Generating 1 Minute Summaries of Day Long Egocentric Videos

Anuj Rathore*
anuj.rathore@research.iiit.ac.in
IIIT Hyderabad

Pravin Nagar*
pravinn@iiitd.ac.in
IIIT Delhi

Chetan Arora
chetan@cse.iitd.ac.in
IIT Delhi

C. V. Jawahar
jawahar@iiit.ac.in
IIIT Hyderabad

ABSTRACT

The popularity of egocentric cameras and their always-on nature has lead to the abundance of day-long first-person videos. Because of the extreme shake and highly redundant nature, these videos are difficult to watch from beginning to end and often require summarization tools for their efficient consumption. However, traditional summarization techniques developed for static surveillance videos, or highly curated sports videos and movies are, either, not suitable or simply do not scale for such hours long videos in the wild. On the other hand, specialized summarization techniques developed for egocentric videos limit their focus to important objects and people. In this paper, we present a novel unsupervised reinforcement learning technique to generate video summaries from day long egocentric videos. Our approach can be adapted to generate summaries of various lengths making it possible to view even 1-minute summaries of one's entire day. The technique can also be adapted to various rewards, such as distinctiveness and indicativeness of the summary. When using the facial saliency-based reward, we show that our approach generates summaries focusing on social interactions, similar to the current state-of-the-art (SOTA). Quantitative comparison on the benchmark Disney dataset shows that our method achieves significant improvement in Relaxed F-Score (RFS) (32.56 vs. 19.21) and BLEU score (12.12 vs. 10.64). Finally, we show that our technique can be applied for summarizing traditional, short, hand-held videos as well, where we improve the SOTA F-score on benchmark SumMe and TVSum datasets from 41.4 to 45.6 and 57.6 to 59.1 respectively.

CCS CONCEPTS

• **Computing methodologies** → **Video summarization**; **Unsupervised learning**; *Reinforcement learning*; • **Applied computing** → Surveillance mechanisms.

KEYWORDS

Egocentric Videos; First Person Videos; Video Summarization

*Both authors contributed equally to this research.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350880

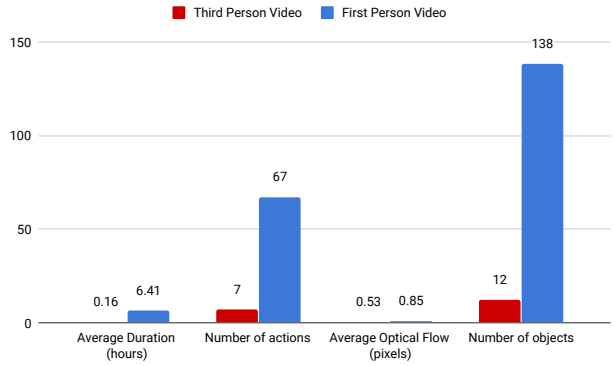

Figure 1: Egocentric videos are characterized by their long, redundant, and extremely shaky nature. The figure shows comparative statistics for benchmark egocentric and third person video. While other statistics are obvious, optical flow indicates frequent sharp changes in viewpoints due to wearer's head motion. The typical characteristics make traditional summarization techniques unsuitable for egocentric videos.

ACM Reference Format:
Anuj Rathore, Pravin Nagar, Chetan Arora, and C. V. Jawahar. 2019. Generating 1 Minute Summaries of Day Long Egocentric Videos. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350880

1 INTRODUCTION

Rapid advancements in technology have made wearable cameras [6, 19, 23] affordable and popular. Apart from recreational purposes, these wearable cameras are increasingly being used in law enforcement, geriatric care (for the old people), and lifelogging applications. The cameras are typically harnessed with head or spectacles and often record day-long visual diaries from a first person perspective in a hands-free mode. The captured videos are highly redundant and extremely shaky, making them difficult to watch from beginning to end, thus necessitating the use of summarization tools for their efficient browsing.

The objective of a video summarization algorithm is to create a compact yet comprehensive summary by selecting appropriate frames from an input video. The problem has been a well-studied area in computer vision. However, most of the work has targeted

Methods	Unsupervised	Scalability to Day Long Videos	Customized Summaries		Head Shake Resistance
			Variable Length	User Specified Saliency	
K-Medoids	✓	✓	✓	✗	✗
DR-DSN[38]	✓	✗	✓	✗	✗
M-AVS[9]	✗	✗	✗	✗	✗
dppLSTM[34]	✗	✗	✗	✗	✗
FFNet[10]	✗	✗	✓	✗	✗
SUM-GAN$_{dpp}$[17]	✗	✗	✓	✗	✗
Ours	✓	✓	✓	✓	✓

Table 1: Comparison of state of the art techniques with proposed method on various criterion important for applicability to egocentric videos. Our approach is specially designed focusing on all the egocentric specific characteristics.

videos from static surveillance cameras [2, 26, 35]. The focus is not mis-placed, since surveillance videos form the majority among all kinds of videos captured and have long uninteresting portions. This makes the use of video summarization attractive. However, from the algorithmic perspective, the summarization problem is much easier for surveillance videos, and can be mostly done by subtracting static background and choosing frames with significant and important foreground objects.

On the other hand videos from point and shoot camera are typically triggered by user interest and do not have long uninteresting portions. In a video captured from a moving camera, the background is also moving, and the task of determining which frames to include in a summary becomes much more challenging. Researchers have suggested various cues to select the summary frames such as motion [36], global image features [9, 17, 38], detecting important events, the presence of salient objects and people [12, 16], as well as role of a frame in a hypothetical storyline [28]. Most of these techniques, give a score to each frame and then use a separate combinatorial algorithm [16, 31] to select the frames that maximize the score in a given summary length constraint. The major shortcoming of these techniques is in their pre-specified saliency definition, restricted capability to model inter-frame interactions for global indicativeness of the summary, and lack of scalability for long videos.

With the success of deep neural networks (DNNs) in learning complex frame and video representations, researchers have proposed various video summarization techniques using DNNs in both supervised [9, 34] as well as unsupervised learning [17, 38] settings. Here, RNNs/LSTMs are typically used to model sequential dependency among frames and score each of them. Given the numerical constraints on the back-propagating gradient, such architectures can not process input videos longer than a couple of thousand frames. Even the Hierarchical RNNs [37] can process sequence length upto 1600 only.

The focus of this paper is on summarizing hours long egocentric videos, containing extreme shakes and long uninteresting portions. Camera wearer often moves in a variety of scenes, and perform various daily activities. The characteristics rule out techniques relying on detection of important pre-specified events or objects. Further, obtaining annotated samples for summarization is hard even for third person video, but is even more harder for egocentric videos, which are often captured in an enhanced privacy scenario. Therefore, in this paper, we propose a novel unsupervised deep reinforcement learning (RL) technique to summarize egocentric videos. While the motivation for unsupervised and DNNs is obvious by now, using an RL based framework allows us to adapt to user specified saliency measures which are difficult to standardize, given the huge variety of contexts in which egocentric videos are typically captured. Moreover, the key features addressed by the proposed approach are compared against SOTA in Table 1. The specific strengths of our work are:

(1) The proposed approach can work with arbitrary long input videos and can be trained to generate summaries of various lengths. We demonstrate it by generating 1, 5, 10 and 15 minutes summaries of day-long egocentric videos from Disney [5], UTE [11, 16], and HUJI [20, 21] datasets. We report Relaxed F-score (explained in Section 4) and BLEU scores of the summary with ground truth. Our approach gives RFS of 32.56 against 19.21 and BLEU score of 12.12 against 10.64 by the SOTA on Disney dataset.
(2) Our approach can focus on various user-specified saliency criterion for the summary, such as distinctiveness, indicativeness, and object, or motion saliency.
(3) Though our focus is on egocentric videos, our technique can summarize hand-held videos as well. We obtain F-score of 45.6 and 59.1 on SumMe [7] and TVSum [27] datasets respectively, against the SOTA score of 41.4 and 57.6 respectively.

2 RELATED WORK

Summarizing Short Hand-Held Videos. Supervised video summarization techniques have dominated the field of short video summarization [7, 9, 34], where variants of submodular function maximization, sequential determinantal point process, and LSTMs have been typically used to maximize various informative measures like representativeness, relevance, and uniformity in the learned summary. Unsupervised video summarization techniques have received more attention in recent years [16, 17, 38], which include low level handcrafted informative measure like visual or motion cues [18, 32] for generating the summary. Higher level informative measure including diversity and representativeness have been proposed recently [27, 31]. Mahasseni et al. [17] use an adversarial

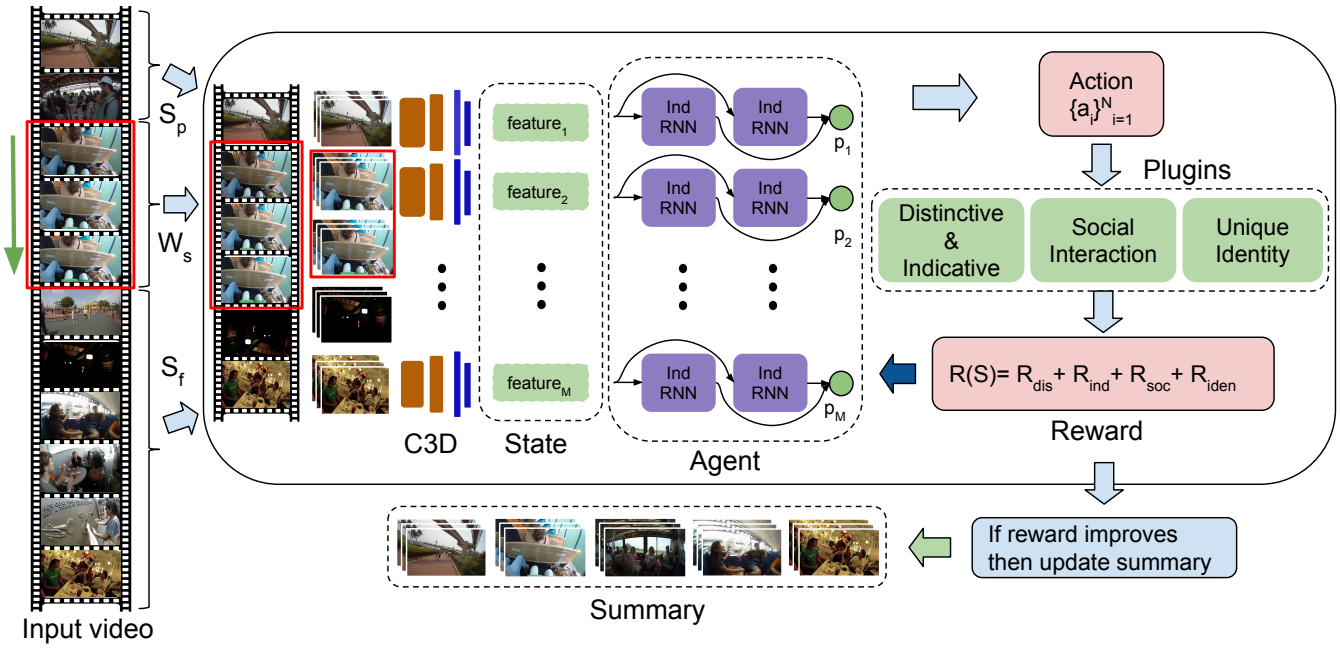

Figure 2: Illustration of the proposed technique to summarize day long egocentric videos based on reinforcement learning (RL) framework. As per the current position of sliding window (W_s) we select a set of segments as a past summary (S_p) and future summary (S_f) (a global representative of input video) from the previously generated summary. The RL agent takes actions on the input (S_p+W_s+S_f) to select the subshots for summary by maximizing the reward in each iteration. The feedback reward R(S), based on various informative measures, assesses the goodness of summary. The figure shows the reward based on distinctiveness, indicativeness, social interaction, and face identity.

learning framework for video summarization. Song et al. [25] proposed an RL technique to extract video category specific keyframes. However, this work requires category information and keyframe labels during training. Zhou et al. [38] have extended the work with a reward function to maximize diversity and representativeness in the summary. This model is unsupervised but does not scale for the videos longer than a few thousand frames.

Egocentric Video Summarization. Egocentric video summarization techniques often rely on important objects and people present in the videos [12, 16]. Xu et al. [31] use gaze information, whereas Lin et al. [15] use context-specific highlight model to generate the summary. Yao et al. [32] generate a summary using a pairwise deep ranking model which give highlight score for each segment of the long video. To overcome the scarcity of the first person labeled data Ho et al. [8] proposed a deep neural network which produces cross-domain feature embedding and transfer highlight across video domain. Lu et al. [16] have proposed story driven summarization which explicitly accounts for connectivity between the important entities. These entities are predefined important objects for the known environment and visual words for the unknown environment. Most of the techniques that are discussed, are specific to a video context (e.g. daily life, cooking video) and fail for the unseen environments.

3 PROPOSED APPROACH

The specific requirements for a practical approach for egocentric video summarization are:

(1) Egocentric videos are recorded in a large variety of scenarios. Also, obtaining annotation is hard due to enhanced privacy concerns. The proposed approach should be unsupervised without any predefined salient events, or objects.
(2) Egocentric videos are captured in an always-on manner and are typically very long. The proposed approach must be able to summarize hours long input videos.
(3) Since saliency measures are often user specific, the proposed approach should be adaptable to user preference for importance as well as for the length of the summary.

3.1 Proposed Formulation

The proposed technique is an encoder-decoder architecture as adopted by various video summarization approaches. Egocentric videos are very shaky and contain abrupt changes so using typical frame level spatial features may not be sufficient. We have used 3D convolutional neural network [29], trained on Sports-1M dataset for feature extraction, which helps us to capture spatio-temporal information. We extract pool5 features, $\{x_i\}_{i=1}^N$, for input video *subshots* (set of 16 temporally adjacent frames). The decoder is a Bidirectional IndRNN [13] with fully connected layers topped by

a sigmoid function. The BiIndRNN takes subshot features as an input and produces an output probability representing the importance score for each subshot. The hidden state ($h_i = h^f \| h^b$) of BiIndRNN encapsulates past and future information of i^{th} subshots using forward and backward stream respectively. The update state of IndRNN for the forward pass is defined as follows:

$$h_i^f = \sigma(W^f x_i + u^f \odot h_{i-1}^f + b^f) \tag{1}$$

where u^f is recurrent weight vector and $\odot$ represents the Hadamard product. W^f and b^f are input weights and the bias of the neurons respectively.

We formulate the summary generation as an RL problem, where the state space comprises of input subshots (x_i) and the action set (a_t) is a binary decision for selecting or not selecting a particular subshot in the summary. To train the summarization agent we use the policy based reinforcement learning to optimize the policy π_θ with parameter θ that maximizes the expected reward:

$$J^\pi(\theta) = \mathbb{E}_{\pi_\theta(a_{1:M}|h_{1:M})}\big[R(S)\big], \tag{2}$$

where M is the number of input subshots to the RL agent and S is the output summary. Probability distribution over the input subshots (M) is denoted by $\pi_\theta(a_{1:M}|h_{1:M})$, where $a_i \in \{0, 1\}$ indicate whether the ith subshot is selected or not. $R(S)$ is the reward function that measures the quality of generated summaries.

Optimization. We use experience replay for robust convergence and speeding up the training process. To perform experience replay we store agent's most recent experience in the memory. During training we apply policy gradient descent update on mini-batch of most recent experience to update the model parameters θ.

$$\theta = \theta - \alpha \sum_{b=1}^{B} \nabla_\theta(J(\theta)), \tag{3}$$

where α is learning rate and B is mini-batch size.

3.2 Capturing Saliency using Rewards

We propose distinctiveness and indicativeness based rewards that allow us to define a measure of goodness of a summary, without any pre-specified important objects or events, in an unsupervised fashion.

Distinctiveness Reward: Let $V = \{1, \cdots, N\}$, represents the set of subshots in the original input video sequence, and $S = \{i \mid i \in \{1, N\}\}$ denotes the set of indices of the subshots included in the summary (here in after called *summary subshots*). Let x_i be the feature representation of i^{th} subshot. Distinctiveness reward is calculated by measuring the degree of distinctiveness among the summary subshots, and computed as the mean of pairwise distinctiveness among the selected video subshots using L2 norm, which is defined as:

$$R_{\text{dis}} = \frac{1}{|S|(|S|-1)} \sum_{i \in S} \sum_{\substack{j \in S, \\ j \neq i}} \left\| x_i - x_j \right\|_2 \tag{4}$$

Indicativeness Reward: The indicativeness reward measures how well the summary subshots represent the original input video. We assume that each subshot in the original video can be described as a linear combination of a few indicative subshots. Therefore we define R_{ind} as:

$$R_{\text{ind}} = -\sum_{i \in V} \min_a (x_i - \sum_{j \in S} a_j x_j)^2 \tag{5}$$

Algorithm 1 Proposed Algorithm

Input $F_{i=1}^N$**:** Video subshots
Output $P_{i=1}^M$**:** Probability scores

```
Freeze the C3D weights and randomly initialize weights of BiIndRNN
for For each epoch do
    for For each video do
        for For each pass do
            for For each sliding window do
                calculate S_p and S_f according to the position of W_s
                Get probability scores from the neural network
                for For each episode do
                    Compute cost and episodic reward
                end for
                if For each mini batch then
                    Back-propagation of expected episodic reward
                end if
            end for
        end for
    end for
end for
```

3.3 Working with Hours Long Videos

The network architecture described above takes few subshots as input and is capable of suggesting the most distinctive and indicative subshots. Note that the proposed technique does not require the input subshots to be temporarily adjacent. Therefore, instead of giving the whole video as input in one shot, we use a sliding window approach. We keep on moving a sliding window (containing continuous subshots) and at any temporal location, we give two sets of input to our model. The first input is subshots covered by current window and second is most recently generated 'indicative subshots' (or the latest summary which is divided into S_p and S_f according to the current position of the sliding window). With these two inputs, we ask the network to pick the most distinctive and indicative subshots.

Based on the trained weights the network outputs probability score corresponding to each subshot. We choose an action sequence of top scoring subshots based on these probability scores to match the desired summary length. We compute the reward in feature space over the action sequence and back-propagate the gradient as per the policy gradient technique. Further, if the selected subshots get a better reward then the previous, we update the 'indicative subshots' of the video according to the current selection. The updated representation is used in the next pass, and the same process is repeated for all sliding windows of the video and over multiple scans of the video. The proposed framework which is further elaborated by Figure 2 and algorithm 1 is unsupervised and can work with arbitrarily long videos while still maintaining the global context for generating a summary.

3.4 Customizing Summaries

Given the unconstrained nature of egocentric videos, it is hard to pre-suppose the saliency criterion. We propose a plugin based architecture where different plugins can bias the generated summaries using appropriate rewards. We propose two novel rewards in this paper.

One of the important events in the egocentric videos is the interaction of the wearer with objects and people around. While, distinctiveness and indicativeness do capture a good summary of the video, many times the focus gets biased towards novelty in the background scene. To bring the focus of the summary on the people to people interactions, we suggest two new rewards functions.

Social Interaction Reward: We propose a new reward emphasizing on the social interactions present in egocentric videos. We integrate a FasterRCNN [22] model, fine-tuned for face detection, into the proposed network. We detect faces in each frame included in the summary and, add the ratio of faces in the summary to the length of the summary, into the reward. We observe that, during social interaction (face^{soc}), faces tend to occupy a larger area ($\text{face}^{\text{area}}$) and also have higher prediction confidence score ($\text{face}^{\text{conf}}$). The smaller faces with low confidence are usually far away from people of no relevance in social interaction. Therefore, we threshold the bounding box area and confidence score, to eliminate the faces with no social interaction with the wearer. We define social interaction reward as

$$R_{\text{soc}} = \frac{\sum_{i \in S} \text{face}_i^{\text{soc}}}{|S|}, \quad \text{where}$$

$$\text{face}_t^{\text{soc}} = \begin{cases} 1, & \text{if face}_t^{\text{conf}} > 98\% \\ & \text{and face}_t^{\text{area}} > 4\% \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Face Identity Reward: We suggest this reward to generate a summary focusing on unique interactions present in a video sequence. To evaluate this reward, we compute OpenFace [1] features of the faces detected by FasterRCNN. However, apart from the usual distinctiveness and indicativeness reward on subshot features, we add an additional reward for the distinctiveness of these face features:

$$R_{\text{iden}} = \frac{1}{|S|(|S|-1)} \sum_{i \in S} \sum_{i' \in S, i' \neq i} \left(1 - \frac{f_i^T f_{i'}}{||f_i||_2 ||f_{i'}||_2}\right),$$

where f_i corresponds to facial feature in i^{th} frame. The reward biases generated summary towards including all the people whom a wearer might have interacted with in the video.

Customizing Summary Length: It is hard to predict the amount of important content in a day-long egocentric video. Therefore, we propose to generate summaries of different lengths to cater to various kinds of content. Since our model is completely unsupervised, all we need to do is to run a different training routine for outputs of different lengths. The different rewards proposed in the earlier sections can all be normalized to the output of different lengths in a straightforward manner. In the experiments section, we demonstrate with output summaries of 1, 5, 10 and 15 minutes. Apart from showing the adaptability of the proposed model, the summaries also demonstrate how well the proposed technique is able to pick up content at different granularity from the input videos.

4 EXPERIMENTS & RESULTS

Datasets: We demonstrate the results on Disney [5], UT Egocentric (UTE) [11, 16], HUJI [20, 21], SumMe [7] and TVSum [27] datasets. Disney, UTE, and HUJI are long duration egocentric video datasets. Disney consists of videos captured at Disney World by 6 individuals for 3 days. Here, we have merged the small video segments following the numbering order provided, into a day-long video for each individual. After merging, we have 8 sequences of 4 to 8 hrs for each individual. For Disney, [33] has provided ground truth text and video summaries of three videos namely Alin Day 1, Alireza Day 1 and Michael Day 2 by three annotators. UTE comprises 4 videos each of 3 to 5 hrs long, and captured in an unconstrained setting. HUJI dataset comprises of 44 egocentric videos of less than 30 minutes duration each and captures daily activities performed by 3 subjects both indoor and outdoor. UTE and HUJI both datasets do not have any ground truth summaries (neither text nor video). To evaluate the proposed approach on UTE we have manually annotated the ground truth by three annotators for all the four vidoes. We will release our annotations post-publication. SumMe and TVSum are benchmark datasets containing small duration video sequence. SumMe consists of 25 video sequences ranging from 1 to 6 minutes videos of various domains such as sports, holidays, etc in both third person and egocentric perspective. It is annotated by 15 to 18 individuals with multiple summaries. TVSum contains 50 video sequences of 2 to 10 minutes, covering news, documentaries etc. It is also annotated by 20 persons with multiple summaries.

Evaluation Methodology: We observe that egocentric videos are highly redundant especially in a temporal neighborhood. Therefore, picking any of the frames from a local neighborhood leads to perceptually similar summaries. However, the commonly used F-score [34], for evaluating summary does not capture this aspect, leading to arbitrary scores with little perceptual correlation. We use the metric proposed by [3] called *Relaxed F-score* (RFS). In Relaxed F-score, given a pair of predicted summary, S and ground truth summary, G; instead of taking exact overlap, we take a fixed temporal relaxation (Δt) around G, while calculating true positive (TP) and then remove these frames from the false positive (FP) and false negative (FN) calculations. The relaxed precision (P_r), recall (R_r) and F-score (F_r) are defined as:

$$P_r = \frac{\text{Relaxed TP}}{\text{Relaxed TP + FP}}, \text{ and } R_r = \frac{\text{Relaxed TP}}{\text{Relaxed TP + FN}}$$

$$F_r = \frac{2 \times P_r \times R_r}{P_r + R_r} \times 100\%$$

For long sequence egocentric videos the semantic information can be more accurately expressed in texts. Therefore, we perform the natural language description based evaluation of video summaries as proposed by [33]. We convert the predicted summary to text using [30] and then use BiLingual Evaluation Understudy (BLEU) [24] score for evaluation. In one more evaluation measure named Average Human Rank (AHR), we asked five volunteers to rank the summaries of various SOTA including the proposed approach on a scale of 10 and reported the Average Human Rank of summaries.

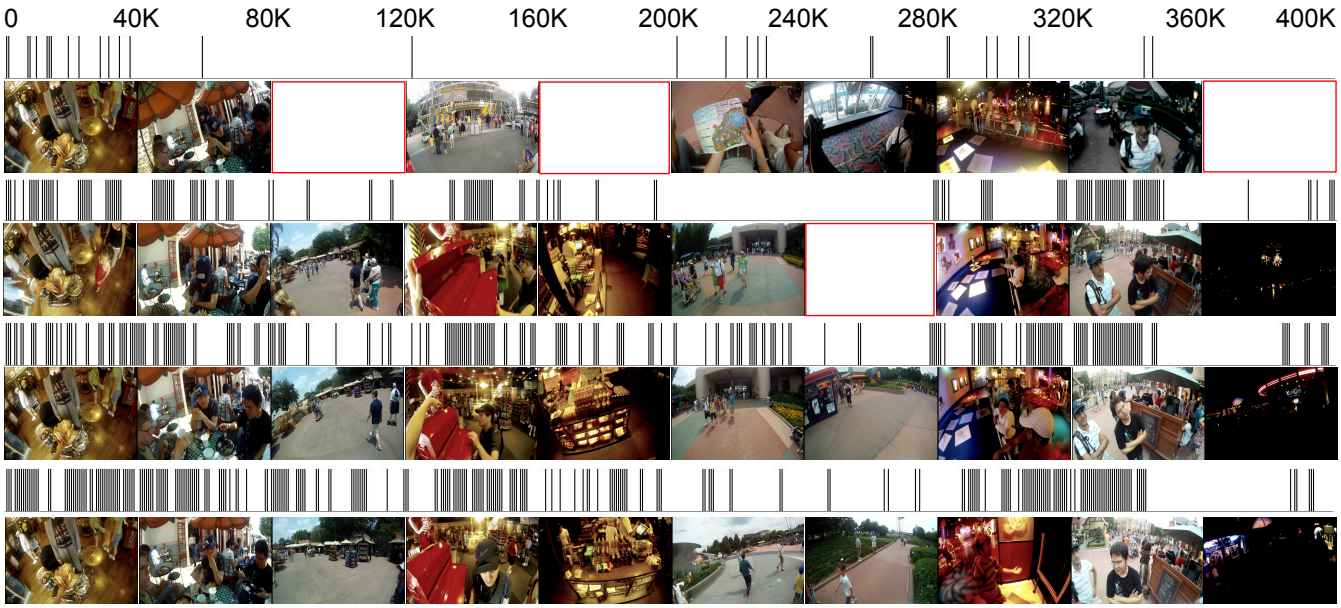

Figure 3: Comparing 1, 5, 10 and 15 minutes summaries (row 1-4) based on distinctiveness-indicativeness reward on 'Michael Day 2' sequence from Disney dataset. The numbers on the top show frame numbers (from 0 to 400K). The pictures show indicative frames in the summary from the corresponding frame range. The blank rectangles indicate no frames were picked from those frame ranges. The black vertical bars in each row indicates a frame was picked from a corresponding temporal window of 700 frames. The bar serves to indicate the distribution of summary frames in the video.

Methods	Alin			Michael			Alireza		
	Relaxed F-Score	BLEU	AHR	Relaxed F-Score	BLEU	AHR	Relaxed F-Score	BLEU	AHR
Uniform samp.	20.6	11.16	7.33	17.23	12.31	6.0	17.05	9.74	5.6
K-medoids	22.08	11.33	4.6	17.73	12.35	4.3	17.84	8.25	3.1
dppLSTM[34]	10.87	7.12	6.31	20.13	10.23	6.0	15.80	5.20	5.8
DR-DSN[38]	11.44	9.81	7.3	16.30	11.23	5.8	16.79	6.31	7.22
FFNet[10]	19.18	6.54	5.8	19.76	9.23	4.3	18.52	9.83	3.96
SUM-GAN[17]	12.27	4.9	7.3	16.53	5.29	7.33	14.14	4.55	6.0
Ours$_{ind}$	31.22	12.30	4.3	27.18	10.92	5.33	20.45	11.54	3.96
Ours$_{dis}$	25.01	12.56	6.0	24.41	10.48	5.6	24.24	10.39	6.0
Ours$_{uni}$	33.84	10.34	4.7	35.22	11.58	4.3	19.44	10.02	5.6
Ours	**34.65**	**12.92**	**2.33**	**35.4**	**12.77**	**4.3**	**27.65**	**10.67**	**3.1**

Table 2: Performance comparison between SOTA and the proposed method on three samples of Disney dataset based on various performance measures such as Relaxed F-score with the temporal relaxation of 50 units (RFS-50), BLEU score, and Average Human Rank (AHR) on a scale of 10 units (low number represents better summary).

For small duration video datasets, we have adopted evaluation method proposed by [34] and use traditional F-score to measure the quality of generated summary (F-score can be defined as RFS with the temporal relaxation of 0). For SumMe and TVSum we generate a summary (S) which is 15% of original video length and report the mean F-score generated from multiple ground truth summaries.

Implementation details: After experiments with a few different sizes, we set sliding window lengths to 25 percent of the desired summary length. We use 3D convolutional features to harness spatio-temporal information for all the four datasets [29]. Our single layer IndRNN contains 256 hidden state units. We set learning rate(α) to 10^{-5}, number of episodes to 5 and mini-batch size to 16. The maximum epochs used to train the network is 20. The proposed technique is implemented in PyTorch and tested on an Nvidia Quadro P5000 GPU. It takes approximately 2 hrs (inclusive of feature extraction) to summarized an 8 hrs long video. The GPU

Figure 4: The figure shows a comparison between baseline [38] and proposed approach for the 10 minutes summaries of 'Michael Day 2' sequence. The 1st row shows the original frames and the numbers on the top show frame numbers (from 140Kth frame to 300Kth in the original video. The 2nd row shows the predicted summary frames by baseline method and 3rd, 4th, and 5th rows show output from the proposed method using distinctiveness-indicativeness, social interaction, and unique identity based rewards respectively. The blank rectangles indicate no frames were picked from those frame ranges. We observe that the baseline approach misses various important events and instead picks clusters of selected frames over two particular locations.

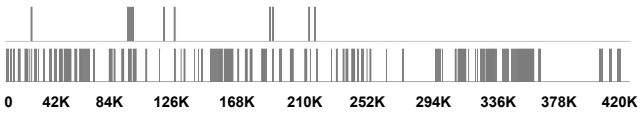

Figure 5: We observed in fig. 4 that baseline approach [38] picks a cluster of frames from a particular location in the summary, whereas the proposed approach effectively distributes the summary frame from all over the video. This figure gives a better visualization by showing the distribution of the summary frames. Each black line indicates a summary frame chosen from a temporal window of 700 frames.

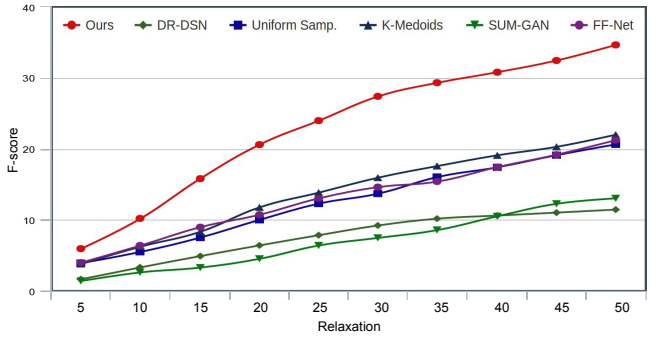

Figure 6: Commonly used F-score do no correlate well with goodness of a summary for long videos. We suggest Relaxed F-score to evaluate the summaries (please see the text for details). The plot above shows Relaxed F-score for different units of temporal relaxation (Δt) for 'Disnet Alin Day 1' video.

memory required to generate 5 minutes summary is approximately 1500MB. The pre-trained model and code will be made available post-publication. [1].

Results on Long Egocentric Videos: Table 2 shows the quantitative evaluation based on RFS, BLEU and AHR based score for Disney dataset. For [38], we unroll the network for the whole video at the test time and generate the probability of picking each frame. Top scoring frames according to the summary length are then outputted as the summary. We notice significant performance improvement over all the SOTA approaches. We report an average of 17% improvement against DR-DSN [38] in relaxed F-score for 50 units of temporal relaxation for three videos of Disney dataset. In Figure 6 we compare various SOTA approaches based on Relaxed F-score for various amounts of temporal relaxation (Δt). As we increase the relaxation, the Relaxed F-score increases linearly for all the methods and from the graph, it is evident that our approach clearly outperforms SOTA approaches by a huge margin for all relaxations. Furthermore, the BLEU score indicates that for all three videos we improve significantly by approximately 2%. The AHR shows significant improvement for all three videos and it's evident that the proposed method consider user aspect of summaries like smooth shots and head shake resistance. Table 3 shows that for UTE dataset, we improve significantly in RFS-50 measure for

[1]Code: https://github.com/anuj-rathore/Generating-One-Minute-Summaries

Method	P01	P02	P03	P04
Uniform samp.	27.78	25.11	36.56	20.79
K-medoids	30.50	22.86	39.66	22.59
FFNet [10]	30.78	19.37	35.92	27.43
SUM-GAN$_{dpp}$ [17]	31.68	10.91	35.85	25.44
dppLSTM [34]	32.47	26.78	41.66	26.93
DR-DSN [38]	36.36	28.21	42.54	**27.81**
Ours	**38.61**	**31.62**	**52.36**	21.66

Table 3: Results on UTE Data with RFS-50

Method	SumMe	TVSum	Category
dppLSTM[34]	38.6	54.7	supervised
SUM-GAN$_{sup}$ [17]	41.7	56.3	supervised
DR-DSN$_{sup}$ [38]	42.1	58.1	supervised
Li et al. [14]	43.1	52.7	supervised
M-AVS [9]	44.4	**61.0**	supervised
H-RNN [37]	44.3	**62.1**	supervised
Uniform samping	29.3	15.5	unsupervised
K-medoids	33.4	28.8	unsupervised
Elhamifar et al. [4]	37.8	42.0	unsupervised
Song [27]	-	50.0	unsupervised
SUM-GAN [17]	39.1	51.7	unsupervised
DR-DSN [38]	41.4	57.6	unsupervised
Ours	**45.6**	**59.1**	unsupervised

Table 4: Though not the focus of this paper, we evaluate our method on short video benchmarks as well, for a thorough comparison. The table shows F-scores for various techniques on SumMe and TVSum datasets. Mentioned results are from respective original papers. Our technique is unsupervised, and improves all unsupervised, and all but one supervised SOTA techniques.

three videos but for one video we perform marginally poor (21.66 vs. 27.81) . Figure 4 shows a qualitative comparison between the baseline method [38] and the summaries generated by our method using distinctiveness-indicativeness, social interaction, and unique identity based rewards. We observe that the baseline often gets biased towards a short temporal segment in the video and all the summary frames are picked from that segment. On the other hand, our distinctiveness and indicativeness reward is able to correctly distribute the summary frames from all over the video. Since its hard to see the clustering in a selection from this figure, we give another visualization in Figure 5, where each bar indicates a frame selected for the summary from a temporal window of 700 frames in the video. Camera ego motion causes significant challenges for summarization algorithm. We observe, brisk head motion captured in the summary generated by various SOTA including the proposed approach when used frame level features. To deal with it we use the spatio-temporal feature which surpasses the effect of brisk head motion in the features by harnessing the temporal information. We

Features	Mode	Dist.	Ind.	Both
CNN	Uni-LSTM	42.12	43.58	43.75
CNN	Bi-LSTM	43.82	44.70	44.23
CNN	Uni-IndRNN	43.84	45.01	45.75
CNN	Bi-IndRNN	45.01	46.01	46.60
C3D	Uni-IndRNN	44.13	43.73	44.80
C3D	Bi-IndRNN	45.20	44.60	45.60

Table 5: Ablation study on SumMe dataset. Please see the text for details

have demonstrated the observation in the supplementary video. Due to lack of annotations, we have shown qualitative results for HUJI dataset in the supplementary material.

We can also observe in Figure 4, that the summaries generated with social interaction and unique identity based reward ignore the video segments like approaching the building, walking over the pool etc., which do not involve social interaction or faces. The summaries are correctly centered towards their desired objective. In Figure 3 we compare 1 minute, 5 minutes, 10 minutes and 15 minutes summaries by our method. As can be seen, our network can adapt to different desired summary lengths. We also observe that, as expected, most of the frames present in the shorter summaries are also present in the longer ones along with some additional frames.

Results on Short Hand-held Videos: Though not the focus of this paper, but for the sake of completion, we also evaluate our method over short hand held videos. Table 4 shows the comparison. Our method outperforms all unsupervised methods. Though the proposed method is unsupervised and comparison with supervised techniques may not be fair, we still did a comparison and except for H-RNN [37] and M-AVS [9], where we perform close, our method improved SOTA supervised techniques as well.

Ablation Study: Table 5 shows various ablations to show the generality of the proposed approach. In all the combinations bidirectional setting always improves by approximately 1% over unidirectional. That validates the hypothesis that the past and future context is essential for summarization. We also validate that together distinctiveness and indicative reward performs better compare to their individual setting. Furthermore, we observe that IndRNN always improves over LSTM, indicates that IndRNN better captures the temporal context. However, Table 5 shows that the CNN feature with IndRNN outperforms spatio-temporal features with IndRNN because SumMe dataset comprises very few egocentric videos.

5 CONCLUSION

In this paper, we proposed a technique to summarize day long egocentric videos. To the best of our knowledge, ours is the first technique with a capability to summarize such long sequences from Disney, UTE and HUJI datasets in a completely unsupervised and end to end manner. To claim the generality of our technique, we have shown robust quantitative and qualitative evaluation and improve SOTA results on long as well as short video datasets.

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* (2016).

[2] Uros Damnjanovic, Virginia Fernandez, Ebroul Izquierdo, and José María Martinez. 2008. Event detection and clustering for surveillance video summarization. In *Image Analysis for Multimedia Interactive Services, 2008.*

[3] Ana Garcia del Molino, Joo-Hwee Lim, and Ah-Hwee Tan. 2018. Predicting Visual Context for Unsupervised Event Segmentation in Continuous Photo-streams. *arXiv preprint arXiv:1808.02289* (2018).

[4] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. 2012. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition.*

[5] Alircza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *CVPR.*

[6] GoPro. [n. d.]. . www.gopro.com. Accessed: 2018-09-03.

[7] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *ECCV.*

[8] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. 2018. Summarizing First-Person Videos from Third Persons' Points of View. In *ECCV.*

[9] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2017. Video summarization with attention-based encoder-decoder networks. *arXiv:1708.09545* (2017).

[10] Shuyue Lan, Rameswar Panda, Qi Zhu, and Amit K Roy-Chowdhury. 2018. Ffnet: Video fast-forwarding via reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 6771–6780.

[11] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *CVPR.*

[12] Yong Jae Lee and Kristen Grauman. 2015. Predicting important objects for egocentric video summarization. *IJCV* (2015).

[13] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2018. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In *CVPR.*

[14] Xuelong Li, Bin Zhao, and Xiaoqiang Lu. 2017. A general framework for edited video and raw video summarization. *TIP* (2017).

[15] Yen-Liang Lin, Vlad I Morariu, and Winston Hsu. 2015. Summarizing while recording: Context-based highlight detection for egocentric videos. In *ICCVW.*

[16] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *CVPR.*

[17] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial LSTM networks. In *CVPR.*

[18] Jingjing Meng, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. 2016. From keyframes to key objects: Video summarization by representative object proposal selection. In *CVPR.*

[19] Pivothead. [n. d.]. . www.pivothead.com. Accessed: 2018-09-03.

[20] Yair Poleg, Chetan Arora, and Shmuel Peleg. 2014. Temporal Segmentation of Egocentric Videos. In *CVPR.*

[21] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. 2016. Compact CNN for Indexing Egocentric Videos. In *WACV.*

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *PAMI* (2017).

[23] SenseCam. [n. d.]. . www.microsoft.com/microsoft-hololens Accessed: 2018-09-03.

[24] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv:1706.09799* (2017).

[25] Xinhui Song, Ke Chen, Jie Lei, Li Sun, Zhiyuan Wang, Lei Xie, and Mingli Song. 2016. Category driven deep recurrent neural network for video summarization. In *Multimedia & Expo Workshops.*

[26] Xinhui Song, Li Sun, Jie Lei, Dapeng Tao, Guanhong Yuan, and Mingli Song. 2016. Event-based large scale surveillance video summarization. *Neurocomputing* (2016).

[27] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. TVsum: Summarizing web videos using titles. In *CVPR.*

[28] Antonio Tejero-de Pablos, Yuta Nakashima, Tomokazu Sato, Naokazu Yokoya, Marko Linna, and Esa Rahtu. 2018. Summarization of User-Generated Sports Video by Using Deep Action Recognition Features. *IEEE TMM* (2018).

[29] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision.*

[30] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *ICCV.*

[31] 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 *CVPR.*

[32] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR.*

[33] Serena Yeung, Alireza Fathi, and Li Fei-Fei. 2014. Videoset: Video summary evaluation through text. *arXiv:1406.5824* (2014).

[34] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *ECCV.*

[35] Shu Zhang, Yingying Zhu, and Amit K Roy-Chowdhury. 2016. Context-Aware Surveillance Video Summarization. *IEEE Trans. Image Processing* (2016).

[36] Yujia Zhang, Xiaodan Liang, Dingwen Zhang, Min Tan, and Eric P Xing. 2018. Unsupervised Object-Level Video Summarization with Online Motion Auto-Encoder. *arXiv:1801.00543* (2018).

[37] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2017. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia.* ACM, 863–871.

[38] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep Reinforcement Learning for Unsupervised Video Summarization With Diversity-Representativeness Reward. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16395