Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks

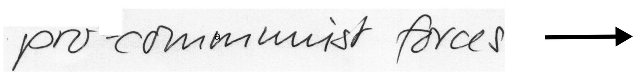
Alex Graves[1] and Jürgen Schmidhuber[1,2]

[1] Technical University of Munich, [2] IDSIA, Lugano

graves@in.tum.de, juergen@idsia.ch

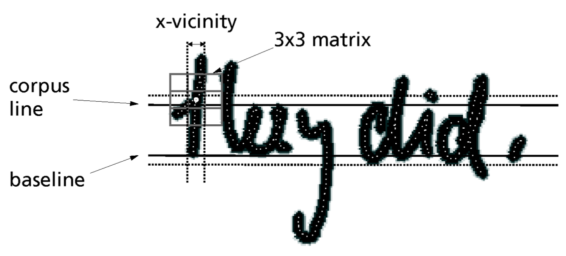
Introduction

Offline handwriting recognition is the automatic transcription of images of handwritten text.



pro-communist forces

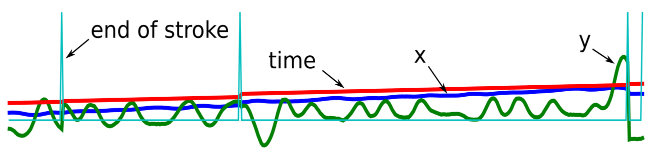
It is usually done by first extracting image features



Then feeding them to a standard sequence labelling algorithm, such as an HMM. However there are several drawbacks to this approach:

- It requires hand designed features for every alphabet
- In the case of HMMs, the features must meet stringent independence assumptions
- The system cannot be globally trained

We previously designed a globally trained, alphabet independent, recurrent network architecture for *online* handwriting recognition.



Once having

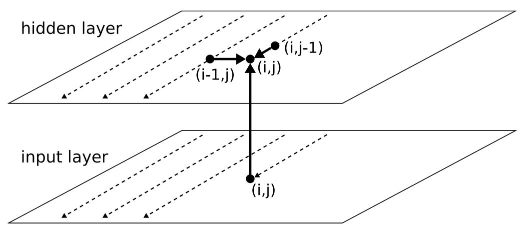
To extend this architecture to *offline* handwriting, we needed to adapt from 1D to 2D input data. Our solution combines three innovations in recurrent network design:

- **Multidimensional recurrent networks**, to scan through the images vertically and horizontally
- **Hierarchical structure**, to incrementally transform the 2D feature maps into a 1D sequence
- **Connectionist temporal classification**, to label the top level sequence without prior segmentation.

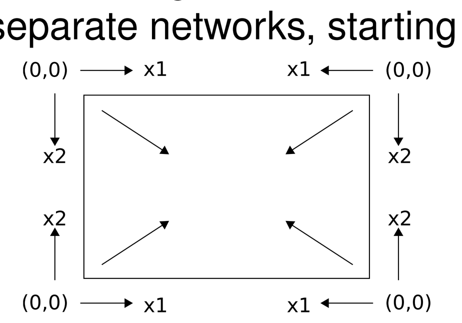
The **complete system** is shown at the bottom of the page.

Multidimensional Recurrent Networks

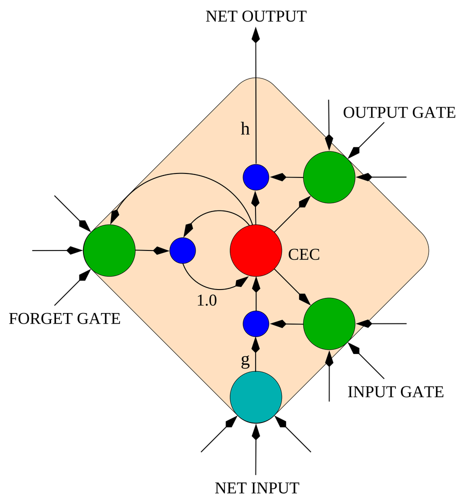
The basic idea of multidimensional recurrent neural networks (MDRNNs) is to replace the single recurrent connection in a standard RNN by a separate connection for each dimension in the data.



This gives the network access to multidimensional context, and makes it robust to local distortions that 'mix' dimensions, such as shears, rotations etc. To get context from all directions, we scan through each n-dimensional data sequence with 2^n separate networks, starting in every corner.



Multidimensional LSTM

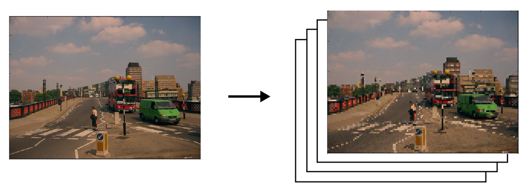


Long Short-Term Memory (LSTM) is a recurrent architecture that uses a linear memory unit surrounded by multiplicative gates to bridge long delays between input events. We have extended LSTM to multidimensional data, thereby giving access to long range context in all input directions.

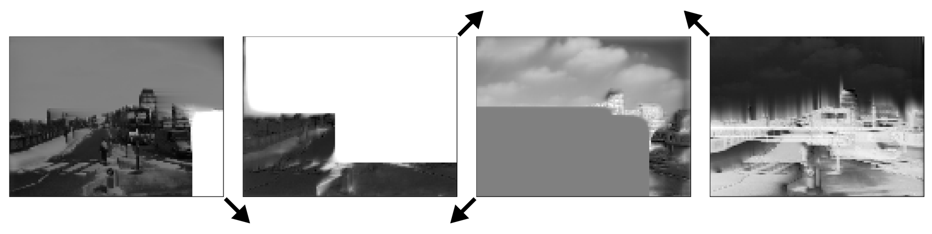
Hierarchical Structure

Hierarchies are often used in computer vision to build complex, high level features out of simple local features in an incremental fashion. We create a hierarchy of MDRNNs with the following steps:

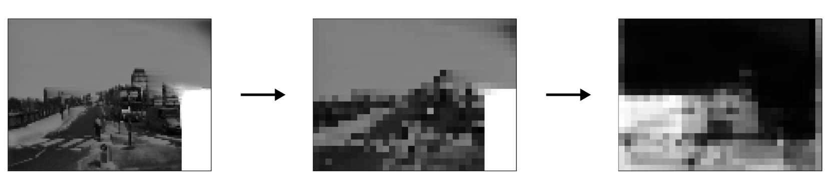
1. The input pixels are gathered together into blocks



2. MDRNNs scan through the resulting 'block' images in all four directions



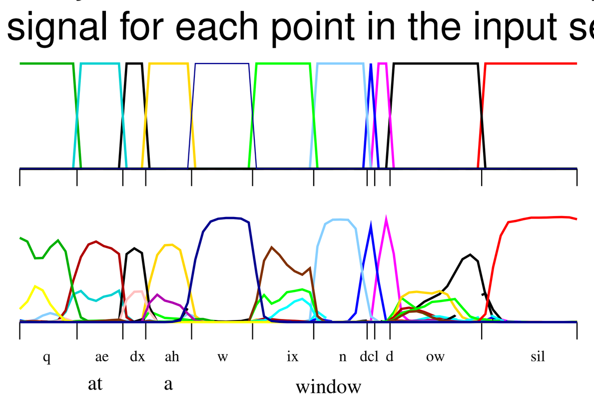
3. The MDRNN output images are gathered into blocks and fed to a feedforward layer



This process is repeated as many times as needed, with the activation of the feedforward layers providing input images for the next level up. Each iteration decreases the effective resolution and increases the number of features. The end result is a 1D sequence of feature vectors that can be labelled by the output layer.

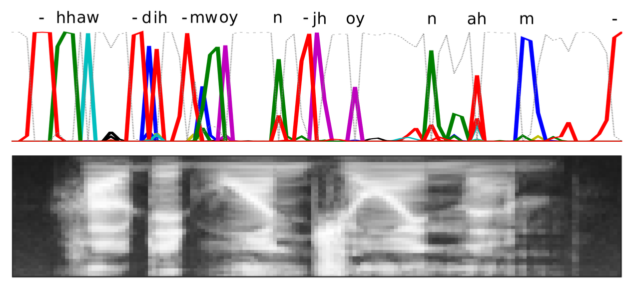
Connectionist Temporal Classification

The standard objective functions for RNNs require a separate training signal for each point in the input sequence.



This is a problem for tasks like cursive handwriting recognition, where the labels are hard to segment.

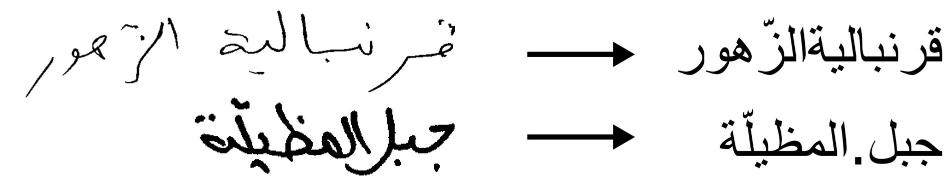
Our system uses a connectionist temporal classification (CTC) output layer. CTC was first applied to speech recognition.



CTC does not require prior segmentation because the network is free to emit the labels at any time, as long as their order is correct. It also allows the labellings to be read straight from the network outputs (follow the spikes).

Arabic Handwriting Recognition

An international Arabic handwriting recognition competition was held at the ICDAR 2007 conference. The goal was to identify Tunisian place names from handwritten forms.

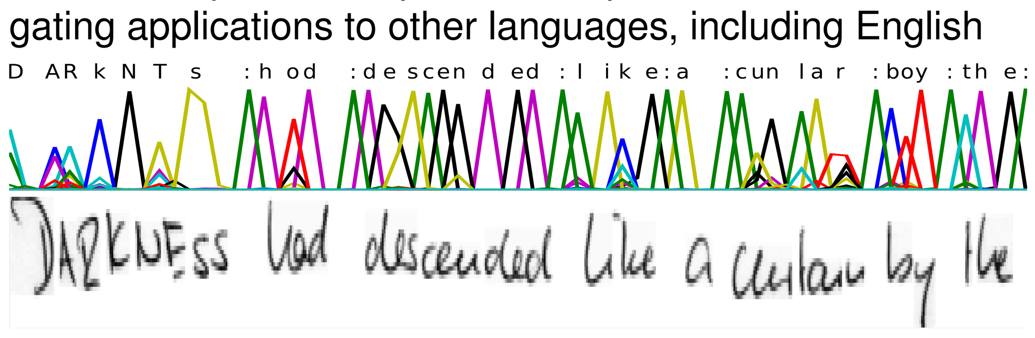


Our system outperformed all entries in the competition, even though we don't know any Arabic.

System	Set f			Set s		
	top 1	top 5	top 10	top 1	top 5	top 10
CACI-3	14.28	29.88	37.91	10.68	21.74	30.20
CACI-2	15.79	21.34	22.33	14.24	19.39	20.53
CEDAR	59.01	78.76	83.70	41.32	61.98	69.87
MITRE	61.70	81.61	85.69	49.91	70.50	76.48
UOB-ENST-1	79.10	87.69	90.21	64.97	78.39	82.20
PARIS V	80.18	91.09	92.98	64.38	78.12	82.13
ICRA	81.47	90.07	92.15	72.22	82.84	86.27
UOB-ENST-2	81.65	90.81	92.35	69.61	83.79	85.89
UOB-ENST-4	81.81	88.71	90.40	70.57	79.85	83.34
UOB-ENST-3	81.93	91.20	92.76	69.93	84.11	87.03
SIEMENS-1	82.77	92.37	93.92	68.09	81.70	85.19
MIE	83.34	91.67	93.48	68.40	80.93	83.73
SIEMENS-2	87.22	94.05	95.42	73.94	85.44	88.18
Ours	**91.43**	**96.12**	**96.75**	**78.83**	**88.00**	**91.05**

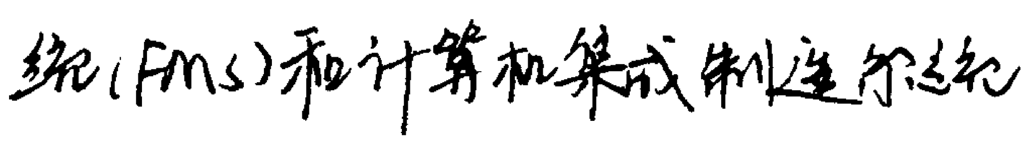
Further Work

Since our system is alphabet independent, we are investigating applications to other languages, including English



And Chinese.

统 (FMS)和 计 算 机 集 成 制 造 系 统



Indeed, since the input can have any number of space time dimensions, the system could be applied to virtually any sequence labelling task, e.g. gesture recognition from video, classification of fMRI sequences etc.

Complete System