Project Proposal: Visual Logger of daily activity

Problem Statement:

Given a video recording of a person's daily activity along with his/her GPS data, the tool should be able to identify salient images that can act as a visual diary for a day. A second goal would be to supplement these images with text that describes the images along with a summary of the day. This would require the use of accelerometer data too. A final goal would be to perform all these tasks on an embedded device.

Recent Work:

There is recent literature in methods to create annotated visual logs of a person's daily activity. In [1], the authors have used CNN and LSTM to create a textual description of salient images in a video. There was also heavy focus on running the tasks of CNN and LSTM on an embedded device within strict power constraints. In [2], the authors have made use of a new dataset for daily activities. This was then used in conjunction with the larger COCO dataset which does not have first person annotation. In [3], a dataset called SenseCam-32 was developed to annotate such daily activities with the specific goal of using it to create a lifelog from video. NeuralTalk [2] by Karpathy has been used in [1] to generate text from images.

Methodology:

Through this project, I want to add the data provided by GPS sensors and perhaps accelerometers, to make the identification of salient images in a video more context aware. Specifically, as opposed to other methods that rely purely on salient image characteristics on identifying which images in a video are unique, this project will also make use off GPS data to identify which video sequences are truly representative of a person's visual log. For example, a 20 minute walk from a person's home to class would warrant as many salient images as a 10 minute conversation in a single GPS location.

Further, the use of GPS and accelerometer can help generate more context aware of textual representations of images. For example, using images alone I could describe an image as 'I am typing on a table'. But using GPS data this could be, 'I went from my house to the POB building. I am typing on my table.' Using accelerometer readings could further this context, through, 'I *sprinted* from my house to the POB building. I am typing on my table.'

To achieve this, I would need to use a pre trained caffe model on the COCO dataset such as in [4] and perform transfer learning for a relevant dataset such as SenseCam [3]. New video would then be passed to the model to get the CNN features corresponding to each of the image frames. GPS data will be used to split the images into different zones based on locations (time slabs where GPS does not change is indicative of no movement). This will ensure that activities such as walking with many salient images does not skew the log. Finally, the selected images will be sent to a LSTM network like NeuralTalk that will be able to create a textual representation for it. GPS and accelerometer data will be used to create text to transition between the description of different frames.

Requirements:

Hardware needed:

1. Accelerometer and GPS (mobile apps)
2. Collecting video data (wearable camera like GoPro)

Datasets:

1. MS COCO dataset
2. SenseCam-32 dataset (not publicly available. I have mailed the authors of [3] to request them to make it public)

References:

[1] Neural Diary: Forming Compressed Visual Stories in Real-Time
[2] DeepDiary: Automatic Caption Generation for Lifelogging Image Streams
[3] In the sight of my wearable camera: Classifying my visual experience
[4] A. Karpathy. Neuraltalk 2. https://github.com/karpathy/neuraltalk2, 2015.