

Predicting Visual Context for Unsupervised Event Segmentation in Continuous Photo-streams

Ana García del Molino*

School of Computer Science and
Engineering, Nanyang Technological
University, Singapore
ana002@e.ntu.edu.sg

Joo-Hwee Lim

Institute for Infocomm Research,
A*STAR, Singapore
joohwee@i2r.a-star.edu.sg

Ah-Hwee Tan

School of Computer Science and
Engineering, Nanyang Technological
University, Singapore
asahtan@ntu.edu.sg

ABSTRACT

Segmenting video content into events provides semantic structures for indexing, retrieval, and summarization. Since motion cues are not available in continuous photo-streams, and annotations in lifelogging are scarce and costly, the frames are usually clustered into events by comparing the visual features between them in an unsupervised way. However, such methodologies are ineffective to deal with heterogeneous events, e.g. taking a walk, and temporary changes in the sight direction, e.g. at a meeting. To address these limitations, we propose Contextual Event Segmentation (CES), a novel segmentation paradigm that uses an LSTM-based generative network to model the photo-stream sequences, predict their visual context, and track their evolution. CES decides whether a frame is an event boundary by comparing the visual context generated from the frames in the past, to the visual context predicted from the future. We implemented CES on a new and massive lifelogging dataset consisting of more than 1.5 million images spanning over 1,723 days. Experiments on the popular EDUB-Seg dataset show that our model outperforms the state-of-the-art by over 16% in f-measure. Furthermore, CES' performance is only 3 points below that of human annotators.

KEYWORDS

Lifelogging; Event Segmentation; Visual Context Prediction

ACM Reference Format:

Ana García del Molino, Joo-Hwee Lim, and Ah-Hwee Tan. 2018. Predicting Visual Context for Unsupervised Event Segmentation in Continuous Photo-streams. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3240508.3240624>

*Also with Institute for Infocomm Research, A*STAR, Singapore.

This work is supported by A*STAR JCO Grant 1335h00098 (REVIVE), IAF-ICP Grant ICP160003 (VInspection), and A*STAR Singapore International Graduate Award (SINGA).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5665-7/18/10...\$15.00

<https://doi.org/10.1145/3240508.3240624>



Figure 1: As humans, we define a new event when the new sequence of frames differs from our understanding of the previous frame sequence. CES models such intuitive framework of perceptual reasoning by predicting the visual context of the photo-stream. At each timestep, it compares the context predicted from the past sequence to the context predicted from the future sequence.

1 INTRODUCTION

Continuously recording our lives in the form of images can be of great usefulness for memory enhancement, tracking of the activities of daily living, and other related healthcare applications. However, lifelogging has an overload problem both in time and space: Lifelogging cameras take a minimum of 2 pictures per minute, which can add to more than 1,000 pictures a day, i.e. 100Gb per year. Such vast load of data requires hours of manual analysis to, for example, select your day's highlights, check what you ate and drank the past month, or monitor your grandparent's routines. Hence, automatic tools to extract highlights and life patterns are needed [18, 21, 22, 39]. However, analyzing lifelogs entails two great challenges related to its Low Time Resolution (LTR) and wearable nature: First, dramatic visual changes between consecutive frames even if these correspond to the same event. Second, a substantial presence of visual occlusions, walls and ceilings in the field of view, and frequent changes of visual orientations.

Extensive research has been conducted to retrieve specific events or obtain summaries from First Person View (FPV) videos [5, 10]. While event segmentation is needed for a complete, informative and diverse summary that includes most life events in the recording, little work has been done to that effect [5, 19]. Many approaches to segment High Time Resolution (HTR) video use motion between frames to infer the wearer's activity, but motion cues are not available in photo-streams. Furthermore, obtaining annotations for such large datasets is very costly. As such, one can only resort to visual

features and sensor metadata, and unsupervised techniques such as K-Means [8] and probabilistic models [11].

Due to these limitations, current automated methods usually fail at modeling the frame sequences. As a consequence, they cannot perceive the overall context in heterogeneous events, and usually misinterpret occlusions and occasional diversions within events as different episodes. Our ambition is to build a segmentation model that mimics the human reasoning, as people can easily detect and discard such noise by comparing the new visual input with their understanding of both the previous and following scene (see Fig. 1).

In this work, we introduce Contextual Event Segmentation (CES), a novel event segmentation technique that, given a sequence of frames, predicts its visual context and then compares it to the context corresponding to the ensuing sequence. An LSTM-based generative model, that we call VCP, is used to predict the visual context. It is able to model our daily activities and learn the associations between different scenes, e.g. a train commute will include corridors, stairs, a platform, the interior of a wagon, etc. To train VCP we introduce *R3*, a novel and vast dataset for unsupervised lifelog analysis. It consists of over 1,5 million images that depict the daily activities of 57 different users over a total of 1,723 days.

The main contributions of this paper are:

- (i) a segmentation approach that mirrors the human perceptual reasoning when segmenting photo-streams into events. In a series of experiments, CES proves to be superior to the state of the art by over 16% in f-measure, and even competitive against manual annotations.
- (ii) an LSTM-based generative model to predict visual context from a sequence of frames. We observe that the model learns event traits in common daily activities.
- (iii) a large-scale lifelogging dataset containing 1,500,890 images from 57 users. To our knowledge, R3 is the largest FPV dataset currently available¹.

2 RELATED WORK

FPV content entails three main challenges: its unconstrained nature, its continuous stream of consecutive events, and its poor visual quality. In particular, the purpose of lifelogging is to have a diary of our lives. However, such huge amount of visual content must be summarized to be of practical use. The summary of these photo-streams should be complete, informative and diverse. When no query is given to constraint the content of the summary, the maximum variety of events should be included. To do so, the content must first be segmented into *subshots* in the case of High Temporal Resolution (HTR) videos, or *events* in the case of Low Temporal Resolution (LTR) videos (or photostreams).

Temporal segmentation in High Temporal Resolution First Person View. Third Person View (TPV) event segmentation approaches typically identify shot boundaries by detecting abrupt changes between consecutive frames [23, 30]. However, FPV content is not comprised of separate shots, but rather a succession of events with smooth transitions, where event boundaries are not well defined.

Most FPV approaches for event segmentation use motion cues, both visual (e.g. optical flow, blurriness) [2, 9, 20, 29, 31–33, 38] and

from sensors [2, 35]. Such features are used to predict the wearer’s activity or attitude patterns using probabilistic models [38] and deep learning [33], to segment the videos accordingly. Other methods resort to visual similarity between groups of frames (e.g. color, GIST, CNN hash) [2, 3, 24, 25, 29, 31, 34, 39, 40]. Temporally constrained clustering [25] and statistical frameworks [34] have been used to determine whether the visual differences correspond to event boundaries or just abrupt head movements.

Temporal segmentation in Low Temporal Resolution First Person View. In the case of lifelog photo-streams, frames can be up to 30 seconds apart. In such low temporal resolution, content may change a lot between consecutive frames even if they are part of the same event, and as Bolanos *et al.* remark in [5], visual motion information is unavailable (sensor information may sometimes be available [13]). Given the limited amount of annotated data, event segmentation is very often unsupervised, performed via K-Means and other hierarchical clustering algorithms on visual cues (e.g. color, CNN hashes) [6, 12, 13, 17, 26, 27, 41]. An exception to these unsupervised methods is [15], in which a personal location classifier is trained for each user, and events are segmented according to changes in the wearer’s location. Since these methods often ignore the semantic nature of the frames, Dimiccoli *et al.* [11] propose defining the frames with semantic and contextual cues defined by CNN features and linguistic information. The relation between frames is assessed using a WordNet [1] based knowledge graph, and the event boundaries are found using a graph-cut algorithm integrating an agglomerative clustering. Such segmentation methodology relies on the cross-analysis of consecutive frames, and cannot detect change points between two events with heterogeneous visual content, nor ignore small and isolated visual changes within an event.

To address this limitation, we present a novel event segmentation paradigm in which each frame is understood as part of a global sequence. As such, the visual context of the upcoming frame can be predicted from the preceding sequence of frames. This prevents the model from detecting false positives due to abrupt changes between consecutive frames, and allows it to understand the nature of heterogeneous events.

Sequence embedding for photo-album summarization and activity classification. Addressing the problem of story-telling from albums of 10 to 50 photos, Yu *et al.* [42] use a Recurrent Neural Net (RNN) to encode the local album context for each photo, so that the best key-frames can be selected. Liu *et al.* [28] use Gated Recurrent Units (GRUs) to align the local storylines into the global sequential timeline. To obtain better event descriptions, they further leverage the semantic coherence in a photo stream by jointly embedding the images and sentences into a common semantic space.

Using video content, Bhatnagar *et al.* [4] obtain good results at describing egocentric motor actions (e.g. *stir*, *fold*, *open*) in HTR videos using an hybrid CNN-LSTM auto-encoder. Similarly, Srivastava *et al.* [36] learn spatio-temporal features using a sequence-to-sequence future prediction model, proving that such an architecture is more efficient than an auto-encoder.

Whereas both [4, 36] learn the spatio-temporal features from the raw frames in HTR video, we propose learning a global semantic visual context from the visual features of LTR frame sequences.

¹The data is accessible from <http://dx.doi.org/10.17632/ktps5my69g.1>

3 CONTEXTUAL EVENT SEGMENTATION

3.1 Overview

Given a continuous stream of photos, we, as humans, would identify the start of an event if the new frame differs from our expectation of what should follow the preceding sequence. We would also check whether that frame is consistent with the subsequent image sequence (or scene). If the new scene spans a very short time and returns to the previous, we would ignore it as an extra event, but rather wrap it within the current event (*e.g.* going for a bottle of water while watching TV). Therefore, we would frequently look forward and backward to verify whether it was a new event, or just a brief diversion or local outlier.

The proposed model is analogous to such intuitive framework of perceptual reasoning. It uses an encoder-decoder architecture to predict the visual context at time t given the images seen before, *i.e.* the past. A second visual context is predicted from the ensuing frames, *i.e.* the future. If the two predicted visual contexts differ greatly, CES will infer that the two sequences (past and future) correspond to different events, and will consider frame_t as a candidate event boundary.

Therefore, CES consists of two modules (*c.f.* Algorithm 1): First, the Visual Context Predictor (VCP), that predicts the visual context of the upcoming frame, either in the past or in the future depending on the sequence ordering. Second, the event boundary detector, that compares the visual context at each time-step given the frame sequence from the past, with the visual context given the sequence in the future.

3.2 Visual Context Predictor

Inspired by [4, 36, 42], we propose predicting the visual context from a sequence of frames with a Long-Short Term Memory network. LSTM networks are a type of Recurrent Neural Network that learn long-time dependencies through four hidden layers, *i.e.* the gates. Thus, LSTMs can aggregate the information they receive by learning to forget. Their mathematical formulation can be expressed as

$$\begin{aligned} \begin{pmatrix} \underline{i} \\ \underline{f} \\ \underline{o} \\ \underline{g} \end{pmatrix} &= \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} \left(\begin{bmatrix} \mathbf{x}_t \\ \mathbf{h}_{t-1} \end{bmatrix} \mathbf{W} \right) \\ \mathbf{c}_t &= \underline{f} \circ \mathbf{c}_{t-1} + \underline{i} \circ \underline{g} \\ \mathbf{h}_t &= \underline{o} \circ \tanh(\mathbf{c}_t), \end{aligned} \quad (1)$$

where \underline{i} , \underline{f} , \underline{o} , and \underline{g} correspond to the four gates of the unit (*input*, *forget*, *output*, and *input modulation*), \circ is the element-wise product, \mathbf{W} are the network weights [16], and \mathbf{c}_t and \mathbf{h}_t are the cell state and the hidden state, respectively, at time-step t .

The sequential and relational nature of lifelogging photo-streams allows us to train the weights of an LSTM-based aggregation network without ground truth annotations. To obtain the weights of our Visual Context Predictor, we train an encoder-decoder architecture that, given a sequence of visual feature vectors, learns to predict the subsequent sequence, as shown in Fig. 2. Since LTR video frames are visually highly different from adjacent ones, the

Algorithm 1: Overview of Contextual Event Segmentation

▷ Get past and future context from the Visual Context Predictor:

$$\begin{aligned} \mathbf{rf}(t-1) &\leftarrow \text{predicted from } [\mathbf{x}_k]_{\forall 0 \leq k < t} \\ \mathbf{rp}(t+1) &\leftarrow \text{predicted from } [\mathbf{x}_k]_{\forall \text{len}[\mathbf{x}] \geq k > t} \end{aligned}$$

▷ Detect boundary candidates:

$$\begin{aligned} \text{pred}(t) &= \text{cos_dist}(\mathbf{rf}(t-1), \mathbf{rp}(t+1)) \\ b &= \{t \mid (\frac{\delta \text{pred}}{\delta t}) = 0\} \end{aligned}$$

▷ Remove noisy candidates:

$$b = \{b_k \mid \text{pred}(b_k) \leq \text{average}(\text{pred}(b))\}$$

model will learn the general context of the event at the same time as the estimation of the visual feature of the upcoming frame.

The auto-encoder is defined as

$$\begin{aligned} \mathbf{r}_t &= \mathbf{h}_{t,\text{encoder}}(\mathbf{x}_t) \\ \hat{\mathbf{x}}_{t+1} &= \mathbf{h}_{t,\text{decoder}}(\mathbf{r}_t), \end{aligned} \quad (2)$$

where \mathbf{x}_t is the deep learned visual feature (*c.f.* Section 5.1) of frame t , \mathbf{r}_t is the predicted visual context at time t , and $\mathbf{h}_{t,\text{encoder}}$ and $\mathbf{h}_{t,\text{decoder}}$ correspond to the models trained to encode and decode the visual feature, respectively. The objective function of the learning process is to minimize the mean squared error of the prediction, *i.e.* $\text{mse}(\mathbf{x}_t, \hat{\mathbf{x}}_t)$.

VCP shares architecture and weights with the encoding model presented above, and is able to encode the visual context of lifelog image sequences both feed forward and backwards, *i.e.* in reverse time order. The chosen architecture for VCP (*i.e.* the encoder) is a single LSTM layer of 1024 neurons. The hidden state is then passed to the decoder, which has a corresponding LSTM layer. The pre-trained model will be made available upon publication.

3.3 Boundary detector

Given a frame \mathbf{x}_t , two different context predictions can be obtained from VCP. The first, the future context \mathbf{rf}_t including the sequence of frames from the past ($\mathbf{x}_k|_{0 \leq k < t}$). The second, the past context \mathbf{rp}_t including the frames in the future ($\mathbf{x}_k|_{T \geq k > t}$), where T is the total length of the lifelog. Thus, at each time-step t , the future context given the past will be \mathbf{rf}_{t-1} , and the past context given the future \mathbf{rp}_{t+1} . Note that the frame \mathbf{x}_t is not seen when predicting the future and past context at time t to avoid overlapping inputs in the prediction.

An event boundary will delimit sequences with very different visual context. Hence, the boundary prediction function is defined

$$\text{pred}(t) = d(\mathbf{rf}_{t-1}, \mathbf{rp}_{t+1}), \quad (3)$$

where $d(\cdot, \cdot)$ is the cosinus distance.

The larger the distance between the two predicted visual contexts, the more likely the upcoming frame will correspond to an event boundary. Since the visual context will change gradually within the vicinity of a boundary, boundary candidates are assigned to the local maxima. Local maximums will also be found for very slight changes in the visual context. Therefore, only the candidates whose prediction value is over the average candidate values are kept as final event boundaries.

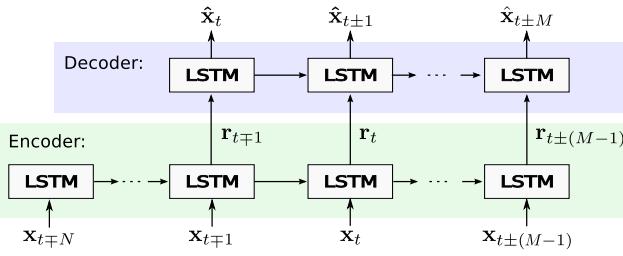


Figure 2: Training of the Visual Context Predictor. Given a sequence of features, the model learns to predict the visual feature of the following frame, either in the future or, if the sequence is in reverse order, in the past. The output of the encoder, r_t , corresponds to the visual context at time-step t .

4 R3 DATASET

A large-scale FPV dataset is needed to train the Visual Context Predictor. Such dataset must consist of continuous LTR streams of images spanning at least a few hours, without the need for any annotation. However, the size of the publicly available LTR datasets is very limited: 170 days in CLEF [8] and NTCIR [19], and 66 in EDUB-Seg [11] and EDUB-SegDesc [7], spanning a total of 2,700 hours and 261,845 images. We can also resort to other popular HTR FPV video datasets such as the First Person Social Interaction Dataset [14], Huji EgoSet [32], and UTEgo-centric [24], that cover 28, 15 and 16 hours, respectively. Down-sampled at $2fpm$, the accumulated length of these datasets is under 10,000 images. This amount of information results insufficient to train efficient deep learning models.

In this work we introduce $R3$, a large scale lifelogging image dataset captured by 57 users during 1,723 days for a total of almost 13,000 hours, resulting in over 1.5 million images. A comparison of the size of $R3$ with respect to the other mentioned datasets is presented in Fig. 3. The users volunteered to capture their daily lives as part of a memory-enhancement user study. They were asked to put on the wearable camera for most of their day during a whole month, and were free to withdraw from the study if they felt that wearing it was disrupting their routines. The volunteers are mostly seniors older than 50 years old, and span a wide range of occupations and lifestyles. To protect their privacy, only the extracted visual features will be released.

5 EXPERIMENTS

5.1 Data setup

The output of the pre-pooling layer of InceptionV3 [37] is used to describe the frames in the lifelog. We use the available lifelogging video data from $R3$, CLEF, NTCIR, and EDUB-Seg to train the VCP model and test our CES framework. EDUB-SegDesc [7] is reserved as validation for further supervised pruning of the prediction obtained from CES.

The datasets are used as follows:

Training of the VCP model: 75% of $R3$ is used as training set for the Visual Context Predictor model. To ensure that the model is not biased toward this dataset, a 20% of both CLEF [8] and NTCIR [19] is also included in the training set. This joined set adds

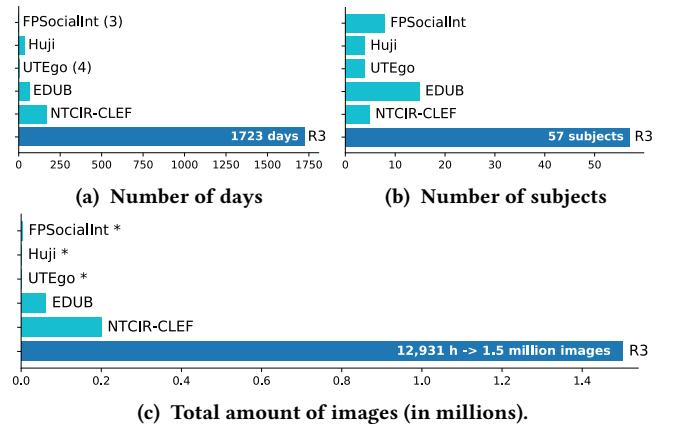


Figure 3: Comparison of $R3$ with respect to other popular FPV datasets. * HTR datasets are down-sampled at $2fpm$.

up to 1,207,483 images. A separate 5% of $R3$ is used to validate the different configurations and select the best hyperparameters.

Testing set for the VCP model: the remaining 20% of $R3$, and 80% of CLEF and NTCIR is kept as test to confirm that VCP is not overfitted toward $R3$ (c.f. 2).

Testing of the CES framework: the semantic features for 12 of the lifelogs in EDUB-Seg [11] have been made available to us. We compare our method to the baselines in two overlapping sets: these 12 lifelogs and the full 20 lifelogs in the dataset.

5.2 Training methodology

We explore several architectures and training parameters for the Visual Context Predictor model. Regarding the architecture, we can modify the number of neurons in the encoding LSTM layer, the number of frames seen before starting the future prediction (N), the amount of frames the decoder needs to predict (M), and whether the prediction will be conditional or not, i.e. whether the model gets further inputs past frame N . We investigate architectures between 256 and 1024 neurons, values of $N = M$ between 10 and 100, and the same range of M for $N = 1$.

Concerning the training parameters, the loss is defined as the mean squared error of the prediction \hat{x}_t , and RMSProp without decay is used as optimizer. The learning rate is randomly set in the range [.0001, .001], and is reduced by half after every 4 epochs without significant improvement in the validation loss. Different batch sizes are used, between 250 and 1000 sequences at a time.

The best configuration is found through a gridsearch on all the different parameters. We find that the best prediction performance (smaller validation loss) is achieved with 1024 neurons on a conditional architecture. The number of frames seen before starting the future prediction is set to $N = 10$, equal to the number of frames to predict ($M = 10$). We observe that training with longer sequences does not improve significantly the model performance (c.f. Table 2), while making the training slower. At test time, one single frame ($N = 1$) is given to start the prediction of the whole day ($M = \text{length(lifelog)} - 1$).

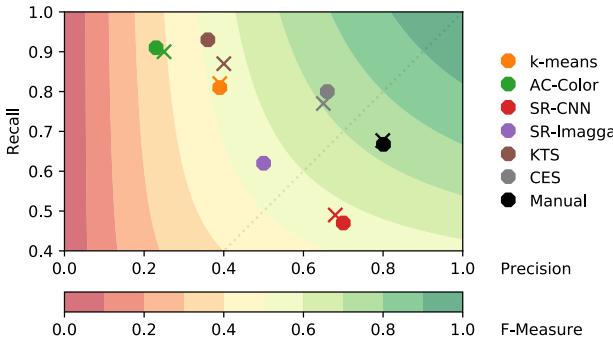


Figure 4: Precision-Recall curve for the tested unsupervised event segmentation algorithms. The corresponding F-measure score is shown in the color space. Results on the smaller set EDUB-Seg12 are represented with an octagon, whereas the larger set EDUB-Seg20 is depicted with an x. (Best viewed in color.)

Other implementation details. We also analyze the possibility of fine-tuning the boundary prediction with supervised learning. For that purpose, we train an SVM with samples from a held-out validation set (EDUB-SegDesc [7]). The SVM evaluates the boundary likeliness from cluster consistency indicators. In particular, two clusters are defined at opposite sides of the candidate boundary, containing the 15 frames that precede or follow it. The indicators used are the correlation between the two clusters, the compactness of each of them and their union, and the BetaCV and Normalized Cut scores [43].

5.3 Evaluation methodology

Following the literature, we report the averaged f-measure, precision and recall for the tested models (Table 1). For our evaluation, a detected boundary is considered a true positive if there is an element in the ground truth within a distance of tolerance, and the ground truth element is not already matched to any other detected boundary. Analogously, all elements in the ground truth for which no detected boundary is found within the tolerance are considered false negatives. This tolerance is set to 5 frames.

We compare the performance of the following baselines on the publicly available *EDUB-Seg* dataset [11]:

- **Smoothed K-Means:** the lifelog is clustered into events using k-means with a fixed $k = 30$. The clustering is then smoothed by assigning each frame to the most common cluster within a window. This operation is done iteratively until no more changes occur. As a result, some clusters may disappear.
- **AC-Color:** Agglomerative Clustering on the color feature of the frames, as done in [25].
- **SR-Clustering:** Semantic Regularized Clustering as described in [11], using only visual features (*CNN*), and also semantic cues (*Imagga*).
- **KTS:** Kernel Temporal Segmentation as described in [34].

method	EDUB-Seg12			EDUB-Seg20		
	F1	Prec.	Rec.	F1	Prec.	Rec.
K-Means smoothed	0.51	0.39	0.81	0.51	0.39	0.82
AC-Color [25]	0.36	0.23	0.91	0.38	0.25	0.90
SR-ClusteringCNN [11]	0.50	0.70	0.47	0.53	0.68	0.49
SR-ClusteringImagga [11]	0.53	0.50	0.62	-	-	-
KTS [34]	0.50	0.36	0.93	0.53	0.40	0.87
CES (with VCP)	0.70	0.66	0.80	0.69	0.66	0.77

Table 1: Comparison to the state of the art. Averaged results (F-measure, Precision and Recall) on the subset of 12 lifelogs with available semantic tags and on the full EDUB-Seg.

Bias in the Ground Truth. Since segmenting lifelogs into events can be a very subjective task, the curators of EDUB-Seg provide in [11] an extensive analysis on the uniformity among the ground truth annotated by different subjects. They conclude that visual lifelog event segmentation can be objectively evaluated, since different people (which are not the camera wearer) tend to segment the lifelogs consistently. For the purpose of our evaluation, we select the ground truth from the first annotator. We use the other annotations as a baseline. For the lifelogs that only included one annotation, we asked independent subjects to annotate the events, so that we would have at least two sets of annotations for each lifelog. We therefore report the performance of the manual annotations as an upper reference in Table 3.

Other implementation details. To find the local maximums in the prediction signal of CES, as well as smoothing the K-Means clustering, a window of size 5 is chosen, so that it is consistent with the ground truth tolerance.

5.4 Results

Table 1 presents the results of CES and the baselines in EDUB-Seg, and a smaller subset (which includes the semantic features needed for *SR-ClusteringImagga*). The position of each method in the Precision-Recall curve is shown in Fig. 4. While most methods fall within the mid-range performance in terms of f-measure, CES stands out of the baselines, improving their performance by over 15%, and positioning itself on the upper range of the absolute spectrum. The performance of CES is even competitive with that of the manual annotations.

We show in Fig. 5 the performance of CES applied to one of the tested lifelogs. We can observe that most elements in the ground truth fall on the spikes of the prediction signal, or very close to them. This confirms the suitability of using the predicted contexts as a boundary cue.

While the baselines fail at detecting boundaries between heterogeneous events, CES is capable of extracting the underlying context of each event, and discern their disparity (e.g. shopping at the supermarket after riding a bike on the street). Moreover, in cases in which the camera wearer orientation changes within a static event (e.g. looking back from your food to your colleagues), traditional segmentation methods detect such view change as an event boundary, whereas CES is able to detect the presence of a

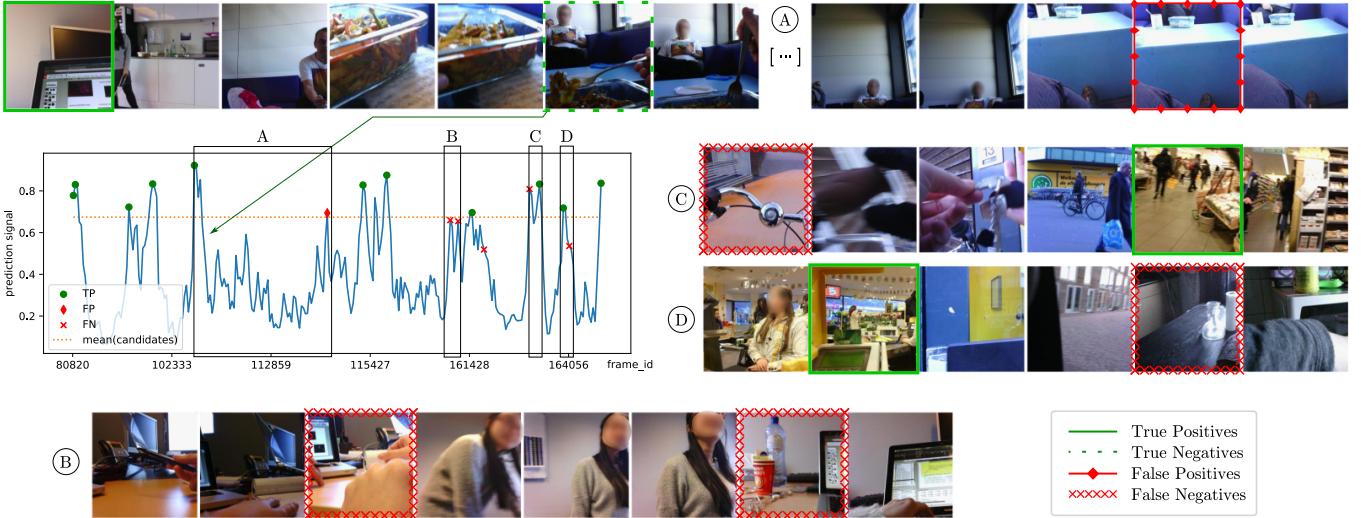


Figure 5: Qualitative example for one of the tested lifelogs. We highlight those frames which are false positives or false negatives in the baselines. We can observe that, unlike the baselines, CES is able to ignore occasional occlusions as long as the different points of view span less frames than CES’ memory span (A). It is also capable of detecting boundaries that separate heterogeneous events such as riding a bike on the street and shopping at the supermarket (C, D). Most of the boundaries not detected by CES correspond to events that take place within the same physical space (B) and short transitions (C, D), e.g. parking the bike. (Best viewed in color.)

common visual path. However, if the view change spans longer than CES memory, CES will not be able to contextualize it within the event. An example for such a situation can be seen in Figs. 6b and 6c. We also note that the ability of CES to detect the general context of the visual sequence and track common cues sometimes misleads the prediction. When the ground truth of a boundary falls within the same physical space, or similar contexts, CES does not perceive their differences, and thus does not detect the boundary. Arguably, such boundaries are also difficult to detect by external viewers. This may also occur when short transitions between events are considered events on their own.

Predicting the context vs predicting the actual frame. One could think that predicting a future frame \hat{x}_t and comparing it to the actual future frame x_t should be better than comparing the visual context. We tested this hypothesis, in which

$$pred(t) = abs(mse(x_t, \hat{x}_{f,t}) - mse(x_t, \hat{x}_{p,t})), \quad (4)$$

where $\hat{x}_{f,t}$ is predicted from $x_{k|0 \leq k < t}$ and $\hat{x}_{p,t}$ from $x_{k|T \geq k > t}$. The intuition behind this formulation is that a local outlier will be badly predicted both from the future and the past, whereas an event change will provide a good prediction only in one direction. This theory proves not precise in practice. The generative model embeds noise into the frame descriptor, and, as expected, generates samples closer to the previous (seen) frame than the (unseen) target. As such, using such a noisy signal is detrimental to the final objective. The performance of such method is reported as CES-error in Table 3.

Informativeness of the Visual Context. To validate the encoding efficiency of VCP and hence the informativeness of the visual context, we have tested CES using two alternative sequence encodings:

trained with N / M :	10 / 10	1 / 40	1 / 100	1 / 1	10 / 1
# neurons :	256 512 1024	512 1024 1024			mean*
mse future pred.:	1.058 1.030 1.024	1.03 1.029 1.028			1.58 1.054
mse past pred.:	1.059 1.029 1.024	1.03 1.029 1.028			

Table 2: Performance of the auto-encoder’s prediction at test time (mean mse amplified $\cdot 10^2$, with $N = 1$, $M = T - 1$ and $T = \text{len}[x]$) for different training configurations of VCP.

*As a reference, we include using the average of the previous N frames as the predicted feature, i.e. $\hat{x}(t) = \sum_{n=1}^N x(t-n)/N$.

first, an average of the previous $N = 10$ frames (or subsequent in the case of the past prediction); second, a PCA time-dimensionality reduction on the aforesaid set. These two variants are reported in Table 3 as CES-mean and CES-PCA, respectively.

We observe that the visual context predicted by VCP results much more informative than any of the other contextual encodings. While the averaged encoding obtains a predictive performance similar to the output of our decoder (c.f. Table 2), the encoding transformation of VCP is superior as a contextual visual feature. Moreover, unlike PCA, which takes the inputs as a set, VCP takes the inputs as a sequence, and is able to learn a more informative context descriptor.

Pruning of the candidate boundaries using supervised learning. For high recall results, false candidate boundaries can be discarded using cluster analysis between the frames that the candidate separates. Having annotated data to train a pruning model can improve the performance of the segmentation algorithm in terms of precision, having minimal impact on the recall. We tested this hypothesis



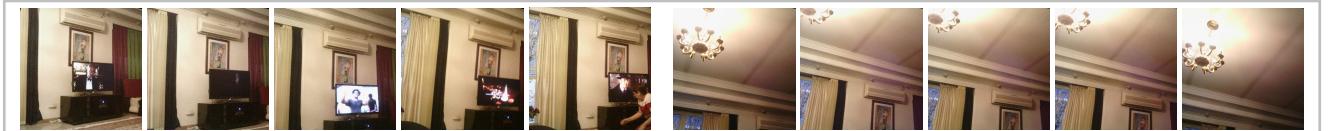
(a) **True Positives:** CES can model public transportation events, as well as street walking.



(b) **True Negatives:** CES remembers previously seen context, and is able to match future and past.



(c) **False Positives:** if the different sight positions span longer than CES' memory span, a false positive will raise.



(d) **False Negatives:** two events taking place in the same location can sometimes be understood as a single one.

Figure 6: Examples of the capacities of CES. The detected events are framed in separate boxes.

training an SVM model to detect false positives. As can be observed in Table 3, such model improves the average precision of CES by 15% (absolute gain of 10%), while recall only decreases by 7.8% (absolute loss of 6%). The benefit of using a supervised SVM pruning is much significant for segmentation algorithms of lower precision, such as k-means, even if coming at a higher recall cost.

Performance of CES relative to manual annotations. Since there is not just one correct way of segmenting video content into events, we have to compare the performance of CES relative to that of the average person. For each lifelog, we average the performance of all available annotations, as evaluated on the selected ground truth. The averaged scores are reported in Table 3. We observe that subjects are, in f-measure, only 3 points better than CES. Even though the precision of the manual annotations is very high, the annotators also obtain worse recall than CES. This is due to some of the subjects selecting very general events, e.g. wrapping all working afternoon within the same event, disregarding the different meetings. Such annotation criteria yields many false negatives, and therefore drops the recall score. Analogously, in some other cases, subjects selected more details than the ground truth. As a result, their rate of false positives is greater than zero.

CES segments, on average, into more events than the annotators. As a result, it is able to detect 13% more true boundaries than the test subjects, but will also find a relative 70% more incorrect events. Such a large increase is to be expected, as the selected ground truth is very exhaustive, and the annotators rarely identify boundaries not present in the ground truth. Overall, we can conclude that CES is a highly precise event segmentation algorithm. Given our ground truth, CES' f-measure is of 96% relative to the manual performance.

	averaged F1	averaged Prec.	averaged Rec.
CES-error	0.42	0.45	0.49
CES-mean	0.52	0.56	0.56
CES-PCA	0.66	0.67	0.69
CES (with VCP)	0.69	0.66	0.77
k-means w/ SVM	0.67	0.70	0.67
CES w/ SVM	0.71	0.75	0.71
Manual segmentation	0.72	0.80	0.68

Table 3: Detailed experiments. Comparison of CES with visual context prediction as opposed to using other feature predictions or aggregations; performance of the SVM pruning; and accuracy of the manual annotations against the selected ground truth. (Evaluated on EDUB-Seg20).

6 CONCLUSIONS

In this paper, we have introduced Contextual Event Segmentation, a novel unsupervised event segmentation method that uses the sequential nature of a photo-stream to infer the presence of event boundaries. At the core of CES is the Visual Context Predictor (VCP), a future sequence generator model that predicts the visual context from a given sequence of frames. The visual context at $t - 1$ given the past is compared to that at $t + 1$ given the future, to determine whether there is a boundary at frame t .

We have also introduced R3, a large scale visual lifelogging dataset depicting a wide variety of events. This dataset is recorded in an unconstrained manner by 57 independent users, who captured

their daily activities morning to evening during over a month. The existence of this dataset has allowed us to train the Visual Context Predictor, which is able to model human activities given sequences of visual features. In a series of experiments, we have proved that the visual context is a strong indicator of event changes. Therefore, we conjecture that the visual context can also be useful for storytelling tasks and tracking of daily activities.

Leveraging on the visual context of the sequences allows CES to detect boundaries between heterogeneous events and ignore local occlusions and brief diversions. CES improves the performance of the baselines by over 16% in f-measure. The performance of CES is competitive with manual annotations, for which the f-measure is only 3% better than CES'. We propose a fully unsupervised pipeline, which results in greater recall than precision. To improve the precision, supervised pruning can be applied to the final detection step by using cluster consistency analysis. Even though further supervised analysis can be performed to improve that performance, it will always be contingent on the ground truth used, which will be inherently subjective.

REFERENCES

- [1] 2010. WordNet. Princeton University. (2010). <http://wordnet.princeton.edu>
- [2] Kiyoharu Aizawa, Kenichiro Ishijima, and Makoto Shiina. 2001. Summarizing wearable video. In *International Conference on Image Processing*, Vol. 3. IEEE
- [3] Vinay Bettadapura, Daniel Castro, and Irfan Essa. 2016. Discovering picturesque highlights from egocentric vacation videos. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016. IEEE, 1–9.
- [4] Bharat Lal Bhatnagar, Suriya Singh, Chetan Arora, and CV Jawahar. 2017. Unsupervised Learning of Deep Feature Representation for Clustering Egocentric Actions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 1447–1453.
- [5] Marc Bolanos, Mariella Dimiccoli, and Petia Radeva. 2017. Toward storytelling from visual lifelogging: An overview. *IEEE Transactions on Human-Machine Systems* 47, 1 (2017), 77–90.
- [6] Marc Bolanos, Ricard Mestre, Estefania Talavera, Xavier Giró-i Nieto, and Petia Radeva. 2015. Visual summary of egocentric photostreams by representative keyframes. In *IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 1–6.
- [7] Marc Bolaños, Álvaro Peris, Francisco Casacuberta, Sergi Soler, and Petia Radeva. 2018. Egocentric video description based on temporally-linked sequences. *Journal of Visual Communication and Image Representation* 50 (2018), 205–216.
- [8] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Giulia Boato, Liting Zhou, and Cathal Gurrin. 2017. Overview of ImageCLEF lifelog 2017: Lifelog Retrieval and Summarization. In *CLEF2017 Working Notes*. Dublin, Ireland.
- [9] Ana Garcia del Molino, Xavier Boix, Joo-Hwee Lim, and Ah-Hwee Tan. 2017. Active Video Summarization: Customized Summaries via On-line Interaction with the User. In *AAAI Conference on Artificial Intelligence*. 4046–4052.
- [10] Ana Garcia del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. 2017. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems* 47, 1 (2017), 65–76.
- [11] Mariella Dimiccoli, Marc Bolaños, Estefania Talavera, Maedeh Aghaei, Stavri G Nikolov, and Petia Radeva. 2017. SR-clustering: Semantic regularized clustering for egocentric photo streams segmentation. *Computer Vision and Image Understanding* 155 (2017), 55–69.
- [12] Aiden R Doherty, Ciarán Ó Conaire, Michael Blighé, Alan F Smeaton, and Noel E O'Connor. 2008. Combining image descriptors to effectively retrieve events from visual lifelogs. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*. ACM, 10–17.
- [13] Aiden R Doherty, Daragh Byrne, Alan F Smeaton, Gareth JF Jones, and Mark Hughes. 2008. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*. ACM, 259–268.
- [14] Alireza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *Computer Vision and Pattern Recognition (CVPR)*, 2012 IEEE Conference on. IEEE, 1226–1233.
- [15] Antonino Furnari, Sebastiano Battiatto, and Giovanni Maria Farinella. 2018. Personal-location-based temporal segmentation of egocentric videos for lifelogging applications. *Journal of Visual Communication and Image Representation*.
- [16] Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*. 1019–1027.
- [17] Ana Garcia del Molino, Mandal Bappaditya, Jie Lin, Joo-Hwee Lim, Subbaraju Vigneshwaran, and Chandrasekhar Vijay. 2017. VC-I2R at ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization. In *CLEF working notes, CEUR*.
- [18] Ana García del Molino and Michael Gygli. 2018. PHD-GIFs: Personalized Highlight Detection for Automatic GIF Creation. In *Proceedings of the 2018 ACM on Multimedia Conference (MM '18)*. ACM, New York, NY, USA.
- [19] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rashmi Gupta, Rami Albatal, Dang Nguyen, and Duc Tien. 2017. Overview of NTCIR-13 lifelog-2 task.
- [20] Michael Gygli, Helmut Grabner, Hayko Riemen Schneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Computer Vision-ECCV*. Springer.
- [21] Morgan Harvey, Marc Langheinrich, and Geoff Ward. 2016. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive and Mobile Computing* 27 (2016), 14–26.
- [22] Graham Healy, Cathal Gurrin, and Alan F Smeaton. 2014. Lifelogging and EEG: utilising neural signals for sorting lifelog image data. (2014).
- [23] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 41, 6 (2011).
- [24] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization.. In *Computer Vision and Pattern Recognition*, Vol. 2. 6.
- [25] Yong Jae Lee and Kristen Grauman. 2015. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision* (2015).
- [26] Jie Lin, Ana Garcia del Molino, Qianli Xu, Fen Fang, Subbaraju Vigneshwaran, and Joo-Hwee Lim. 2017. VC-I2R at the NTCIR-13 Lifelog Semantic Access Task. In *Proceedings of NTCIR-13, Tokyo, Japan*.
- [27] Wei-Hao Lin and Alexander Hauptmann. 2006. Structuring continuous video recordings of everyday life using time-constrained clustering. In *Multimedia Content Analysis, Management, and Retrieval 2006*, Vol. 6073. International Society for Optics and Photonics, 60730D.
- [28] Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. 2017. Let Your Photos Talk: Generating Narrative Paragraph for Photo Stream via Bidirectional Attention Recurrent Neural Networks. (2017).
- [29] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition*. IEEE, 2714–2721.
- [30] Arthur G Money and Harry Agius. 2008. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation* 19, 2 (2008), 121–143.
- [31] Hang Wei Ng, Yasuhito Sawahata, and Kiyoharu Aizawa. 2002. Summarization of wearable videos using support vector machine. In *International Conference on Multimedia and Expo*, Vol. 1. IEEE, 325–328.
- [32] Yair Poleg, Chetan Arora, and Shmuel Peleg. 2014. Temporal segmentation of egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2537–2544.
- [33] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. 2016. Compact cnn for indexing egocentric videos. In *Applications of Computer Vision (WACV)*, 2016 IEEE Winter Conference on. IEEE, 1–9.
- [34] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-specific video summarization. *Computer Vision-ECCV* (2014), 540–555.
- [35] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. 2009. Temporal segmentation and activity classification from first-person sensing. In *Computer Vision and Pattern Recognition Workshops, 2009*. IEEE, 17–24.
- [36] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning*. 843–852.
- [37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [38] Patrizia Varini, Giuseppe Serra, and Rita Cucchiara. 2015. Egocentric Video Summarization of Cultural Tour Based on User Preferences. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*. ACM, 931–934.
- [39] Bo Xiong and Kristen Grauman. 2014. Detecting snap points in egocentric video with a web photo prior. *Computer Vision-ECCV* (2014), 282–298.
- [40] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. 2015. Gaze-Enabled Egocentric Video Summarization via Constrained Submodular Maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2235–2244.
- [41] Shuhei Yamamoto, Takuji Nishimura, Yasunori Akagi, Yoshiaki Takimoto, Taka-fumi Inoue, and Hiroyuki Toda. 2017. Pbg at the ntcir-13 lifelog-2 lat, lsat, and lest tasks. *Proceedings of NTCIR-13, Tokyo, Japan* (2017).
- [42] Licheng Yu, Mohit Bansal, and Tamara L Berg. 2017. Hierarchically-Attentive RNN for Album Summarization and Storytelling. *arXiv preprint arXiv:1708.02977*.
- [43] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.