Chapter 4
Intentional Photos from an Unintentional Photographer: Detecting Snap Points in Egocentric Video with a Web Photo Prior

Bo Xiong and Kristen Grauman

Abstract Wearable cameras capture a first-person view of the world, and offer a hands-free way to record daily experiences or special events. Yet, not every frame is worthy of being captured and stored. We propose to automatically predict *"snap points"* in unedited egocentric video—that is, those frames that look like they could have been intentionally taken photos. We develop a generative model for snap points that relies on a Web photo prior together with domain-adapted features. Critically, our approach avoids strong assumptions about the particular *content* of snap points, focusing instead on their *composition*. Using 17 h of egocentric video from both human and mobile robot camera wearers, we show that the approach accurately isolates those frames that human judges would believe to be intentionally snapped photos. In addition, we demonstrate the utility of snap point detection for improving object detection and keyframe selection in egocentric video.

4.1 Introduction

Photo overload is already well known to most computer users. With cameras on mobile devices, it is all too easy to snap images and videos spontaneously, yet it remains much less easy to organize or search through that content later. This is already the case when the user actively decides which images are worth taking. *What happens when that user's camera is always on, worn at eye-level, and has the potential to capture everything he sees throughout the day?* With increasingly portable wearable computing platforms (like Google Glass, Looxcie, etc.), the photo overload problem is only intensifying.

Of course, not everything observed in an egocentric video stream is worthy of being captured and stored. Egocentric videos contain substantial motion and are often boring to watch. Even though the camera follows the wearer's activity and

B. Xiong (✉) · K. Grauman
University of Texas at Austin, 2317 Speedway, Austin, TX 78712, USA
e-mail: bxiong@cs.utexas.edu

K. Grauman
e-mail: grauman@cs.utexas.edu

G. Hua and X.-S. Hua (eds.), *Mobile Cloud Visual Media Computing*,
DOI 10.1007/978-3-319-24702-1_4

Fig. 4.1 Can you tell which row of photos came from an egocentric camera?

approximate gaze, relatively few moments actually result in snapshots the user would have intentionally decided to take, where he actively manipulating the camera. Many frames will be blurry, contain poorly composed shots, and/or simply have uninteresting content. This prompts the key question we study in this work: can a vision system predict ***"snap points"*** **in unedited egocentric video—that is, those frames that look like intentionally taken photos?**

To get some intuition for the task, consider the images in Fig. 4.1. Can you guess which row of photos was sampled from a wearable camera, and which was sampled from photos posted on Flickr? Note that subject matter itself is not always the telling cue; in fact, there is some overlap in content between the top and the bottom rows. Nonetheless, we suspect it is easy for the reader to detect that a head-mounted camera grabbed the shots in the first row, whereas a human photographer purposefully composed the shots in the second row. These distinctions suggest that it may be possible to learn the generic properties of an image that indicate it is well composed, independent of the literal content.

While this anecdotal sample suggests that detecting snap points may be feasible, there are several challenges. First, egocentric video contains a wide variety of scene types, activities, and actors. This is certainly true for human camera wearers going about daily life activities, and it will be increasingly true for mobile robots that freely explore novel environments. Accordingly, a snap point detector needs to be largely domain invariant and generalize across varied subject matter. Secondly, an optimal snap point is likely to differ in subtle ways from its less-good temporal neighbors, i.e., two frames may be similar in content but distinct in terms of snap point quality. That means that cues beyond the standard texture and color favorites may be necessary. Finally, and most importantly, while it would be convenient to think of the problem in discriminative terms (e.g., training a snap point versus non-snap point classifier), it is burdensome to obtain adequate and unbiased labeled data. Namely, we would need people to manually mark frames that appear intentional, and to do so at a scale to accommodate arbitrary environments.

We introduce an approach to detect snap points from egocentric video that requires no human annotations. The main idea is to construct a generative model of what human-taken photos look like by sampling images posted on the Web. Snapshots that people upload to share publicly online may vary vastly in their content, yet all share the key facet that they were intentional snap point moments. This makes them an ideal source of positive exemplars for our target learning problem. Furthermore, with such

Fig. 4.2 Understandably, while effective for human-taken photos (*left*), today's best object detectors break down when applied to egocentric video data (*right*). Each image displays the person detections by the DPM [9] object detector

a Web photo prior, we sidestep the issue of gathering negatively labeled instances to train a discriminative model, which could be susceptible to bias and difficult to scale. In addition to this prior, our approach incorporates domain adaptation to account for the distribution mismatch between Web photos and egocentric video frames. Finally, we designate features suited to capturing the framing effects in snap points.

We propose two applications of snap point prediction. For the first, we show how snap points can improve object detection reliability for egocentric cameras. It is striking how today's best object detectors fail when applied to arbitrary egocentric data (see Fig. 4.2). Unsurprisingly, their accuracy drops because detectors trained with human-taken photos (e.g., the Flickr images gathered for the PASCAL VOC benchmark) do not generalize well to the arbitrary views seen by an ego camera. We show how snap point prediction can improve the precision of an off-the-shelf detector, essentially by predicting those frames where the detector is most trustworthy. For the second application, we use snap points to select keyframes for egocentric video summaries.

We apply our method to 17.5 h of videos from both human-worn and robot-worn egocentric cameras. We demonstrate the absolute accuracy of snap point prediction compared to a number of viable baselines and existing metrics. Furthermore, we show its potential for object detection and keyframe selection applications. The results are a promising step toward filtering the imminent deluge of wearable camera video streams.

4.2 Related Work

We next summarize how our idea relates to existing work in analyzing egocentric video, predicting high-level image properties, and using Web image priors.

Egocentric video analysis: Egocentric video analysis, pioneered in the 90s [35, 44], is experiencing a surge of research activity, thanks to today's portable devices. The primary focus is on object [29, 38] or activity recognition [6, 8, 24, 29, 37, 39, 43]. Compared with well-posed photographs, egocentric videos contain more uninformative frames, which are often poorly composed and illuminated [11]. Motion cues [38] in egocentric video are useful to segment foreground objects and therefore improve object recognition. Gaze information [29] can also improve both object and activity recognition. No prior work explores snap point detection.

We consider object detection and keyframe selection as applications of snap points for unconstrained wearable camera data. In contrast, prior work for detection in egocentric video focuses on controlled environments (e.g., a kitchen) and hand-held objects (e.g., the mixing bowl) [6, 8, 29, 38, 43]. Nearly, all prior keyframe selection work assumes third-person static cameras (e.g., [31, 32]), where all frames are already intentionally composed, and the goal is to determine which are the representative for the entire video. In contrast, snap points aim to discover intentional-looking frames, not maximize diversity or representativeness. Some video summarization work tackles dynamic egocentric video [27, 34]. Such methods could exploit snap points as a filter to limit the frames they consider for summaries. Our main contribution is to detect human-taken photos, not a novel summarization algorithm.

We are not aware of any prior work using purely visual input to automatically trigger a wearable camera, as we propose. Methods in ubiquitous computing use manual intervention [35] or external nonvisual sensors [15, 16] (e.g., skin conductivity or audio) to trigger the camera. Our image-based approach is complementary; true snap points are likely a superset of those moments where abrupt physiological or audio changes occur.

Predicting high-level image properties: A series of interesting works predict properties from images like saliency [33], professional photo quality [20], memorability [18], aesthetics, interestingness [4, 13], or suitability as a candid portrait [10]. These methods train a discriminative model using various image descriptors, and then apply it to label human-taken photos. In contrast, we develop a generative approach with (unlabeled) Web photos, and apply it to *find* human-taken photos. Critically, a snap point need not be beautiful, memorable, etc., and it could even contain mundane content. Snap points are thus a broader class of photos. This is exactly what makes them relevant for the proposed object detection application. In contrast, an excellent aesthetics detector (for example) would fire on a narrower set of photos, eliminating non-aesthetic photos that could nonetheless be amenable to off-the-shelf object detectors.

Web image priors: The Web is a compelling resource for data-driven vision methods. Both the volume of images as well as the accompanying noisy meta-data open up many possibilities. Most relevant to our work are methods that exploit the biases of human photographers. This includes work on discovering iconic images of landmarks [28, 42, 47] (e.g., the Statue of Liberty) or other tourist favorites [1, 14, 19, 22] by exploiting the fact that people tend to take similar photos of popular sites. Similarly, the photos users upload when trying to sell a particular object (e.g., a used car) reveal that object's canonical viewpoints, which can help select keyframes to summarize short videos of the same object [21]. Event video summarization [23] can also benefit from Web image collections of the same event. Our method also learns about human framing or composition biases, but, critically, in a manner that transcends the specific content of the scene. That is, rather than learn when a popular landmark or object is in view, we want to know when a well-composed photo of *any* scene is in view. Our Web photo prior represents the photos humans intentionally take, independent of subject matter.

Our approach[1] uses a nonparametric representation of snap points, as captured by a large collection of Web photos. At a high level, this relates to work in vision exploiting big data and neighbor-based learning. This includes person detection [46], scene parsing with dense correspondences [30], geographic localization [14], action recognition [5], and pose estimation [40]. Beyond the fact that our task is unique and novel, all these methods assume labels on the training data, whereas our method relies on the distribution of photos themselves.

4.3 Approach

Our goal is to detect snap points, which are those frames within a continuous egocentric video that appear as if they were composed with intention, as opposed to merely observed by the person wearing the camera. In traditional camera-user relationships, this "trigger" is left entirely to the human user. In the wearable camera-user relationship, however, the beauty of being hands-free and always-on should be that the user no longer has to interrupt the flow of his activity to snap a photo. Notably, whether a moment in time is photoworthy is only partially driven by the subject matter in view. The way the photo is composed is similarly important, as is well understood by professional photographers and intuitively known by everyday camera users.

We take a nonparametric, data-driven approach to learn what snap points look like. First, we gather unlabeled Web photos to build the prior (Sect. 4.3.1), and extract image descriptors that capture cues for composition and intention (Sect. 4.3.2). Then, we estimate a domain-invariant feature space connecting the Web and ego sources (Sect. 4.3.3). Finally, given a novel egocentric video frame, we predict how well it agrees with the prior in the adapted feature space (Sect. 4.3.4). Figure 4.3 shows the overview of our approach. To illustrate the utility of snap points, we also explore applications for object detection and keyframe selection (Sect. 4.3.5).

Section 4.4 will discuss how we systematically gather ground truth labels for snap points using human judgments, which is necessary to evaluate our method, but, critically, is *not* used to train it.

4.3.1 Building the Web Photo Prior

Faced with the task of predicting whether a video frame is a snap point or not, an appealing solution might be to train a discriminative classifier using manually labeled exemplars. Such an approach has proven successful for learning other high-level image properties, like aesthetics and interestingness [4, 13], quality [20], canonical views [21], or memorability [18]. This is thanks in part to the availability of relevant meta-data for such problems: users on community photo albums manually score

[1] This chapter expands upon our work as first presented at ECCV 2014 [50].

Fig. 4.3 Overview of our approach. Our method takes an unedited egocentric video as input, and predicts how well it agrees with the Web photo prior in a domain-adapted feature space. We leverage the fact that Internet photos are "free" positives of intentional photos, and so our method does not require any explicitly labeled data

images for visual appeal [4, 20], and users uploading ads online manually tag the object of interest [21].

However, this familiar paradigm is problematic for snap points. Photos that appear human-taken exhibit vast variations in appearance, since they may have almost arbitrary content. This suggests that large-scale annotations would be necessary to cover the space. Furthermore, snap points must be isolated within ongoing egocentric video. This means that labeling *negatives* is tedious—each frame must be viewed and judged in order to obtain clean labels.

Instead, we devise an approach that leverages *unlabeled* images to learn snap points. The idea is to build a prior distribution using a large-scale repository of Web photos uploaded by human photographers. Such photos are by definition human-taken, span a variety of contexts, and (by virtue of being chosen for upload) have an enhanced element of *intention*. We use these photos as a generative model of snap points.

We select the Scene UNderstanding (SUN) database as our Web photo source [48], which originates from Internet search for hundreds of scene category names. Our choice is motivated by two main factors. First, the diversity of photos is high—899 categories in all drawn from 70 K WordNet terms—and there are many of them (130 K). Second, its scope is fairly well matched with wearable camera data. Human- or robot-worn cameras observe a variety of dailylife scenes and activities, as well as interactions with other people. SUN covers not just locations, but settings that satisfy "I am in a *place*, let's go to a *place*" [48], which includes many scene-specific interactions, such as shopping at a pawnshop, visiting an optician, driving in a car, etc. See Fig. 4.4.

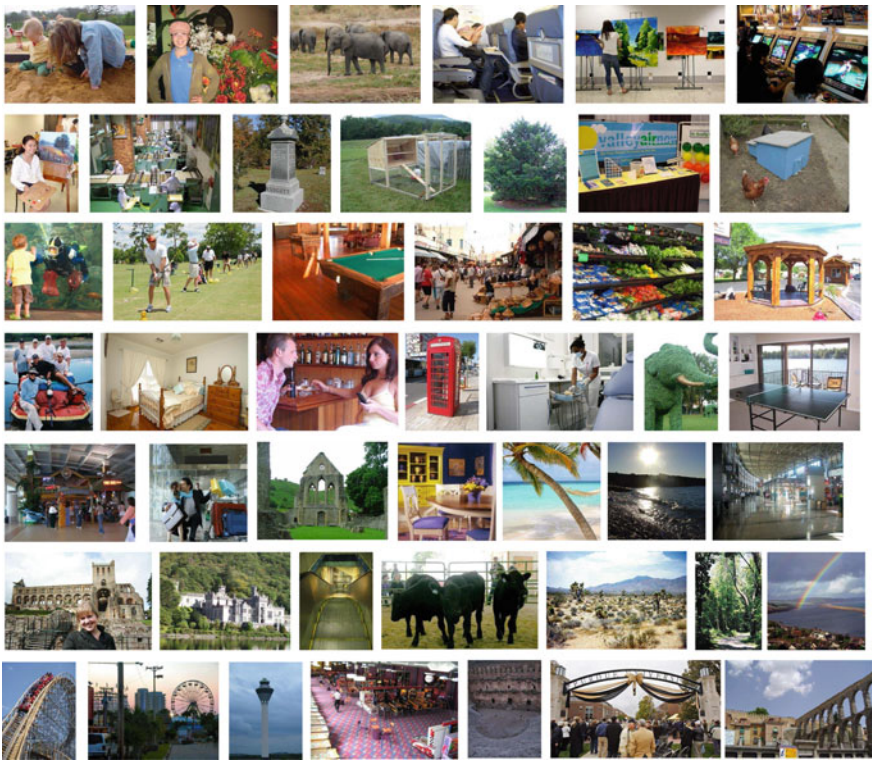

Fig. 4.4 Example images from the SUN dataset [48]. It contains a diverse category of scene types and a wide range of objects

4.3.2 Image Descriptors for Intentional Cues

To represent each image, we designate descriptors to capture intentional composition effects.

Motion: Non-snap points will often occur when a camera wearer is moving quickly, or turning his head abruptly. We therefore extract a descriptor to summarize *motion blur*, using the blurriness estimate of [2]. We also explored flow-based motion features, but found their information to be subsumed by blur features computable from individual frames.

Composition: Snap points also reflect intentional framing effects by the human photographer. This leads to spatial regularity in the main line structures in the image—e.g., the horizon in an outdoor photo, buildings in a city scene, the table surface in a restaurant—which will tend to align with the image axes. Thus, we extract a *line alignment* feature: we detect line segments using the method in [25], and then record a histogram of their orientations with 32 uniformly spaced bins. To capture framing via the 3D structure layout, we employ the geometric class probability map [17]. We

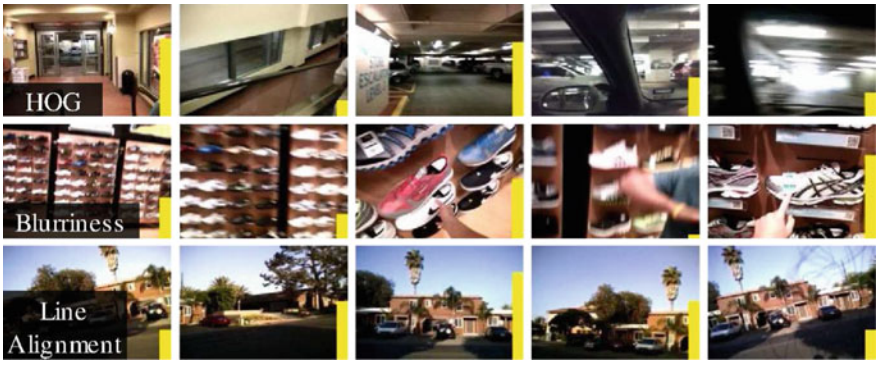

Fig. 4.5 Illustration of HOG, blurriness, and line alignment features on a short sequence of egocentric video frames. Each frame shows a *bar* in *bottom right* indicating how much each descriptor agrees with the Web prior. Here, each frame with the highest bar in each mini-sequence would rate highest as a snap point (if using each feature alone). The line alignment feature helps to find snap points that correspond to the moment when the camera wearer looks straight at the scene. The blurriness feature helps to find clear frames, and the HOG composition feature helps to find semantically meaningful frames

also extract GIST [36], HOG [3], self-similarity (SSIM) [41], and dense SIFT [26], all of which capture alignment of interior textures, beyond the strong line segments. An accelerometer, when available, could also help gage coarse alignment; however, these descriptors offer a fine-grained visual measure helpful for subtle snap point distinctions. See Fig. 4.5.

Feature combination: For all features but line alignment, we use code and default parameters provided by [48]. We reduce the dimensionality of each feature using principal component analysis (PCA) to compactly capture 90 % of its total variance. We then standardize each dimension to ($\mu = 0, \sigma = 1$) and concatenate the reduced descriptors to form a single vector feature space X, which we use in what follows.

4.3.3 Adapting from the Web to the Egocentric Domain

While we expect egocentric video snap points to agree with the Web photo prior along many of these factors, there is also an inherent mismatch between the statistics of the two domains. Egocentric video is typically captured at low resolution with modest quality lenses, while online photos (e.g., on Flickr) are often uploaded at high resolution from high-quality cameras. Egocentric videos often contain frames that are blurry or badly composed. Figure 4.6 shows some typical egocentric frames and images from SUN dataset, both from shopping malls. The examples show that despite some partial overlap in content, there is also a clear domain shift between the two sources of images.

Fig. 4.6 Comparison of shopping mall frames from egocentric video and shopping mall images from SUN dataset

Therefore, we establish a domain-invariant feature space connecting the two sources. Given unlabeled Web photos and egocentric frames, we first compute a subspace for each using PCA. Then, we recover a series of intermediate subspaces that gradually transit from the "source" Web subspace to the "target" egocentric subspace. We use the geodesic flow kernel (GFK) algorithm of [12], an unsupervised domain adaptation method which requires no labeled target data and is kernel-based. The algorithm computes a geodesic flow kernel which can be used to measure similarity in feature space. Since we assume no labels on the target domain and use a kernel-based classifier, this makes GFK a good fit. In contrast, supervised domain adaptation algorithms, which require labels on the target domain, would not be applicable.

Let $\boldsymbol{x}_i, \boldsymbol{x}_j \in X$ denote image descriptors for a Web image i and egocentric frame j. The idea is to compute the projections of an input $\boldsymbol{x}_i$ on a subspace $\phi(t)$, for all $t \in [0, 1]$ along the geodesic path connecting the source and target subspaces in a Grassmann manifold. Values of t closer to 0 correspond to subspaces closer to the Web photo prior; values of t closer to 1 correspond to those more similar to egocentric video frames. The infinite set of projections is achieved implicitly via the geodesic flow kernel [12] (GFK):

$$K_{GFK}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle z_i^{\infty}, z_j^{\infty} \rangle = \int_0^1 (\phi(t)^T \boldsymbol{x}_i)^T (\phi(t)^T \boldsymbol{x}_j) dt, \tag{4.1}$$

where z_i^{∞} and z_j^{∞} denote the infinite-dimensional features concatenating all projections of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ along the geodesic path.

Intuitively, this representation lets the two slightly mismatched domains (Web and ego) "meet in the middle" in a common feature space, letting us measure similarity between both kinds of data without being overly influenced by their superficial resolution/sensor differences.

4.3.4 Predicting Snap Points

With the Web prior, image features, and similarity measure in hand, we can now estimate how well a novel egocentric video frame agrees with our prior. We take a simple data-driven approach. We treat the pool of Web photos as a nonparametric distribution, and then estimate the likelihood of the novel ego frame under that distribution based on its nearest neighbors' distances.

Let $W = \{\boldsymbol{x}_1^w, \ldots, \boldsymbol{x}_N^w\}$ denote the N Web photo descriptors, and let $\boldsymbol{x}^e$ denote a novel egocentric video frame's descriptor. We retrieve the k nearest examples $\{\boldsymbol{x}_{n_1}^w, \ldots, \boldsymbol{x}_{n_k}^w\} \subset W$, i.e., those k photos that have the highest GFK kernel values when compared to $\boldsymbol{x}^e$.[2] Then we predict the snap point confidence for $\boldsymbol{x}^e$:

$$S(\boldsymbol{x}^e) = \sum_{j=1}^{k} K_{GFK}\left(\boldsymbol{x}^e, \boldsymbol{x}_{n_j}^w\right), \tag{4.2}$$

where higher values of $S(\boldsymbol{x}^e)$ indicate that the test frame is more likely to be human-taken. For our dataset of $N = 130\,\text{K}$ images, similarity search is fairly speedy (0.01 s per test case in Matlab), and could easily be scaled for much larger N using hashing or kd-tree techniques.

This model follows in the spirit of prior data-driven methods for alternative tasks, e.g., [14, 30, 40, 46], the premise being to keep the learning simple and let the data speak for itself. However, our approach is label-free, as all training examples are (implicitly) positives, whereas the past methods assume at least weak meta-data annotations.

While simple, our strategy is very effective in practice. In fact, we explored a number of more complex alternatives—one-class SVMs, Gaussian mixture models, nonlinear manifold embeddings—but found them to be similar or inferior to the neighbor-based approach. The relatively lightweight computation is a virtue given our eventual goal to make snap point decisions onboard a wearable device.

4.3.5 Leveraging Snap Points for Egocentric Video Analysis

Filtering egocentric video down to a small number of probable snap points has many potential applications. We are especially interested in how they can bolster object detection and keyframe selection. We next devise strategies for each task that leverage the above predictions $S(\boldsymbol{x}^e)$.

[2] We use $k = 60$ based on preliminary visual inspection, and found that results were similar for other k values of similar order ($k \in [30, 120]$).

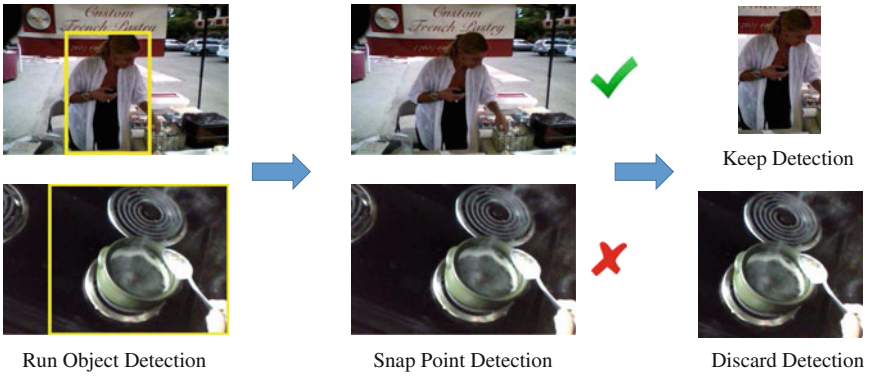

Fig. 4.7 Overview of our approach to improve object detection. We first run a deformable parts object detector trained with intentional (Flickr) photos. Then we run our snap point detection to determine whether we want to trust the object detection results. If the frame does not appear to be a snap point, we will discount the object detector's outputs

Object Detection

In the object recognition literature, it is already disheartening that how poorly detectors trained on one dataset tend to generalize to another [45]. Unfortunately, things are only worse if one attempts to apply those same detectors on egocentric video (recall Fig. 4.2). Why is there such a gap? Precisely because today's very best object detectors are learned from human-taken photos, whereas egocentric data on wearable cameras—or mobile robots—consist of very few frames that match those statistics. For example, a winning person detector on PASCAL VOC trained with Flickr photos, like the deformable parts model (DPM) [9], expects to see people in similarly composed photos, but only a fraction of egocentric video frames will be consistent and thus detectable.

Our idea is to use snap points to predict those frames where a standard object detector (trained on human-taken images) will be most trustworthy. This way, we can improve precision; the detector will avoid being misled by incidental patterns in non-snap point frames. See Fig. 4.7 for an overview of our approach. We implement the idea as follows, using the DPM as an off-the-shelf detector.[3] We score each test ego frame by $S(\boldsymbol{x}^e)$, and then keep all object detections in those frames scoring above a threshold τ. We set τ as 30 % of the average distance between the Web prior images and egocentric snap points. For the remaining frames, we eliminate any detections (i.e., flatten the DPM confidence to 0) that fall below the confidence threshold in the standard DPM pipeline [9]. In effect, we turn the object detector "on" only when it has high chance of success.

[3] http://www.cs.berkeley.edu/~rbg/latent/.

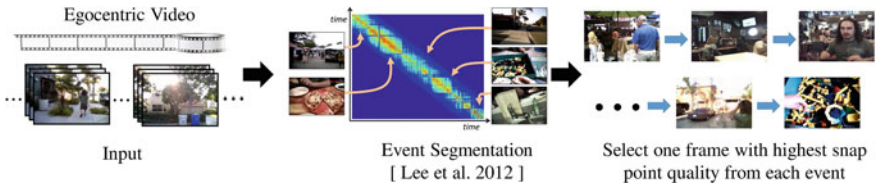

Fig. 4.8 Overview of our keyframe selection method. Given an egocentric video, we first identify temporal event segments [27] and for each such event, we select the frame most confidently scored as a snap point

Keyframe Selection

As a second application, we use snap points to create keyframe summaries of egocentric video. The goal is to take hours of wearable data and automatically generate a visual storyboard that captures key events. We implement a simple selection strategy. First, we identify temporal event segments using the color- and time-based grouping method described in [27], which finds chunks of frames likely to belong to the same physical location or scene. This is done by performing complete-link agglomerative clustering on both global appearance and temporal nearness of all frames of egocentric video. Then, for each such event, we select the frame most confidently scored as a snap point. See Fig. 4.8 for an illustration.

Our intent is to see if snap points, by identifying frames that look intentional, can help distil the main events in hours of uncontrolled wearable camera data. Our implementation is a proof of concept to demonstrate snap points' utility. We are not claiming a new keyframe selection strategy, a problem studied in depth in prior work [27, 31, 32, 34].

4.4 Datasets and Collecting Ground Truth Snap Points

Datasets: We use two egocentric datasets. The first is the publicly available UT Egocentric Dataset (**Ego**),[4] which consists of four videos of 3–5 h each, captured with a head-mounted camera by four people doing unscripted daily life activities (eating, working, shopping, driving, etc.). The second is a mobile robot dataset (**Robot**) newly collected for this project. We used a wheeled robot to take a 25 min video both indoors and outdoors on campus (coffee shops, buildings, streets, pedestrians, etc.). The camera is a FireFly USB 2.0 camera, connected to the robot with a pan-tilt unit. The camera on the robot moves constantly from left to right, pauses, and then rotates back in order to cover a wide range of viewpoints. Our robot was able to take pictures from different viewpoints even when physically located at the same place.

[4] http://vision.cs.utexas.edu/projects/egocentric_data.

Both the human and robot datasets represent incidentally captured video from always-on, dynamic cameras, and unscripted activity. We found other existing ego collections less suited to our goals, either due to their focus on a controlled environment with limited activity (e.g., making food in a kitchen [8, 29]) or their use of chest-mounted or fisheye lens cameras [7, 37], which do not share the point of view of intentional hand-held photos.

Ground truth: Our method requires no labeled data for learning: it needs only to populate the Web prior with human-taken photos. However, to *evaluate* our method, it is necessary to have ground truth human judgments about which ego frames are snap points. The following describes our crowdsourced annotation strategy to get reliable ground truth.

We created a "magic camera" scenario to help MTurk annotators understand the definition of snap points. Their instructions were as follows: *Suppose you are creating a visual diary out of photos. You have a portable camera that you carry all day long, in order to capture everyday moments of your daily life. For instance, you would like to capture scenes such as a dining place where you have dinner with friends, a dog you stopped to pet, children you saw playing in a park, the cashier at the check-out counter, a peaceful street where you took a walk at sunset, or a small but elegant shop that you visited. Unfortunately, your magic camera can also trigger itself from time to time to take random pictures, even while you are holding the camera. At the end of the day, all pictures, both the ones you took intentionally and the ones accidentally taken by the camera, are mixed together.* ***Your task is to distinguish the pictures that you took intentionally from the rest of pictures that were accidentally taken by your camera.***

In Fig. 4.9, we show the instructions that were used on Amazon Mechanical Turk to collect annotations. Workers were required to rate each image into one of the four categories: (a) very confidently intentional, (b) somewhat confident intentional, (c) somewhat confident accidental, and (d) very confident accidental. Since the task can be ambiguous and subjective, we issued each image to 5 distinct workers. We obtained labels for 10,000 frames in the ego data and 2,000 frames in the Robot data, sampled at random.

We devised a scoring system to obtain reliable fine-grained ground truth. Every time a frame receives a rating of category (a), (b), (c), or (d) from any of the 5 workers, it receives 5, 2, −1, −2 points, respectively. Most workers assign category (b) or (c) to the frames and rarely assign category (a), unless they certainly believe that the image is taken intentionally. As a result, if a frame receives a rating of category (a), we reward the frame 5 points. On the other hand, since there are many more negative frames than positive frames, if a frame receives a rating of category (d), it does not get penalized as much (−2). This lets us rank all ground truth examples by their true snap point strength.

To alternatively map these total scores across all 5 annotations to binary ground truth, we threshold a frame's total score: strictly more than 10 points is deemed intentional. This means a frame must receive at least one vote on category (a) in order to be considered an intentional frame (5 votes on category (b) means all workers had some doubt if the frame was intentional; it will receive a total of 10 points but not more than 10 points). If one outlier worker assigns an intentional frame a rating of category (d), as long as the frame receives at least two ratings of category (a) and

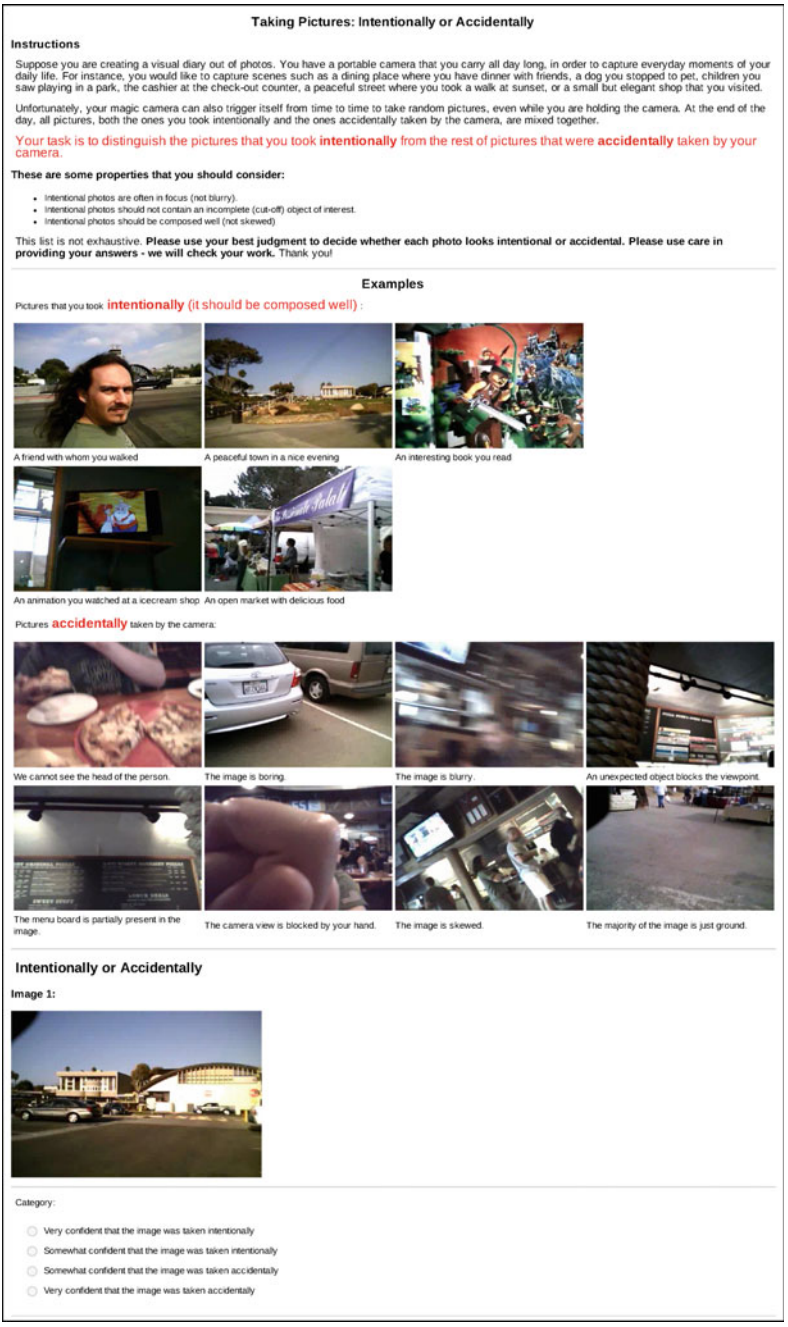

Fig. 4.9 Instructions used on Amazon Mechanical Turk to collect annotations

two ratings of category (b) from the other four different workers, the frame will still be an intentional frame.

Annotators found 14 % of the ego frames and 23 % of the robot frames to be snap points, respectively. The robot data contain more snap points because the robot we used to collect data had less motion compared with human. Out of the 10,000 labeled frames in the ego data, there are 998 frames that all five workers reach consensus on the category, 1748 frames that four workers reach consensus on, and 3871 frames that three workers reach consensus. Out of the 2,000 labeled frames in the robot data, there are 213 frames that all five workers reach consensus on the category, 306 frames that four workers reach consensus on, and 691 frames that three workers reach consensus on. The total MTurk cost was about $500.

Our dataset and software are available online.[5]

4.5 Results

We experiment on the two datasets described above, ego and robot, which together comprise 17.5 h of video. Since no existing methods perform snap point detection, we define several **baselines** for comparison:

- **Saliency** [33]: uses the CRF-based saliency method of [33] to score an image. This baseline reflects that people tend to compose images with a salient object in the center. We use the implementation of [4], and use the CRF's log probability output as the snap point confidence.
- **Blurriness** [2]: uses the blur estimates of [2] to score an image. It reflects that intentionally taken images tend to lack motion blur. Note that blur is also used as a feature by our method; here we isolate how much it would solve the task if used on its own, with no Web prior.
- **People likelihood**: uses a person detector to rank each frame by how likely it is to contain one or more people. We use the max output of the DPM [9] detector. The intuition is people tend to take images of their family and friends to capture meaningful moments, and as a result, many human-taken images contain people. In fact, this baseline also implicitly captures how well-composed the image is, since the DPM is biased to trigger when people are clear and unoccluded in a frame (recall Fig. 4.2).
- **Discriminative SVM**: uses a RBF kernel SVM trained with the ground truth snap points/non-snap points in the ego data. We run it with a leave-one-camera-wearer-out protocol, training on 3 of the ego videos and testing on the 4th. This baseline lets us analyze the power of the unlabeled Web prior compared to a standard discriminative method. Note, it requires substantially more training effort than our approach.

[5] http://vision.cs.utexas.edu/projects/ego_snappoints.

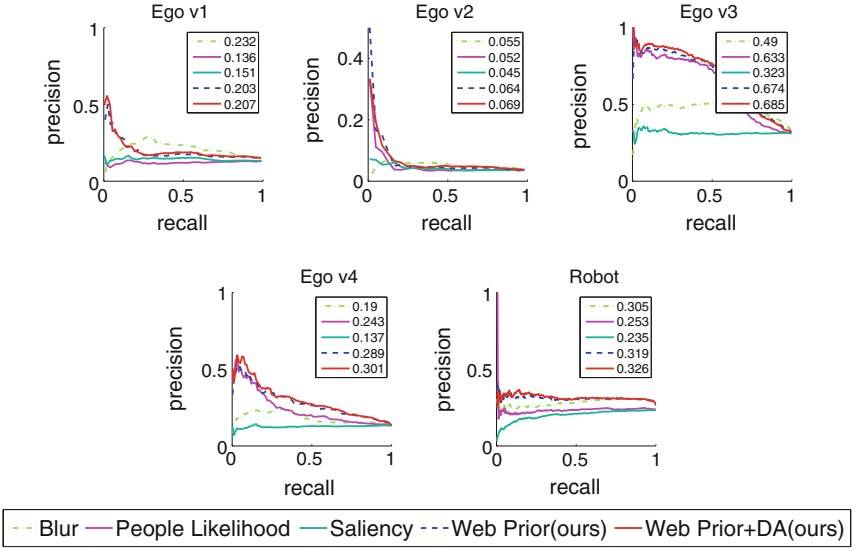

Fig. 4.10 Snap point detection precision/recall on the four ego videos (*top row* and *bottom left*) and the robot video (*bottom right*). Numbers in legend denote mAP. Best viewed in color (color figure online)

4.5.1 Snap Point Accuracy

First, we quantify how accurately our method predicts snap points. Figure 4.10 shows the precision–recall curves for our method and the three unsupervised baselines (saliency, blurriness, people likelihood). Table 4.1 shows the accuracy in terms of two standard rank quality metrics, Spearman's correlation ρ and Kendall's τ. While the precision–recall plots compare predictions against the binarized ground truth, these metrics compare the full orderings of the confidence-valued predictions against the raw MTurk annotators' ground truth scores (cf. Sect. 4.4). They capture that even for two positive intentional images, one might look better than the other to human judges. We show results for our method with and without the domain adaptation (DA) step.

Overall, our method outperforms the baselines. Notably, the same prior succeeds for both the human-worn and robot-worn cameras. Using both the Web prior and DA gives best results, indicating the value of establishing a domain-invariant feature space to connect the Web and ego data.

On ego video 4 (v4), our method is especially strong, about a factor of 2 better than the nearest competing baseline (Blur). On v2, mAP is very low for all methods, since v2 has very few true positives (only 3 % of its frames, compared to 14 % on average for Ego). Still, we see stronger ranking accuracy with our Web prior and DA. On v3, people likelihood fares much better than it does on all other videos, likely because v3 happens to contain many frames with nice portraits. On the robot data,

Table 4.1 Snap point ranking accuracy (higher rank correlations are better)

Methods	Ego v1		Ego v2		Ego v3		Ego v4		Robot	
Rank coefficient	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Blurriness	**0.347**	**0.249**	0.136	0.094	0.479	0.334	0.2342	0.162	0.508	0.352
People likelihood	0.002	0	−0.015	−0.011	0.409	0.289	0.190	0.131	0.198	0.134
Saliency	0.027	0.019	0.008	0.005	0.016	0.011	−0.021	−0.014	−0.086	−0.058
Web prior (Ours)	0.321	0.223	0.144	0.100	**0.504**	**0.355**	**0.452**	0.317	0.530	0.373
Web prior+DA (Ours)	0.343	0.239	**0.179**	**0.124**	0.501	0.353	**0.452**	**0.318**	**0.537**	**0.379**

however, it breaks down, likely because of the increased viewpoint irregularity and infrequency of people.

While our method is nearly always better than the baselines, on v1 Blur is similar in ranking metrics and achieves higher precision for higher recall rates. This is likely due to v1's emphasis on scenes with one big object, like a bowl or tablet, as the camera wearer shops and cooks. The SUN Web prior has less close-up object-centric images; this suggests that we could improve our prior by increasing the coverage of object-centric photos, e.g., with ImageNet-style photos.

Figure 4.11 shows examples of images among those our method ranks most confidently (top) and least confidently (bottom) as snap points, for both datasets. We see that its predictions capture the desired effects. Snap points, regardless of their content, do appear intentional, whereas non-snap points look accidental. Please see our project webpage for more extensive video results.

Figure 4.12 examines the effectiveness of each feature we employ, were we to take them individually. We see that each one has something to contribute, though they are best in combination (Fig. 4.10). HOG on ego is exceptionally strong. This is in spite of the fact that the exact locations visited by the ego camera wearers are almost certainly disjoint from those that happen to be in the Web prior. This indicates that the prior is broad enough to capture the diversity in appearance of everyday environments.

All baselines so far required no labeled images, same as our approach. Next we compare to a discriminative approach that uses manually labeled frames to train a snap point classifier. Figure 4.13 shows the results, as a function of the amount of labeled data. We give the SVM-labeled frames from the held-out ego videos. (We do not run it for the robot data, since the only available labels are scene-specific; it is not possible to run the leave-one-camera-wearer-out protocol.) *Despite learning without any explicit labels*, our method generally outperforms the discriminative SVM. The discriminative approach requires thousands of hand-labeled frames to come close to our method's accuracy in most cases. This is a good sign: while expanding the Web prior is nearly free, expanding the labeled data is expensive and tedious. In fact, if anything, Fig. 4.13 is an optimistic portrayal of the SVM baseline. That is because both the training and testing data are captured on the very same camera; in general scenarios, one would not be able to count on this benefit.

The results above are essential to validate our main idea of snap point detection with a Web prior. Next we provide proof of concept results to illustrate the utility of snap points for practical applications.

4.5.2 Object Detection Application

Today's best object detection systems are trained thoroughly on human-taken images—for example, using labeled data from PASCAL VOC or ImageNet. This naturally makes them best suited to run on human-taken images at test time. Our data statistics suggest only 10–15 % of egocentric frames may fit this bill. Thus,

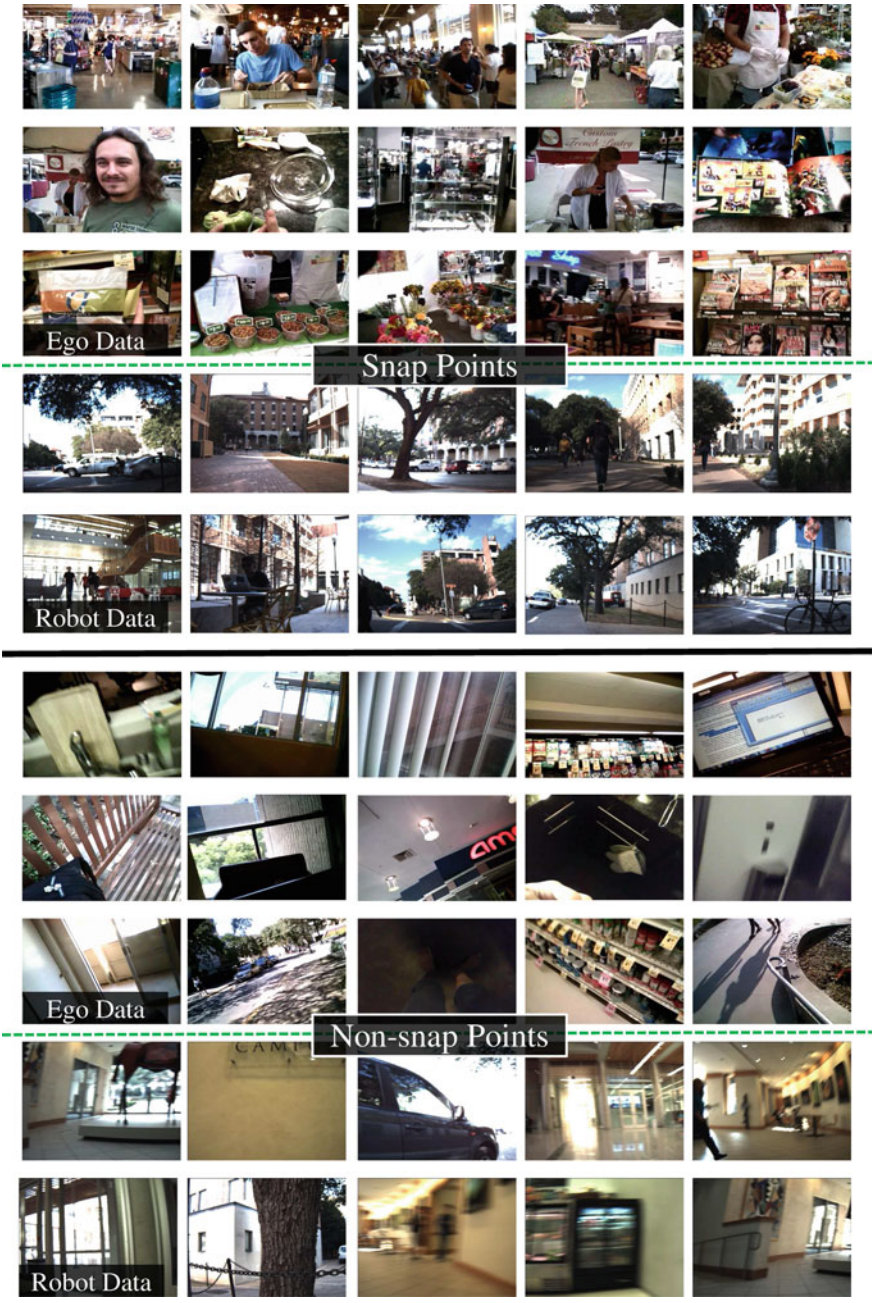

Fig. 4.11 Frames our method rates as likely (*top*) or unlikely (*bottom*) snap points. Our predictions capture the desired effects: snap points appear intentional while non-snap points look accidental

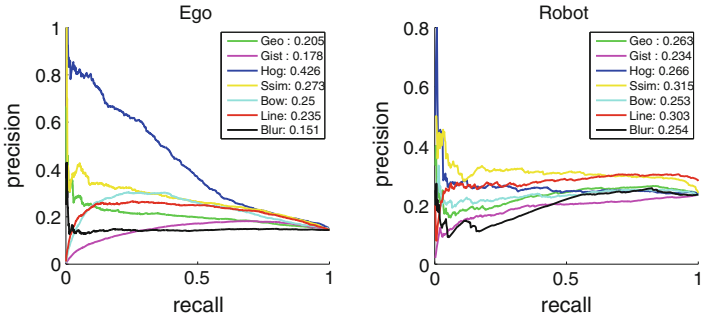

Fig. 4.12 Accuracy per feature if used in isolation. Performance is best when using all features

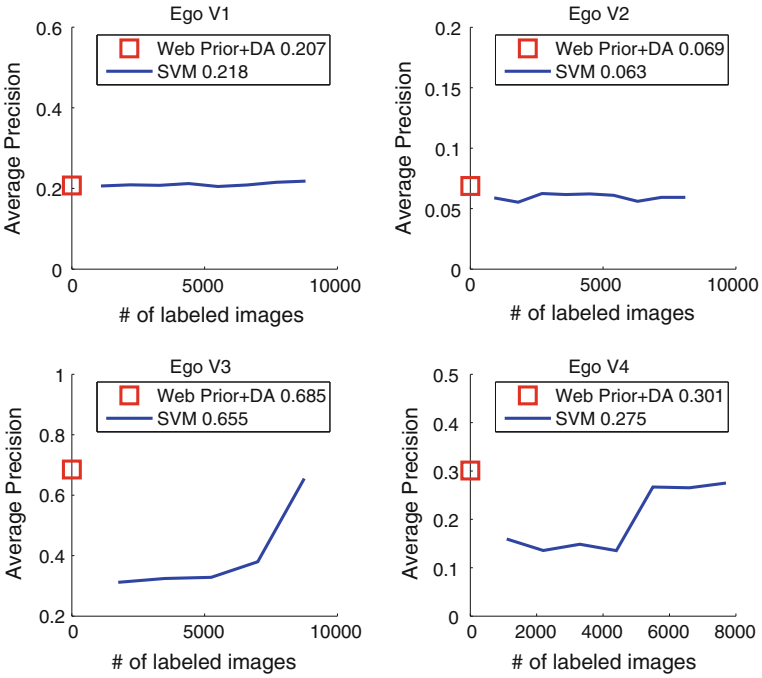

Fig. 4.13 Comparison to supervised baseline. SVM's mAP (legend) uses *all* labeled data

using the method defined in Sect. 4.3.5, we aim to use snap points to boost object detection precision.

We collected ground truth person and car bounding boxes for the ego data via DrawMe [49]. Since we could not afford to have all 17.5 h of video labeled, we sampled the labeled set to cover 50–50 % snap points and non-snap points. We obtained labels for 1000 and 200 frames for people and cars, respectively (cars are more rare in the videos).

Figure 4.14 shows the results, using the PASCAL detection criterion. We see that snap points improve the precision of the standard DPM detector, since they let us ignore frames where the detector is not trustworthy. Of course, this comes at the cost of some recall at the tails. This seems like a good trade-off for detection in video, particularly, since one could anchor object tracks using these confident predictions, and then iteratively refine less confident predictions with object tracks, in order to make up the recall.

Figure 4.15 shows some eliminated person detections of both success and failure cases. While many false positive detections were eliminated, a few true positive detections from non-snap points frames were also eliminated. In these cases, where the detector is robust to the poorly composed frames, our approach can reduce recall.

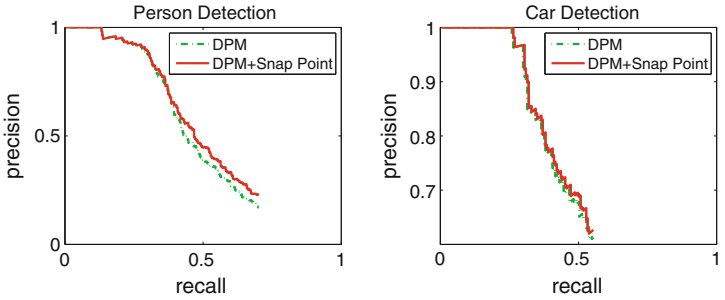

Fig. 4.14 Snap points boost precision for an off-the-shelf object detector by focusing on frames that look human-taken

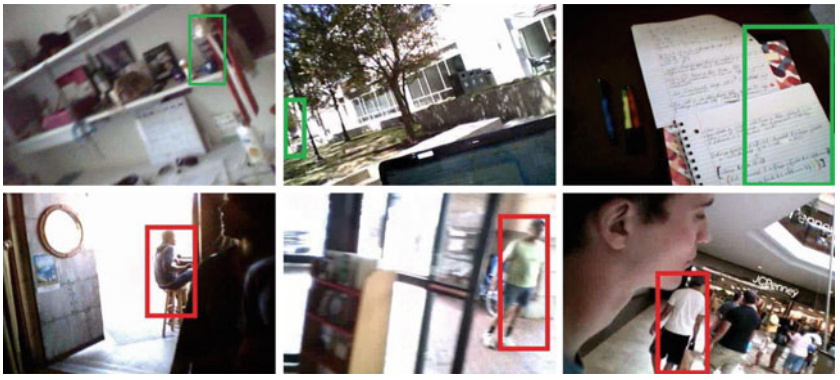

Fig. 4.15 Examples of person detections that are eliminated by our method. The three frames in the *top row* are false detections that are properly eliminated by snap point detection. We also include three failure cases in the *bottom row*, where true positive detections on non-snap point frames are eliminated

4.5.3 Keyframe Selection Application

Keyframe or "storyboard" summaries are an appealing way to peruse long egocentric video, to quickly get the gist of what was seen. Such summaries enable noveleak interfaces to let a user "zoom-in" on time intervals that appear most relevant. As a final proof of concept result, we apply snap points for keyframe selection, using the method defined in Sect. 4.3.5.

Figures 4.16 and 4.17 show example results on the ego data, where the average event length is 30 min, and Fig. 4.18 shows results on the robot data. Keyframe selection requires subjective evaluation; we have no ground truth for quantitative evaluation. We present our results alongside a baseline that uses the exact same event segmentation as [27] (cf. Sect. 4.3.5), but selects each event's frame at random instead of prioritizing snap points. We also show the result of an existing keyframe selection method [32], which selects a sequence of keyframes that maximize diversity.

We see that the snap point-based summaries contain well-composed images for each event. The baseline, while seeing the same events, often uses haphazard shots that do not look intentionally taken. The method of Liu et al. [32] maximizes diversity in the low-level image feature space and often selects semantically uninformative frames that do not look intentionally taken, suggesting it is not a good fit for keyframe selection on egocentric video. While our method generally appears to outperform the baselines, it can make mistakes as well. For example, our method picks a frame of a shopping mall in the first event of the second video (see the first frame in the seventh row in Fig. 4.16), when it would be preferable to pick a frame when a friend was eating as done by the baseline (see the first frame in the ninth row in Fig. 4.16) since the main event was having lunch with a friend. This suggests that our method can be improved by reasoning about importance or human attention, so that it could better select keyframes from important time intervals or when the camera wearer was paying attention.

4.6 Conclusions and Future Work

An onslaught of lengthy egocentric videos is imminent, making automated methods for intelligently filtering the data of great interest. Whether for easing the transfer of existing visual recognition methods to the ego domain, or for helping users filter content to photoworthy moments, snap point detection is a promising direction. Our data-driven solution uses purely visual information and requires no manual labeling. Our results on over 17 h of video show that it outperforms a variety of alternative approaches.

Ultimately, we envision snap point detection being run online with streaming egocentric video, thereby saving power and storage for an always-on wearable device. Currently, a bottleneck is feature extraction. In future work, we will consider ways to triage feature extraction for snap points, and augment the generative model with

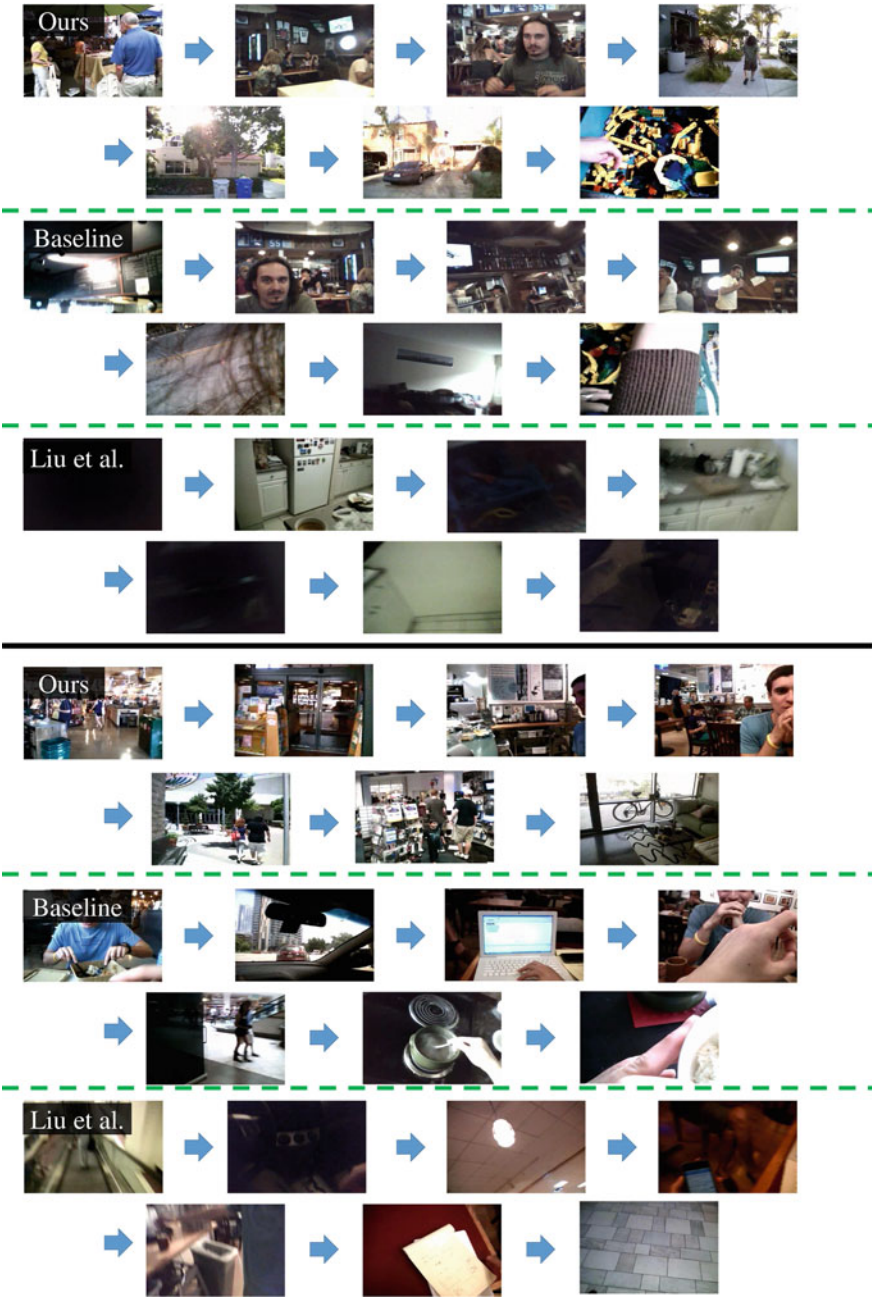

Fig. 4.16 Example keyframe selections for two 4-h ego videos. Top result is produced by our snap point method, middle result is the event segmentation baseline, and bottom result is the existing method of [32]. Our method was able to produce a sequence of informative and well-composed photos. In the first video, we can see that the camera wearer went to a market, had lunch, took a walk, and then went back home to play lego. The other two summaries are less informative

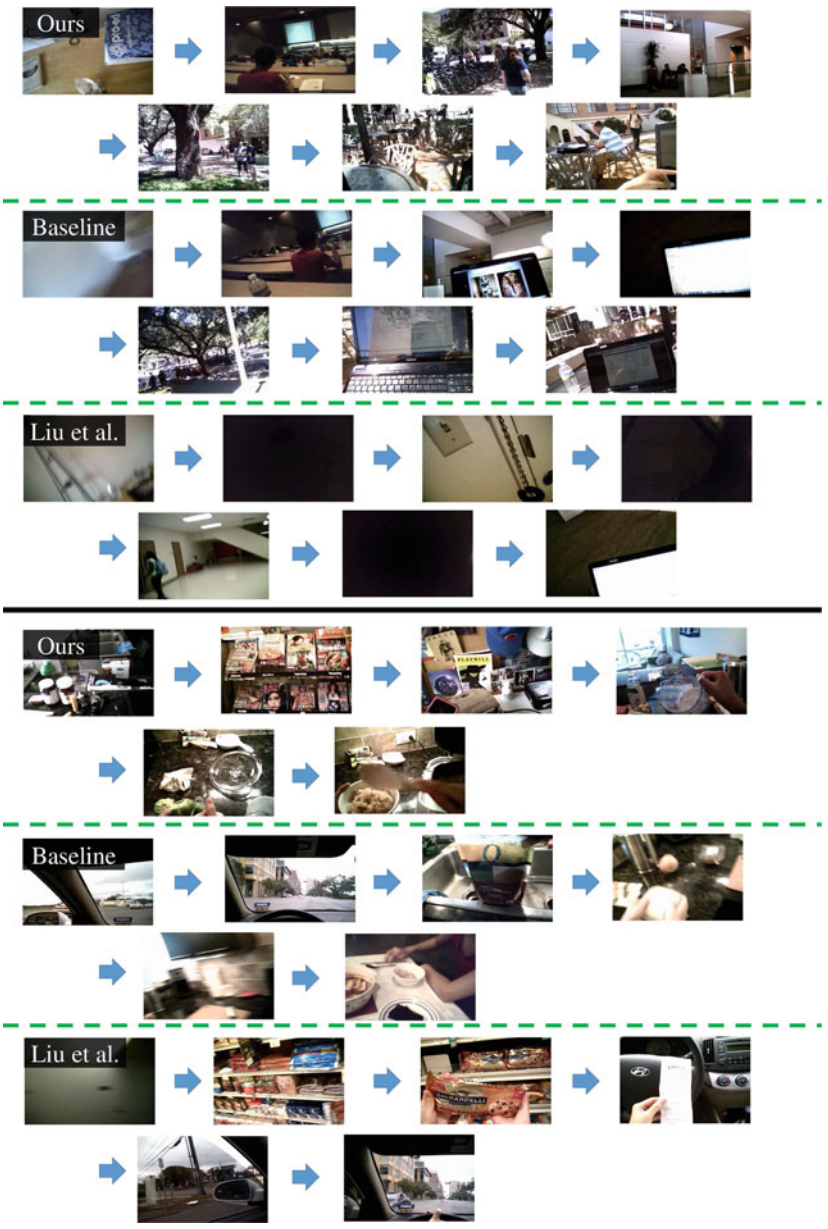

Fig. 4.17 Example keyframe selections for two 4-h ego videos. Top result is produced by our snap point method, middle result is the event segmentation baseline, and bottom result is the existing method of [32]. While our method generally appears to outperform the baselines, it can make mistakes as well. Our method picks a frame of a bunch of books on a bookshelf in the second event of the second video (see the *second frame* in the *seventh row*), when it would be preferable to pick a frame of groceries as done by the method of [32] (see the *second frame* in the *second last row*) since the main event was groceries shopping

Fig. 4.18 Example keyframe selections for the robot video. Top result is produced by our snap point method, middle result is the event segmentation baseline, and bottom result is the existing method of [32]. We see that our snap point-based method was able to pick a representative frame for each location that the robot visited. The other two summaries contain blurry or uninformative frames

user-labeled frames to learn a personalized model of snap points. While we are especially interested in wearable data, our methods may also be applicable to related sources, such as bursts of consumer photos or videos captured on mobile phones.

Acknowledgments This research is sponsored in part by ONR YIP and gifts from Intel and Google. We thank Piyush Khandelwal, Jacob Menashe, and Peter Stone for helping us collect the Robot video. We also thank Yong Jae Lee for helpful discussions and for kindly providing code to generate baseline results.

References

1. Chen, C.Y., Grauman, K.: Clues from the beaten path: location estimation with bursty sequences of tourist photos. In: CVPR (2011)
2. Crete-Roffet, F., Dolmiere, T., Ladret, P., Nicolas, M.: The blur effect: perception and estimation with a new no-reference perceptual blur metric. In: SPIE (2007)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)

4. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: CVPR (2011)
5. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV (2003)
6. Fathi, A., Farhadi, A., Rehg, J.: Understanding egocentric activities. In: ICCV (2011)
7. Fathi, A., Hodgins, J., Rehg, J.: Social interactions: a first-person perspective. In: CVPR (2012)
8. Fathi, A., Rehg, J.: Modeling actions through state changes. In: CVPR (2013)
9. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI **32**(9), 1627–1645 (2010)
10. Fiss, J., Agarwala, A., Curless, B.: Candid portrait selection from video. In: TOG (2011)
11. Flint, A., Reid, I., Murray, D.: Learning texton models for real-time scene context. In: CVPR Workshops (2009)
12. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR (2012)
13. Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., Van Gool, L.: The interestingness of images. In: ICCV (2013)
14. Hays, J., Efros, A.: im2gps: estimating geographic information from a single image. In: CVPR (2008)
15. Healey, J., Picard, R.: Startlecam: a cybernetic wearable camera. In: Wearable Computers (1998)
16. Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., Wood, K.: SenseCam: a retrospective memory aid. In: UBICOMP (2006)
17. Hoiem, D., Efros, A., Hebert, M.: Recovering surface layout from an image. In: IJCV (2007)
18. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What makes an image memorable? In: CVPR (2011)
19. Kalogerakis, E., Vesselova, O., Hays, J., Efros, A., Hertzmann, A.: Image sequence geolocation with human travel priors. In: ICCV (2009)
20. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: CVPR (2006)
21. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-scale video summarization using web-image priors. In: CVPR (2013)
22. Kim, G., Xing, E.: Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines. In: CVPR (2013)
23. Kim, G., Sigal, L., Xing, E.P.: Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In: CVPR (2014)
24. Kitani, K., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: CVPR (2011)
25. Kosecka, J., Zhang, W.: Video compass. In: ECCV (2002)
26. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
27. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: CVPR (2012)
28. Li, X., Wu, C., Zach, C., Lazebnik, S., Frahm, J.M.: Modeling and recognition of landmark image collections using iconic scene graphs. In: ECCV (2008)
29. Li, Y., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: ICCV (2013)
30. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing: label transfer via dense scene alignment. In: CVPR (2009)
31. Liu, D., Hua, G., Chen, T.: A hierarchical visual model for video object summarization. PAMI **32**(12), 2178–2190 (2010)
32. Liu, T., Kender, J.: Optimization algorithms for the selection of key frame sequences of variable length. In: ECCV (2002)
33. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. In: CVPR (2007)
34. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: CVPR (2013)
35. Mann, S.: 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 (1998)

36. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. In: IJCV (2001)
37. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR (2012)
38. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: CVPR (2010)
39. Ryoo, M., Matthies, L.: First-person activity recognition: what are they doing to me? In: CVPR (2013)
40. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: ICCV (2003)
41. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07, pp. 1–8. IEEE (2007)
42. Simon, I., Seitz, S.: Scene segmentation using the wisdom of crowds. In: ECCV (2008)
43. Spriggs, E., la Torre, F.D., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: Workshop on Egocentric Vision, CVPR (2009)
44. Starner, T., Schiele, B., Pentland, A.: Visual contextual awareness in wearable computing. In: International Symposium on Wearable Computers (1998)
45. Torralba, A., Efros, A.: Unbiased look at dataset bias. In: CVPR (2011)
46. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large dataset for non-parametric object and scene recognition. PAMI **30**(11), 1958–1970 (2008)
47. Weyand, T., Leibe, B.: Discovering favorite views of popular places with iconoid shift. In: ICCV (2011)
48. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: CVPR (2010)
49. Xiao, J.: Princeton vision toolkit (2013). http://vision.princeton.edu/code.html
50. Xiong, B., Grauman, K.: Detecting snap points in egocentric video with a web photo prior. In: ECCV (2014)