Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization

Jia Xu[†], Lopamudra Mukherjee[§], Yin Li[‡], Jamieson Warner[†], James M. Rehg[‡], Vikas Singh[†]
[†]University of Wisconsin-Madison, [§]University of Wisconsin-Whitewater
[‡]Georgia Institute of Technology
http://pages.cs.wisc.edu/~jiaxu/projects/ego-video-sum/

Abstract

With the proliferation of wearable cameras, the number of videos of users documenting their personal lives using such devices is rapidly increasing. Since such videos may span hours, there is an important need for mechanisms that represent the information content in a compact form (i.e., shorter videos which are more easily browsable/sharable). Motivated by these applications, this paper focuses on the problem of egocentric video summarization. Such videos are usually continuous with significant camera shake and other quality issues. Because of these reasons, there is growing consensus that direct application of standard video summarization tools to such data yields unsatisfactory performance. In this paper, we demonstrate that using gaze tracking information (such as fixation and saccade) significantly helps the summarization task. It allows meaningful comparison of different image frames and enables deriving personalized summaries (gaze provides a sense of the camera wearer's intent). We formulate a summarization model which captures common-sense properties of a good summary, and show that it can be solved as a submodular function maximization with partition matroid constraints, opening the door to a rich body of work from combinatorial optimization. We evaluate our approach on a new gaze-enabled egocentric video dataset (over 15 hours), which will be a valuable standalone resource.

1. Introduction

The advent of wearable cameras and the ability to record visual data from a first person point of view (namely, egocentric video) has opened the door to a rich trove of computer vision problems. These range from socio-behavioral modeling to analyzing recurring patterns in a person's daily life. Such a wealth of data poses an interesting scientific question — how should one compactly summarize continuous video streams acquired over many hours? A mature body of work on video summarization provides a meaningful starting point, but egocentric videos still pose unique challenges. We want to support continuous egocentric video capture, which will result in long segments, only a few subsets of which will actually contain 'memorable' or 'interesting' content. Further, simple measures of diversity among frames and low-level appearance or flow cues which are useful modules of a classical approach to video summarization may not be informative at all, in fact, even misleading. For example, strong motion cues and potentially strong differences among frames due to background clutter will show up prominently in a sequence of a long walk back from campus. The ideal solution would be to compress such redundant periods but also not leave out anomalies or shorter segments that may be interesting to the camera wearer.

The description above suggests that egocentric video summarization is an ill-posed problem. Indeed, these videos may have poor illumination, camera shake, rapidly changing background, and a spectrum of other confounding factors. Nonetheless, given that the proliferation of wearable image-capture systems will only increase, there is a need for systems that take a long egocentric video and distill it down to its informative parts. They offer the camera wearer the ability to browse/archive his/her daily activities (life log), and review (or search) it in the future. The last two years have seen a number of interesting strategies for this problem. For instance, [17] observed that canonical viewpoints of objects that are relevant for representation in a egocentric summary can be identified by mining large collections of images on the Internet. Very recently, [31] proposed regularizing the summarization process with a so-called "storyline narrative": a coherent (chronological) set of video subshots. Both approaches have been shown to work well but need a nominal amount of training data, which can be very expensive to collect and limited at scale.

Despite the advances described above, the literature on this problem is still in its developmental phase. Approaches so far have not attempted to *personalize* the summary. But, egocentric video summarization *is* subjective and its utility

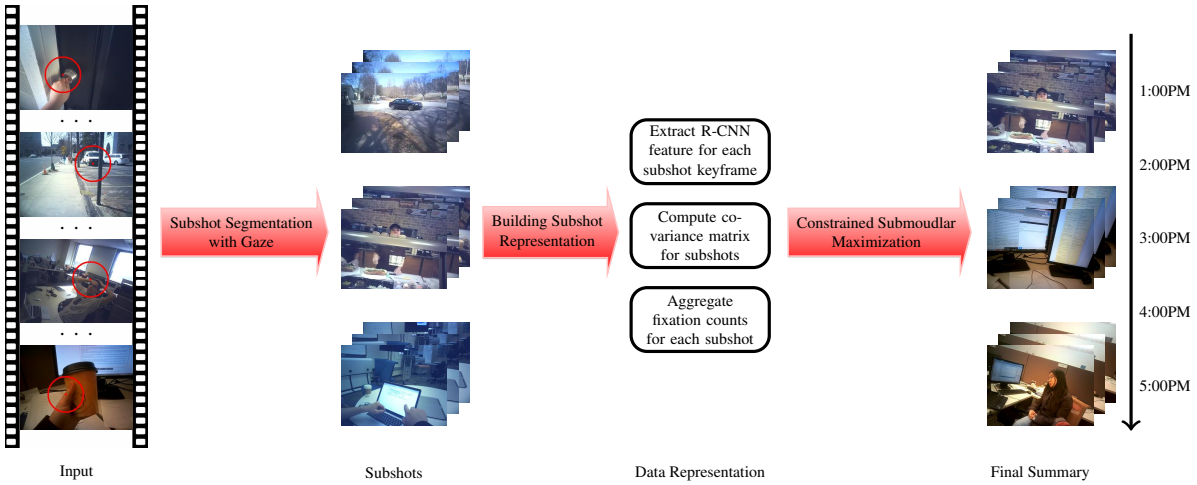

Figure 1. Overview of our summarization algorithm: our approach takes an egocentric video with gaze tracking as input (first column), time windows (last column) as a partition matriod constraint, and produces a compact personalized visual summary: getting lunch, working in an office, and conversation with a colleague.

depends greatly on its relevance to the camera wearer. The challenge is that personalization cannot be accomplished without close involvement of the user. This paper makes the case that a powerful surrogate to personalization is *egocentric gaze*. In fact, how the person views the world through a sequence of gaze measurements conveys a strong sense of his/her intent and interest. In a recent study [46], eye movements were found to inform visual understanding in different but complementary ways. For instance, relative importance of content in an image correlates to how a person's attention was spatially and temporally distributed (the patterns of saccades and fixations). We contend that egocentric gaze measurements are a key missing ingredient in egocentric video summarization – they serve to make the problem well posed, enable comparisons across frames (even with clutter) and provide guidance on which content the user would like to leave out. It turns out that such gaze measurements are now available as wearable devices, as small form-factor attachments (e.g. Pupil Labs [16]) and/or can be predicted in a egocentric sequence via a combination of saliency and machine learning techniques [25].

In this paper, we address the issue of incorporating gaze information to efficiently summarize egocentric videos (Fig. 1 outlines the overview of our algorithm). Our main contributions are: **(i)** We make the first attempt to study the role of gaze in summarizing egocentric videos. To our knowledge, our results are the first to demonstrate that gaze gives the means to 'personalize' the synopsis of a long egocentric sequence which leads to results that are more *relevant* to the camera wearer — arguably, the primary measure of a summary's utility. **(ii)** On the modeling side, gaze helps make the problem well-posed. This leads to a property that is taken as granted in a standard computer vision problem but difficult to achieve with summarization objectives – that a better evaluation of the objective function indeed corresponds to a more meaningful summary. We formulate a summarization model which captures commonsense properties of a good summary: relevance, diversity, fidelity with the full egocentric sequence, and compactness. The optimization scheme is an adaptation of recent work on non-monotone submodular maximization with matroid constraints and comes with approximation guarantees. **(iii)** We introduce a new dataset with 21 egocentric videos. Each video comes with calibrated gaze information, a summary annotation from the wearer as well as human experts.

2. Related Work

We first provide a brief review of literature from a few different lines of work that are related to this paper.

Video Summarization. The problem of video summarization has been studied from various perspectives [43, 21, 34]. Most methods select a sequence of keyframes [43, 12, 17] or subshots [31, 14, 13] to form a visual summary of the most informative parts of a video. Previous summarization techniques are designed for professionally produced videos and rely on low-level features [21] and motion cues [43]. Some recent approaches extract scenes of interest by training a supervised model of important objects [29, 17], attention models [32], user preferences [2], events [42], multi-view [8], and user interactions [12, 36, 4]. These methods are general and usually do not perform well for user-shot videos or egocentric sequences, and so recent works [17, 23, 31] have investigated and offered specialized solutions. Other recent works offering various interesting improvements an/or directions include [47, 37, 45].

Egocentric Video Analysis. Egocentric vision has attracted

a great deal of interest in the last few years for applications such as activity detection and recognition [5, 35, 39], object detection and segmentation [24, 38], temporal segmentation and activity classification [41], and novel event detection [1]. While several works have discussed potential uses of such sequences as a daily-log, summarization strategies have only appeared recently [23, 31]. This paper complements these developments by introducing gaze measurements as an alternative to direct user supervision.

Gaze in Computer Vision. Attention is an integral part of the human visual system and has been widely studied [46]. Previous works have demonstrated the utility of gaze in object segmentation [33, 44, 25], action recognition [6] and action localization [40]. Gaze measurements contain importance cues regarding the most salient objects in the scene [25] and the intent of the camera-wearer. These cues help the task of video analysis and can help overcome poor illumination and background clutter.

Submodular Optimization. Submodular function optimization is a well studied topic in theoretical computer science. It has also been heavily explored, albeit in a specialized form for labeling (energy minimization) problems in vision. Maximization of submodular functions has not found many applications in vision, although it has received much interest recently in machine learning [19, 18, 30, 15, 27, 28]. A small but interesting body of papers has shown how submodular function optimization (either unconstrained or with knapsack constraints) can be used to model problems like sensor placement [19, 18], feature selection [30] and document summarization [27, 28].

3. Submodular Video Summarization

We now introduce our approach to gaze-enabled video summarization via submodular maximization. The starting point is to decompose a continuous video record into subshots which will form the basis for our optimization approach. We perform gaze-enabled subshot selection and extract feature representations for each subshot (we discuss this procedure in detail in our experimental section). Let the set of all subshots be $\mathcal{V} = \{v_1, v_2, \cdots, v_n\}$. Our objective is to choose a subset $\mathcal{S} \subseteq \mathcal{V}$, denoted as the summary of $\mathcal{V}$. In the following section, we formulate the video summarization problem as the maximization of a submodular function subject to some constraints.

As with any summarization task [28], we expect a summary to be a good representative of the video — informative but compact. These goals can be achieved by choosing a subset that maximizes two key properties, namely *relevance* and *diversity*, where the relevance term encourages the inclusion of important events from the larger video (i.e., coverage of the sequence), while diversity reduces redundancy in the summary. We define a few concepts and then give the precise forms of these terms.

Definition 3.1. A set function F is monotone nondecreasing if $\mathcal{A} \subseteq \mathcal{S}$, $F(\mathcal{A}) \leq F(\mathcal{S})$.

Definition 3.2. For any $\mathcal{A} \subseteq \mathcal{S} \subseteq \mathcal{V}$ and $i \in \mathcal{V}, i \notin \mathcal{S}$. $F(\mathcal{S})$ is submodular if $F(\mathcal{S}+i)-F(\mathcal{S}) \leq F(\mathcal{A}+i)-F(\mathcal{A})$.

3.1. Relevance and Diversity Measurement with Mutual Information

Intuitively, we want to select subshots which are most informative with respect to the entire video, that is, if given an ideal summary $\mathcal{S}$, the knowledge of $\mathcal{V}$ is maximized, compared to any other subset of $\mathcal{V}$. A natural notion to quantify this is to minimize the conditional entropy function $H(\mathcal{V}\backslash\mathcal{S}|\mathcal{S})$. Unfortunately, several works [19] have shown that it can sometimes lead to suboptimal results. The reason is that conditional entropy is defined as $H(\mathcal{V}\backslash\mathcal{S}|\mathcal{S}) = H(\mathcal{V}) - H(\mathcal{S})$, therefore optimizing such a function is equivalent to maximizing $H(\mathcal{S})$. So, it only considers the entropy of the selected subshots, rather than taking the coverage over the entire video into account.

Instead, we want a criterion that helps identify the subset of subshots that most significantly reduce the uncertainty about the remainder of the sequence. Mutual information offers precisely this behavior. Specifically, we define our first objective as the mutual information between the sets $\mathcal{S}$ and $\mathcal{V}\backslash\mathcal{S}$,

$$\begin{aligned} M(\mathcal{V}\backslash\mathcal{S};\mathcal{S}) &= H(\mathcal{V}\backslash\mathcal{S}) - H(\mathcal{V}\backslash\mathcal{S}|\mathcal{S}) \\ &= H(\mathcal{V}\backslash\mathcal{S}) + H(\mathcal{S}) - H(\mathcal{V}) \end{aligned} \tag{1}$$

The optimal solution $\mathcal{S}^* = \text{argmax}_S M$ obtains the maximum entropy over both the selected sequence $\mathcal{S}^*$ and the remaining sequence $\mathcal{V}\backslash\mathcal{S}^*$, as desired.

Next, we discuss how to compute this score for a summary $\mathcal{S}$. Let L be the $n \times n$ covariance matrix of the set of subshots $\mathcal{V}$, which we assume are Gaussian random variables. For $\mathcal{S} \subseteq \mathcal{V}$, let $L_\mathcal{S}$ be the principal submatrix of L indexed by $\mathcal{S}$. It is well known that the entropy $H(\mathcal{S})$ of the random variables indexed by $\mathcal{S}$ can be computed as

$$H(\mathcal{S}) = \frac{1+\log(2\pi)}{2}|\mathcal{S}| + \frac{1}{2}\log(\det(L_\mathcal{S})) \tag{2}$$

Then maximizing the mutual information is equivalent to maximizing

$$M(\mathcal{S}) = \frac{1}{2}\log(\det(L_{\mathcal{V}\backslash\mathcal{S}})) + \frac{1}{2}\log(\det(L_\mathcal{S})) \tag{3}$$

as $|\mathcal{S}| + |\mathcal{V}\backslash\mathcal{S}| = n$, and $H(\mathcal{V})$ is constant. Here, the first term of M measures the information we have for the subshots we do not select, which is equivalent to the relevance we want to measure.

Relation to Determinantal Point Process. In the limit, it might seem that the relevance function will encourage coverage of the entire video in the summary because it does not

necessarily preclude inclusion of subshots which are very similar. If this happens, the summary will contain identical or very similar (redundant) segments, which are not indicative of a good summary. However, it turns out because of a special property of our relevance function, inclusion of such redundant frames will be discouraged in the summary. Note that the second term of our objective M has the same functional form as the well-known determinantal point processes (DPPs)[20]. Proposed by Kulesza and Taskar [20], DPPs use the log determinant function to measure the volume spanned by columns of of a subset $\mathcal{S}$ in $\mathcal{V}$. Recent research has also shown encouraging results on standard video summarization by extending it in a sequential manner [13]. It is often used as a means to devise tractable algorithms to measure (and also optimize) diversity in a given set, because maximizing the determinant automatically encourages a bigger volume which in turn implies that the columns of $\mathcal{S}$ are close to orthogonal or uncorrelated which encourages diversity in the elements of $\mathcal{S}$. As a result, our objective function M not only measures relevance but also implicitly encourages diversity in the obtained summary.

3.2. Attention Measurement using Gaze Fixations

For egocentric video, another important source of information is the human point of interest in the video. Past research [23, 31, 14] has developed sophisticated strategies to estimate which regions are important and which frames convey useful information towards a good summary. However, if we have access to gaze information, we have an alternative to such complex preprocessing steps. We explore how the pattern of a subject's gaze can inform the generation of meaningful summarizations. Here, for each subshot, we compute the attention score c_i by counting the number of frames containing fixations. This is similar to the interestingness [14] and importance [23] measure from recent work, though, our proposal is a more natural measurement of interest characterizing how much attention this subshot attracted from the user. We use this to define an additional term $I(\mathcal{S})$ in the objective as follows.

$$I(\mathcal{S}) = \sum_{i \in \mathcal{S}} c_i \tag{4}$$

3.3. Partition Matroid Constraint

In reality, we want our summary to reflect human preference in terms of subshot allocation. For instance, users usually want to allocate more subshots in a summary when more interesting things happen (e.g., in Disneyland) than when less interesting ones happen (e.g. in an office). An ideal summary should respect the compactness while maintaining a user-preferred distribution, if available.

To achieve this goal, we incorporate a partition matroid into our model. First, we partition the video into b disjoint blocks $\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_b$. Given the amount of user preference in each block, we specify an upper bound on the number of sub-shots that can be included from that block. Now, since each block pertains to a subset of subshots, and blocks are mutually disjoint, such a partitioning can be denoted as a matroid, $\mathcal{M} = (\mathcal{V}, \mathcal{P})$ where the set of subshots $\mathcal{V}$ is the ground set, and the blocks are 'independent' subsets of $\mathcal{V}$. We can associate each of the b blocks with an integer, $\{f_1, \cdots, f_b\}$ and ask that no more than f_i subshots be selected from each block. This requirement can be imposed by combinatorial structure known as the partition matroid [9], $\mathcal{I} = \{\mathcal{A} : \quad |\mathcal{A} \cap \mathcal{P}_m| \leq f_m, m = 1, 2, \cdots, b\}$.

3.4. Full Model

Putting all the above pieces together, we can set up a simple model for summarization,

$$\begin{aligned} \max_{\mathcal{S}} \quad & \log(\det(L_{\mathcal{V}\setminus\mathcal{S}})) + \log(\det(L_{\mathcal{S}})) + \lambda \sum_{i \in \mathcal{S}} c_i \\ \text{s.t.} \quad & \mathcal{S} \in \mathcal{I} \end{aligned} \tag{5}$$

where λ is a positive trade-off coefficient and the feasible set $\mathcal{I}$ corresponds to the partition matroid introduced above. We denote our full objective function as $F(\mathcal{S})$.

4. Optimization

Our model in (5) captures various desirable properties of a good summary, but this constrained combinatorial problem is in general difficult to optimize directly. There are three pertinent issues: objective function, monotonicity, and constraints.

First, we analyze our objective function. The mutual information criteria in (1) is difficult to optimize globally. Fortunately, as shown in various recent works on submodular optimization [19], $M(\mathcal{S})$ in (3) is submodular. The other term is a linear sum over positive scalars, hence is also submodular. Since the sum of two submodular functions is also submodular, our objective function in (5) is submodular. Second, when we analyze the monotonicity properties, we see that the mutual information term makes the objective non-monotone. While monotone objectives have better approximation guarantees, non-monotone objectives allow us to provide a richer model. For instance, with a monotone objective, we see that an upper bound constraint will always become tight (even if parts of the summary are redundant) because the model incurs no penalty for including an additional subshot. In contrast, our model prevents adding redundant subshots when the key information is already there (as shown in Fig. 5). Overall, our model is a constrained optimization model. Maximizing submodular functions subject to arbitrary linear constraints is very difficult. However, if the constraints are expressible as a knapsack constraint or a constraint over matroids, recent work

Algorithm 1 Local Search for Constrained Submodular Maximization

Input: $\mathcal{M} = (\mathcal{V}, \mathcal{I}), F, \epsilon \geq 0$
Initialize $\mathcal{S} \leftarrow \emptyset$;
while (Any of the following local operations applies, update $\mathcal{S}$ accordingly) **do**
  Add operation. If $e \in \mathcal{V} \backslash \mathcal{S}$ such that $\mathcal{S} \cup \{e\}) \in \mathcal{I}$ and $F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) > \epsilon$, then $\mathcal{S} = \mathcal{S} \cup \{e\}$.
  Swap operation. If $e_i \in \mathcal{S}$ and $e_j \in \mathcal{V} \backslash \mathcal{S}$ such that $(\mathcal{S} \backslash \{e_i\}) \cup \{e_j\} \in \mathcal{I}$ and $F((\mathcal{S} \backslash \{e_i\}) \cup \{e_j\}) - F(\mathcal{S}) > \epsilon$, then $\mathcal{S} = (\mathcal{S} \backslash \{e_i\}) \cup \{e_j\}$.
  Delete operation. If $e \in \mathcal{S}$ such that $F(\mathcal{S} \backslash \{e\}) - F(\mathcal{S}) > \epsilon$, then $\mathcal{S} = \mathcal{S} \backslash \{e\}$.
end while
return $\mathcal{S}$;

[22, 7] from combinatorial optimization provides a variety of strategies.

Motivated by ideas from [22, 7], we propose a local search algorithm, which requires no rounding. All intermediate solutions are integral. There are three key local operations: add, swap, and delete. The key idea is to iteratively search over these possible operations until no improvement can be made. Alg. 1 outlines our algorithm. Our algorithm achieves an approximation factor with respect to the unknown optimal solution (details in the supplement).

Proposition 4.1. *Alg. 1 achieves a $\frac{1}{4}$-approximation factor for our constrained submodular maximization problem (5).*

It is useful to point out that in [10], an approximation algorithm for maximizing an unconstrained DPP (also non-monotone) was proposed. While the approximation factor in that work was looser than other known results, a salient property was that the algorithm was simple to implement and avoided the usual reduction to a multi-linear relaxation of the submodular function [3]. Algorithm 1 can be viewed as a generalization of the method in [10]. It is equally simple to implement and optimizes a similar (DPP type) objective function but permits inclusion of additional constraints (which is frequently difficult in submodular maximization).

5. Experimental Evaluations

We will begin by describing the datasets used in our experiments and follow it with a discussion of the other baselines used for comparison. Then, we will delve into the qualitative and quantitative evaluations of our approach.

5.1. Dataset Collection and Annotation

Given the type of data (egocentric videos + gaze) we require, there are very few existing publicly available benchmarks that can be used directly for our purposes. Here, we make use of the GTEA-gaze+ dataset which is the only public dataset with video and gaze. In addition, we will also present results on a newly collected dataset ($\sim$ 15 hours).

GTEA-gaze+ dataset. This dataset is designed for action recognition, though it can be used for summarization as well. It consists of 30 videos, each of which includes a meal preparation recording and lasts $12 \sim 20$ minutes. There are fine-grained action annotations available with this dataset. In addition, we ask human experts to generate summaries by grouping those action annotations, and asking them to select $5 \sim 15$ group of consequent segments (referred to as events or blocks), which they think are appropriate to summarize each video. These group level annotations serve as our ground truth summary T.

EgoSum+gaze dataset. We collected a new dataset of videos, acquired by 5 subjects wearing eye tracking devices to record their daily lives in an uncontrolled setting, along with associated gaze information. We used a pair of SMI eye-tracking glasses[1] and a Pupil eye-tracking device[2] to collect gaze information. Our collection has 21 videos, each lasting 15 min $\sim$ 1.5 hour. To facilitate evaluations using this dataset, we obtained human annotations for the summary. To this end, we asked our subjects to select a set of events (blocks) in our videos. Each event constituted a sequence of similar subshots, and we assume that any one of them is an equally good representative of the event. This avoids unnecessarily penalizing an otherwise good summary which differs from the ground truth in the precise frame 'indexes' [45] but is perfectly consistent in a semantic sense. We asked each wearer to select events ($5 \sim 15$) which in their opinion should be included in a good machine generated summary. These annotations serve as the gold standard for our evaluation.

5.2. Subshot extraction and representation

A natural way to represent videos prior to summarization, is to extract subshots and compute their feature descriptors. We found that in general, this is challenging for egocentric videos since such videos are mostly continuous and therefore may not have a clear 'boundary' between shots. Fortunately, gaze turns out to be very useful for egocentric video temporal segmentation (see the first row of Fig. 2) which is used as follows.

We first extracted gaze tracking information (fixation, saccade or blink) for each frame (using our eye tracking device). Next, we removed frames with bad eye tracking data. This procedure provides $6000 \sim 9000$ segments with fixations per one hour of video. We picked the centroid frame as the *key-frame* in each segment, and extracted a feature descriptor around the gaze region (100×100) on this frame using R-CNN [11]. We then computed the cosine similarity

[1] http://www.eyetracking-glasses.com
[2] http://pupil-labs.com

between each key-frame using, $\kappa(i,j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}$ where $\mathbf{x}_i$ refers to the R-CNN feature vector of keyframe i. Next, we grouped consequent segments into subshots (Fig. 2, second row) by thresholding the neighborhood similarity distance at 0.5, which yields around 800 subshots per hour of video. Next, we picked the center key-frame from each subshot, and computed a R-CNN feature descriptor on the whole frame. This is the final descriptor $\mathbf{v}_k$ for subshot k used for our summarization algorithm.

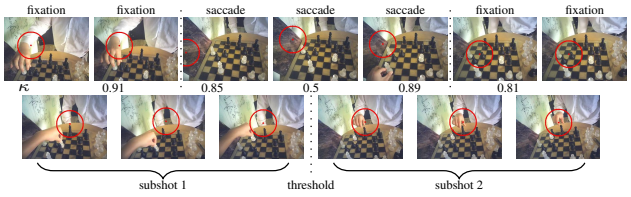

Figure 2. Illustration of our two-stage subshot extraction pipeline. First row: we extract gaze information for each frame, and filter frames with saccades/blinks and isolated fixations. Second row: we group similar neighbor key-frames by thresholding at 0.5 the similarities $\kappa(i, i+1)$ obtained from the gaze region.

5.3. Baselines

We adopted four baseline algorithms that do not require training summaries but are widely used for comparison purposes in video summarization [2, 17]. The first two are uniform sampling and k-means without gaze information. Our next two baselines incorporate *our specific* subshot segmentation and feature extraction processes into the previous two methods, which we refer as uniform (our subshots) and k-means (our subshots). In both cases, a (k subshot) summary is generated by selecting k equally separated subshots in uniform sampling and selecting the subshot closest to each of the k cluster centroid by k-means (reported by average of 20 runs). Unfortunately, direct comparisons to [23] and [31] were not possible due to the lack of available training data.

5.4. Evaluation

For our experiments, we set $\epsilon = 1e^{-6}$ and $\lambda = 0.001$. To enforce the compactness criteria, we divided our video into equal-sized blocks by time (the number of blocks depends on the length of the video), and set $f_i = 1$ for the ith block. Though it worked well in practice, this step can be substituted by more sophisticated procedures if desired. To perform quantitative evaluations, we computed F-measure scores on all summaries by comparing it with ground truth summaries. If a subshot in an output summary lies in a block/event of ground truth subshots, we count it as correct for that algorithm. This allows us to compute precision (P) values as $P = \frac{|\mathcal{S} \cap T|}{|\mathcal{S}|}$, where $\mathcal{S} \cap T$ is the set of subshots from the summary which can be found in a ground truth annotation. Similarly, we computed the recall as $R = \frac{|\mathcal{S} \cap T|}{|T|}$, where $|T|$ is equal to the number of events/blocks in the ground truth. The final F-measure is computed by $F = \frac{2PR}{P+R}$. Finally, note that the running time of our algorithm is 12 frames per second (fps), which is much faster than the time reported in [31]. Next, we analyze the goodness of the summary and the utility of gaze information, using our experimental results.

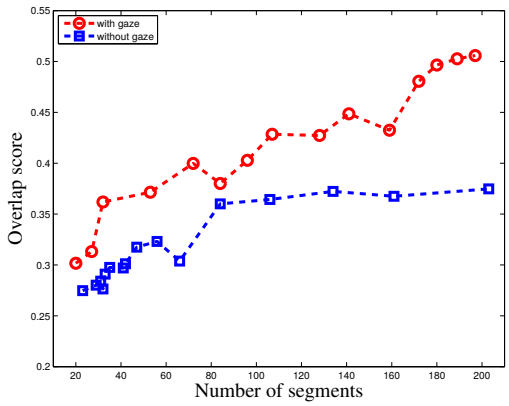

Figure 3. Comparison of temporal segmentation with/without gaze.

Is Gaze useful? We evaluated the utility of gaze in two different contexts, which is described next.

Temporal Segmentation. This is a critical preprocessing step, and is facilitated by the gaze information. To see how the quality of temporal segmentation is affected when gaze is *not* used, we performed temporal segmentation on the GTEA-gaze+ dataset simply by partitioning the video sequence into subshots whenever similarity (using R-CNN from the full frame) between two consecutive frames falls above a threshold. We compared our gaze-based (see Fig. 2) approach with this method, by computing the overlap with ground truth action segmentation provided with this dataset. The results in Fig. 3 show that our gaze-enabled temporal segmentation dominates the baseline which suffers due to the fact that egocentric videos are continuous. In other words, if we agree with the premise that a good temporal segmentation will help *any* egocentric video summarization method, our experiments provide empirical evidence that gaze information will almost certainly improve summarization results (by generating better temporal segments).

Relevance of Summary with and without Gaze. We analyze this issue both qualitatively and quantitatively using GTEA-gaze+ dataset. Fig. 4 shows results for a pizza preparation video summary. As we can see, without gaze information, uniform sampling and k-means pick many saccade frames (column 6 in the first two rows), which do not carry much content at all. However, when using

Figure 4. Results from GTEA-gaze+ data comparing the four baselines to our method (bottom row). This is from a pizza preparation video: cutting vegetables and meat (rows $1-2, 4$), frying (row 3), and adding sauce and toppings (rows $5-8$). Our algorithm picks these important subshots as they are with heavy attention.

Figure 5. Results from our new EgoSum+gaze dataset comparing the four baselines to our method (bottom row). In this video, our subject mixes a shake, drinks it, washes his cup, plays chess and texts a friend.

gaze-enabled subshot segmentation and representation, these baselines benefit significantly. As we can see in rows 3 and 4, each subshot captures a useful amount of information. For a more quantitative measurement, we looked at the F-measure scores in Tab. 1 and 2. Cols $2/4$ and $3/5$, which show these values for both baselines, and our method. There is a significant improvement in the F-measure score whenever gaze is utilized, which is further improved using our proposed algorithm discussed next.

Quality of Our Summarization Algorithm. On the pizza preparation video 4, we observed that our method outperforms the other baselines, even when all methods utilize gaze information. As is evident, both uniform sampling and k-means, still include irrelevant subshots (row $1, 5$). Our summary, on the other hand, constitutes the key stages in the meal preparation procedure: cutting vegetables and meat (rows $1-2, 4$), frying (row 3), and adding sauce and toppings (rows $5-8$). These subshots are selected mainly due to the fact that subjects focused on them, and our objective $I(\mathcal{S})$ picks these important subshots which contained substantial attention. Similar results can be seen with our new dataset (Fig. 5 and 6). As shown in row 2 and 4 in Fig. 5, in this uncontrolled setting, k-means ends up selecting many outliers. Also, row 3 in Fig. 6 shows uniform sampling fails when there are repeated scenes (e.g., washing dishes).

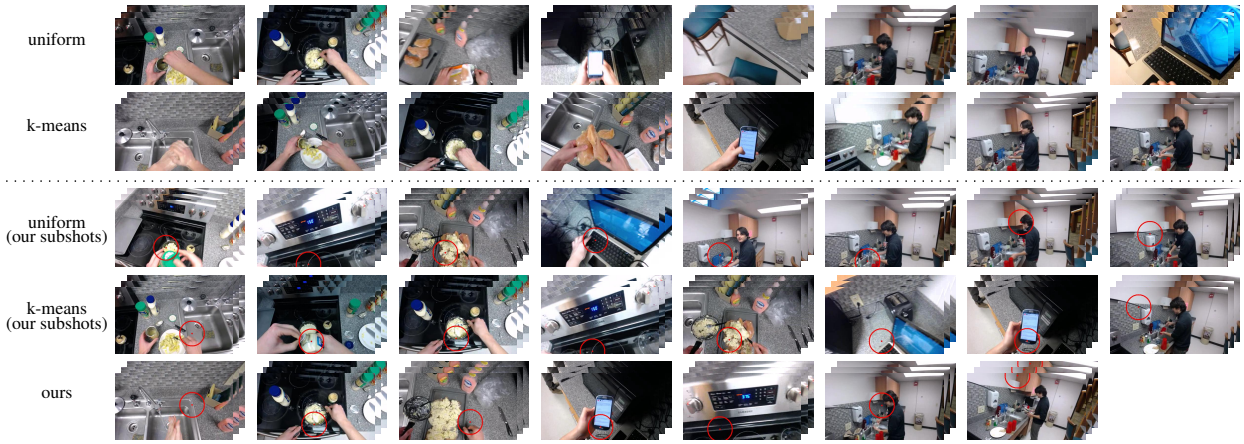

Figure 6. Results from our new EgoSum+gaze dataset comparing the four baselines to our method (bottom row). In this video, our subject is cooking chicken and have a conversation with his roommate.

Method	uniform	kmeans	uniform(our subshots)	kmeans(our subshots)	ours
F-measure	0.161	0.215 ± 0.016	0.526	0.475 ± 0.026	0.621

Table 1. Comparisions of average F-measure on GTEA-GAZE+.

Method	uniform	kmeans	uniform(our subshots)	kmeans(our subshots)	ours
F-measure	0.080	0.095 ± 0.030	0.476	0.509 ± 0.025	0.585

Table 2. Comparisions of average F-measure on our new EgoSum+gaze dataset.

Interestingly, our algorithm also achieves better compactness (fewer subshots in the summary) since our objective is non-monotone. The F-measures in Tab. 1 and 2 also provide evidence that our method using gaze outperforms the other baselines in both datasets.

Comparisons on other measures of quality. Note that we reported on F-measures here since they are widely used in computer vision. Separately, we also performed evaluations using summarization measures developed in the NLP literature such as ROUGE [26]. The main sequence of steps here is for a human to (a) annotate the full video in text and then separately (b) write a summary of the video. Our algorithm generates a sub-shot summary which can be mapped to an English summary using the corresponding sentences in the human's full annotation of the video, i.e., (a) above. Here, we found that if the human's sentences/words are nearly similar between his/her summary and the full annotation (e.g., as between (a) and (b)), then the system generated summary can indeed be meaningfully compared to the human's summary. In this case, the conclusions we can derive from ROUGE scores are similar to the ones we reported here using F-measure in Tab. 1 and 2. On the other hand, if the human's language usage in the full annotation and the summary written by him/her are different (in terms of word usage), then these scores are not very informative.

6. Conclusion

This paper introduced a new approach for egocentric video summarization, utilizing gaze information. We give strong results showing that gaze provides the means to "personalize" the summary, by focusing on what is important to the camera wearer. We formulate the corresponding summarization objective as a submodular function maximization that captures desirable and common-sense requirements of a summary such as relevance and diversity. The compactness property is enforced as a partition matroid constraint, and solved using a simple to implement local search method which offers approximation guarantees. Our experiments show that gaze information universally improves the relevance of a summary. For our experiments, we acquired a large set of gaze enabled egocentric video sequences, which is potentially valuable for future work on this topic.

Acknowledgments: We thank Jerry Zhu for introducing us to the DPP literature. This research is funded via grants NSF RI 1116584, NSF CGV 1219016, NSF Award 0916687, NSF EA 1029679, NIH BD2K award 1U54AI117924 and NIH BD2K award 1U54EB020404. We gratefully acknowledge NVIDIA Corporation for the donation of Tesla K40 GPUs used in this research.

References

[1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *Proc. CVPR*, 2011. 3

[2] J. Almeida, N. J. Leite, and R. da Silva Torres. VISON: VIdeo Summarization for ONline applications. *Pattern Recognition Letters*, 33(4):397–409, 2012. 2, 6

[3] C. Chekuri, J. Vondrák, and R. Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proc. STOC*, 2011. 5

[4] M. Ellouze, N. Boujemaa, and A. M. Alimi. Im(s)^2: Interactive movie summarization system. *JVCIR*, 21(4):283–294, 2010. 2

[5] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *Proc. ICCV*, 2011. 3

[6] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Proc. ECCV*, 2012. 3

[7] Y. Filmus and J. Ward. A tight combinatorial algorithm for submodular maximization subject to a matroid constraint. In *Proc. FOCS*, 2012. 5

[8] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z. Zhou. Multi-view video summarization. *IEEE Transactions on Multimedia*, 12(7):717–729, 2010. 2

[9] S. Fujishige. *Submodular functions and optimization*, volume 58. Elsevier, 2005. 4

[10] J. Gillenwater, A. Kulesza, and B. Taskar. Near-optimal map inference for determinantal point processes. In *Proc. NIPS*, 2012. 5

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014. 5

[12] D. B. Goldman, B. Curless, D. Salesin, and S. M. Seitz. Schematic storyboarding for video visualization and editing. *Proc. SIGGRAPH*, 25(3):862–871, 2006. 2

[13] B. Gong, W. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *Proc. NIPS*, 2014. 2, 4

[14] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *Proc. ECCV*, 2014. 2, 4

[15] R. Iyer and J. Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. In *Proc. NIPS*, 2013. 3

[16] M. Kassner and W. Patera. Pupil: Constructing the space of visual attention. Master thesis, Massachusetts Institute of Technology, Cambridge, MA, 2012. 2

[17] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *Proc. CVPR*, 2013. 1, 2, 6

[18] A. Krause and C. Guestrin. Submodularity and its applications in optimized information gathering. *ACM Transactions on Intelligent Systems and Technology*, 2(4):32, 2011. 3

[19] A. Krause, A. P. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 9:235–284, 2008. 3, 4

[20] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3), 2012. 4

[21] R. Laganire, R. Bacco, A. Hocevar, P. Lambert, G. Pas, and B. Ionescu. Video summarization from spatio-temporal features. In *ACM TRECVid Video Summarization workshop*, 2008. 2

[22] J. Lee, V. S. Mirrokni, V. Nagarajan, and M. Sviridenko. Maximizing nonmonotone submodular functions under matroid or knapsack constraints. *SIAM J. Discrete Math.*, 23(4):2053–2078, 2010. 5

[23] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proc. CVPR*, 2012. 2, 3, 4, 6

[24] C. Li and K. M. Kitani. Model recommendation with virtual probes for ego-centric hand detection. In *Proc. ICCV*, 2013. 3

[25] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *Proc. ICCV*, 2013. 2, 3

[26] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004. 8

[27] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *Proc. ACL*, 2011. 3

[28] H. Lin and J. Bilmes. Learning mixtures of submodular shells with application to document summarization. In *Proc. UAI*, 2012. 3

[29] D. Liu, G. Hua, and T. Chen. A hierarchical visual model for video object summarization. *TPAMI*, 32(12):2178–2190, 2010. 2

[30] Y. Liu, K. Wei, K. Kirchhoff, Y. Song, and J. Bilmes. Submodular feature selection for high-dimensional acoustic score spaces. In *Proc. ICASSP*, 2013. 3

[31] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proc. CVPR*, 2013. 1, 2, 3, 4, 6

[32] Y.-F. Ma, X.-S. Hua, L. Lu, and H. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005. 2

[33] A. K. Mishra, Y. Aloimonos, L. F. Cheong, and A. A. Kassim. Active visual segmentation. *TPAMI*, 34(4):639–653, 2012. 3

[34] C.-W. Ngo, Y.-F. Ma, and H. Zhang. Automatic video summarization by graph modeling. In *Proc. ICCV*, 2003. 2

[35] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Proc. CVPR*, 2012. 3

[36] S. Pongnumkul, J. Wang, and M. F. Cohen. Creating map-based storyboards for browsing tour videos. In *UIST*, 2008. 2

[37] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *Proc. ECCV*. 2014. 2

[38] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *Proc. CVPR*, 2010. 3

[39] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *Proc. CVPR*, 2013. 3

[40] N. Shapovalova, M. Raptis, L. Sigal, and G. Mori. Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In *Proc. NIPS*, 2013. 3

[41] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classication from first-person sensing. In *IEEE Workshop on Egocentric Vision, CVPR*, 2009. 3

[42] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua. Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia*, 14(4):975–985, 2012. 2

[43] W. Wolf. Key frame selection by motion analysis. In *Proc. ICASSP*, 1996. 2

[44] J. Xu, M. D. Collins, and V. Singh. Incorporating User Interaction and Topological Constraints within Contour Completion via Discrete Calculus. In *Proc. CVPR*, 2013. 3

[45] S. Yeung, A. Fathi, and L. Fei-Fei. Videoset: Video summary evaluation through text. *arXiv:1406.5824*, 2014. 2, 5

[46] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *Proc. CVPR*, 2013. 2, 3

[47] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *Proc. CVPR*, 2014. 2