What problem is the paper investigating?

* The paper is attempting at raising the bar of the current state-of-the-art onset detection algorithm, which was based on an RNN (Recurrent Neural Network) model.
* Instead of RNN, the author used CNN (Convolutional Neural Network) model, which is typically used in Computer Vision, for feature detection, such as edge detection. There are overlap between detecting edges in images and detecting vertical edges in spectrograms for onsets. Thus, this technique is being investigated.
  + Each input is consisted of three spectrogram fragments computed with the same hop size (10ms) but different window sizes. This enables the input to have both high time resolution and frequency resolution while maintaining the same frame rate per spectrogram.
  + These spectrograms are then converted to mel-spectrograms, using an 80-band Mel filter. After that, they are normalized so they all have zero mean and unit various for CNN computations.
  + Inside the CNN, the 15 x 80 x 3 (frame \* frequency bands \* channel) are being convoluted with ten 7(frame) \* 3(frequency band) \* 3(channel) filters to form ten (9 \* 78) second hidden layer of inputs, the frequency bands are max-pooled over 3 adjacency bands to form ten 9 \* 26 feature maps. These are then convolved with a 3 \* 3 filter and max-pooled by another 3 \* 1 filter to form twenty 7 \* 8 feature maps. These outputs are finally fed into a fully connected neural network with 256 sigmoid units to form a final classification output. For each convolution layer, the outputs are activated by a tanh function before feeding into the next layer.
  + The dataset used to compute the spectrograms are the same data used in this paper: “S. Böck, F. Krebs, and M. Schedl, “Evaluating the online capabilities of onset detection methods,” in *Proc. of the 13th Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, Porto, Portugal, Oct. 2012.
    - It consists of 102 minutes of music annotated with 25,927 onsets that contains monophonic, polyphonic instrumental recordings as well as popular music excerpts.
  + What they were able to achieve using the method is a 90.3 % f-score, after adjusting the activation function from tanh to ReLu within the convolutional layers. The previous state-of-the-art RNN model achieves a f score of 87.3%
* Later, when they were trying to understand what the model is actually doing, they found out that the model kind of used two different sets of parameters to recognize pitched and percussive onsets. However, it’s hard to replicate the result by using such insights to hand design algorithms for onset detections, as the key reason the model works is still the numerous iterations of parameter changes that are carefully calibrated by the algorithm.
* The paper concluded that machine learning can not only be helpful towards more abstract MIR tasks, but also can outperform traditionally hand designed lower level feature extraction algorithms.
* Limitations and Improvements:
  + There are some discussions about utilizing bagging and dropouts, and widening the final selection sizes to improve results. However, there are not attempts at experimenting with different network architectures, such as changing the convolutional layers, or using a different filter size, adding fully connected layers. The experiment can also probably benefit from data from using other sources to train the model with.