To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

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

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. 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 with 17 hours of egocentric data show the method's promise relative to existing techniques for saliency and summarization.

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.

Existing methods extract keyframes [29, 30, 8], create montages of still images [2, 4], or generate compact dynamic summaries [22, 21]. 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 do not perform *object-driven* summarization and are indifferent to the impact that each object has on generating the "story" of the video.

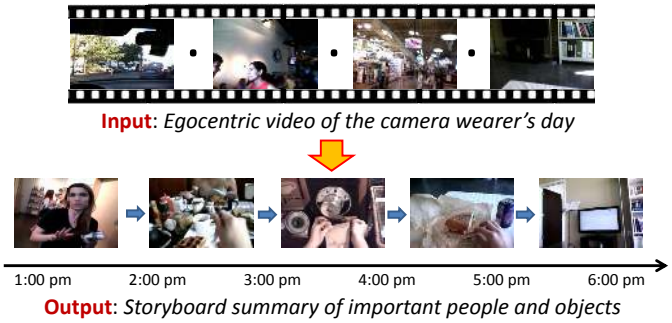

Figure 1. Our system takes as input an unannotated egocentric video, and produces a compact storyboard visual summary that focuses on the key people and objects in the video.

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 higher-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 [9]. 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. 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 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.

Contributions Our main contribution is a novel egocentric video summarization approach that is driven by predicted important people and objects. We apply our method to challenging real-world videos captured by users in uncontrolled environments, and process a total of 17 hours of video—orders of magnitude more data than previous work in egocentric analysis. Evaluating the predicted importance estimates and summaries, we find our approach outperforms state-of-the-art saliency measures for this task, and produces significantly more informative summaries than traditional methods unable to focus on the important people or objects.

2. Related Work

In this section, we review related work in video summarization, saliency detection, and egocentric data analysis.

Video summarization Static keyframe methods compute motion stability from optical flow [29] or global scene color differences [30] to select the frames that go into the summary. The low-level approach means that irrelevant frames can often be selected. 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 [2], or sequentially display automatically selected key-poses [4]. An interactive approach [8] 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 [22, 21]. While the framework aims to focus on foreground objects, it assumes a static camera and is therefore inapplicable to egocentric video. A retargeting approach aims to simultaneously preserve an original video's content while reducing artifacts [24], but unlike our approach, does not attempt to characterize the varying degrees of object importance. In a semi-automatic method [17], 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., [12]). More recent work tries to learn high-level saliency measures, whether for static images [18, 3, 6] or video [16]. 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 [25, 11]. 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 only recently begun to explore egocentric visual analysis. Early work with wearable cameras segments visual and audio data into events [5]. Recent methods explore activity recognition [7], handled object recognition [23], novelty

detection [1], or activity discovery for non-visual sensory data [10]. Unsupervised algorithms are developed to discover scenes [13] or actions [15] based on low-level visual features extracted from egocentric data. In contrast, we aim to build a visual summary, and model high-level importance of the objects present. To our knowledge, we are the first to perform visual summarization for egocentric data.

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.

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 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 based on the predicted important people and objects.

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[1], which captures video at 15 fps at 320 x 480 resolution. It is worn around the ear and looks out at the world at roughly eye-level. We collected 10 videos, each of three to five hours in length (the max Looxcie battery life), for a total of 37 hours of video.

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, and cooking.

3.2. Annotating important regions in training 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.

In order to learn meaningful egocentric properties without overfitting to any particular category, we crowd-source large amounts of annotations using Amazon's Mechanical Turk (MTurk). For egocentric videos, an object's degree of importance will highly 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.

Figure 2. Example annotations obtained using Mechanical Turk.

We carefully design two annotation tasks to capture this aspect. In the first task, we ask workers to watch a three minute accelerated video (equivalent to 10 minutes 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 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 hour video, we obtain roughly 35 text descriptions and 700 object segmentations.

3.3. Learning region importance in egocentric video

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 the segmentation method of [3]. We purposefully 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.[2] We generate roughly 800 regions per frame.

[1] http://looxcie.com/

[2] Indeed, we found KLT tracks to last only a few frames on our data.

Figure 3. Illustration of our egocentric features.

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 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 each feature in detail below.

Egocentric features Fig. 3 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 [14] trained with ground-truth hand annotations on disjoint data. We then classify any superpixel 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 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:

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

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 DoG+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 [19]. We repeat this process for each point in region r, and record their average number of positive matches:

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

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 [19]. 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 minute temporal window.

Object features In addition to the egocentric-specific features, we include three high-level (i.e., object-based) saliency cues. To model **object-like appearance**, we use the learned region ranking function of [3]. 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 this feature estimates how "object-like" a region is, and *not its importance*. It is useful for identifying full object segments, as opposed to fragments.

To model **object-like motion**, we use the feature defined in [16]. It looks at the difference in motion patterns of a region relative to its closest surrounding regions. 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 [27].

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).

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

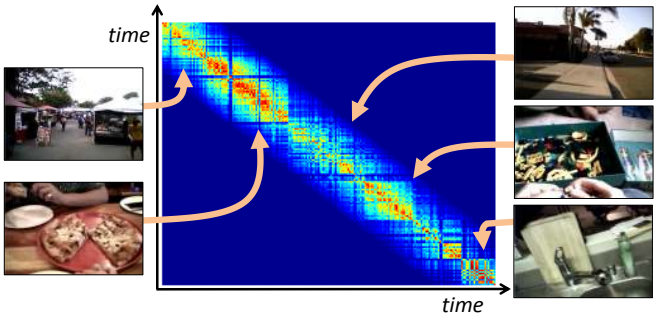

Figure 4. Distance matrix that measures global color dissimilarity between all frames. (Blue/red reflects high/low distance.) The images show representative frames of each discovered event.

it is simply "important" or "not important", so that we can meaningfully adjust the compactness of the final summary (as we demonstrate in Sec. 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 Sec. 3.2. 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 $I(r)$.

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 (e.g., the dog at home in the morning, and then the dog at the park at night).

To detect egocentric 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.[3] 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:

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

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. Since events can overlap, we retain (almost) disjoint events by eliminating those with greater than θ_{event} overlap with events with higher silhouette-coefficients [26] in a greedy manner. Higher/lower θ_{event} leads to more/fewer events in the final summary. See Fig. 4 for the distance matrix computed from one of our subject's day, and the representative frames for each discovered event.

One could further augment the distance in Eqn. 4 with GPS locations, when available (though GPS alone would be insufficient to discriminate multiple indoor positions in the same building).

3.5. Discovering an event's key people and objects

For each event, we aim to select the important people and objects that will go into the final summary, while avoiding redundancy. Given an event, we first score each bottom-up segment in each frame using our regressor. We take the highest-scored regions (where "high" depends on a user-specified summary compactness criterion, see below) and group instances of the same person or object together. 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 high-scoring 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 [20]. 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

[3] Traditional shot detection is impractical for wearable camera data; it oversegments events due to frequent head movements.

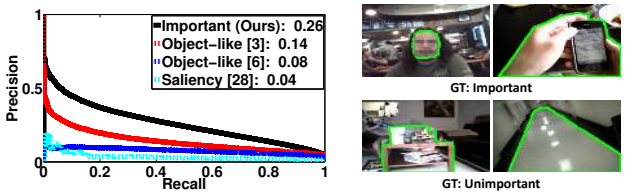

Figure 5. Precision-Recall for important object prediction across all splits, and example selected regions/frames. Numbers in the legends denote average precision. Compared to state-of-the-art high-level [3, 6] and low-level [28] saliency methods, our egocentric approach more accurately discovers the important regions.

affinity to a higher-ranked cluster. Finally, for each remaining cluster, we select the region with the highest importance score as its representative. Note that this grouping step reinforces the egocentric frequency cue described in Sec. 3.3.

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. 8). Each event can display a varying number of frames, depending on how many unique important things our method discovers. We automatically adjust the *compactness* of the summary with selection criteria on the region importance scores and event overlaps, as we illustrate in our results.

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.

4. Results

We analyze (1) the performance of our method's important region prediction, (2) our egocentric features, and (3) the accuracy and compactness of our storyboard summaries.

Dataset and implementation details We collected 10 videos from four subjects, each 3-5 hours long. 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.

We use Lab space color histograms, with 23 bins per channel, and optical flow histograms with 61 bins per direction. We set $t = 27000$, i.e., a 60 minute temporal window. We set $\theta_r = 10000$ 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).

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

Figure 6. Top 28 features with highest learned weights.

Important region prediction accuracy We first evaluate our method's ability to predict important regions, compared to three state-of-the-art high- and low-level saliency methods: (1) the object-like score of [3], (2) the object-like score of [6], and (3) the bottom-up saliency detector of [28]. 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". Since the baselines are all general-purpose metrics (not tailored to egocentric data), 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 Sec. 3.2). Some frames contain more than one important region, and some contain none, simply 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.

Fig. 5 (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.

Fig. 5 (right) shows examples that our method found to be important. The top and bottom rows show correct and incorrect predictions, respectively. Typical failure cases include under-segmenting the important object if the foreground and background appearance is similar, and detecting frequently occurring background regions to be important.

Which cues matter most for predicting importance? Fig. 6 shows the top 28 out of 105 ($= 14 + \binom{14}{2}$) features that receive the highest learned weights. 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

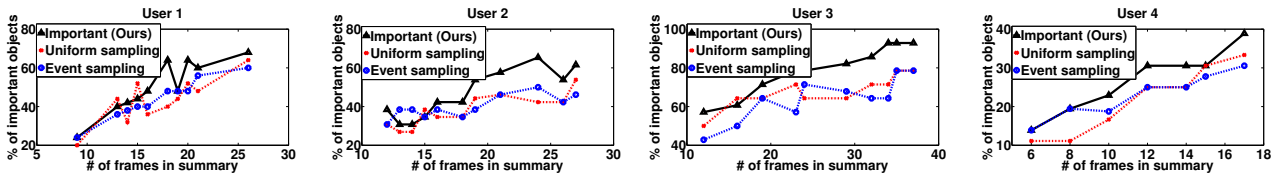

Figure 7. Comparison to alternative summarization strategies, in terms of important object recall rate as a function of summary compactness.

Figure 8. Our summary (top) vs. uniform sampling (bottom). Our summary focuses on the important people and objects.

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-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.

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 (Sec. 3.4), and then divides its keyframes evenly across events. These are natural baselines modeled after classic keyframe and event detection methods [29, 30], and both select keyframes that are "spread-out" across the video.

Fig. 7 shows 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. To vary compactness, our method varies both its selection criterion on $I(r)$ over $\{0, 0.1, \ldots, 0.5\}$ and the number of events by setting $\theta_{event} = \{0.2, 0.5\}$, for 12 summaries in total. We create summaries for the baselines with the same number of frames as those 12. 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 very 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. 8, 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.

Overall, our summaries include more important people/objects with fewer frames. For example, for User 2, our method finds 54% of important objects in 19 frames, whereas the uniform keyframe method requires 27 frames. With very short summaries, all methods perform similarly; the selected keyframes are more spread-out, so they have higher chance of including unique people/objects. With longer summaries, our method always outperforms the baselines, since they tend to include redundant frames repeating the same important person/object. On average, we find 9.13 events/video and 2.05 people/objects per event.

The two baselines perform fairly similarly to one another, though the event-based keyframe selector has a slight edge by doing "smarter" temporal segmentation. Still, both are indifferent to objects' importance in creating the story of the video; their summaries contain unimportant or redundant frames as a result.

Fig. 8 shows an example full summary from our method (top) and the uniform baseline (bottom). The colored blocks for ours indicate the automatically 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

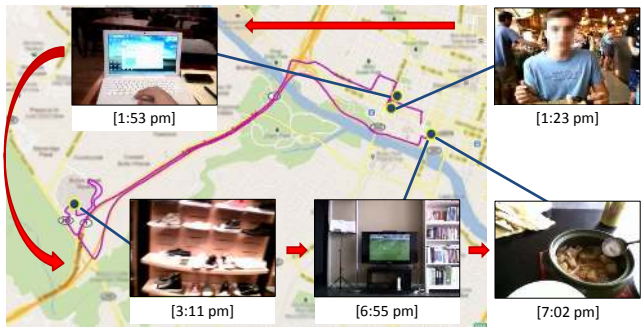

Figure 9. An application of our approach.

	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%

Table 1. User study results. Numbers indicate percentage of responses for each question, always comparing our method to the baseline (i.e., highest values in "much better" are ideal).

in the frame. In this example, our summary more clearly reveals the story: *selecting an item at the supermarket* $\rightarrow$ *driving home* $\rightarrow$ *cooking* $\rightarrow$ *eating and watching tv*.

Fig. 9 shows another example; 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.

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). The camera wearers are the best judges, since they know the full extent of their day that we are attempting to summarize.

We generate four pairs of summaries, 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 short, out of 16 total comparisons, our summaries were found to be better 68.75% of the time. Overall, these results are a promising indication that discovering important people/objects leads to higher quality summaries for egocentric video.

5. Conclusion

We developed an approach to summarize egocentric video. We introduced novel egocentric features to train a regressor that predicts important regions. Using the discovered important regions, our approach produces significantly more informative summaries than traditional methods that often include irrelevant or redundant information.

Acknowledgements Many thanks to Yaewon, Adriana, Nona, Lucy, and Jared for collecting data. This research was sponsored in part by ONR YIP and DARPA CSSG.

References

[1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty Detection from an Egocentric Perspective. In *CVPR*, 2011.
[2] A. Aner and J. R. Kender. Video Summaries through Mosaic-Based Shot and Scene Clustering. In *ECCV*, 2002.
[3] J. Carreira and C. Sminchisescu. Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *CVPR*, 2010.
[4] Y. Caspi, A. Axelrod, Y. Matsushita, and A. Gamliel. Dynamic Stills and Clip Trailer. In *The Visual Computer*, 2006.
[5] B. Clarkson and A. Pentland. Unsupervised Clustering of Ambulatory Audio and Video. In *ICASSP*, 1999.
[6] I. Endres and D. Hoiem. Category Independent Object Proposals. In *ECCV*, 2010.
[7] A. Fathi, A. Farhadi, and J. Rehg. Understanding Egocentric Activities. In *ICCV*, 2011.
[8] D. Goldman, B. Curless, D. Salesin, and S. Seitz. Schematic Storyboarding for Video Visualization and Editing. In *SIGGRAPH*, 2006.
[9] S. Hodges, E. Berry, and K. Wood. Sensecam: A Wearable Camera which Stimulates and Rehabilitates Autobiographical Memory. *Memory*, 2011.
[10] T. Huynh, M. Fritz, and B. Schiele. Discovery of Activity Patterns using Topic Models. In *UBICOMP*, 2008.
[11] S. J. Hwang and K. Grauman. Accounting for the Relative Importance of Objects in Image Retrieval. In *BMVC*, 2010.
[12] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *TPAMI*, 20(11), November 1998.
[13] N. Jojic, A. Perina, and V. Murino. Structural Epitome: A Way to Summarize One's Visual Experience. In *NIPS*, 2010.
[14] M. Jones and J. Rehg. Statistical Color Models with Application to Skin Detection. *IJCV*, 46(1), 2002.
[15] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast Unsupervised Ego-Action Learning for First-Person Sports Video. In *CVPR*, 2011.
[16] Y. J. Lee, J. Kim, and K. Grauman. Key-Segments for Video Object Segmentation. In *ICCV*, 2011.
[17] D. Liu, G. Hua, and T. Chen. A Hierarchical Visual Model for Video Object Summarization. In *TPAMI*, 2009.
[18] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to Detect a Salient Object. In *CVPR*, 2007.
[19] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2), 2004.
[20] P. Perona and W. Freeman. A Factorization Approach to Grouping. In *ECCV*, 1998.
[21] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam Synopsis: Peeking Around the World. In *ICCV*, 2007.
[22] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a Long Video Short. In *CVPR*, 2006.
[23] X. Ren and C. Gu. Figure-Ground Segmentation Improves Handled Object Recognition in Egocentric Video. In *CVPR*, 2010.
[24] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing Visual Data using Bidirectional Similarity. In *CVPR*, 2008.
[25] M. Spain and P. Perona. Some Objects are More Equal than Others: Measuring and Predicting Importance. In *ECCV*, 2008.
[26] Tan, Steinbach, and Kumar. *Introduction to Data Mining*. 2005.
[27] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *CVPR*, 2001.
[28] D. Walther and C. Koch. Modeling Attention to Salient Proto-Objects. *Neural Networks*, 19:1395–1407, 2006.
[29] W. Wolf. Keyframe Selection by Motion Analysis. In *ICASSP*, 1996.
[30] H. J. Zhang, J. Wu, D. Zhong, and S. Smoliar. An Integrated System for Content-Based Video Retrieval and Browsing. In *Pattern Recognition*, 1997.