Master Thesis

Learning Storylines of First-Person Videos with Gaze

(視線を用いた一人称視点映像の
ストーリーラインの学習)

Binhua Zuo

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
The University of Tokyo

July 2018

Abstract

With the widespread access to photo-taking devices and rampant social networking, users are motivated to generate plenty of images or video data. Among this, egcentric videos have become pretty common and popular in our daily lives recently. Since these videos may last even hours, there is a growing demand for representing the video contents in an efficient yet comprehensive way (i.e., a visual image storyline or some causal video subshots). The most related topic, video summarization, has been wildly studied from so many perspectives till now. Most of them try to select some subshots or certain keyframes to generate a visual summary of the most attractive or enlightening parts from the video. However, egocentric videos are usually fixed in a certain area and containing much redundant information, so we have not seen satisfactory performance by applying the common or traditional techniques for video summarization.

Considering the limits for the current video summarization methods, this thesis focuses on the problem of learning the underlying story and temporal dynamics for egocentric videos in an efficient way. In this thesis, we propose a new approach for learning storyline which focuses on gaze region to sample frames for the training of Recurrent Neural Network. Here, we first extract the gaze region to track with important objects for the current action or manipulation of the wearer, by eliminating irrelevant objects in the scene. Then we utilize DPP (a diversity-based sampling method), to help us get better long-term relationship learning among different actions, since we find that egocentric videos usually contain so many repetitive and long-lasting actions.

Finally, we perform two seperate experiments, evaluating storyline and storyline prediction. For the first one, we address it as a video summarization problem, where F-score (temporal overlap between generated and ground truth storyline) is applied as the evaluation metric. The result shows that our method gets the highest F-score of 45.7, compared with other baselines. Then we perform the task of challenging semantic forecasting in storylines, which is to predict the image which represents the next event from the storyline. And our method is able to forecast the next representative event in the storyline with accuracy of 33.5%, much higher than all the other baselines. Both experiments show that our method is capable of generating a storyline with better diversity and longer temporal relationship learning compared with other baselines. We also showed how the gaze region and DPP help us to generate a better storyline in this part.

Contents

Chapter 1

Introduction

1.1 Background

Recently, the widespread access to photo-taking devices and rampant social networking has posed a lot of new problems challenges in the field of computer vision. One among such problem is the overloading of information, for there being too much data available for users, which are full of redundant information and need to be integrated or refined for later use. Plenty of the data, however, containing so many long-lastings, unnecessary and raw contents, for instance, people may take so many similar pictures for a specific object or scenery. Yet the important or attractive information may be possibly ignored or unseen among these redundant sections. Thus, camera users are often overwhelmed by the long lifelogging egocentric videos, and attempting to locate and find those significant or attractive sections, or to clutch those important events or actions, gaining a underlying and comprehensive story for such videos. Hence it is becoming increasingly important to automatically summarize egocentric videos in a more efficient but comprehensive way. The key idea of the solution would be to shrink such redundant parts but remaining anomalies or some interesting segments to the camera user.

From our explanation above, we can see that it's an ill-posed problem to summarize the egocentric video. Actually, since egocentric devices are usually mounted on the head of the user, these egocentric videos are prone to contain some blurry and shaky segments due to the chest or head movement. Besides, egocentric videos are usually taken without a specific topic or concept, so the appropriate structure may be missing for the intention of the video. The user may wear these egocentric devices all the time when they are having fun or enjoying some cool sports or experiments. In such case, most of the recorded videos or images are so proned to be irrelevant or repetitive [24]. Nonetheless, the proliferation of wearable photo-taking devices will only increase, and it is necessary for these systems, which take long egocentric videos, to represent the video information in an efficient yet comprehensive way. They should not only give device users the power to store and concisely view their daily life activities, but also the ablility to look up the important or atractives parts in the future. There have seen so many different and attractive approaches for this problem. For example, [8] produced a concise visual summary that encompasses the key segments of a video, able to locate those important people or significant objects for egocentric video summarization. More recently, [33] made the first effort to made the first attempt to consider the use of gaze to summarize the egocentric

videos, through their results, they showed that the gaze position represents the attractive or personalized information that is interested for the user. So it's necessary and also effective to combine the gaze information to personalize the long sequecne of egocentric videos to get a summary that is more related the camera wearer. This method, however, is supervised by using set of training videos and manually generated summaries, which is tedious and requiring a lot of humman resources.

So given an egocentric video, our goal is to learn the inherent story, namely storyline here. A storyline, which is a subset of video frames containing certain images, normally denotes a sequence of activities or events, which have temporal or causative relations. A sotyline usually give us a brief and effective understanding on contents of the video. As shown in Figure 1.1, this is a storyline containing key steps for making pizza, like preparation, cut mushroom, pepper and hot dog, and so on. Through this storyline, we can have a rough idea of making pizza.

Figure 1.1: A storyline for making pizza.

1.2 Challenges

Since the storyline is such a subjective item, and it's usage varys a lot for different camera users. Thus it's very important to personalize the storyline for distinct users. So the first challenge lies on how to accomplish the personalization for a specific user. As in [33], they showed that the egocentric gaze information is a key point to generate a personalized storyline. Actually, the way a person looks upon the world and what object the user is focused on during the whole event or task in the egocentric video reflects so many clues for his interest or intent, and also showing better causal relationships between objects or person in the video. In the study[35], they found that the gaze motion presents a better visual comprehension, and help to understand the concept of the video in a different perspective. For example, we can find the temporal and spatial distribution for the interest and attention of a person, by looking the relativce importance of different objects or subject in the video frame. Besides, some egocentric videos are usually constrained in thr fixed area or even a specific scene, so some frames in these videos may be very "similar". As we can see in Figure 1.2, these three frames from "making pizza" are pretty similar, for

they share almost the same backgroud, but actually they are three different and necessary steps for making pizza.

Figure 1.2: Some key steps in making pizza, these frames are pretty similar, for that they share almost the same backgroud.

Also, for egocentric video, the background is usually clustered, making it hard to track the important objects for current action. Thus, follow the work in [33], we also believe that gaze information is a fundamental missing component in learning the storyline for egocentric videos – with the help of gaze, we can offer better instructions on those contents or sections which might be attractive to the user, thus eliminating the unrelavant or useless parts. In Figure 1.3, we show A comparision for the original frame and the gaze region (256*256). We can find that gaze region provides more specific information for the current action or manipulation, by eliminating irrelevant objects.

Figure 1.3: A comparision for the original frame and the gaze region. Gaze region provides more specific information for the current action or manipulation.

The other chanlenge for learning storyline for egocentric video is that there are so many repetitive and long-lasting actions, as shown in the Figure 1.4, some actions like cutting mushroom last about 36 seconds, which is very long, considering that for some datasets of action recognition, actions usually have less than 15 seconds. For the current video summarization methods, especially those seqential learning method using RNN, they're proned to be trapped in learning short term relations among the dominant actions, causing it very difficult to learn long-term relationship among different actions compared to the normal video.

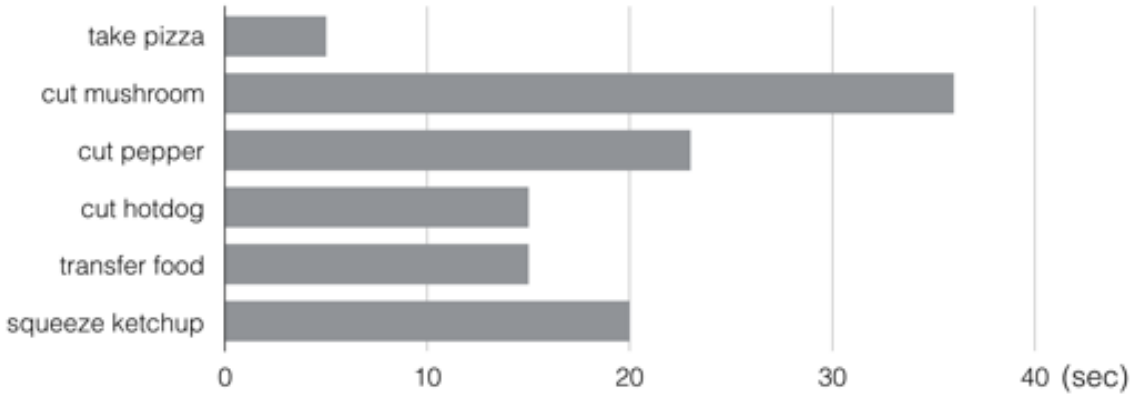

Figure 1.4: Time period distribution for some important actions of making pizza in GTEA Gaze+ dataset.

Chapter 2

Related Works

In this part, we give a brief introduction for some important works that related to our research, consists of learning storyline, video summarization and gaze in egocentric videos.

2.1 Learning storyline

Storyline is a subset of video frames containing certain images, normally denotes a sequence of activities or events, which have temporal or causative relations. A sotyline usually give us a brief and eective understanding on contents of a text, image album or video. The storyline was first used in the field of question answering and text summarization in 1970s. For example, in the work of [28], they generated scripts for the task of text summarization, which is an organized representations of some causal relations or temporal events. Then in [11], Bobick and Yuri proposed to a method which first utilized the statistical techniques to discover the primary components of an event or activity, and also recognize the structure of the sequence simultaneously. However, these methods either based on scripts (an organized representations of some causal relations or temporal events) or stochastic grammar, both require extra human efforts or annotation of videos. So these methods can only be applied to a limited domain. Due to the limitations for those traditional methods, there have arisen so many novel approaches, which are able to learn the underlying storyline and temporal dynamics from an image album or videos automatically. In [16], Gunhee and Eric proposed an automatic storyline-learning method, they formulate this problem as referring a time-varing sparse oriented pictures. Recently, Recurrent neural networks [5] becomes very popular in the field of language modeling [23] and computer vision [34, 40]. Seqeuntial learning is the key idea for RNNs, which disintegrates the probability of a sequence (e.g. image ablum or video frames) into the prediction for the next element from the sequence based on the previous information stored in RNN [14, 3]. So Gunnar *et al* [30] extended this idea to learning the temporal dynamics and the underlying story by introducing Skipping Recurrent Neural Network model, based and refined on RNN. In **??**, it's a visualized storyline for the concept Paris. Given a concept, [30] is able to simultaneously learn the temporal relationships and visual storylines from the album in web.

Figure 2.1: Given a concept, [30] is able to simultaneously learn the temporal relationships and visual storylines from the album. This shows a visualized storyline for the concept Paris. They use arrowed lines to represent the most frequent transitions between different images nodes. The right are three possible storylines (A,B,C) for Paris, all of them containing 10 images.

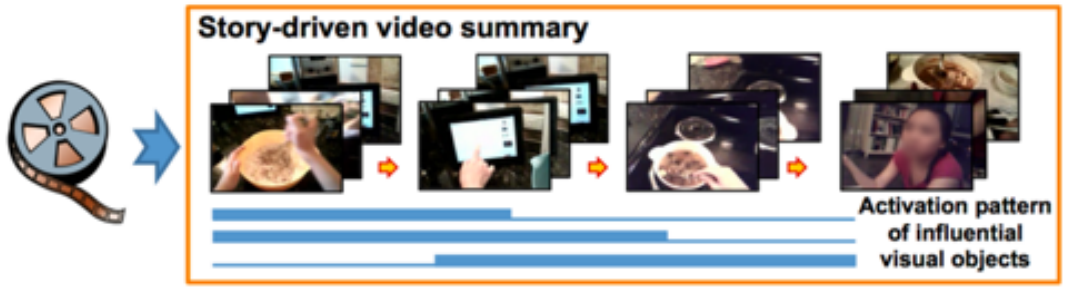

Figure 2.2: [21] creates a summary from an unedited egocentric video. A good storyline is one consists a chain of subshots, which have causal relationships among them.

2.2 Video Summarization

Recently, video summarization is a topic of interest for many researchers, and has been wildly studied and analyzed from many different views [25, 39]. Most of them summary the video by selecting a subset of video subshots or some keyframes to comprise the most attractive and representation parts in the video. Many former approaches have been designed to seek cues from low-level motion and appearances [2, 26]. Then, some new methods focused on the supervised learning of the important objects [15, 20], multiview [7] or user interactions [4], to locate the attractive or important parts in the video. Also, researchers begun to take other external efactors, beside with the story structure into consideration. For instance, in [31] they proposed three different criteria: quality, diversity, and coverage for the summarization task. And later the social relationship, like the aesthetics and characters, is also considered into the framework for summarizing task in [27]. The work in [21] is most similar with our job, which is also trying to learn a underlying story for egocentric videos. In their work, they summary scene of story by selected short subshorts from video, as shown in the Figure 2.2, A good storyline is one consists a chain of subshots, which have causal relationships among them. But first they need to split the video based on the camera movement, which, sometimes cannot seperate the action correctly. Besides, their work relies a lot on the object detection and object coocurrence, also tends to choose a series of video subshots which connectly or mutually influenced, probabaly leads to the missing of some other seperate but also important sections. More recently, in [22], they proposed a new unsupervised video summarization method, by using LSTM [12] to perform a generative adversarial [10] learning. As shown in the Figure 2.3, they use GAN learning to select some key frames, which comprise a similar distribution with the original video. However, they have a main Hypothesis for this work, which is the learned composition for summary video and the input video supposed to be similar. Then it's not suitable for videos with great diversity, (e.g. egocentric video), because if a video is too diverse, containing too many different scenes, it's hard to use limited frames to reconstruct the original video. Considering the video summarization task will be much easier when we have a specific task, we'd like to explore how to summarize the video in a both intelligent and efficient way for some specific task.

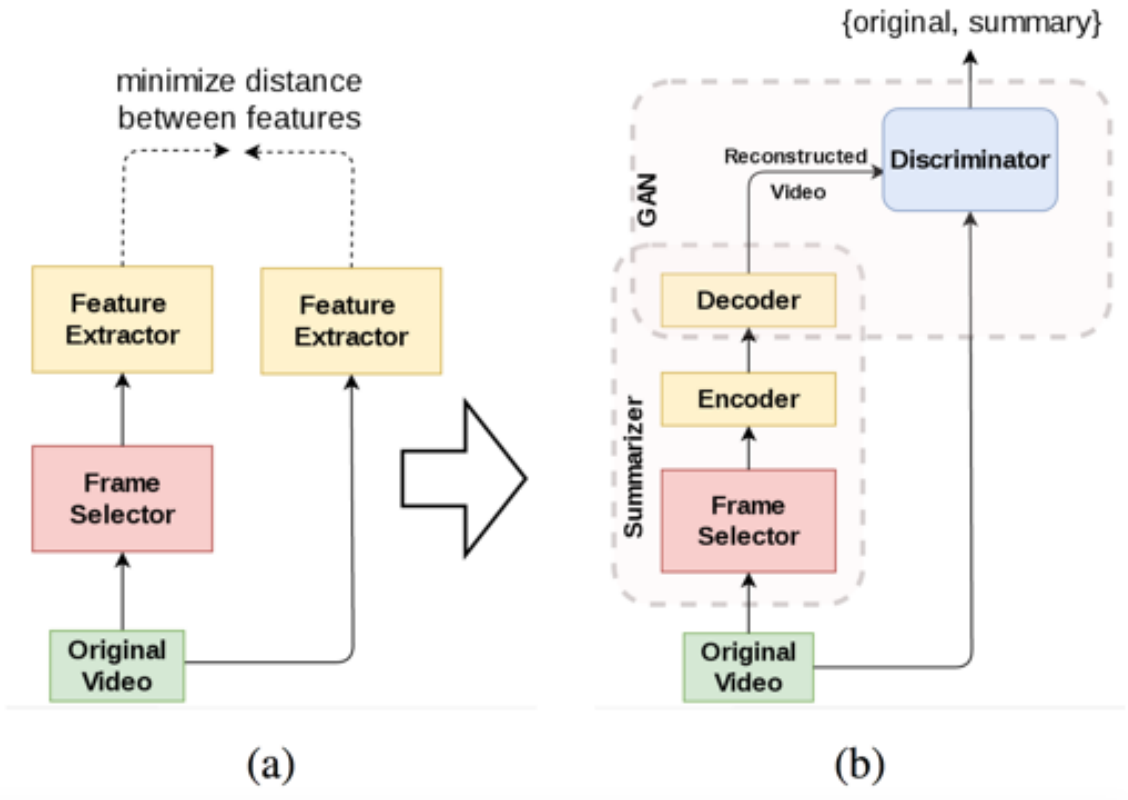

Figure 2.3: [22] use GAN learning to select some key frames, which comprise a similar distribution with the original video.

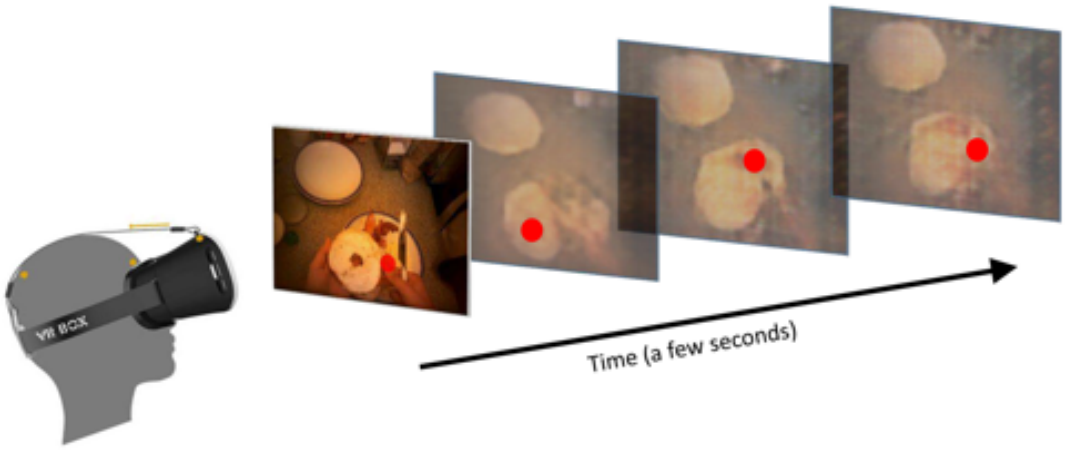

Figure 2.4: Given the current egocentric video frame, [39] can preidct gaze positions for some future frames.

2.3 Gaze in egocentric videos

Gaze is a very important information for egocentric videos, for it usually tells the attention of the wearer, which is one of the human visual system [35]. There has been so many works to utilize the gaze in many traditional computer vision tasks, like action localization [29] and action recognization [6]. For the gaze region shows some clues for the important objects or the intention of the wearer. For egocentric video, the background is usually clustered, making it hard to track the important objects for current action. So in [33], they found that gaze information is a fundamental missing component in learning the storyline for egocentric videos – with the help of gaze, they can offer better instructions on those contents or sections which might be attractive to the user, thus eliminating the unrelavant or useless parts. Also, Considering the current gaze sensors are still expensive and power consuming, some later work decided to predict the gaze position in the egocentric video. In [13], as showin in Figure 2.4, they use low-level features to learn the saliency model directly from human eye movement data. Most recently, Zhang et al. [39] proposed a new GAN model automatically learn the important egocentric clues within the training process, by utilizing two 3D-CNN streams to disentangle foreground and the background motions.

Chapter 3

Proposed Method

In this part, we introduce our proposed approach for learning storyline which focuses on gaze region to sample frames for the training of Recurrent Neural Network. Here, we first extract the gaze region to track with important objects for the current action or manipulation of the wearer, by eliminating irrelevant objects in the scene. Then we utilize DPP (a diversity-based sampling method), to help us get better long-term relationship learning among different actions, since we find that egocentric videos usually contain so many repetitive and long-lasting actions.

3.1 Learning visual storylines

So given an egocentric video, our goal is to learn the inherent story, namely storyline here. A storyline, which is a subset of video frames containing certain images, normally denotes a sequence of activities or events, which have temporal or causative relations. A sotyline usually give us a brief and effective understanding on contents of the video. As shown in Figure 1.1, through a storyline for making pizza, we can have a briefly idea about the key steps for making a pizza, just like a recipe.

In the following sections, we explain our method based on Recurrent Neural Network, a type of neural networks capable of learning sequential transitions, which is trained over all ordered frames from egocentric videos. As shown in Figure 3.1, we first extract gaze region from original video frames, which helps us focus on important objects. Then, in order to address the repetition issue, we utilize DPP (a diversity-based sampling method), to help us get better long-term relationship learning among different actions. In the following sections, we will first give a brief introduction for the basic knowledge about RNN and DPP, and then show that the our method can be accomplished with DPP initialization to the original RNN.

3.2 Recurrent neural network

Recurrent neural networks [5] are popular models which have seen so many promising and attractive applications in the field of language modeling [23] and computer vision. Seqeuntial learning is the key idea for RNNs, which disintegrates the probability of a sequence (e.g. image ablum or video frames) into the prediction for the next element from the sequence based on the previous information stored in RNN.

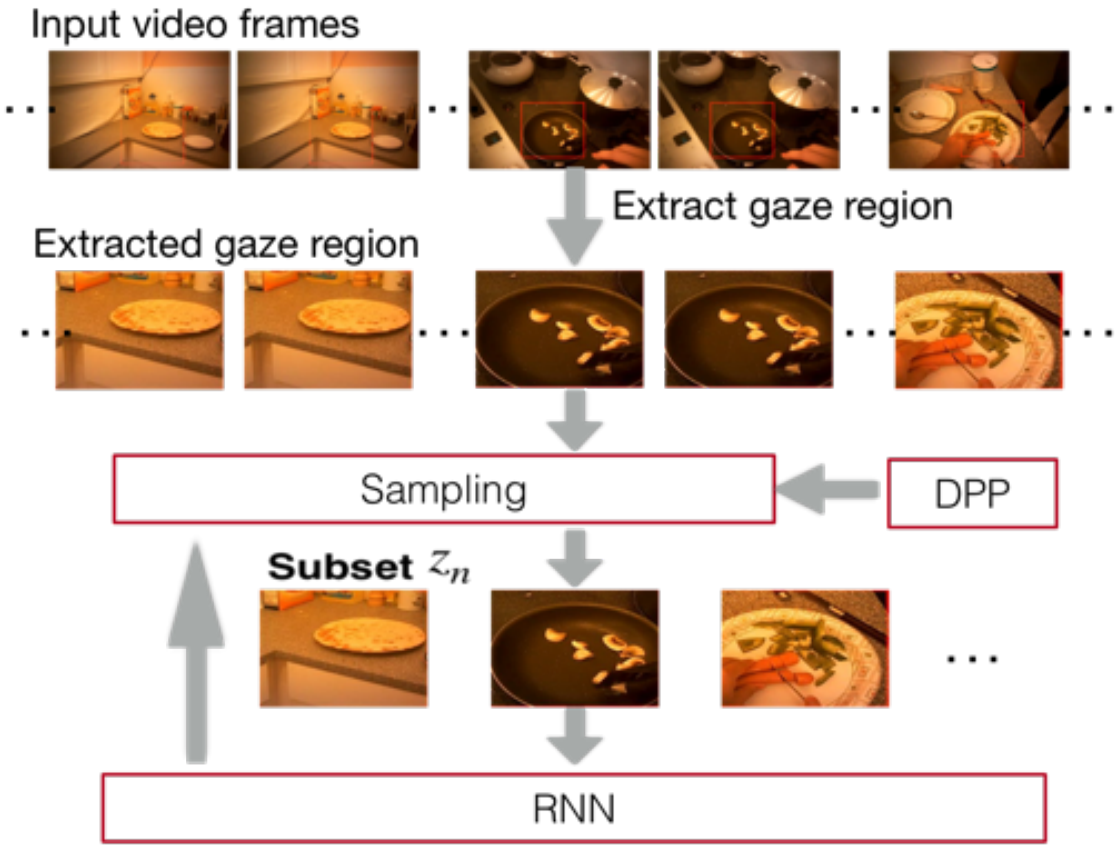

Figure 3.1: Overview of our method. We first extract gaze region from original video frames, which helps us focus on important objects. Then, in order to address the repetition issue, we utilize DPP (a diversity-based sampling method), to help us get better long-term relationship learning among different actions.

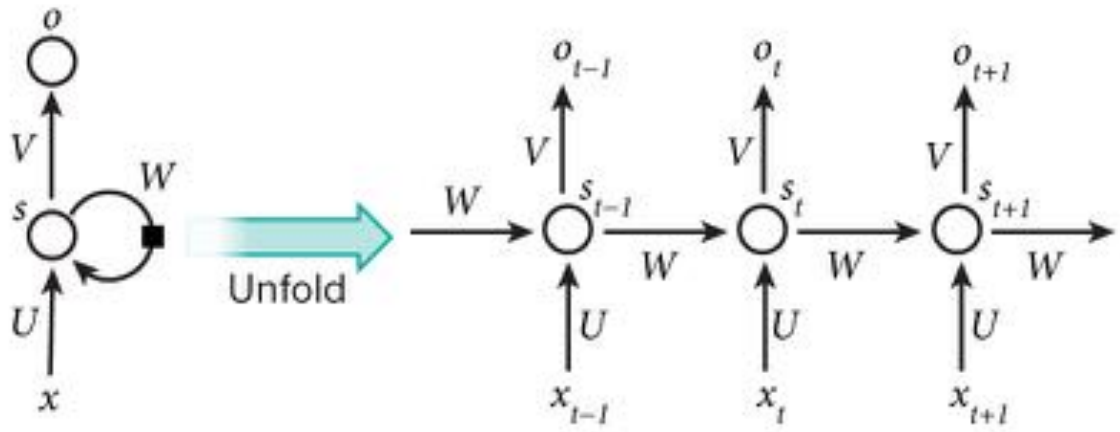

Figure 3.2: A model for RNN, it contains three different layers, including the input layer x_t, the hidden layer s_t and the output layer o_t.

As for using RNN for storyline construction, we suppose a given sequence of T images $\mathbf{x}_{1:T} = x_1, \ldots, x_T$, then the training of RNN is to maximize the likelihood:

$$\begin{aligned} \mathcal{M}^* &= \arg\max_{\mathcal{M}} \log P(\mathbf{x}_{1:T}; \mathcal{M}) - \lambda \mathcal{R}(\mathcal{M}) \\ &\textit{where } \log P(\mathbf{x}_{1:T}; \mathcal{M}) = \sum_t \log P(x_{t+1} | \mathbf{x}_{1:t}; \mathcal{M}). \end{aligned} \tag{3.1}$$

$\mathcal{M}$ here is integrated model parameters for RNN, $\mathcal{R}(.)$ is the regularizer, for example ℓ_2 or ℓ_1. Here, $P(\mathbf{x}_{1:T}; \mathcal{M})$ is the probability of the current sequence. The $P(x_{t+1} | \mathbf{x}_{1:t}; \mathcal{M})$ is a probability depended on the task, for our work of summarization, it could be the output (probability of being selected for the storyline) for the next image x_{t+1}. For optimization of the RNN, the common algorithm is to use Back Propagation Through Time (BPTT) [32], a mathmatical approach for caculating the gradients, which is accumulated along with time seqeunces (e.g. image album).

The RNN model contains three different layers, including the input layer x_t, the hidden layer s_t and the output layer o_t as shown in Figure 3.2.

x_t is the input at time step t, it could be a image or a word depending on the task. The input layer uses it to update the hidden layer s_t with the weights W. And s_t is the reccurent (hidden) layer at time step t, this hidden layer s_t updates itself with the current input x_t and also the previous hidden layer s_{t-1}: $s_t = f(Ux_t + Ws_{t-1})$, and also predicts the output for the current layer $o_t = \sigma(Vs_t)$. Here the function $f()$ and $\sigma()$ are usually a nonlinearity, e.g. tanh, ReLU, softmax or sigmoid, etc. And the output at current time step o_t, for our work of summarization, since we want to predict the next image for our storyline, it could be a probability vector across all the future images from the original image album.

Theoretically, RNN is able to employ the sequential information in any long sequences, but in practice, because of the limited memory capacity of the hidden layer s_t and the vanishing gradients [1], thus we can not back propagate the error efficiently. And this is a significant problem for learning the temporal relationship in some image albums or video frames, because such albums may contain so many similar continous images or frames, so RNN tends to be trapped in learning these

repetitive short term relationships. For example, for an image album containing some specific concepts, many pictures of the same object or scenery may taken within a really short period, and such picitures may look very similar. Also, for our egocentric videos, there may be so many repetitive and long-lasting actions, as shown in the Figure 1.4, some actions like cutting mushroom even last for 36 seconds, which is very long, considering that for some datasets of action recognition, actions usually have less than 15 seconds. So the underlying storylines may be suppressed if we apply RNN to these albums directly, which has such a salient pattern, causing it very difficult to learn long-term relationship among different actions compared to the normal video. One way to solve this problem is to regularize RNN with a diversity term [31], as to learn the long-term relationship, however, it doesn't work well for a single-themed album, since we still need those images with similar visual contents in our storyline.

3.3 Refined RNN

Follow the skipping idea from [30], which is a skipping recurrent neural model, built upon the RNN framework, was proposed a new diversity-based RNN to learn a longterm relationship and the underlying story for a specific task. For the original RNN, the goal of the network is to learn the sequential transition among every item in the sequence, in the work of [30], however, they incorporate the latent valuables z_n, to skip through those repetitive parts in the image album, thus learning a long-term relationship among different activities and the underlying stories in the image album. The key idea is to predict the most possible image, which represent the next important actions or events in the storyline, and then use these selected images to form the transition learning of RNN.

Here, we first crop the gaze region from the video frames as input image sequences. Suppose $\mathbf{x}_{1:T}$ represents the T gaze-centered images in the sequence, $\mathbf{z}_{1:N}$ is the indexes subset which denotes the selected images (representative actions or events in the storyline) from the photo album or video frames. And N is number of images or frames consisted in this storyline, normally, we use 10 - 15 images to construst a storyline. To note here, $N \ll T, z_n \in 1, 2, \ldots, T$, and $\mathbf{z}$ is a subset consists of ordered indexes. The goal of our work is to maximize the marginal probability over the whole album as to learn the optimized likelihood model parameters ($\mathcal{M}$):

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} \log \sum_{\mathbf{z}_{1:N}} P(\mathbf{x}_{1:T}, \mathbf{z}_{1:N}; \mathcal{M}) - \lambda \mathcal{R}(\mathcal{M}) \tag{3.2}$$

Here $\mathcal{R}(.)$ is the regularizer, for example we can use ℓ_2 or ℓ_1. Then by factorizing $P(\mathbf{x}_{1:T}, \mathbf{z}_{1:N}; \mathcal{M})$ as $P(\mathbf{x}_{1:T}|\mathbf{z}_{1:N}; \mathcal{M})P(\mathbf{z}_{1:N})$, here $P(z_{1:N})$ is the prior on z, which is a subset consists of ordered indexes and does not depend on $\mathcal{M}$, and assuming that the probability of the whole image sequence is proportional to the probability of the selected subset of images $\mathbf{x}_\mathbf{z}$ (which means $P(\mathbf{x}_{1:T}|\mathbf{z}; \mathcal{M}) \propto P(\mathbf{x}_z; \mathcal{M})$), then by inincorporating our assumption into Equation 3.2:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} \log \sum_{\mathbf{z}_{1:N}} \big(\prod_n P(\mathbf{x}_{z_{n+1}}|\mathbf{x}_{\mathbf{z}_{1:n}}; \mathcal{M}) \big) P(\mathbf{z}_{1:N}) - \lambda \mathcal{R}(\mathcal{M}) \tag{3.3}$$

We can see that the Equation 3.3 looks very similar to the original RNN in Equation 3.1 by the usage of the chain rule.

Maximizing the Objective. Considering that it is computationally intractable to maximize the marginal probability over all subsets textbfz from the image album. So we decide to utilize the Expectation Maximization (EM) algorithm to help us solve the Equation 3.3 more efficiently. Since our objective is to maximize the likelihood of $x_{1:T}$ over all possible subsets $z_{1:N}$:

$$\begin{aligned}\max_{\mathcal{M}}(P(x_{1:T};\mathcal{M})) &= \max_{\mathcal{M}}\sum_{\mathbf{z}_{1:N}} P(\mathbf{x}_{1:T}|z_{1:N};\mathcal{M})P(\mathbf{z}_{1:N}) \\ &= \max_{\mathcal{M}}\sum_{\mathbf{z}_{1:N}} P(\mathbf{x}_{z_{1:N}}|z_{1:N};\mathcal{M})P(\mathbf{z}_{1:N})\end{aligned} \tag{3.4}$$

Here we presume the prior $P(z_{1:N})$ does not depend on the modal parameters $\mathcal{M}$, and Equation 3.4 can be solved with EM algorithm directly. IN E-step, we sample a subset indexes $\mathbf{z}$ to estimate the expectation:

$$Q(\mathcal{M};\mathcal{M}_\prime) := E_{\hat{z}-q_0}[log(P(\mathbf{x}_{\hat{\mathbf{z}}}|\mathbf{z}_{1:N}=\hat{\mathbf{z}}_{1:N};\mathcal{M})P(\mathbf{z}_{1:N}=\hat{\mathbf{z}}_{1:N})], \tag{3.5}$$

here q_0 is $P(z|x;\mathcal{M}_0)$. Then for one simple sampled $\hat{z}_{1:N}$, the M-step is as follows:

$$\max_{\mathcal{M}} Q(\mathcal{M};\mathcal{M}_\prime) = \max_{\mathcal{M}} log\ P(\mathbf{x}_{\hat{\mathbf{z}}_{1:N}};\mathcal{M}). \quad = \max_{\mathcal{M}}\sum_n log\ P(x_{\hat{z}_{n+1}}|\mathbf{x}_{\hat{\mathbf{z}}_{1:N}};\mathcal{M}). \tag{3.6}$$

So this falls to the original RNN objective over a subset rather than the whole album. In conclusion, the traning process simply samples from $P(\mathbf{z}|\mathbf{x};\mathcal{M}_\prime)$ (E-step in Equation 3.5), and then update $\mathcal{M}$ using Equation 3.6 (M-step). So we just add a sampling step before putting new samples into the RNN.

Softmax Loss. Through the previous introduction, we have already known the way to optimize the objective, the only thing left unsolved is the loss $P(\mathbf{x}_{z_{n+1}}|\mathbf{x}_{z_{1:n}};\mathcal{M})$ (data probability in Equation 3.3). Considering the amount of the possible future images is finite, they are just images after x_{z_n}, denoted as χ_n here, so we apply the softmax over all future images as our loss function:

$$P(\mathbf{x}_{z_{n+1}}|\mathbf{x}_{\mathbf{z}_{1:n}};\mathcal{M}) = \frac{exp(\mathbf{y}_n^T x_{z_{n+1}})}{\sum_{x\in\mathcal{X}_n} exp(\mathbf{y}_n^T x)} \tag{3.7}$$

where $\mathbf{y}_n$ is the output from the network at step n. The Equation 3.7 reprensents the probability for a future image $\mathbf{x}_{z_{n+1}}$ to be selected for our storyline. Actually, this models the negative world as "other feasible options" rather than "anything but the ground truth".

3.4 Diversity via DPP

To futher improve the diversity of the storyline, we utilize the Determinantal point processes (DPPs) [18] mothod to initialize the seleted subset z at first, then use this subset to initialze our network. DPPs, which first occured in random matrix theory and quantum physics, are elegant probabilistic models of repulsion. And recently, some researches, like [9, 37], have employed DPP to some summarization methods, and showing a good result. Also, as we should know, the diversity is measured on a subset of selected or sampled images, rather on independent or sequential images.

Determinantal point processes (DPP)

Here, we give a brief definition of DPP. Suppose we have a gound set $\mathbf{Z}$ of T items, for example it could be all extracted frames from the video. Also we have a T*T kernel matrix $\mathbf{L}$, which records pairwise similarity for each pair of frames. A DPP encodes a discrete probability distribution for all the 2^T subsets from our ground set $\mathbf{Z}$. Then the probability of selecting a subset $\mathbf{z}$ is:

$$P(\mathbf{z} \subset \mathbf{Z}; \mathbf{L}) = \frac{det(\mathbf{L_z})}{det(\mathbf{L} + \mathbf{I})}, \tag{3.8}$$

Here $\mathbf{I}$ is an N x N identity matrix. And $\mathbf{L_z}$ is the principal minor with columns and rows selected referring to the indexes in $\mathbf{z}$. So, if we select a subset $\mathbf{z}$ of two items i and j, we have

$$P(\mathbf{z} = \{i, j\}) \propto L_{ii}L_{jj} - L_{ij}^2. \tag{3.9}$$

If two items i and j are identical, $\mathbf{L_z}$will have identical rows and columns, $L_{ii} = L_{jj} = L_{ij}$, so $P(\mathbf{z} = \{i, j\}) = 0$. In such case, we will get a zero probability for this subset. To summary, a subset with better diversity owns a higher probability.
Through our discussion above, we can see that DPP provides an effective algorithms for sampling or other inference applications. As shown in Figure 1.4, we can find some activities in the egocentric video may persist very long, which is definitely not good for learning storyline, since we want to show the causal activity relations in the video, yet without repetition. So we use the DPP methods to initiate our refined RNN model before training, in order to learn the causal relationships between different activities, rather than trapped into some dominant activies like cutting mushrooms as shown in the Figure 1.4.
In summary, as shown in Figure 3.3, our method introduce the latent variables $\mathbf{z}_n$ and DPP to sample the video frames. And to conquer the regression problem for high-dimensional data, we utilize softmax loss over future images.

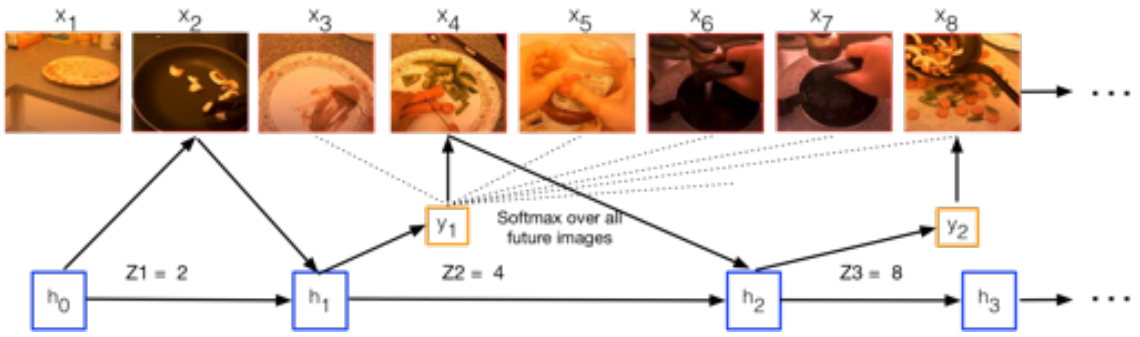

Figure 3.3: Our method, use latent valuable z_n to skip through gaze-centered image sequences, as to extract common latent stories.

Chapter 4

Experiments

Although there have been a lot of works for video summarization, but for the field of storyline learning, there haven't been done a lot yet. And also no established datasets by now, especially for egocentric videos. Therefor, we first introduced GTEA Gaze+ dataset that we used in our experiments, and also how we collect the ground truth storyline for different tasks based on the recipe. Then we performed two seperate experiments to help better evaluate our method. For the first one, we address it as a video summarization problem, where F-score (temporal overlap between generated and ground truth storyline) is utilized as the evaluation metric to evaluate our result on task of video summarization, by the diversity of our generated subset and also the importance of the actions included in our generated storyline. Then we perform the task of challenging semantic forecasting in storylines, which is to predict the image which represents the next event from the storyline.

4.1 Dataset

In this part, We will introduce the GTEA Gaze+ dataset we used in both of our experiments, and also how we collect the ground truth storyline for different tasks.

4.1.1 Related datasets

GTEA Gaze+ Dataset [19]
The GTEA Gaze+ is constructed for action recognition originally, though we can also utilize it for learning storyline. This dataset contains seven different activities of meal preparation, including making American Breakfast, making Pizza, making Snack, making Greek Salad, making Pasta Salad, making Turkey Sandwich and making Cheese Burger. Each task is performed by 4-6 people, and each video lasts for 10 - 20 minutes. Subjects are required to finish these tasks following the steps on given recipes for that task. The videos are recorded at 24 frames per second with frame resolution at 960 x 1280. The gaze information for the subject is also recorded at 30 fps. The gaze distribution for GTEA Gaze+ dataset is less variance and mostly concentrated in the bottom half due to the task at hand (mostly meal preparation). Gaze points are useful if they are reflective of the object being manipulated. As shown in Figure 1.3, we can see that the gaze region usually contains the important objects during the manipulation, thus being helpful to learn the storyline for these videos.

This dataset also have action annotations for each video. The number of the action annotation is around 70 to 200, and each annotation may last for 1 to 40 seconds.

4.1.2 Ground truth storyline

In order to evaluate the generated storyline, we manually collect ground truth storyline for each task, based on the recipe for the task.

Recipe for the task

Every camera wearer is required to complete the task (one task from making American Breakfast, making Pizza, making Snack, making Greek Salad, making Pasta Salad, making Turkey Sandwich and making Cheese Burger) based on the recipe. And each recipe contains 10 - 15 steps. We show the recipe for making snack as following.

Recipe for making snack

- Put a little (around two spoonfuls) peanut butter into a microwave-safe bowl.
- Add a little honey to the peanut butter. Mix to combine the honey and peanut butter.
- Microwave the bowl on high for 20 seconds.
- Add a little (around two spoonfuls) strawberry jam to the bowl and mix to combine.
- Spread the mixture onto a slice of bread. Top with another slice of bread to finish the sandwich.
- Put some water into the kettle and place the kettle on the stove.
- Bring the water in the kettle to a boil.
- Place a tea bag or instant coffe into a coffee/tea cup.
- Pour boiling water into the cup.
- Add sugar to the drink if desired.
- Pour some cereal into a bowl.
- Fill the bowl with milk.
- Add honey or chocolate syrup to sweeten the cereal.

Collect Groud Truth storyline

We collect the groud truth storyline based on the recipe for each task. Basically, we select one image for every step in the recipe. Since we have action annotations for the video frames, we can select a frame which has the reletive action annotations with every step in the recipe. However, in some cases, the step in the recipe may be very general to find a conrresponding action-annotated frame from the video, such as the step 7 (Bring the water in the kettle to a boil.) in the recipe of making snack. So we just ignore these steps. Also, there may consist of more than one actions in the same step, like the step 4 (Add a little (around two spoonfuls) strawberry jam to the bowl and mix to combine.), which contains two actions 'ADD' and 'Mix', actually, from our observation, these actions happened very near, and the image frame of these actions are pretty similar, so we only use collect one image for these kind of steps. In the Figure 4.1, we show a ground truth storyline we collected for the task of making snack.

4.2 Video summarization

For the first experiment, as showin in Figure 4.2, we address our method as an application for video summarization problem, where F-score (temporal overlap between generated and ground truth storyline) is utilized as the evaluation metric to evaluate our result on task of video summarization, by the diversity of our generated subset and also the importance of the actions included in our generated storyline. We use Alexnet[17] to extract the $fc7$ features for each cropped gaze region, then use cross-validation for each task in the GTEA Gaze+ Dataset.

4.2.1 Implementation Details

For every video in GTEA Gaze+ Dataset, we first extract the gaze region for every video frame per second, and use Alexnet[17] to extract the $fc7$ features. For the training, we apply 4:1 cross-validation among each task. And we use DPP to initialize the selected subset, and also use this subset to initialize our modle parameters. For our method, we feed the $fc7$ features to our RNN model directly. The beginning learning rate is set to be 0.05, and it is reduced gradually when the probability on validation set is constrained. And we train our refined RNN with BPTT, by using gradient ascent with the momentum of 0.9. The size of input layer is equal to $fc7$, which is 4096, and 50 is the size of the hidden layer. Considering the recipe for each task contains 8 - 12 key steps, so we keep N = 12 for all the task in the GTEA Gaze+ dataset, which means we take 12 images to contruct our storyline for every task. Also, $\updownarrow_2$ regularizer is used in our network, and the weight decay is set to be 10^{-7}.

We also made comparison with several other approaches to show its effectiveness and accuracy in learning storylines. And here all the other methods used the same $fc7$ features which extracted from AlexNet[17].

Below we list some main baselines which we used to compare with our method.

RNN

We utilize the language model in [23], to predict the cluster of the next image. We

Figure 4.1: Ground truth storyline for the task of making snack.

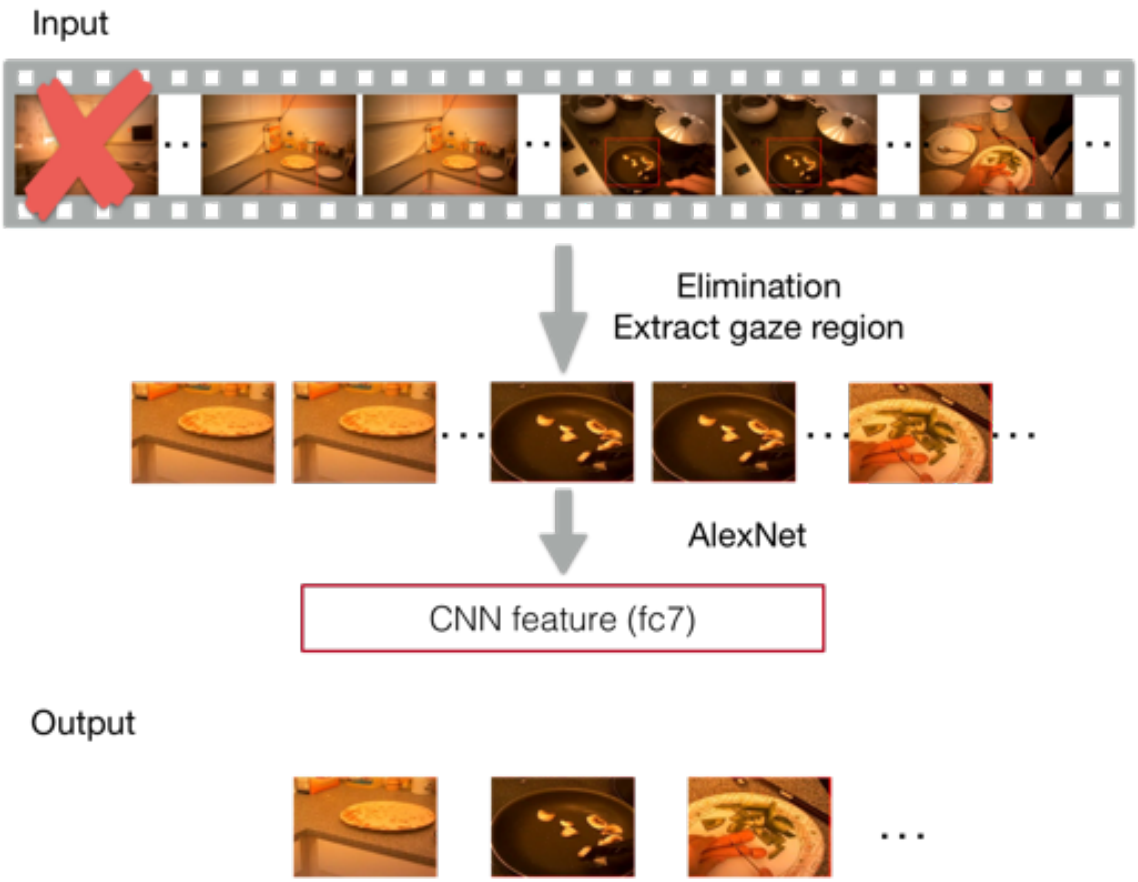

Figure 4.2: Our experiment of video summarization. For the input, we first select one frame per second from each video, eliminate useless frames which have no action annotation, and extract gaze region. Finally we use Alexnet to extract the fc7 features for each gaze-centered image. We perform a 4:1 cross-validation training over each task. The output is subset of input frames containing certain images.

make it a standard application of RNN, by sampling without replacement to create our storyline.

K-Means

We apply K-Means clustering on $fc7$ features for every single video. Then we specify the closest image to every cluster center to get a storyline.

LSTM

Like the RNN baseline, we also trained an LSTM network as in citeKarpathy2015Visualizing.

LSTM-sub

We following the set up in [30], first we train the LSTM as before. But for generating a storyline, we first let the network create a longer sequence (for example, a subset contains 100 images), and then sub-sample the long sequence to the length we want, here we use 12 items.

4.2.2 Evaluation set up

To evaluate our storyline (especially the diversity and importance of the storyline), we utilize the keyshot-based metric proposed in [38] for teh evaluation. Here, Let's suppose A to be the generated storyline and B the ground truth storyline that we collected. We compute the precision and recall according to the amount of *temporaloverlap* between the two as follows:

$$precision(P) = \frac{duration\ of\ overlap\ between\ A\ and\ B}{duration\ of\ A} \tag{4.1}$$

$$recall(R) = \frac{duration\ of\ overlap\ between\ A\ and\ B}{duration\ of\ B} \tag{4.2}$$

also their harmonic mean F-score,

$$F = 100\% * 2P * R/(P + R). \tag{4.3}$$

4.2.3 Results

In the table Table 4.1, we show the results for our method and four other baselines as mentioned in the implementation part. And in the Figure 4.3, we show some examples of storylines created by our method and three other representative baselines RNN, LSTM and K-means, for the video of making snack. And we found that different baseline may have failure in different ways. For instance, K-Means is able to generate a storyline with a more diverse set of images, but it can be easily affected by the inherent noise in the GTEA Gaze+ egocentric video dataset. On the other side, RNN and LSTM are both prone to learn a shorter temporal relationships, thus being easily trapped in the repetitive frames. However, for our method, since we use a sampling architecture combined with DPP initialization, we can easily skip the repitition part, and get a long-term temporal relationship learning, thus generating a more diverse and comprehensive storyline.

4.2.4 Results for different task

As we mentioned before, GTEA Gaze+ Dataset contains seven different activities of meal preparation, including making American Breakfast, making Pizza, making

	RNN	K-Means	LSTM	LSTM-sub	Ours
F score	32.1	39.8	35.3	41.3	45.9

Table 4.1: Results for the video summarization. F-score for our method and four other baselines we have mentioned in the implemention part. Our method shows the best performance.

Figure 4.3: Examples of storylines generated by our method and three representative baselines RNN, LSTM and K-means, for the video of making snack. Compared with other baselines can get a better long-term learning, thus generating a more diverse and comprehensive storyline.

Snack, making Greek Salad, making Pasta Salad, making Turkey Sandwich and making Cheese Burger. Each task is performed by 4-6 people. Subjects are required to finish these tasks following the steps on given recipes for that task. So, in this part, we will present our result for each task.

Task 1: Making American breakfast

This is a task for making American breakfast, which consists of frying eggs, frying bacon and making bagel. The recipe is as follws.

- In a large bowl, crack 2 eggs.
- Add a little milk and salt to taste.
- Beat the egg mixture using a whisk or fork until well blended.
- Pour some oil (olive, vegetable, etc.) into a non-stick frying pan.
- Heat the oil over medium heat.
- Pour the egg mixture into the frying pan and stir frequently.
- When the eggs are done remove the pan from the heat and transfer the eggs to a plate.
- Pour orange juice into a cup.
- Again pour some oil into a non-stick frying pan.
- Heat the oil over medium heat.
- Fry a piece of bacon in the skillet.
- Spread cream cheese onto a bagel half.
- Complete the sandwich with the other half of the bagel.

And in Figure 4.4, we showed the ground truth storyline and our generated storyline for the task of making American breakfast.

Task 2: Making afternoon snack

This is a task for making afternoon snack, which consists of making peanut butter, jelly, hot tea, milk and cereal. The recipe is as follws.

- Put a little (around two spoonfuls) peanut butter into a microwave-safe bowl.
- Add a little honey to the peanut butter. Mix to combine the honey and peanut butter.
- Microwave the bowl on high for 20 seconds.
- Add a little (around two spoonfuls) strawberry jam to the bowl and mix to combine.
- Spread the mixture onto a slice of bread. Top with another slice of bread to finish the sandwich.
- Put some water into the kettle and place the kettle on the stove.

- Bring the water in the kettle to a boil.
- Place a tea bag or instant coffe into a coffee/tea cup.
- Pour boiling water into the cup.
- Add sugar to the drink if desired.
- Pour some cereal into a bowl.
- Fill the bowl with milk.
- Add honey or chocolate syrup to sweeten the cereal.

And in Figure 4.5, we showed the ground truth storyline and our generated storyline for the task of making afternoon snack.

Task 3: Making pizza

This is a task for making pizza, which consists of preparation, frying mushrooms and completing pizza. The recipe is as follws.

- Take the pizza bread from the fridge and let warm on the counter.
- Cut up a hot dog/sausage into 1/2 inch slices.
- Cut up a green bell pepper into bite-sized chunks.
- Cut up a few mushrooms (enough for a pizza).
- Again pour some oil into a non-stick frying pan.
- Heat the oil over medium heat.
- Fry the mushrooms slices in the skillet.
- Put enough ketchup on the pizza crust to thinly cover it.
- Place the hot dog/sausage slices, green bell pepper slices, and fried mushrooms on the pizza.
- Add enough shredded mozzarella cheese to cover the pizza.
- Place the pizza in the pre-heated oven for 20 minutes.
- Remove and let cool on the counter.

And in Figure 4.6, we showed the ground truth storyline and our generated storyline for the task of making pizza. And we can see our generated storyline consists of 5 actions in the ground truth storyline.

Task 4: Turkey Sandwich

This is a task for making Turkey sandwich, which consists of dicing the tomatoes, cutting the lettuce and completing. The recipe is as follws.

- Take an individual tomato and cut it into slices. Set aside and repeat for more tomato slices (1 large tomato cut into slices is usually enough).

Figure 4.4: The ground truth storyline and our generated storyline for the task of making American breakfast.

Figure 4.5: The ground truth storyline and our generated storyline for the task of making afternoon snack.

- If your lettuce is pre-separated in a bag remove a few pieces
- If your lettuce is in a bunch (a head of lettuce) tear off a few pieces of lettuce
- Cut each piece into a manageable size which would fit on a bread slice.
- Take a slice of bread and put it on a plate.
- Place a few slices of turkey on the slice of bread.
- Put the lettuce and tomato slices on top of the turkey.
- Add a cheese slice or two.
- Garnish with ketchup, mustard, or mayo if desired.
- Top with the remaining bread slice to finish the sandwich.

And in Figure 4.7, we showed the ground truth storyline and our generated storyline for the task of making Turkey sandwich. And we can see our generated storyline consists of 5 actions in the ground truth storyline.

Task 5: Greek Salad

This is a task for making Greek salad, which consists of cutting the Vegetables and completing the salad. The recipe is as follws.

- Slightly chop a few pieces of lettuce into bite-sized chunks.
- Dice a few tomatoes. Don't make them too small; the slices should be large enough to pick up with a fork.
- Peel the cucumber and put into " slices.
- Quarter the onion and then separate. Keep only pieces large enough to eat easily.
- In a large bowl add the lettuce, tomato, cucumber, and onion slices.
- Top with feta cheese
- Sprinkle on vinegar, lemon juice, and olive oil to taste.

And in Figure 4.8, we showed the ground truth storyline and our generated storyline for the task of making Greek salad.

Task 6: Pasta Salad

This is a task for making paste salad, which consists of boiling the water, cooking the pasta and completing. The recipe is as follws.

- Fill a small pot with water (about 1/2 full).
- Place the pot on the stove top and set the burner to "high."
- When the water comes to a boil add a cup of macaroni noodles to the water.
- When the noodles become tender remove the pot from the heat and drain.

Figure 4.6: The ground truth storyline and our generated storyline for the task of making pizzz.

Figure 4.7: The ground truth storyline and our generated storyline for the task of making Turkey sandwich.

- Wash the drained noodles under cold water.
- Pour the cooked drained noodles into a bowl.
- Roughly chop a few tomatoes into bite-sized chunks and add to the bowl of noodles.
- Repeat with the green bell peppers, cucumbers, onions, carrots, and black olives. Remember to peel the cucumbers and carrots and quarter the onions.
- Sprinkle with your favorite dressing.

And in Figure 4.9, we showed the ground truth storyline and our generated storyline for the task of making Pasta salad.

Task 7: Cheese Burger

This is a task for making cheese burger, which consists of cooking and completing the burger. The recipe is as follws.

- Pour a little oil into a non-stick frying pan.
- Heat the oil over medium heat on the stove.
- Put a beef patty into the skillet and cook until done.
- Remember to flip the patty every couple of minutes.
- (OPTIONAL) When cooked throughout place a slice of cheese on the patty and let it melt. DO NOT FLIP THE BURGER OVER AT THIS POINT.
- Turn off the stove and remove the patty from the pan using a spatula. Place the burger onto the bottom half of the bun.
- Slice a tomato into fairly thin slices.
- Separate a few pieces of lettuce and cut them in half.
- Put the tomato slices and lettuce slices on top of the burger.
- Garnish with your favorite condiments (ketchup, mayo, mustard).
- Top with the remaining burger half.

And in Figure 4.10, we showed the ground truth storyline and our generated storyline for the task of making cheese burger.

4.2.5 Storyline with and without Gaze

We analyze this issue both qualitatively and quantitatively using GTEA Gaze+ dataset. For our method, we use the original frame and the cropped region (gaze centered region) as input separately, and compare their generalized storyline, we show the result in Table 4.2. And Figure 4.11 shows two storylines expamles for making snack, top side is the result trained with the orignal frames, and bottom side is trained with the cropped gaze region. As we can see, without gaze information,

Figure 4.8: The ground truth storyline and our generated storyline for the task of making Greek salad.

Figure 4.9: The ground truth storyline and our generated storyline for the task of making Pasta salad.

	Original frame	Gaze region
F score	40.5	45.9

Table 4.2: Results for using gaze region and the original frame for video summarization.

Figure 4.10: The ground truth storyline and our generated storyline for the task of making cheese burger.

Figure 4.11: Two storylines for making snack, top side is trained with the orignal frames, and bottom side is trained with the cropped gaze region

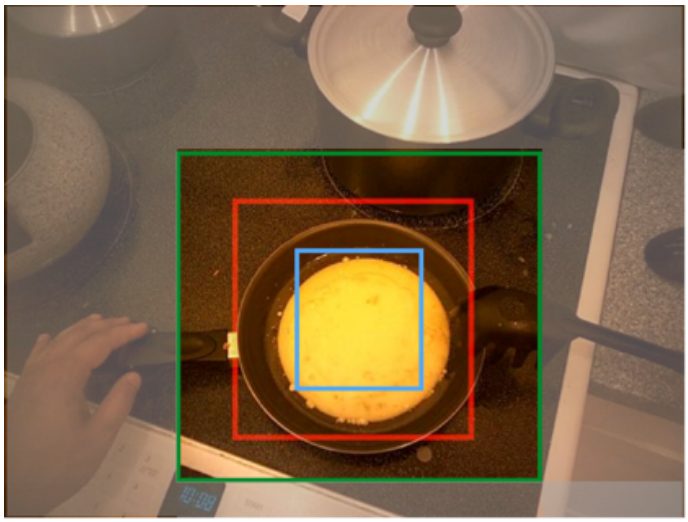

Figure 4.12: Three different size of gaze region we extracted. Green one: 384*384, red one: 256*256, blue one: 128*128

Size of gaze region	F score
128*128	42.4
256*256	44.3
384*384	43.1
Full image	40.5

Table 4.3: Results for three different size of gaze region.

it is more likely to pick some unuseful frames, which do not carry much important information. Besides, we can find that the gaze-enabled storyline shows both more diversity (gaze-enabled storyline tend to select more different activities) and less temporal relevance.

Different gaze size Through our previous results and analysis, we confirmed that we are able to generate a storyline which is more diverse and comprehensive by using only the gaze region. But how should we decide the size of the gaze region? Here we extracted three different size of the gaze region (384*384, 256*256 and 128*128), as shown in the Figure 4.12. To mention here, if the gaze position falls near the edge, we just move the gaze position towards the center of the image, untill we can get the exact square size we want. And the results are shown in the Table 4.3.

4.2.6 Initialization with DPP

As we have introduced in section 3.4, DPP offers efficient and exact algorithms for sampling, conditioning, and other inference tasks. In order to get better performance on the long-term learning among different actions, rather than being trapped into some dominant and long-lasting activies like cutting mushrooms as shown in the Fig.5., we use the DPP methods to initiate the subset sequence before training. Our results showed that if we use DPP as initialization, we gain an F-score of 45.9, much higher than the randomly initialization, whcih is only 41.9. And In Figure 4.13, we showed two example storylines of making snack by using randomly initialization and DPP initialization separately. Through this image, we can comfirm that by using

Figure 4.13: Two storylines for making snack, top side is using a randomly initialization, and bottom side is using the DPP initialization.

the DPP methods to initiate our refined RNN model before training, we can learn a better causal relationships between different activities, rather than trapped into some dominant activies.

4.3 Storyline prediction

In the next experiment, in order to evaluate how our method is capable of learning the long-term temporal dynamics among the different and important actions or events, we perform the task of semantic challenging forecasting for storyline [30]. Given a storyline, which consists of certain images, and one image or several continuous images from this storyline, our goal is to predict a image which represents the next event from the storyline. This is difficult task for that it's hard to use visual feature matching [36] to capature those semantic changing directly. As shown in Figure 4.14, this is a prediction task for the storyline of making pizza. After the step of cutting mushroom and cutting pepper, the correct prediction for the next step or event in this storyline should be cutting hot dog, all the other predictions are wrong.

4.3.1 Experiment set up

We follow the long-term prediction in [30], where the right prediction is next representative action or event in the coming storyline. Here we also pose it as a task of classification, so our goal is to predict the right image, which represents the next event from the storyline, selected from other four images randomly choosed from the same video. Here, we directly use our trained model for each task.

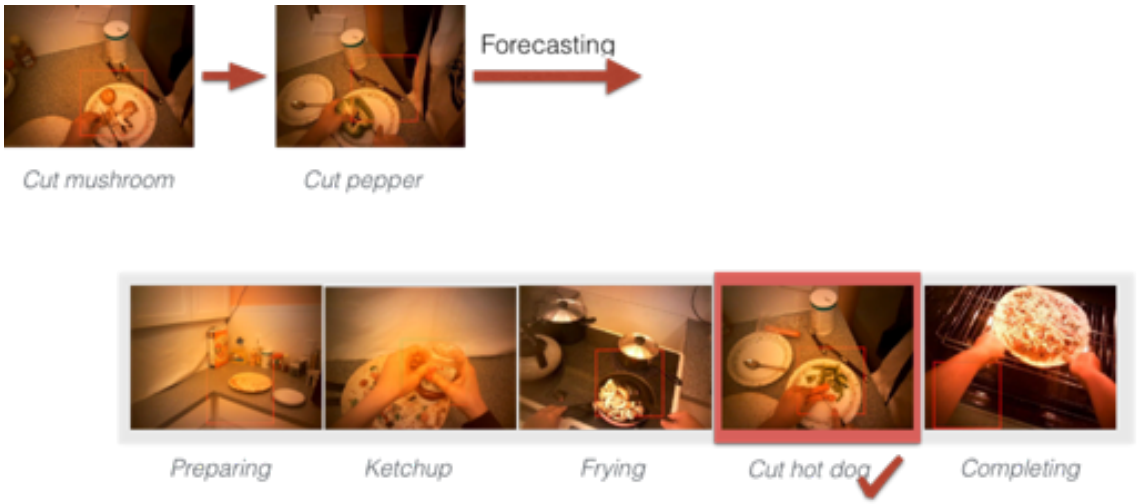

Figure 4.14: Set up for storyline prediction

	RNN	LSTM	LSTM-sub	Ours
Accuracy	18.5%	24.5%	23.3%	33.5%

Table 4.4: Evaluating Storylines. F score for our method and four other baselines we have mentioned in the implemention part. Our method shows the best performance.

Baelines Here we again compare our methods with RNN, LSTM and LSTM-sub as described in the first experiment.

4.3.2 Results

In Table 4.4, we present our results for the storyline prediction. As we can see, our method predicts the next image in the storyline with the accuracy of 33.5%, which is significantly higher than other baselines. In the Figure 4.15, we show the froud truth storyline of making pizza, and in Figure 4.16 we present some prediction examples for this storyline. In Figure 4.16, we use two continous images from the ground truth storyline as input, for this classification task, there is one true image, which represents the next event from the storyline, along with other four images randomly selected from the same video. And image in the red frame is the right prediction, the green one is the wrong prediction.

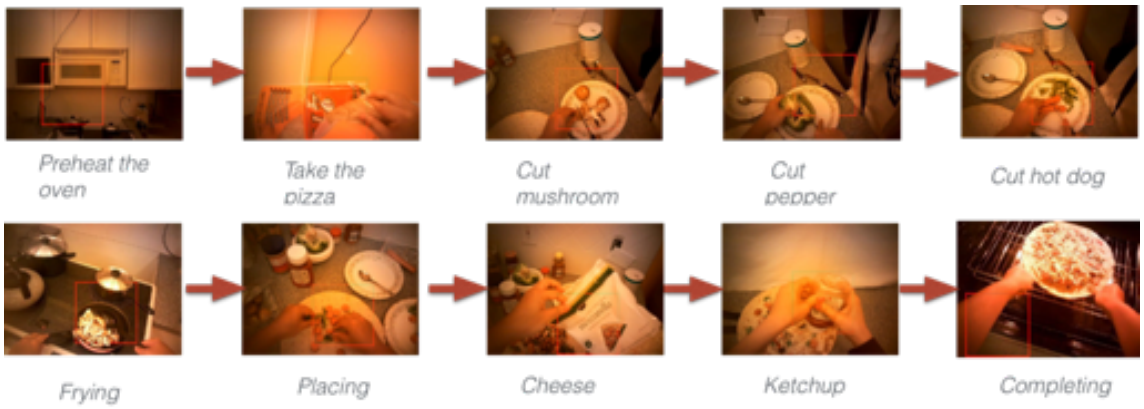

Figure 4.15: Ground truth storyline making pizza.

Figure 4.16: Some examples for storyline prediction in the task of making pizza. Image in the red frame is the right prediction, the green one is the wrong prediction.

Chapter 5

Conclusion

In this thesis, we showed how to utilize gaze position to generalize storyline for egocentric videos since gaze information can provide us more specific clues for the current action of the camera wearer. For that gaze fixation indicates somebody's intention, making it useful for personalization. Moreover, people usually focus on some specific object during a manipulation or activity (e.g. fixing on dishes when making a meal), so it helps for object segmentation and activity recognization, which can be used to construct causal event based storyline.
We first introduced some recent related works, like learning storyline, video summarization, and gaze in egocentric video. Then, we introduced the current challenges for learning storyline in egocentric videos, first one is that the background is usually clustered, which makes it hard to track the important objects for current action. We solve this by focusing on the gaze region. The other one is that there are so many repetitive and long-lasting actions, making it pretty harder to learn long-term relationship among different actions compared to the normal video, which is dealed with DPP sampling in our method.
In the approach part, we first give a briefly introduction about the visual storyline, which is a chain of events that have causal or chronological relations, and being able to give a brief and effective understanding on contents of the video. Then we present a short introduction about Reccurent Neural Network, and also the limits for long-term learning, due to the limited memory capacity and the vanishing gradients. So we utilize DPP mothod to sample the subset first before traing the RNN, which helps to learn the causal relationships between different actions, rather than being trapped into some dominant actions.
Finally, we first introduced the related datasets we used our method, and also how we collected the ground truth storyline for each task based on the recipes. Then, we presents two seperate experiments to help better evaluate our method: evaluating storyline by the diversity and importance, storyline prediction. Both experiments indicate our method show better performance for learning storyline compared with other baselines.
However, there are still some challenges to be overcome in the future. The main challenge would be the personalization of the storyline. Since the evaluation of storyline is so subjective, so a well storyline should achieve to the specific preferences according to that person. Besides, with each methods working with its own dataset and evaluated differently, it is rather hard to make comparisons between different methods among them. Hence, we believe there is a necessarily to build a benchmark

for evaluating storyline-learning techniques for egocentric videos equally.

Bibliography

[1] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 2002.

[2] Daniel DeMenthon, Vikrant Kobla, and David S. Doermann. Video summarization by curve simplification. In *ACM Multimedia*, 1998.

[3] J Donahue, L. A. Hendricks, M Rohrbach, S Venugopalan, S Guadarrama, K Saenko, and T Darrell. *Long-term Recurrent Convolutional Networks for Visual Recognition and Description.* Elsevier,, 2015.

[4] Mehdi Ellouze, Nozha Boujemaa, and Adel M. Alimi. Im(s)2: Interactive movie summarization system. *J. Visual Communication and Image Representation*, 21:283–294, 2010.

[5] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[6] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327, 2012.

[7] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-view video summarization. *IEEE Trans. Multimedia*, 12:717–729, 2010.

[8] Joydeep Ghosh. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, 2012.

[9] Boqing Gong, Wei Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *International Conference on Neural Information Processing Systems*, pages 2069–2077, 2014.

[10] Ian J. Goodfellow, Jean Pougetabadie, Mehdi Mirza, Bing Xu, David Wardefarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3:2672–2680, 2014.

[11] Yuri A. Ivanov and Aaron F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence IEEE Transactions on*, 22(8):852–872, 2000.

[12] N. Jindal. *MIMO Broadcast Channels With Finite-Rate Feedback*. IEEE Press, 2006.

[13] T Judd, K Ehinger, F Durand, and A Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, pages 2106–2113, 2009.

[14] Andrej Karpathy and Fei Fei Li. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664, 2017.

[15] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2698–2705, 2013.

[16] Gunhee Kim and Eric P. Xing. Reconstructing storyline graphs for image recommendation from web community photos. In *Computer Vision and Pattern Recognition*, pages 3882–3889, 2014.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

[18] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5:123–286, 2012.

[19] Yin Li, Alireza Fathi, and James M. Rehg. Learning to predict gaze in egocentric video. In *IEEE International Conference on Computer Vision*, pages 3216–3223, 2014.

[20] David Liu, Gang Hua, and Tsuhan Chen. A hierarchical visual model for video object summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:2178–2190, 2010.

[21] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. pages 2714–2721, 2013.

[22] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[23] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September*, pages 1045–1048, 2010.

[24] Ana Garcia Del Molino, Cheston Tan, Joo Hwee Lim, and Ah Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, PP(99):1–12, 2017.

[25] Chong-Wah Ngo, Yu-Fei Ma, and HongJiang Zhang. Automatic video summarization by graph modeling. In *ICCV*, 2003.

[26] Chong-Wah Ngo, Yu-Fei Ma, and HongJiang Zhang. Video summarization and scene detection by graph modeling. *IEEE Trans. Circuits Syst. Video Techn.*, 15:296–305, 2005.

[27] Pere Obrador, Rodrigo de Oliveira, and Nuria Oliver. Supporting personal photo storytelling for social albums. In *ACM Multimedia*, 2010.

[28] R. C. Schank and R. P. Abelson. Scripts, plans, goals, and understanding. *Readings in Cognitive Science*, pages 190–223, 1988.

[29] N. Shapovalova, M. Raptis, L. Sigal, and G. Mori. Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. *Advances in Neural Information Processing Systems*, july 1971(3):309–319, 2013.

[30] Gunnar A. Sigurdsson, Xinlei Chen, and Abhinav Gupta. *Learning Visual Storylines with Skipping Recurrent Neural Networks.* Springer International Publishing, 2016.

[31] Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. Summarization of personal photologs using multidimensional content and context. In *ICMR*, 2011.

[32] Ronald J Williams and David Zipser. Gradient-based learning algorithms for recurrent networks and their computational complexity. In *Backpropagation*, 1995.

[33] J Xut, L Mukherjee, Y. Li, J Warner, J. M. Rehg, and V Singht. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Computer Vision and Pattern Recognition*, page 2235, 2015.

[34] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. *Computer Science*, pages 21–29, 2015.

[35] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. Studying relationships between human gaze, description, and computer vision. In *Computer Vision and Pattern Recognition*, pages 739–746, 2013.

[36] Kuo Hao Zeng, William B. Shen, De An Huang, Min Sun, and Juan Carlos Niebles. Visual forecasting by imitating dynamics in natural sequences. In *IEEE International Conference on Computer Vision*, pages 3018–3027, 2017.

[37] Ke Zhang, Wei Lun Chao, Fei Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *Computer Vision and Pattern Recognition*, pages 1059–1067, 2016.

[38] Ke Zhang, Wei Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. pages 766–782, 2016.

[39] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. 2017.

[40] Yuke Zhu, Oliver Groth, Michael Bernstein, and Fei Fei Li. Visual7w: Grounded question answering in images. pages 4995–5004, 2015.

List of Figures

List of Tables

Acknowledgements

I would like to extend my sincere thanks to all those who provided me the possibility to complete this thesis. It would not have been possible without the kind support and help of many individuals who provided expertise that greatly assisted the research.
I thank My professor Yoichi Sato and Doctor Mingjie Cai for giving me the opportunity to write this master thesis, and for their professional advise and guidance.
I thank in particular my senior Yifei Huang, who willingly provided information and mentoring at any time.
I would also like to thank all people who supported me in writing this thesis, especially my families and friends.

July 11th, 2018
Binhua Zuo