Predicting Important Objects for Egocentric Video Summarization

Yong Jae Lee · **Kristen Grauman**

Received: 5 January 2014 / Accepted: 4 December 2014 / Published online: 7 January 2015

Abstract We present a video summarization approach for egocentric or "wearable" camera data. Given hours of video, the proposed method produces a compact storyboard summary of the camera wearer's day. In contrast to traditional keyframe selection techniques, the resulting summary focuses on the most important objects and people with which the camera wearer interacts. To accomplish this, we develop region cues indicative of high-level saliency in egocentric video—such as the nearness to hands, gaze, and frequency of occurrence—and learn a regressor to predict the relative importance of any new region based on these cues. Using these predictions and a simple form of temporal event detection, our method selects frames for the storyboard that reflect the key object-driven happenings. We adjust the compactness of the final summary given either an importance selection criterion or a length budget; for the latter, we design an efficient dynamic programming solution that accounts for importance, visual uniqueness, and temporal displacement. Critically, the approach is neither camera-wearer-specific nor object-specific; that means the learned importance metric need not be trained for a given user or context, and it can predict the importance of objects and people that have never been seen previously. Our results on two egocentric video datasets show the method's promise relative to existing techniques for saliency and summarization.

Keywords Egocentric vision · Video summarization · Category discovery · Saliency detection

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. Automatic video summarization methods would be useful for a number of practical applications, such as analyzing surveillance data, video browsing, action recognition, or creating a visual diary of one's personal lifelog video.

Existing methods extract keyframes (Wolf 1996; Zhang et al. 1997; Goldman et al. 2006; Liu and Kender 2002), create montages of still images (Aner and Kender 2002; Caspi et al. 2006), or generate compact dynamic summaries (Rav-Acha et al. 2006; Pritch et al. 2007). Despite promising results, they assume a static background or rely on low-level appearance and motion cues to select what will go into the final summary. However, in many interesting settings, such as egocentric videos, YouTube style videos, or feature films, the background is moving and changing. More critically, a system that lacks high-level information on *which objects matter* may produce a summary that consists of irrelevant frames or regions. In other words, existing methods are indifferent to the impact that each object has on generating the "story" of the video.

Communicated by C. Schnörr.

Electronic supplementary material The online version of this article (doi:10.1007/s11263-014-0794-5) contains supplementary material, which is available to authorized users.

Y. J. Lee (✉)
Department of Computer Science,
University of California, Davis, CA, USA
e-mail: yjlee@cs.ucdavis.edu

K. Grauman
Department of Computer Science,
University of Texas at Austin, Austin, TX, USA
e-mail: grauman@cs.utexas.edu

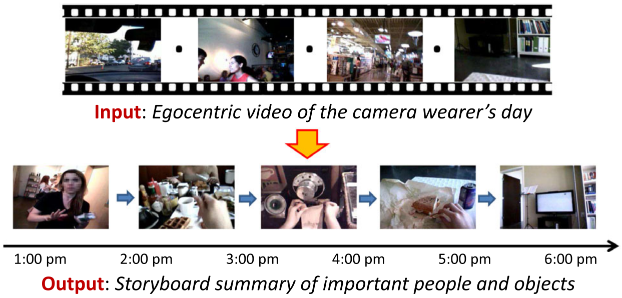

Input: Egocentric video of the camera wearer's day
1:00 pm
2:00 pm
3:00 pm
4:00 pm
5:00 pm
6:00 pm
Output: Storyboard summary of important people and objects

Fig. 1 Given an unannotated egocentric video, our method produces a compact storyboard visual summary that focuses on the key people and objects

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. Not only would recreational users (including "life-loggers") find it useful as a video diary, but there are also high-impact applications in law enforcement, elder and child care, and mental health. For example, the summaries could facilitate police officers in reviewing important evidence, suspects, and witnesses, or aid patients with memory problems to remember specific events, objects, and people (Hodges et al. 2011; Lee and Dey 2007). 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.

Our method works as follows. We first train a regression model from labeled training videos that scores any region's likelihood of belonging to an important person or object. For the input variables, we develop a set of high-level cues to capture egocentric importance, such as frequency, proximity to the camera wearer's hand, and object-like appearance and motion, as well as a set of low-level cues to capture region properties such as size, width, and height. The target variable is the overlap with ground-truth important regions, i.e., the *importance score*. Given a novel video, we use the model to predict important regions for each frame. We then partition the video into unique temporal *events*, by clustering scenes that have similar color distributions and are close in time. For each event, we isolate unique representative instances of each important person or object. Finally, we produce a storyboard visual summary that displays the most important objects and people across all events in the camera wearer's day.

We propose two ways to adjust the compactness of the summary, based on either a target importance criterion or a target summary length. For the latter, we design an energy function that accounts for the importance of the selected frames, their visual dissimilarities, and their temporal displacements, and can be efficiently optimized using dynamic programming.

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. For example, if the camera wearer has lunch with his friend Jill, she would be considered important, whereas people in the same restaurant sitting around them could be unimportant. Then, if they later attend a party but chat with different friends, Jill may no longer be considered important in that context.

Our main contribution is an egocentric video summarization approach that is driven by predicted important people and objects. Towards this goal, we develop two primary technical ideas. In the first, we develop a learning approach to estimate region importance using novel cues designed specifically for the egocentric video setting. In the second, we devise an efficient keyframe selection strategy that captures the most important objects and people, subject to meeting a budget for the desired length of the output storyboard.

We apply our method to challenging real-world videos captured by users in uncontrolled environments, and process a total of 27 h of video—significantly more data than previous work in egocentric analysis. Evaluating the predicted importance estimates and summaries, we find our approach outperforms state-of-the-art high-level and low-level saliency measures for this task, and produces significantly more informative summaries than traditional methods.

This article expands upon our previous conference paper (Lee et al. 2012). In Sects. 3.6.2 and 4.5, we introduce and analyze a novel budgeted frame selection approach that efficiently produces fixed-length summaries. In Sect. 4, we add new comparisons to multiple existing video summarization methods, conduct new user studies with over 25 users to systematically gauge the summaries' quality, and produce new results on the Activities of Daily Living dataset (Pirsiavash and Ramanan 2012).

2 Related Work

Video summarization Static keyframe methods compute motion stability from optical flow (Wolf 1996) or global scene color/texture differences (Zhang et al. 1997; Liu and Kender 2002; Weng and Merialdo 2009) to select the frames that go into the summary. The low-level approach means that irrelevant frames can often be selected, which is particularly problematic for our application of summarizing hours of continuous egocentric video that contain lots of irrelevant data. By generating object-driven summaries, we aim to move beyond such low-level cues.

Video summarization can also take the form of a single montage of still images. Existing methods take a background reference frame and project in foreground regions (Aner and Kender 2002), or sequentially display automatically selected key-poses (Caspi et al. 2006). An interactive approach (Goldman et al. 2006) takes user-selected frames and key points, and generates a storyboard that conveys the trajectory of an object. These approaches generally assume short clips with few objects, or a human-in-the-loop to guide the summarization process. In contrast, we aim to summarize a camera wearer's day containing hours of continuous video with hundreds of objects, with no human intervention.

Compact dynamic summaries simultaneously show several spatially non-overlapping actions from different times of the video (Rav-Acha et al. 2006; Pritch et al. 2007). While that framework aims to focus on foreground objects, it assumes a static camera and is therefore inapplicable to egocentric video. A re-targeting approach aims to simultaneously preserve an original video's content while reducing artifacts (Simakov et al. 2008), but unlike our approach, does not attempt to characterize the varying degrees of object importance. In a semi-automatic method (Liu et al. 2009), irrelevant video frames are removed by detecting the main object of interest given a few user-annotated training frames. In contrast, our approach *automatically* discovers multiple important objects.

Saliency detection Early saliency detectors rely on bottom-up image cues (e.g., Itti et al. 1998; Gao et al. 2007). More recent work tries to learn high-level saliency measures using various Gestalt cues, whether for static images (Liu et al. 2007; Alexe et al. 2010; Carreira and Sminchisescu 2010; Endres and Hoiem 2010) or video (Lee et al. 2011). Whereas typically such metrics aim to prime a visual search process, we are interested in high-level saliency for the sake of isolating those things worth summarizing. Researchers have also explored ranking object importance in static images, learning what people mention first from human-annotated tags (Spain and Perona 2008; Hwang and Grauman 2010). In contrast, we learn the importance of objects in terms of their role in a long-term video's story. Relative to any of the above, we introduce novel saliency features amenable to the egocentric video setting.

Egocentric visual data analysis Vision researchers have recently returned to exploring egocentric visual analysis, prompted in part by increasingly portable wearable cameras. Early work with wearable cameras partition visual and audio data into events (Clarkson and Pentland 1999), or uses supervised learning for specialized tasks like sign language recognition (Starner et al. 1998b) or location recognition within a building (Starner et al. 1998a). Methods in ubiquitous computing use manual intervention (Mann 1998) or external non-visual sensors (Healey and Picard 1998; Hodges et al. 2006) (e.g., skin conductivity or audio) to trigger snapshots from a wearable camera. Others use brain waves (Ng et al. 2002), k-means clustering with temporal constraints (Lin and Hauptmann 2006), or face detection (Doherty and Smeaton 2008) to segment egocentric videos. Recent methods explore activity recognition (Spriggs et al. 2009; Fathi et al. 2011; Pirsiavash and Ramanan 2012; Ryoo and Matthies 2013), handled object recognition (Ren and Gu 2010), novelty detection (Aghazadeh et al. 2011), hand detection (Li and Kitani 2013), gaze prediction (Li et al. 2013), social interaction analysis (Fathi et al. 2012), or activity discovery for non-visual sensory data (Huynh et al. 2008). Unsupervised algorithms are developed to discover scenes (Jojic et al. 2010) and actions (Kitani et al. 2011), or select keyframes (Doherty et al. 2008), based on low-level visual features extracted from egocentric data. In contrast to all these methods, we aim to build a visual summary, and model high-level importance of the objects present.

To our knowledge, we are the first to explore visual summarization of egocentric video by predicting important objects. Recent work (Lu and Grauman 2013) builds on our approach and uses our importance predictions as a cue to generate story-driven egocentric video summarizations.

3 Approach

Our goal is to create a storyboard summary of a person's day that is driven by the important people and objects. The video is captured using a wearable camera that continuously records what the user sees. We define *importance* in the scope of egocentric video: important things are those with which the camera wearer has significant interaction. This is reasonable for the egocentric setting, since the camera wearer is likely to engage in social activities with cliques of people (e.g., friends, co-workers) that involve interactions with specific objects (e.g., food, computer). The camera wearer will typically find these people and objects to be memorable, as we confirm in our user studies in Sect. 4.6.

There are four main steps to our approach: (1) using novel egocentric saliency cues to train a category-independent regression model that predicts how likely it is that an image region belongs to an important person or object; (2) partitioning the video into temporal events. For each event, (3) scoring each region's importance using the regressor; and (4) selecting representative key-frames for the storyboard that encapsulate the predicted important people and objects, either using a user-specified importance criterion or a length budget.

We first describe how we collect the video data and ground-truth annotations needed to train our model. We then describe each of the main steps in turn.

3.1 Egocentric Video Data Collection

We use the Looxcie wearable camera, which captures video at 15 fps at 320 × 480 resolution. It is worn around the ear and looks out at the world at roughly eye-level. We collected 10 videos from four subjects, each three to five h in length (the maximum battery life), for a total of 37 h of video. We call this the UT Egocentric (UT Ego) dataset. Our data is publicly available.[1]

Four subjects wore the camera for us: one undergraduate student, two grad students, and one office worker, ranging in age from early to late 20s and both genders. The different backgrounds of the subjects ensure diversity in the data—not everyone's day is the same—and is critical for validating the category-independence of our approach. We asked the subjects to record their natural daily activities, and explicitly instructed them not to stage anything for this purpose. The videos capture a variety of activities such as eating, shopping, attending a lecture, driving, cooking, and working on a computer.

3.2 Annotating Important Regions in Video

To train the importance predictor, we first need ground-truth training examples. In general, determining whether an object is important or not can be highly subjective. Fortunately, an egocentric video provides many constraints that are suggestive of an object's importance. For example, one can observe the camera wearer's hands, and an object of interest may often be centered in the frame.

In order to learn meaningful egocentric properties without overfitting to any particular category, we crowd-source annotations using Amazon's Mechanical Turk (MTurk). For egocentric videos, an object's degree of importance will depend on what the camera wearer is doing before, while, and after the object or person appears. In other words, 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 3 min accelerated video (equivalent to 10 min of original video) and to describe in text what they perceive to be essential people or objects necessary to create a summary of the video. 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. If none of the described objects are present in a frame, the annotator is given the option to skip it. See Fig. 2 for the two interfaces and example annotations.

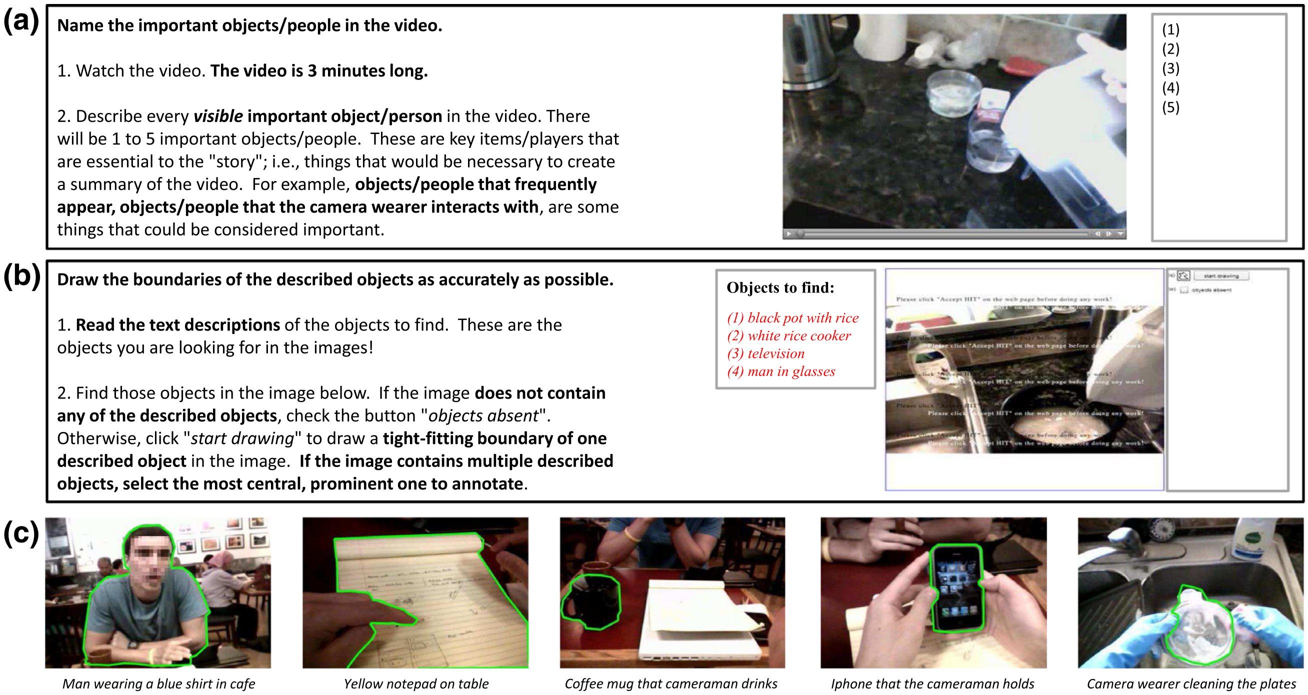
We found this two-step process more effective than a single task in which the same worker both watches the video and then annotates the regions s/he deems important, likely due to the time required to complete both tasks. Critically, the two-step process also helps us avoid bias: a single annotator asked 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. Our setup makes it easy for the first worker to freely describe the objects without bias, since s/he only has to enter text. We found the resulting annotations quite consistent, and only manually pruned those where the region outlined did not agree with the first worker's description. For a 3–5 h training video, we obtain roughly 35 text descriptions and 700 object segmentations.

3.3 Learning Egocentric Region Importance

We now discuss the procedure to train a general purpose category-independent model that will predict important regions in any egocentric video, independent of the camera wearer. Given a video, we first generate candidate regions for each frame using a min-cut method (Carreira and Sminchisescu 2010), which tends to avoid oversegmenting objects. We represent objects at the frame-level, since our uncontrolled setting usually prohibits reliable space-time object segmentation due to frequent and rapid head movements by the camera wearer. We generate roughly 800 regions per frame.

For each region, we compute a set of candidate features that could be useful to describe its importance. Since the video is captured by an active participant, we specifically want to exploit egocentric properties such as whether the object/person is interacting with the camera wearer, whether

[1] http://vision.cs.utexas.edu/projects/egocentric/ Due to privacy issues, we are only able to share 4 of the 10 videos (one from each subject), for a total of 17 h of video. They correspond to the test videos that we evaluate on in Sect. 4.



Name the important objects/people in the video.

1. Watch the video. **The video is 3 minutes long.**

2. Describe every ***visible* important object/person** in the video. There will be 1 to 5 important objects/people. These are key items/players that are essential to the "story"; i.e., things that would be necessary to create a summary of the video. For example, **objects/people that frequently appear, objects/people that the camera wearer interacts with**, are some things that could be considered important.

Draw the boundaries of the described objects as accurately as possible.

1. **Read the text descriptions** of the objects to find. These are the objects you are looking for in the images!

2. Find those objects in the image below. If the image **does not contain any of the described objects**, check the button "*objects absent*". Otherwise, click "*start drawing*" to draw a **tight-fitting boundary of one described object** in the image. **If the image contains multiple described objects, select the most central, prominent one to annotate.**

Objects to find:

(1) black pot with rice
(2) white rice cooker
(3) television
(4) man in glasses

Man wearing a blue shirt in cafe · *Yellow notepad on table* · *Coffee mug that cameraman drinks* · *Iphone that the cameraman holds* · *Camera wearer cleaning the plates*

Fig. 2 Our Mechanical Turk interfaces for important person/object **a** text description and **b** annotation, and **c** example annotations that we obtained. The important people and objects are annotated

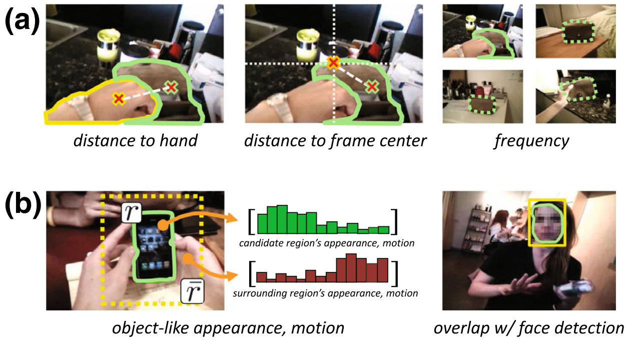

(a)
distance to hand
distance to frame center
frequency
(b)
candidate region's appearance, motion
surrounding region's appearance, motion
object-like appearance, motion
overlap w/ face detection

Fig. 3 Illustration of our **a** egocentric features and **b** object features

it is the focus of the wearer's gaze, and whether it frequently appears. In addition, we aim to capture high-level saliency cues—such as an object's motion and appearance, or the likelihood of being a human face—and generic region properties shared across categories, such as size or location. We describe the proposed features in detail next.

3.3.1 Feature Definitions

Egocentric features Figure 3a illustrates the three proposed egocentric features. To model *interaction*, we compute the Euclidean distance of the region's centroid to the closest detected hand in the frame. Given a frame in the test video, we first classify each pixel as (non-)skin using color likelihoods and a Naive Bayes classifier (Jones and Rehg 2002) trained with ground-truth hand annotations on disjoint data. We then classify any superpixel (computed using Felzenszwalb and Huttenlocher 2004) as hand if more than 25 % of its pixels are skin. While simple, we find this hand detector is sufficient for our application. More sophisticated methods (e.g., Kolsch and Turk 2004) would certainly be possible as well.

To model *gaze*, we compute the Euclidean distance of the region's centroid to the frame center. Since the camera moves with the wearer's head, this is a coarse estimate of how likely the region is being focused upon.

To model *frequency*, we record the number of times an object instance is detected within a short temporal segment of the video. We create two frequency features: one based on matching regions, the other based on matching points. For the first, we compute the color dissimilarity between a region r and each region r_n in its surrounding frames, and accumulate the total number of positive matches:

$$c_{region}(r) = \sum_{f \in \mathcal{W}} \left[\left(\min_n \chi^2(r, r_n^f) \right) \leq \theta_r \right], \qquad (1)$$

where f indexes the set of frames $\mathcal{W}$ surrounding region r's frame, $\chi^2(r, r_n)$ is the χ^2-distance between color histograms of r and r_n, θ_r is the distance threshold to determine a positive match, and $[\cdot]$ denotes the indicator function. The value of c_{region} will be high/low when r produces many/few matches (i.e., is frequent/infrequent).

The second frequency feature is computed by matching Difference of Gaussian SIFT interest points. For a detected point p in region r, we match it to all detected points in each frame $f \in \mathcal{W}$, and count as positive those that pass the ratio test (Lowe 2004). We repeat this process for each point in region r, and record their average number of positive matches:

$$c_{point}(r) = \frac{1}{P} \sum_{i=1}^{P} \sum_{f \in \mathcal{W}} \left[\frac{d(p_i, p_{1*}^f)}{d(p_i, p_{2*}^f)} \leq \theta_p \right], \quad (2)$$

where i indexes all detected points in region r, $d(p_i, p_{1*}^f)$ and $d(p_i, p_{2*}^f)$ measure the Euclidean distance between p_i and its best matching point p_{1*}^f and second best matching point p_{2*}^f in frame f, respectively, and θ_p is Lowe's ratio test threshold for non-ambiguous matches (Lowe 2004). The value of c_{point} will be high/low when the SIFT points in r produce many/few matches. For both frequency features, we set $\mathcal{W}$ to span a 10 min temporal window.

Object features In addition to the egocentric-specific features, we include three high-level (i.e., object-based) saliency cues (see Fig. 3b). To model *object-like appearance*, we use the learned region ranking function of Carreira and Sminchisescu (2010). It reflects Gestalt cues indicative of *any* object, such as the sum of affinities along the region's boundary, its perimeter, and texture difference with nearby pixels. (Note that the authors trained their measure on PASCAL data, which is disjoint from ours.) We stress that while this feature estimates how "object-like" a region is, it does not gauge importance. It is useful for identifying full object segments, as opposed to fragments.

To model *object-like motion*, we develop a key-segments video segmentation descriptor (Lee et al. 2011). It looks at the difference in motion patterns of a region relative to its closest surrounding regions. Specifically, we compare optical flow histograms for the region and the pixels around it within a loosely fit bounding box. Note that this feature is not simply looking for large motions or appearance changes from background. Rather, we are describing how the motion of the region differs from its closest surrounding regions; this allows us to forgo assumptions about camera motion, and also to be sensitive to different magnitudes of motion. Similar to the appearance feature above, 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 $\frac{|q \cap r|}{|q \cup r|}$ between the region r and any detected frontal face q in the frame, using Viola and Jones (2001).

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 expect that important people and objects will occur at non-random scales and locations in the frame, due to social and environmental factors that constrain their relative positioning to the camera wearer (e.g., sitting across a table from someone when having lunch, or handling cooking utensils at arm's length). Our region features capture these statistics.

Altogether, these cues form a 14-dimensional feature space to describe each candidate region (4 egocentric, 3 object, and 7 region feature dimensions).

3.3.2 Regressor to Predict Region Importance

Using the features defined above, we next train a model that can predict a region's importance. The model should be able to learn and predict a region's *degree* of importance instead of whether it is simply "important" or "not important", so that we can meaningfully adjust the compactness of the final summary (as we demonstrate in Sect. 4). Thus, we opt to train a regressor rather than a classifier.

While the features defined above can be individually meaningful, we also 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*:

$$I(r) = \beta_0 + \sum_{i=1}^{N} \beta_i x_i(r) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \beta_{i,j} x_i(r) x_j(r), \quad (3)$$

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 obtained from Sect. 3.2. Thus, regions with perfect overlap with ground-truth will have a target importance score of 1, those with no overlap with ground-truth will have an importance score of 0, and all others will have an importance score in (0, 1). We standardize the features to zero-mean and unit-variance, and solve for the β's using least-squares. For testing, our model takes as input a region r's features (the x_i's) and predicts its importance score $I(r)$. Note that we train and test using video from different users to avoid overfitting our model to any specific camera wearer.

3.4 Segmenting the Video into Temporal Events

Given a new video, we first partition the video temporally into events, and then isolate the important people and objects in each event. Events allow the final summary to include multiple instances of an object/person that is central in multiple contexts in the video. For example, suppose that the camera

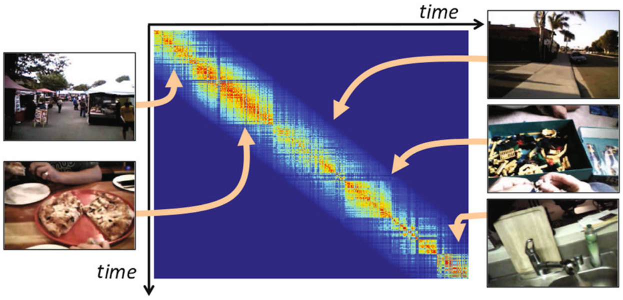


Fig. 4 Distance matrix that measures global color dissimilarity between all frames. (*Blue*/*red* reflects high/low distance.) The images show representative frames of each discovered event. The block structure along the diagonal reveals groups of frames that are close in appearance and time

wearer plays with her dog at home in the morning and later takes the dog out to the park at night. We can treat the two instances of the dog as different objects (since they appear in different events) and include both in the final summary. Moreover, events indicate which selected frames are more related to one another, giving a hierarchical structure to the final summary.

While shot boundary detection has been frequently used to perform event segmentation for videos, it is impractical for our wearable camera data setting. Traditional shot detection generally assumes visual continuity and thus tends to over-segment egocentric events due to frequent head movements. Instead, we detect egocentric events by clustering scenes in such a way that frames with similar global appearance can be grouped together even when there a few unrelated frames ("gaps") between them.

Let $\mathcal{V}$ denote the set of all video frames. We compute a pairwise distance matrix $D_\mathcal{V}$ between all frames $f_m, f_n \in \mathcal{V}$, using the distance:

$$D(f_m, f_n) = 1 - w^t_{m,n} \exp\Big(-\frac{1}{\Omega}\chi^2(f_m, f_n)\Big), \tag{4}$$

where $w^t_{m,n} = \frac{1}{t}\max(0, t - |m - n|)$, t is the size of the temporal window surrounding frame f_m, $\chi^2(f_m, f_n)$ is the χ^2-distance between color histograms of f_m and f_n, and Ω denotes the mean of the χ^2-distances among all frames. Thus, frames similar in color receive a low distance, subject to a weight that discourages frames too distant in time from being grouped.

We next perform complete-link agglomerative clustering with $D_\mathcal{V}$, grouping frames until the smallest maximum inter-frame distance is larger than two standard deviations beyond Ω. The first and last frames in a cluster determine the start and end frames of an event, respectively. Figure 4 shows the distance matrix computed for one subject's day, and the representative frames for each discovered event.

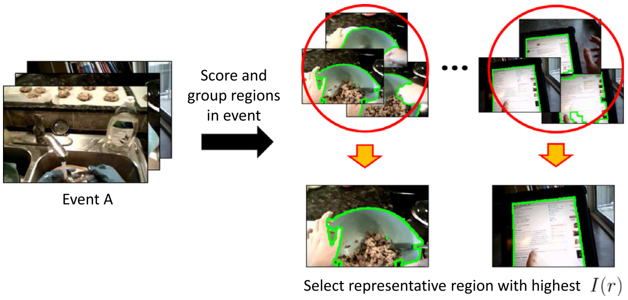

Score and group regions in event
Event A
Select representative region with highest $I(r)$

Fig. 5 Discovering an event's key people and objects. For each event, we group together regions that are likely to belong to the same object, and then for each group, we select the region with the highest importance score as its representative

3.5 Discovering an Event's Key People/Objects

For each event, we aim to select the important people and objects that will go into the final summary, while avoiding redundancy. Recall that objects are represented at the frame-level (Sect. 3.3). Thus, our goal is to group together instances of the same person or object that appear over time in each event.

Given an event, we first score each bottom-up segment in each frame using our regressor. Since we do not know a priori how many important things an event contains, we generate a candidate pool of clusters from the set $\mathcal{C}$ of bottom-up regions, and then remove any redundant clusters, as follows.

To extract the candidate groups, we first compute an affinity matrix $K_\mathcal{C}$ over all pairs of regions $r_m, r_n \in \mathcal{C}$, where affinity is determined by color similarity: $K_\mathcal{C}(r_m, r_n) = \exp(-\frac{1}{\Gamma}\chi^2(r_m, r_n))$, where Γ denotes the mean χ^2-distance among all pairs in $\mathcal{C}$. We next partition $K_\mathcal{C}$ into multiple (possibly overlapping) inlier/outlier clusters using a factorization approach (Perona and Freeman 1998). The method finds tight sub-graphs within the input affinity graph while resisting the influence of outliers. Each resulting sub-graph consists of a candidate important object's instances. To reduce redundancy, we sort the sub-graph clusters by the average $I(r)$ of their member regions, and remove those with high affinity to a higher-ranked cluster. Finally, for each remaining cluster, we select the region with the highest importance score as its representative (see Fig. 5).

3.6 Generating a Storyboard Summary

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. 11). Each event can display a varying number of frames, depending on how many unique important things our method discovers.

We propose two ways to adjust the *compactness* of the summary: (1) according to a target importance criterion, and

(2) according to a target summary length. We describe each process in detail next.

3.6.1 Summarization Given an Importance Criterion

We first describe how to summarize the video given an importance criterion. This allows the system to automatically produce the most compact summary possible that encapsulates only the people and objects that meet the importance threshold.

When discovering an event's key people and objects (Sect. 3.5), we take only those regions that have importance scores higher than the specified criterion to form set $\mathcal{C}$. We then proceed to group instances of the same person or object together in $\mathcal{C}$, and select the frame with the highest scoring region in each group to go into the summary.

3.6.2 Summarization Given a Length Budget

Alternatively, we can summarize the video given a length budget k. This allows the system to answer requests such as, "Generate a 5-min summary." We formulate the objective as a k-frame selection problem and define the following energy function:

$$E(S) = -\sum_{i=1}^{|S|} I(f_{s_i}) + \sum_{i=1}^{|S|-1} \exp\Big(-\frac{1}{\Omega}\chi^2(f_{s_i}, f_{s_{i+1}})\Big) - \sum_{i=1}^{|S|-1} |s_i - s_{i+1}|^{\frac{1}{2}}, \tag{5}$$

where $S = \{s_1, \ldots, s_k\}$ is the set of indices of the k selected frames, and Ω is the mean of the χ^2-distances among all frames.

There are three terms in our energy function. The first term enforces selection of important frames, since we want the summary to contain the discovered important people and objects. We score each frame using the region that has the highest importance score: $I(f_{s_i}) = \max_m I(r_{m,i})$, where $r_{m,i}$ is the mth region in frame i. Our second term enforces visual uniqueness, i.e., that adjacent selected frames contain different objects. We want the summary to avoid including redundant frames. Thus, we compute an affinity based on the χ^2-distance between color histograms of adjacent frames f_{s_i} and $f_{s_{i+1}}$. Finally, our last term enforces selection of frames that are spread out in time such that the summary best captures the entire "story" of the original video. For this, we compute the difference in frame index of the selected frames. Note that $\sum_{i=1}^{|S|-1} |s_i - s_{i+1}|^{\frac{1}{2}}$ achieves a maximum when the temporal distances between all adjacent frames $|s_i - s_{i+1}|$ are equal.

We compute the optimal set S^* of k frames by finding the set that minimizes Eq. 5:

$$S^* = \operatorname*{argmin}_{S \subset V} E(S), \tag{6}$$

where V is the set of frames of the selected important people and objects from Sect. 3.5.

A naive approach for optimizing Eq. 6 would take time $O(\binom{F}{k})$ for $F = |V|$ total frames. Instead, we efficiently find the optimal set S^* using dynamic programming, by exploiting the optimal substructure that exists in the k-frame selection problem.

Specifically, the minimum energy $M(f_n, t)$ of a t-length summary that selects frame f_n at time step t can be recursively computed as follows:

$$M(f_n, t)_{t \le n \le F-k+t} = \begin{cases} -I(f_n), & \text{if } t = 1. \\ -I(f_n) + \min\limits_{p \le m \le q} (e(f_m, f_n) & \\ \quad + M(f_m, t-1)), & \text{if } 1 < t \le k, \end{cases} \tag{7}$$

where $p = t - 1$, $q = F - k + t + 1$, and $e(f_m, f_n) = \exp(-\frac{1}{\Omega}\chi^2(f_m, f_n)) - |m - n|^{\frac{1}{2}}$. We enforce the selected set of frames to be a temporally ordered subsequence of the original video: $s_i < s_{i+1}, \forall i$. Thus, any "path" that does not obey this rule is assigned infinite cost.

Using Eq. 7, we can compute the minimum energy for a k-length summary as $E(S^*) = \min_n M(f_n, k)$, which can be solved in $O(F^2 k)$ time. We retrieve the optimal set of k frames S^* by backtracking from f_n at time k.

3.6.3 Discussion

The two strategies presented above offer certain trade-offs. The importance criterion automatically produces the most compact summary possible that includes all unique instances of the important people and objects; however, it does not give the user direct control on the length of the output summary. In contrast, while the proposed budgeted formulation can return a storyboard of a specified length, it does not permit setting an absolute threshold on how important objects must be for inclusion.

In addition to being a compact video diary of one's day, our storyboard summary can be considered as a *visual index* to help a user peruse specific parts of the video. This would be useful when one wants to relive a specific moment or search for less important people or objects that occurred with those found by our method.

Algorithm 1 recaps all the steps of our approach.

4 Results

In this section we evaluate our approach on our new UT Egocentric (UT Ego) dataset and on the Activities of Daily Living

```
Input: Egocentric video, and importance selection criterion or
       length budget $k$.
Output: Storyboard summary.
1. Train regression model. (Sec. 3.3)
2. Segment video into temporal events. (Sec. 3.4)
For each event,
3. Compute $I(r)$ for all regions. (Sec. 3.3)
4. Group regions that belong to same person/object. (Sec. 3.5)
5. Retain unique clusters, select most important region in each
   group. (Sec. 3.5)
6. Generate storyboard summary that shows selected important
   people/objects. (Sec. 3.6)
```

Algorithm 1: Our summarization approach

(ADL) dataset (Pirsiavash and Ramanan 2012), which consists of 17 and 10 h of egocentric video, respectively. We offer direct comparisons to existing methods for both saliency and video summarization, and we perform a user study with over 25 subjects to quantify the perceived quality of our results. We use UT Ego for the experiments in Sects. 4.2–4.6, and ADL for the experiments in Sect. 4.7.

4.1 Dataset and Implementation Details

For our UT Ego dataset, we collected 10 videos from four subjects, each 3–5 h long.[2] Each person contributed one video, except one who contributed seven. The videos are challenging due to frequent camera viewpoint/illumination changes and motion blur. For evaluation, we use four data splits: for each split we train with data from three users and test on one video from the remaining user. Hence, the camera wearers in any given training set are disjoint from those in the test set, ensuring we do not learn user- or object-specific cues.

ADL contains 20 videos from chest-mounted cameras, each on average about 30 min long. The camera wearers perform daily activities in the house, like brushing hair, cooking, washing dishes, or watching TV. To generate candidate object regions on ADL, we use BING (Cheng et al. 2014), which generates bounding box proposals and is orders of magnitude faster than the min-cut approach of Carreira and Sminchisescu (2010).

We use Lab space color histograms, with 23 bins per channel, and optical flow histograms with 61 bins per direction using Brox and Malik (2011). We set $t = 27,000$ and $t = 2,250$ (i.e., a 60 and 5 min temporal window), for UT Ego and ADL, respectively. We set $\theta_r = 10,000$ and $\theta_p = 0.7$ after visually examining a few examples. We fix all parameters for all results. For efficiency, we process every 15th frame (i.e., 1 fps). For Eq. 5, we standardize each term to zero-mean and unit-variance using training data.

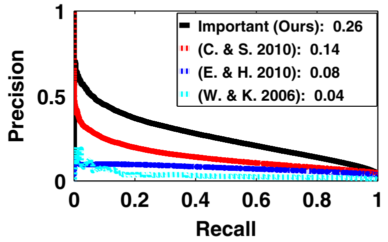


Fig. 6 Precision-Recall for important object prediction. Numbers in the legends denote average precision. By leveraging egocentric-specific cues, our approach more accurately discovers the important regions

4.2 Important Region Prediction Accuracy

We first evaluate our method's ability to predict important regions, compared to three state-of-the-art methods: (1) the object-like score of Carreira and Sminchisescu (2010), (2) the object-like score of Endres and Hoiem (2010), and (3) a bottom-up saliency detector (Walther and Koch 2006). The first two are high-level learned functions that predict a region's likelihood of overlapping a true object, whereas the third is a low-level detector to find regions that "stand-out". They are all general-purpose metrics (not tailored to egocentric data), so they allow us to gauge the impact of our proposed egocentric cues for finding important objects in video.

We use the annotations obtained on MTurk as ground truth (GT) (see Sect. 3.2). Some frames contain more than one important region, and some contain none, depending on what the annotators deemed important. On average, each video contains 680 annotated frames and 280,000 test regions. A region r is considered to be a true positive (i.e., important object), if its overlap score with any GT region is greater than 0.5, following PASCAL convention.

Figure 6 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 (Carreira and Sminchisescu 2010; Endres and Hoiem 2010) can successfully find prominent object-like regions, and so they noticeably outperform the low-level saliency detector. However, by focusing on detecting *any* object, unlike our approach they are unable to distinguish those that may be important to a camera wearer.

Figure 7 shows example important regions detected by each method. The first four columns show examples of correct predictions made by our method. We see that low-level saliency detection (Walther and Koch 2006) is insufficient; its local estimates fail to find object-like regions. For example, it finds a bright blob surrounded by a dark region to be the most salient (first row, fourth column).

The last four columns show examples of incorrect predictions made by our method. The high-level saliency detection methods (Carreira and Sminchisescu 2010; Endres and Hoiem 2010) produce better predictions for these exam-

[2] See Footnote 1

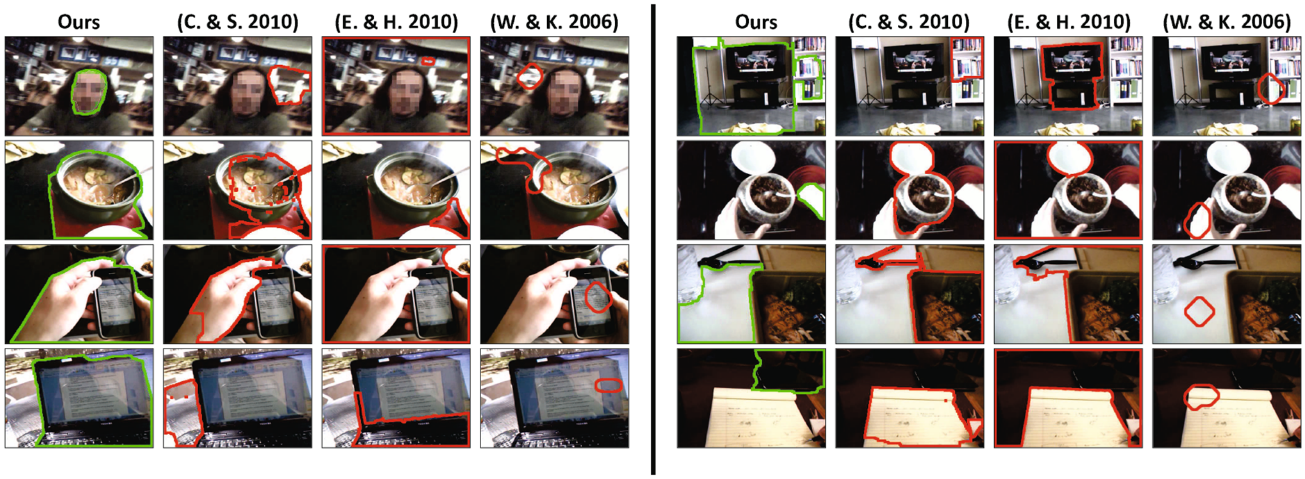


Fig. 7 Example selected regions/frames. The *first four columns* show examples of correct predictions made by our approach, and the *last four columns* show failure cases in which the high-level saliency methods (Carreira and Sminchisescu 2010; Endres and Hoiem 2010) make better predictions

Fig. 8 Top 28 features with highest learned weights

1. size
2. size + height
3. y + face
4. size + pt freq.
5. bbox y + face
6. width
7. size + gaze
8. height
9. pt freq.
10. size + reg freq.
11. gaze
12. face
13. y
14. size + width
15. obj app.
16. x
17. size + x
18. gaze + x
19. obj app. + y
20. x + bbox x
21. y + bbox x
22. bbox x + reg freq.
23. x + reg freq.
24. obj app. + size
25. y + interaction
26. width + height
27. gaze + bbox x
28. bbox y + interaction

ples. In the first example, our method produces an under-segmentation of the important object and includes regions surrounding the television due to the combined region having higher object-like appearance score than the television alone. In the second example, our method incorrectly detects the user's hand to be important, while in the third and fourth examples, it determines background regions to be important due to their high frequency.

We next perform ablation studies to investigate the contribution of the pairwise interaction terms of our importance predictor. Specifically, we compare to a linear regression model and an L1-regularized linear regression model using only the original 14-dimensional features. The average precision of the linear regression model is 0.20, and the average precision of the L1-regularized model ranges from 0.14 to 0.20 depending on the level of sparsity, as enforced by the weight on the regularization term. This result shows that the original features alone are not sufficiently expressive, and that the pairwise terms are necessary to more fully capture the relationship between the features and desired importance values.

4.3 Which Cues Matter Most for Importance?

Figure 8 shows the top 28 out of 105 $\left(= 14 + \binom{14}{2}\right)$ features that receive the highest learned weights (i.e., β magnitudes). Region size is the highest weighted cue, which is reasonable since an important person/object is likely to appear roughly at a fixed distance from the camera wearer. Among the egocentric features, gaze and frequency have the highest weights. Frontal face overlap is also highly weighted; intuitively, an important person would likely be facing and conversing with the camera wearer.

Some highly weighted pair-wise interaction terms are also quite interesting. The feature measuring a region's face overlap *and* y-position has more impact on importance than face overlap alone. This suggests that an important person usually appears at a fixed height relative to the camera wearer. Similarly, the feature for object-like appearance *and* y-position has high weight, suggesting that a camera wearer often adjusts his ego-frame of reference to view an important object at a particular height.

Surprisingly, the pairing of the interaction (distance to hand) and frequency cues receives the lowest weight. A plausible explanation is that the *frequency* of a handled object highly depends on the camera wearer's activity. For example, when eating, the camera wearer's hand will be visible and the food will appear frequently. On the other hand, when grocery shopping, the important item s/he grabs from the shelf will (likely) be seen for only a short time. These conflicting signals would lead to this pair-wise term having low weight. Another paired term with low weight is an "object-

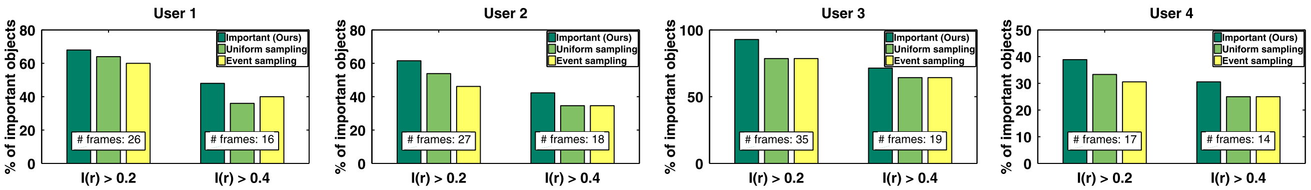


Fig. 9 Comparison to alternative summarization strategies, in terms of important object recall rate. Using the same number of frames, our approach includes more important people and objects

like" region that is frequent; this is likely due to unimportant background objects (e.g., the lamp behind the camera wearer's companion). This suggests that higher-order terms could yield even more informative features.

4.4 Importance-Based Summarization Accuracy

Next we evaluate our method's summarization results using the importance-based criterion, and in the following section we evaluate its budget-based results.

4.4.1 Quantitative Evaluation

The central premise of our work is that day-to-day activity viewed from the first person perspective largely revolves around the important people and objects with which the camera wearer interacts. Accordingly, a good visual summary must capture those important entities. Thus, we analyze the recall rate for our method and two competing summarization strategies. The first is uniform keyframe sampling, and the second is event-based adaptive keyframe sampling. The latter computes events using the same procedure as our method (Sect. 3.4), and then divides its keyframes evenly across events. Both methods are modeled after standard keyframe and event detection methods (Money and Agius Feb 2008; Wolf 1996; Zhang et al. 1997).

Figure 9 shows the results. Each set of bars shows the recall rates for the three methods. Our method varies its selection criterion on $I(r)$ over $\{0.2, 0.4\}$, for two summaries in total for each user. These thresholds are used to cover a broad spectrum (i.e., low and high selection criteria) and are arbitrary; we see consistent relative results for any threshold. To compare our recall rates to those of the baselines, we create summaries for the baselines with the same number of frames as ours.

If a frame contains multiple important objects, we score only the main one. Likewise, if a summary contains multiple instances of the same GT object, it gets credit only once. Note that this measure is favorable to the baselines, since it does not consider object *prominence* in the frame. For example, we give credit for the TV in the last frame in Fig. 10, bottom row, even though it is only partially captured. Furthermore, by definition, the uniform and event-based baselines are likely to get many hits for the most frequent objects. These make the baselines very strong and meaningful comparisons.

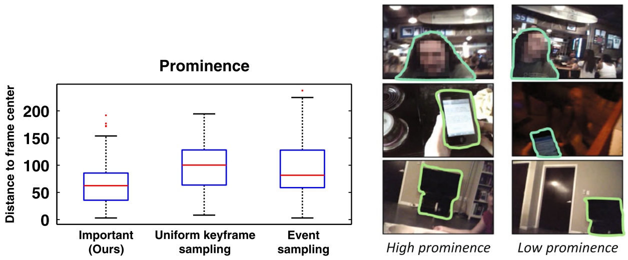


Fig. 10 Comparison to alternative summarization strategies, in terms of the prominence of the objects within selected keyframes. Our summaries more prominently display the important objects

Overall, our summaries include more important people/objects with the same number of frames. For example, for User 2 with selection criterion on $I(r) > 0.2$, our method finds 62 % of important objects in 27 frames, whereas the uniform keyframe and event-based adaptive keyframe sampling methods find 54 and 46 % of important objects, respectively. The lower absolute recall rate for all methods for User 4 is due to many small GT objects that appear together in the same frame (the user was cooking and baking). On average, we find 9.13 events/video and 2.05 people/objects per event.

While Fig. 9 captures the recall rate of the important objects, it does not measure the *prominence* of the objects in the selected frames. An informative summary should include not just any instance of the important object, but frames in which it is displayed prominently (i.e., large and centered). To this end, in Fig. 10, we quantify the prominence of important objects in each method's summaries, in terms of the distance of the region's centroid to the frame center. We see our method better isolates the prominent instances, thanks to its egocentric cues. For example, in the top right example, the TV has high prominence in our summary and low prominence in the uniform keyframe sampling's summary.

4.4.2 Summarization Examples

Figure 11 shows example summaries from our method and the keyframe sampling baseline. The colored blocks on ours indicate the discovered events. We see that our summary not only has better recall of important objects, but it also selects views in which they are prominent in the frame. This helps

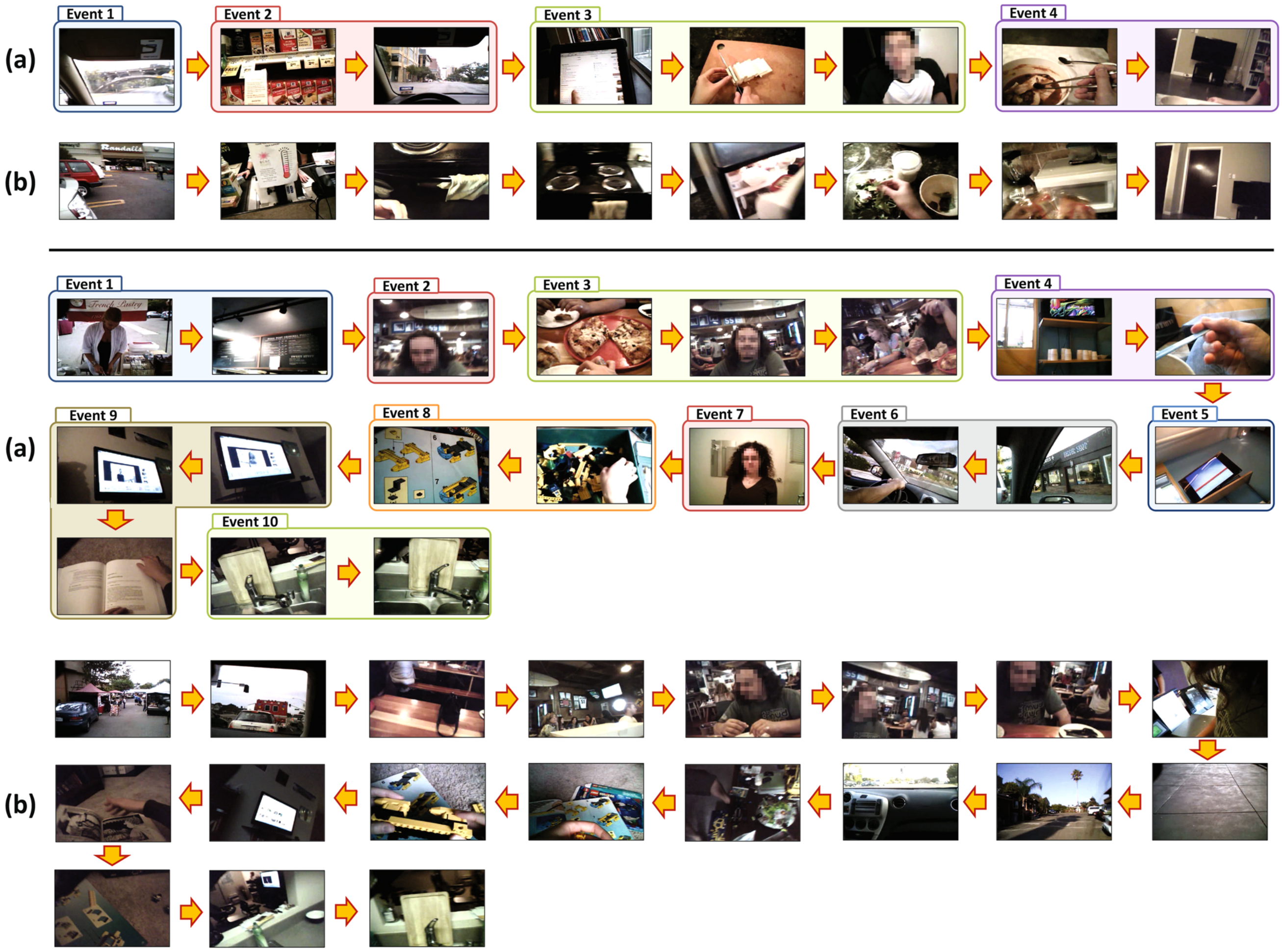


Fig. 11 **a** Our summary versus **b** uniform keyframe sampling for two user videos. The *colored blocks* for ours indicate the discovered events. Our summary focuses on the important people and objects. While uniform keyframe sampling does hint at the course of events, it tends to include irrelevant or redundant frames (e.g., repeated instances of the man in the bottom example) because it lacks a notion of object importance

more clearly reveal the story of the video. For instance, for the top example, the story is: *selecting an item at the supermarket* → *driving home* → *cooking* → *eating and watching TV*. We provide additional summaries at the project webpage.

Figure 11 (bottom) also depicts our method's failure modes. Redundant frames of the same object can appear due to errors in event segmentation (see the man captured in Events 2 and 3) or the candidate important object clustering (the sink is captured twice in Event 10). Adding features like GPS or depth might reduce such errors.

Figure 12 shows another example where we track the camera wearer's location with a GPS receiver, and display our method's keyframes on a map with the tracks (purple trajectory) and timeline. This result suggests a novel multi-media application of our visual summarization algorithm that incorporates location, temporal, and visual data.

In all the results in this section, the two baselines perform fairly similarly to one another; compared to our method, they are prone to choosing unimportant or redundant frames that lack focus on those objects a human viewer has deemed important. This supports our main hypothesis that the traditional low-level cues used in generic video summarization

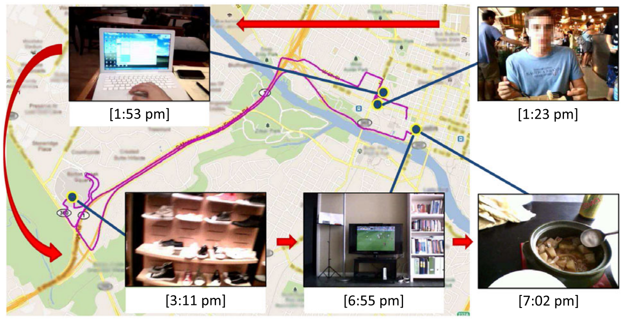


Fig. 12 An application of our approach that shows the GPS tracks of the camera wearer, the important people and objects that s/he interacted with, and their timeline

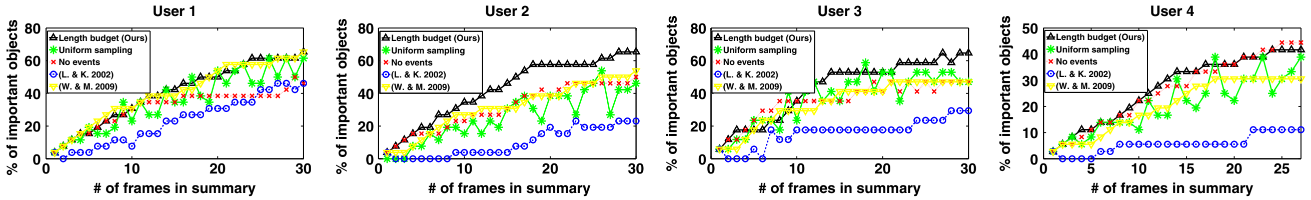


Fig. 13 Comparison to alternative k-frame summarization strategies. Our budgeted frame selection approach produces more informative summaries with fewer frames

methods are insufficient to select keyframes that capture key objects in egocentric video. Building on this finding, the user studies below analyze the impact that including important objects has on perceived summary quality.

4.5 Budgeted Frame Selection Accuracy

We next evaluate our approach for the scenario where we must handle requests such as, "I would like to see a 10-frame summary of the original video".

We compare our budgeted k-frame selection approach to four alternative methods: (1) the state-of-the-art video summarization method of Weng and Merialdo (2009), which selects keyframes that provide maximal content inclusion. Briefly, it iteratively selects the frame that is on average most similar to all remaining frames without being redundant to the frames that have already been chosen.[3] (2) The keyframe selection approach of Liu and Kender (2002), which optimizes an energy function that enforces adjacent frames to be maximally different. For fairest comparison, we use the same χ^2-distance on color histograms used by our method to gauge visual dissimilarity. (3) a side-by-side implementation of our approach without event segmentation and region grouping (i.e., it selects k-frames from all frames of the video), and (4) uniform keyframe sampling. The former two contrast our method with existing techniques that target the generic video summarization problem, highlighting the need to specialize to egocentric data as we propose. The latter two isolate the impact of our importance predictions as well as our event segmentation and region grouping.

4.5.1 Quantitative Evaluation

The plots in Fig. 13 show the results. 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. Each point on the curve shows the result for a different summary of the required length. We score the objects found in the same way as in Sect. 4.4.1.

Our model significantly outperforms the keyframe method (Liu and Kender 2002), which confirms that modeling the importance of the object or person is critical to produce informative summaries for egocentric videos. In fact, the existing method performs even worse than uniform sampling, due to its preference for frames that are maximally dissimilar to their surrounding selected frames. As a result, it tends to select redundant frames containing the same visual elements in an alternating fashion. Our summary does not have this issue since we represent each object in each event with a single region/frame through region clustering.

Our model also outperforms the multi-document method (Weng and Merialdo 2009) on all but one user. While this prior method successfully selects diverse content throughout the video, its reliance on low-level image cues leads to choosing some non-essential frames.

With very short summaries, uniform sampling performs similarly to ours; the selected keyframes are more spread out in time and have a high chance of including unique people/objects. However, with longer summaries, our method always outperforms uniform sampling, since uniform sampling ignores object importance and tends to include frames repeating the same important object.

Our model also outperforms the baseline that selects k-frames from the entire video without event segmentation and region grouping ("No events"). Since this method does not group instances of the same object together, it can select the same important object multiple times.

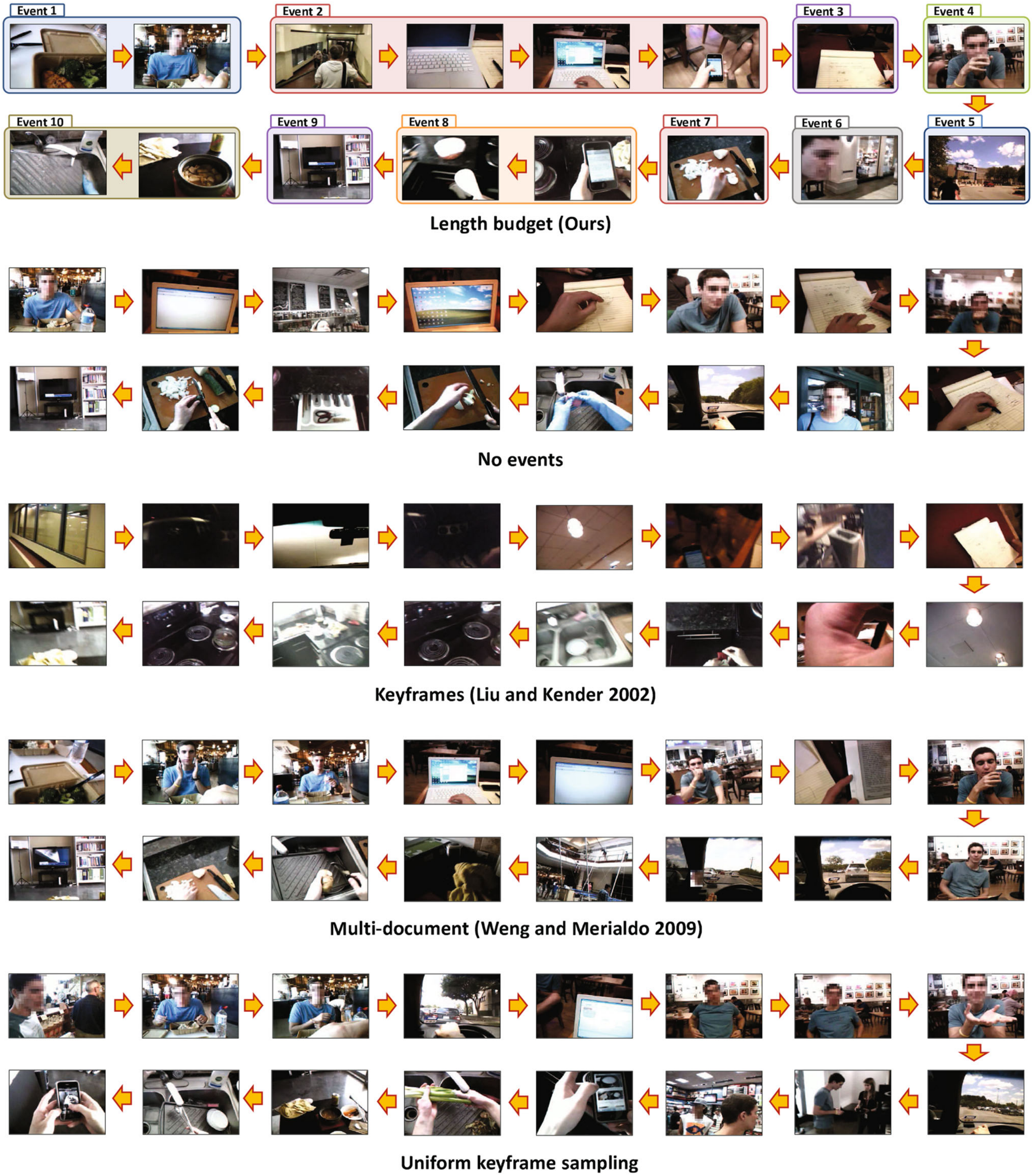
4.5.2 Summarization Examples

Figure 14 shows example summaries created by each method. By focusing on the important people and objects, our method produces the best results.

4.6 User Studies to Evaluate Summaries

We next perform user studies, since ultimately the impact of a summary depends on its value to a human viewer. As subjects, we recruit both the camera wearers as well as 25 subjects uninvolved with the data collection or research in any way. The camera wearers are a valuable resource to discern summary quality, since they alone fully experienced

[3] This method summarizes a collection of videos, so we treat each event in our data as a different video.



Length budget (Ours)

No events

Keyframes (Liu and Kender 2002)

Multi-document (Weng and Merialdo 2009)

Uniform keyframe sampling

Fig. 14 Example summaries per method on UT Ego. The "No events" baseline can include redundant frames because it lacks event segmentation and region clustering to group instances of the same object together (e.g., the yellow notepad and man). Keyframe selection (Liu and Kender 2002) focuses on selecting adjacent frames that are maximally dissimilar, leading it to toggle between highly diverse frames, which need not capture important objects. While the multi-document summarization objective (Weng and Merialdo 2009) overcomes this toggling effect, both it and uniform keyframe sampling tend to select redundant frames (e.g., see repeated instances of the man). Overall, our summary best focuses on the important people and objects. It selects informative frames that convey the chain of events through the objects and people that drive the first person interactions

Table 1 Camera wearer user study results comparing our summaries to uniform keyframe sampling

	Much better (%)	Better (%)	Similar (%)	Worse (%)	Much worse (%)
Imp. captured	31.25	37.5	18.75	12.5	0
Overall quality	25	43.75	18.75	12.5	0

the original content. Complementary to that, the uninvolved subjects are valuable to objectively gauge whether the overall events are understandable—without the implicit benefit of being able to "fill in the gaps" with their own firsthand experience of the events being summarized.

4.6.1 Evaluation by the Camera Wearers

To quantify *perceived* quality, we ask the camera wearers to compare our method's summaries to those generated by uniform keyframe sampling. The camera wearers are good judges, since they know the full extent of their day that we are attempting to summarize.

We generate four pairs of summaries for each user, each of different length. We ask the subjects to view our summary and the baseline's (in some random order unknown to the subject, and different for each pair), and answer two questions: (1) *Which summary captures the important people/objects of your day better?* and (2) *Which provides a better overall summary?* The first specifically isolates how well each method finds important, prominent objects, and the second addresses the overall quality and story of the summary.

Table 1 shows the results, in terms of how often our summary is preferred. In short, out of 16 total comparisons, our summaries were found to be better 68.75 % of the time. We find our approach can fail to produce better summaries than uniform keyframe sampling if the user's day is very simple. Specifically, User 3 was working on her laptop the entire day; first at home, then at class, then during lunch, and finally at the library. For this video, uniform keyframe sampling was sufficient to produce a good summary.

4.6.2 Evaluation by Independent Subjects

Next, to measure the quality of our summary on an absolute scale and to allow independent judges to evaluate a visual summary's informativeness, we ask each camera wearer to provide a "ground-truth" text summary of his/her day. Specifically, we ask the users to provide full sentence descriptions that emphasize the key happenings (i.e., who s/he met, what s/he did, where, and when), and in sequential order as they happened that day. The resulting text summaries are 6-10 sentences long. Here is an example from User 2:

> My boyfriend and I drove to a farmers market in the early afternoon, where we sampled some food. Then (also in the early afternoon) we drove to a pizza place, where we stayed for a while, talked, had pizza, drank beer, and watched TV. After that, in the afternoon, we walked to a frozen yogurt place and split a cup of frozen yogurt, with brief looks at an animation that was playing. Then we walked around for a while, and drove home in the early evening. At home, we played with Legos for a while, in the living room. Then we watched some videos on YouTube. After that we played with Legos some more, and I washed some dishes in the kitchen, in the evening.

See the supplementary file for the remaining text summaries.

We then ask 25 subjects using Mechanical Turk to compare our summary and the baselines' (without knowing which method generated the summary) to the text summary provided by the camera wearer of the corresponding video, and answer: *How well does the visual summary follow the story of the text summary?* On a scale of 1 to 5 (1 being "very well" and 5 being "very poorly"), over all 16 summaries, ours scored 2.61 (±0.97). The prior methods (Liu and Kender 2002; Weng and Merialdo 2009), and uniform sampling scored only 3.43 (±1.05), 3.28 (±1.10), 2.94 (±1.09), respectively. In general, the judges found the longer summaries to better align with the corresponding text summary than the shorter summaries. On some videos, our shorter summaries failed to capture all of the details in the text summary, resulting in poor scores.

While the result above gauges quality on an absolute scale, we also ran a comparative test. Here, we ask the subjects to compare our summary and each baseline's (in random order) to the text summary, and answer: *Which visual summary more closely follows the story of the text summary?* Table 2 shows the accumulated responses from all 25 subjects. Out of 16 total comparisons to each baseline, our summaries were found to be better 48–62 % of the time, and only worse 24–38 % of the time.

4.7 Experiments on ADL

Finally, we perform experiments on ADL, an interesting and complimentary dataset to UT Ego that contains egocentric videos of people performing daily activities in their home (e.g., washing dishes, brushing teeth, etc.). It contains 20 videos, each roughly 30 min in length.

Since this data lacks ground-truth important object annotations, we use it only to evaluate our summaries. We take the importance predictor from UT Ego (trained on all four

Table 2 Mechanical Turk user study results on UT Ego comparing our summaries to baseline summaries

	Much better (%)	Better (%)	Similar (%)	Worse (%)	Much worse (%)
Keyframes (Liu and Kender 2002)	16.43	45.45	13.99	18.88	5.25
Multi-document (Weng and Merialdo 2009)	21.08	36.14	17.47	17.47	7.84
uniform sampling	10.22	37.63	14.52	29.03	8.60

Table 3 Mechanical Turk user study results on ADL comparing our summaries to baseline summaries

	Much better (%)	Better (%)	Similar (%)	Worse (%)	Much worse (%)
Keyframes (Liu and Kender 2002)	26.80	41.24	17.01	11.34	3.61
Multi-document (Weng and Merialdo 2009)	13.90	28.88	25.67	26.20	5.35
uniform sampling	11.95	33.96	20.75	25.79	7.55

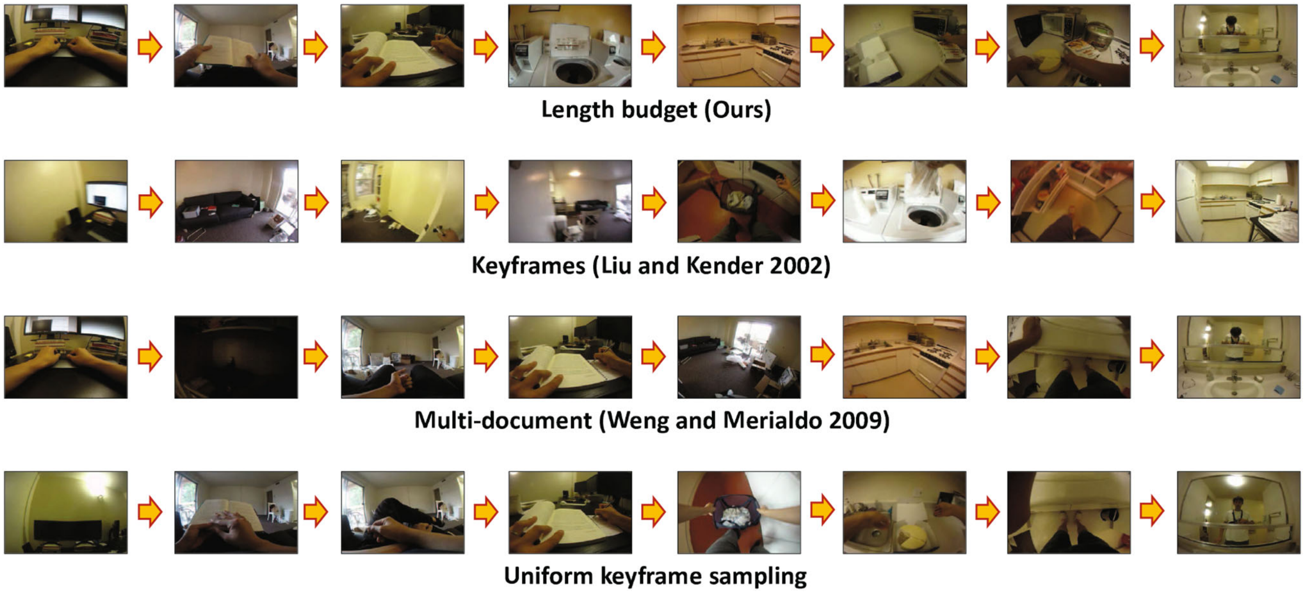

Length budget (Ours)
Keyframes (Liu and Kender 2002)
Multi-document (Weng and Merialdo 2009)
Uniform keyframe sampling

Fig. 15 Example summaries per method on ADL. Keyframe selection (Liu and Kender 2002) focuses on selecting adjacent frames that are maximally dissimilar, leading it to toggle between highly diverse frames, which need not capture important objects. While the multi-document summarization objective (Weng and Merialdo 2009) overcomes this toggling effect, both it and uniform keyframe sampling can select irrelevant frames (e.g., see 2nd and 7th columns). Overall, our summary selects informative frames that best focus on the important objects that drive the first person interactions

videos), and use it to predict region importance on the ADL videos. We use our budgeted frame selection approach and set the summary frame-length to $k = 8$ (an arbitrary but reasonable number given the short length of ADL videos). For each video, we ask an independent subject to watch the video and provide a text summary that emphasizes the key happenings, in the same manner as described in Sect. 4.6.2. The resulting summaries tend to focus on specific actions and are more descriptive than those provided by the camera wearers on UT Ego. We suspect this is due to the relatively short length of each video (∼30 min). Here is an example summary:

A guy brought his laundry basket to the laundry room to do laundry. He poured in the liquid detergent and did his laundry. He then went back home and started to play a video game on TV. The guy went into his room and turned on his laptop computer and looked at a picture of a monkey. The guy went into the bathroom to wash his face and brush his teeth. The guy is now in his kitchen and poured some juice to drink. He's looking at a list and checking off his list. The guy is making tea. The guy went into the bathroom to comb his hair. The guy cleaned his kitchen floor with a broom. He then went into his bedroom and put on his shoes.

See supplementary file for all text summaries.

We then ask 10 Mechanical Turk subjects per video to compare our summaries to those of uniform sampling, keyframes (Liu and Kender 2002), and multi-document (Weng and Merialdo 2009), and ask the same set of questions as in Sect. 4.6.2. Table 3 shows the results. Out of 20 total pairwise comparisons to each baseline, our summaries were found to be better 42–68 % of the time, and worse 15–33 % of the time. In terms of how each method's summary compares to the text summary, ours, Liu and Kender (2002), Weng and Merialdo (2009), and uniform sampling scored 2.71 (±1.02), 3.58 (±0.97), 2.99 (±1.15), 2.89 (±1.06), respectively (recall that lower numbers are better: 1 being "very well" and 5 being "very poorly"). We show clear improvement over keyframes (Liu and Kender 2002), which tends to simply oscillate between bright/dark frames. Our improvements over uniform sampling and multi-document (Weng and Merialdo 2009) are less compared to those on UT Ego. This is likely due to the ADL videos being shorter in length and more structured; in ADL, the camera wearers are given a list of actions they should perform, whereas UT Ego is completely unscripted. Under these conditions, summarization algorithms that aim to select frames that are spread-out over time are likely to select meaningful frames. Still, by focusing on the important objects, our approach produces the best summaries. Figure 15 shows example summaries created by each method. Our method selects the most informative frames.

Overall, the results are a promising indication that discovering important people and objects leads to higher quality summaries for egocentric video. Not only do we better recount those objects that human viewers deem important in the context of the surrounding activity, but we also generate summaries that human viewers prefer to multiple existing summarization approaches.

5 Conclusion and Future Work

We introduced an approach to summarize egocentric video using novel egocentric cues to predict important regions. We presented two ways to adjust summary compactness: given either an importance selection criterion or a length budget. For the latter, we developed an efficient optimization strategy to recover the best k-frame summary. To our knowledge, ours is the first work to summarize videos from wearable cameras by discovering objects that may be important to the camera wearer. Existing summarization techniques rely on static cameras or low-level visual similarity, and so they fail to account for the key objects that drive first person interactions. Through extensive experiments, we showed that our approach produces significantly more informative summaries than prior methods.

Future work can expand this idea in several interesting directions. We assumed that the importance cues can be learned and shared across users, and our experiments confirmed that it is feasible. However, there are also subjective elements; e.g., depending on the user, a person that he has significant interactions with may or may not be considered important. To overcome the subjectivity, one could learn a wearer-specific model that uses input from the wearer for training to complement our wearer-independent model.

Secondly, event segmentation remains a challenge for egocentric data. With the frequent head and body motion inherent to wearable video, grouping frames according to low-level scene statistics is imperfect. In our system, this can sometimes lead to redundant keyframes showing the same object. One way to mitigate this issue is to use a GPS receiver and generate event clusters using both location information and scene appearance. This could provide better separation of events, especially when the scene appearance between two neighboring events is similar. More broadly, more robust detection of event boundaries is needed.

Finally, while our interest lies in the computer vision challenges, other sensing modalities naturally can play a role in egocentric summarization. For example, audio cues could signal person importance based on their speech near the camera, while ambient noise may be indicative of the scene type. Other sensors like an accelerometer can reveal the user's gestures and activity, while GPS coordinates could give real-world location context relevant to which objects are likely important (e.g., a plate in a restaurant, vs. an athlete in a stadium).

References

Aghazadeh, O., Sullivan, J., & Carlsson, S. (2011). Novelty detection from an egocentric perspective. In *CVPR*.

Alexe, B., Deselaers, T., & Ferrari, V. (2010). What is an object? In *CVPR*.

Aner, A., & Kender, J. R. (2002). Video Summaries through mosaic-based shot and scene clustering. In *ECCV*.

Brox, T., & Malik, J. (2011). Large displacement optical flow: Descriptor matching in variational motion estimation. *TPAMI*, *33*(3), 500–513.

Carreira, J., & Sminchisescu, C. (2010). Constrained parametric min-cuts for automatic object segmentation. In *CVPR*.

Caspi, Y., Axelrod, A., Matsushita, Y., & Gamliel, A. (2006). Dynamic stills and clip trailer. *The Visual Computer*, *22*(9), 642–652

Cheng, M.-M., Zhang, Z., Lin, W.-Y., & Torr, P. (2014). BING: Binarized normed gradients for objectness estimation at 300fpsn. In *CVPR*.

Clarkson, B., & Pentland, A. (1999). Unsupervised clustering of ambulatory audio and video. In *ICASSP*.

Doherty, A., & Smeaton, A. (2008). Combining face detection and novelty to identify important events in a visual lifelog. In *International Conference on Computer and Information Technology Workshops*.

Doherty, A., Byrne, D., Smeaton, A., Jones, G., & Hughes, M. (2008). Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In *CIVR*.

Endres, I., & Hoiem, D. (2010). Category independent object proposals. In *ECCV*.

Fathi, A., Farhadi, A., & Rehg, J. (2011). Understanding egocentric activities. In *ICCV*.

Fathi, A., Hodgins, J. K., & Rehg, J. M. (2012). Social interactions: A first-person perspective. In *CVPR*.

Felzenszwalb, P., & Huttenlocher, D. (2004). Efficient graph-based image segmentation. *IJCV*, *59*(2), 167–181.

Gao, D., Mahadevan, V., & Vasconcelos, N. (2007). The discriminant center-surround hypothesis for bottom-up saliency. In *NIPS*.

Goldman, D., Curless, B., Salesin, D., & Seitz, S. (2006). Schematic storyboarding for video visualization and editing. In *SIGGRAPH*.

Healey, J., & Picard, R. (1998). Startlecam: A cybernetic wearable camera. In *Wearable Computers*.

Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., & Wood, K. (2006). SenseCam: A retrospective memory aid. In *UBICOMP*.

Hodges, S., Berry, E., & Wood, K. (2011). Sensecam: A wearable camera which stimulates and rehabilitates autobiographical memory. *Memory*, *19*(7), 685–696.

Huynh, T., Fritz, M., & Schiele, B. (2008). Discovery of activity patterns using topic models. In *UBICOMP*.

Hwang, S. J., & Grauman, K. (2010). Accounting for the relative importance of objects in image retrieval. In *BMVC*.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, *20*(11), 1254–1259.

Jojic, N., Perina, A., & Murino, V. (2010). Structural epitome: A way to summarize one's visual experience. In *NIPS*.

Jones, M., & Rehg, J. (2002). Statistical color models with application to skin detection. *IJCV*, *46*(1), 81–96.

Kitani, K., Okabe, T., Sato, Y., & Sugimoto, A. (2011). Fast Unsupervised Ego-Action Learning for First-Person Sports Video. In *CVPR*.

Kolsch, M., & Turk, M. (2004). Robust hand detection. In *FG*.

Lee, M., & Dey, A. (2007). Providing good memory cues for people with episodic memory impairment. In *ACM SIGACCESS Conference on Computers and Accessibility*.

Lee, Y. J., Kim, J., & Grauman, K. (2011). Key-segments for video object segmentation. In *ICCV*.

Lee, Y. J., Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *CVPR*.

Li, C., & Kitani, K. M. (2013). Pixel-level hand detection for ego-centric videos. In *CVPR*.

Li, Y., Fathi, A. & Rehg, J. M. (2013). Learning to predict gaze in egocentric video. In *ICCV*.

Lin, W., & Hauptmann, A. (2006). Structuring continuous video recordings of everyday life using time-constrained clustering. In *IS&T/SPIE Symposium on Electronic Imaging*.

Liu, T., & Kender, J. R. (2002). Optimization algorithms for the selection of key frame sequences of variable length. In *ECCV*.

Liu, T., Sun, J., Zheng, N., Tang, X., & Shum, H. (2007). Learning to detect a salient object. In *CVPR*.

Liu, D., Hua, G., & Chen, T. (2009). A hierarchical visual model for video object summarization. In *TPAMI*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, *60*(2), 91–110.

Lu, Z., & Grauman, K. (2013). Story-driven summarization for egocentric video. In *CVPR*.

Mann, S. (1998). Wearcam (the wearable camera): Personal imaging systems for long term use in wearable tetherless computer mediated reality and personal photo/videographic memory prosthesis. In *Wearable Computers*.

Money, A., & Agius, H. (2008). Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, *19*(2), 121–143.

Ng, H. W., Sawahata, Y., & Aizawa, K. (2002). Summarizing wearable videos using support vector machine. In *ICME*.

Perona, P., & Freeman, W. (1998). A factorization approach to grouping. In *ECCV*.

Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *CVPR*.

Pritch, Y., Rav-Acha, A., Gutman, A., & Peleg, S. (2007). Webcam synopsis: Peeking around the world. In *ICCV*.

Rav-Acha, A., Pritch, Y., & Peleg, S. (2006). Making a long video short. In *CVPR*.

Ren, X., & Gu, C. (2010). Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*.

Ryoo, M. S., & Matthies, L. (2013). First-person activity recognition: What are they doing to me? In *CVPR*.

Simakov, D., Caspi, Y., Shechtman, E., & Irani, M. (2008). Summarizing visual data using bidirectional similarity. In *CVPR*.

Spain, M., & Perona, P. (2008). Some objects are more equal than others: Measuring and predicting importance. In *ECCV*.

Spriggs, E., la Torre, F. D., & Hebert, M. (2009). Temporal segmentation and activity classification from first-person sensing. In *CVPR Workshop on Egocentric Vision*.

Starner, T., Schiele, B., & Pentland, A. (1998a). Visual contextual awareness in wearable computing. In *ISWC*.

Starner, T., Weaver, J., & Pentland, A. (1998b). Real-time american sign language recognition using desk and wearable computer based video. *PAMI*, *20*(12), 1371–1375.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR*.

Walther, D., & Koch, C. (2006). Modeling attention to salient protoobjects. *Neural Networks*, *19*, 1395–1407.

Weng, F., & Merialdo, B. (2009). Multi-document video summarization. In *ICME*.

Wolf, W. (1996). Keyframe selection by motion analysis. In *ICASSP*.

Zhang, H. J., Wu, J., Zhong, D., & Smoliar, S. (1997). An integrated system for content-based video retrieval and browsing. In *Pattern Recognition*.