Visual Speech Recognition using Hahn Convolutional Neural Networks

Hicham Hammouchi, *TICLab, Université Internationale de Rabat, Morocco* | *hicham.hammouchi@uir.ac.ma*

Introduction

- Lip reading is reading speech from speaker's lips motions.
- Lip reading has many applications such as in medical field, however, it is a challenging computer vision problem.
- Hahn Convolutional Neural Network to visually recognize speech efficiently and with less computation resources.
- Leverage Hahn moments to extract features and perform the recognition with CNN.
- Accurate visual speech recognition can have many implications such as for laryngectomized persons.

Discrete Orthogonal Hahn Moments

- Hahn moments are a set of orthogonal moments based on Hahn polynomials defined on the image coordinates space.
- Discrete orthogonal Hahn moments are descriptors that can extract the main characteristics from image at low orders.
- Hahn Polynomials formula [1]:

  For any integer $x \in [0, N-1] > 0$, Hahn polynomial of order $n, n = 0,1,\cdots,N-1$, is defined as:

  $$h_n^{(\alpha,\beta)}(x,N) = (N+\beta-1)_n(N-1)_n \times \sum_{k=0}^{n}(-1)^k \frac{(-x)_k(-n)_k(2N+\alpha+\beta-n-1)_k}{(N+\beta-1)_k(N-1)_k}\frac{1}{k!}$$

  where $(a)_k = a\cdot(a+1)\cdots(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)}$ is the Pochhammer symbol.
- Hahn moments of order $(n+m)$ of an image with dimensions $N \times M$ is given as follow [1]:

  $$H_{nm} = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1} h_n^{(\alpha,\beta)}(x,N)h_m^{(\alpha,\beta)}(y,N)f(x,y)$$

  where $f(x,y)$ is the image matrix.
- Figure (1) illustrates the polynomials generated for $N = 100, \alpha = \beta = 5$, and $order = 12$ (from 1 to 12)

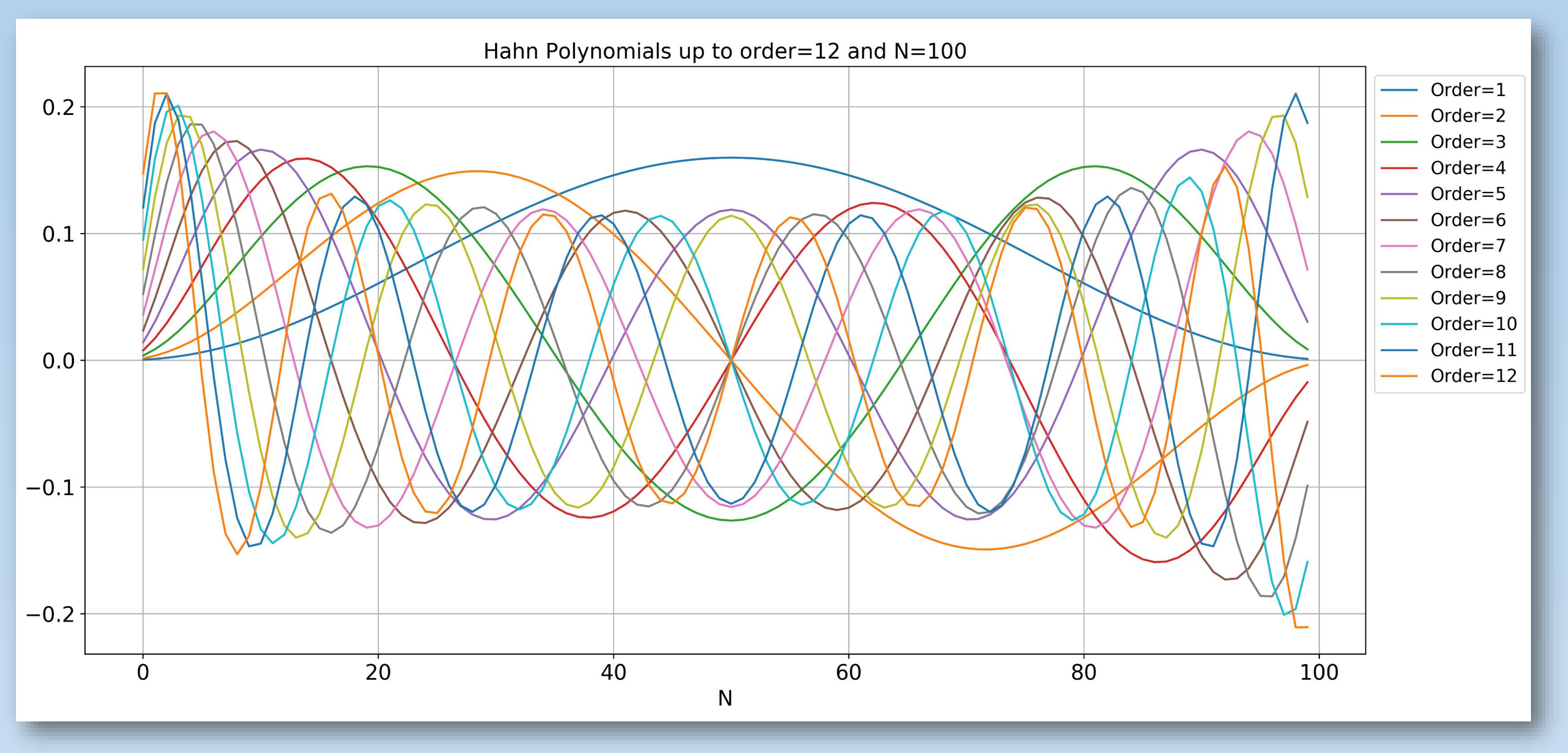
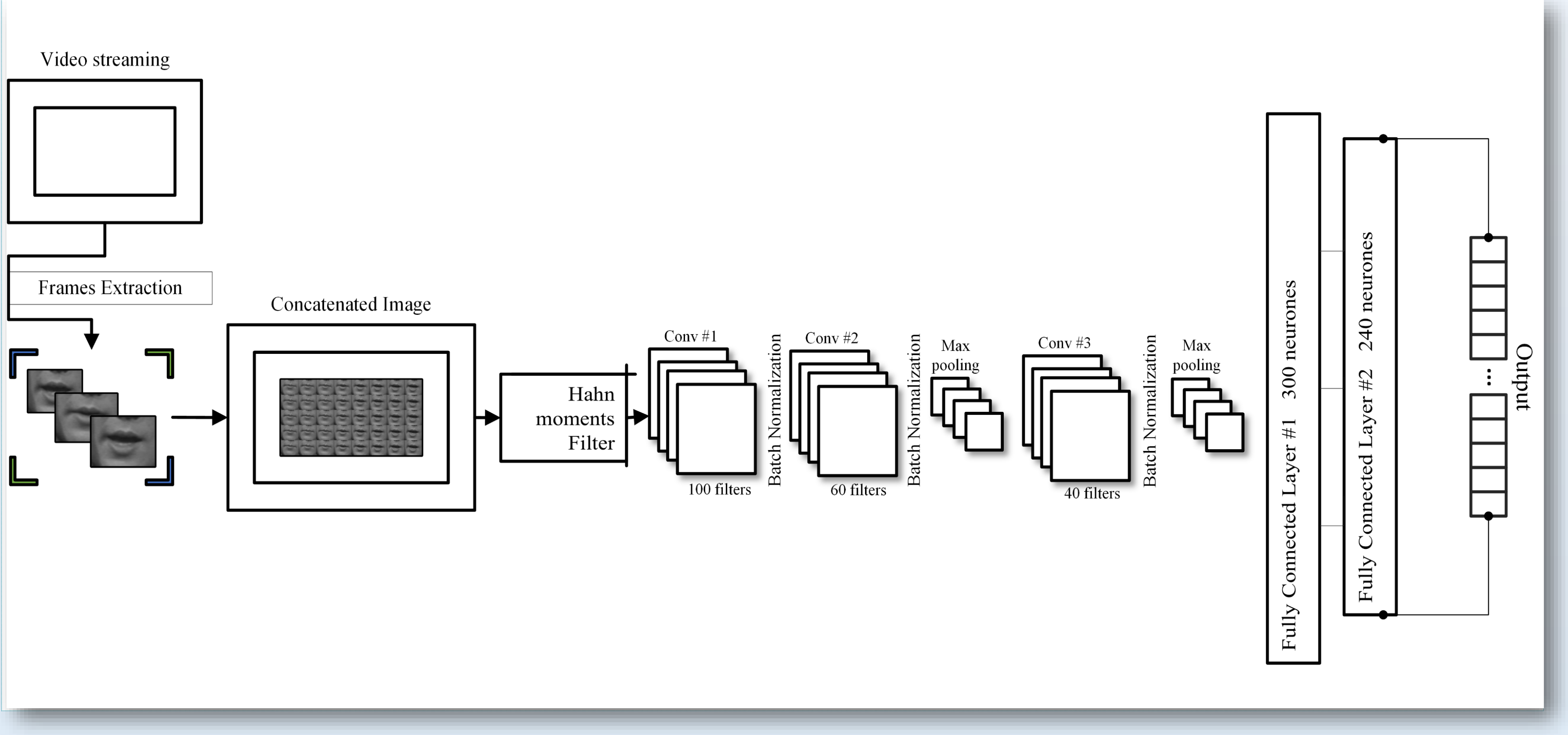


Figure (1): Hahn Polynomials

HCNN Architecture



Data

OuluVS2 Digits sequences [2]: 52 speakers uttering 10 digits sequences:

"1 7 3 5 1 6 2 6 6 7", "4 0 2 9 1 8 5 9 0 4", "1 9 0 7 8 8 0 3 2 8", "4 9 1 2 1 1 8 5 5 1", "8 6 3 5 4 0 2 1 1 2", "2 3 9 0 0 1 6 7 6 4", "5 2 7 1 6 1 3 6 7 0", "9 7 4 4 4 3 5 5 8 7", "6 3 8 5 3 9 8 5 6 5", "7 3 2 4 0 1 9 9 5 0".

Results

Original image size: 800x550

Method	Accuracy
HCNN (order 12)	74.33%
HCNN (order 16)	80.05%
HCNN (order 32)	88.72%
HCNN (order 44)	91.94%
HCNN (order 56)	**93.72%**
HCNN (order 60)	92.66%
CNN	42.27%

Conclusion

- Hahn moments retain the most characteristics of the image and reduce significantly the computation requirements.
- HCNN yields good results with a shallow architecture.
- HCNN can be used efficiently to handle the problem of lip reading and other computer vision problems.

References

[1] Zhou J., Shu H., Zhu H., Toumoulin C., Luo L. (2005) Image Analysis by Discrete Orthogonal Hahn Moments. In Image Analysis and Recognition. ICIAR.

[2] Anina I., Zhou Z., Zhao G., and Pietikäinen M. (2015) OuluVS2: A multi-view audiovisual dataset for non-rigid mouth motion analysis. In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition.

First IndabaX Morocco 2019