Object-based Temporal Segment Relational Network for Activity Recognition

Victor H. C. Melo[1], Jesimon B. Santos[1], Carlos Caetano[1], Jessica Sena[1],
Otavio A. B. Penatti[2], William Robson Schwartz[1]
[1]Smart Sense Laboratory, Department of Computer Science, Universidade Federal de Minas Gerais, Brazil
[2]Advanced Technologies, Samsung Research Institute, Campinas, Brazil

***Abstract*—Video understanding is the next frontier of computer vision, in which activity recognition plays a major role. Despite the recent improvements in holistic activity recognition, further researching part-based models such as context may allow us to better understand what is important for activities and thus improve our current activity recognition models. This work tackles contextual cues obtained from object detections, in which we posit that objects relevant to an action are related to its spatial arrangement regarding an agent. Based on that, we propose Egocentric Pyramid to encode such spatial relationships. We further extend it by proposing a data-centric approach named Temporal Segment Relational Network (TSRN). Our experiments give support to the hypothesis that object spatiality provides an important clue to activity recognition. In addition, our data-centric approach shows that besides such spatial features, there may be other important information that further enhances the object-based activity recognition, such as co-occurrence, relative size, and temporal information.**

I. Introduction

Context plays an important role in activity recognition from videos. Certain activities are strongly tied to its surroundings, which give important priors for activity discrimination. Such contextual cues might include objects, pose, and scene, to name a few of the most evident. Not only that, but characterizing such cues may enable applications in which the relationship between context and activities are important, such as fine-grained activity recognition, video understanding, and retrieval. For instance, in video retrieval, one might be interested in videos that contain one ball, but that are played only with feet, such as soccer, while ignoring videos such as basketball and volleyball. In this kind of task, the spatial location of the ball regarding the agent is one of the most discriminative features. Therefore, characterizing the video surroundings and its synergy with activities is an important task.

With recent advances in image and scene classification, object detection, and pose estimation, there are several context cues that can be gathered for characterizing activities. Jain et al. [1] investigated the importance of context provided by surrounding objects, using an object classifier. Other works also considered objects [2], [3], temporality [4], scene [3], audio [5], and pose [6]. Unlike existing works, our approach turns our attention to spatial cues and its role in activity recognition. For such purpose, we rely on object bounding boxes obtained by state-of-the-art object detectors and their spatial relationships.

In this work, we propose two different approaches to exploit such object detections. Initially, we build object histograms in a pyramid fashion to capture spatial relationships. To reduce issues regarding viewpoint differences, we consider that relevant objects to recognize the activity in execution move along with the performing agent. Thus, we compute the spatial pyramid on top of the most salient agent, which we call *Egocentric Pyramid*. After that, we turn to a data-centric approach, named *Temporal Segment Relational Network (TSRN)*, to learn relationships between Egocentric Pyramids and object detections, while also including temporal features. Our experimental results show that Egocentric Pyramid is able to improve upon spatial pyramid, giving support to our hypothesis. Furthermore, Temporal Segment Relational Network is able to improve over Egocentric Pyramids and reach results similar to others in the state-of-the-art [7], [8].

The contributions of this paper are threefold, namely, (i) we conduct a study on object spatial information for activity recognition; (ii) we propose Egocentric Pyramid, a spatial pyramid that encodes relative spatial relationships regarding the main agent in an activity; and, finally, (iii) we propose the Temporal Segment Relational Network (TSRN), a data-centric approach which leverages both Egocentric Pyramids and other non-explicit relationships between objects and temporal cues. To the best of our knowledge, this is the first approach to focus on spatial and temporal relations between objects for activity recognition, which is also learned through a full end-to-end network architecture.

The remainder of this paper is organized as follows. Section II discusses the related work in contextual information for activity recognition. Section III describes our proposed approaches to leverage contextual cues provided by an object detector. Section IV presents our experimental evaluation of the proposed approach. Finally, in Section V, we present our concluding remarks and future directions of work.

II. Related Work

Contextual cues are auxiliary information to the main task at hand, which in the case of activity recognition from videos, comprises information extracted from video that may not be the activity itself but that may help in the recognition of activities. In particular, Jain et al. [1] thoroughly investigated the use of contextual clues using an object classifier and showed that objects matter for activity recognition. They use

a classifier trained on 15,000 object categories and combine them with HOG [9], MBH, and HOF [10] with improved trajectories [11]. In contrast, in our approach, we expand on this investigation by exploring spatial and temporal cues through an object detector. Additionally, our experimental evaluation (Section IV) showed that using only 200 object categories are enough for the UCF101 Dataset, given that a more accurate detector is employed.

SR-CNN [2] exploits contextual information for activity recognition, similarly leveraging the most salient agent for recognition. SR-CNN is an end-to-end deep learning architecture to integrate object detections with two-stream features. Niebles et al. [12] built concentric ellipsis around an agent performing an action, in addition to other features such as relative size and occupancy. Our approach also builds a pyramid, but as a grid instead. More importantly, we also provide a more thorough evaluation by conducting our experiments on a more challenging dataset, which includes camera movements, more activities, and background objects unrelated to the activity. Furthermore, we also employ a data-centric approach to learn the semantic relationships between other features.

In the literature, several works also exploit context in a variety of tasks. Object bank [13] builds a hierarchical spatial pyramid using the prediction scores of multiple object detectors aiming at scene classification. In contrast to our approach, it employs a regular spatial pyramid which does not focus on the main agent and it is not aimed at activity recognition. Similarly, word spatial arrangement (WSA) [14] encodes the spatial relationship of visual words for image retrieval and classification. Egocentric Pyramids can be seen as WSA employed for encoding the spatial arrangement of object detections. However, WSA is a lower level descriptor modeling keypoints' arrangement, while Egocentric Pyramid is oriented towards higher semantic concepts.

Since relational networks [15] were not originally conceived for activity recognition, adapting it to this task is not trivial. First, we modify it to also gather temporal information using the framework proposed by Temporal Segments Network [8]. Second, we modify the input feature descriptor by augmenting it with Egocentric Pyramids, relative frame position, and also quantizing the object coordinates. The latter is due to object distribution on these videos being more sparse. Thus, we do not need much precision to encode the object coordinate as well as required by other tasks, such as question-and-answering. Third, we replaced the sum pooling operation by averaging, and instead of using ReLU activations, we replace them by SELU [16], as both featured better convergence properties in our experiments. We also tried architecture variations with/without layers and without duplicating the last fully connected layer parameters, and residual connections [17], [18].

Concurrent to our work, Zhou et al. [19] propose an architecture for temporal relational reasoning also based on relational networks [15]. This approach builds a hierarchical relational network pyramid, with varying time scales, to encode temporal information given CNN features. In contrast, our proposed approach models both spatial and temporal relationships between objects given by an object detector. The spatial relationships are obtained by the objects' coordinates and Egocentric Pyramid, while the temporal information is encoded through normalized frame position and the Temporal Segment Networks [8] framework.

III. Contextual Cues

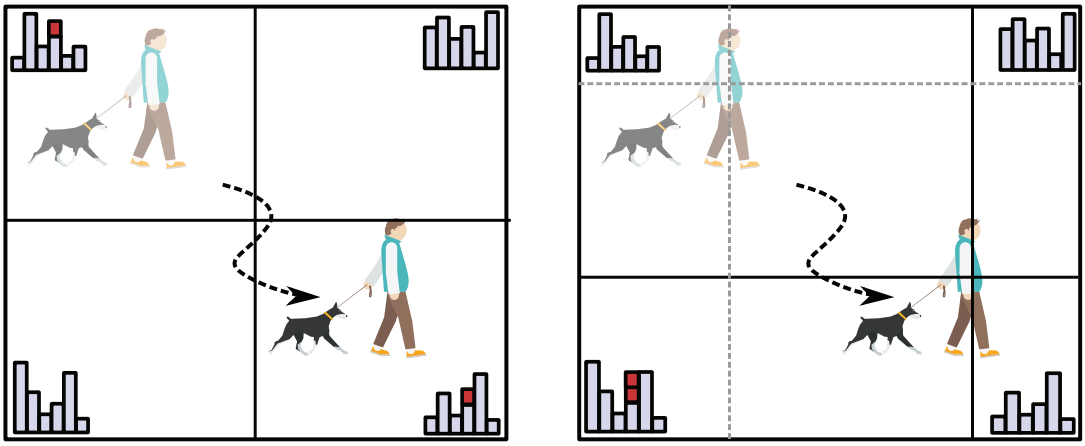
In this section, we describe our proposed approach to take contextual cues into account aiming at activity recognition from videos. Such cues include object spatiality, temporality, and co-occurrence. For encoding such contextual information, we propose two novel approaches, namely *Egocentric Pyramid* and *Temporal Segment Relational Network (TSRN)*. Aiming at understanding spatial cues, we first conduct a case-study in which we design Egocentric Pyramid, a feature descriptor that encodes the spatial arrangement of objects around an agent (see Section III-A). This approach is drawn from the hypothesis that, not only the location of an object matters, but also how they are arranged regarding the agent performing the action.

Afterwards, we turn to a data-centric approach that encodes latent relationships between objects. The proposed approach, Temporal Segment Relational Network (TSRN), is designed to reason about contextual cues through time (Section III-B). Both Egocentric Pyramids and the raw object detections are fed as input to TSRN, enabling it to learn latent features which will be used for classification. Considering that both approaches require object detections, we employ a state-of-the-art object detector [20] to provide bounding boxes for each frame. Each object detection is a 5-tuple comprising a confidence score and the bounding box coordinates $\{(x, y), (w, h)\}$. The following sections detail how these detections are employed in our representation.

A. Egocentric Pyramid

We begin by investigating the objects' spatial arrangement through a spatial pyramid [13]. The spatial pyramid approach builds a pyramid composed of a hierarchy of levels, where a frame in each level is divided into an equally spaced grid. Each quadrant in the grid accounts for the objects that occur within a given region. The grid resolution is determined by the pyramid level, which allows capturing the objects' position with increasing precision as the pyramid level increases, in a coarse-to-fine fashion. For each quadrant in the grid, we compute a histogram of the softmax scores. Then, the histograms across all pyramid levels and quadrants are concatenated into a single vector and presented to a classifier. The resulting dimensionality of the feature vector is a function of the number of pyramid levels and histogram bins, which is given by the number of object categories.

One issue with spatial pyramid is that it is not invariant to the position of the agent performing the action, as we are taking the center of the frame as reference. It can be a problem because it assumes that all activities are always performed at the center of the video, which is not necessarily



(a) (b)

Fig. 1. Difference when computing pyramids through time. Faded-out persons represents someone in the previous frame, while fully-colored represents current frame. (a) Spatial pyramids, in which the pyramid is always at the frame's center; and (b) Egocentric Pyramid, whose center moves according to the agent.

true. For instance, if we are targeting the *walking with dog* activity and the person escorting the dog starts on the upper-left corner of the frame and then moves to the bottom-right corner, the corresponding 'dog' bin will be assigned into the histograms corresponding to the second and fourth quadrants (see Figure 1a). That will generate a different histogram signature for the same activity but in one case that the person with the dog starts at the bottom-left and moves to the bottom-right. However, this could be prevented if we took as reference the agent position instead of the frame's center, since the relevant objects seem to move around the one performing the action.

In human activities, we argue that the most important objects interact by surrounding the person/agent performing the action, hence, moving in relation to this person. Therefore, we propose to extract object spatial pyramids regarding the center of the agent, hence, an Egocentric Pyramid. By taking the agent performing the activity as the reference for the pyramid, we minimize the aforementioned issues. Considering the previous example, in both cases, the 'dog' bin will be assigned to the third histogram. In other words, the dog is always at the bottom-left or bottom-right quadrants regarding the person (Figure 1b).

Before computing the pyramid, we first need to determine who is the agent. We define an *agent* as the person performing the action, and to determine it, we use saliency [2], i.e., the one that is more noticeable. Given this assumption, we adopt a simple heuristic: we consider the agent as the person with the highest score predicted by the object detector. A more sophisticated approach would be to track the main agent. However, the score-based heuristic suffices for our purposes because picking another agent will happen occasionally, as background persons' saliency usually lack consistency through time. In addition, as it is aggregated for multiple frames, it will not affect the descriptor significantly. Finally, in case this heuristic misses an agent, we switch to the frame's center as reference.

As a further enhancement of the descriptor, we evaluate a soft-assignment strategy for bin assignment. While the hard-assignment uses the centroid to determine to which quadrant an object belongs, that will be then assigned to its respective bin, the soft-assignment strategy estimates the area that an object occupies in each quadrant. Such area is then normalized and used to weigh in the scores assigned to each bin. This would allow removing unfairness of assigning an object to a single quadrant, even though it might occupy a large portion of the other three quadrants. However, as we shall see in the experimental evaluation, such strategy worsened the recognition accuracy.

B. Temporal Segment Relational Network

So far, we have described context through object spatiality alone. As aforementioned, our goal is also to explore other contextual features from object detections that are discriminative for activity recognition. Hand-crafting each of these features might be a daunting task and we might miss important correlations between certain features. Therefore, in this section, we investigate a data-centric approach to encode meaningful context features which complements Egocentric Pyramids. For this task, we propose a temporal model for relational reasoning, named Temporal Segment Relational Network (TSRN), based on relational networks [15] and Temporal Segment Networks [8].

Relational networks (RN) compute relations by taking pairs of features as input, which is performed for the cartesian product of all pairs of objects. This phase can be seen as encoding all pairs of relations into a single, small representation, from which we will make predictions. However, RN were not originally designed to handle temporal data. An important principle to have in mind is how to encode object consistency across the temporal dimension. With this intention, TSRN resorts to the framework proposed by Temporal Segment Networks to leverage such consistency by sampling multiple snippets from a video. In each of these snippets, we extract

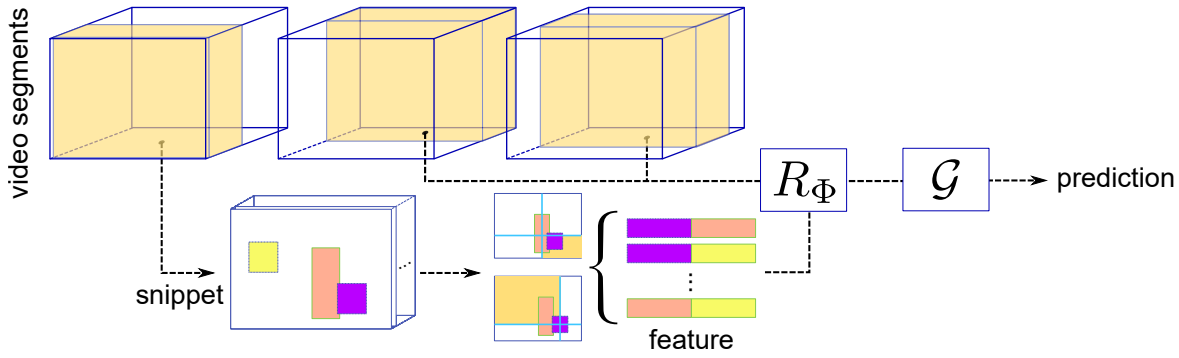


Fig. 2. Overall workflow of the proposed Temporal Segment Relational Network (TSRN). At training time, the input video is split in t segments, from which snippets of length $|S_i|$ are sampled. Each snippet representation is composed of object detections augmented with Egocentric Pyramid, which are then encoded by R_Φ. The relational encodings of each snippet are then merged in the consensus layer $\mathcal{G}$, and we get the predictions after softmax. At test time, snippets are uniformly sampled throghout the whole video [8].

features—in our case, object detections—which are fed to the relational network. The relational network yields activations for each snippet, which are then combined in a segmental consensus layer.

Formally, given a sequence of snippets $\mathbf{S} = \{S_1, S_2, \ldots, S_t\}$, we define TSRN as

$$\text{TSRN}(\mathbf{S}) = \mathcal{G}(R_\Phi(S_1), R_\Phi(S_2), \ldots, R_\Phi(S_t)) \qquad (1)$$

where S_t is a video snippet, R_Φ is a relational network with parameters Φ, and $\mathcal{G}$ is a pooling operation. In particular, a relational network R_Φ, given parameters $\Phi = [\phi_1, \phi_2]$, is defined as

$$R_\Phi(O) = f_{\phi_1}\left(\frac{1}{n^2}\sum_{o_i,o_j} g_{\phi_2}(o_i, o_j)\right). \qquad (2)$$

Here, $O = \{o_i\}_{i=1}^n$ represents an input set of n objects, where o_i is the i^{th} object such that $o_i \in \mathbb{R}^f$; and functions f_{ϕ_1} and g_{ϕ_2} are stacked multilayer perceptrons (MLP) parameterized by parameters ϕ_1 and ϕ_2, respectively. Notice that in this formulation we replaced the original sum operation by averaging, since it showed better convergence in our experiments.

Figure 2 depicts TSRN overall workflow. During training, the video is first split into t segments of equal size T. From each segment, a random snippet S_i is sampled with length $|S_i|$ such that $|S_i| \leq T$. The object detections within a snippet are combined in all pairs possible and then presented to the relational network. Notice that object pairings are only performed within a segment, not across them to avoid a large dimensionality. In this training setting, sampling random snippets is a data augmentation where every time a different snippet is seen by the network. At the same time, we ensure that the video is seen as a whole, according to the number of segments and the snippet length. For instance, if we choose three segments, then we ensure that the network will see data from the beginning, middle, and end of the video. The consensus layer then pushes the network to learn weights that favors consistency across them.

As a well-established principle in neural networks, part of architecture design lies in building invariance properties into its structure [21]. One common design principle is in efficiently reusing weights. For instance, convolutional layers share weights between locations, while recurrent neural networks reuse weights through time. Similarly, relational networks share weights between object pairings through g_{ϕ_2}, while the temporal segments reuses R_Φ through time. This imposes constraints that act as regularizers, while also reducing the amount of parameters [22].

Egocentric Pyramids (Section III-A) and TSRN are orthogonal contributions. In fact, we can augment the input of relational networks with spatial or Egocentric Pyramids—although should be noted that there is a trade-off regarding the representation size due to pyramids having a considerable dimensionality size starting from three levels. In our experiments, we augment the feature descriptor with Egocentric Pyramids applied to every object pairing. Given that there are only two objects concatenated, Egocentric Pyramid does not require bins for all object categories in each quadrant, thus, it will only require four dimensions, each corresponding to a quadrant. Differently from Egocentric Pyramids, here we are computing the spatial arrangement regarding all objects instead of a single main agent, which may decrease the contribution of noisy detections.

IV. Experimental Evaluation

This section evaluates the proposed approaches for activity recognition based on context. We first evaluate Egocentric Pyramid, which allows us to understand how spatiality is relevant for activity recognition. Afterwards, we evaluate different TSRN architectures and combine them with a state-of-the-art architecture for action recognition using a late-fusion strategy.

Experimental setup. We conduct our experiments regarding Egocentric Pyramid and TSRN on the well-known UCF101 Dataset [23]. The object detections are computed by the SSD detector [20], with 500×500 resolution. The Egocentric Pyramid is evaluated with 2 levels and compared to a spatial pyramid with up to three levels. In addition, we propose

TABLE I

ARCHITECTURES USING SHORTHAND NOTATION. NUMERIC VALUES SPECIFIES THE AMOUNT OF NEURONS IN EACH MLP; $*$ STANDS FOR ALPHA-DROPOUT OF 0.5; AND SUBSCRIPTS INDICATE RESIDUAL CONNECTIONS (f_i FROM THE i^{th} MLP LAYER, $\mathcal{G}$ FROM POOLING).

NAME	ACTIV.	ARCHITECTURE
		(f_1, f_2, f_3, f_4)-$\mathcal{G}$-(g_1, g_2, g_3)
TSRN-R	ReLU	$(32, 32, 32, 32)$-sum-$(32, 64^*, 101)$
TSRN-A	SELU	$(32, 32, 32, 32)$-avg-$(32, 64^*, 101)$
TSRN-B	SELU	$(64, 64, 64, 64)$-avg-$(64, 64^*, 101)$
TSRN-C	SELU	$(128, 64, 128_{f_1}, 128_{f_1,f_2})$-avg-$(128, 128^*_{\mathcal{G}}, 101)$
TSRN-D	SELU	$(128, 128, 128_{f_1}, 128_{f_1})$-avg-$(128, 128^*_{\mathcal{G}}, 101)$

a simple extension to the contextual descriptor for action recognition proposed by Jain et al. [1] by incorporating the number of occurrences of objects, which here is only possible because we are using an object detector instead of an object classifier. Occurrence is computed similarly to scores [1], i.e., given a video V with n frames, let $\mathbb{1}$ be the indicator function and H a histogram with D bins (object categories), then the q^{th} bin is defined by $H_q = 1/n \sum \mathbb{1}(p_i \leq \theta)$ for an object detection with score p_i and a threshold θ.

TSRN is trained using the top-20 object detections extracted from 10 frames and 3 temporal segments. As features, we use the relative frame position, prediction score, coordinates $\{(x, y), (w, h)\}$ quantized by a factor of 8, label in one-of-k encoding (200 object categories); Egocentric Pyramid between objects, i.e., relative quadrant between two objects. Therefore, the total number of feature dimensions is 210, which is doubled when presented as input due to concatenation of two objects. The coordinates are quantized because the object positions are very sparse in human action recognition, hence this task does not require fine-grained resolution. The factor of 8 was obtained experimentally.

We train TSRN for 200 epochs, with the initial learning rate set to 0.01 and decreased by 10^{-1} at epochs $\{35, 90, 135\}$ and a batch size of 22. As data augmentation, we employ mirroring along the x-axis. For normalized coordinates, the operation is done as $1 - x - w$ (clipped within $[0, 1]$). We also tried adding noise to coordinates and scores, however, it was omitted as it did not provide significant gains. During the test, we sample 25 equally spaced segments and fuse their predictions. Architectures are detailed in Table I.

Discussion. We begin by conducting preliminary experiments on the 1st split of the UCF101 dataset to evaluate Egocentric Pyramid and its baselines, namely, spatial pyramid, object scores as reported by Jain et al. [1], our implementation using an object detector, and our extension based on occurrences. Table II shows that using the SSD detector [20] to reproduce the baseline yields a gain of 5 percentual points (p.p.) when compared to the original report by Jain et al. In addition, by evaluating the representation of number of occurrences alone, we achieve a similar result to the baseline (65%). It shows that occurrences by itself are not as representative as object scores, despite obtaining similar results to Jain et al.'s [1] due to a better object detector. Combining both scores and occurrences yields an accuracy of 72%, demonstrating that there is spatial information that is not captured by scores alone. We further delve into this hypothesis by evaluating the spatial and Egocentric Pyramids. The improvement obtained by the spatial pyramid over #occurrences + scores shows that more fine-grained spatial relationships may be relevant for action recognition. In addition, Egocentric Pyramid slightly improves upon spatial pyramids, suggesting that it addresses the aforementioned spatial pyramid issues. We also tried Egocentric Pyramid with a soft-assignment approach whose object positions are weighted according to the area it occupies in each quadrant, however, it yields less accuracy.

TABLE II

RESULTS ON THE UCF101 1ST SPLIT. ALL MODELS WERE EXECUTED USING AN OBJECT DETECTOR COMPRISING 200 CLASSES, EXCEPT FOR JAIN ET AL. [1] WHICH USED A 15000 OBJECT CLASSIFIER.

APPROACH	1ST SPLIT
scores [1]	65%
scores	70%
#occurrences	65%
#occurrences + scores	**72%**
spatial pyramid (2 levels)	73.37%
spatial pyramid (3 levels)	73.52%
egocentric	**73.98%**
egocentric (soft)	70.47%
TSRN-R	—
TSRN-A (no mirroring)	73.56%
TSRN-A	74.96%
TSRN-B	**75.90%**
TSRN-C	70.12%
TSRN-D	75.38%

Afterwards, we evaluate different TSRN architectures coupled with the egocentric representation. Initially, we evaluated TSRN using the same sum pooling and ReLU activation as the original RN [15], which we named TSRN-R (Table II). However, it was unable to converge. By iteratively tweaking components, we found out that replacing the 'sum' operation by 'avg', and then changing the activations to SELU [16] helped convergence. Thus, the remaining architectures all followed these modifications. In our first experiment, we see that data augmentation using mirroring is important, showing an improvement of 1.4 p.p., as demonstrated by TSRN-A with/without mirroring. We therefore train architectures TSRN-B, TSRN-C, and TSRN-D also with mirroring. TSRN-B, featuring slightly more capacity, shows an improvement of 0.94 p.p. over TSRN-A. Comparing these results with the Egocentric Pyramid, we see that there is an improvement in all architecture variations, which suggest that there are other non-explicit contextual features that might be exploited besides spatial arrangements, such as temporal cues from relative frame position and multi-snippets, and other spatial cues such as size and fine-grained localization. In TSRN-C and TSRN-D, we tried to increase TSRN capacity and added residual connections to handle convergence during training. However, it did not provide significant gains regarding TSRN-B, possibly

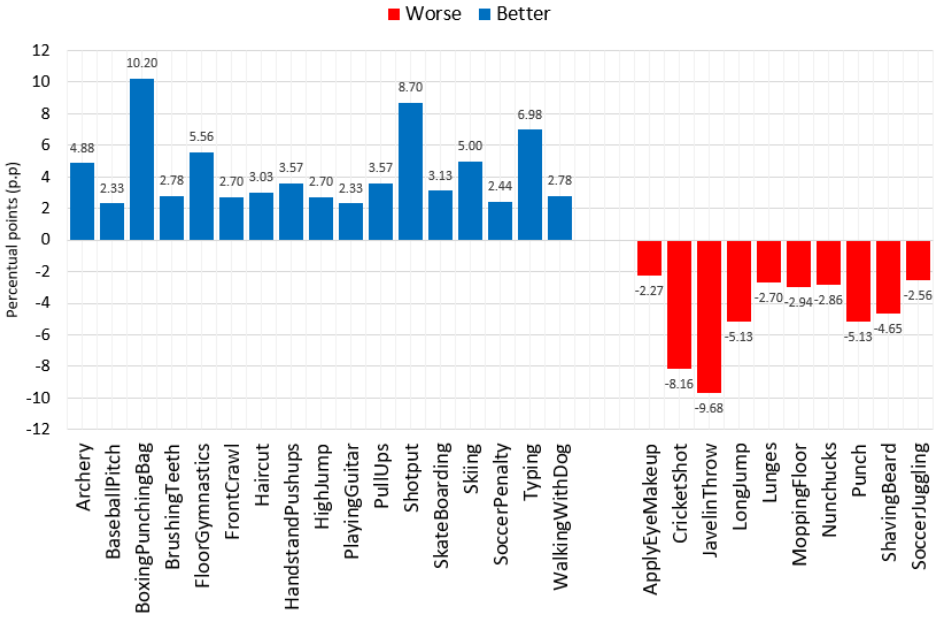


Fig. 3. Accuracy difference between TSRN and TSRN + TSN two-stream. Improvements are shown in blue.

related to the amount of data used. As TSRN-B performed better, we used it in our remaining experiments, referring to it simply as TSRN.

Table III presents TSRN evaluated over the three UCF101 splits, where the last column is the average of the accuracies. We compare these results with the well-known two-stream networks [24], with modifications by Wang et al. [7]. Results of Wang et al. [7] were obtained by running the code provided by the authors. We also include the Temporal Segment Networks [8], whose results on each UCF101 split are available at the authors webpage[1]. We see that, compared to the spatial stream, TSRN obtains a similar result, which might suggest that it is encoding part of the necessary spatial information encoded by the spatial stream. However, it still lacks other cues such as appearance and scene, which explains why it was unable to overcome the spatial stream. When the spatial stream is also imbued in the TSN framework [8], the gap between the two approaches increases. The reason might be that TSN enables the spatial ConvNet to learn temporally consistent visual patterns that are not available from object detections alone, such as scene/background and pose cues. Fusing TSRN predictions with temporal stream yields an improvement close to two-streams itself, suggesting complementarity between the two modalities. Comparing to two-stream alone, we see that our fusion of TSRN + two-stream is able to slightly improve recognition (1.34 p.p). However, this gain is smaller when fused with TSN (0.04 p.p.), showing a smaller complementarity between the two approaches.

To better understand how TSRN and TSN affect each other, we analyze the difference in accuracy for each activity class regarding the fusion of TSRN + TSN two-stream. Figure 3 shows a summary of the scenarios that TSRN + TSN two-stream performed better (blue) and worse (red) than TSN two-stream alone. We see that activities that have objects easier to recognize, such as *archery*, *boxing punching bag*, *shotput*, and *typing* performed better, while it performed worse for classes that objects are difficult to detect (*apply eye makeup*), or that appearance and/or motion plays a major role (*long jump*, *punch*), or objects are absent among the detector categories (*javelin throw*). Comparing TSRN alone with TSN, we see that TSRN only performed better in situations that objects played an important role, such as *playing guitar* or *horse riding*. Still, appearance and motion perform better in the majority of classes, as expected. However, as we have seen in Figure 3 and Table III, there are activity categories that benefit from fusing it with TSRN, suggesting that there are contextual cues that

[1]http://yjxiong.me/others/tsn/

TABLE III
TEMPORAL RELATIONAL NETWORKS COMPARED TO STATE-OF-THE-ART ARCHITECTURES ON THE UCF101 DATASET.

APPROACH	SPLIT 1	SPLIT 2	SPLIT 3	AVG
spatial stream [7]	79.8%	77.3%	77.8%	78.4%
temporal stream [7]	85.7%	88.2%	87.4%	87.0%
two-stream [7]	90.9%	91.6%	91.4%	91.4%
TSN spatial [8]	85.5%	84.9%	84.5%	85.1%
TSN temporal [8]	87.6%	90.2%	91.3%	89.7%
TSN two-stream [8]	**93.5%**	**94.3%**	**94.5%**	**94.0%**
TSRN	75.90%	76.34%	73.51%	75.25%
TSRN + spatial	88.77%	87.31%	87.47%	87.85%
TSRN + temporal	87.91%	90.98%	89.70%	89.53%
TSRN + two-stream	92.33%	93.04%	92.85%	92.74%
TSRN + TSN spatial	86.14%	85.09%	84.26%	85.16%
TSRN + TSN temporal	88.01%	91.05%	90.95%	90.00%
TSRN + TSN two-stream	**93.64%**	**94.24%**	**94.23%**	**94.04%**

may be exploited by action recognition architectures besides appearance/motion.

V. Conclusions

In this work, we studied the role of spatial cues provided by object detections in activity recognition. We proposed two approaches to leverage object detections, namely, Egocentric Pyramid and Temporal Segment Relational Network (TSRN). The former captures spatial relationships centred around an agent, while the latter is an end-to-end network architecture to learn relations between object detections and egocentric pyramids itself. Our experimental results showed that spatiality is significant in activities, even in challenging datasets like the UFC101. Egocentric pyramid obtained better results than spatial pyramid, which suggests that our hypothesis of agent-centred activities might be valid. TSRN improved on both approaches, showing that it uses other features not originally encoded by the pyramid, such as temporality and co-occurrence. TSRN yields similar results to spatial stream [7], but slightly worse than TSN spatial stream [8]. This might be due to TSN spatial stream encoding more information than available from objects. Finally, fusing TSRN with TSN shows that there are activity categories that benefit from object spatiality/temporality, which shreds light on future directions to improve two-stream architectures.

As future work, we will evaluate TSRN on datasets that are more object-centric than UCF101, such as the Something-Something [25] and the EPIC-KITCHENS datasets [26]. Another issue concerns the higher computational cost, because it has the preprocessing step of running an object detector. We posit that coupling egocentric pyramids and TSRN on top of feature maps extracted from spatial and temporal streams may encode spatial relationships without the need of an object detector, reducing the computational cost of this approach.

Acknowledgments

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project). Part of the results presented in this paper were obtained through research on a project titled "HAR-HEALTH: Reconhecimento de Atividades Humanas associadas a Doenças Crônicas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

References

[1] M. Jain, C. G. M. Snoek, and J. C. V. Gemert, "What do 15,000 object categories tell us about classifying and localizing actions?" in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2015, pp. 46–55.

[2] Y. Wang, J. Song, L. Wang, L. V. Gool, and O. Hilliges, "Two-Stream SR-CNNs for Action Recognition in Videos," in *British Machine Vision Conference*, 2016, pp. 108.1–108.12.

[3] L. S. Zuxuan Wu, Yanwei Fu, Yu-Gang Jiang, "Harnessing Object and Scene Semantics for Large-Scale Video Understanding," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3112–3121. [Online]. Available: https://doi.org/10.1109/CVPR.2016.339

[4] Y. Sun, Z. Wu, X. Wang, H. Arai, T. Kinebuchi, and Y.-G. Jiang, "Exploiting objects with LSTMs for video categorization," in *ACM Multimedia*, 2016, pp. 142–146. [Online]. Available: http://doi.acm.org/10.1145/2964284.2967199

[5] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 352–364, 2018.

[6] G. Ch'eron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," in *IEEE Intl. Conference on Computer Vision*, 2015, pp. 3218–3226.

[7] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards Good Practices for Very Deep Two-Stream ConvNets," *CoRR*, jul 2015. [Online]. Available: http://arxiv.org/abs/1507.02159

[8] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in *European Conference on Computer Vision*, 2016. [Online]. Available: http://arxiv.org/abs/1608.00859

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[10] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*, 2006, pp. 428–441. [Online]. Available: http://dx.doi.org/10.1007/11744047_33

[11] H. Wang and C. Schmid, "Action recognition with improved trajectories," *IEEE Intl. Conference on Computer Vision*, pp. 3551–3558, 2013. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2013.441

[12] V. Escorcia and J. C. Niebles, "Spatio-temporal human-object interactions for action recognition in videos," in *IEEE Intl. Conference on Computer Vision Workshops*, 2013, pp. 508–514.

[13] L.-j. Li, H. Su, E. P. Xing, and L. Fei-fei, "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification," in *Neural Information Processing Systems*, 2010, pp. 1–9.

[14] O. A. B. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, and R. da S. Torres, "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognition*, vol. 47, no. 2, pp. 705–720, 2014.

[15] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. P. Lillicrap, "A simple neural network module for relational reasoning," in *CoRR*, vol. abs/1706.01427, 2017. [Online]. Available: http://arxiv.org/abs/1706.01427

[16] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-Normalizing Neural Networks," in *Neural Information Processing Systems*, 2017, pp. 971–980. [Online]. Available: http://arxiv.org/abs/1706.02515

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2016, pp. 171–180.

[18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2017.

[19] B. Zhou, A. Andonian, and A. Torralba, "Temporal Relational Reasoning in Videos," in *arXiv preprint arXiv:1711.08496*, 2017. [Online]. Available: http://arxiv.org/abs/1711.08496

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*, 2016.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[23] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *CoRR*, vol. abs/1212.0402, 2012.

[24] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," in *arXiv preprint arXiv:1406.2199*, 2014, pp. 1–11.

[25] R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," jun 2017. [Online]. Available: http://arxiv.org/abs/1706.04261

[26] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset," *arXiv preprint arXiv:1804.02748*, 2018.