Received August 20, 2021, accepted September 19, 2021, date of publication September 29, 2021, date of current version October 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3116882

OSVidCap: A Framework for the Simultaneous Recognition and Description of Concurrent Actions in Videos in an Open-Set Scenario

ANDREI DE SOUZA INÁCIO[1,2], MATHEUS GUTOSKI[1], ANDRÉ EUGÊNIO LAZZARETTI[1], (Member, IEEE), AND HEITOR SILVÉRIO LOPES[1]

[1]Graduate Program in Electrical Engineering and Industrial Informatics, Federal University of Technology—Paraná, Curitiba 80230-901, Brazil
[2]Federal Institute of Santa Catarina, Gaspar, Santa Catarina 89111-009, Brazil

Corresponding author: Andrei de Souza Inácio (andrei.inacioo@gmail.com)

The work of Andrei de Souza Inácio was supported in part by the UNIEDU/FUMDES Program for the Ph.D. scholarship. The work of Matheus Gutoski was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the Scholarship under Grant 141983/2018-3. The work of Heitor Silvério Lopes was supported in part by the CNPq and Fundação Araucária under Grant 311785/2019-0 and Grant PRONEX 042/2018.

ABSTRACT Automatically understanding and describing the visual content of videos in natural language is a challenging task in computer vision. Existing approaches are often designed to describe single events in a closed-set setting. However, in real-world scenarios, concurrent activities and previously unseen actions may appear in a video. This work presents the OSVidCap, a novel open-set video captioning framework that recognizes and describes, in natural language, concurrent known actions and deal with unknown ones. The OSVidCap is based on the encoder-decoder framework and uses a detection-and-tracking-object-based mechanism followed by a background blurring method to focus on specific targets in a video. Additionally, we employ the TI3D Network with the Extreme Value Machine (EVM), which learns representations and recognizes unknown actions. We evaluate the proposed approach on the benchmark ActivityNet Captions dataset. Also, an enhanced version of the LIRIS human activity dataset was proposed by providing descriptions for each action. We also provide spatial, temporal, and caption annotations for existing unlabeled actions in the dataset – considered unknown actions in our experiments. Experimental results showed our method's effectiveness in recognizing and describing concurrent actions in natural language and the strong ability to deal with detected unknown activities. Based on these results, we believe that the proposed approach can be potentially helpful for many real-world applications, including human behavior analysis, safety monitoring, and surveillance.

INDEX TERMS Video captioning, open-set recognition, deep learning.

I. INTRODUCTION

Video understanding is a challenging issue in computer vision. It requires sophisticated techniques to process the diversity of humans and objects appearances in different environments and their relationships over time.

The ability to detect and identify specific events is also a critical step towards video understanding. Video events are high-level semantic concepts perceived by humans in a video sequence [1]. Each event is composed of one or more meaningful objective actions, such as walking or jumping, and interaction with objects, such as typing a computer or handshaking [2]. Each perceived concept consists of an entity (human, object, action, or scene attributes) that occupies a specific position in a frame and may vary in size, color, shape or other specific attributes.

Video description (also called video captioning) is one of the many problems under video understanding. It has become a hot topic in computer vision and deep learning [3] and requires solving many different tasks simultaneously, including object detection and classification, action detection and recognition, and visual relationships among humans and objects. A video description approach may be employed in various applications such as human-robot interaction, video indexing, assistance to the visually impaired, understanding sign language, and video surveillance.

The associate editor coordinating the review of this manuscript and approving it for publication was Khoa Luu.

Current deep learning techniques are effective to learn discriminative spatio-temporal features from raw data. They are used to solve several complex tasks, such as object detection and classification [4], human action recognition [5], [6], video summarization [7], semantic image segmentation [8], and video understanding [9]. However, a step beyond the simple categorical classification of actions in scenes is to describe events in a human-comprehensible language. To accomplish this, it is crucial to understand the semantics of a given video scene.

Despite the efforts and progress that have been made in the video description task, it is still an open problem and has attracted much attention [3]. Existing approaches are limited to the fixed list of activities in the training corpus and have focused on generating a holistic description of short-length videos with only one main action happening in the video. However, in practical applications, such as safety monitoring and surveillance, videos may have concurrent activities, and humans can perform many different actions and even create new movements and hand gestures at will.

A more realistic approach is to assume an open-set scenario for describing actions. Open-set classifiers allow performing classification by enclosing each class in the feature space and reserving space for new classes to emerge, unlike closed-set classifiers, which assign infinite spaces to training classes. This strategy allows rejecting data from previously unknown classes instead of wrongly assigning the class label with the highest probability value [10].

Following this idea, a video captioning approach in an open-set scenario can adequately describe known actions and deal with unknown ones. Thus, it is essential to detect if the performed action was seen during the training step to correctly describe known actions or activities and avoid generating wrong descriptions of new detected actions.

Based on that, this work presents a novel open-set video captioning framework that aims to describe, in natural language, not only single but also concurrent events occurring in a video. The proposed approach uses an open-set action recognition model to detect unknown actions, thus avoiding incorrect descriptions and hallucinations. Some recent works have successfully performed video action recognition in an open-set scenario [10], [11]. However, to the best of our knowledge, this is the first time such properties are explored in the video captioning task.

The proposed representation learning approach is based on the encoder-decoder framework and uses a detection-and-tracking-object-based mechanism followed by a background blurring method to define the targets and recognize the concurrent actions to be described. Additionally, we employ the Triplet Inflated 3D Neural Network recently proposed by [11], which uses Deep Metric Learning and the Extreme Value Machine (EVM) [12] as the open-set classifier. The main contributions of this paper can be summarized as follows:

- We propose a novel video captioning framework to recognize and describe concurrent actions/activities performed by humans in an open-set scenario;
- We present a novel open-set mechanism to detect out-of-domain videos of unseen activities;
- We present extensive experiments and analysis, using 2D and 3D feature representations, demonstrating the effectiveness of our approach.

The remainder of this paper is organized as follows. Section II presents a brief description of related works. In Section III, we present the theoretical aspects related to the proposed method for open-set action recognition. In Section IV, we describe in detail the proposed framework. Next, in Section VI, we present the experimental settings, their results, and a discussion. Finally, in Section VII we present the conclusions and suggestions for future research directions.

II. RELATED WORKS

Early proposed methods for the video description task started with template-based methods in which the Subject (S), Verb (V), and Object (O) were detected and then, used in a sentence template [3]. Although these methods could generate descriptions based on grammar, they did not take into account the spatial and temporal associations between entities and suffered from the lack of diversity of generated sentences. Inspired by the rapid development of deep learning techniques in the Computer Vision and Natural Language Processing area, video description research has recently become a hot topic.

The video description approaches based on deep learning methods are mainly designed in the encoder-decoder architecture [3], [13]. The encoder is usually a combination of 2D and/or 3D CNN and LSTM that converts the input into a feature vector representation of fixed length. The decoder is usually an LSTM or GRU that generates a sequence of words.

Pre-trained deep learning models, such as VGGNet [14] or ResNet [15], are commonly used to extract spatial features from frames. These features are usually combined across the frames by an average pooling or max-pooling operation, resulting in a single fixed-length feature vector representation for a short video clip. Besides, the C3D [16] or I3D [17] models, pre-trained in a large dataset such as the Sports-1M dataset [18] or Kinetics dataset [19], are used to extract temporal features. The use of pre-trained models on large datasets provides a strong visual representation of objects, actions, and scenes depicted in the video [20].

Reference [20] proposed the first end-to-end learning approach based on deep neural networks for the video captioning task. A variant of AlexNet pre-trained on a subset of the ImageNet [21] dataset was used to extract visual features from frames. Then, the mean pooling method was employed, resulting in a single vector representing the entire video. Finally, two stacked LSTM was used to generate the sentence.

Since then, many approaches have been proposed to use attention mechanisms to dynamically select spatial and temporal features focusing on important frames and regions inside them, providing meaningful visual evidence for caption generation [22]–[26]. The use of attention mechanism has improved the video captioning task suggesting that the this method can efficiently improves the descriptions, especially in discontinuous videos, by focusing on specific parts of the visual input.

Considering that open-domain videos cover a broad range of topics, such as sports, music, food, and so on, some approaches have been proposed to generate sentences guided by latent topics [27] and semantic attributes [28]. The use of multimodal data, such as visual, audio, motion, and textual information, was also explored in some works [29], [30]. The combination of audio, movement, and visual information has been shown to play an important role in the description generation process.

The dense-captioning events task was proposed by [31] and consists of detecting and identifying all events in a given video and describe them in natural language. Their proposed approach uses DAPs [32] to localize temporal event proposals and a caption module based on LSTM to generate a sentence for each event proposal. Reference [33] also propose a unified end-to-end approach for video dense captioning. However, instead of using RNN for description generation, the authors used Transformers [34]. Their proposed approach is composed of three components: a video encoder, a proposal decoder, and a captioning decoder. The video encoder is composed of multiple self-attention layers. The Temporal Action Proposal (TAP) is based on ProcNets [35], which was designed to detect actions in long videos. Moreover, the captioning decoder module uses Transformers to generate the sentence for each event proposal.

Despite achieving promising results, these approaches often fail to describe concurrent activities happening in a video. Also, the datasets used to evaluate these approaches are created with videos extracted from movies or YouTube videos. Such videos cover a broad range of topics, such as sports, music, food, and so on, and a wide variety of different individual and collective actions performed by humans, animals, and even moving cartoon objects. These videos also present specific challenges, including the presence of discontinuity points between frames, as reported by [36], which may result in inadequate temporal representation features.

Besides the limitations presented above, the lack of well-labeled data is a crucial problem in the deep learning area. The zero-short learning task has been studied to classify actions with no or few examples during the training step [28], [37]. Some approaches have been proposed for visual descriptions task [38], [39] to describe novel objects not presented in paired image sentence dataset. The zero-shot video captioning task [40] focuses on describing out-of-domain of a novel activity without paired captions, but with the knowledge of the activity.

The approaches presented so far assume that all possible classes are already known during the train or test phase. However, new classes emerge as time passes in the real-world dynamic environments. An open-set Human Action Recognition approach requires the classifier to accurately classifies known classes seen during the training stage and deals with unknown classes, which are unseen and with no semantic information provided during the training stage [10]. In this work, we also exploit the nature of the open-set recognition problem to propose a framework to describe videos in an open-set scenario. As previously stated, to the extent of our knowledge, there is a lack of related works in this approach in the literature, being the main original contribution of the present work.

III. THEORETICAL ASPECTS

This Section presents the fundamentals of the methods used in our open-set recognition module: the Extreme Value Machine and the Triplet Inflated 3D Neural Network.

A. THE EXTREME VALUE MACHINE

The Extreme Value Machine (EVM) was initially proposed by [12] to perform open-set classification. In the EVM, the modeling of each class in the training set is based on a set of extreme vectors, which are associated to a Probability of Sample Inclusion (Ψ).

The key concept of EVMs is the use of margin distributions, which is the distribution of the half margin distances of the training data. In the original formulation, one can consider $\mathbf{x}_i$ as a training sample and y_i the corresponding label. Considering $\mathbf{x}_i$ and $\mathbf{x}_j$, where $\forall j, y_j \neq y_i$, $\mathbf{x}_j$ can be considered the nearest point to $\mathbf{x}_i$ and, in this case, the margin estimate for the pair $(\mathbf{x}_i, \mathbf{x}_j)$ is given by $\mathbf{m}_{ij} = ||\,\mathbf{x}_i - \mathbf{x}_j\,||/2$.

The $\mathbf{m}_{ij}$ value can be computed for the τ nearest points and the distribution of the margins is estimated with those points using the Extreme Value Theorem (EVT). The EVT states that the minimum values of $\mathbf{x}_i$ is given by a Weibull distribution [12]. The probability of inclusion Ψ for a point $\mathbf{x}'$ is given by

$$\Psi(\mathbf{x}_i, \mathbf{x}', \kappa_i, \lambda_i) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}'\right\|}{\lambda_i}\right)^{\kappa_i}, \tag{1}$$

in which $\left\|\mathbf{x}_i - \mathbf{x}'\right\|$ is the distance between $\mathbf{x}'$ and $\mathbf{x}_i$, λ_i and κ_i are the Weibull's shape and scale parameters.

Each Ψ is considered an EVT rejection model and $\Psi(\mathbf{x}_i, \mathbf{x}', \kappa_i, \lambda_i)$ corresponds to the probability that a sample is not beyond the negative margin. Even though a sample has zero probability around the margin, the model can also be extended to support soft margins. The probability that a point $\mathbf{x}'$ belongs to class C_l, where l is the class index, is given by Equation 2:

$$\hat{P}(C_l|\mathbf{x}') = \text{argmax}_{i:y_i=C_l}\Psi(\mathbf{x}_i, \mathbf{x}', \kappa_i, \lambda_i). \tag{2}$$

Finally, the classification function is:

$$y^* = \begin{cases} \text{argmax}_{i:y_i=C_l} \hat{P}(C_l|\mathbf{x}'), & \text{if } \hat{P}(C_l|\mathbf{x}') \geqslant \delta \\ \text{unknown}, & \text{otherwise}, \end{cases} \quad (3)$$

in which δ is a threshold responsible for defining the boundary between known and open-space.

In order to reduce the size of the model, many redundant $[\mathbf{x}_i, \Psi(\mathbf{x}_i, \mathbf{x}', \kappa_i, \lambda_i)]$ pairs can be discarded with minimal impact on performance. Details of this procedure can be found in [12].

B. TRIPLET INFLATED 3D NEURAL NETWORK (TI3D)

The TI3D is a Deep Metric Learning Neural Network introduced in [11]. It uses the I3D as the base model to build a cosine triplet loss network. The TI3D learns a feature mapping such that intra-class distances are small and inter-class distances are large.

The TI3D takes three inputs: Anchor, Positive, and Negative. For the human action recognition task, the Anchor (a) represents a video of any given action, the Positive (p) represents a video of the same action, and the Negative (n) represents a video of a different action, both w.r.t. the anchor. Given N (a, p, n) triplets, the Triplet loss function L is defined by:

$$L_\Theta = \sum_{i=1}^{N} \left[\Theta(f(\mathbf{x}_i^a), f(\mathbf{x}_i^p)) - \Theta(f(\mathbf{x}_i^a), f(\mathbf{x}_i^n)) + \alpha\right]_+ . \quad (4)$$

in which i is the triplet index, $f(\mathbf{x}^a)$, $f(\mathbf{x}^p)$, $f(\mathbf{x}^n)$ are the Anchor, Positive and Negative embeddings, respectively, α is the margin parameter, and Θ denotes the cosine distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$\Theta(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\| \mathbf{x}_i \| \| \mathbf{x}_j \|}. \quad (5)$$

Additionally, the symbol $+$ indicates the operator $max(\beta, 0)$, for $\beta = \Theta(f(\mathbf{x}_i^a), f(\mathbf{x}_i^p)) - \Theta(f(\mathbf{x}_i^a), f(\mathbf{x}_i^n)) + \alpha$, which imposes $L_\Theta \geq 0$ for every $f(\mathbf{x}_i^a), f(\mathbf{x}_i^p)$ and $f(\mathbf{x}_i^a), f(\mathbf{x}_i^n)$ pairs, since $max(\beta, 0) = 0, \forall \beta \in \mathbb{R} \mid \beta < 0$. This loss function attempts make the cosine distance between Anchor and Positive samples smaller than the distance between the Anchor and Negative instances by, at least, a margin of α. Alternatively, it will force examples of the same class to be mapped closer than examples of different classes (or even previously unknown examples).

We employ the TI3D with its default parameters and use hard and semi-hard triplet mining, as shown by [11]. Semi-hard triplets are defined as triplets in which the distance between the Anchor and Positive is smaller than the distance between the Anchor and Negative videos, but this distance is smaller than the margin parameter, i.e., $\Theta(f(\mathbf{x}^a), f(\mathbf{x}^p)) < \Theta(f(\mathbf{x}^a), f(\mathbf{x}^n)) < \Theta(f(\mathbf{x}^a), f(\mathbf{x}^p)) + \alpha$. Hard triplets are defined as triplets in which the distance between the Anchor and Positive is larger than the distance between the Anchor and Negative, i.e., $\Theta(f(\mathbf{x}^a), f(\mathbf{x}^p)) > \Theta(f(\mathbf{x}^a), f(\mathbf{x}^n))$. This triplet mining strategy ensures that only triplets with a positive loss w.r.t. Eq. 4 are used during training.

IV. METHODS

In this section, we present the OSVidCap framework for video captioning. It consists of five main modules: Target Detection and Localization (TDL), Features extraction, Open set module, Encoder, and Caption Generation. The overall architecture of OSVidCap is presented in Figure 1 and detailed as follows.

A. TDL MODULE

Detecting multiple concurrent events in a given video is essential to describe them in natural language adequately. The Target Detection and Localization (TDL) module consists of a mechanism designed to detect and track significant moving objects in a given video, which are considered the main concepts of the event. The output of this module consists of video segments for each moving object detected with a blurred background.

More specifically, the TDL module detects and tracks humans but is easily adaptable for other moving objects (such as animals and vehicles). We employ the Yolo-v4 [4] to detect humans and track them using the Deep SORT method [41]. The human-human or human-objects interaction is captured when they overlap in consecutive frames. In such cases, the entities are considered a single region of interest in the final video segment.

Finally, inspired by [42], we use a background blur method to guide the sentence generator module to focus on each region of interest in each video segment during the generation of the sentences.

B. FEATURES EXTRACTION

When human actions are described, it is important to consider details of the person, place, and action [43]. Thus, the Encoder module comprises four main classes of features extracted from a given input video as shown in Figure 1. All these features were extracted using off-the-shelf models, pre-trained on large datasets, which proved to be beneficial for video captioning tasks [20], detailed as follows:

- **Scene type features**: A sample of 16 evenly-spaced frames per video was used to extract the max-pooling features from the last convolutional layer using the VGG model pre-trained on the Places365 dataset.[1] The final representation is a 512-dimensional feature vector.
- **Spatial Features**: For extracting spatial features, we used the ResNet-101 model [15], pre-trained on the Imagenet dataset. From a sample of 16 equally spaced frames, we extracted a 2048-dimensional semantic feature vector of each frame from the last pooling layer. Then, an average pooling operation was performed, resulting in the final feature vector of dimension 2048.
- **Temporal Features**: The ResNeXt-101 with 3D convolutions [44], pre-trained on the Kinetics dataset [19], was used to extract a 2048-dimensional semantic feature

[1]Weights available at https://github.com/GKalliatakis/Keras-VGG16-places365.git

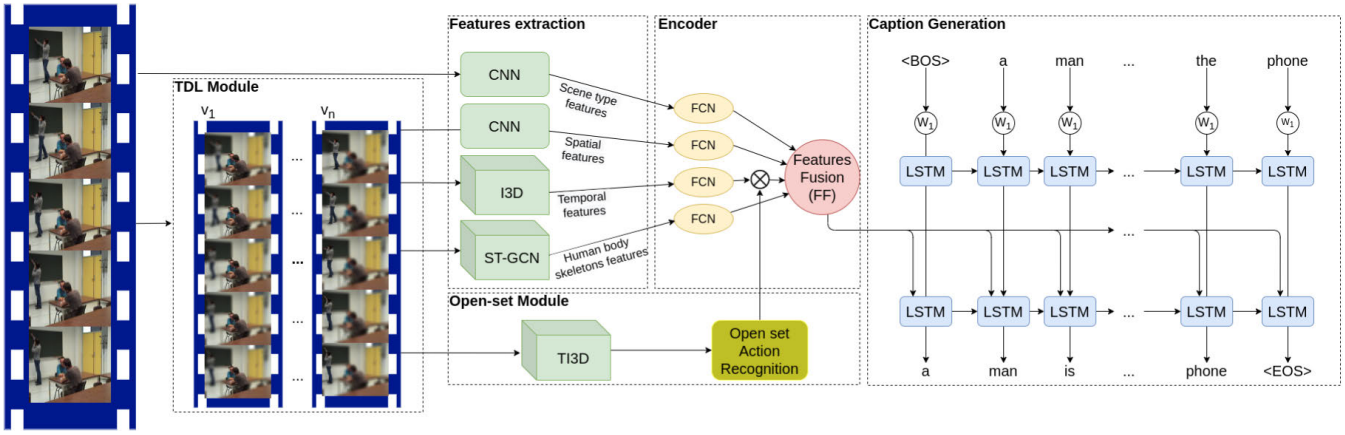


FIGURE 1. An overview of the OSVidCap framework.

vector for every 16 frames (with 50% of overlap). Then, followed an average pooling to obtain a final vector with 2048 features.

- **Human body skeleton features**: We used the ST-GCN model [45], pre-trained on the Kinetics dataset, to extract significant complementary information for the spatial and temporal features. This is a graph-based model for modeling dynamic skeletons extracted with the Openpose toolbox [46]. It is aimed to capture motion information in dynamic skeleton sequences. We performed a global max-pooling operation over all skeleton sequences to obtain a single 256-dimension feature vector for a given video. The combination of skeleton features with spatial and temporal features was intended to improve the performance in action recognition and, consequently, in the descriptions of the videos [47].

Except for the scene type features extracted from the original video frames, all other features were computed with the video segment processed by the TDL module. All these features are used in the encoder model to compute the feature final vector representation.

C. OPEN SET MODULE

The TI3D was initialized using the weights of the I3D and trained according to Section III-B. Then, it was used to extract features from both training and test videos. The features are used to train the EVM classifier, which predicts each action in the test set as known or unknown. The output of the module supports the caption generation by signalling whether the action belongs to a known or unknown class.

The TI3D was trained for 20 epochs, updating the triplets every epoch using the hard and semi-hard triplet mining strategy proposed by [11]. The learning rate was set to 0.02, the margin parameter to 0.2, and the batch size to 256. For the EVM, we set the tail size τ to 10% of the number of samples in the train set, the cover threshold for model reduction was set to 0.5, and the probability of inclusion (δ) to 0.5. These parameters were empirically set, based on previous experiments on the LIRIS dataset [48] used in this work.

D. ENCODER

This block aims to derive a feature vector representing the essential concepts to predict the next word for describing the ongoing action in the video. All the previous features extracted from the video were mapped into a common high-level abstract space by a feedforward network (FCN) with ReLU activations, as depicted in Figure 1.

Before Features Fusion (FF) step, we fuse the output processed by the Open Action Recognition Module with the processed Temporal Features (F_{tp}) to consider the unknown action information. Notice that the processed Place-type features (F_p), Spatial features (F_{sp}), and Human body skeleton features (F_{sk}) were remained to preserve essential information for caption generation, such as information about the place-type and number of people detected in the scene.

The output calculation of the encoder module provided by the FF can be formulated as follows:

$$F_p = \Phi(W_1 * U_p + b_1), \tag{6}$$
$$F_{sp} = \Phi(W_2 * U_{sp} + b_2), \tag{7}$$
$$F_{sk} = \Phi(W_3 * U_{sk} + b_3), \tag{8}$$
$$F_{tp} = \Phi(W_4 * U_{tp} + b_4) \otimes O_{uk}, \tag{9}$$
$$FF = F_p \odot F_{sp} \odot F_{sk} \odot F_{tp}, \tag{10}$$

in which W_1, W_2, W_3, and W_4 are weight matrices; U_p, U_{sp}, U_{sk}, and U_{tp} are features from the input modules: scene type, spatial, human body skeleton, and temporal, respectively; b_1, b_2, b_3, and b_4 are the bias vectors; Φ denotes the ReLU activation function; $\otimes$ denotes element-wise multiplication operator; $*$ is the convolution; $\odot$ is the concatenation operator; and O_{uk} denotes the feature vector provided by the TDL module.

E. CAPTION GENERATION

This module consists of the sentence generation and uses two Long Short-Term Memory (LSTM), a variant of Recurrent Neural Network (RNN), which works better with long-term dependencies. The first LSTM encodes the preceding

sequence of words $S = s_0, s_1, \ldots, s_{t-1}$. The second LSTM predicts the next word based on the output of the first LSTM combined with visual features computed by the Encoder module. The LSTM calculation formula used in this work is given by the following equations:

$$h_t = tanh(C_t) * o_t, \tag{11}$$
$$C_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t), \tag{12}$$
$$\tilde{C}_t = tanh(x_t U^g + h_{t-1} W^g), \tag{13}$$

in which U^g and W^g are weight matrices; x_t is the input at time t; h_{t-1} is the previous state; and f_t, i_t, and o_t are the forget, input and output gates, respectively. The calculations of unit gates are:

$$f_t = \sigma(x_t U^f + h_{t-1} W^f + b_f), \tag{14}$$
$$i_t = \sigma(x_t U^i + h_{t-1} W^i + b_i), \tag{15}$$
$$o_t = \sigma(x_t U^o + h_{t-1} W^o + b_o), \tag{16}$$

in which U^f, U^i, U^o, W^f, W^i, and W^o are weight matrices, b_f, b_i and b_o are bias vectors, and σ denotes the sigmoid activation function.

V. DATASET

There are a few datasets publicly available for video captioning task [3]. The most used datasets in the literature are MSVD [49] and MSR-VTT [50], containing a wide variety of open domain short videos. Each video has only a single main activity and multiple sentences with different details describing the video.

Despite the availability of annotated datasets for the video captioning task, none of them contain specific information about the action performed in each video, such as an action categorization. This information is essential in detecting and recognizing known and unknown events in an open-set scenario. Also, they do not contain concurrent events happening in the same video.

To overcome the above-mentioned limitations, we improved the LIRIS human activities dataset with captions and temporal annotations of new actions. Furthermore, we evaluate the generalization of our method on the large-scale ActivityNet Captions dataset. Both datasets are detailed as follows and are made available for further studies.[2]

A. LIRIS CAPTIONS DATASET

It was designed for recognizing complex and realistic actions in videos and made available for the ICPR-HARL'2012 competition. The full dataset contains 828 actions (including discussing, telephone calls, giving an item, etc.) performed by 21 different people in 10 different classes. Each action performed in a video contains spatial annotations in a bounding box and temporal information (the beginning and end of action). It was organized into two independent subsets: the D1 subset, with depth and grayscale images, and the D2 subset, with color images. The dataset also has unannotated actions, such as walking, running, whiteboard writing, book leafing, etc.

[2]http://labic.utfpr.edu.br/datasets/UTFPR-OSVidCAP.html

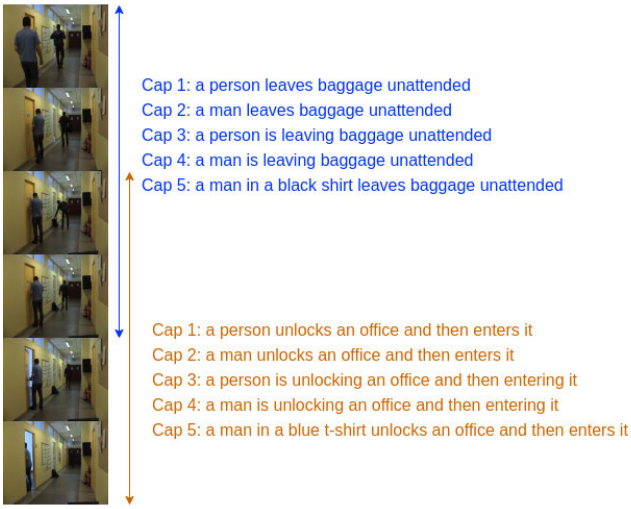


FIGURE 2. Example of a video clip and the ground-truth sentences created for each human activity in the LIRIS human activities dataset. Blue and Brown captions correspond to two different concurrent activities performed by different actors.

In this work we used the D2 subset that contains 367 annotated actions from 167 videos. Each action consists of one or more people performing one or more different activities. Besides, we extract 116 video segments in 15 different unannotated actions from the original videos to be used as unknown classes. Each new video segment was also annotated with spatial, temporal, and description information.

Reference [51] suggested that the number of reference sentences directly affects the accuracy of automated metrics. Also, those authors affirm that using five sentences models obtain a substantial boost in performance compared with only one sentence. Following this work, we improved the LIRIS human activity dataset with five different descriptions for each action, as shown in Figure 2.

B. ACTIVITYNET CAPTIONS DATASET

The ActivityNet Captions dataset [31] is a large dataset proposed for dense-captioning events, which involves both detecting and describing events in a video.

It contains 20,000 videos split into around 50%, 25%, 25% for training, validation, and testing set, respectively. All videos were taken from the ActivityNet Dataset [52], a benchmark for video classification and detection, which covers 200 classes of activities. The dataset also has an overlap of 10% of the temporal descriptions, thus indicating the presence of concurrent events. Each video is annotated with a series of temporally localized descriptions.

Although the ActivityNet Captions dataset is available for download as a collection of Youtube video links, many of these videos are no longer available for download, as reported in previous works [53], and only the pre-computed C3D

features provided by the authors are not helpful in our experiments. Thus, we used 12,714 videos that were still available for download. Videos shorter than 3 seconds were disregarded due to the small number of extracted frames. As our approach focused on describing entire videos and not detecting a series of events, we used the ground-truth event proposals to extract 34,934 video clips for each temporarily localized description provided in the annotations.

While ActivityNet Captions was originally designed for video dense captioning, we adapt it to our task by including action annotations to evaluate the generality of the proposed method in a large-scale dataset. Due to the considerable effort required to annotate each video clip manually, these annotations were collected from the ActivityNet dataset based on the video name, which is the same in both datasets. Each resulted action class contains, on average, 114 videos for training and 55 videos for testing. The action annotations were used to split videos into known and unknown classes for the detection of known and unknown actions.

VI. EXPERIMENTS

A. IMPLEMENTATION DETAILS

The proposed OSVidCap framework uses an encoder-decoder architecture. Therefore, both the encoder and caption generation modules (decoder) were trained in an end-to-end way. Before training, all captions were tokenized and converted to lowercase. Sparse words occurring less than three times in the training set were replaced with the unknown token. The *fasttext* [54] word embedding pre-trained on the Common Crawl Corpus was used to embed features into a 300-dimensional feature vector. It provides much more powerful and effective low-dimensional word representations for video captioning than other techniques such as sparse one-hot encoding vectors [55].

During the training step, a begin-of-sentence and end-of-sentence token were added to the sentence to deal with varying lengths. Also, an unknown tag was used to replace sparse words. We input the begin-of-sentence token into our Caption Generation Module to start the description generation process during the test step. Then, previously generated words are used as input to produce the following words until the max sentence length or the end-of-sentence token is achieved. In our experiments, the max sentence length was set as 19 and 25 for the Liris dataset and ActivityNet Captions dataset, respectively. Zero padding is applied if the sentence is shorter than the max number of words. The Beam Search method was employed to select the best sentence and avoid local optima. In our experiments, the beam size k was set to 3.

We empirically set the hidden state LSTM with 512 units and applied dropout with a rate of 0.5 on the input and output of the LSTM. The Adam algorithm, with a learning rate of 5×10^{-5} was used for optimization. The cross-entropy loss was used to train our model. All experiments were implemented using Tensorflow and Keras library.

To demonstrate the effectiveness of the proposed method, we have conducted two experiments to analyze the influence of the open set module and compare the video caption performance with related works.

1) EXPERIMENTS ON THE LIRIS DATASET

Due to the small number of videos and known actions in the Liris dataset, we performed a 5-fold cross-validation procedure to assess the OSVidCap performance. The same training and testing set of each cross-validation fold was used to train the open set module. In addition, to evaluate the effectiveness of the proposed approach in detecting unknown events, we include in the testing set 116 videos with unknown actions as described in Section V-A.

2) EXPERIMENTS ON THE ActivityNet CAPTIONS DATASET

The OSVidCap performance to generate captions of known events was performed using the standard data split.[3] Since this dataset was made available as a challenge, the test set was not provided with the ground truth. Thus, we follow the previous works [33], [53] and report the results on the validation set. The effectiveness of the proposed approach in detecting unknown events was performed using a 5-fold cross-validation procedure. Each fold contains known videos of 40 actions for the training and testing set, as explained in Section V-B. We also included in the testing set v_r random videos from other classes as unknown actions. The v_r was defined as the same number of videos presented in the training set to avoid imbalanced data.

B. EVALUATION METRICS

The captions generated by the proposed framework were evaluated according some metrics, frequently used in the area: BLEU [56], METEOR [57], ROUGE-L [58], and CIDEr [51]. All metrics were computed using the COCO-caption API [59].

BLEU is a metric based on n-grams precision modified and measures the predicted sentence proximity with one or more reference descriptions. Following most previous works for video captioning [3], we used four-grams with the BLEU metric, which is referred as BLEU-4. METEOR is based on the precision, recall, and harmonic mean and consists of creating an alignment between uni-grams from candidate and reference sentences. The word matching supports morphological variants including stemming and synonyms. CIDEr is a consensus-based metric and measures the similarity of a generated sentence against a majority of a set of ground-truth sentences. It employs morphological variations by changing each word in their stem (or root form) to resolve word-level correspondences. ROUGE-L computes the recall and precision scores using the longest common subsequences (LCS) technique and tends to reward long sentences with high recall. In our experiments, BLEU, METEOR, and ROUGE metrics were normalized to range from 0 to 100, with 100 as identical

[3] https://cs.stanford.edu/people/ranjaykrishna/densevid/

to the reference sentence. CIDEr ranges from 0 to 1000, with 1000 as identical to the reference.

C. QUANTITATIVE RESULTS

In this section, the performance evaluation of the proposed method is presented and compared with two recent existing approaches.

SGN [60] exploits the use of semantic groups based on meanings such as people, objects, or actions, rather than frame by frame for understanding a video. It is comprised of four main components: (i) a Visual Encoder component that aims to extract visual features from video frames; (ii) a Phrase Encoder which produces phrase representations from words by using the self-attention mechanism; (iii) a Semantic Grouping which employs a semantic aligner to align the video frames with phrases; and (iv) a Decoder based on LSTM with temporal attention.

Non-Autoregressive Coarse to-Fine (NACF) model [61] proposes a coarse-to-fine captioning procedure using a bi-directional self-attention-based network as caption generator. For improving caption quality, the decoder method is decomposed into two stages. First, a coarse-grained "template" is generated. Then, dedicated decoding algorithms generate fine-grained descriptions by filling in the generated "template" with suitable words and modifying inappropriate phrasing via iterative refinement.

For a fair comparison, all the methods utilize the ResNet-101 and ResNext-101 features as input, and the reported results were obtained using Microsoft COCO caption evaluation tool [59]. Furthermore, all approaches were set with the same maximum sentence length and minimum word frequency during training.

Table 1 presents a comparison performance of the OSVidCap with existing approaches on LIRIS dataset. It can be noticed that our model OSVidCap$_{(S+T)}$ achieved better performance in terms of Rouge-L and CIDEr and competitive performance in terms of Bleu and Meteor. Also, our model OSVidCap$_{(S+T+SK+P)}$ surpasses the compared approaches by 4.9% of BLEU-4, 5.1% of METEOR, 4.3% of ROUGE-L, and 9.3% of CIDEr. This suggests that our approach can better describe concurrent events in videos. In addition to spatial (S) and temporal (T) features, the model considered Human body skeleton (SK) extracted from human movements and Place-Type (P) features extracted from places. This points out that specialized features can be essential to better describe similar actions or actions according to the context (place). Such feature enrichment provides essential information to distinguish some actions, such as shaking hands and giving a small item to a second person. Also, the place type gives meaningful semantic information, as some actions tend to happen in specific places.

Table 2 presents the video captioning comparison on ActivityNet Captions dataset. It can be noticed that the proposed approach also achieved better or competitive results across all metrics, showing robust generalization to other contexts and scenarios. It is also noteworthy that the values of the metrics presented in Table 2 are significantly lower than those presented in Table 1 due to the complexity of the datasets, as reported in section V. The performance reported on this dataset is similar to those reported in recent literature [53], [62]. Note that, despite having used the same dataset to report the results, they are not comparable with the presented approach, as the videos and features used for training, validation, and testing are different.

TABLE 1. Comparison performance of video captioning on the LIRIS human activities dataset. 5-fold cross-validation results are presented in terms of BLEU-4 (B), METEOR (M), ROUGE-L(R), and CIDEr (C). S denotes spatial features. T denotes temporal features. SK denotes skeleton features. P denotes places features.

Model	B	M	R	C
NACF [61]	66.27	46.94	80.52	323.66
SGN [60]	62.08	44.38	76.95	298.06
OSVidCap $_{(S+T)}$	65.28	46.49	80.69	330.65
OSVidCap $_{(S+T+SK)}$	69.50	49.31	**84.05**	351.19
OSVidCap $_{(S+T+SK+P)}$	**69.54**	**49.34**	83.78	**354.04**

FIGURE 3. Example of events temporally localized in the video with independent start and end times, resulting in some events occurring concurrently in the ActivityNet Captions dataset.

TABLE 2. Video captioning performance on the ActivityNet captions validation set. Results are presented in terms of BLEU-4 (B), METEOR (M), ROUGE-L(R), and CIDEr (C). S denotes Spatial features, T denotes temporal features, SK denotes skeleton features, and P denotes places features.

Model	B	M	R	C
NACF [61]	2.20	8.16	20.44	23.78
SGN [60]	3.90	9.72	20.38	29.69
OSVidCap $_{(S+T)}$	4.19	9.71	20.98	28.54
OSVidCap $_{(S+T+SK)}$	4.30	9.96	21.26	**30.50**
OSVidCap $_{(S+T+SK+P)}$	**4.32**	**9.98**	**21.33**	29.84

TABLE 3. Open-set module on Liris captions dataset.

		F1-Score		Precision		Recall	
		Unknown	Known	Unknown	Known	Unknown	Known
1	89.0%	92.0%	85.0%	86.00%	100.0%	100.0%	74.0%
2	86.0%	90.0%	81.0%	84.0%	96.0%	98.0%	70.0%
3	83.0%	90.0%	77.0%	81.0%	100.0%	100.0%	63.0%
4	88.0%	92.0%	84.0%	85.0%	100.0%	100.0%	76.0%
5	85.0%	91.0%	80.0%	83.0%	100.0%	100.0%	67.0%
AVG	86.2%	91.0%	81.4%	83.8%	99.2%	99.6%	70.0%

TABLE 4. Open-set module on ActivityNet captions dataset.

		Average F1-Score	Average F1-Score		Average Precision		Average Recall	
			Unknown	Known	Unknown	Known	Unknown	Known
10	20	79.80%	79.20%	80.60%	84.10%	76.95%	75.15%	85.15%
20	10	77.10%	76.20%	78.10%	81.40%	74.00%	71.80%	82.90%
25	8	75.50%	75.38%	76.30%	78.63%	73.75%	69.00%	82.63%
40	5	73.60%	72.60%	74.60%	77.20%	71.00%	68.60%	79.00%
50	4	72.50%	71.25%	73.75%	75.50%	69.50%	67.50%	77.50%

In both datasets, the use of Place-type features did not show significant improvements. This may indicate that previously used features can also describe this visual information or are irrelevant for the video description task.

In Table 3, one can observe the evaluation performance of the open-set module in detecting known and unknown actions on the Liris Dataset. Results are presented in a 5-fold cross-validation procedure. The proposed method achieved satisfactory results in detecting known and unknown classes with an average F1-Score of 86.2%.

Table 4 shows the evaluation performance of the open-set module in detecting known and unknown actions on the ActivityNet Captions dataset. Five experiments with different numbers of the known classes in a cross-validation procedure were performed. The proposed method achieved satisfactory results in detecting known and unknown classes with an average F1-Score of 79.80% when ten classes were considered as known actions.

In Table 4, it can also be seen that the average precision of the unknown class is about 9% higher than the known class, and the average recall of the known class is 13% higher than the unknown class. This shows that the proposed approach achieves better results in detecting unknown classes than known classes. The automatic annotation process of video actions on the ActivityNet Captions dataset, as described in section V-B, also produced some annotation noises during the training and testing process. These noises can be a performed action with a different label or even a video without human actions. Figure 3 depicts an example of a video presented in the dataset. It can be observed that the video has different events with different start and end times. The automatic annotation process set the action class ''Removing ice from car'' to all video clips. However, in this example, only two video clips are related to the annotated action. Therefore, the degradation in the average precision metric of the known class may have been caused by the presence of these annotation noises. When considering new actions as known classes, the average F1-Score decreased due to the cumulative annotation errors provided by the automatized annotation process, as reported below.

Table 5 reports the impact of the open-set component on the video descriptions generated by the proposed approach. The results reported in the Liris dataset used the same data in a cross-validation procedure, as used in Table 3. For reporting the results on ActivityNet Captions Dataset, we used the 5-fold cross-validation applied in Table 4.

These results are significantly higher when compared with those reported in Tables 1 and 2 because, in this experiment, we considered videos in the test set with unknown activities. For these videos, the model is supposed to generate descriptions such as ''a person is performing an unknown action''.

The experiments with unknown actions in the testing set suggested that Place-type features did not lead to a significant improvement. However, these features are important to understand scenes in which the information about the place type is relevant, for example, to describe whether the person is entering or leaving an office or writing a whiteboard in a classroom. In the testing set used to report the experiments in Table 5, several videos from unknown classes were included to evaluate the proposed open set module. Therefore, the overall influence of the Place-type features has quantitatively decreased due to the small number of sentences that require such features. To the best of our knowledge, this is the first work to address the video captioning task in an open-set world by generating captions of known events present in the training set and dealing with unknown events not previously seen.

D. QUALITATIVE RESULTS

In Figure 4, we illustrate three examples of video descriptions generated by the baselines method SGN and NACF and the proposed OSVidCap. Figure 4a depicts a scene with two sequential actions. First, a man in a striped t-shirt talks to

TABLE 5. Influence of the open set module in the OSVidCap approach. S denotes spatial features. T denotes temporal features. SK denotes Skeleton features. P denotes places features.

Dataset	Model	B	M	R	C	F1-score
Liris dataset	OSVidCap (S+T)	77.8	56.1	87.9	381.1	86.2%
	OSVidCap (S+T+SK)	**80.9**	**57.3**	**89.1**	385.8	
	OSVidCap (S+T+SK+P)	80.2	57.2	**89.1**	**385.9**	
ActivityNet Captions dataset	OSVidCap (S+T)	16.2	16.8	38.7	68.0	73.6%
	OSVidCap (S+T+SK)	15.6	16.8	39.1	70.0	
	OSVidCap (S+T+SK+P)	**16.5**	**16.9**	**39.2**	**72.9**	

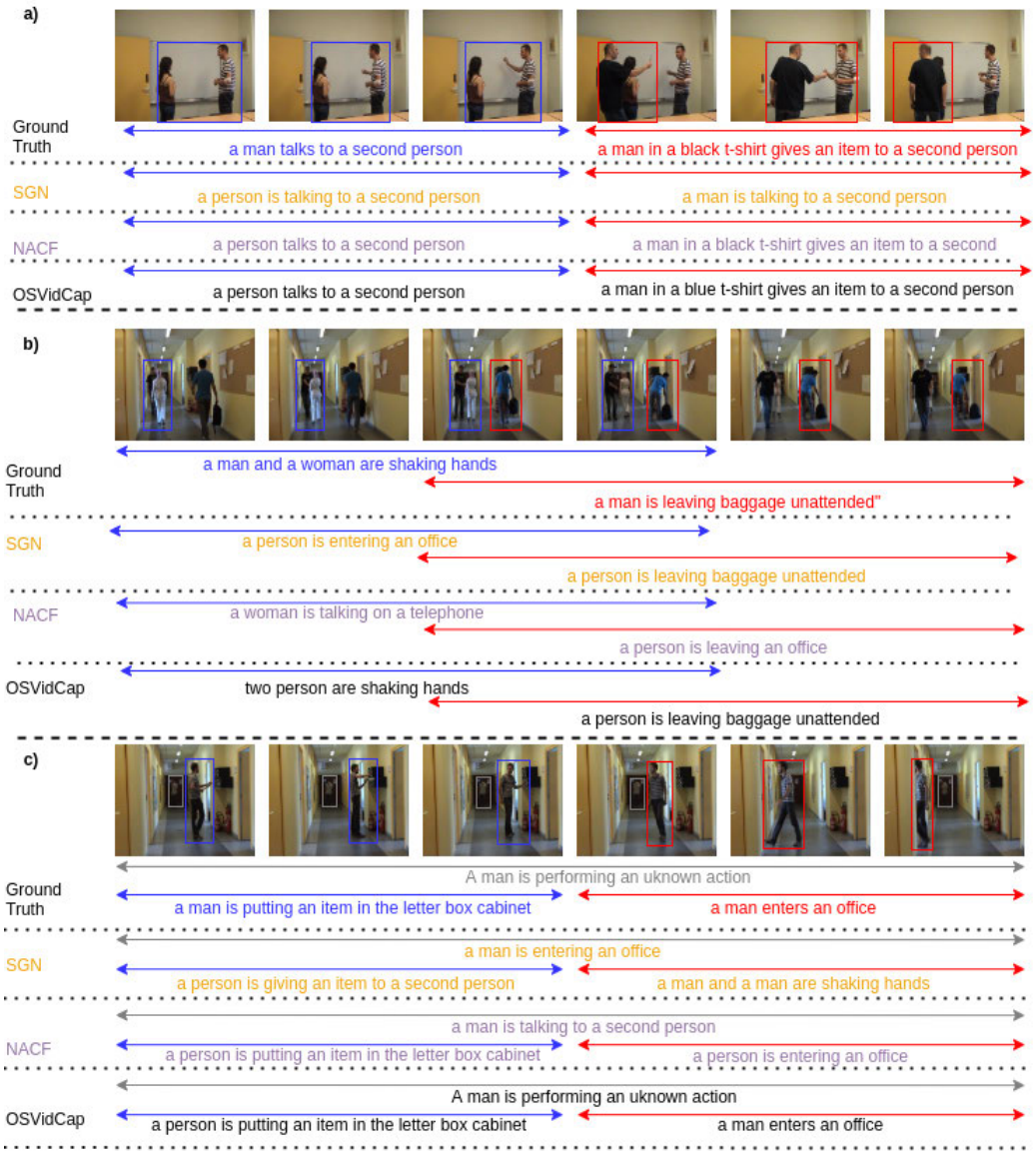


FIGURE 4. Qualitative example of generated descriptions on the Liris dataset.

a woman in front of a whiteboard. Then, another man in a black t-shirt enters the room and gives an item to the man in a striped t-shirt. Figure 4b shows two concurrent events. While a man and a woman are handshaking, another man is leaving baggage unattended. Finally, in Figure 4c, three events take place in the video. At the same time, a man is performing an unknown action. Another man leaves an item in the letterbox cabinet and then enters the room.

For the examples of Figure 4, our approach described concurrent actions better than the baselines. In Figure 4a, the OSVidCap correctly described the ongoing action but wrongly represented the color of the t-shirt, suggesting that

the model did not learn this information from the input features. Possibly, more specific features should fix this issue.

In Figure 4b, we can observe that the compared approaches could not detect the shake hands action, suggesting the importance of using human body features in describing human action videos. Also, they fail to detect and describe concurrent actions in videos.

We can realize the importance of the open set module in the situation considered in Figure 4c). While the OSVidCap detected an unknown action performed by a man and correctly described it as such, the compared approaches generated a wrong description. It is worth highlighting that this action was previously labeled as unknown and did not appear in the training set.

VII. CONCLUSION AND FUTURE WORKS

The majority of artificial intelligence methods rely on the closed-set world assumption. The same holds for the specific case of automatic video captioning systems. Existing methods based on a closed-set world can adequately describe only the temporal events previously seen during the training step. Unless they are trained with all existing events and actions of interest, they will not be able to recognize unknown events found in videos in the wild. Furthermore, most current approaches for video description focus only in single actions occurring at a time, while in the real-world, concurrent events may take place. To address the above-mentioned issues, in this paper we proposed the OSVidCap framework, that can detect and describe concurrent events in an open-set world scenario. From a given input video, the TDL module detects and tracks humans and outputs a set of video segments to be described. Then, spatial and temporal features are extracted from each video segment. Also, the open set module, built upon the TI3D metric learning approach coupled with an extreme value machine (EVM), classifies each detected action as a known or unknown class. Then, the Encoder module computes the features and generate a fixed-length vector that represents the whole video content. Finally, the caption generation module, based on the LSTM, generates the descriptions in a human-comprehensible form.

Experimental results demonstrate the effectiveness of the framework in describing concurrent events in a given video. Also, the open-set module allows the framework to describe unknown events. Our experiments also show that different features such as the Human body skeleton and Place-type features are quite relevant to understand fine-grained actions, frequently performed in specific environments. Such features enrichment provides a better video representation for generating a more detailed description. Furthermore, due to the lack of specific datasets for evaluating concurrent events in an open-set scenario, we have contributed new annotations of unknown actions in the LIRIS human activity dataset that can be used as a benchmark for the proposed task.

Despite the excellent results achieved by OSVidCap, we observed that it could provide a more detailed description of people, for instance, including the type and color of the clothes. This enrichment of details can plays an important role for applications in surveillance. The TDL module is capable of capturing individual humans or objects of interest and simple interactions between them, by capturing the overlapping region among objects. However, the proposed module may fail to capture more complex human interactions.

Therefore, in future work, the proposed framework will be evolved by enriching the description of people in the scene, as well as to improve the detection of events involving persons that interact at a distance, such as watching TV or throwing an object to another person. Another future work involves providing a human evaluation over a subset of testing data, as existing metrics used for automatic evaluation of video captioning may not properly correlate with human judgment.

ACKNOWLEDGMENT

The authors would like to thank NVIDIA Corporation for the donation of the Titan-Xp GPUs used in this work.

REFERENCES

[1] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video," *IEEE Trans. Syst., Man, C, Appl. Rev.*, vol. 39, no. 5, pp. 489–504, Sep. 2009.

[2] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Integrated semantic-syntactic video modeling for search and browsing," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 839–851, Dec. 2004.

[3] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Comput. Surv.*, vol. 52, no. 6, pp. 115:1–115:37, 2019.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*. [Online]. Available: http://arxiv.org/abs/2004.10934

[5] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[6] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, pp. 1005:1–1005:20, 2019.

[7] R. Sobue, M. Nakazawa, Y. Chae, B. Stenger, T. Yamashita, and H. Fujiyoshi, "Cooking video summarization guided by matching with step-by-step recipe photos," in *Proc. 16th Int. Conf. Mach. Vis. Appl. (MVA)*, May 2019, pp. 1–6.

[8] A. De Souza Inacio and H. S. Lopes, "EPYNET: Efficient pyramidal network for clothing segmentation," *IEEE Access*, vol. 8, pp. 187882–187892, 2020.

[9] I. Budvytis, M. Teichmann, T. Vojir, and R. Cipolla, "Large scale joint semantic re-localisation and scene understanding via globally unique instance coordinate regression," in *Proc. 30th Brit. Mach. Vis. Conf.*, 2019, pp. 1–13.

[10] C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021.

[11] M. Gutoski, A. E. Lazzaretti, and H. S. Lopes, "Deep metric learning for open-set human action recognition in videos," *Neural Comput. Appl.*, vol. 33, no. 4, pp. 1207–1220, Feb. 2021.

[12] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 762–768, Mar. 2018.

[13] M. Nabati and A. Behrad, "Video captioning using boosted and parallel long short-term memory networks," *Comput. Vis. Image Understand.*, vol. 190, Jan. 2020, Art. no. 102840.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. IEEE 6th Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[17] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[19] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: http://arxiv.org/abs/1705.06950

[20] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics*, 2015, pp. 1494–1504.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, vol. 1, 2012, pp. 1097–1105.

[22] A. Liu, Y. Qiu, Y. Wong, Y. Su, and M. Kankanhalli, "A fine-grained spatial-temporal attention model for video captioning," *IEEE Access*, vol. 6, pp. 68463–68471, 2018.

[23] H. Xiao and J. Shi, "Video captioning with adaptive attention and mixed loss optimization," *IEEE Access*, vol. 7, pp. 135757–135769, 2019.

[24] Y. Tu, X. Zhang, B. Liu, and C. Yan, "Video description with spatial-temporal attention," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1014–1022.

[25] D. Francis and B. Huet, "L-STAP: Learned spatio-temporal adaptive pooling for video captioning," in *Proc. 1st Int. Workshop AI Smart TV Content Prod., Access Del. (AI4TV)*, 2019, pp. 33–41.

[26] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 229–241, Feb. 2020.

[27] S. Chen, Q. Jin, J. Chen, and A. G. Hauptmann, "Generating video descriptions with latent topic guidance," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2407–2418, Sep. 2019.

[28] Y. Zhu, J. Xie, Z. Tang, X. Peng, and A. Elgammal, "Semantic-guided multi-attention localization for zero-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14943–14953.

[29] A.-A. Liu, N. Xu, Y. Wong, J. Li, Y.-T. Su, and M. S. Kankanhalli, "Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language," *Comput. Vis. Image Understand.*, vol. 163, pp. 113–125, Oct. 2017.

[30] L. Sun, B. Li, C. Yuan, Z. Zha, and W. Hu, "Multimodal semantic attention network for video captioning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1300–1305.

[31] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 706–715.

[32] V. Escorcia, F. Caba Heilbron, J. C. Niebles, and B. Ghanem, "DAPs: Deep action proposals for action understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 768–784.

[33] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8739–8748.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[35] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7590–7598.

[36] T. Chen, Q. Zhao, and J. Song, "Boundary detector encoder and decoder with soft attention for video captioning," in *Proc. Asia–Pacific Web (APWeb) Web-Age Inf. Manage. (WAIM) Joint Int. Conf. Web Big Data*. Cham, Switzerland: Springer, 2019, pp. 105–115.

[37] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–37, Feb. 2019.

[38] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–10.

[39] Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu, and Y. Yang, "Cascaded revision network for novel object captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3413–3421, Oct. 2020.

[40] X. Wang, J. Wu, D. Zhang, Y. Su, and W. Y. Wang, "Learning to compose topic-aware mixture of experts for zero-shot video captioning," in *Proc. of the AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8965–8972.

[41] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[42] J.-K. Tsai, C.-C. Hsu, W.-Y. Wang, and S.-K. Huang, "Deep learning-based real-time multiple-person action recognition system," *Sensors*, vol. 20, no. 17, p. 4758, Aug. 2020.

[43] Y. Shigeto, Y. Yoshikawa, J. Lin, and A. Takeuchi, "Video caption dataset for describing human actions in Japanese," in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, May 2020, pp. 4664–4670.

[44] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[45] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[46] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[47] J. Li, X. Xie, Q. Pan, Y. Cao, Z. Zhao, and G. Shi, "SGM-Net: Skeleton-guided multimodal network for action recognition," *Pattern Recognit.*, vol. 104, Aug. 2020, Art. no. 107356.

[48] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, and B. Sankur, "Evaluation of video activity localizations integrating quality and quantity measurements," *Comput. Vis. Image Understand.*, vol. 127, pp. 14–30, Oct. 2014.

[49] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics*, 2011, pp. 190–200.

[50] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.

[51] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[52] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.

[53] V. Lashin and E. Rahtu, "Multi-modal dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 4117–4126.

[54] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[55] N. Aafaq, N. Akhtar, W. Liu, and A. Mian, "Empirical autopsy of deep video captioning encoder-decoder architecture," *Array*, vol. 9, Mar. 2021, Art. no. 100052.

[56] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

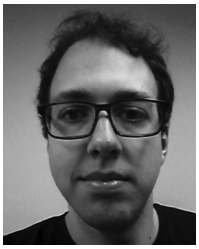
[57] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*, 2007, pp. 228–231.

[58] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[59] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*. [Online]. Available: http://arxiv.org/abs/1504.00325

[60] H. Ryu, S. Kang, H. Kang, and C. D. Yoo, "Semantic grouping network for video captioning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2514–2522.
[61] B. Yang, Y. Zou, F. Liu, and C. Zhang, "Non-autoregressive coarse-to-fine video captioning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3119–3127.
[62] C. Deng, S. Chen, D. Chen, Y. He, and Q. Wu, "Sketch, ground, and refine: Top-down dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 234–243.

ANDREI DE SOUZA INÁCIO received the B.Sc. and M.Sc. degrees in computer science from the Federal University of Santa Catarina (UFSC), in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Federal University of Technology—Paraná, Paraná, Brazil. Since 2014, he has been a Lecturer at the Federal Institute of Santa Catarina (IFSC). He has professional experiences in information systems design, web development, and IT project management. His research interests include but not limited to computer vision, machine learning, and data mining.



MATHEUS GUTOSKI received the bachelor's degree in information technology from Santa Catarina State University, Brazil, in 2015, and the M.Sc. degree in computer engineering from the Federal University of Technology—Paraná (UTFPR), Brazil, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include deep learning, pattern recognition, and computer vision.

ANDRÉ EUGÊNIO LAZZARETTI (Member, IEEE) received the B.Sc., M.Sc., and D.Sc. degrees in electrical engineering from the Federal University of Technology—Paraná, in 2007, 2010, and 2015, respectively. He is currently a Professor with the Department of Electronics, Federal University of Technology—Paraná. His research interests include machine learning, computer vision, and digital signal processing.

HEITOR SILVÉRIO LOPES received the B.Sc. and M.Sc. degrees in electrical engineering from the Federal University of Technology—Paraná (UTFPR), Curitiba, in 1984 and 1990, respectively, and the Ph.D. degree from the Federal University of Santa Catarina, in 1996. In 2014, he spent a sabbatical year at the Department of Electrical Engineering and Computer Science, University of Tennessee, USA. Since 2003, he has been a Research Fellow in the area of computer science at the Brazilian National Research Council. He is currently a tenured Full Professor with the Department of Electronics and the Graduate Program on Electrical Engineering and Applied Computer Science (CPGEI), UTFPR. He is the Founder and the current Head of the Bioinformatics and Computational Intelligence Laboratory (LABIC). His major research interests include computer vision, deep learning, evolutionary computation, and data mining.