Understanding Event Boundaries for Egocentric Activity Recognition from Photo-Streams

Alejandro Cartas[1*][0000−0002−4440−9954], Estefania Talavera[2][0000−0001−5918−8990], Petia Radeva[1][0000−0003−0047−5172], and Mariella Dimiccoli[3][0000−0002−2669−400X]

[1] Faculty of Mathematics and Computer Science, University of Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain
{alejandro.cartas,petia.ivanova}@ub.es
[2] Bernoulli Institute, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands
[3] Institut de Robòtica i Informàtica Industrial (CSIC- UPC), C/ Llorens i Artigas 4-6, 08028, Barcelona, Spain

Abstract The recognition of human activities captured by a wearable photo-camera is especially suited for understanding the behavior of a person. However, it has received comparatively little attention with respect to activity recognition from fixed cameras. In this work, we propose to use segmented events from photo-streams as temporal boundaries to improve the performance of activity recognition. Furthermore, we robustly measure its effectiveness when images of the evaluated person have been seen during training, and when the person is completely unknown during testing. Experimental results show that leveraging temporal boundary information on pictures of seen people improves all classification metrics, particularly it improves the classification accuracy up to 85.73%.

Keywords: Egocentric Action Recognition · Egocentric Vision · Lifelogging

1 Introduction

Behavior understanding plays a crucial role in improving the habits of people. The activities that people perform in their daily living help in describing their lifestyle. Therefore, automatically discovering their activities is an important step towards understanding their behavior. Several approaches have addressed this problem in the literature [1,2,3]. However, their performance is not close to being precise and automatic.

More recently, the recognition of activities from wearable photo-cameras has gained increasing attention. These devices autonomously capture images at regular intervals of 30 seconds from the first-person perspective, also known as egocentric photo-streams. Since this kind of camera can be worn everywhere and are able to collect sequences over long periods of time such as days, they are well-suited not only for activity recognition but also for understanding different socio-behavioral aspects of a person [4,5].

In this work, we take a step forward in this direction by investigating the questions: *How important are event boundaries for activity recognition from egocentric photo-streams?*, and *Does the temporal coherence of segmented events from egocentric photo-streams improves the activity recognition performance at the frame-level?*. In [6], it was

*Corresponding author.

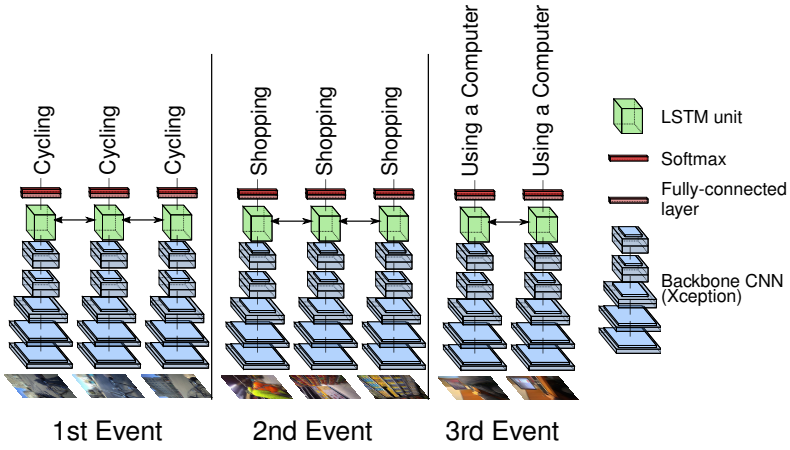

Figure 1. Event-based activity recognition. We first extract event boundaries from a photo-stream sequence that clusters images with similar contextual and semantic features. These boundaries determine the starting and ending frames of a CNN+BLSTM architecture.

demonstrated that the training strategy directly affects the performance of the activity classifier. The core contribution of this work is to evaluate a training strategy for activity recognition based on temporal boundaries that define events. We believe this is a relevant problem to address in the field of egocentric vision, more specifically when analyzing egocentric photo-streams. Our work presents a rich ablation study that defines the basis for future works in the field. Our proposed model (see Fig. 1) allows evaluating our hypothesis which states that we can obtain a more robust classification of the activity occurring in the scene by the inclusion of temporal borders estimated automatically during the training process.

Since egocentric photo-streams describe what the users see throughout their daily routine, they tend to present visual patterns when performing their activities in consecutive days. This is the reason why the same location, people, and objects might appear in several photo-streams when captured by the same person. However, the images that describe such visual settings are not exactly the same, as they are not collected from the exact viewpoint and not only people but objects change over time. *Personalized* learning over the egocentric photo-streams consists of training a model using single frames or full-sequences from a set of users and later on evaluate them on the same set of users. This kind of approach has achieved high classification performance on previous works [7,8,9]. However, *generic* approaches are desired in order to avoid the need for the training phase. We consider the generalization capacity of the algorithm that is when the model is trained with data from a collection of users and is applicable to different ones. In this paper, we evaluate both approaches for the task of activity classification.

The remainder of the paper is organized as follows. In Section 2 we describe the relevant works in the field. Then, in Section 3 we present our event-based approach for daily activity recognition from egocentric photo-streams. Later, in Sections 4 and

5 we describe the experimental setup and results, respectively. Finally, in Section 6 we outline our conclusions and possible future work.

2 Related work

Egocentric vision has shown to be a rich source of information for the understanding of the behavior of the camera wearer. It has allowed the description of social behavior [10,11], food-related scenes [12], and routine [13], among others.

The detection of event boundaries in egocentric videos has been an object of investigation in recent years [14,15]. *Events* are generally understood as a group of sequential images that are homogeneous with respect to a given criterion. What does the criterion specify typically depends on the application at hand. Events were considered as temporal segments characterized by the same global motion and partitioned egocentric videos based on motion-features in [14]. In [15] events are intended as groups of images highlighting the presence of personal locations of interest specified by the end-user. In the domain of egocentric photo-streams events are defined as temporal semantic segments sharing semantic and contextual information [16,17].

The classification of activities has been studied through the analysis of *egocentric videos* [18,19,20]. In these works, the authors addressed the classification of *atomic* actions that describe a more detailed activity. For example the activity *preparing a sandwich* is composed of actions such as *get bread*, *put ham*, and *put mayonnaise*. However, the approaches addressing this problem typically rely on information such as motion and attention patterns that cannot be reliably estimated in photo-streams due to the very low frame rate (1-2 fpm).

The classification of activities from *egocentric photo-streams* has been previously addressed in multiple occasions [7,8,9]. In [7,9], the authors addressed the classification based on information from a single frame, by leveraging semantic and contextual features estimated via a convolutional neural network (CNN). The availability of sequences of images captured at regular intervals was later explored by integrating the temporal information in the classification by using a long short-term memory (LSTM) on the top of a CNN. In [8,21], a learning strategy based on sliding windows over the image sequence improved the testing performance of the classification model.

There have been several multi-modal approaches in the field. For instance, in [1,22,23], the authors proposed the classification of the performed activities by analyzing data collected by different sensors or audio-visual data by using different fusion strategies. The work in [24] presented a multi-modal dataset from sensor data and proposed two methods, one using crafted features and the other using deep learning.

3 Activity Recognition from Event Boundaries

In this paper, we aim to analyze and introduce a simple, but robust activity recognition pipeline for the analysis of given collections of photo-streams.

Figure 2. One hour photo-stream sequence clustered into four consecutive event segments. Each row shows the first 15 frames of each segment.

3.1 Boundaries Detection

Events in egocentric photo-streams correspond to temporally adjacent images that share contextual and semantic features, as defined in [16]. This method relates sequential images represented as a combination of semantic and visual features extracted with a CNN. Temporal boundaries are detected when combining the grouping results obtained by under- and over-segmentation clustering methods, which are combined with graph cuts for energy optimization. We relied on such an approach to extract event boundaries from the daily visual lifelogs or photo-streams in our dataset [8], as shown in Fig. 2. As it can be observed, these events constitute a good basis for activity recognition, since typically, when the user is engaged in an activity, such as *cooking* contextual and semantic features have little variation.

3.2 Event-Based Activity Recognition

In order to exploit the temporal boundaries determined by the event segmentation, we proposed to use a recurrent neural network variant as a temporal learning mechanism. We combined the encoding produced by a CNN with a bidirectional LSTM (BLSTM) [25]. This recursive neural network evaluates a sequence in forward and backward order and merges the result. Thus, it captures patterns that might have been missed by the unidirectional version and it obtains potentially more robust representations [26]. The pipeline of our approach is shown in Fig. 1.

4 Experimental Setup

We train our models in a single training split and evaluate them in two test settings, namely *Generic* and *Personalized*. With *Generic*, we test the generalization capability of our model in images from *unseen* users during training. We want to highlight the difficulty of this task since what we consider the same class environment can be represented by completely different objects and descriptors. In contrast, in the *Personalized* setting the model is tested using images from *seen* users during training, but from different collected days.

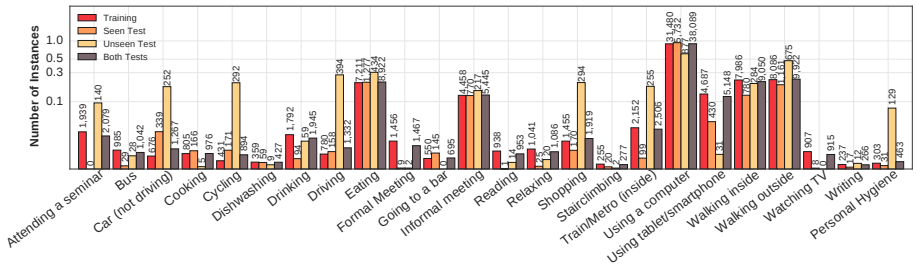

Figure 3. Number of training/testing instances in the data splits. Note that the histograms are normalized and the vertical axis has a logarithm scale, but their corresponding value appears at top of each column.

4.1 Dataset

We carried out our experiments on the ADLEgoDataset [6], a visual lifelogging dataset collected using the Narrative Camera. This dataset consists of 125 egocentric photo-streams with 35 activity categories recorded by 15 students on their daily routine. In this dataset, most of the sequences were labeled by the camera wearer himself and the annotation process showed them consecutive frames instead of single frames to be labeled.

In order to test generalization capabilities, we divided the data into training and the two testing split sets, i.e. Generic and Personalized. These testing splits contain full-day sequences not present in the training split, and their data percentage for the unseen and seen users was around 5% and 10%, respectively. In contrast to [6], we discarded the categories that were only performed by one participant, as the model would probably overfit that category. Moreover, we also removed the categories that had less than 200 instances, since we considered that they had a few instances for training a convolutional model. This resulted in a total number of 24 categories. The number of training, seen and unseen test sequences are 91, 15, and 19, respectively. The resulting histogram of the number of photos per category and split is shown in Fig. 3.

4.2 Implementation

In order to measure the performance of our proposed pipeline for the definition of a robust activity classification model, we perform an ablation study. To this end, we trained the models CNN+RF+LSTM[21] and CNN+LSTM[27], and their bidirectional versions using daily sequences of egocentric images. For comparative purposes, we used as a baseline to train all temporal models the Xception network [28], and left its convolutional layers frozen.

Static-image level The following two models were trained for static-image level classifiers training:

Table 1. Classification performance of the proposed model and the defined baseline models. We present results when the users in the test sets have been seen during training (personalized), when were hidden (generic), and their overall results. The best result is shown in bold, and the best result other than the groundtruth boundaries is highlighted in blue.

		STILL-IMAGE LEVEL		IMAGE-SEQUENCE LEVEL															
		CNN	CNN +RF	CNN+RF+LSTM				CNN+RF+BLSTM				CNN+LSTM				CNN+BLSTM			
	Measure	Xception	Avg. pool+pred.	No segmentation	CES [32] segmentation	Our segmentation	GT Boundaries	No segmentation	CES [32] segmentation	Our segmentation	GT Boundaries	No segmentation	CES [32] segmentation	Our segmentation	GT Boundaries	No segmentation	CES [32] segmentation	Our segmentation	GT Boundaries
Personalized	Accuracy	79.27	82.09	82.54	81.62	81.93	82.64	80.63	80.72	80.89	81.30	83.71	85.73	83.01	**86.32**	84.27	83.63	84.26	83.74
	mAP	58.08	60.22	53.41	45.93	46.64	51.63	48.91	49.09	48.89	53.93	67.33	67.55	65.39	**70.11**	67.72	67.48	67.83	64.50
	Macro precision	54.74	**64.67**	58.36	47.19	47.31	56.67	52.29	48.79	48.90	53.74	59.33	**64.12**	59.13	61.95	60.78	62.62	62.73	59.55
	Macro recall	51.23	45.87	46.07	44.20	42.50	48.88	37.13	36.81	37.16	41.95	62.44	**66.81**	62.84	61.52	63.72	62.48	64.52	62.10
Generic	Accuracy	72.56	74.76	73.79	71.91	72.59	75.16	65.22	66.64	66.40	70.62	82.21	81.92	80.07	**83.07**	79.21	78.31	78.49	77.92
	mAP	47.80	55.59	47.07	41.74	46.76	53.97	40.92	40.51	42.92	43.59	61.77	57.78	62.58	**67.25**	59.97	61.75	60.87	56.19
	Macro precision	44.10	52.21	48.40	36.10	38.69	44.41	31.87	31.91	31.87	47.01	59.69	51.15	55.55	**63.77**	57.60	57.16	53.01	53.36
	Macro recall	43.61	41.62	40.10	35.67	36.58	54.85	32.32	32.41	33.11	38.40	55.22	53.11	52.45	**60.57**	54.53	52.86	50.62	52.50
Both	Accuracy	77.43	80.08	80.14	78.95	79.37	80.58	76.40	76.86	76.91	78.37	83.30	84.69	82.20	**85.43**	82.88	82.17	82.68	82.14
	mAP	51.39	55.96	48.33	42.12	44.18	48.67	43.95	43.84	43.98	47.24	64.17	61.65	63.61	**66.77**	63.02	62.97	63.21	59.52
	Macro precision	54.63	63.98	59.77	43.60	47.31	51.98	48.49	45.33	45.27	55.22	66.64	62.16	63.01	**68.44**	63.45	65.06	63.99	60.96
	Macro recall	48.00	41.71	41.70	39.78	40.19	49.75	33.20	32.98	33.22	38.01	61.00	59.68	59.29	**61.39**	56.40	55.25	56.98	56.39

CNN We used Xception [28] as backbone CNN and we replaced the top layer with a fully-connected layer of 24 outputs. The fine-tuning procedure used Stochastic Gradient Descent (SGD) and a class-weighting scheme based on [29] to handle class imbalance. The CNN initially used the weights of a pre-trained network on ImageNet [30] that was fine-tuned. During the first 2 epochs, only the fully connected layers were optimized using a learning rate $\alpha = 1 \times 10^{-1}$, a momentum $\mu = 0.9$, and a weight decay equal to $\alpha = 5 \times 10^{-6}$. For the last epoch, the last 2 separable convolutional layers from the exit flow were also fine-tuned and the learning rate changed to $\alpha = 1 \times 10^{-3}$. In addition, the data augmentation consisted of randomly applying horizontal flips, translation and rotation shifts, and zoom operations at the frame level.

CNN+RF One random forest (RF) having a different number of trees (100, 200, . . ., 500) was trained using output layers from Xception network. Specifically, the RF was trained using as input the features extracted from the average pooling (avg. pooling) and fully-connected (FC) layers. The random forest used the Gini impurity criterion [31]. The best configuration resulted in using a number of trees equal to 200.

Image-sequence level The image-sequence level models took into account temporal information and used as a backbone the previously trained models. Our event boundaries were segmented using the SR-Clustering [16]. In order to measure the importance of the temporal information in the models, we used boundaries from three other settings. As a lower bound, the first setting considered the full-day sequence (no segmentation). The second setting used event boundaries segmented using the contextual event seg-

Table 2. Mean average precision as the overlap of different IoU thresholds θ. We present results when the users in the test sets have been seen during training (personalized), when were hidden (generic), and their overall results. The best result is shown in bold, and the best result other than the groundtruth boundaries is highlighted in blue.

		STILL-IMAGE LEVEL		IMAGE-SEQUENCE LEVEL															
		CNN	CNN +RF	CNN+RF+LSTM				CNN+RF+BLSTM				CNN+LSTM				CNN+BLSTM			
	θ	Xception	Avg. pool+pred.	No segmentation	CES [32] segmentation	Our segmentation	GT Boundaries	No segmentation	CES [32] segmentation	Our segmentation	GT Boundaries	No segmentation	CES [32] segmentation	Our segmentation	GT Boundaries	No segmentation	CES [32] segmentation	Our segmentation	GT Boundaries
Personalized	0.5	07.49	08.37	10.20	09.23	08.66	21.11	07.61	08.55	07.71	09.40	13.28	12.36	11.70	**24.14**	17.21	14.99	15.55	14.11
	0.4	08.49	09.90	11.99	11.77	10.35	22.26	10.33	10.71	10.78	11.13	14.62	14.35	12.70	**24.80**	18.03	17.10	16.47	15.63
	0.3	09.53	11.10	13.98	12.39	12.26	22.83	11.12	11.79	11.86	13.28	16.01	16.71	14.02	**25.83**	19.51	18.59	18.40	16.63
	0.2	11.60	12.69	15.13	14.35	14.27	23.04	12.68	13.54	12.67	14.31	18.36	18.96	15.34	**27.68**	21.11	20.45	20.13	17.69
	0.1	12.04	13.71	17.68	16.47	16.93	23.17	14.80	14.67	14.63	15.58	21.31	21.41	18.42	**28.15**	21.96	21.39	20.87	19.07
Generic	0.5	04.52	05.88	12.30	08.72	11.98	**32.73**	06.73	06.34	06.49	05.65	15.21	15.74	15.05	27.34	12.66	12.93	18.27	14.80
	0.4	06.01	07.55	13.26	12.20	13.08	**34.25**	07.57	07.03	07.27	07.11	19.71	18.02	18.39	30.34	18.36	18.61	19.72	16.89
	0.3	07.85	08.49	16.76	13.63	14.62	**34.41**	08.07	07.73	07.99	08.28	24.77	21.95	21.76	33.28	23.42	22.69	24.30	19.56
	0.2	08.71	09.06	17.85	16.56	17.96	**35.02**	09.10	09.01	09.31	08.77	27.78	25.40	23.84	33.74	26.30	24.40	25.41	21.12
	0.1	09.78	10.65	18.49	18.22	18.97	35.02	11.19	09.76	09.87	12.15	33.85	26.15	25.68	**35.16**	26.87	26.06	27.37	22.47
Both	0.5	04.50	05.83	10.25	08.21	08.14	21.23	06.35	07.03	06.24	06.67	12.51	13.61	10.84	**22.81**	13.62	12.75	14.99	12.94
	0.4	05.86	07.33	11.62	10.72	09.23	22.40	07.72	08.01	07.79	08.26	14.46	15.44	12.76	**23.95**	15.89	15.45	15.93	14.32
	0.3	07.12	08.63	14.28	11.77	11.29	22.86	08.55	09.08	08.98	10.42	16.99	18.66	15.20	**26.20**	18.93	18.30	18.89	15.97
	0.2	08.26	09.53	15.38	13.77	13.30	23.13	09.45	10.05	09.85	11.10	19.64	21.09	16.88	**27.24**	20.66	19.77	20.30	17.16
	0.1	08.97	10.81	17.24	15.47	15.55	23.24	11.46	11.24	11.00	12.90	26.23	22.60	19.95	**27.89**	21.67	21.30	21.55	18.57

mentation (CES) algorithm [32] trained over the R3 dataset. As an upper bound, the last setting used the groundtruth activity boundaries.

With the purpose of making a fair comparison, the weights and outputs of the backbone models were frozen during training. All the day and event photo-stream sequences were considered as full sequences during training. All the models were trained using the SGD optimization algorithm using different learning rates, but the same momentum $\mu = 0.9$, weight decay equal to $\alpha = 5 \times 10^{-6}$, batch size of 1, and a timestep of 5.

CNN+LSTM and CNN+BLSTM Both models removed the top layer of the Xception network and respectively added an LSTM and BLSTM layer having 256 units, followed by a fully-connected layer of 24 outputs. For both models, the learning rates were $\alpha = 1 \times 10^{-2}$ and $\alpha = 1 \times 10^{-3}$, respectively.

CNN+RF+LSTM and CNN+RF+BLSTM These models were trained using as input the prediction of the *CNN+RF* model. Both models added an LSTM and BLSTM layer having 30 units, followed by a fully-connected layer of 24 outputs. The learning rate for both models was $\alpha = 1 \times 10^{-3}$.

4.3 Evaluation Metrics

We considered that the sequential classification of frames from a photo-stream can be seen as an action recognition and detection tasks. Therefore, we used a specific set of

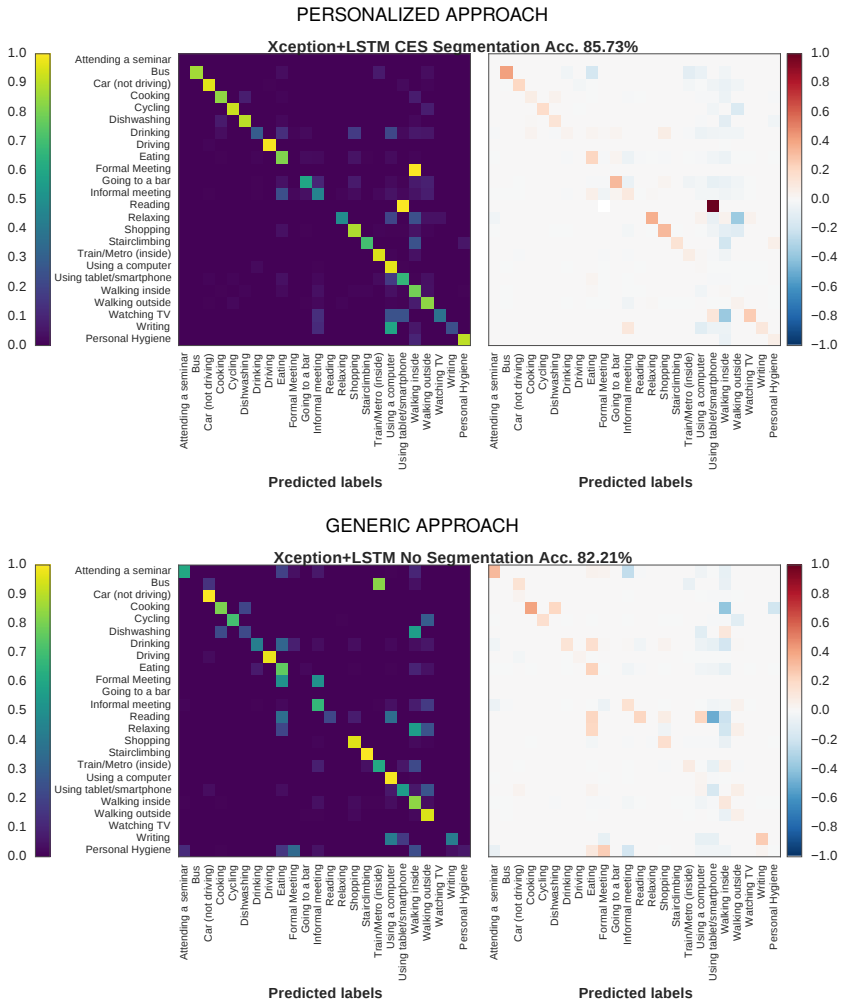

Figure 4. Normalized confusion matrices of the best models for the seen and unseen test sets and their difference with respect to the CNN model. The increase and decrease of confidence is represented by the intensity of red and blue colors.

metrics for each task. In the case of action recognition, we considered that using only the accuracy for measuring the model performance would be misleading as the testing splits are highly imbalanced. Therefore, we also used mean average precision (mAP) and other macro metrics for precision and recall as in [6]. In the case of action detection, we measured the mAP as the overlap of intersection over the union (IoU) with different thresholds as defined in [33].

Since the event segmentation clusters contextual consecutive images and not activity boundaries, we measured their homogeneity and completeness and summarized them using the V-measure. We also used the adjusted Rand index (ARI) to measure how close to the real activity segments are.

5 Results

In Tables 1 and 2, we present the classification performance and mAP overlap of IoU for all the static and temporal models, respectively. The next subsections discuss the results in detail.

5.1 Generic vs Personalized learning

Since the test categories have different proportions in the personalized and generic users splits, a straight performance comparison per category between them cannot be made. Nevertheless, the temporal models can be compared with respect to their static models, as illustrated in Fig. 4. It shows the confusion matrices for the best temporal models of each test split and their difference with respect to the CNN model. It can be observed a low performance for the categories *Formal Meeting*, *Cooking*, and *Relaxing*. They might be due to the large intra-class variability of the category (*Relaxing*), the social context ambiguity (*Formal* and *Informal meeting*), and to the fact that same activities occur on very similar places (*Cooking* and *Dishwashing*). Additionally, the performance of these classes does not increase even using temporal methods. A comparison of all the models over the test set is presented in Table 1. It shows that the best recall scores are obtained using fully deep temporal models (CNN+LSTM and CNN+BLSTM).

5.2 Random forest based models vs. deep models

The results show that the deep models have a better and more robust performance than the RF based models. Although the CNN+RF improved the overall accuracy and precision of the Xception network, it decreased the rest of the evaluated macro metrics. Particularly, the recall decreased in both test splits, thus it missed a large number of test images. Furthermore, its temporal models CNN+RF+LSTM and CNN+RF+BLSTM performed consistently worse in both splits. This contrasts the results previously obtained in [21] using another dataset and it is likely due to the fact that here we are using unseen users in our test set.

5.3 LSTM vs. BLSTM temporal models

The RF based and fully deep temporal models have contrasting results, as seen in Table 1. In the case of the RF based models, the bidirectional models performed worse than the unidirectional counterparts and even than its baseline model CNN+RF, thus indicating that the RF output did not provide enough information for generalization. In the case of the fully deep temporal models (CNN+LSTM and CNN+BLSTM), they showed improvement in all metrics and were the best in general terms. The unidirectional models achieved the highest classification accuracy in both test splits and the rest of the metrics showed a more solid classification, specifically achieving an accuracy of 85.73% and 82.21% for the Personalized and Generic test sets, correspondingly.

Figure 5. Example of qualitative results obtained from personalize and generic test sets. Each row shows the first 12 frames of segmented events having only one activity. Full-day and Events refers to no-segmentation and SR-Clustering from photo-streams, correspondingly. False and true activity labels are marked in **red** and **green**, respectively.

5.4 Event clustering vs. No segmentation

We present the results of event segmentation with respect to the groundtruth activity boundaries in Table 3. They show that the event segments obtained with the CES algorithm have slightly less mixed categories with a V-measure equal to 0.936. Additionally, it shows that its clusters are closer to actual activity boundaries by having an ARI of 0.679. Their similarity of clustering performance also was reflected in achieving similar classification scores. Nevertheless, the CES algorithm obtained better performance than our segmentation algorithm for the best model CNN+LSTM (84.69% accuracy for both test splits).

The overall performance shows that the most robust performance is achieved using the CNN+LSTM with no segmentation, as shown in Table 1. Nonetheless, the results indicate that the events segmentation helps the classification when the users have been previously seen during training, and especially when they are used in conjunction with unidirectional LSTMs (achieving 85.73% accuracy as seen in Table 1). Moreover, their effect over the CNN+LSTM consistently has a better performance for higher IoU thresholds ($\theta \in \{0.2 - 0.5\}$) in both test splits, as seen in Table 2. The results show that bidirectional models are not benefited from activity boundaries provided by the groundtruth, as seen from the results presented in Table 2. Specifically, having no segmentation at all for the CNN+BLSTM model is better than the groundtruth. This might

Table 3. Activity events clustering performance. The best result is shown in bold.

Method	Homogeneity	Completeness	V-measure	Adjusted Rand index
SRclustering [16]	0.897	0.953	0.924	0.620
CES[32]	**0.905**	**0.969**	**0.936**	**0.679**

be explained as the BLSTM having a smoothing effect over mixed input categories. As a classification example over sequences, Fig. 5 shows some qualitative results.

6 Conclusions

This paper addresses the effect of event boundaries for activity recognition in egocentric photo-streams, a poorly investigated topic. By using a recently published egocentric dataset acquired from 15 users, our contributions are the following. First, we propose an event-based architecture for a robust activity recognition in photo-streams. This architecture automatically segments egocentric photo-streams into subsequences with similar contextual and semantic features. These segments define the training and testing event boundaries for a CNN+BLSTM model. Second, in order to determine the effect of event boundaries, we provide a rich ablation study of different state-of-the-art methods comparing them with groundtruth boundaries and their lack of. Additionally, these tests thoroughly evaluate the generalization capabilities of these methods on a *generic* and a *personalized* training. Our results show that event boundaries benefit activity recognition performance of the CNN+LSTM when tested on users previously seen during training, thus achieving a classification accuracy of 85.73%. Moreover, our tests show that actual activity segments in the photo-stream are better classified using event boundaries for higher IoU thresholds ($\theta \in \{0.2 - 0.5\}$) using this architecture. The results also point that event boundaries make more robust the activity classification and detection performance of the CNN+BLSTM than not using them, but their improvement is not as good as its unidirectional counterpart. Finally, the results also indicate that event boundaries improved the detection of activity segments for temporal RF based architectures but failed to improve their classification baseline. Since activity recognition from egocentric photo-streams could be posed as a multi-classification problem, future research lines should consider the context ambiguity. For example, a person might be *eating* something while *reading* a book inside a *train*.

Acknowledgment

This work was partially funded by projects RTI2018-095232-B-C2, SGR 1742, CERCA, Nestore Horizon2020 SC1-PM-15-2017 (n 769643), Validithi EIT Health Program, and the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (MINECO/ERDF, EU) through the program Ramon y Cajal, the national Spanish project PID2019-110977GA-I00 and the Spanish national network

RED2018-102511-T. A. Cartas supported by a doctoral fellowship from the Mexican Council of Science and Technology (CONACYT) (grant-no. 366596). The authors acknowledge the support of NVIDIA Corporation for hardware donation.

References

1. A. Cartas, J. Luque, P. Radeva, C. Segura, and M. Dimiccoli, "Seeing and hearing egocentric actions: How much can we learn?" in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
2. L. Chen and C. D. Nugent, *Human Activity Recognition and Behaviour Analysis*. Springer, 2019.
3. R. de Jong, "Multimodal deep learning for the classification of human activity: Radar and video data fusion for the classification of human activity," 2019.
4. M. Bolaños, M. Dimiccoli, and P. Radeva, "Toward storytelling from visual lifelogging: An overview," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 77–90, 2017.
5. M. Aghaei, M. Dimiccoli, and P. Radeva, "All the people around me: face discovery in egocentric photo-streams," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1342–1346.
6. A. Cartas, P. Radeva, and M. Dimiccoli, "Activities of daily living monitoring via a wearable camera: Toward real-world applications," *IEEE Access*, vol. 8, pp. 77 344–77 363, 2020.
7. D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa, "Predicting daily activities from egocentric images using deep learning," pp. 75–82, 2015.
8. A. Cartas, M. Dimiccoli, and P. Radeva, "Batch-based activity recognition from egocentric photo-streams," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2347–2354.
9. A. Cartas, J. Marín, P. Radeva, and M. Dimiccoli, "Recognizing activities of daily living from egocentric images," *Pattern Recognition and Image Analysis*, pp. 87–95, 2017.
10. M. Aghaei, M. Dimiccoli, C. C. Ferrer, and P. Radeva, "Towards social pattern characterization in egocentric photo-streams," *Computer Vision and Image Understanding*, vol. 171, pp. 104–117, 2018.
11. E. S. Aimar, P. Radeva, and M. Dimiccoli, "Social relation recognition in egocentric photostreams," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3227–3231.
12. E. Talavera, M. Leyva-Vallina, M. K. Sarker, D. Puig, N. Petkov, and P. Radeva, "Hierarchical approach to classify food scenes in egocentric photo-streams," *IEEE journal of biomedical and health informatics*, 2019.
13. E. Talavera, C. Wuerich, N. Petkov, and P. Radeva, "Topic modelling for routine discovery from egocentric photo-streams," *Pattern Recognition*, p. 107330, 2020.
14. Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2537–2544, 2014.
15. A. Furnari, G. M. Farinella, and S. Battiato, "Temporal segmentation of egocentric videos to highlight personal locations of interest," *European Conference on Computer Vision(ECCV)*, pp. 474–489, 2016.
16. M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, S. G. Nikolov, and P. Radeva, "Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation," *Computer Vision and Image Understanding*, 2017.
17. C. Dias and M. Dimiccoli, "Learning event representations by encoding the temporal context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

18. H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2847–2854.
19. S. Sudhakaran and O. Lanz, "Attention is all we need: Nailing down object-centric attention for egocentric activity recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.
20. G. García Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
21. A. Cartas, J. Marín, P. Radeva, and M. Dimiccoli, "Batch-based activity recognition from egocentric photo-streams revisited," *Pattern Analysis and Applications*, May 2018. [Online]. Available: https://doi.org/10.1007/s10044-018-0708-1
22. H. Yu, W. Jia, Z. Li, F. Gong, D. Yuan, H. Zhang, and M. Sun, "A multisource fusion framework driven by user-defined knowledge for egocentric activity recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, no. 1, p. 14, 2019. [Online]. Available: https://doi.org/10.1186/s13634-019-0612-x
23. H. Yu, G. Pan, M. Pan, C. Li, W. Jia, L. Zhang, and M. Sun, "A hierarchical deep fusion framework for egocentric activity recognition using a wearable hybrid sensor system," *Sensors*, vol. 19, no. 3, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/3/546
24. S. Song, V. Chandrasekhar, B. Mandal, L. Li, J.-H. Lim, G. Sateesh Babu, P. Phyo San, and N.-M. Cheung, "Multimodal multi-stream deep learning for egocentric activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 24–31.
25. A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602 – 610, 2005, iJCNN 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608005001206
26. F. Chollet, *Deep Learning with Python*, 1st ed. Greenwich, CT, USA: Manning Publications Co., 2017, pp. 219–221.
27. J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Computer Vision and Pattern Recognition*, 2015.
28. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," pp. 1800–1807, 2017.
29. G. King and L. Zeng, "Logistic regression in rare events data," *Political analysis*, vol. 9, no. 2, pp. 137–163, 2001.
30. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.
31. L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
32. A. Garcia del Molino, J.-H. Lim, and A.-H. Tan, "Predicting Visual Context for Unsupervised Event Segmentation in Continuous Photo-streams," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 10–17.
33. Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.