LIP READING BY LEVERAGING HAHN CONVOLUTIONAL NEURAL NETWORK IN LOW-RESOURCED ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Lipreading or Visual speech recognition is the process of decoding speech from speaker's mouth movements. It is used for people with hearing impairment, to understand patients attained with laryngeal cancer, people with vocal cord paralysis and in noisy environment. In this paper we present a novel architecture called **HCNN** based on Hahn moments as first layer in the Convolutional neural network architecture, to tackle the problem of visual speech recognition. Because of the high computational cost of the standard CNN and its time consuming training, we propose HCNN to reduce the dimensionality of images while preserving the main characteristics to gain training time and to improve the CNN capabilities in extracting features and patterns with Hahn moments. HCNN as a small architecture is cost-effective and yields an outstanding performance on images. We report results on OuluVS2 dataset, and HCNN succeed to outperform the existing works in the literature with an accuracy of 93.72%. In addition, HCNN proves its effectiveness for the visual speech recognition and it may be used in other applications. literature.

1 INTRODUCTION

Lipreading is a challenging process for humans especially when the context is absent. It requires special qualities for experts to follow lips movements, tongue articulations and teeth. Another confusing issue is the similarity between phonemes explained by Fisher in 1968 Fisher (1968). Additionally, the differences between each speaker's mouth shape, mustache, or the effect of makeup can make the task of lipreading more complicated. To face these issues a robust lipreading system is needed to differentiate all these variations. In the comparison between human and machine lipreading performance conducted by Hilder et al. in 2009 Hilder et al. (2009), the experiments showed that machine lipreading has outperformed the human lipreading and therefore an automated lipreading system is indispensable to solve the issue.

Toward building an automated lipreading system, several approaches were proposed and tested on several datasets, especially on AVLetters dataset Matthews et al. (2002), for example an approach using Active Shape Models (ASM) Cootes et al. (1995) and Active Appearance Models (AAM) Cootes et al. (1998) was conducted by Matthews *et al.* in 2002 [5] to extract features from lips images, and train a model using Hidden Markov model (HMM) classifier, this method obtained 44.6% accuracy. Zhao *et al.* in 2009 Zhao et al. (2009) proposed a lip-reading method using local spatiotemporal descriptors, in which they represented the isolated phrase sequences by extracting the spatiotemporal local binary patterns (LBP) from mouth region. The best performance attained was 58.85% using Support Vector Machine (SVM) classifier. A method based on Deep bottleneck features extraction directly from pixels was introduced by Petridis and Pantic, 2016 Petridis & Pantic (2016), where the authors trained a model using Long-Short Term Memory (LSTM), this method achieved 58.1% accuracy. Bakry and Elgammal in 2016 Bakry & Elgammal (2016), conducted a comparison between manifold kernels in Manifold Kernel Partial Least Squares(MKPLS). Their approach consists of using distances such Euclid distance between images and LBP to extract features. Another method proposed by Tian and Weijun Tian & Ji (2017), in which they introduced an auxiliary multimodal LSTM (am-LSTM) that aims to combine audio-visual data at the same time. It learns from both audio and video modalities and uses a pre-trained VGG-16 model to extract features and PCA to reduce dimensionality. On cross modality protocol, which means the audio and video are used for training and only the video is used for testing, the performance obtained was 88.83%. As for the OuluVS2 dataset, it was first proposed by Annie *et al.* in 2015 Anina et al. (2015) to address the problem of non-rigid mouth motion analysis. The provided baseline performance was 41% accuracy on the frontal view. On the same dataset Joon Son Chung and Andrew Zisserman conducted in Chung & Zisserman (2016) a method called SyncNet to synchronize mouth motion and speech in a video. The proposed model is a mixture of LSTM and CNN, where the LSTM model ingests the visual features produced by the CNN image by image until the end of the sequence. The model was 92.8% accurate on OuluVS2 fixed digits. Further, Saitoh *et al.* in Saitoh et al. (2016) propose a method called CFI-based CNN to represent the spatiotemporal aspect in the video for visual speech recognition. They evaluated the method on OuluVS2 dataset and the performances achieved on the OuluVS2 digits (frontal view) were 61.7% using Netwotk in Network (NiN) model with data augmentation (DA), and an accuracy of 89.4% with DA and using GoogLeNet model.

In this paper we propose a novel architecture called HCNN based on Hahn moments and convolutional neural network (CNN). The new architecture lies on the mixture of Hahn moments with its ability to hold and extract the most useful information in images with effectiveness and less redundancy, and the performance of the convolutional neural networks in learning pattern and image classification. The Hahn moments are used as the first layer of our architecture to extract the moments and feed them to the CNN. To the best of our knowledge this is the first time, moments will be used as a filter in a CNN architecture applied to lipreading. Hahn moments were chosen over other discrete orthogonal moments like chebyshev and Krawtchouk moments, because Hahn moments cover all the properties of both moments, and because of their great ability to represent image with less redundancy in the amount of information. Furthermore, they can be parameterized to retain the global or local characteristics of the image in the lowest orders. In this work we propose a solution that encompasses several issues. The main contributions of this paper are: 1) further improve the performance of the CNN architecture and customize it for a better features extraction and better patterns learning, 2) reduce significantly the dimensionality of images by integrating the Hahn moments as first layer which leads to decrease the computational cost, 3) present a cost-effective solution to the Lip reading problem.

2 Two-dimensional Hahn Moments

Hahn moments are a set of orthogonal moment based on the discrete Hahn polynomials defined over the image coordinate space. Their implementation does not involve any numerical approximation. In this section, we will give brief formulation of 2D weighted Hahn moments including polynomials and we will describe their capacity to capture significant features from image with significant reduce of dimensionality.

2.1 Hahn polynomials

Hahn polynomials of one variable x, with the order n, defined in the interval $[0, N-1]$ as given in Zhou et al. (2005), respect the following equation:

$$h_n(\alpha,\beta,N|x) = {}_3F_2\left(\begin{array}{c} -n, n+\alpha+\beta, -x \\ \alpha+1, -N \end{array}\middle|\, 1\right) \tag{1}$$

with $n, x = 0, 1, \cdots, N-1$
where $\alpha\, and\, \beta$ are free parameters, and ${}_3F_2$ is the generalized hyper-geometric function given by :

$${}_3F_2\left(\begin{array}{c} a_1, a_2, a_3 \\ b_1, b_2 \end{array}\middle|\, z\right) = \sum_{k=0}^{\infty} \frac{(a_1)_k (a_2)_k (a_3)_k}{(b_1)_k (b_2)_k k!} z^k \tag{2}$$

Hahn polynomials satisfy the orthogonal property:

$$\sum_{x=0}^{N-1} h_n\left(\ \alpha,\beta,N \mid x\ \right) h_m\left(\ \alpha,\beta,N \mid x\ \right) \omega_h(x) = \rho_h(n)\delta_{mn} \tag{3}$$

where $w_h(x)$ is the weighting function given by

$$\omega_h(x) = \frac{(\alpha+1)_x(\beta+1)_{N-x}}{(N-x)!x!} \tag{4}$$

while ρ_h is the squared-norm expressed by

$$\rho_h(n) = \frac{(-1)^n n!(\beta+1)_n(\alpha+\beta+n+1)_{N+1}}{(-N)_n(2n+\alpha+\beta+1)N!(\alpha+1)_n} \tag{5}$$

3 Proposed Architecture HCNN

The novel architecture HCNN as shown in fig. 1 aims to solve the problem of Lip reading by processing and recognizing lips images rapidly and efficiently. It is a combination of the method of discrete orthogonal moments and the convolutional neural network.

The HCNN architecture comes to surmounts the high computational costs and the sophisticated hardware resources required by the CNN. Furthermore, HCNN enhance the quality of features extraction and the assimilation of patterns incorporated in the image. Indeed, the Hahn moments as a powerful descriptors to retrieve the most useful information in the image, with the property of covering global, local and both features at the same time with efficiency. This advantage can be achieved by tuning the suitable values of α and β parameters as detailed in the Hahn moments section above. Based on the study conducted in Mesbah et al. (2016) we have set these parameters to $\alpha = \beta = 5$, so the moments retrieved can encompass the whole image with the potential to apprehend its global features.

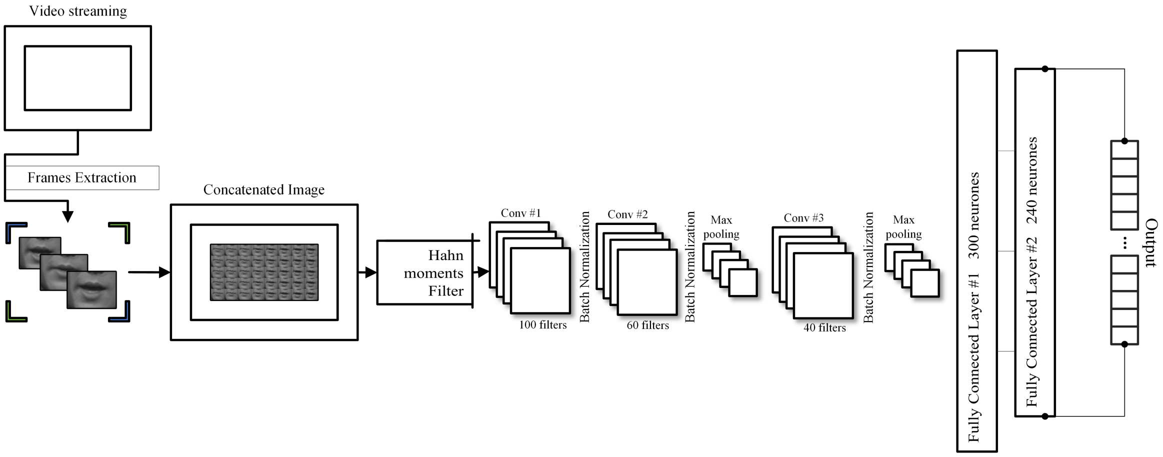

Figure 1: HCNN model parameters: Hahn moments up to desired order, first convolution (kernel 3x3 and 100 filters), second convolution (kernel 3x3 and 60 filters), first max pooling (pool size 3x3). Third convolution (kernel 3x3 and 40 filters). Second max pooling (pool size 3x3). First fully connected layer (300 neurons), and a second layer (240 neurons)

4 Experiments & Results

4.1 Datasets

In order to evaluate the classification performance of the proposed architecture (HCNN), we conduct experiments on two lip reading datasets, nemely, AVLetters and OuluVS2.

- **AVLetters:** a dataset that contains 780 videos for 10 speakers, every speaker utters the 26 alphabet letters three times, which results in 78 videos for each speaker's mouth. each frame is of a dimension of 80x60.

- **OuluVS2:** a dataset that contains 52 speakers uttering 10 digits sequences with three repetitions each. The dataset is provided with cropped mouth region and with multiple views. In our experiments we use the frontal view, by resizing the extracted images to 50×50.

4.1.1 Data Augmentation

In order to conduct a fair comparison with other works in the literature, especially those who worked on OuluVS2 dataset, we perform a data augmentation (DA) by applying rotations with angle degrees [-15, -10, 10, 15], on each frame of each video.

4.2 Results and Discussion

The recognition rate of our model in comparison with the previous works on AVLetters dataset are shown in table 3. Our method clearly perform better than the methods compared to, which shows the effectiveness of using Hahn moments to capture the global features although we use a simple CFI to represent a whole sequence. Indeed, using Hahn moments we achieve 20% absolute improvement over CNN. As for the OuluVS2 fixed digits dataset we report in table 2 two works for comparison, SyncNet Chung & Zisserman (2016) and CFI-based CNN Saitoh et al. (2016), where the first employ both CNN and LSTM for recognition in a speaker independent (SI) manner, while the CFI-based CNN lies on a frames concatenation method for modeling the sequences and uses very deep pre-defined CNN architecture such as GoogleNet, AlexNet and Network in Network (NIN) for the recognition task. It can be clearly seen that our shallow HCNN model outperforms the two related works in terms of classification accuracy and reduces enormously the complexity. Similarly to AVLetters dataset, adding Hahn moments filter achieves over than 50% improvement over CNN. Furthermore, in our experiments with only rotation data augmentation we can achieve better than CFI-based CNN and SyncNet, in which, extensive data augmentation like translation, rotation, flipping and color shift were used.

Table 1: Obtained results on AVLetters with different Hahn moments orders

Order	16	32	52	56	60	64	72
Accuracy	49.61%	53.41%	**59.23%**	55.76%	56.63%	57.69%	56.15%

Table 2: Obtained results on OuluVS2 fixed digits with different Hahn moments orders using SI protocol with DA

Order	12	16	32	44	56	60
Accuracy	74.33%	80.05%	88.72%	91.94%	**93.72%**	92.66%

Table 3: Obtained results on AVLetters in comparison with other methods

Method	Accuracy
HMM Matthews et al. (2002)	44.6%
LBP-SVM Zhao et al. (2009)	58.85%
LSTM Petridis & Pantic (2016)	58.1%
HCNN	**59.23%**
CNN Without Hahn	39.23%

Table 4: Obtained results on OuluVS2 Digits (frontal view) in comparison with other methods

Method	Accuracy
CFI-based CNN (GoogLeNet) +DA Saitoh et al. (2016)	89.4%
SyncNet (CNN+LSTM) +DA Chung & Zisserman (2016)	92.8%
HCNN +DA (SI)	**93.72%**
CNN Without Hahn +DA (SI)	42.27%

5 Conclusion

In this paper we introduced HCNN, a novel architecture based on Hahn moments and Convolutional Neural Networks. The proposed method provides a powerful solution to overcome the highly computation requirements of CNN, by extracting the main and useful characteristics of the image to perform the classification with effectiveness. With a shallow architecture such ours, we have demonstrated the effectiveness of HCNN on the two datasets.

References

Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pp. 1–5. IEEE, 2015.

Amr Bakry and Ahmed Elgammal. Manifold-kernels comparison in mkpls for visual speech recognition. *arXiv preprint arXiv:1601.05861*, 2016.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *European conference on computer vision*, pp. 484–498. Springer, 1998.

Cletus G Fisher. Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4):796–804, 1968.

Sarah Hilder, Richard W Harvey, and Barry-John Theobald. Comparison of human and machine-based lip-reading. In *AVSP*, pp. 86–89, 2009.

Iain Matthews, Timothy F Cootes, J Andrew Bangham, Stephen Cox, and Richard Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002.

Abderrahim Mesbah, Aissam Berrahou, Mostafa El Mallahi, and Hassan Qjidaa. Fast and efficient computation of three-dimensional hahn moments. *Journal of Electronic Imaging*, 25(6):061621, 2016.

Stavros Petridis and Maja Pantic. Deep complementary bottleneck features for visual speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2304–2308. IEEE, 2016.

Takeshi Saitoh, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Concatenated frame image based cnn for visual speech recognition. In *Asian Conference on Computer Vision*, pp. 277–289. Springer, 2016.

Chunlin Tian and Weijun Ji. Auxiliary multimodal lstm for audio-visual speech recognition and lipreading. *arXiv preprint arXiv:1701.04224*, 2017.

Guoying Zhao, Mark Barnard, and Matti Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265, 2009.

Jian Zhou, Huazhong Shu, Hongqing Zhu, Christine Toumoulin, and Limin Luo. Image analysis by discrete orthogonal hahn moments. In *International Conference Image Analysis and Recognition*, pp. 524–531. Springer, 2005.