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

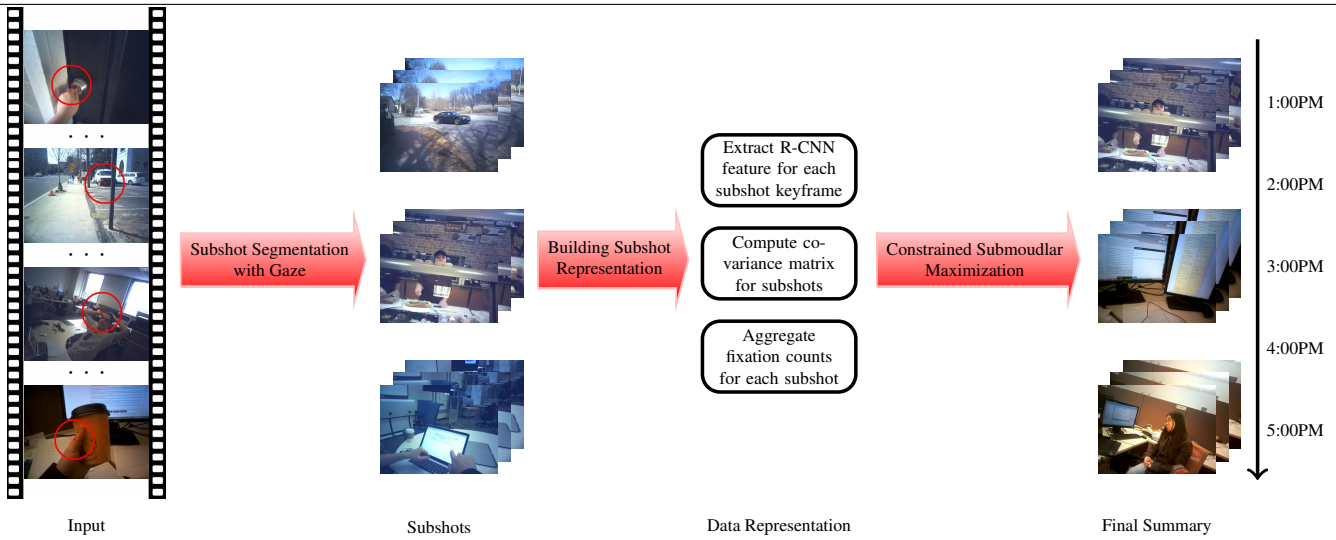

Figure 1: Overview of our summarization algorithm: our approach takes an egocentric video with gaze tracking as input (first column), time windows (last column) as a partition matriod constraint, and produces a compact personalized visual summary: getting lunch, working in an office, and conversation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

With the proliferation of wearable cameras, the number of videos from users documenting their personal lives using such devices is rapidly increasing. Since such videos can span hours, there is an important need for mechanisms to represent the information content in a compact form (i.e.,summary). 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, which makes standard video summarization tools yield unsatisfactory performance. Recent research shown promising results for this task [3, 6, 8], but they need a nominal amount of training data, which can be very expensive to collect and limited at scale.

In this paper, we address these issues by incorporating gaze information. We make the first attempt to study gaze in summarizing egocentric videos, motivated by recent research progress on gaze[7, 9]. To our knowledge, our results are the first to demonstrate that gaze provides the means to 'personalize' the synopsis of a long egocentric sequence, leading to results more *relevant* to the camera wearer — arguably, the primary measure of a summary's utility.

Our approach starts with temporally segmenting a whole egocentric video into subshots. 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 in this task, as human fixations are not continuous. We hence facilitate this preprocessing step by the use of gaze information.

Then, for each subshot, we compute the attention score c_i by counting the number of frames containing fixationx. This is similar to the interestingness [2] and importance [6] measure from recent work, though, our proposal is a more natural measurement of interest characterizing precisely how much attention this subshot attracted from the user. This also provides a form of weak supervision on the fly.

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 $\mathcal{V}$. Here we use the mutual information between the sets $\mathcal{S}$ and $\mathcal{V}\backslash\mathcal{S}$. We note one component of our objective has the same functional form as the well-known determinantal point processes (DPPs)[4], which use the log determinant function to measure the volume spanned by columns of a subset $\mathcal{S}$ in $\mathcal{V}$. It is often used as a means to devise tractable algorithms to measure (and also optimize) diversity in a given set. As a result, our objective function not only measures relevance but also implicitly encourages diversity in the obtained summary.

Next, we want our summary to reflect human preference in terms of subshot allocation. To achieve this goal, we employ a partition matroid into our model: $\mathcal{I} = \{\mathcal{A} : \quad |\mathcal{A} \cap \mathcal{P}_m| \leq f_m, m = 1, 2, \cdots, b\}$, which limits the overlap of the summary $\mathcal{A}$ with a given partition $\mathcal{P}_m$ of the video. Putting all these components together, we have a summarization model which captures common-sense properties of a good summary: relevance, diversity, fidelity with the full egocentric sequence, and compactness,

$$\max_{\mathcal{S}\in\mathcal{I}} \quad \log(\det(L_{\mathcal{V}\backslash\mathcal{S}})) + \log(\det(L_{\mathcal{S}})) + \lambda \sum_{i\in\mathcal{S}} c_i \tag{1}$$

Taking a close look at our model (1), we note our objective function in (1) is submodular, while non-monotone. Motivated by ideas from [1, 5], we propose a local search algorithm, which requires no rounding. We further prove our algorithm achieves an approximation factor with respect to the optimal solution. Our experiments show that gaze information universally improves the relevance of a summary. For our experiments, we acquire a large set of gaze enabled egocentric video sequences (15 hours), which is potentially valuable for future work on this topic.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

[1] Yuval Filmus and Justin Ward. A tight combinatorial algorithm for submodular maximization subject to a matroid constraint. In *FOCS*, 2012.
[2] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *Proc. ECCV*, 2014.
[3] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *Proc. CVPR*, 2013.
[4] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3), 2012.
[5] Jon Lee, Vahab S. Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. Maximizing nonmonotone submodular functions under matroid or knapsack constraints. *SIAM J. Discrete Math.*, 23(4):2053–2078, 2010.
[6] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Proc. CVPR*, 2012.
[7] Yin Li, Alireza Fathi, and James M. Rehg. Learning to predict gaze in egocentric video. In *Proc. ICCV*, 2013.
[8] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proc. CVPR*, 2013.
[9] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. Studying relationships between human gaze, description, and computer vision. In *Proc. CVPR*, 2013.