Face and body detection in images of crowds using Recurrent Neural Networks

Manali Diwakar Trivedi, student, *Arizona State University*

Abstract—This proposal highlights the objective of using recurrent neural networks in order to detect and highlight faces in images. The speciality about this method is that compared to other classification methods, this method does not require post processing and classifying but it straightaway gives the output by highlighting the faces in test images. Special elements called LSTM (Long Short Term Memory) are used in this method to give the best training and test accuracy. Furthermore, the final aim of this project is to use Gated Recurrent Units in place of LSTMs to compare the final accuracies.

Index Terms- Convolutional Neural Networks, Face Detection, LSTMs.

1 Research Question

This proposal aims to optimize traditional face detection algorithms and I am using [1] as my base paper to implement. The system that I aim to implement uses a combination of convolutional neural networks and a set of recurrent LSTM units that produce variable length outputs [1] and based on the input of previous LSTMS, updation of output boundaries will be done. The LSTMs carry information from previous steps to generate boundaries. The big picture of the project is given in the image Fbelow (Fig 1), taken from [1]

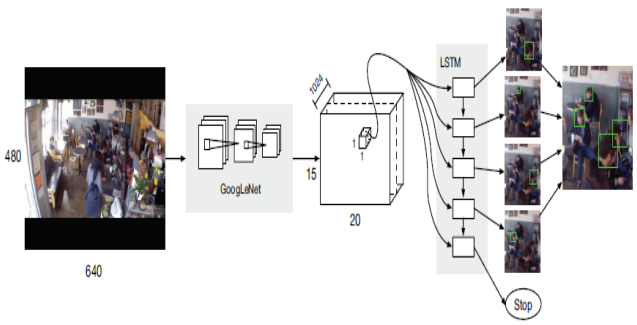

Fig. 1

2 Expected Experiments

The experiments that will be performed to evaluate how well the network is performing, I will use two similar datasets consisting of images of crowds from a public cam, and try different schemes like using one dataset for training and a subset of the second dataset for testing, or even a hybrid dataset consisting of a mix of the types of images in these datasets.
As proposed in [1], I will be using the TUD-Crossing [2] dataset as one of my datasets.

3 Tentative Implementation Plan

The implementation that I will follow is as described in [1]. The first step would be to train the convolutional neural network that would give high level descriptors rich in information about a region. Then, this information is utilized by LSTM units for decoding. The output of each LSTM is fed to the next LSTM and the history of previous stages is fed this way. After this stage, the output consists of bounding boxes and confidence score for each box. The confidence score indicates that a face or body is bounded by that box. The loss function used to calculate gradients, as given in [1] is:

Given f, we define a loss function on pairs of sets G and C as

$$L(G, C, f) = \alpha \sum_{i=1}^{|G|} l_{pos}(\mathbf{b}_{pos}^{i}, \tilde{\mathbf{b}}_{pos}^{f(i)}) + \sum_{j=1}^{|C|} l_c(\tilde{\mathbf{b}}_{c}^{j}, y_j) \tag{1}$$

Fig. 2

where $l_{pos} = \|\mathbf{b}_{pos}^{i} - \tilde{\mathbf{b}}_{pos}^{f(i)}\|_1$ is a displacement between the position of ground-truth and candidate hypotheses, and l_c is a cross-entropy loss on a candidate's confidence that it would be matched to a ground-truth. The label for this cross-entropy loss is provided by y_j. It is defined from the matching function as $y_j = \mathbb{1}\{f^{-1}(j) \neq \varnothing\}$. α is a term trading off between confidence errors and localization errors. We set $\alpha = 0.03$ with cross validation. Note that for

Finally, after convergence, the bounding boxes are stitched together using the bipartite matching problem [4]. Additionally, my contribution to this would be to modify this algorithm. As long as I manage to complete the implementation of the algorithm as mentioned in the base paper, the modification to it would be to substitute LSTMs with Gated Recurrent Units (GRU) [3] and evaluate the resulting network. As explained in [3], a GRU would expose the earlier states entirely without needing to use memory units unlike LSTMs which use memory units to control access to information.

4 TIMELINE OF DEVELOPMENT

Since I am the only team member, I plan to consistently work on the elements of this project. The division of this project is as follows in Fig. 3:

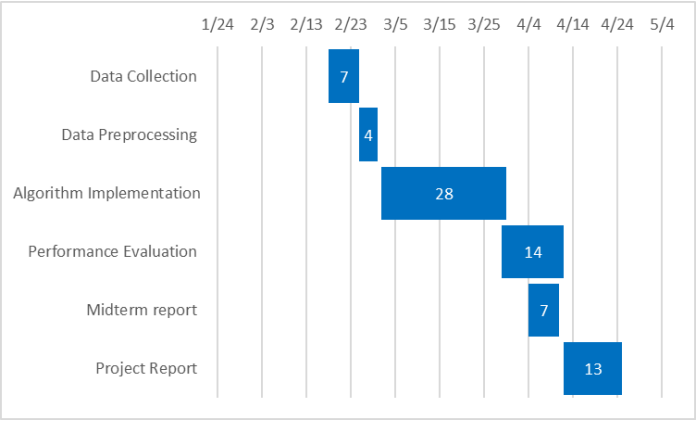

Fig. 3

REFERENCES

[1] Russell Stewart and Mykhaylo Andriluka. End-to-end people detection in crowded scenes. arXiv preprint-arXiv:1506.04878, 2015.

[2] M. Andriluka, S. Roth, and B. Schiele. People-tracking-bydetection and people-detection-by-tracking. In *CVPR 2008*.

[3] Chung, Junyoung, et al. "Empirical evaluation of gated recur rent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

[4] Karp, Richard M., Umesh V. Vazirani, and Vijay V. Vazirani. "An optimal algorithm for on-line bipartite match-ing." *Proceedings of the twenty-second annual ACM symposium on Theory of computing*. ACM, 1990.