DYNAMIC HETEROGENEOUS SENSOR REGISTRATION FOR VEHICLE PERCEPTION VIA DEEP NEURAL NETWORKS

Michael Giering, Vivek Venugopalan, Kishore Reddy

United Technologies Research Center
Decision Support and Machine Intelligence Group
411 Silver Lane, East Hartford CT 06108

ABSTRACT

When performing multi-modal fusion to perform an analytic task, spatio-temporal alignment of the incoming signals is often a key to the stability of all methods subsequently applied to the fused data. Lidar-Video systems like those many driverless cars are a common example of where keeping the Lidar and video channels registered to common physical features is important. We develop a deep learning technique that takes multiple channels of heterogeneous data, to test whether we can detect misalignment of the Lidar-video inputs. A number of variations were tested on the Ford LV driving test data set.

1. MOTIVATION

Navigation and situational awareness of optionally manned vehicles requires the integration of multiple sensing modalities such as LIDAR and video, but could just as easily be extended to Radar, SWIR and GPS. Spatio-temporal registration of information from multi-modal sensors is technically challenging in its own right. For many tasks such as pedestrian and other object detection tasks that make use of multiple sensors, decision support methods rest on the assumption of proper registration. Most approaches in LIDAR-video for instance, build separate vision and lidar feature extraction methods and then try to identify common anchor points in both. The ability to dynamically register information from the available data channels for perception related tasks can alleviate this achielles ankel of multi-modal sensor fusion.

Deep neural networks lend themselves in a seamless manner for data fusion on these types of problems. By building a single model on all information being collected at once, it has been shown that features generated on the fused information [][] can provide insight that neither input alone can. From a more applied perspective, it is possible to create such systems with far less overhead. The need for domain experts and hand-crafted feature design are lessened, thereby allowing more rapid prototyping and testing.

The trained nnets easily run within the real-time constraints of common frame rates and lidar data collection.

The generalization of autoregistration across multiple assets is clearly a path to be explored.

By including dynamic flow as input channels, we imbue the nnet with information on the dynamics observed across time steps.

2. PREVIOUS WORK

Need some references to define the state of the art

3. PROBLEM STATEMENT

need detailed description and citation regarding the ford data set need greater clarity on the optical flow method used.

Being able to detect and correct the misalignment among sensors of the same or different kinds is critical when operating on the fused information emanating from the sensors. For this work DCNN's were implemented for the detection of small spatial misalignments in Lidar and Video frames. The data was collected from a driverless car was chosen as the multi-modal fusion test case. LV is a common combination for providing perception capabilities to many types of ground and airborne robots including driverless cars [google, ford].

3.1. Ford LV data set and experimental setup

Detailed description of the ford data set [], our test and training and the justifications for it. a brief description of the hardware used. As shown in diagram-n, we divided the data set into training and testing sections A and B. They were chosen in a manner that minimizes the likelihood of contamination between training and testing. Because of this, the direction of the lighting is source is never the same in the testing and training sets. This provides an additional measure of testing the generalizability of our models.

3.2. Preprocessing

The inputs to our model consisted of C-channels of data with C ranging from 3-6 channels. Channels consisted of inputs that included greyscale and (R,G,B)-video channels, horizontal and vertical optical flow and Lidar depth information. Each channel was cropped to a uniform 800x??? pixels. Each time step has an 800 x ??? x C array of integer values.

These arrays were subdivided into p x p x C patches at a prescribed stride. For any experiment we can denote the preprocessing parameters

- R,G,B — Frame color channels.
- U,V — optical flow channels.
- L — lidar depth channel.
- C — number of input channels.
- p — patch size.
- s — stride.

For a given frame of size 800 x h there are approximately n= (800 x h)/s patches (exact number?). The training and test sets had X and Y frames respectively, therefore the entire data set consists of N = n x X inputs of the patch-size dimension.

Preprocessing is repeated O times, where O is the number of offset classes. For this work we used two setups. A 5 class, linearly distributed set of offsets and a 9 class eliptically distributed set of offsets. (see figure x) For each offset class, **Kishore explain how you generated the data.**

4. MODEL DESCRIPTION

need to describe the parameters, pre-processing, labeling, post-processing classification metric for each patch a table with common params for the experiments would help voting scheme

Our model consists of a CNN classifier that estimates the offset between the LV inputs at each time step. To

5. EXPERIMENTS AND POST-PROCESSING

Need a complete list of the experiments run images to visualize the frame level results please place any confusion matrices and your comments on what you think the results say. feel free to suggest any tables or other visuals to include.

6. CONCLUSION AND FUTURE WORK

7. REFERENCES

populate the papers to be cited in the folder and if possible the bib file

We propose to utilize Deep Convolutional Neural Networks (DCNNs) that have multiple hidden layers for feature extraction and task discrimination enabling perception and decision support for autonomous vehicles. [1] , [2]. The inputs to the model may be from multiple sensors of the same or different kind as well as other non-sensor information. Examples of sensor modalities could include Radar, LIDAR, video, SWIR and GPS.

DCNN's have recently surpassed other state of the art methods in image analysis, audio analysis and other domains. DCNN's are a specific type of *deep neural networks (DNNs)* [3][**?**] that are able to take advantage of local structure in data. These are high dimensional encodings that preserve most information and hence are excellent for tasks such as structure detection and disambiguation.

A DNN is a feedforward artificial neural network that has more than one layer of hidden units between its inputs and outputs. Each hidden unit, j, uses the nonlinear mapping function, often the logistic function, to map its total input from the layer below, x_j, to the scalar state, y_j, that it sends to the layer above,

$$y_j = \frac{1}{1+e^{-x_j}}, \quad x_j = b_j + \sum_i y_i w_{ij},$$

where b_j is the bias of unit j, i is an index over units in the layer below, and w_{ij} is the weight to unit j from unit i in the layer below.

For CNN's convolution is performed at the convolutionsal layers to extract local structure features from the features of the previous layer. Additive bias is applied at this point. This is followed by a local pooling step. A nonlinear mapping (most often a sigmoid) is applied after either the convolution or pooling layer and varies by practioner. Iteratively repeating these convolution and pooling steps results in a CNN architecture [**?**] [**?**].

Image ofD CNN

The value for each spatial point (*x,y*) on the *j*th feature map in the *i*th layer denoted as v_{ij}^{xy} is

$$v_{ij}^{xy} = tanh(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}),$$

where b_{ij} is the bias for this feature map, m indexes over the set of feature maps in the (i-1)th layer connected to the current feature map, w_{ijm}^{pq} is the value at the position *(p,q)* of the kernel connected to the *k*th featuremap, and P_i and Q_i are the height and width of the kernel respectively.

The target layer of the DCNN can be chosen to represent the degree to which sensor information is misaligned. This infomation can in turn be used to properly register sensor data by physical manipulation of the sensors or within the system software.

Methods for improving the accuracy of DCNN's such as dropout may be applied. It is especially useful for applications where the amount of available data is of marginal size to learn the number of necessary parameters in the DCNN. Most consider it analogous to averaging the results over multiple lower dimensional versions of the model.

With ground truth In the event that you have initial data with sensors properly registered, a DCNN can be trained on versions of the data at various known offsets. The input to this model are a matrix representation of two sensors and any supplemental information nodes. (E.g. LIDAR - Video - Optical Flow) A DCNN is created with the standard iterative layers of convolution and pooling, terminating in a soft-max layer for classification of any input sensor pairings as one of the known offsets. The softmax layer enables the user to interpret the offset prediction as a distribution or as a discrete classification.

With NO ground truth Given multi-modal sensor data with no prior knowledge of the system, it is often still possible to register their data streams. For illustration we use the LIDAR-video registration example. Creating an overconstrained DAC (Deep auto-encoder) DEFINE, we can drive the autoencoder to capture mutual information in both the LIDAR and video by reducing the rand of the DAC bottleneck layer well beyond the rank at which optimal reconstruction occurs. Minimizing the reconstruction error with respect to relative shifts of the LIDAR video data reflects that the current alignment of the sensor data has the greatest correlation possible (smallest misalignment). This method can be applied for both spatial and temporal registration.

8. BENEFITS OF THE INVENTION

As the number of channels and modalities of information increase modeling perception systems becomes difficult if not impossible in part due to the large overhead of creating and operating registration methods, especially for real time streaming applications. This invention removes the need for timely expert based feature creation and implicitly generates expressive data features which have been demonstrated to be state of the art in machine learning []. In addition, once trained DCNN's are extremely fast with low memory and computational requirements, making them ideal for real time processing in energy constrained embedded systems. The large volumes of data typically collected in perception applications is very suitable for training the large numbers of parameters deep neural networks such as this have.

DNN's have been shown [] to be able to make more effective use of the information present in the data for discriminative tasks.

9. CURRENT PATENT STRATEGY

We will move forward capturing Deep Learning IP under UTRC. This is in response to very active IP filing in this domain, slow response of BU's to adopt and the likelyhood of being able to license this technology for noncompetitive applications.

A less inclusive submission on Deep Learning for accelerometers has been chosen for patent application by Sikorsky.

The use of deep learning methodologies for feature generation and event detection for PHM is unclaimed IP as far as we know at this time.

The IP landscape in the area of deep neural networks is extremely active. Time is of the essence in protecting this IP if deemed in the interest of UTC.

10. LICENSING OPPORTUNITIES

- Any multi-sensor perception or PHM application
- dynamic registration of multiple sensors (UAV's, satelites) and modalities (multispectral cameras, audio, high frequency time series)
- ability to properly integrate sensor information as they appear or disappear.
- autoregistration of imagery from various platforms across differing spectral bands.
- ability to dynamically temporally shift sensor data to remove time inherent lags between sensors. e.g. Multiple microphones receiving sounds from a single event at different times.

11. REFERENCES

[1] K. Verma, V.K. Gupta, M. Sharma, and R.K. Sevakula, "Intelligent condition based monitoring of rotating machines using sparse auto-encoders," *IEEE*, 2013.

[2] V.T. Tran, F. AlThobiani, and A. Ball, "An approach to fault diagnosis of reciprocating compressor valves using teager-kaiser energy operator and deeep belief networks," *Expert Systems with Applications*, 2014.

[3] G Hinton, L Deng, D Yu, G Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, T Sainath, and K Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Proc. Magazine*, Nov 2012.