Video Description using Deep Learning

Summer Undergraduate Research Award

Suyash Agrawal
2015CS10262
Computer Science
CGPA: 9.91
Mob: 9717060183
cs1150262@iitd.ac.in

Madhur Singhal
2015CS10235
Computer Science
CGPA: 8.66
Mob: 9540972599
cs1150235@iitd.ac.in

Supervisor:-
Subhashis Banerjee
Professor
Department of CSE
suban@cse.iitd.ac.in
IIT Delhi

Prof. S. Arun Kumar
Head of Department
Department of CSE
sak@cse.iitd.ernet.in

1 Introduction

Video Description is the process of discovering knowledge, structures, patterns and events of interest in the video data and describing them in natural language. Video Description is an incredibly hard problem in computer vision and currently the only source of video description is manual labour.
Video Description has wide variety of applications. It can help visually impared people "see" the world by describing the scene around them. It has also use in automated survelliance by analysing the videos in real time and reporting criminal activites. Also, it can be used to efficiently index large video databases based upon their content for ease of accessibility.

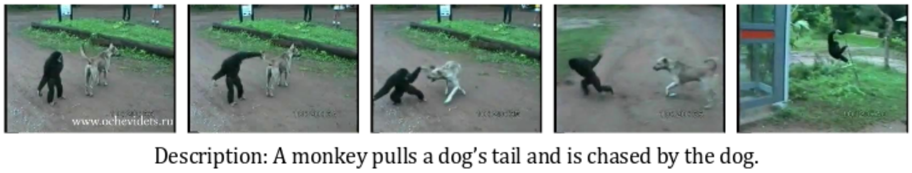
Description: A monkey pulls a dog's tail and is chased by the dog.

Figure 1: Sample Video Description

Figure **??** shows a possible description of a sample video. The *traditional pipeline* is shown in Figure 2.

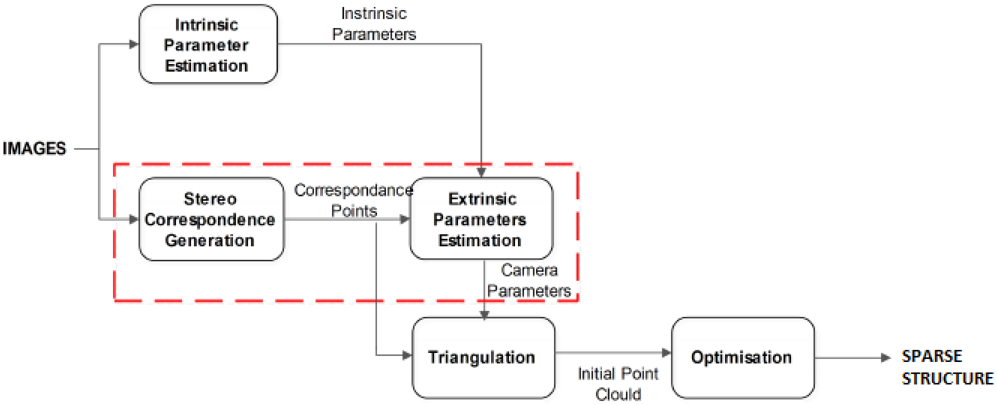

Figure 2: Traditional pipeline

The red-highlighted part of the pipeline is ***computationally expensive***. Thus, our project aims to reduce this computation and perform 3D reconstruction in near real-time.
The processing parts are:

- *Intrinsic and extrinsic parameters*: The camera projection matrix is a

3×4 matrix which represents the pinhole geometry of a camera for mapping 3D points in the world coordinates to 2D points on images. This matrix depends on extrinsic and intrinsic parameters. The intrinsic parameters mainly comprises of focal length, image sensor format, and principal points. The extrinsic parameters define the position of the camera center and the camera's heading in world coordinates in terms of a rigid rotation and translation.

- *Stereo correspondence generation*: Given two or more images of the same 3D scene, taken from different points of view, the correspondence problem refers to the task of finding a set of points in one image which can be identified as the same points in another image. To do this, points or features in one image are matched with the corresponding points or features in another image. The images can be taken from a different point of view, at different times, or with objects in the scene in general motion relative to the camera(s).

- *Triangulation*: Triangulation refers to the process of determining a point in 3D space given, its projections onto two or more images and their corresponding camera projection matrices. This point is found as the intersection of the two or more projection rays formed from the inverse projection of the 2D image points representing that 3D point in space.

- *Initial point cloud and 3D sparse reconstruction*: As the word suggests, *3D sparse construction* is done for only some set of data points in the given coordinate system called *initial point cloud.* Figure **??** illustrates a 3D sparse construction of a chariot. Figure **??** illustrates 3D dense construction of the same initial point cloud.

2 Objectives

Our main objective is to perform 3D reconstruction in near real time using a mobile device. This can be further be subdivided into following points:

1. To get accurate position and orientation estimate based on readings of IMU sensors in smart-phones.

2. To use the camera feed in smart-phones to enhance the position estimate based on visual tracking of objects.

3. To do sparse 3D reconstruction based on sensor fusion data and computer vision techniques.

4. To enhance the quality and efficiency of 3D reconstruction by adding more details and moving towards dense 3D reconstruction.

5. We will ultimately be fusing digital signal processing and computer vision based techniques that will enable us to perform near real time 3D reconstructions on mobile or hand-held devices.

3 Basic Concepts

3.1 Deep Learning

Deep Learning is a branch of machine learning in which multiple parameter based models are used in series. In a deep network, there are many layers between the input and output, allowing the algorithm to use multiple processing layers, composed of multiple linear and non-linear transformations. At each layer, the signal is transformed by a processing unit, like an artificial neuron, whose parameters are 'learned' through training. Deep Learning has been shown to excel in tasks where the goal is to find intuitive patterns in the data. In particular, in the field of Computer Vision, deep networks are increasingly used to extract feature descriptions and interrelationships from images.

3.2 Convolutional Neural Networks

Convolutional Neural Networks (CNN, or ConvNet) are a type of feed-forward artificial neural network in which the connectivity pattern between the neurons is inspired by the organization of the animal visual cortex.Individual cortical neurons respond to stimuli in a restricted region of space known as the receptive field. The receptive fields of different neurons partially overlap such that they tile the visual field. The response of an individual neuron to stimuli within its receptive field can be approximated mathematically by a convolution operation. A Convolutional Neural Network consists of the following layers.

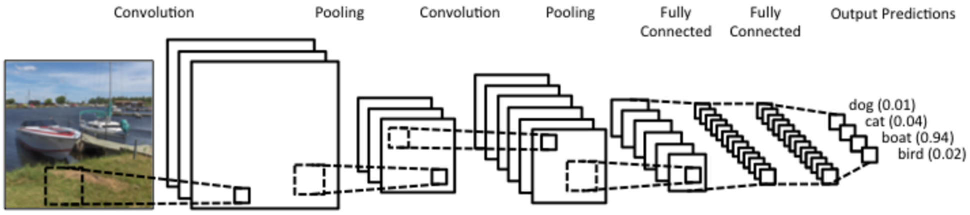

Figure 3: A Typical Comvolutional Neural Network

3.3 Convolutional Layer

The convolutional layer is the core building block of a CNN. The layer's parameters consist of a set of learnable filters (or kernels), which have a small receptive field, but extend through the full depth of the input volume. During the forward pass, each filter is convolved across the width and height of the input volume, computing the dot product between the entries of the filter and the input and producing a 2-dimensional activation map of that filter. As a result, the network learns filters that activate when it detects some specific type of feature at some spatial position in the input.

3.4 Max Pooling Layer

It is common to periodically insert a Pooling layer in-between successive Conv layers in a ConvNet architecture. Its function is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence to also control overfitting. The Pooling Layer operates independently on every depth slice of the input and resizes it spatially, using the max operation.

3.5 Fully-Connected Layer

Finally, after several convolutional and max pooling layers, the high-level reasoning in the neural network is done via fully connected layers. Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset.

3.6 Long Short Term Memory Networks

3.7 Training Data

3.8 Finetuning

3.9 Camera calibration

The camera parameters can further be subdivided into intrinsic and extrinsic parameters. **Camera intrinsic parameter** K is dependent on the focal length of the camera and principal point (which in most cases is the center of the image). The **camera extrinsic parameter** is composed of the rotation R and translation t between camera coordinate system and the world coordinate system. Together they form the camera projection matrix P, a 3×4 matrix which describes the mapping of a pinhole camera from 3D points in the world to 2D points in an image.

$$P = K[R|t] \tag{1}$$

3.10 Sparse 3D reconstruction

Given two different images of the same scene from different angles, the position of a 3D point can be found as the intersection of the two projection rays which is commonly referred to as **triangulation**. For this first point correspondences have to be established. Then using this point correspondences a Random Sampling Consensus (RANSAC) based voting framework is used to estimate the camera intrinsic and extrinsic parameters. Finally, a joint non-linear optimization is used to further refine the camera parameters and the 3D points in a

bundle adjustment framework. This method is computationally very expensive and hence done only for very sparse set of points. This is known as sparse 3D reconstruction.

4 Mobile IMU sensors

IMU (Inertial Measurement Unit) sensors are on-chip devices embedded in most of the smart phones or hand-held devices today. It mainly consists of a series of motion sensors: accelerometer, gyroscope, magnetometer and gravitation sensor. The data from these sensors can be fused to obtain the orientation and the position of the device in the world coordinate system.

5 Conventional versus mobile 3D reconstruction

In the case of a smart-phone or any hand-held device having a camera and IMU sensors, we wish to use the IMU sensors to obtain extrinsic camera parameters in real time. This will help in reducing the load on conventional 3D reconstruction methods and get it in near real time.

6 Approach to the project

First, we shall be using sensor fusion to obtain accurate estimates for camera position and orientation of the mobile device. Then we will move on to 3D reconstruction, which further has two parts: i.e. sparse 3D reconstruction and then use tracking to obtain dense correspondence of points for dense 3D reconstruction.

1. Position and orientation estimation

   (a) Get accelerometer data and orientation data at real time using the IMU sensors like accelerometer, gyroscope, gravity sensor and magnetometer present on the smart phone. This data is highly noisy. Figure 4 shows the position estimate from accelerometer data across various devices.

   As evident from the graph, this data cannot be directly used for calculation of position and orientation. Figure 5 shows the integration of static accelerometer data to obtain velocity and displacement.

   The graph of both velocity and displacement shows significant deviation from the actual value which is zero. Thus, signal processing and smoothening is required to get a better estimate.

   (b) Making the orientation data more accurate by infusing the higher frequency components from the gyroscope orientation after drift correction.

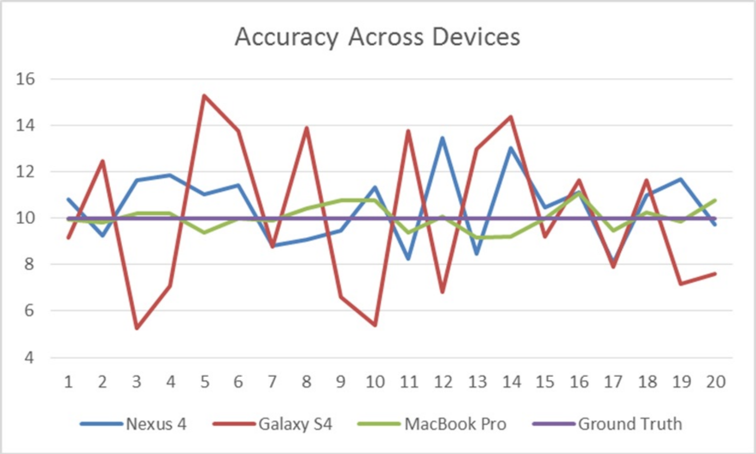

Figure 4: Accuracy of accelerometer data across different devices (scale cm)

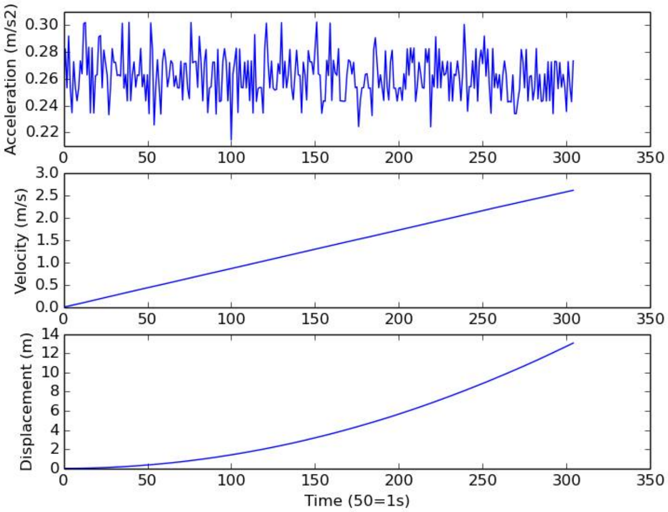

Figure 5: Obtaining velocity and displacement from static accelerometer data

(c) Obtaining the displacement and orientation data from the camera feed on the device using visual tracking methods.

(d) A comparative study is to be done between the position estimates obtained by the two methods along with ground truth and fusing the results to obtain an enhanced position and orientation estimate.

2. 3D reconstruction

(a) Obtain sparse 3D reconstruction based on camera rotation and position parameters obtained previously.

(b) Use tracking data from different tracking methods like “Good features to track” or “KL tracker” for obtaining dense correspondence of points.

(c) Use guided matching by indirect computation of fundamental matrix from estimated camera motion from sensors to further enrich the correspondences.

(d) Triangulate the dense correspondences and do a final global refinement.

3. Further possibilities

   - Getting a more detailed texture mapping of the object.
   - Making an object recognition software on the basis of this 3D reconstruction.
   - Improving the algorithm for a quicker and more efficient 3D reconstruction.
   - Releasing applications for Apple, Android and Windows platforms for near real time 3D reconstruction on the device itself.

7 Uses and applications

- Using the device as an accurate measuring device. This can be of particular interest to blind as they will be able to measure distances and angles accurately with great ease.

- Doing real time dense 3D reconstructions on mobile phones and other hand-held devices.

- Allowing the user to generate a 3D printable file on his mobile device. As 3D printers are becoming cheaper and more common, this feature will reduce the need of the person to use a 3D scanner to be able to generate prototypes of objects. This will allow engineers and students to work more efficiently as they can generate copies of 3D objects easily.

- This project can have applications in the field of archeology. It can be used to generate replica of artifacts and fragile objects for further studies, without harming its integrity.

- Our approach can also be applied in the field of medical sciences, especially for orthopedics and joint replacement surgery. The part to be replaced can be made with high accuracy using this project.

- The method can also be used by the astronauts up in space. With the help of our approach the parts to be changed can be easily made using a 3D printer.

- Localization at tourist sites and providing real time directions to landmark locations. This will involve the use of GPS (Global positioning system) as well to get a rough location of the user.

8 Budget, duration and facilities

8.1 Budget

Rs. 25,000 will be needed to purchase an android smart phone having high quality sensors and a high resolution camera.

8.2 Duration

We will try to complete this project by the end of the summer break i.e. the end of July, 2015.

8.3 Facilities

- Access to the vision lab.