Storyline Representation of Egocentric Videos with an Application to Story-based Search

Bo Xiong
The University of Texas at Austin
bxiong@cs.utexas.edu

Gunhee Kim
Seoul National University
gunhee@snu.ac.kr

Leonid Sigal
Disney Research
lsigal@disneyresearch.com

Abstract

Egocentric videos are a valuable source of information as a daily log of our lives. However, large fraction of egocentric video content is typically irrelevant and boring to re-watch. It is an agonizing task, for example, to manually search for the moment when your daughter first met Mickey Mouse from hours-long egocentric videos taken at Disneyland. Although many summarization methods have been successfully proposed to create concise representations of videos, in practice, the value of the subshots to users may change according to their immediate preference/mood; thus summaries with fixed criteria may not fully satisfy users' various search intents. To address this, we propose a storyline representation that expresses an egocentric video as a set of jointly inferred, through MRF inference, story elements comprising of actors, locations, supporting objects and events, depicted on a timeline. We construct such a storyline with very limited annotation data (a list of map locations and weak knowledge of what events may be possible at each location), by bootstrapping the process with data obtained through focused Web image and video searches. Our representation promotes story-based search with queries in the form of AND-OR graphs, which span any subset of story elements and their spatio-temporal composition. We show effectiveness of our approach on a set of unconstrained YouTube egocentric videos of visits to Disneyland.

1. Introduction

The recent emergence of lightweight and wearable egocentric cameras allows collection of vast volumes of data that visualize our experiences and social interactions [8]. The point-of-view data are potentially valuable as they have been shown to be useful in estimating gaze [25] or eye contact [44], and simplifying recognition of activities [7, 18, 27] and object-person interactions [9, 27, 31]. Despite its appeal, however, the challenge with egocentric data is that it consists of long and unstructured content, most of which is boring for the users. Hence, recently, there has been a large focus on efficient summarization techniques for egocentric videos [20, 23]. However, such methods may be required to make user- and intent-agnostic decisions about importance of objects [20] (through pre-trained detectors) or events in videos [3] and select either keyframes or subshots that maximize the *perceived* utility and diversity.

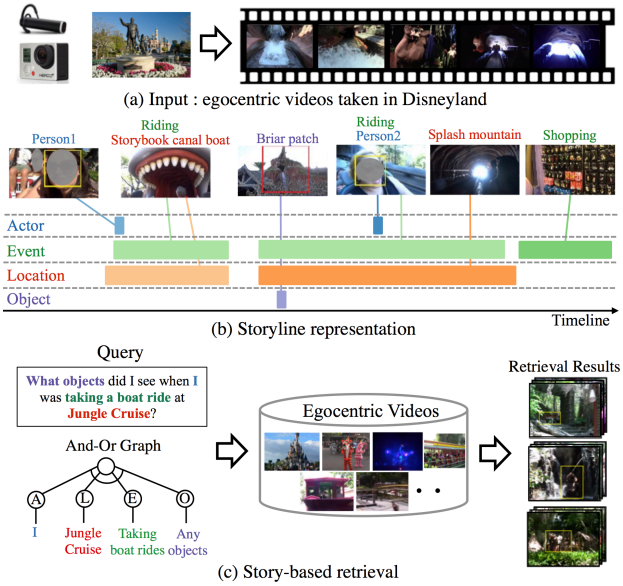


Figure 1. Motivation for storyline representation of egocentric videos. (a) An input video. (b) Storyline representation with four story elements (*i.e.* actors, events, locations, objects) on a timeline. (c) Story-based video search for a query represented by an AND-OR graph that combines any subsets of story elements.

In this paper, our objective is somewhat different; we aim at developing an approach for *representing* a given egocentric video using a storyline representation, as is exemplified with trip to *Disneyland* in Fig.1. Inspired by story research in psychology and AI [37], we extract four story elements: {*Actors*, *Location*, *Supporting objects*, *Events*} on a *timeline*. We infer these story elements jointly, using an MRF formulation, based on weakly-supervised online data (as opposed to annotations, *e.g.* [20]), and automatically account for the co-occurrences between the various story elements such as watching wild animals (supporting

object) while taking a boat ride (event) in *Jungle Cruise* (location) in *Disneyland*. The core application of the storyline representation is the focused semantic story-based video search and retrieval that cannot be achieved by conventional keywords, content, or objects-based searches (See Fig.1(c)). We define spatio-temporal queries using simple AND-OR graphs [41] that combine any subsets of story elements (*e.g.* At what attractions did we meet princesses *Tiana*? Show me our walk to the castle? What did we do after eating lunch?) Compared to summarization, we provide user-specific and intent-centric approach for retrieving desired content from the unstructured videos using semantically structured queries.

Although our approach is general enough to work for any domain with sufficient amount of online data and egocentric videos (*e.g.* city tours or museum tours), here we focus on *visiting Disneyland* in particular. Any theme park and/or traveler destination, including *Disneyland*, is an appealing domain for a few key reasons: (1) people are more willing to instrument themselves with egocentric cameras when they embark on special experiences with their family and friends; and (2) storyline representation can play a central role in such settings for the purpose of personalization, recommendation, and video search. Theme parks provide different sets of attractions and entertainment, only parts of which individual visitors can experience in a single day. Through story-based queries, users can preview attractions or paths that are taken by other families with similar demographics and interests. For example, story-based queries of the following form can be informative for new visitors: *To what attraction do families with 4 year-old boys most frequently go to after visiting a castle*? In summary, our storyline representation potentially enables a variety of applications, including virtual exploration, path planning, and travel recommendations among many others.

1.1. Related work

We discuss representative previous works from two lines of research that are closely related to our work. We then discuss the relation of some components of our framework to recent works in computer vision.

Storyline Representation. Storylines have been used as an important concept to summarize video clips, or allow users to intuitively explore or edit them. The stories are implemented in different forms, for example, annotated schematic storyboards [12], cartoon pictures with word balloons [42], and AND-OR graphs of human actions [13]. Our work differs in two key ways. First, we deal with a set of unstructured egocentric videos that are contributed by general users. Most previous works use structured domains including movies [4], TV series [40], or sports broadcasts [13]. Second, our storyline visualizes temporal changes of interactions between a rich set of story elements.

The work of [23] is similar to ours in that it also addresses the story-based summary from user-centric videos. However, the final output of [23] is chains of coherent video subshots that are largely based on low-level image features and object detection results. In contrast, we jointly infer and visualize four high-level story elements. The work of [40] creates *StoryGraphs* that visualize the storyline of TV episodes on the timeline. However, it only deals with interactions between actors for structured videos like TV series. In their follow-up work [39], story-based video search is also discussed, but it requires textual supervision of plot synopses, subtitles, and transcripts, all of which are not available for general users' egocentric videos. Our work is also related to [17] that leverages a large set of Flickr images and YouTube videos to build storylines. However, in [17], videos are simply summarized with a small set of selective keyframes, and their story graphs are chronologically connected image clusters obtained using only low-level image features. Finally, in [16] photo streams are leveraged to create storylines in *Disneyland*, but video use, including of first-person videos, is not discussed.

Egocentric Vision Research. Recent studies on egocentric video have explored the tasks of object recognition [30], activity recognition [7] and novelty detection [1]. Some works also aim to efficiently visualize egocentric videos, such as summarization [20, 23] and snap points detection [43]. The work of [20] outputs a sequence of frames containing important objects and people from egocentric videos based on image region ranking obtained from low-level cues. A subshot-based summarization of [23] considers the influence between subshots in order to capture event connectivity. Other researches aim to recognize social interaction [8], location [2] and human motion [29]. In contrast, our method jointly models multiple high-level semantic story elements. Instead of learning how to make decision on inherent importance of subshots, as is the goal in much egocentric summarization, we shift the focus to providing a semantically rich representation that can help search effectively. Thus, a variety of semantic queries can be answered without re-training or re-tuning of the model.

Image-based localization. Location, which is one of our story elements, estimation is related to the broader task of image-based localization (*e.g.* IM2GPS [14]). Most image-based localization approaches employ structure-from-motion algorithms [33] to reconstruct geometry of the scene and then proceed with 2D-to-3D image-to-geometry matching. This can be both expensive and require dense imagery of locations of interest, which is inappropriate for our task. Instead, we build on purely image-based localization techniques that amount to matching a query image to a gallery of images representing location directly [14, 34].

Object discovery. Our unsupervised supporting object discovery is related to recent works in mid-level object mining

that attempt to discover image patches that correspond to semantic objects or object parts. Most such works focus on discriminative clustering with pruning [15, 21, 35]. A notable departure is never-ending image learner [5] where objects and their sub-categories are learned in a weakly supervised manner from text web search. The difference with [5] is that, instead of using object names as search queries, since we do not know what supporting objects appear or exist beforehand, we search for locations, from which supporting objects are discovered in an unsupervised manner.

1.2. Contributions

Our contributions can be summarized as follows.

(1) To the best of our knowledge, our work is the first attempt to automatically create storyline representation, and support story-based semantic retrieval for unstructured egocentric videos, particularly in the tourism domain.

(2) We define four sets of story elements and propose a joint inference algorithm to infer and visualize them on a timeline. We also formulate the story-based retrieval based on simple AND-OR graphs through marginalization using the Metropolis-Hastings algorithm.

(3) We evaluate our approach with a Disneyland dataset collected from YouTube. We show that our joint inference approach outperforms alternative baselines for the discovery of story elements and story-based retrieval.

2. Problem Formulation

Given an input egocentric video, our goal is to jointly recognize and visualize story elements: *actors*, *locations*, *events*, and *supporting objects* (defined in detail in Section 2.2) on a *timeline*. We first obtain various training data sources from the Internet in order to train classifiers that assign confidence to presence of each story element at each frame (Section 2.1). We then jointly infer these story elements over time, taking into account temporal continuity and mutual context of co-occurrences between them (Section 3). Finally, we present a visual summary of interactions between story elements on the timeline and show how our storylines support search and retrieval with various story-based queries (Section 4).

2.1. The Data

The Disneyland park consists of multiple districts, each of which includes a sets of attractions, dining, and other entertainment (*e.g.* parades and stage shows). For simplicity, we hereafter use the term *attractions*, without distinction, to indicate the locations corresponding to all of the attractions, dining, and entertainment. We denote the set of attractions by $\mathcal{L}$. We obtain the list of attractions from Disney's official maps, resulting in $|\mathcal{L}| = 64$ attractions. For dining, we consider a *character dining* as a separate attraction, and all the other dining experiences as a single *restaurant* attraction. The *character dining* (*e.g.* *Goofy's Kitchen*) is a set of restaurants where visitors can take pictures with characters.

Training sets of images and videos. Once list of attractions is defined, we crawl two types of training data from the Internet: sets of attraction images and videos. We query each attraction name to download at maximum 100 and $1,000$ top-ranked images from Google and Flickr[1], respectively. We then manually remove irrelevant images. These images are used to learn location classifiers and to discover supporting objects and learn corresponding detectors.

Training videos are downloaded from YouTube by querying the same attraction names. Training videos are not restricted to be egocentric. Most training videos are short, around 2-3 minutes long, of a subject participating in one event at one attraction. After manual cleanup of irrelevant videos, we obtain at maximum 100 training videos per attraction, which are used for training event classifiers.

Test Set of Egocentric Videos. We collect test egocentric videos captured in the target domain (*i.e.* Disneyland park) from YouTube[2]. Since we are interested in the interaction between different story elements, we impose several requirements on the selected test videos; (i) each video should be longer than 5 minutes, and (ii) contain recordings of multiple *actors* (*i.e.* travelers with family or friends) who visit multiple attractions (*i.e.* at least 3 different attractions) and participate in more than one event (*e.g.* walking, shopping, dining and taking rides).

Consequently, we download 10 egocentric videos from YouTube. The number of videos is relatively small, mainly because we limited ourselves to collection of egocentric videos under above mentioned terms[3]. However, our videos are sufficient representative of general egocentric videos on YouTube. While the dataset of [8] is also Disneyland-related, we exclude it here because it focuses on interactions among visitors and lacks Disney attraction experiences.

2.2. Definition of Story Elements

In the following, we discuss each of four story elements: {*Actors*, *Locations*, *Supporting objects*, *Events*}.

Actors. We define *actors* as friends or family members who interact with the subject recording the egocentric video. The number of actors is not fixed beforehand, and is driven by the face detection and association in the video. In order to reliably discriminate actors from bystanders appearing accidentally in the frame, we perform the following procedure: (i) we run face detection using the Fraunhofer

[1] We download only creative-commons licensed images to ensure we have the rights to use them (original image content was not altered).

[2] Videos are, again, collected under creative-commons licensing terms, or with explicit permission of the users, to ensure we have the right to use or display them (original content of videos was not altered).

[3] A majority of YouTube videos are not creative-commons licensed.

engine [32] to detect actor candidates, (ii) we perform temporal clustering and filtering, informed by face bounding box sizes, to extract tracks of faces that are sufficiently long and at a fairly high resolution. We discuss details of actor detection in Section 3.1.

Locations. It is useful to find at which location each frame is likely to have been taken, because visitors' activities and stories can be segmented according to attractions. To this end, we perform *attraction-based localization* in which we assign likelihood to each frame, in the video, of it being captured at each of the defined attractions. In this paper, we use terms *attractions* and *locations* interchangeably. We present the algorithm in detail in Section 3.2.

Supporting objects. Supporting objects that interact with actors play an important role not only in storyline representation but also in story element estimation. The detection of characteristic objects can hint at the location, or at the event, that actors must be enjoying (*e.g.* *Lightning McQueen* character is most often encountered in the *Car's land* district). Unlike the attractions, we do not have a readily available prior information about object lists or named entities, which would allow even weak supervision. Therefore, we propose a fully unsupervised algorithm for object discovery (*i.e.* detecting supporting objects and training detectors for them), which we explain in Section 3.4.

Events. We define *events* as various activities in which visitors participate in Disneyland. We enumerate the following common events: *walking*, *shopping*, *dining*, *watching performance*, *posing with Disney characters*, *getting on ground transportation*, *taking a boat*, and *going on rides*.

2.3. Other Domains Beyond Disneyland

Our approach is easily extendible to other domains of tourism. Among the four types of story elements, only class list of *locations* and *events* should be defined *a priori*, the discovery of *actors* and *objects* are unsupervised. Since plenty of online public information is available for all popular tourist destinations, it is typically not challenging to define *locations* and *events*. Here, for example, we leverage the official visitors' maps for the Disneyland park.

Given a new domain of interest (*e.g.* city tour of Paris), we perform two pre-requisite steps before applying our approach: (i) constructing list of *locations* and *events* using the official guide books, maps, or webpages, and (ii) collecting training sets of images and videos from Google/Flickr and YouTube, as done in previous section. Once the input data are ready, we can use exactly the same algorithms, without modification, to extract story elements from egocentric videos of the new domain.

Note that although we focus on the Disneyland park here, our method is applicable as long as one can list classes of locations and events, and collect training images and videos.

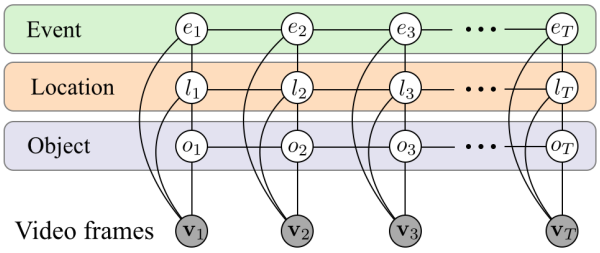


Figure 2. The graphical model for joint inference of story elements between locations, events, and objects.

3. Joint Inference of Story Elements

Given an unlabeled egocentric video of T frames, our task is to recognize actors $a_t \in \mathcal{A}$, locations $l_t \in \mathcal{L}$, events $e_t \in \mathcal{E}$ and supporting objects $o_t \in \mathcal{O}$ in every frame $t \in [1, T]$. Our key idea is to jointly infer the story elements, which often show temporal smoothness and co-occurrence relations. Particularly, events, objects, and locations are highly correlated among each other. Detection of a specific object can be a strong clue for localization (*e.g.* if a frame includes an elephant-like object, it is likely to be taken at the location *Dumbo the Flying Elephant*). In a reverse direction, localization can significantly reduce the search space for object detection. Events and locations are also interconnected; for instance, the event *taking a boat* is highly likely at the attraction *Jungle Cruise*[4].

On the other hand, the *actor* story elements have different correlation behavior, because actors can appear anywhere at any time as they potentially share every experience with the subject. As such, the presence/absence of actors is presumably independent of all the other story elements and would not benefit from joint modeling.

Therefore, we jointly model locations, events and objects (*i.e.* all the story elements except actors) with a Markov random field as shown in Fig.2. The vertices in the graph represent story elements at each frame t, and edges indicate the dependencies among the variables. In particular, Fig.2 illustrates that each element at time t is conditionally independent of those at $t \pm 2$, given the elements of the same type at $t \pm 1$ (*i.e.* the 1st-order Markovian assumption). The model also encodes that supporting objects are conditionally independent of events given the location information.

The joint probability of the model for the egocentric video $\mathbf{V}$ is encoded by:

$$p(\boldsymbol{e}, \boldsymbol{l}, \boldsymbol{o}, \mathbf{V}) = \frac{1}{Z}\phi(\boldsymbol{e}, \boldsymbol{l}, \boldsymbol{o}, \mathbf{V})\psi(\boldsymbol{e}, \boldsymbol{l}, \boldsymbol{o}) \qquad (1)$$

where Z is the normalizing constant, $\boldsymbol{e} = \{e_t\}_{t=1}^T$, $\boldsymbol{l} = \{l_t\}_{t=1}^T$, and $\boldsymbol{o} = \{o_t\}_{t=1}^T$. The unary potential $\phi(\cdot)$ and the pairwise potential $\psi(\cdot)$ are

[4] Similar correlations exist in other domains, for example, in *Paris*, the event *taking a panoramic view* is likely to be associated with the location *Eiffel Tower*, but not with *Champs Elysees*.

$$\phi(\boldsymbol{e},\boldsymbol{l},\boldsymbol{o},\mathbf{V}) = \prod_{t=1}^{T} \phi_e(e_t,\mathbf{v}_t)\phi_l(l_t,\mathbf{v}_t)\phi_o(o_t,\mathbf{v}_t), \quad (2)$$

$$\psi(\boldsymbol{e},\boldsymbol{l},\boldsymbol{o}) = \left[\prod_{x\in\{e,l,o\}}\prod_{t=1}^{T-1}\psi_x(x_t,x_{t+1})\right]\left[\prod_{x\in\{e,o\}}\prod_{t=1}^{T}\psi_{x,l}(x_t,l_t)\right]$$

where $\mathbf{v}_t$ it the t-th frame of the video $\mathbf{V}$.

While we defer the details of the unary terms to the following sections, we here focus on the pairwise potential $\psi_x(x_t, x_{t+1})$, which models the transition probability of each story element (*i.e.* $x = \{e, l, o\}$: event, location, object). We use a Potts-like potential to encourage the elements of neighboring frames to have similar values:

$$\psi_x(x_t,x_{t+1}) = \begin{cases}\gamma_x & \text{if } x_t = x_{t+1} \\ 1 & \text{if } x_t \neq x_{t+1}\end{cases}, \quad x \in \{e,l,o\}, \quad (3)$$

where $\gamma_x \geq 1$ is a parameter to learn. A large γ_x encourages the story elements to persist at the same values for longer.

The pairwise potential $\psi_{x,l}(x_t, l_t)$ models correlations among the location and other story elements of events and supporting objects. Since the object discovery is unsupervised, we let $\psi_{o,l}(o_t, l_t) = \eta_{o_t,l_t}$, where the parameter η_{o_t,l_t} is learned by co-occurrence statistics of locations and objects in the training images. We might do the same with the pairwise potential between locations and events, if we had sufficient video training data. However, since we do not, we rely on prior information and encourage locations and event variables to take compatible values:

$$\psi_{e,l}(e_t,l_t) = \begin{cases}\nu & \text{if } e_t \text{ and } l_t \text{ are compatible} \\ 1 & \text{otherwise}\end{cases} \quad (4)$$

The value of $\nu \geq 1$ is a parameter that we learn. By sharing the parameters, we can keep the number of parameters in our model small, which helps avoid overfitting with a small number of training video instances.

Learning: We learn γ_x and ν using cross validation, and $\eta_{o,l}$ is set to a uniform prior plus the approximation of $p(l|o)$. The uniform prior ensures each location have nonzero probability of every object. The approximated $p(l|o)$ is computed by counting the fraction of detections for object o that co-occur with location l. For example, if an object $o = A$ co-occurs a half of the time with location $l = B$ and the other half with $l = C$, then $p(l = B|o = A) = p(l = C|o = A) = 0.5$.

Inference. We use loopy belief propagation [24], which does not guarantee convergence to the global optimum, but we observe that this approximate inference works well in practice for our problem.

3.1. Actor Detection

We recognize actors a_t in an unsupervised manner as follows. We first detect faces in each video frame $\mathbf{v}_t$ using Fraunhofer face detector [32], which returns bounding boxes containing faces along with confidence scores. The detector is also capable of returning estimates of age, face orientation and emotion, which can further be useful in estimating the mood of actors such as happiness and surprise. Due to a large amount of motion and other noise in egocentric videos, many false positives are unavoidable even with the state-of-the-art face detector [32]. In addition, it is challenging to distinguish true actors from accidental bystanders that may appear in the vicinity, given that Disneyland is densely populated. We observe that true actors should appear for larger and more frequently, compared to bystanders. Therefore, we retain high confidence detections that are temporally consistent (*i.e.* appear for at least 1 second in the video) and larger than 60-by-60 pixels in size.

3.2. Image-based Localization

We define unary localization potential, $\phi_l(l_t, \mathbf{v}_t)$, by recognizing in which attraction each frame $\mathbf{v}_t$ is recorded. Recall that we collect training images per location from Google/Flickr, which are used to train attraction-based location classifiers. We leverage the images instead of videos, because it is easier to obtain weakly-labeled images than videos. For example, if we query the attraction name *Mad Tea Party*, almost all top-ranked images obtained by the search engine are relevant, whereas the returned videos may include lots of noisy and irrelevant parts.

We extract dense SIFT features and use Improved Fisher Vectors (IFV) [26] for feature encoding. In order to preserve weak spatial information of each attraction, we also use the spatial histogram [19]. Each image is divided into 1×1 and 2×2 regions, and the total of 5 spatial regions are all encoded by IFV and then concatenated to obtain the final feature $f_l(\mathbf{v}_t) \in \mathbb{R}^{81,920}$. Then we train a linear SVM for each attraction, where a score of frame t belonging to the attraction l is computed by $\mathbf{w}_l f_l(\mathbf{v}_t) - b_l$, where $\mathbf{w}_l$ is a learned weight vector and b_l is a bias. Since each linear SVM is trained independently, we found it useful to calibrate the different 1-vs-all SVM scores by fitting a sigmoid function to the output of each attraction SVM [28]. The final potential is then represented by

$$\phi_l(l_t,\mathbf{v}_t) = \sigma_{l_t}\left(\mathbf{w}_{l_t} f_l(\mathbf{v}_t) - b_{l_t}\right), \quad (5)$$

where $\sigma_{l_t}(\cdot)$ is the learned sigmoid for each attraction l_t.

3.3. Event Recognition

Both visual content and motion information are excellent cues for event recognition. For example, the motion pattern of a walking person is clearly different from that of

a person who is watching a performance. In the same vein, some prior works [23, 29] leverage the camera egomotion to recognize activities. However, multiple activities may have similar motion patterns; for example, the camera motion is relatively stationary in both *watching performance* and *dining* events. Hence, visual content is appended to disambiguate such cases (*e.g.* the presence of menu, plates and silverware would strongly indicate the *dining* event).

Based on the above observation, we exploit both visual and motion features to train event classifiers. As a visual feature, we exploit Bag-of-Words (BoW) and spatial histogram to encode dense-SIFT features. As a motion feature, we extract dense optical flow [22], and quantize the flow angles and magnitudes into 16-bins. We also compute the mean and variance of both flow angles and magnitudes. The visual and motion features are then concatenated to produce a feature vector $f_e(\mathbf{v}_t) \in \mathbb{R}^{15,036}$. Similar to localization, we train 1-vs-all SVM for each event and fit a sigmoid function to obtain a probability distribution over each event, resulting in the following unary potential:

$$\phi_e(e_t, \mathbf{v}_t) = \sigma_{e_t}\left(\mathbf{w}_{e_t} f_e(\mathbf{v}_t) - b_{e_t}\right). \quad (6)$$

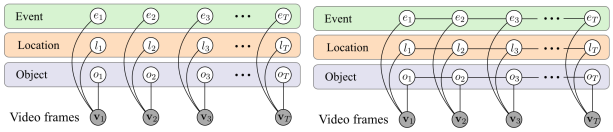
3.4. Supporting Object Detection and Recognition

One possible approach to discover supporting objects is to let a human annotator identify a list of objects of interest, and manually label them. However, this would require excessive human effort, and thus is expensive and time-consuming. Therefore, we implement an unsupervised approach that discovers supporting objects from training images of each attraction, and then trains an object detector for each discovered object. Given a novel egocentric video, we can then apply the trained object detectors to each frame to detect the most salient supporting object in that frame.

We first cluster training images from the same attraction, using two global features of GIST and RGB histograms, with Gaussian mixture models (GMM) and affinity propagation [10]. For each image cluster, we then use an algorithm similar to [38] that localizes one common object from each cluster. To ensure the quality of the discovered objects, we also manually rule out inconsistent or poor discovered objects. However, note that this is done by a binary selection on the level of discovered objects, and hence requires little effort. We then train object detectors with linear SVM for each discovered object. The final object unary potential is given in the form similar to locations and events:

$$\phi_o(o_t, \mathbf{v}_t) = \sigma_{o_t}\left(\mathbf{w}_{o_t} f_o(\mathbf{v}_t) - b_{o_t}\right). \quad (7)$$

Note that we are interested in discovering iconic objects that are dedicated and characteristic to each attraction, as opposed to generic object categories (*e.g.* ImageNet [6]). For instance, we want to detect *Dumbo* (*i.e.* an elephant-like ride vehicle) in the attraction *Dumbo the Flying Elephant*, but not generic chair in the attraction *Enchanted Tiki Room*.



(a) Frame-Based (b) Frame+HMM

Figure 3. The graphical models of the two baselines.

Although we mainly use SVM classifiers with standard image features for the story element detection, our framework is orthogonal to the choice of classifiers and descriptors, which can be replaced by any state-of-the-arts methods (*e.g.* deep learning features [11]).

4. Story-based Retrieval

Wearable cameras allow users to record the videos of everything they experience throughout a day. However, it still remains challenging to search and retrieve relevant video segments since egocentric videos are long and taken in an unconstrained environment. In this section, our goal is to perform story-based retrieval that enabled users to easily review locations and events, and find objects or people of interest, by leveraging our model of jointly inferred story elements. We use a simple AND-OR graph to describe the query syntax, which consists of combination of story elements (*e.g.* search for a subshot taken after my daughter (*actors*) danced with (*events*) Chip 'n' Dale (*supporting objects*)). In Section 5.2 we illustrate typical examples of queries written in the form of AND-OR graphs and their corresponding retrieval results.

Once each query is encoded by an AND-OR graph, we then use a sampling method to compute the marginal probability that satisfies the query. As a simple example, if we want to query subshots of event A at location B, the AND-OR graph looks like [(e_t=A) AND (l_t=B)]. We sample our Markov random field with the Metropolis-Hastings algorithm to compute $p(e_t = A, l_t = B)$. This is done by first drawing a large number of samples from the Markov random field, and then for each subshot i that is represented by e_i, l_i and o_i, we compute the proportion of number of samples that agree with the queried evidence to the total number of samples. The AND-OR graph can also encode temporal queries. For instance, if we want to search for subshots after event A at location B, then the AND-OR syntax is [($e_{t-1} = A$) AND ($l_{t-1} = B$)]. Likewise, we rank each subshot at t by $p(e_{t-1} = A, l_{t-1} = B)$. For OR syntax, we use max pooling of all possible branches. If we want to query subshots of event A at location B_1 or B_2, the AND-OR graph can be expressed as [($e_t = A$) AND (($l_t = B_1$) OR ($l_t = B_2$))], then we rank the subshots by $\max(p(e_t = A, l_t = B_1), p(e_t = A, l_t = B_2))$.

Methods	Localization		Event recognition
Metrics	Top-1	Top-5	Top-1
Chance	0.016	0.078	0.125
Frame-based	0.243	0.390	0.356
Frame+HMM	0.315	0.464	0.392
Our Method	**0.323**	**0.473**	**0.439**

Table 1. Performance of attraction-based localization and event recognition. We report top-1/top-5 attraction accuracies for localization and top-1 accuracies for event recognition.

5. Experiments

We evaluate our proposed method with a new Disneyland dataset of egocentric videos collected from YouTube (see Section 2.1). Since this is the first work that aims to predict story elements from egocentric videos, there are no existing baselines to compare against. Instead, we define the following baselines for the performance comparison of both story element recognition and story-based retrieval.

1. **Frame-Based**: predicts story elements only based on each frame without considering any temporal smoothness or joint co-occurrence inference.
2. **Frame+HMM**: models temporal continuity with an HMM for each story element separately. The main difference with our method is that our method also models the relations among different story elements.

These two baselines can be viewed as simplified variants of our model. We use the same features and the same SVM classifiers, but different model structure. Fig.3 illustrates the graphical representation for each baseline.

We perform experiments on the two key tasks that require building and using the storyline representation: 1) joint inference of story elements, and 2) story-based egocentric video retrieval. It is challenging to quantitatively evaluate the storyline representation itself, because it is a form of mid-level data structure that is designed to facilitate visualization and other high-level tasks. Therefore, we evaluate the estimation accuracies of story elements and the performance of story-based retrieval instead.

5.1. Results of Joint Estimation of Story Elements

We quantitatively evaluate the accuracies of our joint inference model for the location and event recognition. To obtain groundtruth, we let human labelers, with expert knowledge of *Disneyland*, annotate the locations and events at every 10 frames of test egocentric videos. We report the recognition rate, which is a fraction of predicted labels that match the groundtruth labels. Table 1 shows the prediction accuracies for both location and event recognition. For attraction-based localization, we report top-1 and top-5 accuracies. Our method outperforms all the baselines for the both tasks. Among baselines, the frame+HMM baseline leads a better performance than purely frame-based variant, which demonstrates that modeling temporal smoothness benefits

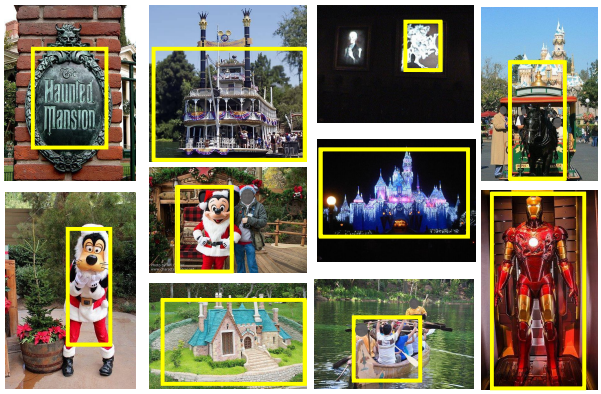


Figure 4. Examples of discovered supporting objects that occur frequently in individual attractions.

location and event recognition in egocentric videos. Our method is better than the frame+HMM baseline, showing the value of joint inference of story elements. In particular, our method significantly improves event recognition performance; we attribute this to strong correlations that exist between events and locations.

Fig.4 illustrates some examples of supporting object detections. Notice that our algorithm is able to discover different types of objects that occur frequently in individual attractions, including buildings, transportation vehicles, and Disney characters. However, due to large image appearance variations, which comes from the nature of theme park images, we also obtain lots of inconsistent object clusters.

Fig.5 shows an example of storyline representation of an input egocentric video, which displays four story elements on a timeline. Each instance of story elements is represented by a horizontal bar with a unique color. The classes of the same story element (*e.g.* *watching* and *walking*) are represented with similar but different colors. The story-based visualization can help summarize egocentric videos better and provide users with an overview of various correlations between story elements. Fig.5 clearly shows that objects, locations and events often co-occur; for example, we can expect to find the supporting object *birds* in the location *Enchanted Tiki Room*.

5.2. Results of Story-based Retrieval

We also evaluate the performance of semantic story-based retrieval. All egocentric videos in the test set are segmented into subshots of 2-second long each, which constitute the database for retrieval. To respond to a query such as *what objects did I see at the Enchanted Tiki Room*, we first retrieve subshots which were taking place at the *Enchanted Tiki Room*, then return detected objects from the retrieved subshots. In order to evaluate the retrieval, we use the average normalized rank of relevant subshots as used in the *video google* [36],

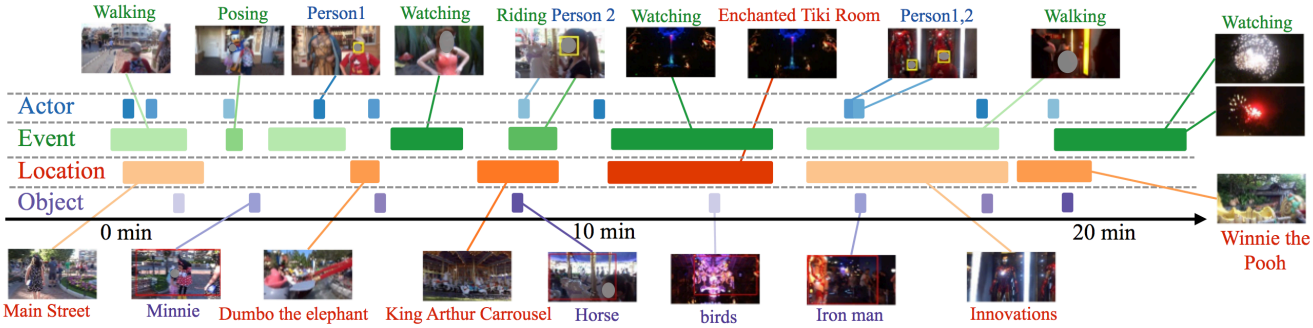


Figure 5. Storyline visualization for an egocentric video. We present four story elements and their co-occurrences on a timeline. We show some selected frames with inferred labels for the detection of story elements.

Method	Event	Location	Event & Location
Frame-based	0.179	0.266	0.217
Frame+HMM	0.176	0.228	0.201
Our Method	**0.170**	**0.217**	**0.192**

Table 2. Evaluation of retrieval performance with average normalized ranks of relevant images in Eq.(8). Lower is better.

$$\widetilde{r} = \frac{1}{NN_{rel}}\left(\sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}(N_{rel}+1)}{2}\right) \quad (8)$$

where N_{rel} is the number of relevant subshots for a story-based query, N is the number of the total subshots, and R_i is the rank of the i-th relevant subshots. The value of $\widetilde{r}$ is from 0 to 1 and 0.5 corresponds to a random retrieval. A lower rank value is better.

For test, we experiment on three different types of queries: query by location only (*i.e.* find subshots that occur in a particular place), query by event only (*i.e.* find subshots of a particular event), and query by both location and event (*i.e.* find subshots of a particular event that happens in a particular place). Queries are automatically generated from groundtruth labels for all possible locations, events and combinations of both. We report the retrieval performance in Table 2, which clearly show that our method outperforms both baselines in all cases.

Fig.6 shows two retrieval examples with queries and their corresponding AND-OR graphs. For the first query, our system searches for objects at the *Enchanted Tiki Room* while the subject does *watching performance*, and returns the top-three object detections. Since the database contains segmented subshots of 2 seconds of egocentric videos, the top-ranked results often originate from the same video. Our algorithm can find birds during the show at the *Enchanted Tiki Room*, though the localization of bounding boxes may be sometimes inaccurate due to weak object proposals. The second query can successfully retrieve the girl riding the merry-go-round at the *King Arthur Carrousel*.

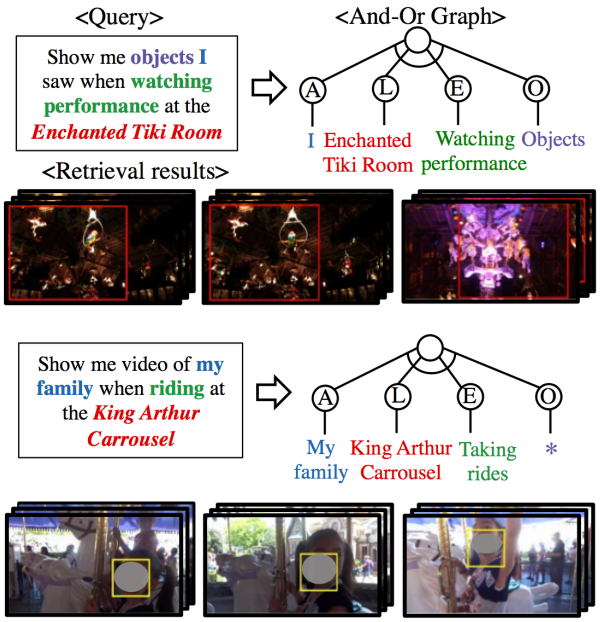


Figure 6. Two examples of retrieval results with queries and their corresponding AND-OR graphs. We show the frames from top-three ranked subshots for the queries.

6. Conclusion

In order to provide users with concise but meaningful visualization, our work explicitly defines story elements and proposes to use a new storyline representation to summarize a large set of unstructured egocentric videos. We also propose a novel joint inference method to jointly recognize all story elements. Through experiments on YouTube egocentric videos, we show that our approach outperforms alternative baselines. We also demonstrate that story-based retrieval indeed helps users easily search and organize unstructured egocentric video content.

Acknowledgements. Gunhee Kim is partially supported by Basic Science Research Program through National Research Foundation of Korea (2015R1C1A1A02036562).

References

[1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty Detection from an Ego-Centric Perspective. In *CVPR*, 2011. 2

[2] H. Altwaijry, M. Moghimi, and S. Belongie. Recognizing Locations with Google Glass: A Case Study. In *WACV*, 2014. 2

[3] I. Arev, H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. In *ACM SIGGRAPH*, 2014. 1

[4] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *ICCV*, 2013. 2

[5] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 3

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 6

[7] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *CVPR*, 2011. 1, 2

[8] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social Interactions: A First-Person Perspective. In *CVPR*, 2012. 1, 2, 3

[9] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 1

[10] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972—976, 2007. 6

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*, 2014. 6

[12] D. B. Goldman, B. Curless, D. Salesin, and S. M. Seitz. Schematic Storyboarding for Video Visualization and Editing. In *SIGGRAPH*, 2006. 2

[13] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009. 2

[14] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008. 2

[15] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013. 3

[16] G. Kim and L. Sigal. Discovering collective narratives of theme parks from large collections of visitors' photo streams. In *KDD*, 2015. 2

[17] G. Kim, L. Sigal, and E. P. Xing. Joint Summarization of Large-scale Collections of Web Images and Videos for Storyline Reconstruction. In *CVPR*, 2014. 2

[18] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 1

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 5

[20] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering Important People and Objects for Egocentric Video Summarization. In *CVPR*, 2012. 1, 2

[21] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013. 3

[22] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, MIT, 2009. 6

[23] Z. Lu and K. Grauman. Story-Driven Summarization for Egocentric Video. In *CVPR*, 2013. 1, 2, 6

[24] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, 1999. 5

[25] H. S. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. In *NIPS*, 2012. 1

[26] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 5

[27] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 1

[28] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, 1999. 5

[29] Y. Poleg, C. Arora, and S. Peleg. Temporal Segmentation of Egocentric Videos. In *CVPR*, 2014. 2, 6

[30] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010. 2

[31] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *CVPRW*, 2009. 1

[32] T. Ruf, A. Ernst, and C. Küblbeck. Face Detection with the Sophisticated High-speed Object Recognition Engine (SHORE). *Microelectronic Systems*, 2011. 4, 5

[33] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *ICCV*, 2011. 2

[34] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007. 2

[35] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 3

[36] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 7

[37] I. Swartjes and M. Theune. A Fabula Model for Emergent Narrative. In *TIDSE*, 2006. 1

[38] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014. 6

[39] M. Tapaswi, M. Bauml, and R. Stiefelhagen. Story-based Video Retrieval in TV series using Plot Synopses. In *ICMR*, 2014. 2

[40] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. StoryGraphs: Visualizing Character Interactions as a Timeline. In *CVPR*, 2014. 2

[41] K. Tu, M. Meng, M. W. Lee, T. Choi, and S. Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70, 2014. 2

[42] M. Wang, R. Hong, X.-T. Yuan, S. Yan, and T.-S. Chua. Movie2Comics: Towards a Lively Video Content Presentation. *IEEE T. Multimedia*, 14(3):858–870, 2012. 2

[43] B. Xiong and K. Grauman. Detecting Snap Points in Egocentric Video with a Web Photo Prior. In *ECCV*, 2014. 2

[44] Y. L. Zhefan Ye, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg. Detecting eye contact using wearable eye-tracking glasses. In *PETMEI*, 2012. 1