Vision-based holistic scene understanding for context-aware human-robot interaction

Giorgio De Magistris[2][0000−0002−3076−4509], Riccardo Caprari[2], Giulia Castro[2], Samuele Russo[2], Luca Iocchi[1][0000−0001−9057−8946], Daniele Nardi[1][0000−0001−6606−200X] and Christian Napoli[1][0000−0002−3336−5853]

[1] Department of Computer, Automation and Management Engineering, Sapienza University of Rome, via Ariosto 25 Roma 00185, Italy
{nardi, cnapoli}@diag.uniroma1.it
[2] Sapienza University of Rome, piazzale Aldo Moro 5, Roma 00185, Italy
{giorgio.demagistris, samuele.russo}@uniroma1.it

Abstract. Human anctivity recognition systems from static images or video sequences are becoming more and more present in our life. Most computer vision applications such as human-computer interaction, virtual reality, public security, smart home monitoring, or autonomous robotics, to name a few, highly rely on human anctivity recognition. Of course, basic human activities, such as "walking" and "running", are relatively easy to recognize. On the other hand, identifying more complex activities is still a challenging task that could be solved by retrieving contextual information from the scene, such as objects, events, or concepts. Indeed, a careful analysis of the scene can help to recognize human activities taking place. In this work, we address a holistic video understanding task to provide a complete semantic level description of the scene. Our solution can bring significant improvements in human anctivity recognition tasks. Besides, it may allow equipping a robotic and autonomous system with contextual knowledge of the environment. In particular, we want to show how this vision module can be integrated into a social robot to build a more natural and realistic context-based Human-Robot Interaction. We think that social robots must be aware of the surrounding environment to react in a proper and socially acceptable way, according to the different scenarios.

Keywords: human anctivity recognition · holistic video understanding · Human-Robot Interaction.

1 Introduction

Human anctivity recognition (HAR) has the aim of identifying human actions from the most simple ones such as gestures or atomic actions like "walking" or "sitting" to the most complex ones like behaviors or events, using sensory data information.

On the field of robotics and human-robot interaction, especially when elderly people are involved, HAR can represent an important moment that produces a

feeling of comfort and reassurance. This can in turn: reduce the risk factors connected to the feeling of being "useless" and "incapable", increasing, instead, the positive feelings of self-efficacy, since the person can feel dependent on other humans. Moreover also a sense of empowerment can follow and allow the elderly person to feel more understood and at the same time less dependent on caregivers. On some occasions this interaction can also constitute a moment perceived as "company" that can reduce the feeling of loneliness, real or fantasized, which often accompanies the elderly person's experience.

HAR from static images or video sequences has experienced significant growth over the last decade in the scientific areas of computer vision. As a consequence, a lot of applications in a wide spectrum of domains greatly rely on HAR systems. Few examples are human-computer interaction, augmented reality, intelligent home monitoring, or also video assistance and surveillance in public security, where crowds' movements are tracked to detect violent or criminal situations. More complex applications also concern advanced robotics, including mobile robot navigation or human-robot cooperation, and it also touches the medical environment to ensure surgical operations or continuous patient monitoring.

Of course basic human activities, such as "walking" and "running," are quite easy to recognize, but identifying more complex activities is still a challenging task, due to intraclass and interclass similarities problems. Namely, the same action can be expressed differently by diverse body movements of different users, and on the contrary different types of actions may show the same information or very similar features. Other common problems when dealing with HAR tasks are also related to complex background, lightness, scaling, and point of view that may represent significant limitations as well. In all these cases, only the contextual information extracted from the background and the detection of objects in the same scene may help to better understand the ongoing event and then the ongoing human activity. For this reason, in this work, we want to address a holistic video understanding (HVU) task, which is a multi-label and multi-task learning problem introduced in [1].

First of all, *Holism* is a theoretical position according to which the properties of a system cannot be explained just by its singular components, but capturing also their relative overall connections. It is a very common theory and widely used approach when we need to describe real-world phenomena, just look at the fundamental sciences of medicine or physics. Following this new current of thought, the aim of the holistic video understanding is not only to recognize specific and individual actions but also to provide a semantic level description of the scene describing the higher level connections among objects. Of course, the HVU task we propose includes human anctivity recognition and, at the same time, it provides valuable information on other multiple semantic categories. To naturally capture contextual information from dynamic real-world scenarios it's reasonable to use objects, scenes, attributes, concepts, actions, and events. First of all, the fact that we can identify together multiple semantic categories from the scene may be very useful for the recognition of specific and advanced actions, even in complex or cluttered background, or in a crowded environment

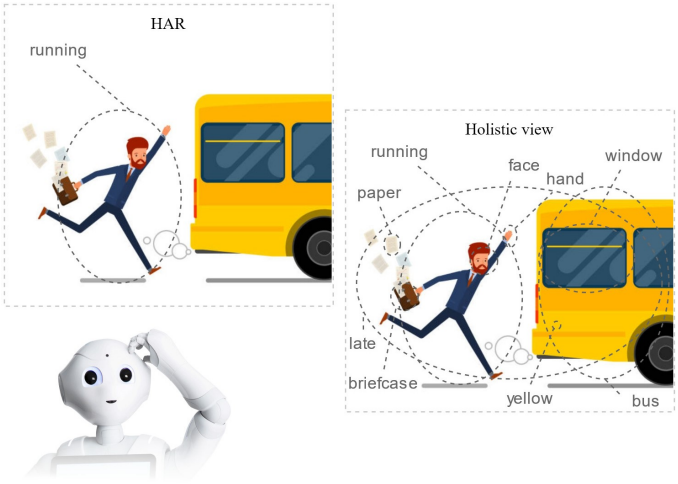

Fig. 1. An example of Human Action Recognition (HAR) vs *holistic* scene understanding.

with multiple subjects and actions, that is one of the biggest challenges in HAR. At the same time, this new holistic approach will allow capturing the contextual knowledge of the environment, not only to acquire information about the ongoing action but also to know where, why, and how that particular action is performed. In this way, for example, a robot could infer also the reasons and causes, which are hidden inside the scene, that led to perform a specific action, as shown in figure 1. The ability to fast recognizing contextual information is a very strong and highly desirable feature for any intelligent robotic system that needs to react online to the dynamic situation evolving in time. Particularly, for social and assisting robots in public and private environments, the use of contextual knowledge can be a key factor for achieving higher flexibility and adaptability to environmental conditions.

The remainder of this paper is structured as follows. Section 2 analyzes existing literature and discusses the state-of-the-art in human anctivity recognition, holistic video understanding, and context-aware robots. Section 3 reviews some of the most influential HAR datasets over the last decade. Section 4 formalizes the problem statement. Section 5 explains our proposed system architecture and methodology with particular attention to Convolutional Neural Networks and Recurrent Neural Networks. Section 6 presents experiments, implementation details, and their relative results. Section 7 shows how this vision-based holistic scene understanding module can be integrated to build an efficient context-aware human-robot interaction and possible application scenarios. Finally, in Section 8, we discuss conclusions and future directions.

2 Related Work

Human anctivity recognition has been widely explored in the last decade, given the growing technological progress in the field. Several methods and approaches have been studied in literature starting from a multi-modal human action analysis from gestures poses, facial expressions, and audio signals [2]. More advanced human anctivity recognition systems use depth cameras to create a more informative 3D representation of the human body as in [3], [4] or also human body parts motion analysis from skeletal poses as in [5], [6].

Only in the very recent years, research has focused on how HAR tasks can be expanded from single-label classification problems towards a more comprehensive understanding of image and video sequences. A first attempt was provided with the SOA dataset (Scene-Object-Actions) in [7], that first analyzed the possibility of applying the information learned from one task to improve the others. Only the last year, the authors of [1] provided a public available multi-label and multi-task video dataset intending to promote new research ideas and further works in the field of the holistic video understanding (HVU). This dataset strongly differs from the most influential ones in the HAR field, since it provides a significant increase both in the number of semantic categories (Scene-Object-Actions-Attribute-Concept-Event) and both in the corresponding number of labels per category. Indeed, the most common HAR video datasets, from the earliest ones such as UCF101 [8] and HMDB [21] to more recent and largest works such as ActivityNet [9] and Kinetics [10] are targeting human action or sport recognition in non complex background which makes them non-applicable in real-world applications. HVU dataset, instead, contains about 572K videos with approximately 9 million annotations spanning over 3142 labels among 6 different semantic categories. Given the significant improvements and the great potential of this very innovative dataset, the main goal of this work is to contribute and enlarge research in the holistic video understanding field by proposing a new model able to capture the whole information of a video. Our work is based on the idea proposed in [1] but it differentiates by introducing a new spatio-temporal network architecture: a Convolutional Neural Network (CNN) is used to acquire spatial information, combined with a Recurrent Neural Network (RNN) for capturing temporal relationships. Moreover, differently from previous works that just analyze possible solutions for recognizing human activities or the whole video content, our project wants to show how the holistic video understanding task can be successfully exploited to build a more natural context-aware human-robot interaction. Indeed it is another area that still needs to be explored to let many intelligent and robotic systems be operational in many advanced application domains.

The necessity of having complementary holistic learning is also validated in [11], in which the authors show the substantial importance of having contextual knowledge as "the information that surrounds a situation of interest in the world", as anticipated in [12]. They identified 3 main benefits an intelligent system can acquire from the ability to quickly recognize the context: *(i)* robustness to complete the required tasks, namely the performance of the systems, as well

as its sake of applicability *(ii)* adaptability to multiple operational conditions and application domains *(iii)* flexibility in tackling the main goal of a robot.

In the last years, many approaches have just proven the benefits of using contextual information for solving robot navigation problems. For example, the authors of [24] develop an intelligent mobile robot system that understands the semantics of human environments and the spatial relationships with and between humans. Context-awareness for person following is considered in [25], while in [26] the robot's speed is adjusted when it is in a hallway setting. To the best of our knowledge, instead, very few experiments have been conducted in the field of context-aware human-robot interaction [27]. Even fewer works have shown how crucial contextual information can be leveraged to make the robot aware of the environment, and also reactive to various situations when interacting with real people. For this reason, the main goal of our work is not only to promote new ideas in the field of human anctivity recognition and holistic video understanding but also to demonstrate how they can be exploited to provide key information for the development of active and assistive social robots. This will allow not only to make the robot operative in variable conditions but also to avoid prefixed models that may result in non-natural robot behaviors and interactions, since models that are based solely on predefined user scenarios and action scripts may not be able to take into account the uncertainty introduced by variations in the environment or unclear expectations from the user [27].

3 Datasets

The most popular and commonly used video datasets in the field of human anctivity recognition are strictly targeting highly specific human actions or sports recognition. As shown in table 1 early datasets in the HAR field like Hollywood [13] and UCF101 [8] were simple and were completely scripted datasets filmed in very ideal and fully controlled conditions. Moreover, they include very little variation in the ambiance parameters such as lighting, occlusion, and viewpoints. In most cases, the non-complex backgrounds and the non-intraclass variations in human movements make these datasets non-applicable for real-world applications.

More recent datasets like Something-Something [14], ActvityNet [9] and Kinetics [10] typically consider unconstrained videos, which emulate real environments. Concurrently, being datasets with million-scale samples, they provide a great increase in the number of labels and videos. However, they are always limited to a single semantic category, allowing to recognize human actions only, while leaving a significant gap towards describing the overall content of a video.

Table 1 shows as research has turned over the years towards a more global and comprehensive video dataset that can understand all the multiple aspects of reality including not only human activities but also different classes as object or scene. This holistic spirit is first observable in the SOA dataset [7] and YouTube-8M [15] that is the largest multi-label video classification dataset, composed of $\sim$ 8 million videos to recognize several visual concepts. Above all, the HVU dataset

Table 1. Recent evolution of modern datasets for human anctivity recognition.

Dataset	Year	#Videos	Total Labels	Focus
Hollywood [13]	2008	663	101	Basic Human Actions
HMDB [21]	2011	7K	51	Facial and Body Actions
UCF101 [8]	2012	13K	101	Sports
Sports1M [22]	2014	1 million	487	Sports
ActivityNet [9]	2015	20K	203	Complex Human Activities
Charades [23]	2016	10K	157	Person–Object Actions
Something Something [14]	2017	108K	174	Person–Object Actions
Kinetics600 [10]	2017	500K	600	Group Action, Person–Object Actions
SOA [7]	2018	562K	553	Scenes, Objects, Action
YouTube-8M [15]	2018	7 million	4716	Scenes, Objects, Actions, Events
HVU [1]	2020	572K	3142	Scenes, Objects, Actions, Events, Concept, Attribute

[1] has brought in last year more attention to holistic video understanding as a comprehensive and multi-faceted problem, as it encompasses the largest and most comprehensive list of semantic categories. HVU consists of 572k real-world trimmed video clips whose duration can vary from a minimum of 2s length to a maximum of 10s. Each video sample is associated with a set of labels (or tags). Each one of the tags can belong to 6 main semantic categories: scene, object, action, event, attribute, and concept, that are able to naturally capture real-world scenarios. There are in total 3142 labels with, on average, $\sim$ 2112 annotations per label between training, validation, and test set. The dataset is not manually annotated since it would require a vast amount of time due to the considerable number of labels and videos. The automatic annotation mechanism uses the Google Vision API [17] and Sensifai Video Tagging API [16], providing relatively coarse and approximate results and allowing to select exactly 30 tags for each video. Then the tags are adjusted and arranged manually to remove amiss labels. For a better understanding of the dataset, an example of a training data sample is shown in figure 2.

4 Problem statement

We model the problem of HAR as a Supervised Learning (SL) classification task over the labels of different categories C_i with $i \in [1, 6]$, more precisely, $C \in \{\text{object, action, concept, scene, attribute, event}\}$. For each category i, we assume there is a set of binary labels L_i with a cardinality that is category-dependent. In particular, a value $y \in L_i$, $i \in [1, 6]$, is equal to 1 if the corresponding label appears in the video, 0 otherwise.

fun,games,dance,girl,performing_arts,team,leisure_centre,
recreation,sport,entertainment,indoor_games_and_sports,youth,
leisure,performance,choreography,sport_venue,competition

Fig. 2. Video frame sample from HVU datset with corresponding set of tags from different semantic categories.

In the training phase, the vision-based system receives, as input, an unordered set of N videos $\{v_n\}_{n=0}^{N}$ of dimension $[H\text{x}W\text{x}C]$, with C fixed to 3 as we work with RGB data and H, W, the height and width of the videos, respectively. The length of the videos can be variable and depends on the specific dataset. Assigning a set of binary labels for each video and setting $L = \sum_{i=0}^{6} L_i$, we can write the dataset as $D = \left\{ \left(v_n, L_n \right)_{n=0}^{N} \right\}$. In the test phase, given a set of videos, we want to classify each video with its labels accordingly.

Finally, the vision module is exploited in a HRI setting to detect which scenario is more plausible in a real-world designed application and how the robot should act in line with the visual context.

5 Methodology

In this section, we first introduce the concept of Convolutional Neural Networks (CNNs) and how they work with data of different dimensionality. Next, we proceed by explaining the idea behind Recurrent Neural Networks (RNNs) and by reviewing the state of the art of this class of Artificial Neural Networks (ANNs). Finally, our model architecture is introduced to tackle the problem of classification while satisfying its requirements.

5.1 Convolutional Neural Network

Convolutional Neural Network is a deep learning model for processing image data in grid-shape matrices, with the primary goal of identifying features and extracting specific and even complex patterns from an image. A CNN can successfully capture the pixel spatial dependencies and recognize more sophisticated feature schemes to achieve effective results in classification and object recognition tasks than a simple feed-forward network. A CNN architecture comprises

multiple convolution layers, followed by pooling layers and fully connected layers. This connectivity structure reduces the size of the image and thus the number of parameters to be optimized during the training process, still keeping salient features crucial to accomplish the intended visual task. The core layer of a CNN is the convolutional one. In order to capture salient information, the image, i.e., a tensor, is convolved with a set of learnable filters, also known as kernels or weights. The filters are stacked together as multiple stages of feature extractor: earlier stages compute basic features, higher stages focus on more global and invariant features.

Given an input image I of dimension $H \times W$, and a squared filter K of dimension $F \times F$, a 2D convolution operation can be mathematically formalized as:

$$O[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} K[m,n] \cdot I[i-m, j-n] \tag{1}$$

where O is the output feature map, and each pixel location $O[i,j]$ is computed as the weighted sum of the original pixel and the eight nearby ones.

5.2 Recurrent Neural Network

Recurrent Neural Network (RNN) is introduced in [28]. This type of neural network is commonly used to process sequential data that show a temporal correlation. Its functioning consists of maintaining internal memory states, called *hidden states*, while handling data sequences of variable length, just like videos. The process of a Recurrent Neural Network carrying memory can be written as:

$$h_t = \phi(Kx_t + Th_{t-1}) \tag{2}$$

The formula states that the hidden state at time t depends on the previous hidden state at $t-1$ multiplied by a transition matrix T and the input at time step t multiplied by a weight matrix K. This sum is then given as input to an activation function ϕ that is typical $tanh$ or $ReLU$. Equation 2 can be seen as a loop in which each hidden state h_t maintain the memory of the state up to $t-1$, as long as this can be traced. Indeed, some of the drawbacks of RNNs are the vanishing and exploding gradient problems, resulting in major difficulties in keeping track of long-term dependencies.

Most of the RNN disadvantages have been solved with the introduction of a variation called Long Short-Term Memory (LSTM) [29]. This neural network, in fact, is capable of capturing long-term dependencies. LSTMs present a different and more sophisticated structure, starting from the addition of the cell states C_t as shown in figure 3. Moreover, an LSTM regulate its information flow using three different gates:

– Input gate

$$i_t = \sigma(W_i\,[h_{t-1}, x_t] + b_i)\tilde{C}_t = tanh(W_c\,[h_{t-1}, x_t] + b_c) \tag{3}$$

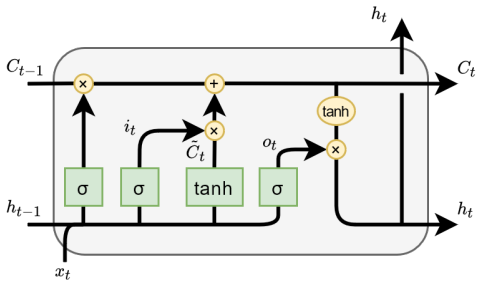

Fig. 3. Overview of a LSTM repeating module.

– Forget gate

$$f_t = \sigma(W_f\,[h_{t-1}, x_t] + b_f) \tag{4}$$

– Output gate

$$C_t = f_t\,C_{t-1} + i_t\,\tilde{C}_t o_t = \sigma(W_o\,[h_{t-1}, x_t] + b_o) h_t = o_t\,tanh(C_t) \tag{5}$$

These gates together completely solve the vanishing gradient problem that occurs in the vanilla RNNs. While LSTMs are well suited for capturing temporal correlations along an input sequence, they require much memory and may take longer to train.

In the last years, a new class of RNN has been introduced in [30], namely, Gated Recurrent Unit (GRU). GRUs inherits LSTMs structure with the difference that the output gate is discarded. Indeed, the hidden state is completely exposed, without any control, while the cell state C is not used anymore. In this way, the structure of GRUs results less complex and so computationally more efficient than LSTMs. Another note in favor of GRUs is that they perform better on smaller datasets and almost equal to LSTMs on bigger amounts of data [31].

5.3 CNN + RNN Architecture

While CNNs are powerful feed-forward artificial neural networks suitable for spatial data, on the other side, sequential data like videos represent temporal information of arbitrary length which is better handled by Recurrent Neural Networks. However, the authors of [1] show how Convolutional Neural Networks that work in 3 dimensions (3D CNNs), in combination with 2D CNNs, can capture both spatial and temporal details from sequential data. Moreover, previous works like [32] and [33], already proposed the use of 3D ConvNets in HAR achieving promising outcomes. This simple architecture consists of expanding the 2D Convolution explained in section 5.1 with a third dimension that represents the time steps. In this way, videos can be processed as a series of 3D volumes. Furthermore, recent works have reached state-of-the-art results combining 3D CNNs with RRNs ([34], [35], [36]) in human anctivity recognition and prediction tasks.

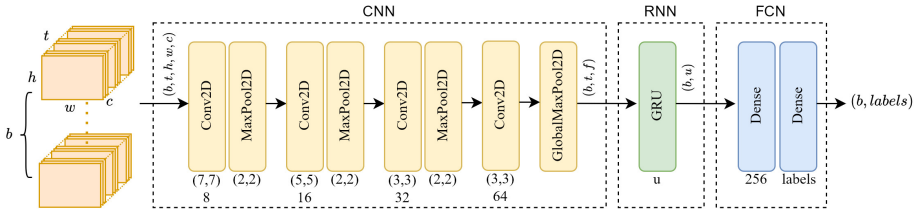

Fig. 4. Illustration of our model architecture. The annotations under the layers represent the number of filters, units or, kernel size, depending on layer type.

Although higher dimension ConvNets help to catch richer motion information, they bring a substantial number of parameters that make the computation more complex. In addition, the purpose of this work is to design a robotics application that most likely has limited computational resources which are in contrast with 3D CNNs characteristics.

A reasonable architecture that meets the requirements of our application can be a simple concatenation of 2D Convolutional layers with an RNN module. In particular, as mentioned in section 5.2, GRUs are, in terms of performance, comparable to LSTMs while also exploiting a much lighter structure. For this reason, we propose the use of a CNN-GRU architecture to solve the first task of video labels classification problem.

As shown in figure 4, our model architecture is composed of several layers. The input of the network is a batch of videos that can be seen as a set of images (frames) sequences or, formally, as a tensor of dimension (b, t, h, w, c) where:

- b: batch size
- t: time steps (frames in the video)
- h: height of the images
- w: width of the images
- c: images channels

Input data flows into a CNN block made of groups of 2D Convolutional layers and 2D Max Pooling operations. As explained in subsection 5.1, this process reduces input size while expanding its depth in terms of feature extraction. Afterward, the data tensor dimension is reduced using a 2D Global Max Pooling operation and its shape becomes (b, t, f) with f the final number of features, in this case, 64. The three-dimension CNN output is straightly given as input to the GRU block. This latter looks for temporal correlations in the features of the videos and returns a matrix of dimension (b, u) which are the hidden states of each video at their last timestep. In fact, u is the number of units of the GRU which also corresponds to hidden states length. A set of fully-connected layers is then placed after the recurrent action. In particular, we use a *Sigmoid* activation function for the very last *Dense* layer, in order to predict a set of probabilities in the [0,1] range for each label in all categories. Therefore, the number of units of this last layer depends on the total number of labels in the dataset.

6 Experiments

The model architecture introduced in the previous section is evaluated with a subset of the HVU [1] dataset. We first introduce how we select a portion of the data and which metrics are used to carry out this task. We proceed by listing some of the implementation details of our work. In the end, we analyze and discuss the training and test phases of our model.

6.1 Dataset

We collected part of the data from the Large Scale holistic video understanding dataset instead of using it entirely. Indeed, two strong reasons led us not to use the entire dataset. The first motivation is that the HVU dataset has more than 3k labels and not all of them might be of interest in designing an HRI application. Selecting a subset of labels, consequently, reduces the number of annotations, and so videos. The second reason is that a large dataset of 572k videos requires a model with many parameters and, in consequence, a huge number of computational resources.

The first metric we adopt to lighten the dataset is to re-order the labels of each category by the number of their annotations in the videos. In this way, we have a clear view of the most frequent labels. Next, we pick a subset of labels for each category from the ordered list. Moreover, we manually add some labels which we think to be helpful in an HRI scenario. These two actions lead us to a total number of 80 selected labels, about 13 per category. After an initial screening of the labels, we select the videos to be used in the training and test steps. To accomplish the task, we use an effective metric, which we name discard factor (d_f). This parameter has the task of discarding videos that contain a low number of activations, i.e., the number of 1s in the binary vector of labels describing the video, as explained in section 4. Given a video, we compare the discard factor with the ratio between the number of labels present in the video that belong to our 80 selected labels and the number of its total annotations. If this ratio is below d_f then the video is discarded. Next, we drop from the video the labels different from the selected ones and ensure each video to have at least 3 activations. In this manner, we get many samples with multiple labels present at the same time, ensuring the model recognizes multiple entities together. This is also done to increase as much as possible the ratio between 1s and 0s in the binary vectors of labels. Indeed, using a really small number of non-weighted activations could lead the model to always predict zeros. The other important reason we use the discard factor is to obtain a set of videos that don't refer to an excessively complex context that differs too much from the background areas described by the selected labels. In terms of our implementation, we have set d_f=0.68. Finally, we further limit the number of annotations for each of the 80 selected labels to be 4000. In this manner, we reduce the imbalance of video annotations among the selected labels, avoiding too many activations for a set of them. In particular, given the annotations of a video, if a label has reached its limit and is among them, we discard the entire video.

We prefer to have a balanced number of labels for each category instead of an equal number of annotations for each label. In this way, we show the model's fair ability to recognize different labels from different categories. An alternative would be to balance both things setting this up an optimization problem. This, however, could result in a long and complex procedure. This problem can be resumed as a choice between innovation and performance. In the first case, we push the model to recognize diverse categories of entities. In the second, we aim at a balanced number of annotations and so enough data for each label not to compromise the model's capability while renouncing on recognizing different context areas.

After these preprocessing steps, we obtain a dataset of $\sim$ 10k videos. Since the original HVU dataset is composed of videos of different aspect ratios, including vertical videos, we decide to apply a center cropping to obtain a fixed aspect ratio. During the experiments, this operation doesn't show notable disadvantages in model performance. In particular, we use 0.66 as aspect ratio and resize the videos to be $(vid_w, vid_h, vid_c) = (150, 100, 3)$. Finally, we discard videos having less than 60 frames, which corresponds to 2 seconds if the video is recorded at 30 fps (like the majority of the videos from the HVU dataset), because during training each sample is clipped at 60 frames from the start in order to have sequences with the same length inside a batch. Thereby, we obtain a final dataset of 8k videos.

6.2 Implementation Details

For the implementation, we highly rely on TensorFlow 2.4 framework. In particular, we build an efficient data pipeline to process videos as a set of images using *TFRecords* format [18] and *tf.data* API [19]. This significantly increases performance during the training process.

We train the network from scratch using mini-batch gradient descent with a batch size of b=32. We normalize the dataset dividing each video frame by 255 to transform values in [0,1] range and help the convergence of the model. Moreover, we cut videos from the start to have exactly 60 frames and process simultaneously multiple batches (batch, frames, height, width, channels). However, we would like to point out that, at inference time, the model is still capable of classifying videos of variable length. The validation set is composed of the 20% of the dataset, and we use Adam optimizer with learning rate lr = 1e-4, and decay equal to 0.9. We decide to fix the number of GRU units to 64 and use a dropout value of 0.2 to prevent overfitting. All these values come from a fine-tuning procedure. For the loss, we use binary cross-entropy (BCE) since our ground truth labels are either zeros or ones. In particular, we implement a modified version of BCE, which weighs the values 1 in the labels 2.5 times more than the zeros. In this way, we adjust the balance of annotations in the videos. The evaluation metrics that we use are binary accuracy, precision, and recall, all with a threshold of 0.5. Our network has 11 layers for a total number of 77k parameters, which makes it pretty light.

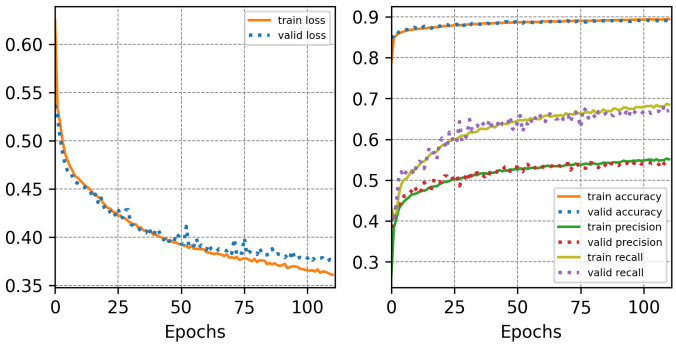

Fig. 5. (Left) Training and validation curves for weighted binary cross-entropy loss. (Right) Training and validation curves for accuracy, precision, and recall.

To accomplish our experiments we make use of an NVIDIA RTX 3060 12GB GPU, while the mean training time for one epoch (validation step included) is about 270 sec.

6.3 Result analysis

The performance results of our model on the parsed dataset are shown in figure 5. We do not report HATNet [1] as a baseline since no source code has been made available by the authors. Moreover, not enough implementation details were provided to reproduce the network and use it on our dataset.

The presented results refer to a training procedure of 110 epochs. The left plot shows the loss curve, which is a weighted binary cross-entropy. As evidenced by the trend, the network presents minor issues of underfitting in the first part caused by the small dropout in the GRU layer. In the end, there is a small gap of overfitting between the training and validation losses. Instead, the accuracy on the right plot presents a stable and slowly-growing behavior and reaches, at the last epoch, a score of 0.89. Conversely, the precision reaches a lower value of ~ 0.55 while also showing no overfitting tendency. The highlighted difference between accuracy and precision could be due to the unbalanced distribution of the labels' video annotations. The recall, instead, reaches a value of ~ 0.68 both for training and validation. In figure 6, we show some video frames from the test set and their relative output labels predicted by the network. In the first frame, the model perfectly recognizes all the featured labels. In the second example, it is possible to see a missing label, i.e., *infant*, from the predicted ones. Moreover, one may notice that some of the predicted labels (in light red) agree with the context of the frame while not being included in the ground truth. This occurs because, as explained in section 3, the HVU dataset uses APIs for automatic annotation of the videos and, just partially, human verification to add correct labels and remove wrong ones. The third video frame also demonstrates that some labels, which are considered false positives, are instead present in the

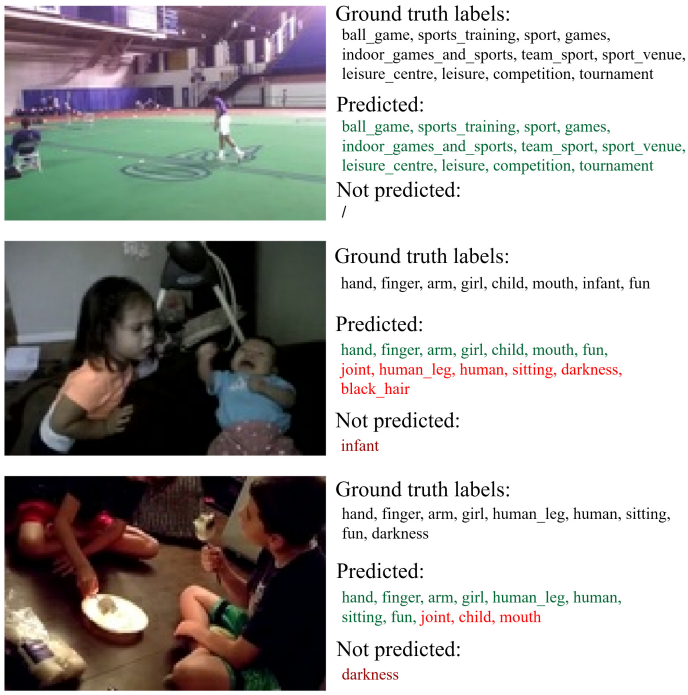

Fig. 6. Model predictions on some test video frames. We report well predicted (green), not predicted (dark red), and false-positive (light red) labels.

video and would become true positives in a more accurately annotated dataset. Thus, while HVU videos might be poorly noted, we show that the model can learn labels' features and predict them even when the ground truth is wrong. On the other hand, wrong ground truth data penalizes the network's learning process and shows performance results different from reality.

In summary, the network can recognize the labels in the video and perceive many true positives. Instead, it suffers from false positives due to the missing labels of the HVU dataset. This can also be noticed by the recall being higher than the precision. These results confirm that the model can capture the spatial and temporal information in the videos and predict in a good way ground truth labels.

7 Context-aware human-robot interaction

In this chapter, we discuss and analyze a possible HRI application that exploits the holistic scene understanding vision module. The aim is to elaborate on how the robot's interaction performance can be improved when contextual knowledge is provided to the embedding system. Indeed, human-robot interaction can take several advantages from being context-aware [27]. First of all, the ability to

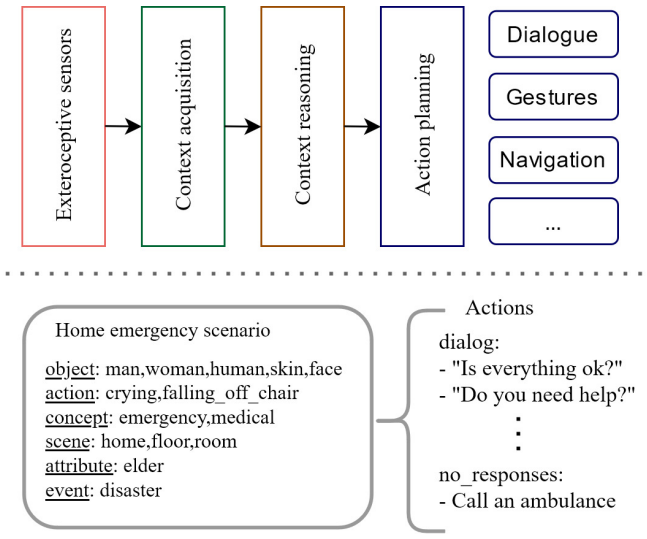

Fig. 7. (Up) A robot framework in an HRI setting. (Down) Example of context scenario.

be operative and deployable in different application domains, and adaptable to multiple operational conditions, which is a key factor for achieving precision and robustness when accomplishing the desired task. Secondly, the information extracted from the context can be extremely useful to control the interaction so that it appears much more natural according to the particular situation. Moreover, this induced behavior could also be an important milestone in terms of the robot's social acceptance.

We propose a complete architectural framework for developing a context-based human-robot interaction. It includes three main components: (*i*) a vision-based context acquisition module with the aim of extracting contextual information labels from the environment by using a holistic scene understanding; (*ii*) a context reasoning module for translating contextual knowledge into agent's behaviors; (*iii*) an action planner module to control the interaction employing motion, gestures, communication, and other means. A complete sketch of the framework is shown in the upper part of figure 7.

7.1 Application

We present a proof of concept on a possible application that makes use of the holistic scene understanding capabilities. In its general form, an HRI framework should be able to perceive environmental information. In particular, we require the robot to possess a vision system, such as an accessible low-resolution RGB camera. For the moment, we exclude the use of advanced and expensive sensors like LiDARs. Since we are in a HRI setting, the robot must have means of communication, such as automatic speech recognition and Text-To-Speech (TTS).

This is necessary to expose a realistic behavior to the user while not making him disdain the technological capabilities of the robot. Another attribute of interest, especially for social robots, is to be capable of gesturing and, eventually, to move in the environment.

When a robot approaches a human, it needs to have some sort of context reasoning and context representation. Specifically, when we refer to the context we point out a set of high-level features that improve robots' capability to adapt in the real world while standing beside humans. In our application, we use different categories labels from the real world as high-level features for robots' understanding. Once we have the representation of the context, we move forward with the reasoning part. As a result, we introduce the concept of context scenario which is defined as *an unordered set of category labels that, put together, refer to a particular situation*. Scenarios are, in general, hand-crafted and customizable by the final user. The aim of the context reasoning part in the framework is to select the Most Likely Scenario (MLS) using the probabilities of the labels extracted from the vision module. In this way, is possible to link a set of actions to the various scenarios and let the robot plan an action, based also on humans' requests and needs. As shown in the bottom part of figure 7, we provide an example of context scenario. In the first part, we select a set of the proposed labels which comes divided into 6 categories. Subsequently, a group of possible actions is linked to the scenario, letting the robot plan among them. For this purpose many types of planning can be used, like the one based on user interactions as proposed before.

Among the possible test cases, one of the most interesting that could happen and has to be analyzed is when multiple scenarios occur. To solve this issue we propose priority-level context scenarios. As said before, a scenario takes place when it's considered the most likely one. This imposes the check of scenario's labels probabilities and a threshold value that establishes when a particular situation is happening. However, this doesn't guarantee the recognition of a single scenario, especially in a wide multi-tasking environment. Apart from the likelihood, we consider also a priority level which makes the difference in multiple-scenarios settings. In this way, urgent scenarios can take place safely while same-priority scenarios can be selected randomly when overlapping happens.

8 Conclusion and further work

In this work, we address the problem of Human Action Recognition (HAR) and expand its horizon presenting holistic approach for video understanding in Human-Robot Interaction (HRI). This new perspective consists of recognizing human activities while also considering other key factors like the scene, the objects, or the concept. We propose a CNN-RNN architecture to solve the problem of multi-labeled video classification. The results show that the proposed methodologies can be well suited to a medium-sized dataset. We further design an HRI scenario-based application showing its possible benefits and test cases.

In the future, we want to focus on different application fields suitable for our conceptual HRI pipeline. In the first place, the development of a scenario-based application in a social robot like Pepper from SoftBank Robotics [39]. This, together with human questionnaire analysis [38], would show the effects of context acquisition on robot behaviors, which we think could be particularly beneficial.

In certain circumstances, such as in a nursing home, the constant presence of the operator is not always possible and this can result in possible episodes or moments in which the elderly person remains alone when he decides to take an action. The presence of the robot would therefore allow to offer a logistic support and perform functions in order to help the elder. For example, when an elder is alone and wants to reach an object far from his reach, the robot recognizing the action of the person can help the elderly in carrying at the end of his behavior aimed at a purpose. The same scenario can be applied for youngsters and children, enforcing a list of allowed and forbidden objects to make reachable. In facts, this same principle can be extended into other contexts where there are different types of frailty. In this way the fragile individual could enjoy a moment in which his request is listened to and welcomed, giving the child a feeling of gratification and interaction.

A further improvement for the HRI system could also include more advanced forms of automatic reasoning. As already done in [27] and [37], robot's interaction behaviors can be modeled using, specific to a given context, a Partially Observable Markov Decision Process (POMDP). This will allow us to condition on the observations, that is the contextual information, the future actions of the robot.

Another possible direction of future work could explore a similar strategy for developing a visual context-aware Automatic Speech Recognition system (VC-ASR). The basic idea is exploiting visual signals and contextual knowledge to improve the robustness and reliability of ASR, with particular attention to grounding and re-ranking. Indeed, we can leverage the acquired visual context to re-rank the lists of transcriptions, and to ground text hypothesis from the first pass of ASR.

References

1. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L., *Large Scale holistic video understanding*, European Conference on Computer Vision (ECCV), 2020.
2. Jaimes, A., Sebe, N., *Multimodal human–computer interaction: A survey*, Computer vision and image understanding, Elsevier - 2007.
3. Aggarwal, J.K., Xia, L. *human anctivity recognition from 3D data: A review*, The University of Texas at Austin, USA, 2014.
4. Li, W., Zhang, Z., Liu, Z., *Action recognition based on a bag of 3D points*, Computer Vision and Pattern Recognition (CVPR), 2010
5. Chen, L., Wei, H., Ferryman, J., *Tracking-based 3D human skeleton extraction from stereo video camera toward an on-site safety and ergonomic analysis*, Computational Vision Group, School of Systems Engineering, University of Reading, UK, 2013.

6. Liu, M., Han, S., Lee, S., *A survey of human motion analysis using depth imagery*, Construction Innovation, Vol. 16 No. 3, pp. 348-367, 2016.
7. Ray, J., Wang, H., Tran, D., Wang, Y., Feiszli, M., Torresani, L., Paluri, M., *Scenes-Objects-Actions: A Multi-Task, Multi-Label Video Dataset*, European Conference on Computer Vision (ECCV), 2018.
8. Soomro, K., Zamir, A.R., Shah, M., *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*, Center for Research in Computer Vision (CRCV), 2012.
9. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C., *ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
10. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A., *The Kinetics Human Action Video Dataset*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
11. Bloisi, D.D., Nardi, D., Riccio, F., Trapani, F., *Context in Robotics and Information Fusion*, IEEE CSpringer International Publishing, 2016.
12. Snidaro, L., García, J., Llinas, J. *Context-based Information Fusion: A survey and discussion*, Information Fusion, 2015.
13. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B., *Learning realistic human actions from movies*, 26th IEEE Conference Computer Vision Pattern Recognition, pp. 1–8. (CVPR), 2008.
14. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al., *The "something something" video database for learning and evaluating visual common sense*, IEEE International Conference of Computer Vision (ICCV), 2017.
15. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S., *Youtube-8m: A large-scale video classification benchmark.*, CoRR, abs/1609.08675, 2016.
16. Sensifai video tagging API: www.sensifai.com
17. Google vision AI API: cloud.google.com/vision
18. TFRecord TensorFlow Tutorial: www.tensorflow.org/tutorials/load_data/tfrecord
19. tf.data TensorFlow API: www.tensorflow.org/api_docs/python/tf/data
20. Beddiar, D.R., Nini, B., Sabokrou, M., Hadid, A., *Vision-based human anctivity recognition: a survey*, Multimedia Tools and Applications 79, 30509–30555, 2020.
21. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., *HMDB: a large video database for human motion recognition*, Proceedings of the IEEE International Conference on Computer Vision, pp. 2556–2563, 2011.
22. Karpathy, A. et al., *Large-scale video classification with convolutional neural networks*, 2014.
23. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A., *Hollywood in homes: crowdsourcing data collection for activity understanding*, Lecture Notes in Computer Science, vol. 9905, pp. 510–526.LNCS, 2016.
24. Cosgun, A., Christensen, H.I., *Context-aware robot navigation using interactively built semantic maps*, Paladyn Journal of Behavioral Robotics, 2018.
25. Zender, H., Jensfelt, P., Kruijff, G., *Human-and situation-aware people following*, 16th IEEE International Symposium on Robot and Human interactive Communication (RO-MAN), 2007, pages 1131–1136.
26. Pacchierotti, E., Christensen, H.I., Jensfelt, P., *Human-robot embodied interaction in hallway settings: a pilot user study*, IEEE International Workshop on Robot and Human Interactive Communication (ROMAN), 2005, pages 164–171.

27. Quintas, J., Martins, G.S., Santos, L., Menezes, P., Dias, J., *Toward a Context-Aware Human–Robot Interaction Framework Based on Cognitive Development*, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, 2018.
28. Rumelhart, D., Hinton, G., and Williams, R., *Learning representations by back-propagating errors*, nature 323, 6088 (1986), 533, 1986
29. Hochreiter, S., and Schmidhuber, J., *Long Short-term Memory*, Neural computation, 1997, pages 1735-80.
30. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y., *Learning phrase representations using RNN encoder-decoder for statistical machine translation*, arXiv preprint arXiv:1406.1078, 2014
31. Chung, J., Gulcehre, C., Cho, K., and Bengio, Y., *Empirical evaluation of gated recurrent neural networks on sequence modeling*, arXiv preprint arXiv:1412.3555, 2014
32. Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M., *Learning Spatiotemporal Features with 3D Convolutional Networks*, IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489-4497.
33. Ji, S., Xu, W., Yang, M., and Yu, K., *3D Convolutional Neural Networks for Human Action Recognition*, IEEE TPAMI, 35(1):221–231, 2013.
34. Yao, L., and Qian, Y., *DT-3DResNet-LSTM: An Architecture for Temporal Activity Recognition in Videos*, Advances in Multimedia Information Processing - PCM 2018, Springer International Publishing, pp. 622-632, 2018.
35. Umamakeswari, A., Angelus, J., Kannan, M., and Bragadeesh, S. A., *Action Recognition Using 3D CNN and LSTM for Video Analytics*, International Conference on Intelligent Computing and Communication. Springer, Singapore, 2020.
36. Alfaifi, R., and Artoli, A. M., *Human Action Prediction with 3D-CNN*, SN Computer Science, 1.5: 1-15, 2020.
37. Kim, J., *POMDP-based Human-Robot Interaction Behavior Model*, Journal of Institute of Control, 20(6):599-605, 2014.
38. Bartneck, C., Croft, E., Kulic, D. and Zoghbi, S., *Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots*, International Journal of Social Robotics, 1(1) 71-81, 2009.
39. Pandey, A. K., and Gelin, R., *A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind*, IEEE Robotics & Automation Magazine, 2018.