Multimedia Tools and Applications (2019) 78: 5919–5939
https://doi.org/10.1007/s11042-018-6875-7

Fusing depth and colour information for human action recognition

Danilo Avola[1,2] · **Marco Bernardi**[2] · **Gian Luca Foresti**[1]

Received: 4 October 2017 / Revised: 8 November 2018 / Accepted: 13 November 2018 /
Published online: 27 November 2018

Abstract

In recent years, human action recognition systems have been increasingly developed to support a wide range of application areas, such as surveillance, behaviour analysis, security, and many others. In particular, data fusion approaches that use depth and colour information (i.e., RGB-D data) seem to be particularly promising for recognizing large classes of human actions with a high level of accuracy. Anyway, existing data fusion approaches are mainly based on feature fusion strategies, which tend to suffer of some limitations, including the difficult of combining different feature types and the management of missing information. To address the two problems just reported, we propose an RGB-D data based human action recognition system supported by a decision fusion strategy. The system, starting from the well-known Joint Directors of Laboratories (JDL) data fusion model, analyses human actions separately for each channel (i.e., depth and colour). The actions are modelled as a sum of visual words by using the traditional Bag-of-Visual-Words (BoVW) model. Subsequently, on each channel, these actions are classified by using a multi-class Support Vector Machine (SVM) classifier. Finally, the classification results are fused by a Naive Bayes Combination (NBC) method. The effectiveness of the proposed system has been proven on the basis of three public datasets: UTKinect-Action3D, CAD-60, and LIRIS Human Activities. Experimental results, compared with key works of the current state-of-the-art, have shown that what we propose can be considered a concrete contribute to the action recognition field.

Keywords Human action recognition · Decision level fusion · Bag-of-visual-word · Naive bayes combination · Support vector machine · RGB-D

✉ Danilo Avola
avola@di.uniroma1.it

Marco Bernardi
bernardi@di.uniroma1.it

Gian Luca Foresti
gianluca.foresti@uniud.it

[1] Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy

[2] Department of Computer, Sapienza University, Rome, Italy

1 Introduction

Vision-based human action recognition is a fascinating and challenging field of the modern computer vision. Systems designed to support this field interpret automatically human actions from a set of observations acquired by different sensors, such as RGB cameras, thermal cameras, Time-of-Flight (ToF) cameras, and others. In the last two decades, these systems have been increasingly used in various application areas, including monitoring and surveillance [28, 57, 71], healthcare and rehabilitation [7, 22], robotics [11, 55], and many others, as reviewed in [1, 2, 67]. Most of the existing works generally use one or more conventional RGB cameras [18, 73]. These devices are extremely effective for this kind of applications, but they have different well-known limitations, such as sensitivity to the illumination changes, inability to provide useful information on overlapping subjects, difficulty of handling the viewpoint variations, management of the background clutter, and so on. The recent widespread of low-cost sensors able to provide a depth map of an area of interest (e.g., Microsoft Kinect, Asus Xtion Pro) has led to consider also their use within human action recognition systems [2, 60]. These sensors are able to address some of the previously introduced limitations (e.g., overlapping subjects) but, at the same time, they suffer of other limitations, such as sensitivity to the sunlight, interference issues due to the cameras placed one facing each other, error measurements introduced by the reflective materials, and so on. In addition, depth maps are not able to provide the aspect of the surface of an object or a subject. To overcome the limitations of the RGB cameras and depth sensors, respectively, and to join their main strengths, many researches, in the last years, have been focused on the design of approaches to efficiently fuse depth and colour information [47, 65].

Multi-sensor data fusion is an open problem in computer vision aimed to coherently merge together different data types to provide a better interpretation of actions or events that occur within a video sequence [37]. One of the most popular models that describes the process of combining different data coming from various sources is the Joint Directors of Laboratories (JDL) data fusion model [26]. This model (shown in Fig. 1) can be considered

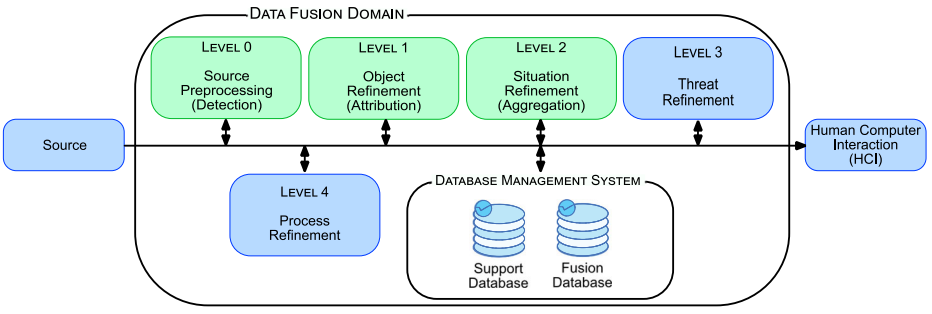

Fig. 1 JDL data fusion model. Source: Input data streaming (i.e., RGB-D data); Level 0: Pre-processing of the acquired video sequences (e.g., pixel adjustment, noise reduction, and so on); Level 1: Estimation and prediction of the subjects' state on the basis of the inferences due to the observations (i.e., SIFT and SURF feature extraction from the RGB and depth streams and clustering by using the BoVW); Level 2: Estimation and prediction of the subjects' state on the basis of the inferred relations among them (i.e., the sets of visual words are classified, on each stream, by using a multi-class SVM and then fused together by using a NBC); Level 3: Estimation of the current actions/events and prediction of the next actions/events on the basis of the available information; Level 4: Meta-level that monitors the data fusion process to assess the real-time system performance; Support Database: Maintaining of native information; Fusion Database: Maintaining of the fused information; HCI: Interaction between the system and one or more subjects

Table 1 Association between the JDL data fusion model and the system's functionalities

Level	Main process	Main methods	Main functionalities
0	Detection	RGB and depth streams	Data acquisition
1	Attribution	Descriptor extraction (SIFT, SURF) bag of visual world	Feature extraction clustering
2	Aggregation	Multi-class support vector machine bayes naive fusion	Information fusion action recognition

a functional framework to define the set of functions by which to implement any kind of data fusion system. In our context, we have adopted part of this model to define and develop the proposed human action recognition system. In particular, we have customized the first three levels of the model to fit the proposed system and removed the last two levels since they are neither suitable nor necessary to the functionalities of what we propose.

Table 1 summarizes how the implemented processes and techniques have been mapped within the first three levels of the JDL data fusion model. The multi-modal data acquisition is performed at Level 0. Subsequently, at Level 1, the Scale-Invariant Feature Transform (SIFT) [48] and Speeded Up Robust Features (SURF) [9] descriptors are extracted from the RGB and depth streams, respectively. In addition, these descriptors are clustered by using the Bag-of-Visual-Words (BoVW) model [20]. Finally, at Level 2, the sets of visual words are classified, on each stream, by using a multi-class Support Vector Machine (SVM) classifier [16]. Finally, the classification results are fused together by using a Naive Bayes Combination (NBC) [45] strategy. From a practical viewpoint, the proposed customization of the JDL data fusion model performs, at Level 0, the detection of subjects within the area of interest, while their interpretation is achieved at Level 1. Finally, at Level 2, the final decision on the recognized actions is taken.

According to the JDL data fusion model, there are, in general, three main strategies to fuse information (Fig. 2): *raw level fusion* (also named *low level fusion*), *feature level fusion* (also named *medium level fusion*), and *decision level fusion* (also named *high level fusion*). In the first strategy, the sensors directly provide data as an input to the fusion process, which derives a more accurate aggregate data than the individual sources. The feature extraction and classification processes are performed uniquely on this latter data. In the second strategy, different feature vectors (e.g., shapes, colours, trajectories) are fused to obtain more complex features on which a final classification step is carried out. Finally, in the last strategy, each sensor individually acquires data, extracts features, and classifies them. A final decision is taken by evaluating together the different classifications. Usually, raw level fusion is used when data come from homogeneous multi-sensors (e.g., two or more RGB cameras) [39]. In this case, techniques for raw data fusion typically involve well-known estimation methods, such as Kalman filtering and many others, as summarized in [26]. When, instead, the system is composed of heterogeneous multi-sensors (e.g., RGB-D cameras), strategies based on feature level fusion or decision level fusion are preferred. The choice between the last two strategies depends on several aspects, including kind of area of interest and assigned task. The feature level fusion strategies are often supported by robust techniques (e.g., multiple kernel learning [50], canonical correlation analysis [15]) that are used to combine different classes of features in an unique pattern by which to identify a specific human action. Although widely used, these strategies have many limitations. In particular, a first critical aspect regards the combination of different feature types, and a second is related to the achievement of a final decision when the information of one or more sensors is

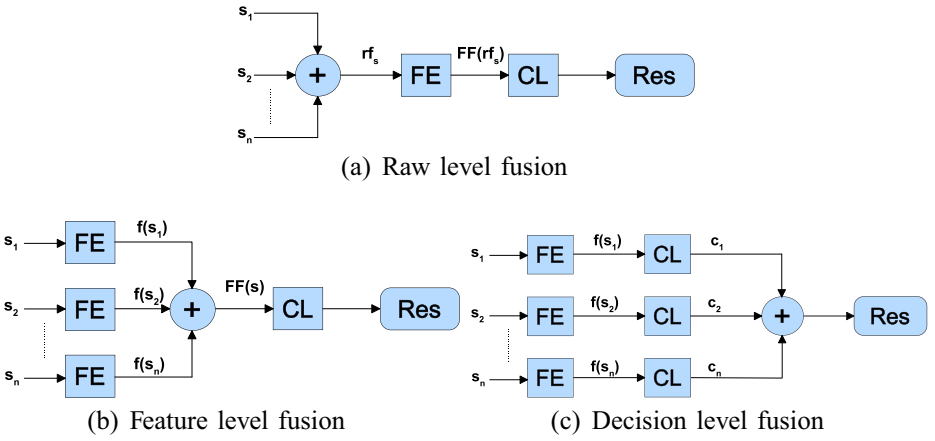

Fig. 2 Typical fusion strategies. In each figure, $s_i \ \forall i \in [1..n] \subset \mathbb{N}$ is a specific source (e.g., an RGB stream), the FE block is the feature extraction process, the CL block is the classification process, the Res block is the result's container. In **a**, rf_s and $FF(rf_s)$ are the univocal representations of the fused data and the result of the feature extraction process on it, respectively. In **b** and **c**, the $f(s_i)$ is the result of the feature extraction process on each s_i. In **b**, the $FF(s)$ is the univocal representation of the fused features, while in **c**, each c_i $\forall i \in [1..n] \subset \mathbb{N}$ is the result of the classification process related to each source s_i

missing. Decision level fusion strategies can offer a good compromise to overcome the two limitations just reported. They are based on the assumption that each sensor can contribute to make more or less reliable the classification of a specific human action. These strategies are often based on the customization of different probabilistic methods (e.g., NBC [45], Dempster-Shafer's approach [39]) designed to produce a ranking of the most plausible recognized actions. In our context, we have considered this last class of strategies to implement a robust decision mechanism. In the next subsection, motivation and contribution of the proposed paper are highlighted.

1.1 Motivation and contribution

The system proposed in this paper presents some novelties compared to the current state-of-the-art:

- It has been developed by fitting the implemented processes and techniques within the first three levels of the JDL data fusion model (see Table 1). This aspect allows the system to inherit the basic characteristics of the model, including versatility and extendibility.
- It is the first action recognizer that uses a decision fusion strategy based on an NBC technique supported by both a BoWV based clustering and a multi-class SVM on each acquired stream (see Section 3).
- It is robust to the different lengths of the acquired video sequences. In fact, thanks to the use of both keypoint descriptors and BoVW, the system processing is independent from the video frame rate. This aspect allows to recognize human actions even in subsequences of arbitrary length. In addition, thanks to the use of the keypoint descriptors, it is possible to detect better the local image details, even in presence of scale, rotation, and translation, thus improving the recognition rate of the actions. Despite the

method we propose does not treat explicitly the temporal sequences, the advantages just reported above, including the temporal independence, and the ability of the method to synthesize an action occurred over time, allow to the system to achieve very good performance. Actually, this feature has been inspired by some works in the current literature that, for similar goals, have used temporal independence strategies [21, 75].

– Most of the works in human action recognition are focused on feature level fusion strategies [47, 51] and often remarkable results are obtained [62]. Despite this, these strategies present different limits [59, 63]. First of all, the fusion of heterogeneous feature vectors can be considered a very hard task. In addition, when ad-hoc algorithms are designed to manage together these vectors, the extendibility property is not usually maintained. Finally, when a classifier is designed to work on multi-modal data, the unavailability of a single source can provoke the failure of the entire recognition process. To overcome these problems, the proposed system is based on a decision fusion strategy able to support the consistent treatment of the incoming data since it is processed separately for each source. Thanks to this last aspect, the fusion process can be performed by well-known statistical approaches [24]. The system improves the recognition performance with respect to the single modalities and, at the same time, it is effective and robust with respect to the dual modality (see Section 3).
– In the current literature, this kind of systems is evaluated by using either the 2-Fold Cross-Validation (2FCV) in which the best performance is sought through the most appropriate division, in two equal parts, i.e., training and evaluation, of the videos contained inside the dataset, or the Leave-One-Out Cross-Validation (LOOCV) in which the same goal is pursued, but looking for an optimal division based on different percentages, i.e., ninety percent for the training and ten percent for the evaluation [32]. To provide a wide comparison with the current state-of-the-art, we have reported results with both metrics by using the UTKinect-Action3D dataset, a very popular collection of videos containing different action types [70]. In both cases, the results have shown that the proposed system can be considered a concrete contribution to the current literature in human action recognition. To confirm the validity of the obtained results, we have also tested the system with other two datasets: LIRIS Human Activities [69] and CAD-60 [64, 65]. The first is a collection of challenging videos also oriented to surveillance aims. The second is a collection of RGB-D video sequences of humans performing daily activities. All the additional tests have confirmed the good performance of the system (see Section 4).

The rest of the paper is structured as follows. In Section 2, selected key works of the current state-of-the-art that use a single or a dual modality are presented. Section 3 describes the proposed method underlining the logical RGB-D based architecture. Section 4 reports the experimental results and discusses the comparison of the proposed system with respect to the selected key works of the current literature. Finally, Section 5 concludes the paper and highlights key considerations.

2 Related work

As depicted in Fig. 3, human action recognition systems based on an RGB camera, a depth sensor, or an RGB-D camera can be divided into two main classes. The first class (single modality) can be in turn divided into two sub-classes containing the colour-based systems [8, 10, 14, 46] and the depth-based systems [25, 32, 49, 58, 70], respectively. The

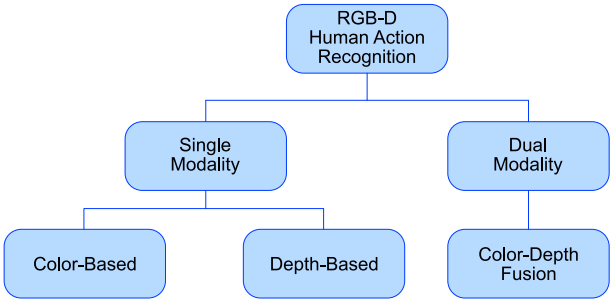

Fig. 3 A Taxonomy for human action recognition systems. Single modality (i.e., an RGB camera or a depth sensor). Dual modality (i.e., an RGB-D camera)

second class (dual modality) contains instead all those systems that fuse together colour and depth information to achieve a better recognition performance [17, 47, 50, 51, 65].

The human action recognition by color-based approaches is a very active research area and different methods, in recent years, have been proposed. A first reference work is shown in [46], where the authors use a set of Space-Time Interest Points (STIPs) to recognize realistic human actions in unconstrained videos. The classification process is subsequently performed by using a multi-class SVM. Other works have instead used the BoVW model achieving very interesting results. A first example is reported in [8], where the authors uses it to represent actions as a sequence composed of histograms of visual features. Each sequence is treated as a string where each histogram is considered as a character. The classification process is performed by using different SVM classifiers with a string kernel that uses the Needleman-Wunsch edit distance. Another example is presented in [14], where the authors propose a BoVW-like model in combination with a 2D descriptor (Harris-SIFT) and a motion vector histogram for the detection of human actions in video streams. The last two works show how the clustering of visual features (or, in general, visual words) can be profitably used to reach high performance even in complex video sequences. In particular, the works highlight how the well-known issue of the BoVW model, i.e., the management of the temporal information between consecutive frames, can be suitably faced. Anyway, the use of only one RGB channel can present different limitations due to several aspects, such as illumination changes, overlapped subjects, and many others. For these reasons, recently, other sensors have been explored to support this application area.

The depth-based approaches [19, 23, 53] adopt different sets of features compared with those obtained by the RGB cameras [2, 44]. In particular, these approaches are often based on the measurement of specific parts (i.e., joints) of the human skeleton models that are derived by the processing of the depth maps [3, 5]. An example of these approaches is reported in [70], where the authors use a global feature representation of the entire sequence. More specifically, the authors propose an action recognition system that uses the histograms of 3D joint locations as a compact representation of postures. The temporal evolutions of these latter, represented by visual words, are modelled by discrete Hidden Markov Models (HMMs). This approach allows authors to manage the different lengths of the video sequences, which is one of the main issues during the classification process. However, also these systems present several limitations, such as the misinterpretation of some movements or the sharing of specific movements in different action classes [56]. Like for the

RGB cameras, also the action recognition by depth sensors can fail in crowded or challenging scenarios due to the occlusions. To support these issues, the work presented in [41] describes a BoVW based approach that combines motion and 3D information to improve the action recognition performance with a low movement rate. Anyway, the systems implemented by consumer depth sensors tend to suffer of different limitations, including sensitivity to the sunlight, error measurements related to the reflective materials, and many others.

To take advantages from the classes of techniques that treat the single modalities, i.e., colour-based or depth-based, recent years have seen several research groups intensify the efforts on fusion techniques. The most popular solutions are based on feature level fusion strategies. A first key work is described in [47], where the authors propose a human action recognition system based on a coupled hidden conditional random fields model. In [50], instead, the authors present a system whose fusion of colour and depth information is performed by a Multiple Kernel Learning (MKL) approach. Starting from the idea of using the fusion step from a deeper level, the work proposed in [62] describes a new hierarchical Bag-of-Words (BoWs) feature fusion technique based on multi-view structured sparsity learning to merge atomic features from RGB and skeletons. As previously introduced (see Section 1.1) and remarked in [43], the feature level fusion strategies have several limitations, including difficulties in merging heterogeneous features, difficulties in managing the lack of information, and many others.

In the context reported above, a particular reference is due to the use of the Convolutional Neural Networks (CNNs) [19, 23]. In very recent years, some researchers have started to explore their use to address the human action recognition area, thus obtaining remarkable results [17, 31, 36]. In [31], for example, the authors define a new representation able to provide emphasis to the key poses associated with each action. The features obtained from motion, in RGB and depth video streams, are given as input to a CNN to learn the discriminative features. In [17], instead, the authors propose an action recognition method based on a mixture of RGB and depth skeletons, supported by a deep CNN. Finally, in [36], the authors propose an original framework that combines CNN streams of RGB-D and skeletal data for the recognition of challenging human actions.

Unlike the previous works, this paper proposes an RGB-D human action recognition system based on a decision level fusion strategy. The proposed architecture based on the customization of the JDL data fusion model provides an alternative method and a concrete contribution to the current literature.

3 RGB-D based human action recognition system

In Fig. 4, the logical architecture of the proposed system is shown. Each action is represented by a video sequence that may differ in length depending on the frame rate and time duration of the action itself. The system is composed of two main sub-modules. The first sub-module, i.e., action recognition, extracts from each RGB and depth sequence of frames a set of visual features. Then, each set is synthesized by means of a statistical model. Finally, each model is labelled by using a multi-class classifier. In the training stage, each class, i.e., C_{rgb} and C_{depth}, is empirically evaluated and associated with a numeric weight to set its reliability level in recognizing the related action. In the second sub-module, i.e., fusion, the reliability of the two classes is merged through a simple voting method.

Fig. 4 Architecture of the proposed human action recognition system. In the action recognition phase, from each RGB and depth stream a set of SIFT and SURF descriptors are extracted, respectively. Each set is synthesized by means of a BoVW model and labelled (i.e., C_{rgb} and C_{depth}) by means of a multi-class SVM classifier. In the fusion phase, the numerical weights of the classes are used as input of an NBC classifier to provide the result

3.1 Action recognition phase

Initially, from each frame of the RGB and depth streams, which represent the same action, the SIFT and SURF local descriptors are extracted, respectively. Then, each frame is modelled by a BoVW technique to obtain a set of visual words. In addition, for each stream, the different BoVW histograms are pooled to provide a unique representation of the action in terms of sum of visual words. Finally, a multi-class SVM based on Radial-Bases Function (RBF) kernel is used to classify each action. During the training stage, a set of videos is used both to train the system and to compute a set of numerical weights representing the reliability of the system itself in recognizing the different action classes. In the evaluation stage, the weights are used by means of an NBC classifier to merge the decisions of the two channels (i.e., RGB and depth) in a unique final result. In Fig. 5 a running example focused on depth video streams is depicted. As just reported in Section 1.1, the method we propose, inspired by the works [21, 75], is independent from the frame rate of the video sequences.

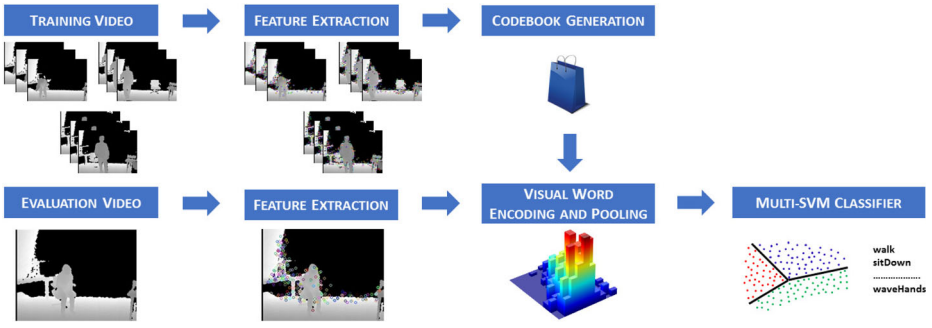

Fig. 5 Running example of the action recognition phase focused on depth video streams. During the training stage, from each frame that composes a video sequence (i.e., an action) the SURF descriptors are extracted and clustered by means of a BoVW technique. The different histograms are pooled in an unique BoVW to obtain a compact representation of the video. Finally, a multi-class SVM is used to label each action

This is due to the fact that the descriptors collected by the BoVW are synthesized in a unique measure independently from their amount. This simple, but effective approach, also allows the system to suitably manage video sub-sequences of arbitrary length.

3.1.1 Feature extraction

The set of training video sequences produces a codebook obtained by the quantization of the local descriptors related to each action. In other words, each frame of a video sequence generates a visual word frequency vector and the pooling of these vectors contributes to form an element of the codebook. In this work, two different feature extraction and description methods are used: SIFT and SURF. The first method has been chosen to analyse RGB video streams thanks to its high level of reliability in recognizing human actions on this kind of media [8, 52]. Recent works of the current literature have shown that the SURF method is especially suitable to manage the depth video streams [66]. In particular, this method has proven to be extremely fast and robust in analysing different kinds of depth maps. In this work, we have used the keypoint descriptors because they allow to distinguish better local image details than low-level features (e.g., texture, colours, edges) [6]. Moreover, they enable the recognition of actions even in presence of slightly differences in terms of scale, rotations, and translations.

3.1.2 Bag of visual words

For each channel, the descriptors extracted from each frame that composes a video sequence representing an action are clustered by means of a BoVW approach. This technique is utilized to easily synthesize the informative content of each frame without using complex data, such as shape and skeleton of the subjects in RGB images and depth maps, respectively. More specifically, the BoVW technique adopts a K-Means strategy [27] to partition the descriptors of each frame, thus providing an abstraction of it in terms of visual words. A visual word is nothing more than a sparse vector of occurrences, i.e., a histogram of the different descriptors within a frame. All the visual words of a same video are further summarized by using another BoVW whose purpose is to provide a semantic representation of the analysed video. All the obtained representations contribute to form the codebook of the system through which to classify the different actions. The pooling method related to an action can be expressed as follows [54]:

$$w_k = \sum_{i=1}^{n} w_{k,i} \tag{1}$$

where, w_k is an element of the codebook, $w_{k,i}$ is the sparse vector of a frame i, and n is the amount of sparse vectors for an action.

3.1.3 Multi-class support vector machine

The SVMs are state-of-the-art classifiers that have gained great popularity in a wide range of application areas [8, 60]. In this work, a multi-class SVM classifier for human action recognition is used. The chosen classifier utilizes the one-against-all method with an RBF kernel [61]. The latter, given x and y as feature vectors, can be defined as follows:

$$K(x, y) = \exp(-\gamma \|x - y\|^2) \tag{2}$$

where, γ is the spread of the kernel. To estimate the best value of this parameter, an automatic selection based on the minimization of the similarity between the x and y vectors is used. Notice that, the γ parameter sets the width of the bell-shaped curve. The larger the value of γ the narrower will be the bell. Moreover, small values of γ yield wide bells. In other words, a suitable value of γ can increase the overall accuracy of the multi-class SVM classifier [13].

3.2 Fusion phase

The fusion phase is performed by a decision level strategy in which the results of the classifiers linked to each channel are used as input for another classifier to obtain a final prediction. In the proposed system, this last step is carried out by an NBC technique [45] due to its high reliability in ensuring the minimum error rate during the classification process [38]. Moreover, this simple technique is particularly suitable for the required task.

The NBC fusion technique assumes that the classifiers are mutually independent with respect to the given class labels. This means that for each classifier C_i, an $n \times n$ confusion matrix CM_i can be computed by applying C_i to the training dataset. More specifically, let (k, s) a cell of this matrix, then $cm^i_{(k,s)}$ is the number of elements whose true class label is k and whose class is s assigned by the classifier C_i. In this context, $N_{i,s}$ can be defined as the total number of elements labelled by the classifier C_i into the class s (this value is calculated as the sum of the s^{th} column of CM_i). By using these parameters, an $n \times n$ label matrix LM_i can be derived, in which the $(k,s)^{th}$ entry, i.e., $lm^i_{(k,s)}$, is an estimate of probability (i.e., the numerical weight) that the true label k is actually associated to the label s for the class C_i. In addition, let v a feature vector, then $lm^i_{(k,s)}$ can be computed as follows:

$$lm^i_{(k,s)} = P(k|C_i(v)) = \frac{cm^i_{(k,s)}}{N_{i,s}} \tag{3}$$

Finally, let $s_1, \ldots, s_L$ be the class labels assigned to v by the classifiers $C_1, \ldots, C_L$. Then, thanks to the independence assumption, the estimate of the probability that the true class label is k, can be determined as follows:

$$\mu_k(v) = \prod_{i=1}^{L} P(k|C_i(v)) = \prod_{i=1}^{L} lm^i_{k,s} \tag{4}$$

A running example of the proposed NBC fusion algorithm is shown in Fig. 6. Simplifying, the reliability of the classes during the evaluation stage is derived by a pre-assessment phase in which the same videos used for the training stage are used to compute a pre-evaluation stage. The proposed approach presents several advantages. First, more sensors, i.e., channels, can be added without altering the proposed architecture. Second, thanks to the scalability of the JDL model and thanks to the simple fusion approach the system can monitor very wide areas. Finally, thanks to the basic information required by the method, i.e., the keypoints, the recognition can be done in real-time.

4 Experimental results

In this section, the evaluation of the proposed system and a comparison with selected key works of the current state-of-the-art are reported. Initially, the system was tested through the 2FCV and LOOCV approaches using the UTKinect-Action3D dataset [70]. Subsequently,

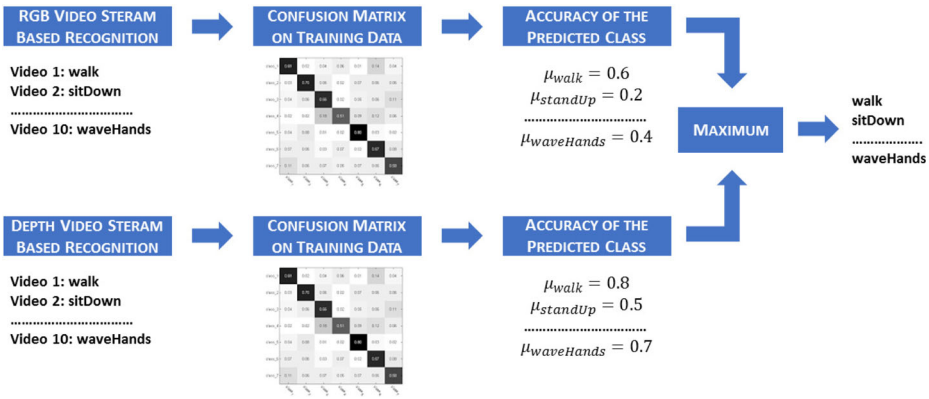

Fig. 6 Running example of the fusion phase. During the evaluation stage, an NBC classifier performs the fusion operation. A numerical weight representing reliability in recognizing the related action is associated to each predicted hypothesis. The final result is established by choosing the majority likelihood for each couple of hypotheses

to confirm the validity of the obtained results and to stress the proposed method with other challenging video sequences, the system was also tested, through the LOOCV approach, using two other datasets: CAD-60 [64, 65] and LIRIS Human Activities [69].

4.1 Dataset description and experimental protocol

The UTKinect-Action3D dataset was created by using a single stationary Microsoft Kinect with Kinect for Windows SDK Beta Version. The dataset contains 10 action types: walk (WK), sit down (SD), stand up (SU), pick up (PU), carry (CR), throw (TW), push (PS), pull (PL), wave hands (WH), and clap hands (CH). The different video sequences were acquired through 10 subjects, where each subject performed each action twice. The acquisition was carried out on three synchronized channels, i.e., RGB, depth, and skeleton joint locations, of which only the first two channels were used to test the proposed system. The video sequences were acquired at 30 frames-per-second (fps), however, since the dataset authors recorded frames only when the skeleton was tracked, the final frame rate of each sequence is about 15 fps. The dataset is an unique collection of videos and users can choose how splits it in training and evaluation sequences. The UTKinect-Action3D dataset was mainly chosen because it is one of the few datasets also used in other works to treat the topic reported in the present paper, thus allowing a comparison with other systems. In Fig. 7 some examples of dataset frames are shown.

The LIRIS Human Activities dataset was created by using both a Microsoft Kinect and a Sony Consumer Camcoder. The dataset contains several video sequences showing people performing various activities taken from daily life, such as: discussing, telphone calls, giving an item, and so on. Also in this dataset the acquisition was performed on three channels, RGB, depth, and gray, and we have used only the first two channels. The dataset is fully annotated, where the annotations not only report information on the action class but also its spatial and temporal positions in the video. Moreover, the video sequences are already organized into training and evaluation partitions. The LIRIS human activities dataset was chosen because allows the proposed system to be tested with challenging videos. In particular, we selected the following three videos of interest: baggage, door, and object. The first simulates

Fig. 7 Examples of actions of the UTKinect-Action3D dataset

the abandonment of a baggage in a corridor, the second captures the attempt of opening a closed door, and the last shoots an exchange of an object between two subjects. All three human activities are of great interest in challenging contexts, such as video surveillance and behaviour analysis. In Fig. 8 some frames of the three actions are shown.

The CAD-60 dataset is an RGB-D dataset acquired with a Microsoft Kinect sensor at 30 fps, whose videos have a resolution of 640×480 pixels. The dataset contains 14 daily human activities performed indoors by 4 subjects (two males and two females). The total number of frames for each activity of each person is about one thousand. For the experiments, we have used the setting "New Person" (a sort of LOOCV approach) described in [64]. In this setting, the data of 3 subjects were used for training and the remaining one subject for testing. The CAD-60 dataset was chosen because it is one of the most challenging dataset used, as for the UTKinect-Action3D dataset, in other works related to the human action recognition field, thus allowing a comparison on different types of video sequences. In Fig. 9 some dataset frames are shown.

The proposed approach was implemented in C++ by VS 2017 IDE and OpenCV 3.1 framework. The implementation of the multi-class SVM is based on LibSVM [12]. Finally,

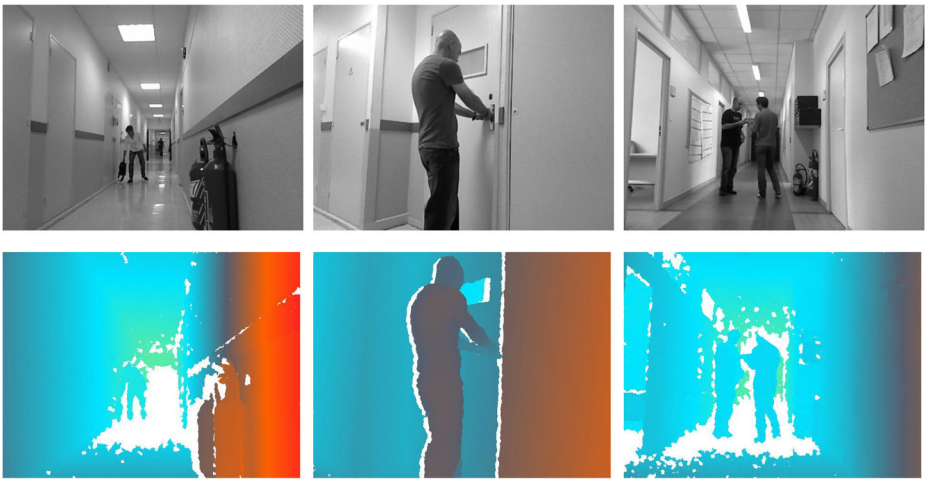

Fig. 8 Examples of actions of the LIRIS human activities dataset

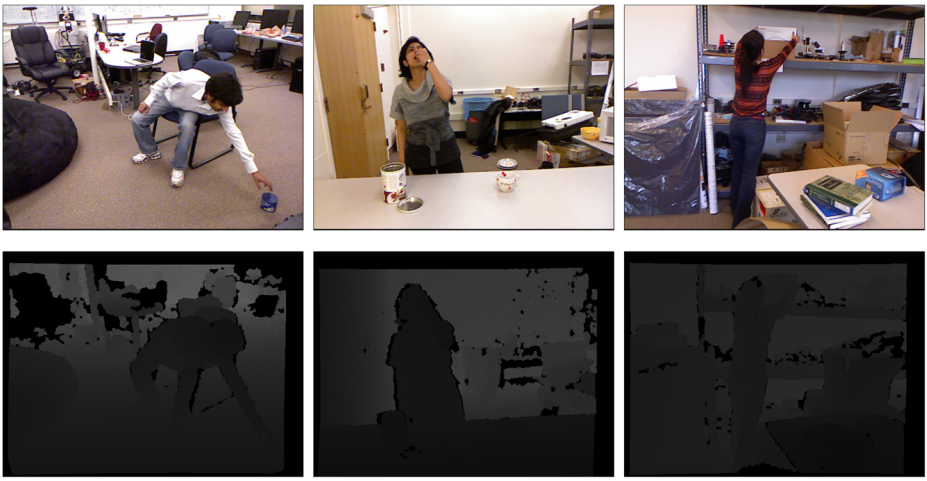

Fig. 9 Examples of actions of the CAD-60 dataset

the running environment was an Intel® Core^{TM} i7-4720HQ Processor, 6M cache, up to 3.60 GHz, 32 GB of RAM.

4.2 System evaluation

The system evaluation was performed in three steps. In the first, by using random sets of videos coming from both UTKinect-Action3D and LIRIS Human Activities datasets, the best codebook size, in terms of classification accuracy, was established. In the second, the UTKinect-Action3D and CAD-60 datasets were used to show the effectiveness of the

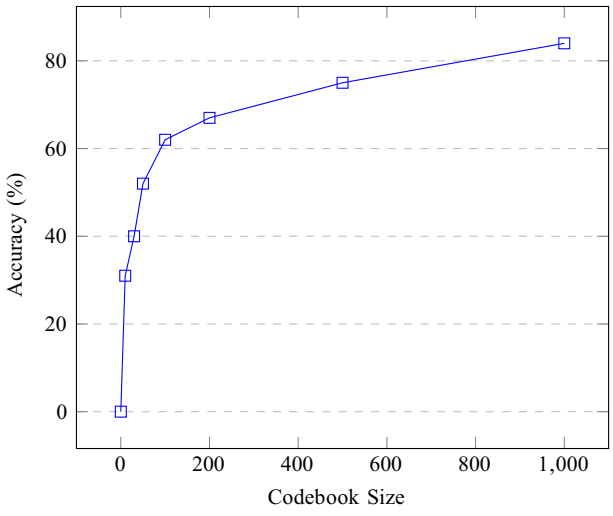

Fig. 10 In ordinate and abscissa the mean value of the system accuracy and the number, K, of established clusters are reported, respectively

Table 2 Accuracy results for human action recognition in UTKinect-Action3D dataset by using the 2FCV approach

Method	Accurancy
Zhong et al. [74]	34%
Yan et al. [72]	34%
Jia et al. [34]	35.5%
Jia et al. [33]	38.5%
He et al. [29]	47.5%
Jia et al. [32]	48%
Our	51%

system in comparison with different key works of the current literature. Finally, in the last, the LIRIS Human Activities dataset was used to prove the effectiveness of the system in very challenging video sequences.

4.2.1 Codebook size setting

The codebook size is derived by the number of clusters established during the training stage. This can be considered the most critical aspect of the proposed method. In fact, on the one hand, the creation of a small codebook allows the system to be trained quickly, albeit at the expense of the discriminative performance of the recognizer since two different descriptors may be assigned to the same cluster. On the other, a large codebook is time-consuming and may be less generalizable. Anyway, the agreement among discriminative performance, time-consuming, and generalization depends on the video content and on the video domain. In our context, we have established the number of clusters according to a set of empirical experiments performed by using all videos in the UTKinect-Action3D dataset. In Fig. 10, the increment of the system accuracy according to the increment of the number of cluster, K, is reported. In particular, the graph shows that the accuracy grows sharply until then the number of clusters increases up to about 200. After this value, the system has an increasingly limited benefit in accuracy in view of the computational effort required to manage an high number of clusters during the training and evaluation stages. In our experiments, we have fixed the codebook size K to 1000 clusters to obtain an excellent trade-off between accuracy and performance.

4.2.2 Evaluation on the UTKinect-Action3D dataset

The experimental sessions by the UTKinect-Action3D dataset were performed by using both approaches discussed in [32], i.e., 2FCV and LOOCV. Beyond the accuracy, also the following well-known metrics were computed to highlight the goodness of the obtained results [4], i.e., Precision ($Prec$), Recall (Rec), and F-measure ($F1$). In this context, the

Table 3 Accuracy results for human action recognition in UTKinect-Action3D dataset by using the LOOCV approach

Method	Accuracy
Zhu et al. [76]	80 %
Our	84%
Xia et al. [70]	90.9%
Gupta et al. [25]	96%

Table 4 Confusion Matrix for human action recognition in UTKinect-Action3D dataset by using the LOOCV approach

	WK	SD	SU	PU	CR	TW	PS	PL	WH	CH
WK	1.0	0	0	0	0	0	0	0	0	0
SD	0	0.90	0.10	0	0	0	0	0	0	0
SU	0	0.20	0.80	0	0	0	0	0	0	0
PU	0	0	0	1.0	0	0	0	0	0	0
CR	0	0	0	0	1.0	0	0	0	0	0
TW	0	0	0	0.25	0	0.75	0	0	0	0
PS	0	0	0	0	0	0	1.0	0	0	0
PL	0	0	0	0	0	0	0.30	0.70	0	0
WH	0	0	0	0	0	0	0	0	1.0	0
CH	0	0	0	0	0	0.40	0	0	0	0.60

$Prec$ points out the level of probabilistic proximity between a recognized action and its membership class, the Rec highlights the sensitivity level of the clustering, finally, the $F1$ measures the effectiveness of the classifier. By using the results presented in [32] on the UTKinect-Action3D dataset with the 2FCV approach, in Table 2 a comparison of the accuracy of the proposed system with different key works of the current state-of-the-art is reported.

The proposed system, under the mentioned experimental settings, is the one with the highest percentage of accuracy. Since the 2FCV approach divides the dataset in two equal parts, all the permutations of the video sequences were considered and the mean values were computed. Since each video contains 10 actions performed twice by 10 users, we separated them, thus obtaining 200 source sequences. The system also shows remarkable performance in $Prec = 0.667$, $Rec = 0.693$, and $F1 = 0.677$. Unfortunately, the selected key works do not report this kind of values thus preventing a comparison with them.

Regarding the LOOCV approach, in Table 3 the comparison of the proposed system with the accuracies of other key works of the current literature is reported. In addition, in Table 4, the related confusion matrix is shown. Notice that, for the two experimental approaches we had to consider different sets of works because they utilized only either an experimental way or the other. As in the previous case, we considered all the permutations of the source video sequences and computed the mean values.

Table 5 Accuracy results for human action recognition in CAD-60 dataset by using the "New Person" settings

Method	Accurancy
Zhu et al. [77]	62.50 %
Karpathy et al. [35]	65.30%
Koppula et al. [42]	71.40%
Our	82.60%
Wang et al . [68]	74.70%
Hu et al. [30]	84.10%
Koperski et al. [40]	80.36
Das et al. [17]	95.58

Table 6 Accuracy results for human action recognition in LIRIS human activities dataset

Video sequence	Accuracy
Unattended baggage	75.5%
Force a closed door	72.2%
Exchange of object	40.5%

The accuracy results by the LOOCV approach show that the proposed system is comparable with these different key works achieving satisfactory performance. Confronting these results with the previous ones it is worth noting that the system learns quickly but requires of a more sophisticated clustering. Anyway, the system works properly as also pointed out by the computed additional measures, i.e., $Prec = 0.850$, $Rec = 0.816$, and $F1 = 0.821$.

The proposed system presents other advantages beyond the accuracy performance. First, the JDL model provides a full support with respect to the generalization and the scalability of the system. Second, the original pipeline allows the system to be robust to the number of heterogeneity sensors. Third, since the system adopts a decision level strategy, it is also robust to the sensor failures. Finally, the system manages light information, i.e., the keypoints, allowing a reduction of the computation both in training and evaluation stages.

4.2.3 Evaluation on the CAD-60 dataset

In this section, we report the overall accuracy of the proposed system when applied on the CAD-60 dataset. Also in this case, the proposed approach obtains comparable results, in terms of accuracy, with key works of the current literature, as reported in Table 5. The training stage, with this dataset, was performed by using all the training video sequences acquired by the Microsoft Kinect. The method we propose presents some weaknesses when the scene contains interactions among subjects and objects. This is due to the fact that the current clustering processes the descriptors without providing a pre-labelling that could support a better management of the codebook.

4.2.4 Evaluation on the LIRIS human activities dataset

To evaluate the system in complex environments, a further experimental session was performed. In particular, by using the LIRIS Human Activities dataset, three challenging video sequences were tested. In Table 6, the accuracy results on the selected videos are reported.

The training stage, with this dataset, was performed by using all the 111 training video sequences acquired by the Microsoft Kinect. Also in this case, the system presents a high level of accuracy despite various adverse factors, such as: illumination changes, distance of subjects, and occlusions. As shown for CAD-60 dataset, the proposed system presents some weaknesses when the scene contains several subjects close to each other.

5 Conclusion

This paper presents a system for human action recognition supported by a decision level fusion for depth and colour information. The proposed system presents some novelties compared with the current state-of-the-art, including an original pipeline based on the customization of the JDL data fusion model. Unlike other works, whose effectiveness is

tested by using only either the 2FCV or the LOOCV, we have used both strategies on the UTKinect-Action3D dataset. In addition, to confirm the obtained results, we have also performed experimental sessions by using two further datasets containing challenging video sequences: LIRIS Human Activities and CAD-60. The tests have shown that the system can be compared with selected key works of the current literature, thus providing a concrete contribute to the human action recognition field. The JDL guarantees a complete system scalability, while the decision level fusion ensures an extendibility of the system with any set of heterogeneous sensors. The method we propose is also able to mange video sub-sequences due to its temporal independence strategy. Despite this, we are aware that a certain amount of semantic meaning of the video sequences is lost during the BoVW pooling. For this reason, with the intent to maintain the temporal independence property, but also to preserve the whole information contained in a video sequence, we are currently engaged in developing an additional set of temporal features compatible with the BoVW mechanism.

Acknowledgments This work was supported in part by the MIUR under grant "Departments of Excellence 2018-2022" of the Department of Computer Science of Sapienza University.

References

1. Aggarwal J, Ryoo M (2011) Human activity analysis: a review. ACM Comput Surv 43(3):16, 1–16, 43
2. Aggarwal J, Xia L (2014) Human activity recognition from 3D data: a review. Pattern Recogn Lett 48:70–80
3. Avola D, Cinque L, Levialdi S, Placidi G (2013) Human body language analysis: a preliminary study based on kinect skeleton tracking. In: Proceedings of the international conference on image analysis and processing (ICIAP), pp 465–473
4. Avola D, Bernardi M, Cinque L, Foresti GL, Massaroni C (2018a) Combining keypoint clustering and neural background subtraction for real-time moving object detection by PTZ cameras. In: Proceedings of the international conference on pattern recognition applications and methods (ICPRAM), pp 638–645
5. Avola D, Bernardi M, Cinque L, Foresti GL, Massaroni C (2018b) Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. IEEE Transactions on Multimedia, pp P–P (in press)
6. Avola D, Cinque L, Foresti G, Martinel N, Pannone D, Piciarelli C (2018c) Low-level feature detectors and descriptors for smart image and video analysis: a comparative study. In: Bridging the semantic gap in image and video analysis, pp 7–29
7. Avola D, Cinque L, Foresti GL, Marini MR, Pannone D (2018d) VRheab: a fully immersive motor rehabilitation system based on recurrent neural network. Multimedia Tools and Applications 77(19):24, 955–24, 982
8. Ballan L, Bertini M, Del Bimbo A, Serra G (2010) Video event classification using string kernels. Multimedia Tools and Applications 48(1):69–87
9. Bay H, Ess A, Tuytelaars T, Gool LV (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110(3):346–359
10. Benmokhtar R (2014) Robust human action recognition scheme based on high-level feature fusion. Multimedia Tools and Applications 69(2):253–275
11. Canal G, Escalera S, Angulo C (2016) A real-time human-robot interaction system based on gestures for assistive scenarios. Comput Vis Image Underst 149(C):65–77
12. Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27
13. Chathuramali KGM, Rodrigo R (2012) Faster human activity recognition with SVM. In: Proceedings of the international conference on advances in ICT for emerging regions (ICTer), pp 197–203

14. Cámara-Chávez G, de Albuquerque Araújo A (2009) Harris-SIFT descriptor for video event detection based on a machine learning approach. In: Proceedings of the IEEE international symposium on multimedia (ISM), pp 153–158
15. Correa NM, Adali T, Li YO, Calhoun VD (2010) Canonical correlation analysis for data fusion and group inferences. IEEE Signal Proc Mag 27(4):39–50
16. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
17. Das S, Koperski M, Bremond F, Francesca G (2017) Action recognition based on a mixture of RGB and depth based skeleton. In: Proceedings of the IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–6
18. Duta IC, Uijlings JRR, Ionescu B, Aizawa K, Hauptmann AG, Sebe N (2017) Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information. Multimedia Tools and Applications 76(21):22, 445-22, 472
19. Eigen D, Fergus R (2015) Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 2650–2658
20. Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), vol 2, pp 524–531
21. Foggia P, Percannella G, Saggese A, Vento M (2013) Recognizing human actions by a bag of visual words. In: Proceedings of the IEEE international conference on systems, man, and cybernetics (SMC), pp 2910–2915
22. Gao Y, Xiang X, Xiong N, Huang B, Lee HJ, Alrifai R, Jiang X, Fang Z (2018) Human action monitoring for healthcare based on deep learning. IEEE Access 6:52, 277–52, 285
23. Garg R, BG VK, Carneiro G, Reid I (2016) Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: Proceedings of the european conference on computer vision (ECCV), pp 740–756
24. Gunatilaka AH, Baertlein BA (2001) Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection. IEEE Trans Pattern Anal Mach Intell 23(6):577–589
25. Gupta K, Bhavsar A (2016) Scale invariant human action detection from depth cameras using class templates. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 38–45
26. Hall DL, Llinas J (1997) An introduction to multisensor data fusion. Proc IEEE 85(1):6–23
27. Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C Appl Stat 28(1):100–108
28. He C, Shao J, Sun J (2018) An anomaly-introduced learning method for abnormal event detection. Multimedia Tools and Applications 77(22):29, 573–29, 588
29. He X, Cai D, Niyogi P (2006) Tensor subspace analysis. In: Advances in neural information processing systems, pp 499–506
30. Hu J, Zheng W, Lai J, Zhang J (2017) Jointly learning heterogeneous features for RGB-d activity recognition. IEEE Trans Pattern Anal Mach Intell 39(11):2186–2200
31. Ijjina EP, Chalavadi KM (2017) Human action recognition in RGB-d videos using motion sequence information and deep learning. Pattern Recogn 72:504–516
32. Jia C, Fu Y (2016) Low-rank tensor subspace learning for RGB-d action recognition. IEEE Trans Image Process 25(10):4641–4652
33. Jia C, Kong Y, Ding Z, Fu YR (2014a) Latent tensor transfer learning for RGB-D action recognition. In: Proceedings of the ACM international conference on multimedia (MM), pp 87–96
34. Jia C, Zhong G, Fu Y (2014b) Low-rank tensor learning with discriminant analysis for action classification and image recovery. In: Proceedings of the AAAI conference on artificial intelligence (CAI), pp 1228–1234
35. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 1725–1732
36. Khaire P, Kumar P, Imran J (2018) Combining cnn streams of RGB-D and skeletal data for human activity recognition. Pattern Recognition Letters pp P–P (in press)
37. Khaleghi B, Khamis A, Karray FO, Razavi SN (2013) Multisensor data fusion: a review of the state-of-the-art. Information Fusion 14(1):28–44
38. Kim TY, Ko H (2005) Bayesian fusion of confidence measures for speech recognition. IEEE Signal Process Lett 12(12):871–874
39. Klein LA (2004) Sensor and data fusion: a tool for information assessment and decision making. SPIE Press, Bellingham

40. Koperski M, Bremond F (2016) Modeling spatial layout of features for real world scenario RGB-D action recognition. In: Proceedings of the IEEE international conference on advanced video and signal based surveillance (AVSS), pp 44–50
41. Koperski M, Bilinski P, Bremond F (2014) 3D trajectories for action recognition. In: Proceedings of the IEEE international conference on image processing (ICIP), pp 4176–4180
42. 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
43. Kosmopoulos DI, Doliotis P, Athitsos V, Maglogiannis I (2013) Fusion of color and depth video for human behavior recognition in an assistive environment. In: Proceedings of the internation conference on distributed, ambient, and pervasive interactions (DAPI), pp 42–51
44. Kumar P, Mittal A, Kumar P (2006) Fusion of thermal infrared and visible spectrum video for robust surveillance. In: Proceedings of the indian conference on computer vision, graphics and image processing (ICVGIP), pp 528–539
45. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, New York
46. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1–8
47. Liu AA, Nie WZ, Su YT, Ma L, Hao T, Yang ZX (2015) Coupled hidden conditional random fields for RGB-d human action recognition. Signal Process 112:74–82
48. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
49. Miranda L, Vieira T, Martínez D, Lewiner T, Vieira AW, Campos MFM (2014) Online gesture recognition from pose kernel learning and decision forests. Pattern Recogn Lett 39:65–73
50. Ni B, Nguyen CD, Moulin P (2012) RGBD-camera based get-up event detection for hospital fall prevention. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1405–1408
51. Ni B, Pei Y, Moulin P, Yan S (2013) Multilevel depth and image fusion for human activity detection. IEEE Transactions on Cybernetics 43(5):1383–1394
52. Oneata D, Verbeek J, Schmid C (2013) Action and event recognition with Fisher vectors on a compact feature set. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1817–1824
53. Padhy RP, Chang X, Choudhury SK, Sa PK, Bakshi S (2018) Multi-stage cascaded deconvolution for depth map and surface normal prediction from single image. Pattern Recognition Letters pp P–P (in press)
54. Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Comput Vis Image Underst 150:109–125
55. Piyathilaka L, Kodagoda S (2013) Human activity recognition for domestic robots. In: Proceedings of the international conference on field and service robotics (FSR), pp 395–408
56. Presti LL, Cascia ML (2016) 3D skeleton-based human action classification: a survey. Pattern Recogn 53:130–147
57. Rahmani H, Mian A, Shah M (2018) Learning a deep model for human action recognition from novel viewpoints. IEEE Trans Pattern Anal Mach Intell 40(3):667–681
58. Raman N, Maybank S (2015) Action classification using a discriminative multilevel HDP-HMM. Neurocomputing 154:149–161
59. Ross AA, Govindarajan R (2005) Feature level fusion of hand and face biometrics. In: SPIE proceedings, pp 196–204
60. Sanchez-Riera J, Hua KL, Hsiao YS, Lim T, Hidayati SC, Cheng WH (2016) A comparative study of data fusion for RGB-d based visual recognition. Pattern Recogn Lett 73:1–6
61. Scholkopf B, Sung KK, Burges CJC, Girosi F, Niyogi P, Poggio T, Vapnik V (1997) Comparing support vector machines with Gaussian kernels to radial basis function classifiers. IEEE Trans Signal Process 45(11):2758–2765
62. Shahroudy A, Wang G, Ng TT (2014) Multi-modal feature fusion for action recognition in RGB-D sequences. In: Proceedings of the international symposium on communications, control and signal processing (ISCCSP), pp 1–4
63. Sharma P, Kaur M (2013) Multimodal classification using feature level fusion and SVM. Int J Comput Appl 76(4):26–32
64. Sung J, Ponce C, Selman B, Saxena A (2011) Human activity detection from RGBD images. In: Proceedings of the AAAI conference on plan, activity, and intent recognition (PAIR), pp 47–55
65. Sung J, Ponce C, Selman B, Saxena A (2012) Unstructured human activity detection from RGBD images. In: Proceedings of the IEEE international conference on robotics and automation (ICRA), pp 842–849

66. Sykora P, Kamencay P, Hudec R (2014) Comparison of SIFT and SURF methods for use on hand gesture recognition based on depth map. AASRI Procedia 9:19–24
67. Tripathi RK, Jalal AS, Agrawal SC (2018) Suspicious human activity recognition: a review. Artif Intell Rev 50(2):283–339
68. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1290–1297
69. Wolf C, Mille J, Lombardi E, Celiktutan O, Jiu M, Dogan E, Eren G, Baccouche M, Dellandrea E, Bichot CE, Garcia C, Sankur B (2014) Evaluation of video activity localizations integrating quality and quantity measurements. Comput Vis Image Underst 127:14–30
70. Xia L, Chen CC, Aggarwal JK (2012) View invariant human action recognition using histograms of 3D joints. In: Proceedings of the IEee conference on computer vision and pattern recognition workshops (CVPRW), pp 20–27
71. Xian Y, Rong X, Yang X, Tian Y (2017) Evaluation of low-level features for real-world surveillance event detection. IEEE Trans Circuits Syst Video Technol 27(3):624–634
72. Yan S, Xu D, Yang Q, Zhang L, Tang X, Zhang HJ (2005) Discriminant analysis with tensor representation. In: Proceedings of the conference on computer vision and pattern recognition (CVPR), vol 1, pp 526–532
73. Yao T, Wang Z, Xie Z, Gao J, Feng DD (2017) Learning universal multiview dictionary for human action recognition. Pattern Recogn 64:236–244
74. Zhong G, Cheriet M (2014) Large margin low rank tensor analysis. Neural Comput 26(4):761–780
75. Zhou X, Zhuang X, Yan S, Chang SF, Hasegawa-Johnson M, Huang TS (2008) SIFT-bag kernel for video event analysis. In: Proceedings of the ACM international conference on multimedia (MM), pp 229–238
76. Zhu Y, Chen W, Guo G (2013) Fusing spatiotemporal features and joints for 3D action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 486–491
77. Zhu Y, Chen W, Guo G (2014) Evaluating spatiotemporal interest point features for depth-based action recognition. Image Vis Comput 32(8):453–464

Danilo Avola received the M.Sc. degree in Computer Science from Sapienza University, Rome, Italy, in 2002 and the Ph.D. degree in Molecular and Ultrastructural Imaging from University of L'Aquila, L'Aquila, Italy, in 2014. Since 2015 he is postdoc researcher at the Department of Mathematics, Computer Science and Physics (DMIF), University of Udine, Udine, Italy, and member of the AVIRES Lab at the same University. Previously, he was research engineer at the Department of Computer Science, Sapienza University and at the Multimodal & Multimedia Lab of the National Research Council, Rome, Italy. His research interests include Human Computer Interaction, Computer Vision, Signal Processing, Machine Learning, Deep Learning, Image and Video Processing, Multimodal Systems, and Pattern Recognition. He serves on the Steering Committee of selected International Conferences and is an Editorial Board member of different International Journals. Danilo Avola is author or coauthor of more than 80 papers in International Journals, refereed International Conferences, and International Book Chapters. Since 2011, Danilo Avola is member IEEE, member of IAPR, and member of CVPL.

Marco Bernardi received the B.Sc and the M.Sc. (cum laude) in Computer Science from Sapienza University of Rome, Rome, Italy, in 2012 and 2016, respectively. Since 2016 he is a Ph.D. Student in Computer Science and he is a member of the Computer Vision Laboratory in the Department of Computer Science at the same University. His research interests include Data Fusion, Human Activity Recognition, Pattern Recognition, Machine Learning, Human Computer Interaction. He is student member of IEEE, member of IAPR and member of CVPL.

Gian Luca Foresti is Full Professor of Computer Science at the University of Udine and Deputy Director of the Department of Mathematics, Computer Science and Physics. Since 2006, he is visiting professor in Artificial Vision at University of Klagenfurt. From 2000 to 2009, he was the appointed Italian member of the NATO RTO Information System Technology Panel. He was Finance Chair of the 11th IEEE Conference on Image Processing (ICIP05), General Chair of the 16th International Conference on Image Analysis and Processing (ICIAP11) and of the 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS11). His main interests involve Computer Vision and Image Processing, Multisensor Data and Information Fusion, Pattern Recognition, and Neural Networks. He is author of more than 300 papers published in International Journals and International Conferences, and he has been co-editor of several Books in the field of Multimedia and Video Surveillance. He has been Guest Editor of a Special Issue of the Proceedings of the IEEE on "Video Communications, Processing and Understanding for Third Generation Surveillance Systems". In 2002, he has been awarded of best IEEE Vehicular Electronics paper. Gian Luca Foresti is Senior member of IEEE, member of IAPR, and member of CVPL.