# Generating Personalized Summaries of Day Long Egocentric Videos

Pravin Nagar, Anuj Rathore, C.V. Jawahar, Chetan Arora

**Abstract**—The popularity of egocentric cameras and their always-on nature has lead to the abundance of day long first-person videos. The highly redundant nature of these videos and extreme camera-shakes make them difficult to watch from beginning to end. These videos require efficient summarization tools for 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. This paper presents a novel unsupervised reinforcement learning framework to summarize egocentric videos both in terms of length and the content. The proposed framework facilitates incorporating various prior preferences such as faces, places, or scene diversity and interactive user choice in terms of including or excluding the particular type of content. This approach can also be adapted to generate summaries of various lengths, making it possible to view even 1-minute summaries of one's entire day. 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). The quantitative comparisons on the benchmark Disney dataset show that our method achieves significant improvement in Relaxed F-Score (RFS) ($29.60$ compared to $19.21$ from SOTA), BLEU score ($0.68$ compared to $0.67$ from SOTA), Average Human Ranking (AHR), and unique events covered. Finally, we show that our technique can be applied to summarize traditional, short, hand-held videos as well, where we improve the SOTA F-score on benchmark SumMe and TVSum datasets from $41.4$ to $46.40$ and $57.6$ to $58.3$ respectively. We also provide a Pytorch implementation and a web demo at https://pravin74.github.io/Int-sum/index.html.

## 1 Introduction

RAPID advancements in technology have made wearable cameras [1], [2], [3] 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 with two styles for generated summary: *keyframes* and *video skims*. In the keyframes-based output, the summary is represented by a set of salient frames of the original video sequence. This is also called *still image abstract* or *static storyboard*. A video skim-based summary is generated as the collection of video segments extracted from the original video sequence. This is also called the *moving image abstract*, or *moving storyboard*. The focus of this paper is on generating video skims.

The majority of the summarization techniques include predefined events/criteria such as action scenes and loud music for movies, anomaly detection in the surveillance video, and specific events in a sports video. On the other hand, videos from point and shoot cameras are typically triggered by user interest and do not have long uninteresting portions. However, in a video captured using 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 [4], global image features [5], [6], [7], detecting important events, the presence of salient objects and people [8], [9],

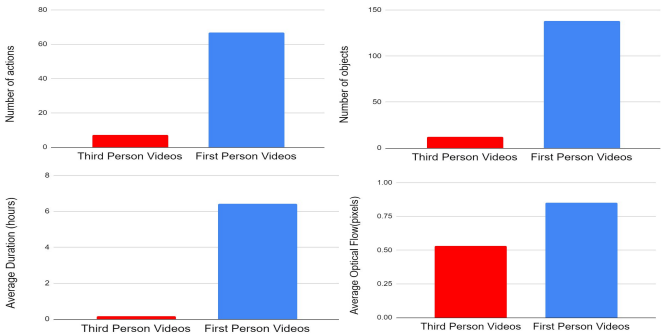


Fig. 1: Egocentric videos are characterized by their long, redundant, and extremely shaky nature. The figure shows comparative statistics for benchmark egocentric and third person video. We use Disney, HUJI, and UTE datasets for first-person and TVSum and SumMe for third-person datasets to calculate the statistics. While other statistics are obvious, optical flow indicates frequent sharp changes in viewpoints due to the wearer's head motion. The typical characteristics make traditional summarization techniques unsuitable for egocentric videos.

- *Pravin Nagar is with Indraprastha Institute of Information Technology Delhi, New Delhi, India, 110020. E-mail: pravinn@iiitd.ac.in*
- *Anuj Rathore and C.V. Jawahar are with International Institute of Information Technology Hyderabad, Hyderabad, Telangana, India, 500032*
- *Chetan Arora is with Indian Institute of Technology Delhi, New Delhi, Delhi, 110016.*

| Methods | Unsup | Scalable | Customization | | | SR |
|---|---|---|---|---|---|---|
| | | | VL | US | Int | |
| K-Medoids | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ |
| DR-DSN [6] | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ |
| M-AVS [7] | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| dppLSTM [12] | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| FFNet [13] | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| SUM-GAN$_{dpp}$ [5] | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| Ours | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

TABLE 1: Comparison of SOTA techniques with the proposed method on various criteria important for applicability to egocentric videos. Abbreviations: Unsup = Unsupervised, VL: Variable Length, US: User Saliency, Int: Interactive, SR: Shake Resistance.

as well as role of a frame in a hypothetical storyline [10]. Most of these techniques give a score to each frame and then use a separate combinatorial algorithm [9], [11] to select the frames that maximize the score in a given summary length constraint. The major shortcomings of these techniques are in their pre-specified saliency definition, the restricted capability to model inter-frame interactions for global indicativeness of the summary and lack of scalability and customization for long videos.

The success of deep neural networks (DNNs) in learning complex frames and video representations has paved the way for supervised [7], [12] and unsupervised [5], [6] summarization techniques. Here, RNNs/LSTMs are typically used to model sequential dependency among frames. Given the numerical constraints on back-propagating gradients over many recurrent connections, such architectures can not process input videos longer than a few hundred frames. Even hierarchical approaches [14] can handle up to 1600 frames only.

Egocentric videos contain extreme shakes and long uninteresting portions (see Fig. 1). The camera wearer often moves in a variety of scenes and performs various daily activities. These characteristics rule out techniques that rely on the detection of important pre-specified events or objects. Moreover, the task of obtaining annotated samples for summarization for third-person videos is hard. It is even harder for egocentric videos, which are often captured in enhanced privacy-sensitive scenarios. This rules out the supervised approach, rendering many SOTA techniques unsuitable [5], [7], [12], [13].

While generating visually diverse summaries, it is observed that the summarization criteria are inherently personal. Specifically, in the day long life-logging videos, the same user may want to explore the summary focusing on the different types of events like social interaction, having food, walking, etc. Hence, a key requirement of a summarization framework for egocentric videos is to personalize summaries by interactively collecting user feedback on the fly.

In this work, we formulate video summarization as a sequential decision making process over video frames, where each decision is binary (whether to include the frame in summary or not). The setup requires a sequential model to capture the temporal dependencies, which has been addressed using a bidirectional LSTM based architecture. The quality of the summary is available only for the whole set and not for individual frames. Hence we find the RL framework, which works with sparse rewards, suitable to solve this problem. Our experiments also show an ablation study with various RL optimization algorithms viz policy gradient, Q-Learning, and Actor-Critic styles. The key strengths of our approach are shown in Table 1. The specific contributions of our work are: 1) We propose an interactive summarization framework that can personalize summaries based on the length, content as well as interactive feedback from the user. 2) Our framework 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 several benchmark datasets [9], [15], [16], [17], [18]. 3) Our approach can focus on various user-specified saliency criteria for the summary, such as distinctiveness, indicativeness, and object, or motion saliency. 4) We achieve state-of-the-art performance on benchmark egocentric video datasets. We report Relaxed F-score (explained in Section 4) of $29.60$ against $19.21$ from the SOTA [6]. We also report BLEU score of $11.55$ from our approach in comparison to $10.64$ by the SOTA on the Disney dataset [15]. 5) Though our focus is on egocentric videos, our technique can summarize hand-held videos as well. We obtain F-score of $46..40$ and $58.3$ on SumMe [19] and TVSum [20] datasets respectively, against the SOTA scores of $41.4$ and $57.6$ respectively. A previous version of this work that appeared in [21] only demonstrates the naive RL framework, namely policy gradient, to summarize day long egocentric videos. The current version contains the following core contributions: 1) We propose an interactive summarization framework that can personalize summaries based on the feedback (video exemplars) provided by the user. 2) Advance RL frameworks, namely Q Learning and AC framework, are introduced with various plugins such as distinctiveness, indicativeness, and object or motion saliency.

## 2 Related Work

**Video Summarization:** The majority of keyframe extraction techniques identify events using salient objects and video dynamics from various viewpoints and different degrees [22]. Zhang et al. [23] identify the content change in the video segment to extract keyframes. De et al. [24] find a cluster centroids as a representative of each cluster, which eventually derives the keyframes . However, video datasets exhibit lower inter-class and higher intra-class variance leading to difficulty in defining these clusters. Liu and Kender [25] have used a sequence reconstruction measure (SRM) to measure the degree to which selected keyframes can reconstruct the original video sequence. Dementhon et al. [26], and Latecki et al. [27] pick salient points of manifold formed by the representation of input frames as the keyframes. Dufaux [28] selects keyframes by considering high-level semantic criteria such as high motion, spatial activity, and the likelihood of having people. In contrast, Kang and Hua in [29] used attention, context dominance, and frame quality. These techniques work well for the targeted domain but do not generalize since the heuristic for frame selection is drawn from empirical observations. Video skims based summary generation typically require

high-level context analysis and can be divided into four basic categories: (1) Redundancy elimination in a video by selecting a set of continuous frames that exhibit maximum similarity with input videos [30]. (2) Event/highlight detection and localization techniques which identify and locate the pre-defined events in a video sequence, such as sports videos, e.g. baseball [31], athletics [32], and cricket [33]. (3) Skim curve formulation techniques generate a curve that shows the likelihood of each base unit to include in the skim with respect to some user criteria. A threshold is used on the generated curve, and the segments above the threshold are assembled to form a final skim [34]. (4) Query context personalization which incorporates user feedback, either as a query or a personalized profile, e.g., [35] use human face, and caption text, and [36] use favorite players or a team preferred by the user.

**Summarizing Short Hand-Held Videos:** Supervised video summarization techniques have dominated the field of short video summarization [19]. Variants of submodular function maximization, sequential determinantal point process, and LSTMs have been used to maximize various informative measures like representativeness, relevance, and uniformity in the learned summary. Lu and Grauman [9] proposed unsupervised techniques that include low-level handcrafted informative measures like visual or motion cues [37] for generating the summary. Higher level informative measures including, diversity and representativeness have been proposed recently [20]. Mahasseni et al. [5] use an adversarial learning framework for video summarization. An RL technique proposed by Song et al. [38] extract video category-specific keyframes. Zhou et al. [6] extend their work with a reward function to maximize diversity and representativeness in summary. This model does not scale for videos longer than a few hundred frames.

**Egocentric Video Summarization:** Egocentric video summarization techniques often rely on important objects and people present in the videos [8], as well as gaze [11]. Lin et al. [39] use context-specific highlight model to generate the summary. Yao et al. [40] uses a pairwise deep ranking model to give a highlight score for each segment of the input video. To overcome the scarcity of the first-person labeled data, Ho et al. [41] propose a deep neural network that produces cross-domain feature embedding and transfer highlight across the video domain. Lu et al. [9] propose 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 discussed above are specific to a video context (e.g., daily life or kitchen videos) and fail for the unseen environments.

**Customizing Video Summaries:** The summarization criteria are often user-specific viz inclusion of predefined object or event, presence of audio, duration of summary, etc. Hence generating customized summaries is an important sub-area of video summarization. Malino et al. [42] propose an interactive summarization framework that collects feedback from the user over the most frequent item in the original video. Then it iteratively refines the summary by a question asking interface. A probabilistic framework called active inference in the conditional random field (CRFs) is used to infer the summary preferred by the user. This work fine-tunes CNN on Places dataset [43] to detect most frequent objects or events, which is not feasible for the egocentric setting. Other works take user feedback in the form of natural language queries and use a mapping mechanism to bridge the gap between visual and language to personalize the summarization [44], [45]. Zhang et al. [44] select diverse sub-shots of a video that are representative of the whole video and yet related to a given user query in the natural language. They use a mapping network to connect visual and query space. This mapping network uses a relatedness reward to measure the distance between the predicted and ground truth query embedding for personalization. Similarly, Yousefi and Kuncheva [45] find all the frames related to the query using a semantic concept search. Jin et al. [46] segment video by analyzing visual features and speech detection and assign an importance score to each segment. It uses a variant of the knapsack problem to find an optimal video summary by fast-forwarding or removing unimportant segments. Han et al. [47] represent video by manifold embedding and assigns weights to each frame. Visual saliency features are applied between each pair of frames to learn the inherent video structure. Darabi and Ghinea [48] use predefined categories to score each video segment using Scale Invariant Feature Transform (SIFT) features. The user feedback towards the high-level visual concepts is recorded in the vector form for personalization. After combining these two groups of data highest score video segments reflecting the user priority are returned. We emphasize that the techniques proposed in this work do not rely on the predefined objects or events and take user feedback in the form of video clips instead of text to reduce the overhead resulting from the use of cross-modality.

## 3 Proposed Approach

The specific objectives of the proposed summarization approach are as follows:

1) **Unsupervised:** To handle enhanced privacy concerns.
2) **Scalable:** To handle day long egocentric videos.
3) **Customizable:** To handle vast variety of contexts in the *wild* egocentric videos.
4) **Interactive:** To accommodate user preferences.

To simplify the exposition, we first describe our architecture to generate summaries for short videos in an unsupervised manner. We then explain to scale-up of the architecture for day long videos, followed by the modifications required for customization and interactive summary generation.

### 3.1 Architecture

The proposed framework uses 3D convolutional neural networks (CNNs) for capturing spatio-temporal features from an egocentric video. We have used 3D CNN model [49], called *C3D* hereon, trained on Sports-1M dataset for feature extraction in our design. Other 3D CNN models such as [50], [51], [52] can be used as well. We first divide our video into *sub-shots* of 16 non-overlapping frames and extract 512 dimension features from pool5 layer: $\{x_t\}_{t=1}^{T}$ for each sub-shot from C3D. Here $T$ denotes the total sub-shots extracted

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3118077, IEEE Transactions on Pattern Analysis and Machine Intelligence



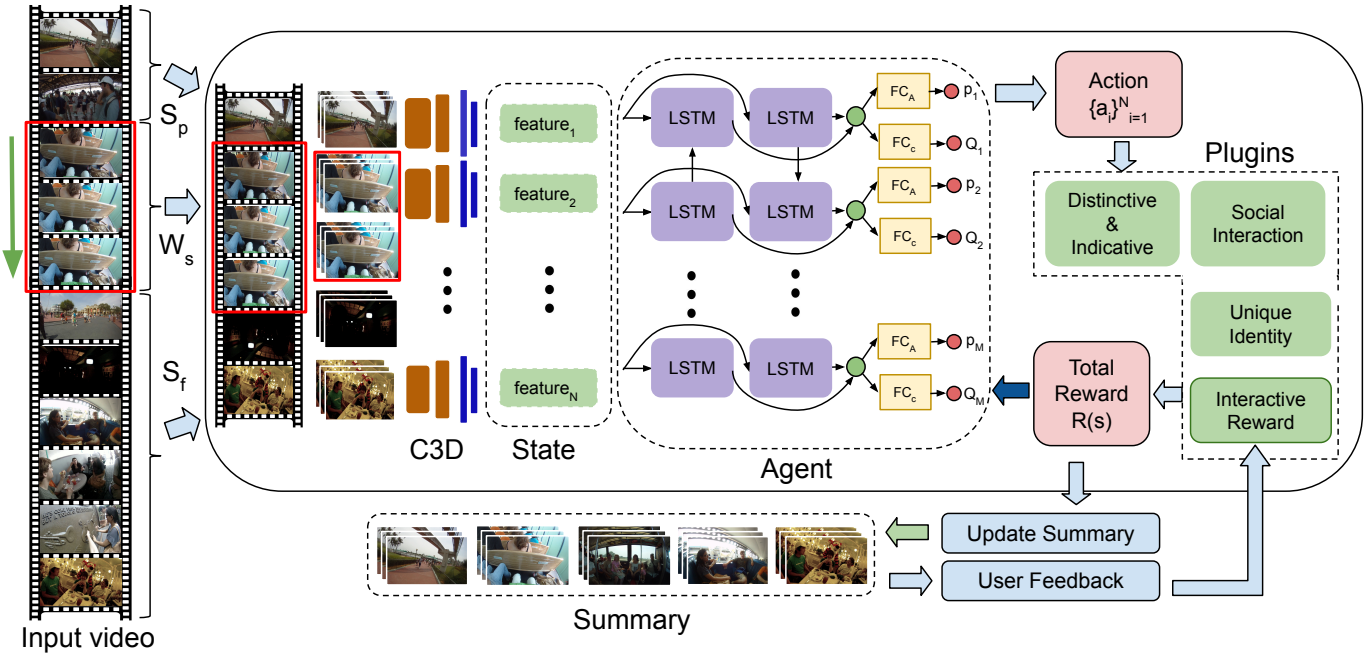


Fig. 2: Illustration of the proposed technique to summarize day long egocentric videos based on reinforcement learning (RL). As per the current position of the sliding window ($W_s$) we select a set of segments as a past summary ($S_p$) and future summary ($S_f$) (a global representative of input video) from the previously generated summary. The first column to the left C3D shows the representation of past, current, and future segments of the video. The past and future segments are represented by their sub-shots in the current summary. Further, each sub-shot in the representation (whether coming from past, current, or future segments) is essentially a set of 16 consecutive frames from which we evaluate the C3D features. The second column to the left of C3D features indicates these sub-shots/sets. The RL agent takes actions on the input ($S_p$+$W_s$+$S_f$) to select the sub-shots for summary by maximizing the reward in each iteration. Based on various informative measures, the feedback reward R(S) assesses the goodness of the summary. The figure shows the reward based on distinctiveness, indicativeness, social interaction, and face identity.

from a video. The extracted features are inputted to the reinforcement learning agent, which uses a bidirectional long short-term memory network (BiLSTM). The hidden state ($h_i = h^f \| h^b$) of BiLSTM encapsulates past and future information of $i^{th}$ sub-shot using forward and backward stream respectively. Here $h^f$ and $h^b$ are hidden states of forward and backward layers of BiLSTM, respectively, and $\|$ indicates the concatenation of the two. We unroll the BiLSTM network $M$ times for the training and give a sub-shot as input to each BiLSTM unit. .

### 3.2 Formulation

We formulate the summary generation as a Reinforcement Learning (RL) problem, where the state space comprises of input sub-shots features $\{x_m\}$, and the action set $\{a_m\}$ is a binary decision for selecting or not selecting a particular sub-shot in summary. To train the summarization agent, we experiment with the following RL optimization strategies: 1) Policy Gradient, 2) Q Learning, and 3) Actor-Critic.

**Summarization with Policy Gradient:** For the policy gradient framework, we design the agent as a BiLSTM network followed by a fully connected (FC) layer for final prediction. The BiLSTM takes C3D features $\{x_m\}_{m=1}^M$ as input and produces corresponding hidden states $\{h_m\}_{m=1}^M$. In the end, the FC layer is followed by a sigmoid function to predict the probability score $\{p_m\}_{m=1}^M$ corresponding to each sub-shot. The output summary corresponding to the input video is then selected by sampling each sub-shot based on the probability outputted by each LSTM unit. The reward for the agent is the score of the overall summary based upon the pre-specified or user-defined scoring functions as described later in Section 3.3, Section 3.5, and Section 3.6.

To train the summarization agent, we use the policy-based reinforcement learning to optimize the policy $\pi_\theta$ with parameter $\theta$ that maximizes the expected reward:

$$J^{\pi}(\theta) \;=\; \mathbb{E}_{\pi_\theta(a_{1:M}|h_{1:M})}\left[R(\mathcal{S})\right], \qquad (1)$$

where $\mathcal{S}$ denotes the output summary. $\pi_\theta(a_{1:M}|h_{1:M})$ denotes probability distribution over the input sub-shots ($M$), where $a_m \in \{0,1\}$ indicates whether the $m^{\text{th}}$ sub-shot is selected or not. $R(\mathcal{S})$ is the reward function that measures the quality of generated summaries.

It can be shown that the derivative of objective function w.r.t. parameters $\theta$ is given as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{p_\theta(a_{1:M})}\left[R(\mathcal{S})\sum_{m=1}^{M}\nabla_\theta \log \pi_\theta(a_m|h_m)\right], \qquad (2)$$

where $p_\theta(a_{1:M})$ denotes probability distribution over possible action sequence. Since we calculate the expectation over the action sequence, which is difficult to compute directly. We approximate it by sampling actions for $E$ episodes on

the same input and output probability distribution and then calculate the average gradient:

$$\nabla_\theta J(\theta) = \frac{1}{E}\sum_{e=1}^{E}\sum_{m=1}^{M} R(\mathcal{S}_e)\nabla_\theta \log \pi_\theta(a_m|h_m), \quad (3)$$

where $R(\mathcal{S}_e)$ is the reward computed for summary a in the $e^{\text{th}}$ episode. The high variability in cumulative reward and log probabilities make the network hard to converge. We use a common countermeasure to ensure smaller and stable gradient, which is to subtract a baseline, $\mathcal{B}$, from the cumulative reward:

$$\nabla_\theta J(\theta) = \frac{1}{E}\sum_{e=1}^{E}\sum_{m=1}^{M} (R(\mathcal{S}_e) - \mathcal{B})\nabla_\theta \log \pi_\theta(a_m|h_m) \quad (4)$$

where $\mathcal{B}$ is computed as the moving average of rewards experienced so far.

Policy gradient is a naive RL framework that uses the baseline function to calculate the episodic reward. The baseline functions are not learnable, which leads to high variance across video samples. We introduce the Q learning and AC framework that uses a Q value network that leads to a stable gradient across video samples. On the other end, the extra parameters required more training samples. For the proposed framework, each position of the sliding window (refer Section 3.4) constitutes one training sample, so we generate sufficient training samples (especially for day long videos) to train the Q learning and AC frameworks.

**Summarization using Q Learning:** In Q learning, instead of predicting the confidence score, $p_m$, we predict the Q values for selecting or not selecting a sub-shot for a particular state. The objective function of Q learning is to minimize the mean squared error between the target Q value and the approximate Q value with parameter $\theta$ over the input sequence:

$$J^\pi(\theta) = \mathbb{E}_\pi\left[\left(Q^\pi(s,a) - Q^\pi_\theta(s,a)\right)^2\right]. \quad (5)$$

Here $Q^\pi(s,a)$ and $Q^\pi_\theta(s,a)$ is the target Q value and approximate/predicted Q value. As suggested in [53, Ch. 6], we use TD target to approximate the target Q values i.e $Q^\pi(s_m,a_m) = r + \gamma Q^\pi_{\theta^-}(s_{m+1},a_{m+1})$, where $r$ is the current reward, $\gamma$ is the discount factor, and $Q^\pi_{\theta^-}$ is the Q value of the target with parameters updated in the alternate epochs. With the approximation, the weight update is given by:

$$\Delta\theta = \alpha \times \delta_m \times \nabla_\theta Q^\pi_\theta(s_m,a_m), \quad (6)$$

where $\delta$ is the TD error computed as:

$$\delta_m = r + \gamma Q^\pi_{\theta^-}(s_{m+1},a_{m+1}) - Q^\pi_\theta(s_m,a_m) \quad (7)$$

We adopt the idea proposed by [54] to calculate the reward for 'm' steps of an episode, and calculate TD error for the entire episode as:

$$\delta_{1:M} = \sum_{m=1}^{M-1}\left[r_m + \gamma\, Q^\pi_{\theta^-}(s_{m+1},a_{m+1}) - Q^\pi_\theta(s_m,a_m)\right]$$

$$\delta_{1:M} = R(\mathcal{S}) + \gamma\sum_{m=1}^{M-1} Q^\pi_{\theta^-}(s_{m+1},a_{m+1}) - \sum_{m=1}^{M-1} Q^\pi_\theta(s_m,a_m)$$

where $R(\mathcal{S}) = \sum_{m=1}^{M-1} r_m$, is the total reward. And the weight update is given as:

$$\Delta\theta = \alpha \times \delta_{1:M} \times \sum_{m=1}^{M} \nabla Q_\theta(s_m,a_m), \quad (8)$$

where $\alpha$ is the learning rate for the parameters.

**Summarization using Actor-Critic Framework:** For the Actor-Critic framework, we propose a common BiLSTM network, with tied weights, followed by two separate fully connected layers for Actor and Critic as shown in Fig. 2. The common BiLSTM reduces the parameters and ensures fast convergence. The basic policy gradient in an actor-critic framework is given as follows:

$$\nabla_\theta J(\theta) = \mathbb{E}_{p_\theta(a_1:M)}\left[R(\mathcal{S})\sum_{m=1}^{M}\nabla_\theta \log \pi_\theta(a_m|h_m)\right] \quad (9)$$

The actor policy is denoted by $\pi_a$ and its parameters $\theta$ are updated as follows:

$$\Delta\theta = \alpha_a \sum_{m=1}^{M} Q_c(s_m,a_m)\nabla_\theta \log \pi_a(s_m,a_m), \quad (10)$$

where $Q_c$ is the Q-value for the state-action pair given by the critic, and $\alpha_a$ is the learning rate of the actor. Denoting critic parameters by $w$, we update the critic parameters using TD target and calculate the TD error in the same way as done for Q learning:

$$\delta_{1:M} = R(S) + \gamma\sum_{m=1}^{M-1} Q_{w^-}(s_{m+1},a_{m+1}) - \sum_{m=1}^{M-1} Q_w(s_m,a_m).$$

where $Q_{w^-}$ indicates the Q value returned by the critic for the target. With the TD error computed as above, the weight update for the critic is given by:

$$\Delta w = \alpha_c \times \delta_{1:M} \times \sum_{m=1}^{M} \nabla Q_w(s_m,a_m) \quad (11)$$

### 3.3 Scoring a Summary and Basic RL Rewards

The proposed RL framework requires a summary scoring mechanism to compute the goodness of a summary. This score is used as a reward to train the agent using any of the training methodologies (policy gradient, Q learning, or actor-critic) discussed in the previous section. Though we describe many rewards to customize the summaries in the next section, three basic rewards are common to all the summaries produced by our framework. Note that all these rewards do not require the notion of any pre-specified important objects or events.

**Distinctiveness Reward:** Let $\mathcal{V} = \{1,\ldots,M\}$, represents the set of input sub-shots, and $\mathcal{S} = \{i \mid i \in \mathcal{V}\}$ denotes the set of indices of the sub-shots included in the summary (hereinafter called *summary sub-shots*). Let $x_m$ be the feature representation of $m^{\text{th}}$ sub-shot. *Distinctiveness* reward measures the degree of uniqueness among the summary sub-shots, and is computed as the mean of pairwise distance among the selected video sub-shots using $\ell_2$ norm:

$$R_{\text{dis}} = \frac{1}{|\mathcal{S}|(|\mathcal{S}|-1)}\sum_{i\in\mathcal{S}}\sum_{\substack{j\in\mathcal{S},\\ j\neq i}} \|x_i - x_j\|_2 \quad (12)$$

**Indicativeness Reward:** The indicativeness reward measures how well the summary sub-shots represent the original input video. Here the assumption is that each input sub-shot can be described as a linear combination of a small subset of indicative sub-shots. Hence, we define $R_{\text{ind}}$ as:

$$R_{\text{ind}} = -\frac{1}{|V|}\sum_{i\in V}\min_{\mathbf{b}}\left(x_i - \sum_{j\in\mathcal{S}} b_j^i x_j\right)^2, \tag{13}$$

where $V$ indicates the set of input sub-shots in the whole video and each variable $b_j^i$ denotes the weight corresponding to sub-shot $x_j$ in the summary, to best reconstruct an input sub-shot $x_i$. The set of weights $\mathbf{b} = \{b_j^i\}$ are found as the ones maximizing the indicativeness reward for a summary set $\mathcal{S}$.

**Summary Length Reward:** A trivial way to generate a summary that maximizes distinctiveness and indicativeness is to choose all the input sub-shots in the output summary. To prevent such a trivial solution and keep the summary concise, we introduce an additional constraint penalizing the length of the summary. We propose the following reward to generate a summary of the desired length:

$$R_{\text{length}} = -\left(\frac{1}{M}\sum_{m=1}^{M} p_m - \epsilon\right)^2, \tag{14}$$

where $p_m$ denotes the probability outputted by our framework for selecting sub-shot $m$, and $\epsilon$ denotes the desired percentage of sub-shots (given as input to our system) to be selected in the summary.

### 3.4 Scalability to Day Long Egocentric Videos

The proposed technique, as described above, does not require the input sub-shots to be temporarily adjacent. Therefore, to scale it to long videos, instead of giving the whole video as input in one go, we use a sliding window approach. We keep on moving a sliding window (containing temporally adjacent sub-shots), and at any temporal location, we give two sets of input to our model. The first input is sub-shots covered by the current window, and the second is the most recently generated 'indicative sub-shots' (or the latest summary generated by our method). We divide the indicative sub-shots into $\mathcal{S}_p$ and $\mathcal{S}_f$ according to the current position of the sliding window. We use the model described in the previous section to pick the most distinctive and indicative sub-shots with these two inputs.

Based on the trained weights, the network outputs probability scores corresponding to each sub-shot. We choose an action sequence of top-scoring sub-shots 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 one of the RL techniques viz Policy Gradient, Q Learning, or Actor-Critic. Further, if the selected sub-shots get a better reward than the previous summary, we update the 'indicative sub-shots' of the video according to the current selection. The updated representation is then used in the next pass for the next sliding window, and the same process is repeated for all sliding windows of the video. We move the sliding window from the beginning to the end of any day long egocentric video. We call this one scan, and then we repeat this multiple times to better assimilate the information from all parts of the video. Furthermore, we observed no significant systematic bias in the output summary due to the initialization because of multiple scans.

The proposed framework is visually described in Fig. 2. The proposed algorithm can work with arbitrarily long videos while still maintaining the global context for generating a consistent and concise summary.

### 3.5 Customizing Summaries

The unconstrained nature of egocentric videos makes it hard to pre-suppose the saliency criterion. We propose a plugin-based architecture where different plugins can bias the generated summaries using appropriate rewards. Apart from distinctiveness and indicativeness, we propose following two novel rewards, especially for the first-person context:

**Social Interaction Reward:** We propose a new reward emphasizing the social interactions present in egocentric videos. We integrate a FasterRCNN [55] 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, as the reward. We observe that, during social interaction 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 the wearer and are irrelevant from a social interaction perspective. Therefore, we threshold the bounding box area and confidence score, to eliminate the faces with no social interaction with the wearer. With this, we define social interaction reward as:

$$R_{\text{soc}} = \frac{\sum_{m\in\mathcal{S}} \text{face}_m^{\text{soc}}}{|\mathcal{S}|}, \quad \text{where}$$

$$\text{face}_m^{\text{soc}} = \begin{cases} 1, & \text{if face}_m^{\text{conf}} > 98\%, \text{ and face}_m^{\text{area}} > 4\% \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

**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 [56] features of the faces detected by Faster-RCNN. However, apart from the usual distinctiveness and indicativeness reward on sub-shot features, we propose an additional reward for the distinctiveness of face features:

$$R_{\text{iden}} = \frac{1}{|\mathcal{S}|(|\mathcal{S}|-1)}\sum_{i\in\mathcal{S}}\sum_{\substack{j\in\mathcal{S},\\ j\neq i}}\left(1 - \frac{f_i^T f_j}{\|f_i\|_2\|f_j\|_2}\right),$$

where $f_i$ corresponds to the facial features from the $i^{\text{th}}$ sub-shot. The reward biases generated summary towards including all the people, with whom a wearer might have interacted within 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, we merely need to change the desired percentage of sub-shots (epsilon) and retrain the

network to output different length summaries. In the experiments section, we demonstrate the capability by outputting summaries of $1$, $5$, $10$ and $15$ minutes for hours long videos. Apart from showing the adaptability of the proposed model, the summaries also demonstrate how well the proposed technique select content at different granularity from the input videos.

### 3.6 Interactive Summarization

The variety of contexts in which an egocentric video is captured ensures that, despite the various customization proposed for the summary generation in the last few sections, a user may still find some interesting portions not included or some redundant portions included in the summary. Therefore, we propose to introduce a new module in our framework that can interact with the user in an online manner and personalize the summaries by collecting the feedback provided by the user.

After generating the initial summary as described in the last few sections, we ask the user to pick the sub-shots which the user certainly wants in summary. We call such sub-shots positive sub-shots($\mathcal{S}_+$). Similarly, we collect in negative sub-shots($\mathcal{S}_-$), the sub-shots which the user dislikes. Based upon the sets of positive and negative sub-shots, we define the *interactive* reward as follows:

$$R_{\text{int}} = \frac{A}{|\mathcal{S}||\mathcal{S}_+|}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_+}\frac{x_i^T x_j}{\|x_i\|\|x_j\|} + \frac{B}{|\mathcal{S}||\mathcal{S}_-|}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_-}\left(1-\frac{x_i^T x_j}{\|x_i\|\|x_j\|}\right) \tag{16}$$

where A, and B are the weights to fine-tune the impact of the user feedback. We use the interactive reward just as the other rewards in our RL based summarization framework.

## 4 Experiments & Results

### 4.1 Datasets

We demonstrate the results on Disney [15], UT Egocentric (UTE) [9], [16], HUJI [17], [18], SumMe [19] and TVSum [20] datasets. Disney, UTE, and HUJI are long duration egocentric video datasets. Disney consists of videos captured at Disney World by six individuals for three days. Here, we have merged the small video segments, following the numbering order provided by their authors, into a day long video for each individual. After merging, we have eight sequences of $4$ to $8$ hrs for each individual. For Disney, Yeung et al. [57] have 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 four videos, each of 3 to 5 hrs long, and captured in an unconstrained setting. To evaluate the proposed approach on UTE, we have used the annotations provided by Yeung et al. [57]. The HUJI dataset comprises $44$ egocentric videos of less than $30$ minutes each and captures daily activities performed by three subjects, both indoor and outdoor. HUJI dataset do not have any ground truth summaries (neither text nor video).

SumMe and TVSum are benchmark datasets containing small-duration video sequences. 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 perspectives. 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.

### 4.2 Evaluation Methodology

To prove the efficacy of the proposed framework, we use four evaluation measures. 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 [12] for evaluating summary does not capture this aspect, leading to arbitrary scores with little perceptual correlation. In the first evaluation measure, we use the metric proposed by Molino et al. [58], called *Relaxed F-score* (RFS). In Relaxed F-score, given a pair of predicted summary, $\mathcal{S}$ and ground truth summary, $\mathcal{G}$; instead of taking exact overlap, we take a fixed temporal relaxation ($\Delta t$) around $\mathcal{G}$, 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:

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

$$F_r = \frac{2 \times P_r \times R_r}{P_r + R_r} \times 100\% \tag{17}$$

For long sequence egocentric videos, the semantic information can be more accurately expressed in texts [57]. Therefore, in the second evaluation measure, we perform the natural language description based evaluation of video summaries as proposed by [57]. We convert the predicted summary to text using the text description provided for the entire video by [57] and then use BiLingual Evaluation Understudy (BLEU) [59] score for evaluation.

In the third evaluation named Average Human Rating (AHR), we follow [34], [60], [61] to rate the summary based on *informativeness* and *enjoyability* with a confidence score by 10 participants. The participants were recruited using *purposive sampling* [62], where the participants have a different background, with three of them having professional experience in recording videos. The demographic information about the participants is given in the supplementary material. The *informativeness* and *enjoyability* of each participant are weighted by the normalized confidence score, and the average over participants is reported. To conduct the user study, we have shared an information sheet (shown in the supplementary material) with the users, explaining the detailed evaluation procedure to them.

In the last evaluation measure, we score the generated summary based on the number of unique events captured and the jerks present. To calculate the unique events, we have used the text description of the input videos (three videos of the Disney dataset) provided by Yeung et al. [57]. The consecutive sentences are merged if the BLEU score between them is greater than $0.5$. Each unique sentence is then identified as a unique event. To calculate the number

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3118077, IEEE Transactions on Pattern Analysis and Machine Intelligence



| Methods | Alin | | | | Michael | | | | Alireza | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RFS | BLEU | AHR | | RFS | BLEU | AHR | | RFS | BLEU | AHR | |
| | | | INF | ENJ | | | INF | ENJ | | | INF | ENJ |
| Uniform samp. | 20.60 | 0.76 | 2.95 | 1.91 | 17.23 | 0.69 | 2.62 | 1.64 | 17.05 | 0.56 | 2.48 | 1.65 |
| K-medoids | 22.08 | 0.74 | 2.82 | 2.53 | 17.73 | 0.71 | 2.32 | 2.22 | 17.84 | **0.57** | 2.68 | 2.28 |
| dppLSTM [12] | 10.87 | 0.63 | 2.42 | 2.68 | 20.13 | 0.58 | 2.73 | 2.01 | 15.80 | 0.44 | 3.12 | 2.50 |
| DR-DSN [6] | 11.44 | 0.76 | 2.53 | 2.75 | 16.30 | **0.74** | 2.63 | 2.86 | 16.79 | 0.53 | 2.44 | 3.04 |
| FFNet [13] | 19.18 | 0.59 | 1.91 | 1.91 | 19.76 | 0.70 | 2.80 | 2.88 | 18.52 | 0.26 | 2.33 | 2.62 |
| SUM-GAN [5] | 12.27 | 0.53 | 1.17 | 2.26 | 16.53 | 0.64 | 2.14 | 2.48 | 14.14 | 0.41 | 3.18 | 2.78 |
| $\text{Ours}_{\text{PG}}$ | 32.59 | 0.72 | 2.88 | **3.22** | 25.40 | **0.74** | 2.86 | 2.75 | **27.65** | 0.54 | 2.68 | 3.17 |
| $\text{Ours}_{\text{Q}}$ | 30.38 | **0.77** | 3.26 | 2.66 | 23.89 | 0.72 | 2.93 | **3.00** | 23.89 | 0.56 | 3.46 | **3.55** |
| $\text{Ours}_{\text{AC}}$ | **35.65** | 0.74 | **3.68** | 2.74 | **30.00** | 0.73 | **3.46** | 2.95 | 23.16 | **0.57** | **4.06** | 2.90 |

TABLE 2: Performance comparison between SOTA approaches and the variations of the proposed method on the three samples of Disney dataset. We compare various performance measures such as Relaxed F-score (RFS) with the temporal relaxation of 50 units (RFS-50), BLEU score, and Average Human Rating (AHR) using the basic RL rewards. PG, Q, AC show our framework trained with Policy Gradient, Q Learning, and Actor-Critic learning techniques, respectively.

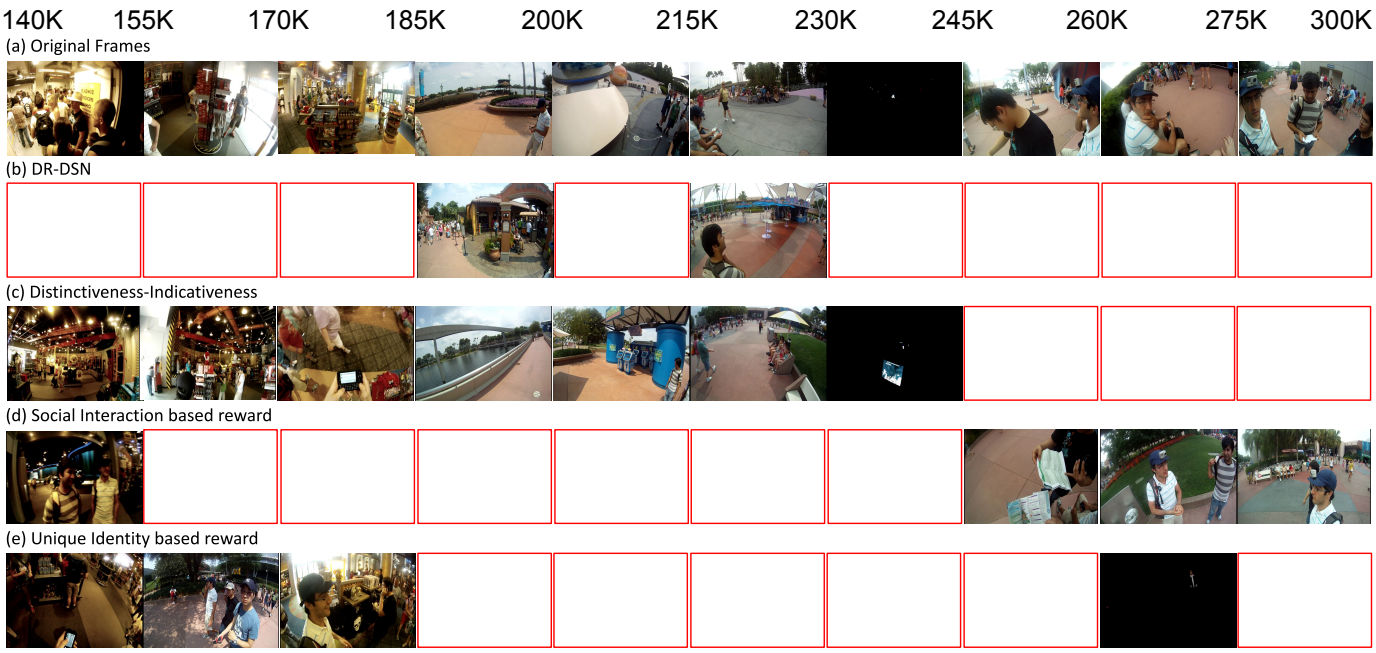


Fig. 3: The figure shows a comparison between DR-DSN [6] and proposed approach for the **10** minutes summaries of 'Michael Day 2' sequence using basic RL rewards. The 1st row shows the original frames, and the numbers on the top show frame numbers (from **140Kth** frame to **300Kth** in the original video. The 2nd row shows the predicted summary frames by the DR-DSN method. The 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 that no frames were picked from those frame ranges. We observe that DR-DSN misses various important events and instead picks clusters of selected frames over two particular locations.

of jerks, we count the number of temporally discontinuous shots in the summary. The final score is calculated as:

$$\text{Score}_{ue} = \text{Unique Events} - \alpha_j \times \text{Number of Jerks} \quad (18)$$

where $\alpha_j$ is weight to penalize unique events by the number of jerks. We use $\alpha_j = 0.3$ in our experiments.

For small duration video datasets, we follow [12] and use traditional F-score to measure the quality. Note that the traditional F-score can also be seen as a special case of Relaxed F-score (RFS) with temporal relaxation of 0. For SumMe and TVSum, we generate a summary ($\mathcal{S}$) which is $15\%$ of original video length, and report the mean F-score generated from multiple ground truth summaries.

As suggested by [42], we did a qualitative evaluation of personalized summaries in two scenarios by 10 participants. In the first scenario, a participant was asked to evaluate a system-generated summary while being unaware of the video content. Here, the system iteratively personalized the generated summary by taking into account the participant's feedback. In the second scenario, we assume that the user is aware of the video content (*e.g.*, the user may be the camera wearer) *a priori*. The detailed procedure is described in the information sheet given to the user and attached in the supplementary material. Once the personalized summary is generated, then the participants rate the summary by the quality of personalization compared to the default summary on the Likert scale (1: very poor, 2: poor, 3: ok, 4: good, 5: excellent) along with their confidence score (1: Not confident to 5: Completely confident).

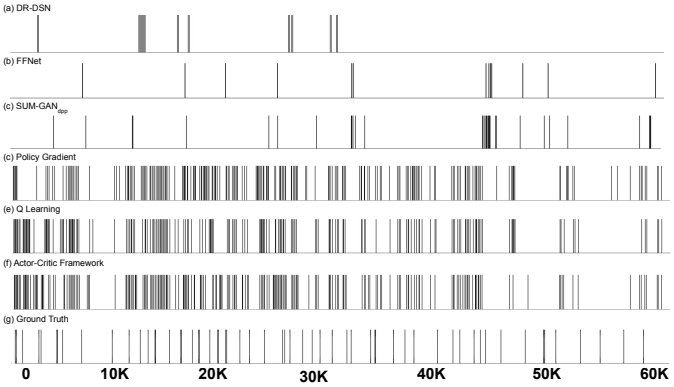


Fig. 4: We observed in Fig. 3 that DR-DSN [6] picks a cluster of frames from a particular location in summary, whereas the proposed frameworks effectively distribute the summary frame from all over the video, same as ground truth. This figure gives a better visualization by showing the distribution of the summary frames with respect to the ground truth summary for various frameworks, including ours for the full video. The figure also indicates that most of the selected summary frames are common despite using different RL frameworks as the reward is the same for all the frameworks.

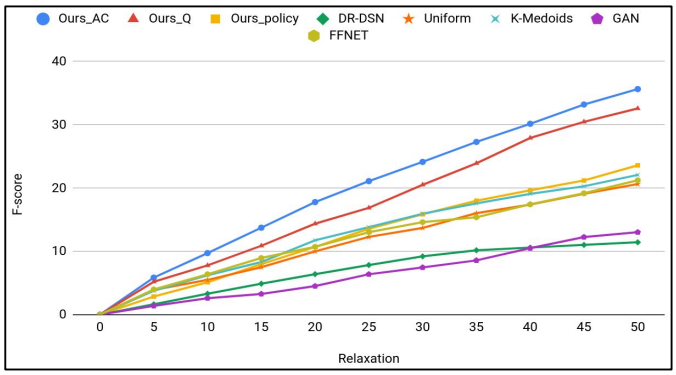


Fig. 5: Commonly used F-score do not correlate well with goodness of a summary for long videos. We suggest Relaxed F-score to evaluate the summaries. The plot above shows Relaxed F-score for different units of temporal relaxation ($\Delta$t) for 'Alin Day 1' video sequence of Disney dataset.

### 4.3 Implementation details

After experiments with a few different sizes, we set sliding window lengths to 25 percent of the desired summary length. For all the frameworks, we set the learning rate ($\alpha$) to $10^{-5}$, learning rate decay to 0.1, number of episodes to 5, number of sliding window pass per video to 4, $\epsilon$ to 0.5, hidden units in the BiLSTM to 256, and mini-batch size to 16. We set the discount factor ($\gamma$) to .99 for Q learning and AC framework. The actor ($\alpha_\theta$) and critic ($\alpha_w$) learning rate are set to $10^{-3}$. The maximum epochs used to train the network is 20. We also add $l_2$ regularization on the weights to avoid overfitting.

The proposed technique is implemented in PyTorch and tested on a regular workstation containing Nvidia Quadro P5000 GPU. It takes approximately 2 hrs (inclusive of feature extraction) to summarize an 8 hrs long video. The GPU memory required to generate a 5 minutes summary is approximately 1500MB.

### 4.4 Results on Long Egocentric Videos

Table 2 shows the quantitative evaluation based on RFS, BLEU, and AHR based scores for the Disney dataset. For comparison with DR-DSN [6], 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 10% improvement against DR-DSN [6] in relaxed F-score for 50 units of temporal relaxation for three videos of the Disney dataset. We perform only marginally better in terms of BLEU score because, for many events, the text description of visually different events overlapped. For example, *"My friends and I walked through the park"* and *"My friends and I walked through the line"* are two visually different events but exhibit close BLEU score. Hence, even if our technique picks more unique events, the BLEU score is marginally better. However, the AHR shows significant performance improvement for all the videos in terms of *informativeness* and *enjoyability* score. Our visualization in Fig. 4 shows that the SOTA approaches typically pick a cluster of frames in summary from the same location, which lowers the *informativeness* and *enjoyability* score compared to the proposed framework. The same is validated through our user study as well. Similarly, the summaries generated by uniform sampling and K-medoids show sudden changes that lead to poor comprehension and lower the *informativeness* and *enjoyability* score. The detailed discussion on the results shown and the comments by the participants are given in the supplementary material.

Table 3 shows the summary score for the unique events covered by 1, 2.5, and 5 minutes summaries. The numbers show that the proposed approach significantly improves compared to all the SOTA approaches for all cases except for one case of where uniform sampling performing better for 'Alireza Day 1' video when the summary length is 2.5.

In Fig. 5, we compare various SOTA approaches based on Relaxed F-score for various amounts of temporal relaxation ($\Delta t$). As we increase the relaxation, the Relaxed F-score increases linearly for all the methods, and from the graph, it is evident that our techniques outperform SOTA approaches by a huge margin for all relaxations.

The UTE dataset comprises small video sequences ($<$ 5 hrs) and is less complex than the Disney dataset. Due to the aforementioned reason, Table 4 shows significant improvement over SOTA in terms of RFS-50 measure for all the UTE videos.

Fig. 3 shows a qualitative comparison between DR-DSN [6] and the summaries generated by our method using distinctiveness-indicativeness, social interaction, and unique identity based rewards on the Disney Dataset. We observe that, due to the specific rewards used, the summaries generated by our technique 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. We also observe in our experiments that the SOTA 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

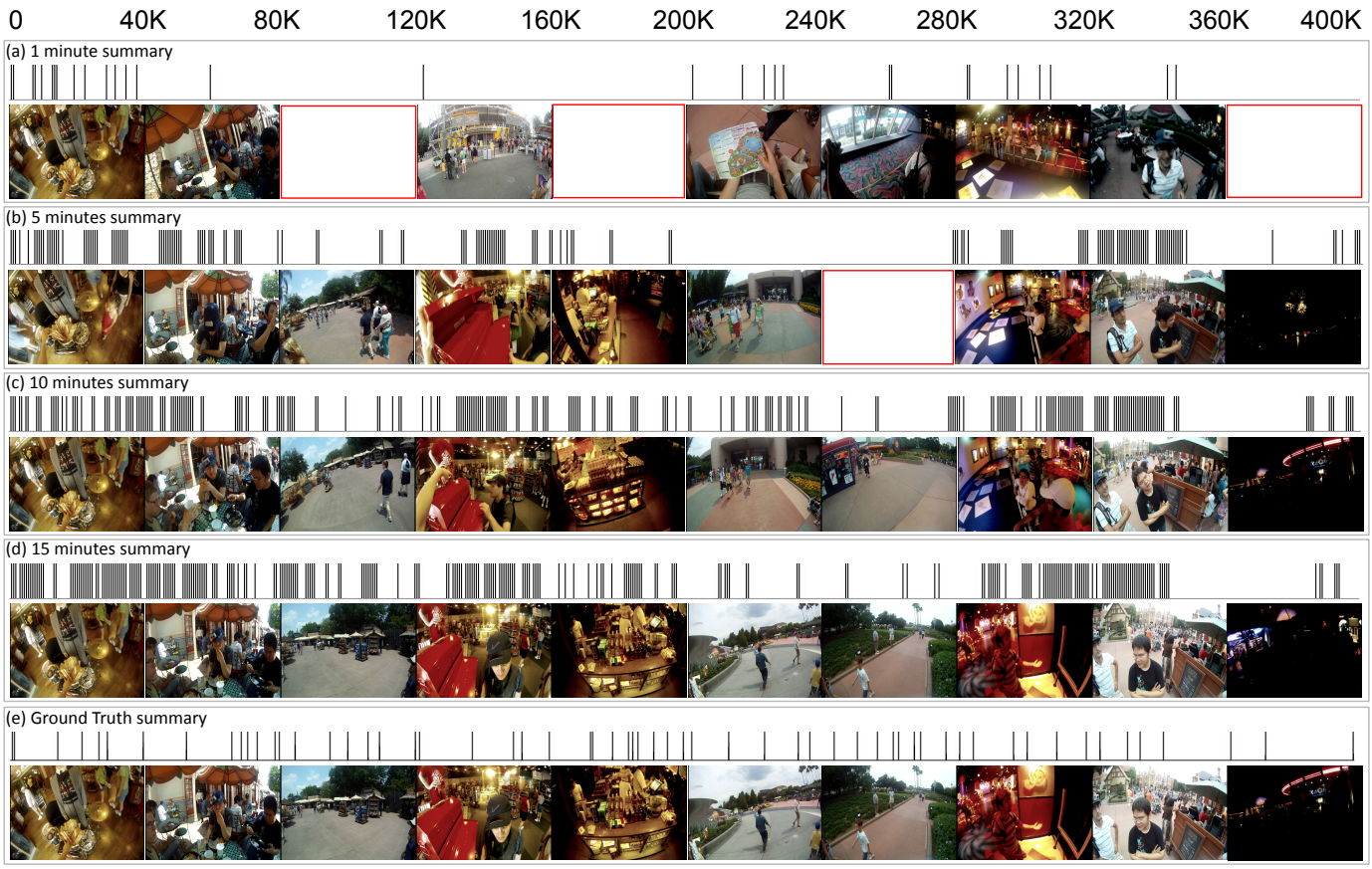


Fig. 6: Comparing 1, 5, 10, and 15 minutes summaries (row 1-4) based on the basic RL rewards using Policy Gradient framework on 'Michael Day 2' sequence from Disney dataset with the ground truth summary (row 5). Note that the ground truth summary length is approximately 5 minutes. The numbers on the top show frame numbers (from 0 to 400K). The pictures show indicative frames in summary from the corresponding frame range. The blank rectangles indicate no frames were picked from those frame ranges. The black vertical bars indicate a frame was picked from a corresponding temporal window of 70 frames in each row. The bar serves to indicate the distribution of summary frames in the video.

| Methods | 1 minute | | | 2.5 minutes | | | 5 minutes | | |
|---|---|---|---|---|---|---|---|---|---|
| | Al | Mi | Az | Al | Mi | Az | Al | Mi | Az |
| Uniform samp. | 21 | 30 | 27 | 40 | 52 | **60** | 38 | 56 | 70 |
| K-medoids | 25 | 28 | 27 | 32 | 48 | 46 | 19 | 49 | 66 |
| FFNet [13] | 21.4 | 14.4 | 10.9 | 20.5 | 43 | 4.7 | 13.3 | 0.5 | 6.7 |
| DR-DSN [6] | 17.5 | 21.5 | 20.2 | 19.1 | 15.7 | 22.8 | 5.2 | 14.4 | 20.9 |
| $\text{Ours}_{\text{PG}}$ | 27.6 | 28.9 | 31.1 | 48.6 | 57.6 | 49.9 | 41.2 | 58.5 | 63.1 |
| $\text{Ours}_{\text{Q}}$ | 28.4 | **43** | 30.9 | 42.2 | 66.6 | 48.6 | 56.6 | 62.5 | 69 |
| $\text{Ours}_{\text{AC}}$ | **33.7** | 33 | **33.4** | **57.7** | **74.8** | 56.6 | **70.4** | **99.9** | **75.2** |

TABLE 3: Performance comparison between SOTA and the variations of the proposed method for the number of unique events covered. We demonstrate the results for 1, 2.5, and 5 minute summaries on the three samples of the Disney dataset using basic rewards (distinctiveness, indicativeness, and summary length).

| 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 [13] | 30.78 | 19.37 | 35.92 | 27.43 |
| $\text{SUM-GAN}_{dpp}$ [5] | 31.68 | 10.91 | 35.85 | 25.44 |
| dppLSTM [12] | 32.47 | 26.78 | 41.66 | 26.93 |
| DR-DSN [6] | 36.36 | 28.21 | 42.54 | 27.81 |
| $\textbf{Ours}_{pol}$ | 43.64 | 46.39 | 51.16 | 39.41 |
| $\textbf{Ours}_{Q}$ | 41.94 | **48.24** | 48.47 | 39.65 |
| $\textbf{Ours}_{AC}$ | **47.50** | 36.26 | **58.86** | **48.10** |

TABLE 4: Comparison on UTE dataset based on basic RL rewards using RFS-50 metric.

| Subjects | Video Name | Dataset | Events Included | Events Excluded | Score (1 to 5) |
|---|---|---|---|---|---|
| S01-S1 | Alin | Disney | 'Dinner' | 'In Dark' | 3 |
| S03-S1 | Alin | Disney | 'Dinner' | 'Tram ride' | 5 |
| S02-S2 | P01 | UTE | 'Driving' | 'Prep. Food' | 4 |
| S01-S2 | Yair | HUJI | 'Driving' | 'Sitting' | 4 |

TABLE 5: The table shows the Likert score when specific events are included or excluded in summary. S0X-SY represents subject 'X' in scenario 'Y'. The detailed results for all 10 participants are shown in the supplementary material.

indicativeness reward is able to distribute the summary frames from all over the video correctly. Since it's hard to see the clustering in a selection from Fig. 3, we give another vi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3118077, IEEE Transactions on Pattern Analysis and Machine Intelligence



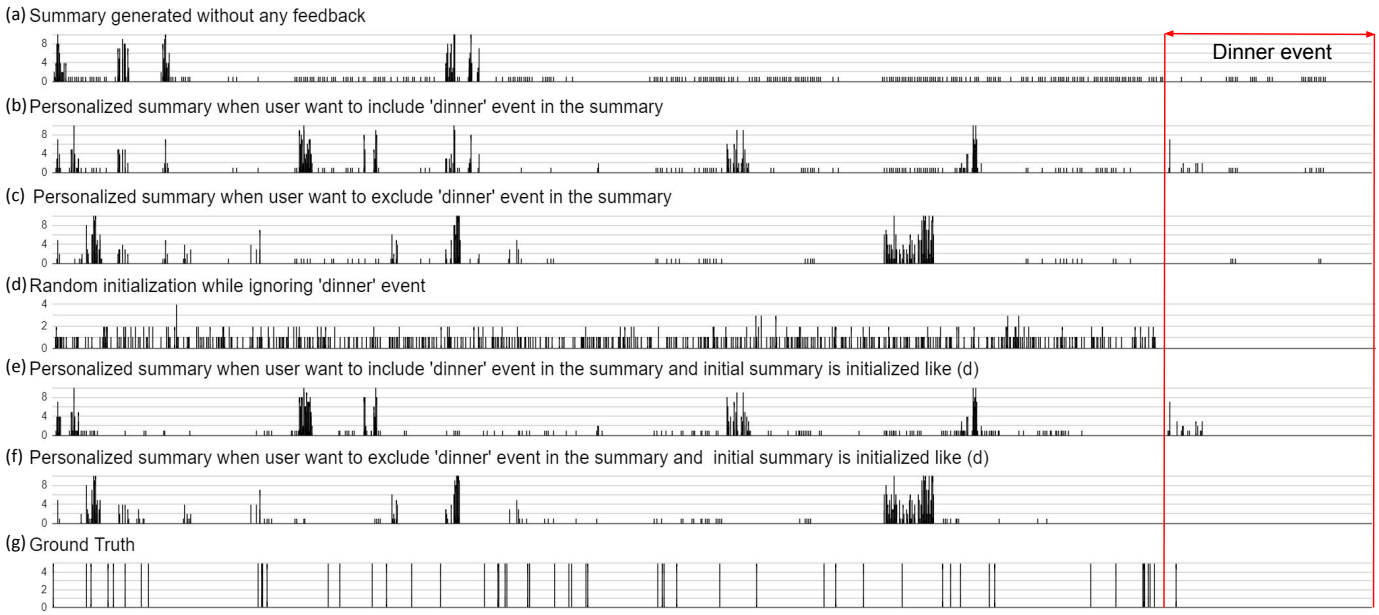


Fig. 7: The figure demonstrates the visualization of the interactive summarization of the 'Alin Day 1' video sequence of the Disney dataset for 10 minutes summaries. Each bar represents 10 seconds of a time interval. (a)-(f) shows different summaries when the user asks to exclude/include 'dinner' event in summary, and (g) shows the ground truth summary distribution. We observe that (b) shows big peaks in the 'dinner' event area, whereas (c) shows very few spikes because of the negative feedback. As an ablation study, we initialized the summary by random frames but not included any frame from the 'dinner' event in the initialization, as shown in (d). When we personalized the summary to include the 'dinner', with the initialization as done in (d), we observe that the summary changes to select sub-shots from the 'dinner' event as shown in (e).

| Method | SumMe | TVSum | Category |
|---|---|---|---|
| dppLSTM [12] | 38.6 | 54.7 | supervised |
| SUM-GAN$_{sup}$ [5] | 41.7 | 56.3 | supervised |
| DR-DSN$_{sup}$ [6] | 42.1 | 58.1 | supervised |
| Li et al. [63] | 43.1 | 52.7 | supervised |
| M-AVS [7] | 44.4 | **61.0** | supervised |
| H-RNN [14] | 44.3 | **62.1** | supervised |
| Uniform samping | 29.3 | 15.5 | unsupervised |
| K-medoids | 33.4 | 28.8 | unsupervised |
| Elhamifar et al. [64] | 37.8 | 42.0 | unsupervised |
| Song [20] | - | 50.0 | unsupervised |
| SUM-GAN [5] | 39.1 | 51.7 | unsupervised |
| DR-DSN [6] | 41.4 | 57.6 | unsupervised |
| **Ours**$_{pol}$ | 44.48 | 56.40 | unsupervised |
| **Ours**$_{Q}$ | 44.56 | 56.44 | unsupervised |
| **Ours**$_{AC}$ | **46.40** | **58.30** | unsupervised |

TABLE 6: 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 using basic RL rewards. Mentioned results are from respective original papers. We choose 5 fold validation (fixed five splits of both the dataset by the script provided by [6]) and reported an average F-score for all the proposed frameworks.

sualization in Fig. 4 to highlight the same. Each bar indicates a frame selected for the summary from a temporal window of 70 frames in the video. We show qualitative results for the HUJI dataset in the supplementary material.

In Fig. 6, we compare 1 minute, 5 minutes, 10 minutes and 15 minutes summaries generated by our framework using the policy gradient method. As can be seen, our network can adapt to different desired summary lengths. We observe, and as expected, most of the frames present in the shorter summaries are also present in the longer ones along with some additional frames. In the supplementary material, we show comparative results on different length summaries generated by the Actor-Critic framework for the 'P04' sequence of the UTE dataset.

Fig. 7 shows the qualitative analysis of the interactive summarization using Interactive Summarization reward along with the basic RL rewards. From the visualization, it is evident that the summary is indeed biased towards user feedback. Similarly, we demonstrate the interactive summarization framework on the 'P01' video sequence of the UTE dataset in the supplementary material.

Table 5 shows the results from a user study as discussed in the evaluation section. It is evident that the users like personalized summaries generated by our method.

### 4.5 Results on Short Hand-held Videos

Though not the focus of this paper, we also evaluate our method over short hand-held videos. Table 6 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 made a comparison and except for H-RNN [14] and M-AVS [7], where we perform close, our method improved SOTA supervised techniques as well.

Comparing the performance of three configurations of our technique corresponding to different RL optimization techniques, we observe that Q learning performs better than the policy gradient, and the actor-critic performs better than Q learning. The policy gradient uses a baseline function that reduces the cumulative reward variance and leads to smaller gradients. In contrast, the Q learning and actor-critic techniques use a Q-value network instead of a baseline

| Datasets | SumMe | | | TVSum | | |
|---|---|---|---|---|---|---|
| Methods | DIST | IND | Both | DIST | IND | Both |
| Policy Gradient | 44.5 | 44.74 | **44.76** | 56.1 | 56.3 | **56.40** |
| Q Learning | 45.1 | 45.2 | **45.62** | 55.72 | 55.72 | **56.44** |
| Actor-Critic | 46.36 | 46.48 | **46.40** | 55.77 | 56.66 | **58.30** |

TABLE 7: The table shows the F-scores measure of different techniques for various combinations of rewards for SumMe and TVSum datasets. DIST and IND represent the Distinctiveness and Indicativeness rewards, respectively. We choose 5 fold validation (fixed five splits of both the dataset by the script provided by [6]) and reported an average F-score for all the experiments.

| Datasets | Disney | | | UTE | | |
|---|---|---|---|---|---|---|
| Methods | DIST | IND | Both | DIST | IND | Both |
| Policy Gradient | 26.77 | 27.23 | **28.54** | 42.87 | 43.4 | **45.15** |
| Q learning | 24.24 | 25.77 | **26.05** | 41.91 | 42.39 | **44.57** |
| Actor-Critic | 27.36 | 28.99 | **29.60** | 45.27 | 45.79 | **47.68** |

TABLE 8: The table shows the average RFS-50 (Relaxed F Score with temporal relaxation of 50) for three video sequences of Disney and UTE datasets for different rewards. DIST and IND represent the Distinctiveness and Indicativeness rewards, respectively. Note that the summary length reward is fixed to generate 5 minutes summary for all the experiments.

function to calculate TD error. This ensures higher gradients across multiple video samples, leading to better and faster reward maximization. To support the above claim, we show training plots corresponding to different RL techniques in the supplementary material. We also give a detailed comparison and discuss the stability of the experimented RL techniques in the supplementary material.

### 4.6 Ablation Study using various rewards

We have conducted extensive experiments to demonstrate the contribution of each reward in the final summary. We consider two basic rewards, namely distinctiveness, and indicativeness rewards, and did all the ablation for small and day long datasets in Table 7 and 8 respectively. The results show that both rewards individually cater complementary information, and when used together, we get performance improvement in all the experimental setups. For other plugins such as social interaction and interactive summarization, we did an extensive qualitative analysis. Furthermore, user feedback for interactive summarization is inherently subjective and dynamic, so we can not demonstrate any quantitative analysis.

## 5 Conclusion

In this paper, we have proposed a reinforcement learning based technique to generate personalized summaries of day long egocentric videos. Ours is the first technique with the capability to summarize such long sequences. We train our model end-to-end in a completely unsupervised manner and demonstrate the scalability of our technique on Disney, UTE, and HUJI datasets. To claim the superiority of our technique, we have performed extensive quantitative and qualitative evaluation, demonstrating significant improvement over SOTA results on long and short video sequences. Our framework allows the inclusion of various kinds of rewards in a plug-and-play manner, which can influence the selection of frames for the summary. We have shown the performance of our framework using visual diversity, representativeness, social saliency, faces, and summary length-based rewards. We also demonstrated how these rewards could be exploited to incorporate exemplar-based user preferences. The major limitation of our approach seems to be its over-reliance on visual diversity and representativeness rewards. In contrast, humans seem to rely more on the surprise element of a sub-shot for it to be included in a summary. The natural effect of a sub-shot even if visually diverse, is often not asked to be included in a summary. In the future, we would like to bring in more causal aspects into the summary, which are sometimes brought in by story-driven summarization techniques.

## Acknowledgments

We express our gratitude to the participants and reviewers who spared time for this work, even during the pandemic. Their critical feedback was crucial in strengthening our paper. We also acknowledge Anupriya Tuli for her support and counsel in designing the user study. This work was supported in part by the DST, Government of India, under project id T-138. Pravin is supported by Visvesvaraya PhD fellowship from the Government of India.

## References

[1] GoPro, www.gopro.com. Accessed: 2018-09-03.
[2] Pivothead, www.pivothead.com. Accessed: 2018-09-03.
[3] SenseCam, www.microsoft.com/microsoft-hololens Accessed: 2018-09-03.
[4] Y. Zhang, X. Liang, D. Zhang, M. Tan, and E. P. Xing, "Unsupervised object-level video summarization with online motion auto-encoder," *Pattern Recognition Letters*, 2018.
[5] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *CVPR*, 2017.
[6] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *AAAI*, 2018.
[7] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *IEEE TCSVT*, 2019.
[8] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *IJCV*, 2015.
[9] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *CVPR*, 2013.
[10] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE TMM*, 2018.
[11] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *CVPR*, 2015.
[12] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *ECCV*, 2016.
[13] S. Lan, R. Panda, Q. Zhu, and A. K. Roy-Chowdhury, "Ffnet: Video fast-forwarding via reinforcement learning," in *CVPR*, 2018.
[14] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *ACMMM*, 2017.
[15] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *CVPR*, 2012.
[16] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *CVPR*, 2012.
[17] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, "Compact CNN for indexing egocentric videos," in *WACV*, 2016.
[18] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in *CVPR*, 2014.
[19] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *ECCV*, 2014.
[20] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVsum: Summarizing web videos using titles," in *CVPR*, 2015.

[21] A. Rathore, P. Nagar, C. Arora, and C. Jawahar, "Generating 1 minute summaries of day long egocentric videos," in *ACMMM*, 2019.
[22] Y. Zhang, X. Liang, D. Zhang, M. Tan, and E. P. Xing, "Unsupervised object-level video summarization with online motion auto-encoder," *Pattern Recognition Letters*, 2020.
[23] X.-D. Zhang, T.-Y. Liu, K.-T. Lo, and J. Feng, "Dynamic selection and effective compression of key frames for video abstraction," *Pattern recognition letters*, 2003.
[24] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, 2011.
[25] T. Liu and J. R. Kender, "An efficient error-minimizing algorithm for variable-rate temporal video sampling," in *Proceedings. IEEE International Conference on Multimedia and Expo*, 2002.
[26] L. J. Latecki, D. de Wildt, and J. Hu, "Extraction of key frames from videos by optimal color composition matching and polygon simplification," in *IEEE Workshop on Multimedia Signal Processing*, 2001.
[27] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," in *ACMMM*, 1998.
[28] F. Dirfaux, "Key frame selection to represent a video," in *ICIP*, 2000.
[29] H.-W. Kang and X.-S. Hua, "To learn representativeness of video frames," in *ACMMM*, 2005.
[30] H. Sundaram and S.-F. Chang, "Video skims: Taxonomies and an optimal generation framework," in *ICIP*, 2002.
[31] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," in *ICIP*, 2002.
[32] N. Peyrard and P. Bouthemy, "Motion-based selection of relevant video segments for video summarization," *Multimedia Tools and Applications*, 2005.
[33] H. Tang, V. Kwatra, M. E. Sargin, and U. Gargi, "Detecting highlights in sports videos: Cricket as a test case," in *IEEE ICME*, 2011.
[34] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *ACMMM*, 2002.
[35] L. Shi, I. King, and M. R. Lyu, "Video summarization using greedy method in a constraint satisfaction framework," in *ICDS*, 2003.
[36] N. Babaguchi, "Towards abstracting sports video by highlights," in *ICME*, 2000.
[37] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, "From keyframes to key objects: Video summarization by representative object proposal selection," in *CVPR*, 2016.
[38] X. Song, K. Chen, J. Lei, L. Sun, Z. Wang, L. Xie, and M. Song, "Category driven deep recurrent neural network for video summarization," in *Multimedia & Expo Workshops*, 2016.
[39] Y.-L. Lin, V. I. Morariu, and W. Hsu, "Summarizing while recording: Context-based highlight detection for egocentric videos," in *ICCVW*, 2015.
[40] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *CVPR*, 2016.
[41] H.-I. Ho, W.-C. Chiu, and Y.-C. Frank Wang, "Summarizing first-person videos from third persons' points of view," in *ECCV*, 2018.
[42] A. G. del Molino, X. Boix, J.-H. Lim, and A.-H. Tan, "Active video summarization: Customized summaries via on-line interaction with the user," in *AAAI*, 2017.
[43] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.
[44] Y. Zhang, M. Kampffmeyer, X. Zhao, and M. Tan, "Deep reinforcement learning for query-conditioned video summarization," *Applied Sciences*, 2019.
[45] P. Yousefi and L. I. Kuncheva, "Selective keyframe summarisation for egocentric videos based on semantic concept search," in *IPAS*, 2018.
[46] H. Jin, Y. Song, and K. Yatani, "Elasticplay: Interactive video summarization with dynamic time budgets," in *ACMMM*, 2017.
[47] B. Han, J. Hamm, and J. Sim, "Personalized video summarization with human in the loop," in *WACV*, 2011.
[48] K. Darabi and G. Ghinea, "Personalized video summarization using sift," in *ACM Symposium on Applied Computing*, 2015.
[49] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *CVPR*, 2015.
[50] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.
[51] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *PAMI*, 2012.
[52] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *PAMI*, 2017.
[53] C. Szepesvári, "Algorithms for reinforcement learning," *Morgan and Claypool*, 2009.
[54] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *ICML*, 2016.
[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *PAMI*, 2017.
[56] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, 2016.
[57] S. Yeung, A. Fathi, and L. Fei-Fei, "Videoset: Video summary evaluation through text," *arXiv:1406.5824*, 2014.
[58] A. Garcia del Molino, J.-H. Lim, and A.-H. Tan, "Predicting visual context for unsupervised event segmentation in continuous photo-streams," in *ACMMM*, 2018.
[59] S. Sharma, L. E. Asri, H. Schulz, and J. Zumer, "Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation," *arXiv:1706.09799*, 2017.
[60] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Automatic video summarization by graph modeling," in *ICCV*, 2003.
[61] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE TMM*, 2005.
[62] M. D. C. Tongco, "Purposive sampling as a tool for informant selection," *Ethnobotany Research and applications*, 2007.
[63] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE TIP*, 2017.
[64] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *CVPR*, 2012.

**Pravin Nagar** is a Ph.D. candidate at Computer Vision and Machine Learning (CVML) Lab, IIIT Delhi, India. His dissertation is focused on analyzing day long egocentric videos using deep learning frameworks. His interest lies in Video Analytics using various frameworks including LSTM, Memory Network, and Deep Reinforcement Learning. He has received Visvesaraya Ph.D. Fellowship from the Government of India.

**Anuj Rathore** is currently pursuing a Masters by Research in Computer Science at the Center for Visual Information Technology (CVIT) lab in IIIT Hyderabad, India. He holds B.Tech (Hons) in Computer Science from IIIT Hyderabad, where he worked in the fields of Computer Vision (CV) and Machine Learning (ML). His research interests include - Egocentric Vision, Deep Learning, Reinforcement Learning, Computer Vision.

**C.V. Jawahar** is a professor at IIIT Hyderabad, India. He received his PhD from IIT Kharagpur and has been with IIIT Hyderabad since Dec. 2000. At IIIT Hyderabad, Jawahar leads a group focusing on computer vision, machine learning and multimedia systems.

**Chetan Arora** Chetan Arora received his Bachelor's degree in Electrical Engineering in 1999 and the Ph.D. degree in Computer Sc. in 2012 both from IIT Delhi. From 2000-2009 he was an entrepreneur involved in setting up companies working on various computer vision based products. From 2012 to 2014 he was a post doctoral researcher at Hebrew University. He is currently an Associate Professor at IIT Delhi where he leads a group focusing on computer vision and machine learning systems.