Discovering Important People and Objects for Egocentric Video Summarization

Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman
University of Texas at Austin
yjlee0222@utexas.edu, joydeep@ece.utexas.edu, grauman@cs.utexas.edu

1. Introduction

The goal of video summarization is to produce a compact visual summary that encapsulates the key components of a video. Its main value is in turning hours of video into a short summary that can be interpreted by a human viewer in a matter of seconds.

Existing methods extract keyframes [14, 15, 6], create montages of still images [1, 3], or generate compact dynamic summaries [11]. Despite promising results, they assume a static background or rely on low-level appearance and motion cues to produce the final summary. However, in many interesting settings, such as egocentric or YouTube style videos, the background is moving and changing. Furthermore, existing methods do not perform *object-driven* summarization and are indifferent to the impact that each object has on generating the "story" of the video.

In this work, we are interested in creating object-driven summaries for videos captured from a wearable camera. An egocentric video offers a first-person view of the world that cannot be captured from environmental cameras. For example, we can often see the camera wearer's hands, or find the object of interest centered in the frame. Essentially, a wearable camera focuses on the user's activities, social interactions, and interests. We aim to exploit these properties for egocentric video summarization.

Good summaries for egocentric data would have wide potential uses: They could facilitate police officers in reviewing important evidence, suspects, and witnesses, or aid patients with memory problems to remember specific events, objects, and people [7]. Furthermore, the egocentric view translates naturally to robotics applications—suggesting, for example, that a robot could summarize what it encounters while navigating unexplored territory, for later human viewing.

Motivated by these problems, we propose an approach that learns category-independent *importance* cues designed explicitly to target the *key objects and people* in the video. The main idea is to leverage novel egocentric and high-level saliency features to train a model that can predict important regions in the video, and then to produce a concise visual summary that is driven by those regions (see Fig. 1). By learning to predict important regions, we can focus the visual summary on the main people and objects, and ignore irrelevant or redundant information.

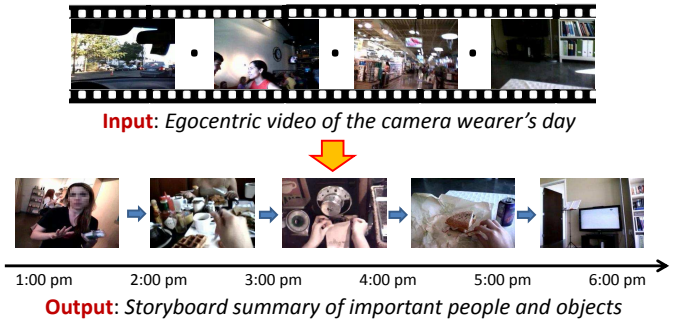

Figure 1. Our system takes as input an unannotated egocentric video, and produces a compact storyboard visual summary that focuses on the key people and objects in the video.

We emphasize that we do not aim to predict importance for any specific category (e.g., cars). Instead, we learn a general model that can predict the importance of any *object instance*, irrespective of its category. This category-independence avoids the need to train importance predictors specific to a given camera wearer, and allows the system to recognize as important something it has never seen before. In addition, it means that objects from the same category can be predicted to be (un)important depending on their role in the story of the video.

While recent work on egocentric visual analysis has shown many interesting applications (e.g., activity recognition [5], object recognition [12], and action clustering [8]), to our knowledge, we are the first to perform visual summarization for egocentric data. Please see our full CVPR 2012 paper and project page (http://vision.cs.utexas.edu/projects/egocentric/) for more algorithmic details and results.

2. Approach

Our goal is to create a storyboard summary of a person's day that is driven by the important people and objects. We define *importance* in the scope of egocentric video: important things are those with which the camera wearer has sig-

Figure 2. Illustration of our egocentric features.

nificant interaction.

Egocentric video data collection and annotation We use the Looxcie wearable camera, which is worn around the ear and captures video at 15 fps at 320 x 480 resolution. We collected 10 videos from four subjects, each of three to five hours in length, for a total of 37 hours of video. The videos capture a variety of activities such as eating, shopping, attending a lecture, driving, and cooking.

To train the importance predictor, we crowd-source large amounts of annotations using Amazon's Mechanical Turk (MTurk). For egocentric videos, the object must be seen in the context of the camera wearer's activity to properly gauge its importance.

We carefully design two annotation tasks to capture this aspect. In the first task, we ask workers to watch a three minute accelerated video and to describe in text what they perceive to be essential people or objects necessary to create a summary. In the second task, we display uniformly sampled frames from the video and their corresponding text descriptions *obtained from the first task*, and ask workers to draw polygons around any described person or object. This two-step process helps us avoid bias: a single annotator asked to complete both tasks at once may be biased to pick easier things to annotate rather than those s/he finds to be most important. For a 3-5 hour video, we obtain roughly 35 text descriptions and 700 object segmentations.

Learning and predicting region importance Given a video, we first generate candidate regions for each frame. For each region, we compute a set of candidate features that could be useful to describe its importance:

Egocentric features Fig. 2 illustrates the three proposed egocentric features. To model **interaction**, we compute the L2-distance of the region's centroid to the closest detected hand. To model **gaze**, we compute the L2-distance of the region's centroid to the frame center. To model **frequency**, we record the number of times an object instance is detected within a short temporal segment of the video.

Object features We include three high-level saliency cues. To model **object-like appearance**, we use the learned region ranking function of [2]. It reflects Gestalt cues indicative of *any* object and is useful for identifying full object segments, as opposed to fragments. To model **object-like motion**, we use the feature defined in [9]. It looks at the difference in motion patterns of a region relative to its closest surrounding regions. It is useful for selecting object-like regions that "stand-out" from their surroundings. To model the **likelihood of a person's face**, we compute the maximum overlap score between the region and any detected frontal face in the frame.

Region features Finally, we compute the region's **size**, **centroid**, **bounding box centroid**, **bounding box width**, and **bounding box height**. They reflect category-independent importance cues and are blind to the region's appearance or motion.

We next train a model that can learn and predict a region's *degree* of importance. While the features defined above can be individually meaningful, we expect significant interactions between the features. For example, a region that is near the camera wearer's hand might be important only if it is also object-like in appearance. Therefore, we train a linear regression model with pair-wise interaction terms to predict a region r's *importance score*:

$$I(r) = \beta_0 + \sum_{i=1}^{N} \beta_i x_i(r) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \beta_{i,j} x_i(r) x_j(r), \quad (1)$$

where the β's are the learned parameters, $x_i(r)$ is the ith feature value, and $N = 14$ is the total number of features.

For training, we define a region r's target importance score by its maximum overlap $\frac{|GT \cap r|}{|GT \cup r|}$ with any ground-truth region GT in a training video. We solve for the β's using least-squares. For testing, our model takes as input a region r's features and predicts its importance score $I(r)$.

Generating a storyboard summary We first partition the video temporally into events. We cluster scenes in such a way that frames with similar global appearance can be grouped together even when there are a few unrelated frames ("gaps") between them. Specifically, we perform complete-link agglomerative clustering with a distance matrix that reflects color similarity between each pair of frames weighted by temporal proximity.

Given an event, we first score each region in each frame using our regressor. We take the highest-scored regions and group instances of the same person or object together using a factorization approach [10]. For each group, we select the region with the highest score as its representative.

Finally, we create a storyboard visual summary of the video. We display the event boundaries and frames of the selected important people and objects (see Fig. 3). We automatically adjust the *compactness* of the summary with selection criteria on the region importance scores and number of events, as we illustrate in our results.

3. Results

We analyze (1) the performance of our method's important region prediction, and (2) the accuracy and compactness of our storyboard summaries. For evaluation, we use

Figure 3. Our summary (**a**) vs. uniform sampling (**b**). Our summary focuses on the important people and objects.

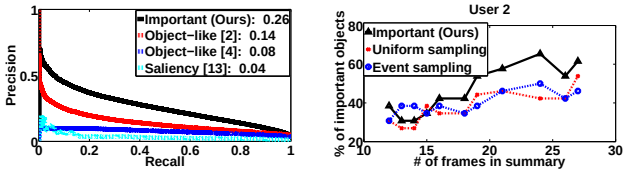

Figure 4. (**left**) Precision-Recall for important object prediction across all splits, and example selected regions/frames. Numbers in the legends denote average precision. (**right**) Comparison to alternative summarization strategies, in terms of important object recall rate as a function of summary compactness.

four data splits: for each split we train with data from three users and test on one video from the remaining user. For efficiency, we process every 15th frame.

Important region prediction accuracy We first evaluate our method's ability to predict important regions, compared to (1) the object-like score of [2], (2) the object-like score of [4], and (3) the bottom-up saliency detector of [13]. The first two are learned functions that predict a region's likelihood of overlapping a true object, whereas the low-level detector aims to find regions that "stand-out".

Fig. 4 (left) shows precision-recall curves on all test regions across all train/test splits. Our approach predicts important regions significantly better than all three existing methods. The two high-level methods can successfully find prominent object-like regions, and so they noticeably outperform the low-level saliency detector. However, by focusing on detecting *any* prominent object, unlike our approach they are unable to distinguish those that may be important to a camera wearer.

Egocentric video summarization accuracy Next we evaluate our method's summarization results. We compare against two baselines: (1) uniform keyframe sampling, and (2) event-based adaptive keyframe sampling. The latter computes events using the same procedure as our method, and then divides its keyframes evenly across events. These are natural baselines modeled after classic keyframe and event detection methods [14, 15], and both select keyframes that are "spread-out" across the video.

Fig. 4 (right) shows an example result. We plot *% of important objects found* as a function of *# of frames in the summary*, in order to analyze both the recall rate of the important objects as well as the compactness of the summaries. To vary compactness, our method varies both its selection criterion on $I(r)$ and the number of events, for 12 summaries in total. We create summaries for the baselines with the same number of frames as those 12.

Overall, our summaries include more important people/objects with fewer frames. For example, our method finds 54% of important objects in 19 frames, whereas the uniform keyframe method requires 27 frames. Fig. 3 shows an example full summary from our method (a) and the uniform baseline (b). Our summary more clearly reveals the story: *selecting an item at the supermarket* $\rightarrow$ *driving home* $\rightarrow$ *cooking* $\rightarrow$ *eating and watching tv*.

User studies to evaluate summaries To quantify the *perceived* quality of our summaries, we ask the camera wearers to compare our method's summaries to those generated by uniform keyframe sampling (event-based sampling performs similarly). In short, out of 16 total comparisons, our summaries were found to be better 68.75% of the time.

References

[1] A. Aner and J. R. Kender. Video Summaries through Mosaic-Based Shot and Scene Clustering. In *ECCV*, 2002.
[2] J. Carreira and C. Sminchisescu. Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *CVPR*, 2010.
[3] Y. Caspi, A. Axelrod, Y. Matsushita, and A. Gamliel. Dynamic Stills and Clip Trailer. In *The Visual Computer*, 2006.
[4] I. Endres and D. Hoiem. Category Independent Object Proposals. In *ECCV*, 2010.
[5] A. Fathi, A. Farhadi, and J. Rehg. Understanding Egocentric Activities. In *ICCV*, 2011.
[6] D. Goldman, B. Curless, D. Salesin, and S. Seitz. Schematic Storyboarding for Video Visualization and Editing. In *SIGGRAPH*, 2006.
[7] S. Hodges, E. Berry, and K. Wood. Sensecam: A Wearable Camera which Stimulates and Rehabilitates Autobiographical Memory. *Memory*, 2011.
[8] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast Unsupervised Ego-Action Learning for First-Person Sports Video. In *CVPR*, 2011.
[9] Y. J. Lee, J. Kim, and K. Grauman. Key-Segments for Video Object Segmentation. In *ICCV*, 2011.
[10] P. Perona and W. Freeman. A Factorization Approach to Grouping. In *ECCV*, 1998.
[11] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a Long Video Short. In *CVPR*, 2006.
[12] X. Ren and C. Gu. Figure-Ground Segmentation Improves Handled Object Recognition in Egocentric Video. In *CVPR*, 2010.
[13] D. Walther and C. Koch. Modeling Attention to Salient Proto-Objects. *Neural Networks*, 19:1395–1407, 2006.
[14] W. Wolf. Keyframe Selection by Motion Analysis. In *ICASSP*, 1996.
[15] H. J. Zhang, J. Wu, D. Zhong, and S. Smoliar. An Integrated System for Content-Based Video Retrieval and Browsing. In *Pattern Recognition*, 1997.