Multi-stage RGB-based Transfer Learning Pipeline for Hand Activity Recognition

Yasser Boutaleb[1,2], Catherine Soladie[2], Nam-Duong Duong[1], Jérôme Royan[1] and Renaud Seguier[2]
[1]*IRT b-com, 1219 Avenue des Champs Blancs, 35510 Cesson-Sevigné, France*
[2]*IETR/CentraleSupelec, Avenue de la Boulaie, 35510 Cesson-Sevigné, France*

Keywords: First-person Hand Activity Recognition, Transfer Learning, Multi-stream Learning, Features Fusion.

Abstract: First-person hand activity recognition is a challenging task, especially when not enough data are available. In this paper, we tackle this challenge by proposing a new low-cost multi-stage learning pipeline for first-person RGB-based hand activity recognition on a limited amount of data. For a given RGB image activity sequence, in the first stage, the regions of interest are extracted using a pre-trained neural network (NN). Then, in the second stage, high-level spatial features are extracted using pre-trained deep NN. In the third stage, the temporal dependencies are learned. Finally, in the last stage, a hand activity sequence classifier is learned, using a post-fusion strategy, which is applied to the previously learned temporal dependencies. The experiments evaluated on two real-world data sets shows that our pipeline achieves the state-of-the-art. Moreover, it shows that the proposed pipeline achieves good results on limited data.

1 INTRODUCTION

Understanding first-person hand activity is a challenging problem in computer vision, that has attracted much attention due to its wide research and practical applications, such as Human-Computer Interaction (Sridhar et al., 2015), Humanoid Robotics (Ramirez-Amaro et al., 2017), Virtual/Augmented Reality (Surie et al., 2007), and Multi-media for automated video analysis (Bambach, 2015).

Recent advances in embedded technologies, such as wearable cameras which provide low-cost data such as RGB image sequences, have allowed more widespread machine-learning-based egocentric activity recognition (EAR) methods (Tadesse and Cavallaro, 2018). In addition to its low-cost, RGB image sequences take into consideration both appearance and motion information unlike depth maps or 3D skeletal data which focus more on the motion. Yet, the majority of egocentric activities are centered around hand-object interactions and appearance is highly important to perform inter-objects and inter-scenarios differentiation.

To this end, many RGB-based approaches have been proposed. Most of them are based on end-to-end Deep Learning (DL) (Kondratyuk et al., 2021) which has been proven to be effective when a large amount of data is available. However, for some industrial applications, providing large-scale labeled data is still hard and expansive to achieve due to the manual data annotation process. On the other hand, recent advances in DL benefit greatly from problems such as image classification (He et al., 2016; Xie et al., 2017) and object detection (Liu et al., 2016; He et al., 2020) which can be exploited as an alternative to overcome the data scarcity in EAR problems e.g. transfer learning (TL) techniques.

A particular branch of DL approaches focused on observation and exploration of spatial attention

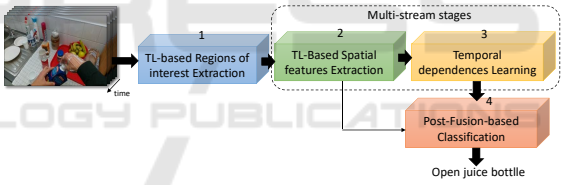

Figure 1: Our proposed learning pipeline for RGB-based first-person hand activity recognition. For a given RGB images activity sequence, in the first stage, the regions of interest are extracted using a pre-trained NN. Then, in the second stage, high-level spatial features are extracted using pre-trained deep NN. Sequentially, in the third stage, the temporal dependencies are learned. In the last stage, a hand activity sequence classifier is learned, using a post-fusion strategy, which is applied to the previously learned temporal dependencies.

Boutaleb, Y., Soladie, C., Duong, N., Royan, J. and Seguier, R.
Multi-stage RGB-based Transfer Learning Pipeline for Hand Activity Recognition.
DOI: 10.5220/0010856200003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 839-848
ISBN: 978-989-758-555-5; ISSN: 2184-4321

through deep neural networks (NNs) to recognize activities based on visual information (Sudhakaran et al., 2019; Sudhakaran and Lanz, 2018). However, the learned spatial attention is not fully confident, since it is learned in an unsupervised manner while training a supervised EAR model. This has led some researchers to supervise spatial attention learning by using Gaze information (Min and Corso, 2020) or by manually annotating the data (Ma et al., 2016) which is more expensive. In all cases, this has confirmed that, in first-person hand activity recognition problems, the visual points of interest are concentrated around the hands and manipulated objects. This relevant information can be used to design more robust EAR algorithms.

Motivated by all these observations, we introduce in this paper, a new learning pipeline for RGB-based first-person hand activity recognition, that aims at overcoming the data scarcity problem while ensuring a low-cost good and accurate recognition. It is a novel four-stage learning pipeline, such as each stage is described as follows: (1) Regions of Interest Extraction (RoIE). Unlike existing methods that use DL-based visual attention and require a large amount of data, we propose to directly use the right and left hands as pertinent regions of interest that give information about manipulated objects and actions being performed. These regions of interest are extracted using a TL technique. Our experiments showed that this information is the key to first-person hand activity recognition. In order to robustify the recognition model, we propose a data augmentation process, which is specifically adapted to these regions of interest. (2) Spatial Features Extraction (SFE). Here, we also use TL instead of end-to-end DL methods. This stage exploits the visual information of the resulted regions of interest from the previous stage. Adapting TL for RoIE and SFE allows learning with a limited number of training samples while providing a good accuracy score. Furthermore, it decreases the training cost, since the transferred NN are already pre-trained. (3) Temporal Dependencies Learning (TDL). For each extracted deep visual descriptor (right and left) resulting from the previous stage, we learn the temporal dependencies in a multi-stream manner (Boutaleb et al., 2021) which also avoids the over-fitting problem. (4) Post-Fusion classifier (PFC). This last stage is a classifier that learns activity classes (Boutaleb et al., 2021).

The remainder of this paper is organized as follows. After giving a review on the related work in Section 2, we describe our proposed pipeline for RGB-based hand activity recognition in Section 3. Then, we show the benefits of the proposed approach by presenting and discussing the experimental results in Section 4. Section 5 concludes the paper.

2 RELATED WORK

First-Person hand activity recognition using visual data that provide motion and appearance information has attracted a lot of attention over the last few years.

Aiming at exploiting the motion information, many approaches make use of optical flow as the main source of motion features (Tadesse and Cavallaro, 2018). Optical flow can be obtained using direct motion estimation techniques (Irani and Anandan, 1999) to achieve frames/sub-frames sub-pixel accuracy resulting in a dense representation. Yet, this representation has a high-computational cost and suffers from redundancy. This has lead (Abebe et al., 2016; Poleg et al., 2014) to use grid (spars) representation of the optical flow. Sparse optical flow gains in computational cost. However, it suffers from an information leak and have limited discriminative capabilities as specific motion characteristics (e.g. magnitude) are not exploited (Tadesse and Cavallaro, 2018).

In order to exploit the appearance information, many works traditionally used local visual features such as HOF (Laptev et al., 2008), MBH (Wang et al., 2012), 3D SIFT (Scovanner et al., 2007), HOG3D (Kläser et al., 2008), and extended SURF (Willems et al., 2008) to encode appearance information, so that it can be used as feature descriptors to recognize activities. On the other hand, DL NNs have been successful in learning high-level appearance features for image classification (Rawat and Wang, 2017). This has attracted a lot of interest in the EAR area (Karpathy et al., 2014; Tran et al., 2015; Ji et al., 2013; Taylor et al., 2010). Recently, (Singh et al., 2016), proposed a two-stream DL architecture, 2D and 3D CNNs fed by egocentric cues (hand Mask, head Motion, and saliency Map). The two-streams networks are followed by class score fusion strategy to classify activities. To make use of the temporal dimension, they added a temporal stream that uses stacked optical flow as input to capture motion information. However, these egocentric cues are not always available. Similarly, (Ma et al., 2016) proposed a two-stream architecture: an appearance stream for object classification task by applying hand segmentation and object location; and a motion stream for action classification using optical flow. Finally, the activity class label is given by the concatenation of the action and the object class labels. Therefore, a heavy manual data annotation was necessary for object region localization and hand segmentation. Moreover, a single RGB image is

used for encoding appearance without considering the temporal ordering. As an alternative to optical-flow-based motion information, which is also interpreted as temporal dependencies features, (Ryoo et al., 2015) extracted features from a series of frames to perform temporal pooling with different operations, including max pooling, sum pooling, or histogram of gradients. Then, a temporal pyramid structure allows the encoding of both long-term and short-term characteristics. However, these methods do not take into consideration the temporal order of the activity sequence frames.

Furthermore, to better exploit information in the temporal dimension, many other works focused on Recurrent Neural networks (RNNs) equipped with Long Short Term Memory (LSTMs) cells (Cao et al., 2017; Verma et al., 2018), and Convolutional Long Short-Term Memory (ConvLSTM) (Sudhakaran and Lanz, 2017; Sudhakaran and Lanz, 2018) for their capabilities of reasoning along the temporal dimension to learn the temporal dependencies in the respect of temporal order. This has motivated (Sudhakaran et al., 2019) to propose an customized LSTM unit in order to learn visual attention along the activity sequence jointly with the temporal dependencies. However, attention-based methods still have some limitations as we mentioned in section 1. In contrast, we propose to directly extract the regions of interest and their associated spatial features, then we learn the temporal dependencies in a multi-stream manner.

3 PROPOSED METHOD

This section details our proposed pipeline following the illustration of Figure 1. In the first stage, we extract the regions of interest (Sec 3.1). Then, in the second stage, we extract the spatial features (Sec 3.2). In the third stage, we learn the temporal dependencies (Sec 3.3). Once the temporal learning is ended, in the last stage, we transfer and exploit the knowledge from the previous stage to learn to classify activities (Sec 3.4).

3.1 TL-based Regions of Interest Extraction (RoIE) and Data Augmentation

Our pipeline uses as unique input a sequence of images (frames) representing a first-person hand activity, that we denote by $S = \{I_1, I_2, .., I_T\}$, where I_t is an image frame at time-step t and T the sequence max length.

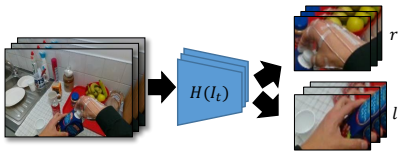

Figure 2: The first stage of the pipeline: TL-based Regions of interest Extraction (RoIE). Each image frame I_t is fed into a pre-trained NN $H(I_t)$ resulting in two hand region sequences l and r that refer to the left and the right-hand regions respectively.

As we mentioned in section 1, the main focus on the first-person hand activity is centred around the hands and manipulated objects. To this end, we propose to directly extract and use the left and the right hand regions as regions of interest. Let denoting $H(I_t) = \{h_t^{left}, h_t^{right}\}$ where $H(.)$ is the pre-trained NN that takes an image frame I_t as an input and outputs two sub-images h_t^{left} and h_t^{right} that refers to the left and the right hand respectively. So, by applying this to all image frames, the activity sequence will be reformulated by two sequences l and r that belong to the left and right hand respectively, such as:

$$l = \{h_t^{left}\}_{t=1:T} \text{ and } r = \{h_t^{right}\}_{t=1:T} \quad (1)$$

Figure 2illustrates the hand region extraction process. The proposed regions of interest characterize the hand activity sequence in a relevant way since the visual information from the hands contains information about the type of grasp and the shape of objects being manipulated (noun) e.g. "Juice bottle". Moreover, passing this information through the time dimension allows retrieving relevant information about the performed action (verb) e.g. "Open". In Section 4.4, we quantitatively show the efficiency of the proposed regions of interest. On the other hand, unlike visual attention methods based on end-to-end NN (Sudhakaran et al., 2019), using TL to extract the regions of interest helps to avoid the over-fitting problem and allows training with a limited number of samples while ensuring a good accuracy score. In section4.2, we give details about the adopted pre-trained NN.

In daily/industrial hand activities, one of the two hands, left or right, can be dominant. It depends on whether the participant is right- or left-handed. This may cause an imbalance in the training data-set and make the model less generalizable. To this end, we proposed an adapted data augmentation process in order to balance the training data-set. It is applied to the RoIE stage's outputs. If only one hand is detected for e.g. left hand, we augment the extracted sub-image of the right-hand h_t^{right} with the mirror effect of the de-

tected left-hand h_t^{left}. The figure 3illustrates the data augmentation process.

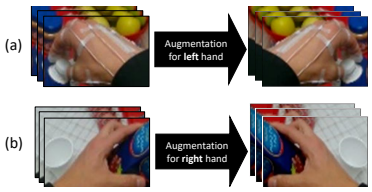

Figure 3: Illustration of our data augmentation process. (a) the mirror effect of extracted right-hand sub-images h_t^{right} are used as augmentation for those of the left hand. (b) the mirror effect of extracted left-hand sub-images h_t^{left} is used as augmentation for those of the right hand.

In section 4.5, we show quantitatively the effectiveness of this proposed data augmentation process.

3.2 TL-based Spatial Features Extraction (SFE)

One of the problems where deep learning excels is image classification (Xie et al., 2017). The goal in image classification is to classify a specific picture according to a set of possible categories by deeply exploring and learning the spatial information. This motivated us to uses a pre-trained NN classifier to extract learned spatial features from the sub-images Eq.1 resulted from the previous stage.

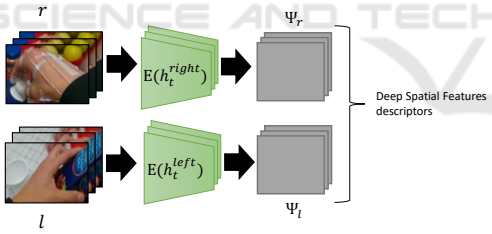

Figure 4: The second stage of the pipeline: TL-Based Spatial features Extraction (SFE). Each extracted sub-image $h_t^{left} \in l$ and $h_t^{right} \in r$ is fed into a pre-trained NN $E(.)$. This stage results two deep spatial feature descriptor sequences Ψ_l and Ψ_r for right and left hand respectively.

We denote by $E(.)$ this pre-trained NN. And we formulate the spatial feature descriptor sequences by Ψ_l and Ψ_r referring to the left and the right hands regions as follow:

$$\Psi_l = \{E(h_t^{left})\}_{t=1:T} \text{ and } \Psi_r = \{E(h_t^{right})\}_{t=1:T} \quad (2)$$

This stage allows to exploit the hands visual information resulted from the previous stage. Using a sophisticated pre-trained NN reduces the dimension while keeping a pertinent high-level spatial features. Adding to that all TL benefits, it decreases the learning cost and avoids the over-fitting problem while learning on a limited number of training samples. In section 4.2, we gives details about the adopted pre-trained NN.

3.3 Temporal Dependencies Learning (TDL)

Learning long and complex activities requires considering the temporal dimension to make use of the long-term dependencies between sequence time-steps. As we do not have a learned NN for this very specific task, we train a LSTM-based NN for its great success and capabilities to learn these long/short term dependencies. Moreover, in contrast to traditional RNNs, LSTMs overcome the vanishing gradient problem by using a specific circuit of gates (Hochreiter and Schmidhuber, 1997).

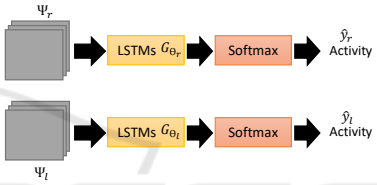

Figure 5: The third stage of the pipeline: Temporal dependencies Learning. For each feature descriptor sequence Ψ_l and Ψ_r a NN composed of stacked LSTM layers followed by softmax layer are trained independently to learn temporal dependencies by classifying activities.

(Avola et al., 2019; Liu et al., 2019) concatenate different types of feature spaces as one input vector, which may complicate the input and confuse the NN. In contrast, similarly to (Boutaleb et al., 2021), for each spatial feature descriptors Ψ_l and Ψ_r (seen in Sec 3.2), we train separately a simple NN that consists of staked LSTM layers followed by a softmax layer to classify activities. Therefore, in total, we train two NN separately as shown in Figure 5.

More formally, for each descriptor sequence Ψ_l and Ψ_r, we model the temporal dependencies with a composite function $G_{\theta_l}(\Psi_l)$ and $G_{\theta_r}(\Psi_r)$ respectively, where $G_{\theta_.}(\cdot)$ is a LSTM network with θ_l and θ_r learnable parameters, while the output of $G_{\theta_.}(\cdot)$ refers to the last hidden state of the last LSTM unit. For each network we define a cross entropy loss functions $\mathcal{L}_l$ and $\mathcal{L}_r$ as follows:

$$\mathcal{L}_l = -\sum_{c=1}^{N} y^c log(\hat{y}_l^c) \text{ and } \mathcal{L}_r = -\sum_{c=1}^{N} y^c log(\hat{y}_r^c) \quad (3)$$

where N is the number of classes and y^c the target label. The $\hat{y}_l^c$ and $\hat{y}_r^c$ are the softmax outputs that refers to the predicted label using left and right hand

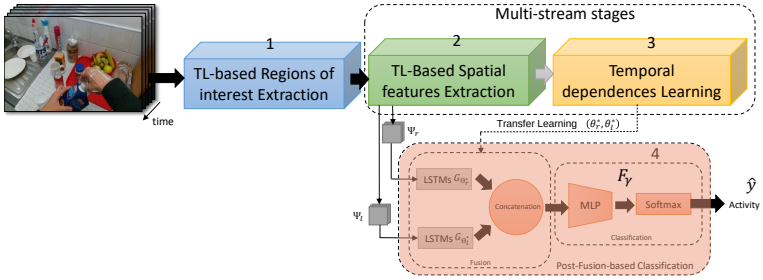

Figure 6: The fourth stage of the pipeline: Post-Fusion-based Classification. Once the temporal dependencies are learned in the third stage. The LSTM layers are transferred to the fourth stage with fixed parameters θ_l^* and θ_r^*. Their outputs are concatenated and fed into a MLP+softmax for the final classification.

descriptor sequence respectively. The temporal learning parameters are optimized by minimizing over a labeled data set:

$$\theta_l^* = \arg\min_{\theta_l} \mathcal{L}_l(y, \hat{y}_l) \text{ and } \theta_r^* = \arg\min_{\theta_r} \mathcal{L}_r(y, \hat{y}_r) \quad (4)$$

At the end of the pre-training, as a result, we have a set of two trained stacked LSTM layers, with optimised parameters θ_l^* and θ_r^*:

$$G_{\theta_l^*}(\Psi_l), G_{\theta_r^*}(\Psi_r) \quad (5)$$

We note that the purpose of this third stage is to learn the temporal dependencies, and all the classification results $\hat{y}_l$ and $\hat{y}_r$ are ignored. Only the results shown in Eq.5 are needed for the next stage.

This pre-training strategy of multiple networks avoids the fusion of different features spaces, which reduces the input complexity and the noise learning. It also allows the LSTM to focus only on learning over one specific descriptors sequence Ψ_l or Ψ_r independently, which also helps to avoid the over-fitting problem (Ying, 2019; Boutaleb et al., 2021).

3.4 Post-Fusion-based Classification (PFC)

Once the temporal dependencies are learned, we proceed to the final classification. To this end, similarly to (Boutaleb et al., 2021), we train another multi-input NN that exploits the resulted two pre-trained stacked LSTM layers introduced in (Sec 3.3) that we transfer with a fixed optimized parameters θ_l^* and θ_r^* as illustrated in Figure 6.

Seeking to ensure the best classification accuracy, the two parallel output branches of the transferred LSTMs are concatenated, then fed into a Multi Layers Perceptron (MLP) that consists of two Fully Connected (FC) layers, followed by a softmax layer (Figure 6). We model this network as shown in Eq.6, where F_γ is a MLP+softmax with learnable parameters γ, and C is the concatenation function:

$$F_\gamma(C(\{G_{\theta_l^*}, G_{\theta_r^*}\})) \quad (6)$$

The learnable parameters γ are optimized using the same loss function as in the previous stage (Sec 3.3) by minimizing over the same training data set.

This post-fusion strategy aims at ensuring a good accuracy score by tuning between the pre-trained LSTMs outputs.

4 EXPERIMENTS

4.1 Data Sets

Several large-scale data-sets have been proposed for EAR, e.g. EGTEA (Sigurdsson et al., 2018) and CharadesEgo (Fathi et al., 2011). In this work, we try to solve a sub-problem of EAR, namely first-person hand activity recognition, while activities are supposed to be performed with the hands, which is not the case for some activity categories of these data-sets. To this end, to validate our approach, we used the following real-world data sets:

FPHA Data Set. Proposed by (Garcia-Hernando et al., 2018). It provides RGB and depth images with annotations (associated activitie labels). It is a diverse data set that includes 1175 activity videos belonging to 45 different activity categories, in 3 different scenarios performed by 6 actors with high inter-subject and intra-subject variability of style, speed, scale, and viewpoint. It represents a real challenge for activity recognition algorithms. For all the experiments, we used the setting proposed in (Garcia-Hernando et al., 2018), with exactly the same distribution of data: 600 activity sequences for training and 575 for testing.

EgoHand Data Set. Proposed by (Bambach et al., 2015). It has 48 videos recorded with a Google glass. Each video has two actors doing one of the 4 activities: playing puzzle, cards, jenga or chess. These videos are recorded in 3 different environments: office, courtyard and living room. We chose this data set to evaluate our method in case there is not enough

training data. We used the setting proposed by (Bambach et al., 2015) that randomly splits these videos into 36 samples for training, 4 for validation and 8 for the test.

4.2 Implementation Details

Regions of Interest Extraction. To this end, we used a pre-trained NN proposed by (Shan et al., 2020), which is based on Faster R-CNN (Ren et al., 2015). We have mainly chosen this NN for its great hand detection accuracy providing similar performance on the same and cross-data set as reported in (Shan et al., 2020). On the other hand, Faster R-CNN combines a region proposal network (RPN) based on the CNN model with the R-CNN (Girshick et al., 2014). This combination allowed to reduce the computational cost while achieving efficient object detection. This NN is pre-trained on 100K frames of 100DOH data-set (Shan et al., 2020) and 56.4K frames sub-sets of (Damen et al., 2018; Sigurdsson et al., 2018; Fathi et al., 2011). It achieves 90.46% of hand detection accuracy on the 100DOH data set. Detectron2 (Wu et al., 2019) is used for the implementation.

The pre-trained NN predicts bounding boxes for all detected hands in the image frame with a confidence score between 0 and 1. We accept boxes with a confidence score above 0.8. We assign each box to the left or the right hand of the user according to the coordinates of the box center in the image frame. If a third-person hand is detected (more than two hands), we only consider the largest boxes (the closest to the camera) as the user's hands. Finally, for the frames with no available detection, we assume a hand position below the field of view.

Spatial Features Extraction. For this purpose, we deliberately chose VGG16 (Liu and Deng, 2015) for its widespread use as a standard foundation for TL (Tammina, 2019) and domain adaptation (Chaves et al., 2020). It is a powerful convolutional neural network, mainly designed for large-scale image recognition. VGG16 model contains a stack of convolutional layers which capture basic features like spots, boundaries, and colors pattern followed by three fully-connected layers (FCL) that provides complex higher-level feature patterns. To this end, we extracted features from the last FCL, which provides an output vector of dimension 1x4096. VGG16 has shown good results. However, it is highly computational due to its complex architecture and a large number of parameters. Moreover, the size of its last FCL output is very large, and multiplying this size by the length of the activity sequence results in a large input dimension (200x4096) for the LSTM network. This requires high computing resources and time for the training process. Indeed, we experimented with a lighter pre-trained model, namely MobileNetV2 based on an inverted residual structure (Sandler et al., 2018). Table 2 shows the comparison between VGG16 and MobileNetV2. By using MobileNetV2, the accuracy dropped by 1.5% but we achieved gain in inference/training time and computational resources. The two models VGG16 and MobileNetV2 are pre-trained for image classification tasks on the ImageNet data-set (Russakovsky et al., 2015) achieving 92.7% and 90% accuracy respectively. Keras framework is used for the implementation.

Temporal Dependencies Learning. For each spatial descriptors sequence that refers to the right and the left hands, we trained different configurations of separated NNs that consist of 1, 2, 3, and 4 staked LSTM layers followed by a softmax. We selected the best configuration that gives the best accuracy score: 2 staked LSTM layers of 100 units. We set the probability of dropout to 0.5 (outside and inside the LSTM gates). We used Adam with a learning rate of 0.003 for the optimization. All the networks are trained with a batch size of 64 for 400 epochs. We also padded all sequence lengths to 200 and 100 time-steps per sequence for the FPHA and EgoHand data sets respectively.

Post-Fusion-based Classification. Once all the temporal dependencies are learned (end of stage 3), in the PFC stage, we recover the pre-trained LSTM networks, we fix all their weights and discard softmax layers. Then, the two outputs branches from the two parallel transferred LSTMs are concatenated and followed by a MLP that consists of two dense layers of 256 and 128 neurons respectively, equipped with a relu activation function. At the end of the network, a softmax layer is used for the final classification. This network is trained until 100 epochs, with the same batch size and optimization parameters as the previous networks. The implementation is based on Keras framework.

4.3 State-of-the-Art Comparison

Table 1 shows the accuracy of our approach compared with state-of-the-art methods on the FPHA data set. The best performing approach among state-of-the-art methods is Tear (Li et al., 2021), a transformer-based that consists of two modules, inter-frame attention encoder, and mutual-intentional fusion block. By exploiting RGB and depth modalities they achieved 97.04% of accuracy, which is equivalent to our achievement (97.91%) while using the RGB modality only. The approach proposed by Boutaleb et al.

Table 1: Activity recognition accuracy comparison of our proposed approach and the state-of-the-art on the FPHA data set. Our method outperforms all RGB-based methods including end-to-end visual attention methods.

Methods	Year	Modality	Accuracy(%)
Two stream-color (Feichtenhofer et al., 2016)	2016	RGB	61.56
H+O (Tekin et al., 2019)	2019	RGB	82.26
Rastgoo et al. (Rastgoo et al., 2020)	2020	RGB	91.12
Trear (Li et al., 2021)	2021	RGB	94.96
HON4D (Oreifej and Liu, 2013)	2013	Depth	59.83
HOG2-depth (Ohn-Bar and Trivedi, 2014)	2014	Depth	70.61
Novel View (Rahmani and Mian, 2016)	2016	Depth	69.21
Trear (Li et al., 2021)	2021	Depth	92.17
Lie Group (Vemulapalli et al., 2014)	2014	3D Pose	82.69
Gram Matrix (Zhang et al., 2016)	2016	3D Pose	85.39
TF (Garcia-Hernando et al., 2018)	2017	3D Pose	80.69
Nguyen et al. (Nguyen et al., 2019)	2019	3D Pose	93.22
Boutaleb et al. (Boutaleb et al., 2021)	2020	3D Pose	**96.17**
HOG2-depth+pose (Ohn-Bar and Trivedi, 2014)	2014	Depth+3D Pose	66.78
JOULE-all (fang Hu et al., 2015)	2015	RGB+Depth+3D Pose	78.78
Tear (Li et al., 2021)	2021	RGB+Depth	**97.04**
Our	-	RGB	**97.91**

Table 2: Performance comparison of our method on FPHA data-set using two different pre-trained NNs for spatial features extraction, namely VGG16 and MobileNetV2.

Model	Inference time (ms)	Parameters (millions)	Last FCL size	Acc.(%)
VGG16	5.17	138	1x4069	96.52
MobileNetV2	3.34	3.5	1x1028	95.01

Table 3: Activity recognition accuracy results on EgoHand data-set that contains only 48 samples. Results show that our method performs better on a limited amount of data.

Method	Acc (%)
Khan et al. (Khan and Borji, 2018) + Ground truth hand mask	71.1
Khan et al. (Khan and Borji, 2018)	68.4
Bambach et al. (Bambach et al., 2015) + Ground truth hand mask	**92.9**
Bambach et al. (Bambach et al., 2015)	73.4
Babu et al. (Babu et al., 2019)	89.0
Our	**98.79**

(Boutaleb et al., 2021) gives good results, but they used the ground truth of 3D hand joints, which is not always available. This may conclude that RGB image sequences can provide the necessary elements to recognize hand activities.

Table 3 shows the accuracy of our approach compared to state-of-the-art methods on the EgoHand data set. The proposed work by (Khan and Borji, 2018) and (Bambach et al., 2015) was more focused on hand segmentation in an egocentric viewpoint. Nevertheless, they used the estimated and ground-truth hand masks to recognize activities. We outperformed their results by more than 5% of accuracy, confirming the effectiveness of the proposed regions of interest over the hand mask. Since the EgoHand contains only 48 samples, this can also prove the ability of our method to learn on a limited amount of data.

4.4 Contribution of Proposed Regions of Interest

To better show the contribution of left and right hands regions of interest, we skipped the RoIE stage. Instead, we used the full-image frames. As expected, results presented in table 4 shows that without our regions of interest, the accuracy dropped by more than 14%, which confirms RoIE effectiveness. Moreover, by using only the right hand as the region of interest, we overcome most state-of-the-art methods.

Table 4: Activity recognition accuracy results on FPHA data-set with and without using our proposed regions of interest. Results show the significant impact of these regions of interest.

Extracted region of interest	Acc.(%)
Full image	82.01
Left hand bounding box	85.00
Right hand bounding box	**91.82**
Left+Right hands bounding boxes	**96.52**

As we mentioned in section 3.1, highly relevant information related to manipulated objects (nouns) e.g. "juice bottle" can be derived from the visual data of the hand boxes, such as grasp type and object shape. Furthermore, by learning the temporal dependencies through this information, we can also relevantly characterize the actions (verbs) e.g. "open". For more ablation studies, we experimented our method on object and action recognition. Table 5 shows that our proposed method gives a good object and action recognition score by achieving 97.56% and 94.26% of accuracy respectively.

Table 5: Object (noun) and Action (verb) recognition accuracy on FPHA data-set using our proposed pipeline. The accuracy results show that the proposed regions of interest allow object and action recognition which facilitates the hand activity recognition.

Task	Number of classes	Region of interest	Acc(%)
Objects (nouns)	27	Left hand	88.69
		Right hand	95.82
		Left+Right hands	**97.56**
Actions (verbs)	27	Left hand	85.56
		Right hand	92.17
		Left+Right hands	**94.26**

4.5 Data Augmentation

The results in Table 6 show that the accuracy is significantly increased by 1.39% when we used our adapted data augmentation process. Furthermore, using only the right-hand regions of interest, we outperforms most state-of-the-art methods by achieving 94.26% of accuracy.

Table 6: Activity recognition accuracy results on FPHA data-set. (*) without data augmentation, (**) using data augmentation.

Extracted region of interest	Acc(*) (%)	Acc(**) (%)
Left hand bounding box	85.00	88.00
Right hand bounding box	91.82	94.26
Left+Right hands bounding boxes	96.52	**97.91**

5 CONCLUSION

In this paper, a novel learning pipeline for first-person hand activity recognition has been introduced. The proposed pipeline is composed of four stages. In the first stage, we presented our TL-based regions of interest extraction, the left, and right hands regions, which has proven to be effective. The second stage is the TL-based deep spatial feature extraction method that exploits the regions of interest visual information. To manage the temporal dimension, in the third stage we trained temporal NNs in a multi-stream manner. Then, in the last stage, we applied a post-fusion strategy to classify activities. The pipeline is evaluated on two real-world data sets and showed good accuracy results.

As future improvements, we plan to exploit other regions of interest, for e.g. the manipulated object regions, in order to avoid the ambiguous case of high intra-class dissimilarity, where manipulated objects in the same activity class may have different shapes, grip types, and colors, which may be challenging for our proposed regions of interest that focus only on hands' motion and appearance.

REFERENCES

Abebe, G., Cavallaro, A., and Parra, X. (2016). Robust multi-dimensional motion features for first-person vision activity recognition. *Comput. Vis. Image Underst.*, 149:229–248.

Avola, D., Bernardi, M., Cinque, L., Foresti, G. L., and Massaroni, C. (2019). Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. *IEEE Transactions on Multimedia*, 21:234–245.

Babu, A. R., Zakizadeh, M., Brady, J., Calderon, D., and Makedon, F. (2019). An intelligent action recognition system to assess cognitive behavior for executive function disorder. *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pages 164–169.

Bambach, S. (2015). A survey on recent advances of computer vision algorithms for egocentric video. *ArXiv*, abs/1501.02825.

Bambach, S., Lee, S., Crandall, D. J., and Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1949–1957.

Boutaleb, Y., Soladié, C., Duong, N.-D., Kacete, A., Royan, J., and Séguier, R. (2021). Efficient multi-stream temporal learning and post-fusion strategy for 3d skeleton-based hand activity recognition. In *VISIGRAPP*.

Cao, C., Zhang, Y., Wu, Y., Lu, H., and Cheng, J. (2017). Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3783–3791.

Chaves, E., Gonçalves, C. B., Albertini, M., Lee, S., Jeon, G., and Fernandes, H. (2020). Evaluation of transfer learning of pre-trained cnns applied to breast cancer detection on infrared images. *Applied optics*, 59 17:E23–E28.

Damen, D., Doughty, H., Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., and Wray, M. (2018). Scaling egocentric vision: The epic-kitchens dataset. *ArXiv*, abs/1804.02748.

fang Hu, J., Zheng, W.-S., Lai, J.-H., and Zhang, J. (2015). Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*.

Fathi, A., Ren, X., and Rehg, J. M. (2011). Learning to recognize objects in egocentric activities. *CVPR 2011*, pages 3281–3288.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941.

Garcia-Hernando, G., Yuan, S., Baek, S., and Kim, T.-K. (2018). First-person hand action benchmark with rgb-d videos and 3d hand pose annotations.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2020). Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

Irani, M. and Anandan, P. (1999). About direct methods. In *Workshop on Vision Algorithms*.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:221–231.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Khan, A. U. and Borji, A. (2018). Analysis of hand segmentation in the wild. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4710–4719.

Kläser, A., Marszalek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC*.

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., and Gong, B. (2021). Movinets: Mobile video networks for efficient video recognition. *ArXiv*, abs/2103.11511.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Li, X., Hou, Y., Wang, P., Gao, Z., Xu, M., and Li, W. (2021). Trear: Transformer-based rgb-d egocentric action recognition. *ArXiv*, abs/2101.03904.

Liu, S. and Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. (2016). Ssd: Single shot multibox detector. In *ECCV*.

Liu, Y., Jiang, X., Sun, T., and Xu, K. (2019). 3d gait recognition based on a cnn-lstm network with the fusion of skegei and da features. *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8.

Ma, M., Fan, H., and Kitani, K. M. (2016). Going deeper into first-person activity recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1894–1903.

Min, K. and Corso, J. J. (2020). Integrating human gaze into attention for egocentric activity recognition. *ArXiv*, abs/2011.03920.

Nguyen, X. S., Brun, L., Lézoray, O., and Bougleux, S. (2019). A neural network based on spd manifold learning for skeleton-based hand gesture recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12028–12037.

Ohn-Bar, E. and Trivedi, M. M. (2014). Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 15:2368–2377.

Oreifej, O. and Liu, Z. (2013). Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723.

Poleg, Y., Arora, C., and Peleg, S. (2014). Temporal segmentation of egocentric videos. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2544.

Rahmani, H. and Mian, A. S. (2016). 3d action recognition from novel viewpoints. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1506–1515.

Ramirez-Amaro, K., Beetz, M., and Cheng, G. (2017). Transferring skills to humanoid robots by extracting semantic representations from observations of human activities. *Artif. Intell.*, 247:95–118.

Rastgoo, R., Kiani, K., and Escalera, S. (2020). Hand sign language recognition using multi-view hand skeleton. *Expert Syst. Appl.*, 150:113336.

Rawat, W. and Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29:2352–2449.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252.

Ryoo, M., Rothrock, B., and Matthies, L. (2015). Pooled motion features for first-person videos. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 896–904.

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. *Proceedings of the 15th ACM international conference on Multimedia*.

Shan, D., Geng, J., Shu, M., and Fouhey, D. F. (2020). Understanding human hands in contact at internet scale.

2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9866–9875.

Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., and Karteek, A. (2018). Charades-ego: A large-scale dataset of paired third and first person videos. *ArXiv*.

Singh, S., Arora, C., and Jawahar, C. V. (2016). First person action recognition using deep learned descriptors. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2620–2628.

Sridhar, S., Feit, A. M., Theobalt, C., and Oulasvirta, A. (2015). Investigating the dexterity of multi-finger input for mid-air text entry. In *CHI '15*.

Sudhakaran, S., Escalera, S., and Lanz, O. (2019). Lsta: Long short-term attention for egocentric action recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9946–9955.

Sudhakaran, S. and Lanz, O. (2017). Convolutional long short-term memory networks for recognizing first person interactions. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2339–2346.

Sudhakaran, S. and Lanz, O. (2018). Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. *ArXiv*, abs/1807.11794.

Surie, D., Pederson, T., Lagriffoul, F., Janlert, L.-E., and Sjölie, D. (2007). Activity recognition using an egocentric perspective of everyday objects. In *UIC*.

Tadesse, G. and Cavallaro, A. (2018). Visual features for ego-centric activity recognition: a survey. *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*.

Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International journal of scientific and research publications*, 9:9420.

Taylor, G. W., Fergus, R., LeCun, Y., and Bregler, C. (2010). Convolutional learning of spatio-temporal features. In *ECCV*.

Tekin, B., Bogo, F., and Pollefeys, M. (2019). H+o: Unified egocentric recognition of 3d hand-object poses and interactions. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4506–4515.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595.

Verma, S., Nagar, P., Gupta, D., and Arora, C. (2018). Making third person techniques recognize first-person actions in egocentric videos. *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2301–2305.

Wang, H., Kläser, A., Schmid, C., and Liu, C. (2012). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103:60–79.

Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2.

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995.

Ying, X. (2019). An overview of overfitting and its solutions.

Zhang, X., Wang, Y., Gou, M., Sznaier, M., and Camps, O. (2016). Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.