A multimodal approach for human activity recognition based on skeleton and RGB data

Pattern Recognition Letters
journal homepage: www.elsevier.com

A multimodal approach for human activity recognition based on skeleton and RGB data

Annalisa Franco[a,**], Antonio Magnani[a], Dario Maio[a]

[a]*Department of Computer Science and Engineering, University of Bologna, Via dell'Universitá 50, Cesena 47521, Italy*

ABSTRACT

Human action recognition plays a fundamental role in the design of smart solution for home environments, particularly in relation to ambient assisted living applications, where the support of an automated system could improve the quality of life of humans trying to interpret and anticipate user needs, recognizing unusual behaviors or preventing dangerous situations (e.g. falls). In this work the potentialities of the Kinect sensor are fully exploited to design a robust approach for activity recognition combining the analysis of skeleton and RGB data streams. The skeleton representation is designed to capture the most representative body postures, while the temporal evolution of actions is better highlighted by the representation obtained from RGB images. The experimental results confirm that the combination of these two data sources allow to capture highly discriminative features resulting in an approach able to achieve state-of-the-art performance on public benchmarks.

1. Introduction

The continuous advances in sensing technologies and networking infrastructures enable the development of intelligent software which can provide a real-time analysis of specific situations of interest in a home environment, with the aim of enhancing the quality of life of the occupants. For instance in the field of health-care, particular attention is generally devoted to systems able to detect and recognize dangerous situations and to provide prompt alarms (Cardinaux et al. (2011)); other interesting applications are abnormal human behavior detection or human-computer interfaces.
Our proposal focuses on the use of vision-based techniques which guarantee a higher degree of unobtrusiveness with respect to sensor-based approaches; the last ones in fact are more invasive in our opinion for the necessity of wearing sensors of different nature and not always appropriate for some categories of users (e.g. elderly people). Moreover, even being aware of the great success of neural networks coupled with deep learning techniques in many applications, we choose to design an activity recognition approach based on hand-crafted features. In the specific context of this work, in fact, the acquisition of a large amount of training data typically needed for network training is quite difficult and unlikely. The home environment is usually characterized by a very limited number of users, and also most of the reference benchmarks for activity recognition reproduce a "small-size" scenario, with few users and few activity samples per user. We are confident that in this scenario also "traditional" computer vision techniques can achieve good results and real time processing capabilities even with limited computational power.
The paper is organized as follows: section 2 discusses the state of the art of vision-based activity recognition techniques, section 3 describes the multimodal approach proposed in this work, section 4 summarizes the experiments carried out to evaluate our proposal on public datasets and finally section 5 draws some conclusions and outlines future research directions.

2. Related work

Human activity recognition (HAR) is a very active research area and summarizing the existing approaches is a quite hard task. Focusing on vision-based approaches, good reviews of the literature are provided in the recent surveys by Kong and Fu (2018) and Herath et al. (2017). A first criterion to categorize the existing approaches is the input data type; most of the

**Corresponding author: Tel.: +39-0547-338847;
e-mail: annalisa.franco@unibo.it (Annalisa Franco)

works exploit either the RGB images or skeleton information (provided for instance by the Kinect sensor). In this work we propose a multimodal system combining both aforementioned sources of information. Due that, in the following we discuss the main approaches related to. It is worth noting that, actually, the two categories are overlapped to some extent; pure methods exploiting a single data category are quite rare and many works combine different information to improve robustness. Each method is then included in the category related to the main information exploited.

2.1. Action recognition from RGB images

Many works in the literature adopt a representation of human actions based on a 3D volume, where the human pose and its variations are described. The 3D volume is then encoded in different ways. Gorelick et al. (2007) and Yilmaz and Shah (2005) use shape features, other approaches are based on optical-flow representation (e.g. Wang and Mori (2011)). Finally, many works adopt local representations in place of holistic descriptors to better deal with noise. Space-time interest points have been used in several works and represent an interesting category of approaches, which demonstrated a good robustness to image variations. Different techniques for keypoint detection have been proposed (see for instance Laptev (2005); Scovanner et al. (2007)) as well as different approaches for descriptor computation such as histograms of optical flow (Laptev et al. (2008)) and HOG features (Kläser et al. (2008); Wang et al. (2009)).
Several recent approaches exploit the potentialities of deep learning for activity recognition. Often the concept of 3D convolution (see Ji et al. (2013)) is used to capture temporal dynamics in a short period of time; other works model temporal dynamics by using multiple streams (Carreira and Zisserman (2017); Simonyan and Zisserman (2014); Feichtenhofer et al. (2016); Girdhar et al. (2017)). Few works suggest the combination of RGB, depth and skeletal data to improve action recognition accuracy (Khaire et al. (2018); Qi et al. (2018)).

2.2. Action recognition from skeleton data

Most of the approaches based on RGB-D data perform a skeleton analysis, adopting different representations of the set of joints. Some works exploit simple joint coordinates, normalized according to some body reference measure (see Gaglio et al. (2015); Shan and Akella (2014)) or joint distances (e.g. Cippitelli et al. (2016)). EigenJoints are proposed in Yang and Tian (2014) where PCA is applied to static and dynamic posture features to create a motion model. Histograms of 3D joints are described in Xia et al. (2012), while Zhang and Tian (2012) suggest the use of kinematic features, obtained observing the angles between couples of joints. Other works propose alternative representations: Gaussian Mixture Models representing the 3D positions of skeleton joints in Piyathilaka and Kodagoda (2013), Dynamic Bayesian Mixture Model of 3D skeleton features in Faria et al. (2014) or spatio-temporal interest points and descriptors derived from the depth image in Zhu et al. (2014). Another common approach is to adopt a hierarchical representation where an activity is composed of a set of sub-activities, also called *actionlets* (see Sung et al. (2012); Wang et al. (2012, 2014); Koppula et al. (2013)). In the recent work by Qi et al. (2018) an automatic joint configuration learning method, based on dictionary learning and sparse representation. The interaction of humans with objects is analyzed in a few works. The authors of Koppula et al. (2013) adopt a Markov random field where the nodes represent objects and sub-activities and the edges represent the relationships between object affordances, their relations with sub-activities, and their evolution over time; in Koppula and Saxena (2013) the authors propose a graph-based representation.
Also for skeletal data classification some techniques based on neural networks have been proposed. Long short-term networks are well suited to this aim for their capabilities of processing changes across time (see Battistone and Petrosino (2019)).

3. A multimodal system for activity recognition

The Kinect sensor provides parallel access to different data streams; in this work we are interested in coupling information from both skeleton and RGB images. We will define an activity as a sequence S of L data frames, $S_t, t = 1, .., L$; each element $S_t = (F_t, SK_t)$ includes F_t, the RGB frame acquired at time t (of size $W \times H$), and SK_t which is the corresponding skeleton. In practice the two data streams could be slightly misaligned, mainly due to the skeleton extraction and serialization procedures which are not always able to work at the same frame rate the data are provided. This misalignment was observed in several databases available for research purposes, for example in the well-known NTU RGB+D (Shahroudy et al., 2016) this phenomenon has lead to the loss of skeletal information in many sequences[1]. Nevertheless, its impact on our approach is negligible since the contribution of the two information is combined at decision-level.

3.1. Skeleton

We recently proposed in Franco et al. (2017) an activity recognition approach based on skeleton joint orientations. Many works in the literature based on skeleton only exploit joint positions to describe human postures; since Kinect provides for each joint also the estimated orientation, we decided to explore the robustness of this information. We therefore derived a posture representation based exclusively on angle information, derived from both the joint position and orientation. The great advantage of angle features derived from skeletons is that they are intrinsically normalized and independent from the user's physical structure. A good degree of invariance with respect to pose and view changes is also achieved since all the angles are computed with respect to the subject's coordinate system.
Each frame of a video sequence is represented by a set of angles derived from the human skeleton, which summarize the positions of the different body parts. The Kinect SDK represents the human skeleton as a set of d joints $J = \{j_1, j_2, ..., j_d\}$;

[1] NTU RGB+D Repository.

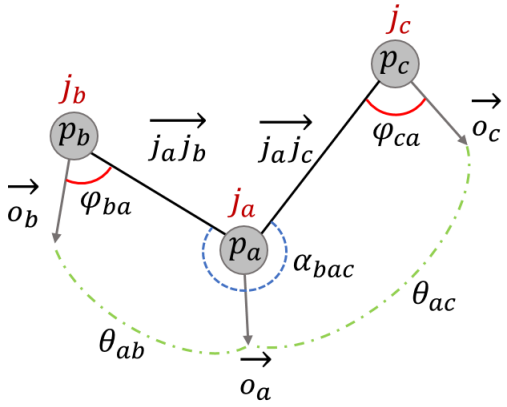


Figure 1: Representation of a subset of joints $j_a = (p_a, \overrightarrow{o_a})$, $j_b = (p_b, \overrightarrow{o_b})$ and $j_c = (p_c, \overrightarrow{o_c})$ and related angles θ, φ and α.

each joint $j_i = \left(\mathbf{p_i}, \overrightarrow{\mathbf{o_i}}\right)$ is described by its 3D position $\mathbf{p_i}$ and its orientation $\overrightarrow{\mathbf{o_i}}$ with respect to the sensor coordinate system[2] (X grows to the sensors left, Y grows up and Z grows out in the direction the sensor is facing). To encode the user posture, we defined three types of angles:

- θ_{ab}: angle between the orientations $\overrightarrow{\mathbf{o}_a}$ and $\overrightarrow{\mathbf{o}_b}$ of joints j_a and j_b (green angles in Figure 1). θ_{ab} angles are computed from a set of m couples of joints A_θ ($m = 8$).
- φ_{ab}: angle between the orientation $\overrightarrow{\mathbf{o}_a}$ of j_a and the segment $\overrightarrow{j_a j_b}$ connecting j_a to j_b (red angles in Figure 1). φ_{ab} angles are computed from a set of n couples of joints A_φ ($n = 16$).
- α_{bac}: angle between the segment $\overrightarrow{j_a j_b}$ connecting j_a to j_b and $\overrightarrow{j_a j_c}$ that connects j_a to j_c (blue angles in Figure 1). α_{bac} angles are computed from s triplets of joints A_α ($s = 4$).

Unfortunately the skeleton estimation provided by Kinect is not always accurate. The reliability is generally good for the joints of the upper part of the body, which contains most of the information needed for activity recognition. Legs are generally quite unreliable, but in many cases they are occluded or almost static and do not provide significant contribution for activity recognition. For this reason only a subset of the possible angles is considered, mainly obtained from the joints of the upper part of the body. All the details about the selected angles (A_θ, A_φ and A_α) can be found in Franco et al. (2017).
Each skeleton SK_t of the video sequence is represented by a vector obtained as the ordered concatenation of the values of $\theta_i \mid i \in A_\theta$, $\varphi_j \mid j \in A_\varphi$, $\alpha_k \mid k \in A_\alpha$

$$\mathbf{v}_i = (\theta_1, ..., \theta_m, \varphi_1, ...\varphi_n, \alpha_1, ..., \alpha_s)$$

of size $(m + n + s)$ where $m = |A_\theta|$, $n = |A\varphi|$ and $s = |A_\alpha|$.
The complete video sequence S is finally encoded using a Bag of Word model (BoW, Wang et al. (2009)) where each activity is represented as an histogram of occurrences of some reference postures. The skeleton BoW representation allows to effectively represent the main postures assumed by the human body during activities, but the final representation does not capture the temporal evolution of the body movement (due to the global nature of the histogram representation). The temporal images described in the following subsection allow to better represent this aspect and provide a complementary representation with respect to the skeleton information.

[2]Coordinate mapping - Microsoft docs.

3.2. HOG features from temporal images

In order to improve the recognition capabilities of the previously described approach, we developed a technique based on the analysis of RGB images with a two-fold objective: i) better encoding the temporal evolution of the activity, needed to discriminate between actions characterized by similar postures but presented in a different order (e.g. sit down and get up); ii) capture to some extent the user interaction with objects which could help to classify the activity. The feature extraction approach can be summarized into three main steps, described in the following subsections.

3.2.1. Construction of the temporal images

We can represent a sequence of frames F_t with $t = 1, .., L$ as a volume image V (see Fig. 2a), i.e. a parallelepiped in a 3D space (x, y, t), where the first two coordinates refer to the spatial coordinates of the frame pixels and the third one represents time. To achieve independence from the body position in the images, each frame is cropped to a fixed-size window (25% of the frame width) centered on the spine mid joint. The width of the region of interest has been empirically determined, based on a rough analysis of the training set, and is not always accurate for the test sequences; however it represents a good trade-off between computational complexity and accuracy.
Our representation is obtained by a slicing operation of the volume V at predefined values of the y-coordinate (see Fig.2), properly selected to capture the body motion during the activity. In particular, a set $T = \{T_{y_1}, T_{y_2}, ..., T_{y_M}\}$ of M temporal images of size $W \times L$ will be computed from V; the generic element of T is defined as: $T_{y_i}(r, c) = V(r, y_i, c)$ with $r \in [1, .., W]$ and $c \in [1, .., L]$. Examples of temporal images at fixed values of y are given in Fig. 2b. As clearly visible in the example, the temporal image highlights the specific movement of a body region during time; the slice at the level of hands will show a very typical periodic movement originated by the steering action performed. Other temporal images, for instance from the leg region, will be more static for this specific action. As expected, the selection of the sections to analyze (y values) has an important impact on the accuracy of representation. We evaluated two strategies: i) y value of the main skeleton joints; ii) uniform sampling along the skeleton. The two approaches will be compared in the experiment section.

3.2.2. Temporal image gradient computation

Looking at the temporal images, it's easy to observe that the relevant information for activity recognition is represented by the dynamic elements, the variations observed across time; the constant regions of the image are not interesting and must not be encoded. For this reason we convert each temporal image

$T_{y_i} \in T$ in a grayscale image and we compute the gradient moduli G_{y_i} using the Sobel operator (see Fig. 2c). Even if the RGB frames look quite defined, an analysis of the gradient images reveals the presence of a significant noise component that must be removed for reliable feature extraction. A denoising operation is therefore applied both before and after gradient computation to reduce the effects of inter-frame variations due to the sensor, thus obtaining a regularized image $\tilde{G}_{y_i}$. The technique used for denoising is non-local means denoising (Buades et al. (2005)), widely used in the literature and adopted in this work since it allows to preserves the main image characteristics without introducing a noticeable blurring effect, typical of other denoising approaches (e.g. bilateral filter).

3.2.3. HOG encoding of gradient images.

Each regularized gradient image $\tilde{G}_{y_i}$ is finally encoded by HOG descriptors proposed in Dalal and Triggs (2005). The length of the different sequences could be different of course, thus determining temporal images of different size. We need however a fixed-length descriptor to train a classifier, so each image is partitioned into a fixed number of overlapping blocks and the final descriptor is obtained by the concatenation of the block descriptors. The OpenCV implementation of HOG descriptor computation has been used here; in particular, best results were achieved with a window of 4x8 cells. The size of the cells for a specific sequence obviously depends on the size of the input temporal image. A L2 normalization is carried out on blocks made of 4x4 cells. The adoption of a histogram-based representation allows to further reduce the influence of noise.

3.3. Activity classification

The two techniques discussed in the previous sections are quite complementary and their fusion can be useful to achieve good recognition accuracy. As shown in Figure 3, two classifiers are trained using the features extracted from skeleton and RGB images respectively. As for the skeleton data, we used the same configuration described in our previous work (Franco et al. (2017)) with the training of a Random Forest classifier. The second classifier consists of a set of M linear Support Vector Machines where each SVM represents a T_y slice, i.e. each model is trained on a specific volume slice; the classification of a particular activity is carried out by the fusion of results obtained from the individual SVMs.
The outputs of the two classifiers are then combined for the final result; among the existing combination strategies, the decision-level fusion is the most suited in this case due to the possible misalignment of the RGB and the skeleton streams which makes difficult a fusion at feature level. The two classifiers are equally weighted for the computation of the combined score, obtained by a simple sum rule. In our internal experiments, the typical fusion rules (max, sum, prod) have been evaluated. Overall, the sum rule provided better results, probably because the two approaches are quite complementary and their sum results in a more robust estimation. The max rule provides the worst results meaning that, in some cases, one of the two methods provides the wrong class with a high confidence value and this problem is amplified by the max rule.

4. Experiments and results

The proposed approach has been evaluated over three different public benchmarks (CAD-60 and CAD-120 from Cornell University[3] and OAD dataset[4] internally acquired), each of them including different sets of activities. In order to evaluate the effectiveness of the proposed approaches we had to focus on datasets providing both RGB frames as well as information about joint orientations. Most of the existing datasets provide only one of the two categories of data, so we finally selected two well-known public benchmarks (CAD-60 and CAD-120) and we extended the experiments to a dataset internally collected. The performance are reported in terms of confusion matrix, where the rows correspond to the real activity label and the columns to the estimated one. Moreover, we report precision P and recall R values, computed as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}$$

where *TP, FP* and *FN* represent respectively the True Positives, False Positives and False Negatives which can be easily derived from the extra-diagonal elements of the confusion matrix.

4.1. Results on CAD-60

The Cornell Activity Datasets (CAD-60 and CAD-120) are two of the most well-known and complete datasets in the field of HAR. Despite of the number of HAR benchmarks available, to the best of our knowledge, only CAD-60 and CAD-120 provide joint orientations as well as 3D position. The first one consists of 60 RGB-D videos performed by four different subjects in five environments (*office, bathroom, bedroom, living room, kitchen*). The actors are two males and two females, one subject is left-handed. The authors identified three to four activities for each environment, providing a total of 12 unique activities (*rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, opening pill container, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard, working on computer*). One of the main characteristics of the dataset is the explicit splitting into different environments. Indeed, some activities are common to several rooms (*e.g., drinking water*), others are peculiar to a specific environment (*e.g., cooking*). The actors are typically placed in the center of the scene, without relevant occlusions. For the same activity the point of view is always the same regardless of the actor. The CAD-60 provides 3D positions of 15 joints and orientations of 11 of them.

All experiments were conducted on the basis of the "new person" setting introduced in Sung et al. (2012), which consists in a leave-one-out cross validation with rotation of the test subject. As mentioned in Section 3.2, we evaluated two different strategies for the selection of y-values, both based on skeleton information. In the first one, volume slices are extracted in correspondence of the position of the 15 joints describing the skeleton *(RGB - joint-based selection)*; the second one simply applies a uniform slice sampling along the whole skeleton *(RGB*

[3] CAD-60 and CAD-120 datasets.
[4] OAD dataset.

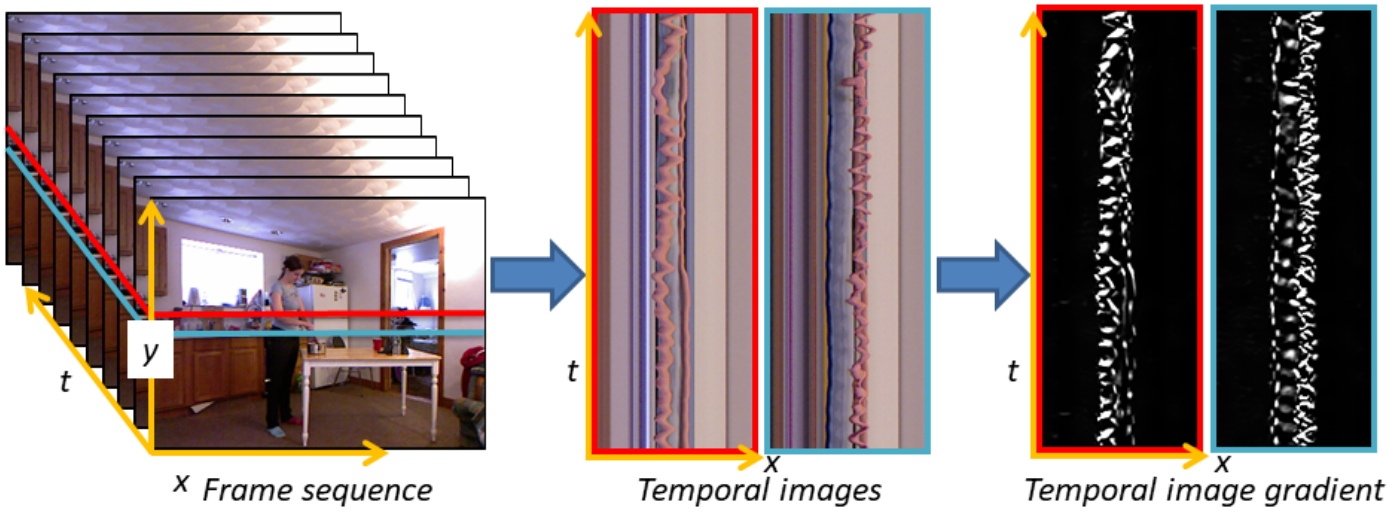


Figure 2: Representation of the feature extraction approach from RGB images. The temporal image (b) is a "slice" of the 3D volume representing the frame sequence (a). Relevant changes in time are well highlighted in the gradient image (c) extracted from (b).

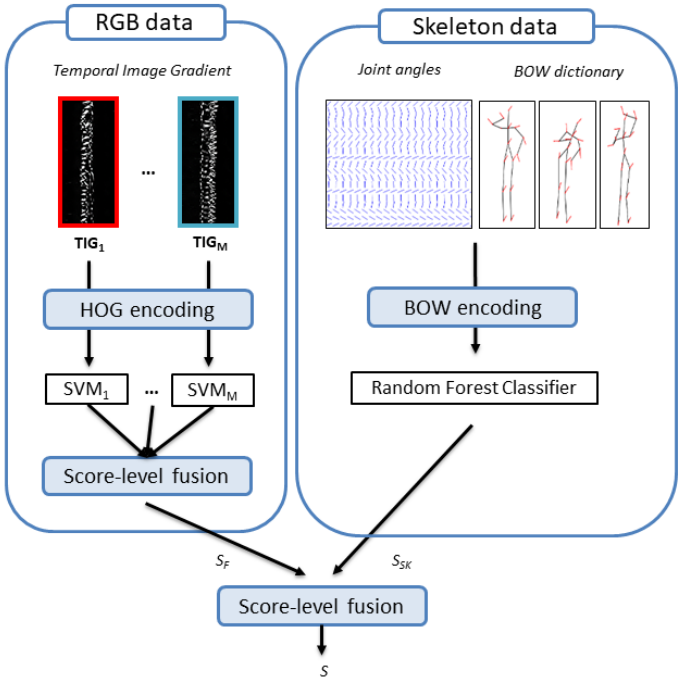


Figure 3: Schema of the proposed approach. The final score is obtained by a score-level fusion of the output of the two modules based on RGB and Skeleton data respectively.

- *uniform selection)*. A comparison between the two strategies on the CAD-60 dataset is given in Table 1 which also reports the results of other methods in the literature. Besides precision and recall, for each method an indication about the Kinect data exploited is given (*Sk*: skeleton, *RGB*: color frames, *De*: depth frames). It is worth noting that the skeleton information is derived by Kinect SDK from depth data, but we checked the *De* column only when the approach directly exploits depth images for feature extraction (different from skeleton).

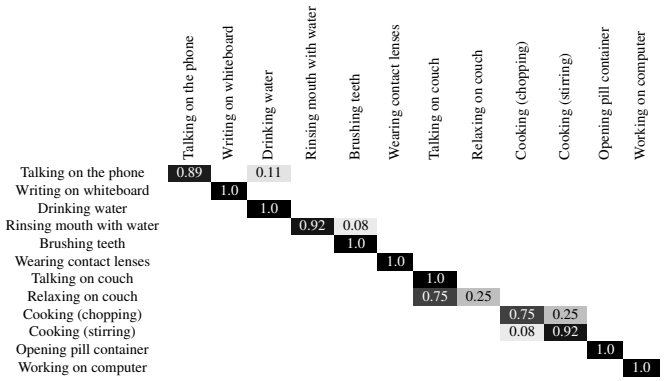
Table 1: Precision (*P*) and recall (*R*) of the proposed approaches on CAD-60, compared to the state-of-art results. For each method, the indication about the Kinect data exploited is also given: *Sk*: skeleton, *RGB*: color frames, *De*: depth frames.

Algorithm	*Sk*	*RGB*	*De*	*P*	*R*
Sung et al. (2012); Sung et al.	✓	✓	✓	67.9	55.5
Koppula et al. (2013)	✓	✓	✓	80.8	71.4
Zhang and Tian (2012)	✓			86.0	84.0
Ni et al. (2013)		✓	✓	75.9	69.5
Gupta et al. (2013)			✓	78.1	75.4
Yang and Tian (2014)	✓		✓	71.9	66.6
Zhu et al. (2014)	✓	✓	✓	93.2	84.6
Faria et al. (2014)	✓			91.1	91.9
Shan and Akella (2014)	✓			93.8	94.5
Gaglio et al. (2015)	✓			77.3	76.7
Parisi et al. (2015)	✓		✓	91.9	90.2
Cippitelli et al. (2016)	✓			93.9	93.5
Urbano Nunes and Peixoto (2017)	✓			81.83	80.02
Franco et al. (2017)	✓			**95.0**	**95.0**
Qi et al. (2018)	✓			90.18	92.9
Khaire et al. (2018)	✓	✓	✓	93.06	90.0
Battistone and Petrosino (2019)	✓	✓		94.4	93.7
RGB - joint based selection	✓	✓		**87.4**	**86.3**
RGB - uniform selection	✓	✓		**92.5**	**89.4**
Proposed approach	✓	✓		**98.8**	**98.3**

The results show that the uniform sampling is more effective, probably because the initial join position in some cases (e.g. hands) is not significant. Moreover, the information related to specular joints (e.g., shoulders, pelvis, knees, elbows) is redundant and not informative, thus making us lean towards uniform sampling along the entire skeleton. The confusion matrix shown in Table 2, allows to analyze the results obtained with uniform sampling with 20 different slices.

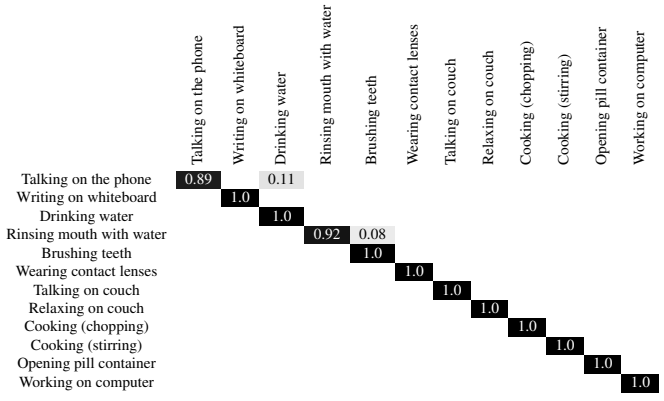
Table 3 reports the confusion matrix obtained by the combination of RGB and skeletal representations; excellent results are observed, compared to existing approaches, both in terms

Table 2: Confusion matrix of the RGB-based approach (using 20 uniform slices) on CAD-60.



of precision and recall.

Table 3: Confusion matrix using the score-level fusion approach on CAD-60.



4.2. Results on CAD-120

As the name suggests, CAD-120 consists of 120 videos of human activities where the subjects perform 10 high-level activities (*making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects, having a meal*). As for CAD-60, four subjects were considered, each of which performs each activity three times. Several elements make CAD-120 a more challenging dataset: in particular almost all activities exhibit relevant occlusions and the point of view varies depending on the actor. Moreover, only one environment is considered. As for CAD-60, even CAD-120 provides the 3D positions of 15 joints and the orientations of 11 of them. Different protocols are available for this benchmark; the most feasible for our evaluation is referred to as *Activity classification without ground-truth segmentation* on the CAD-120 website.

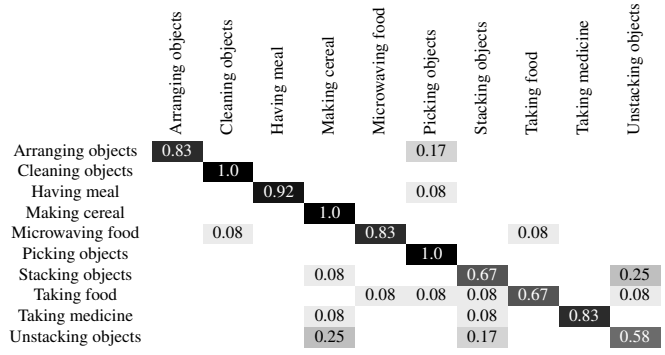
The results on CAD-120 for the proposed approach are shown in Table 4 and 7. It is possible to observe in Table 7 that temporal images alone do not provide satisfactory results on CAD-120. This is probably due to the complexity of the dataset and in particular the frequent occlusion of subjects (typically through motionless objects that hinder the production of temporal images). The results obtained from skeletal information are consistently better; however it clearly emerges that the two approaches are quite independent and their score-level fusion allows to significantly increase precision and recall (see Table 7). Due to space constraints, only the confusion matrix describing the results obtained by merging the two techniques is shown in Table 5.

Overall the results are encouraging, even if the method by Koppula and Saxena slightly outperforms our approach on this database. In our opinion, there are two main reasons for this behavior. First, they perform a hierarchical analysis, identifying both high-level and low-level activities, and the information from low-level analysis can be very useful to improve recognition. Second, their graph-based representation explicitly models objects and interactions with objects, while in our approach these aspects are only indirectly represented by observing their effects of this interaction on the subject's movements in RGB frames. The explicit knowledge about the objects in the scene allows to better deal with activities where the interaction with objects is a fundamental aspect (e.g. stacking or unstacking objects). Based on these considerations, we plan to explore possible improvements in our future research focusing on a better analysis of the context.

Table 4: Precision (P) and recall (R) of the proposed approaches on CAD-120, compared to the state-of-art results.

Algorithm	*P*	*R*
Koppula et al. (2013)	81.8	80.0
Koppula and Saxena (2013)	87.0	82.7
Prop. appr. (RGB and skeleton fusion)	85.4	83.3

Table 5: Confusion matrix using the proposed approach on CAD-120.



4.3. Results on Office Activity Dataset

Finally, the third dataset used for testing is the extended version of the Office Activity Dataset (OAD) presented in Franco et al. (2017). It includes a total of 560 video sequences of 14 activities (*drinking, getting up, grabbing an object from a shelf, pour a drink, scrolling book pages, sitting, stacking items, take objects from a shelf, talking on the phone, throwing something in the bin, waving hand, wearing coat, working on computer,*

writing on paper); each activity is performed twice by 20 subjects in a different environment from several perspectives based on the activity being performed. It is worth noting that the execution of the different activities was loosely supervised, just giving to the subjects a generic definition of the activity without specific indications on how it should be carried out. The skeletal data provided by OAD are composed of the 3D positions of 25 tracked joints and the orientations of 19 of them. Also OAD adopts the "new person" protocol introduced in Section 4.1. RGB and depth images will be released with permission and in accordance with the General Data Protection Regulation (GDPR, EU no. 2016/679).

It is important to underline that the subjects were free to carry out the activities in the way they thought to be more appropriate. This results in a significant intra-class variability. From the confusion matrix shown in Table 6, it can be seen that the most critical activity is *"throw something in the bin"*. Indeed, this is probably due to the high variability in the execution of the action by different subjects. Many of them interpreted the activity as a sequence comprising the approach to bin, bending down and finally the release of the object. Others preferred a literal interpretation of the label name and performed the action by throwing the object from a distance. This explains some of the errors due to the misclassification with *"grab object from the ground"*. The misclassification of *"waving"* and *"drinking"* is mainly due to the similar configuration of a significant portion of the angles between these two activities. Despite of some errors in specific activities, the good behavior of the proposed approach is confirmed in this test as well.

5. Conclusions

Human activity recognition has been addressed in this work by a multimodal approach based on the combination of skeletal information and HOG descriptors derived from RGB frames and designed to capture the temporal evolution of actions. Of course, combining different modalities increments the computational effort, in particular when dealing with RGB images. The cost of processing skeleton information is, in fact, negligible, i.e. a few milliseconds to process a whole activity; processing RGB frames for gradient, denoising and HOG features extraction is quite expensive, but overall the system is able to operate in real time since the recognition of a sequence (including RGB and skeleton data processing and their fusion) requires about 0.5 seconds using non optimized Python and C# code on an Intel Core i7-2600. We believe that the increment of computational effort is fully justified by the considerable improvement in recognition accuracy, in particular on the most difficult datasets. Of course, the deployment of this approach on embedded systems with reduced computational resources would require ad hoc optimizations, but this goes beyond the scope of this work. The results on public benchmarks confirm the complementarity of the two information, leading to a significant improvement of classification performance with respect to the single techniques.

An error analysis allows to identify possible future research directions. In particular, the problem of body occlusion must be seriously considered to improve the robustness of the approach; another important aspect to investigate is the explicit modeling of user interaction with objects which could represent a precious source of information for activity comprehension. Finally, the proposed approach is clearly designed for indoor activities with limited dynamics; the extension to other, more general scenarios, will require further analysis and the definition of ad hoc representations.

References

Battistone, F., Petrosino, A., 2019. TGLSTM: A time based graph deep learning approach to gait recognition. Pattern Recognition Letters 126, 132–138.

Buades, A., Coll, B., Morel, J., 2005. A non-local algorithm for image denoising, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, San Diego, CA, USA, pp. 60–65.

Cardinaux, F., Bhowmik, D., Abhayaratne, C., Hawley, M.S., 2011. Video based technology for ambient assisted living: A review of the literature. J. Ambient Intell. Smart Environ. 3, 253–269.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 2017, pp. 4724–4733.

Cippitelli, E., Gasparrini, S., Gambi, E., Spinsante, S., 2016. A human activity recognition system using skeleton data from RGBD sensors. Comp. Int. and Neurosc. vol. 2016, article id 4351435, 1–14.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, San Diego, CA, USA, pp. 886–893.

Faria, D.R., Premebida, C., Nunes, U., 2014. A probabilistic approach for human everyday activities recognition using body motion from RGB-D images, in: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, IEEE RO-MAN 2014, Edinburgh, UK, 2014, pp. 732–737.

Feichtenhofer, C., Pinz, A., Zisserman, A., 2016. Convolutional two-stream network fusion for video action recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 2016, pp. 1933–1941.

Franco, A., Magnani, A., Maio, D., 2017. Joint orientations from skeleton data for human activity recognition, in: Image Analysis and Processing - ICIAP 2017 - 19th International Conference, Catania, Italy, Proceedings, Part I, pp. 152–162.

Gaglio, S., Re, G.L., Morana, M., 2015. Human activity recognition process using 3-d posture data. IEEE Trans. Human-Machine Systems 45, 586–597.

Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.C., 2017. Actionvlad: Learning spatio-temporal aggregation for action classification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 2017, pp. 3165–3174.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R., 2007. Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. 29, 2247–2253.

Gupta, R.K., Chia, A.Y.S., Rajan, D., 2013. Human activities recognition using depth images, in: ACM Multimedia Conference, MM '13, Barcelona, Spain, 2013, pp. 283–292.

Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition. Image Vision Comput. 60, 4–21.

Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35, 221–231.

Khaire, P., Kumar, P., Imran, J., 2018. Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognition Letters 115, 107–116.

Kläser, A., Marszalek, M., Schmid, C., 2008. A spatio-temporal descriptor based on 3d-gradients, in: Proceedings of the British Machine Vision Conference 2008, BMVC08, Leeds, UK, pp. 1–10.

Kong, Y., Fu, Y., 2018. Human action recognition and prediction: A survey. CoRR abs/1806.11230.

Koppula, H.S., Gupta, R., Saxena, A., 2013. Learning human activities and object affordances from RGB-D videos. I. J. Robotics Res. 32, 951–970.

Koppula, H.S., Saxena, A., 2013. Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation, in: Proceed-

Table 6: Confusion matrix using the score-level fusion between the two classifiers on OAD.

	Drinking	Getting up	Grabbing obj.	Pouring a drink	Scrolling book	Sitting	Stacking items	Taking objects	Talking on phone	Throwing something	Waving	Wearing coat	Working on computer	Writing on paper
Drinking	0.88				0.03			0.02	0.07					
Getting up		0.88				0.10		0.02						
Grabbing obj.		0.06	0.82							0.12				
Pour a drink	0.09			0.88	0.03									
Scrolling book				0.03	0.97									
Sitting						1.0								
Stacking items							1.0							
Taking objects								1.0						
Talking on phone	0.05				0.02				0.90		0.03			
Throwing something			0.16		0.07		0.05			0.70	0.02			
Waving	0.20			0.05						0.03	0.72			
Wearing coat										0.07		0.93		
Working on computer													1.0	
Writing on paper							0.03							0.97

Table 7: Summary of the performance obtained on the three testing datasets.

Dataset	Approach	Precision	Recall
CAD-60	Skeleton	95.0	95.0
	RGB (20 sectors)	92.5	89.4
	Score-level fusion	98.8	98.3
CAD-120	Skeleton	77.6	73.1
	RGB (20 sectors)	61.1	59.3
	Score-level fusion	85.4	83.3
OAD	Skeleton	80.6	80.5
	RGB (20 sectors)	85.8	85.9
	Score-level fusion	90.6	90.4

ings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 2013, pp. 792–800.

Laptev, I., 2005. On space-time interest points. International Journal of Computer Vision 64, 107–123.

Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic human actions from movies, in: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2008, Anchorage, Alaska, USA.

Ni, B., Pei, Y., Moulin, P., Yan, S., 2013. Multilevel depth and image fusion for human activity detection. IEEE Trans. Cybernetics 43, 1383–1394.

Parisi, G.I., Weber, C., Wermter, S., 2015. Self-organizing neural integration of pose-motion features for human action recognition. Frontiers in Neurorobotics 9, 1–4.

Piyathilaka, L., Kodagoda, S., 2013. Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features, in: 2013 IEEE 8th Conference on Industrial Electronics and Applications, ICIEA 2013, Melbourne, VIC, Australia, 2013, pp. 567–572.

Qi, J., Wang, Z., Lin, X., Li, C., 2018. Learning complex spatio-temporal configurations of body joints for online activity recognition. IEEE Trans. Human-Machine Systems 48, 637–647.

Scovanner, P., Ali, S., Shah, M., 2007. A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th International Conference on Multimedia 2007, Augsburg, Germany, 2007, pp. 357–360.

Shahroudy, A., Liu, J., Ng, T., Wang, G., 2016. NTU RGB+D: A large scale dataset for 3d human activity analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 1010–1019.

Shan, J., Akella, S., 2014. 3d human action segmentation and recognition using pose kinetic energy, in: 2014 IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO 2014, Evanston, IL, USA, 2014, pp. 69–75.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, pp. 568–576.

Sung, J., Ponce, C., Selman, B., Saxena, A., . Human activity detection from RGBD images, in: Proceedings of the 16th AAAI Conference on Plan, Activity, and Intent Recognition, San Francisco, California, USA, 2011, pp. 47–55.

Sung, J., Ponce, C., Selman, B., Saxena, A., 2012. Unstructured human activity detection from RGBD images, in: IEEE International Conference on Robotics and Automation, ICRA 2012, St. Paul, Minnesota, USA, pp. 842–849.

Urbano Nunes, D.R.F., Peixoto, P., 2017. A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier. Pattern Recognition Letters 99, 21–31.

Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C., 2009. Evaluation of local spatio-temporal features for action recognition, in: British Machine Vision Conference, BMVC 2009, London, UK. Proceedings, pp. 1–11.

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012, Providence, RI, USA, 2012, pp. 1290–1297.

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2014. Learning actionlet ensemble for 3d human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 36, 914–927.

Wang, Y., Mori, G., 2011. Hidden part models for human action recognition: Probabilistic versus max margin. IEEE Trans. Pattern Anal. Mach. Intell. 33, 1310–1323.

Xia, L., Chen, C., Aggarwal, J.K., 2012. View invariant human action recognition using histograms of 3d joints, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, pp. 20–27.

Yang, X., Tian, Y., 2014. Effective 3d action recognition using eigenjoints. J. Visual Communication and Image Representation 25, 2–11.

Yilmaz, A., Shah, M., 2005. Actions sketch: A novel action representation, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, San Diego, CA, USA, pp. 984–989.

Zhang, C., Tian, Y., 2012. Rgb-d camera-based daily living activity recognition. Journal of Computer Vision and Image Processing 2, 1–12.

Zhu, Y., Chen, W., Guo, G., 2014. Evaluating spatio-temporal interest point features for depth-based action recognition. Image and Vision Computing 32, 453 – 464.