A CNN Based Framework for Unistroke Numeral Recognition in Air-Writing

Prasun Roy, Subhankar Ghosh, and Umapada Pal
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute
203, B. T. Road, Kolkata, WB - 700108, India

Motivation

Air-writing refers to virtually writing text through hand gestures in three dimensional space. Existing approaches to build air-writing recognition systems use depth and motion sensors such as **Kinect**, **Leap Motion** and **Myo Armband**. As these sensors are not widely available in commonly used devices, a generic video camera based approach can be highly beneficial. In this paper a generic video camera dependent CNN based air-writing recognition system is proposed.

Proposed Approach

A. Marker Segmentation

$$g(x,y) = \begin{cases} 1, & \text{if } f(x,y) = I_m \\ 0, & \text{otherwise} \end{cases}$$

Assuming ***f(x, y)*** and ***g(x, y)*** to be pixel values at position ***(x, y)*** of the originally captured video frame and segmented frame and I_m being the threshold for segmentation.

B. Marker Tip Identification

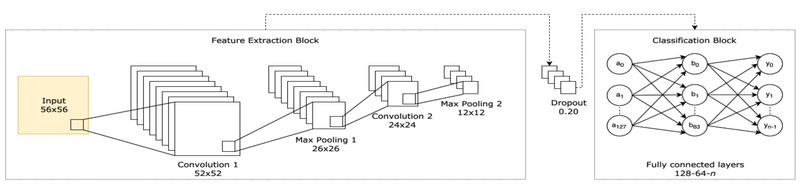
The contour with largest area in the segmented image is assumed as the marker. Tip of the marker is estimated as the top-most point on this contour boundary with lowest ***y***-coordinate value.

C. Trajectory Approximation

$$N_{FPS} = \frac{1}{t_{update}} \qquad dx = \frac{1}{N_{FPS}} \sum_{t}^{t+N_{FPS}} \Delta_x \qquad dy = \frac{1}{N_{FPS}} \sum_{t}^{t+N_{FPS}} \Delta_y$$

Where t_{update} is the time required to process the last video frame. Assuming Δ_x and Δ_y as changes in position of marker tip along ***x*** and ***y*** direction respectively between two consecutive frames. The starting and ending of continuous trajectory during air writing is decided by comparing ***dx*** and ***dy*** with a empirically proposed velocity threshold V_T. When both ***dx*** and ***dy*** are below V_T the marker is assumed to be static (pen-up state). Otherwise the marker is assumed to be in motion (pen-down state).

D. Character Recognition



CNN architecture used for the proposed method.

The architecture of the employed network includes a feature extraction block followed by a classification block. The feature extraction block takes a 56X56 grayscale image as input and it consists of two convolution layers each followed by a pooling layer. The first convolution layer uses 32 convolution filters and a 5X5 convolution kernel. The second convolution layer uses 16 convolution filters and a 3X3 convolution kernel. Both convolution layers use rectified linear units (ReLU) as activation function. The classification block consists of three fully connected layers having 128, 64 and n computing units (neurons) respectively where n is the numbers of output classes corresponding to the character set under consideration. The first two fully connected layers use ReLU as activation function while the final layer uses normalized exponential function (softmax) for the same purpose. To reduce the possibility of overfitting during training, dropout technique is employed between the two blocks.

Dataset

DATASET DISTRIBUTION FOR ENGLISH NUMERALS

Dataset	Type	Source Dataset	#Instance
TS-A	Training Set A	Air-Writing	6000
TS-B	Training Set B	MNIST	70000
EVAL	Test Set	Air-Writing	4000

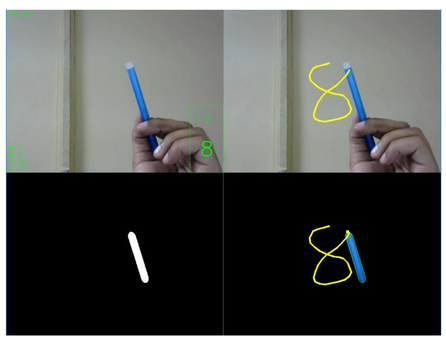
DATASET DISTRIBUTION FOR BENGALI NUMERALS

Dataset	Type	Source Dataset	#Instance
TS-A	Training Set A	Air-Writing	6000
TS-B	Training Set B	Bengali	14650
EVAL	Test Set	Air-Writing	4000

DATASET DISTRIBUTION FOR DEVANAGARI NUMERALS

Dataset	Type	Source Dataset	#Instance
TS-A	Training Set A	Air-Writing	6000
TS-B	Training Set B	Devanagari	22546
EVAL	Test Set	Air-Writing	4000

Results



Top-Left: Original video frame. **Top-Right:** Original video frame with approximate marker trajectory overlay. **Bottom-Left:** Segmentation mask. **Bottom-Right:** Segmented marker and approximate marker trajectory.

TEST ACCURACY FOR ENGLISH NUMERALS

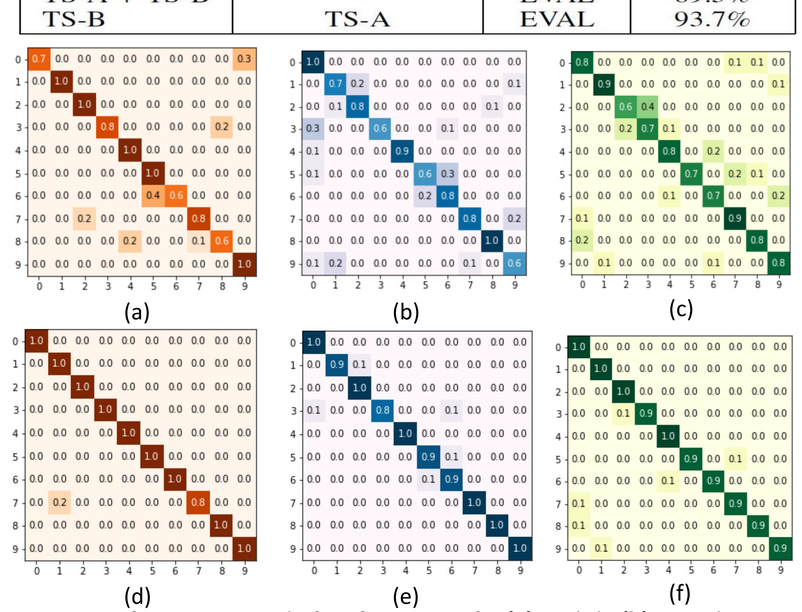
Training Set	Fine-Tuning Set	Test Set	Accuracy
TS-A	–	EVAL	81.8%
TS-B	–	EVAL	86.4%
TS-A + TS-B	–	EVAL	95.5%
TS-B	TS-A	EVAL	97.7%

TEST ACCURACY FOR BENGALI NUMERALS

Training Set	Fine-Tuning Set	Test Set	Accuracy
TS-A	–	EVAL	78.1%
TS-B	–	EVAL	81.1%
TS-A + TS-B	–	EVAL	92.5%
TS-B	TS-A	EVAL	95.4%

TEST ACCURACY FOR DEVANAGARI NUMERALS

Training Set	Fine-Tuning Set	Test Set	Accuracy
TS-A	–	EVAL	77.3%
TS-B	–	EVAL	82.4%
TS-A + TS-B	–	EVAL	89.5%
TS-B	TS-A	EVAL	93.7%



Confusion matrices before fine-tuning for **(a)** English, **(b)** Bengali **(c)** Devanagari and after fine-tuning for **(d)** English, **(e)** Bengali, **(f)** Devanagari

Conclusion

Here a generic video camera based air-writing recognition system is proposed and **97.7%**, **95.4%** and **93.7%** recognition rates over English, Bengali and Devanagari numerals, respectively are achieved. This method is fully independent on any depth or motion sensor such as Kinect, LEAP Motion, Myo etc., which is the main advantage of this framework over existing approaches.