtual reality, and 3D simulation.

Mohamed Ridha Madi has received an engineering and Master's degree in computer science from Ecole Militaire Polytechnique, in 2020, Algeria. His current research interests include pattern recognition, data fusion, and deep learning.

Said Daikh has recently received an engineering and Master's degree in computer science from Ecole Militaire Polytechnique, in 2020, Algeria. His current research interests include video classification, image processing and neural network architecture design.