Lip Reading with Hahn Convolutional Neural Networks moments

Abderrahim Mesbah, Hicham Hammouchi, Aissam Berrahou, Hassan Berbia, Hassan Qjidaa, Mohamed Daoudi

Lip Reading with Hahn Convolutional Neural Networks

Abderrahim Mesbah[1¶], Hicham Hammouchi[1,2¶], Aissam Berrahou[2], Hassan Berbia[2], Hassan Qjidaa[1], Mohamed Daoudi[3]

[1] *Sidi Mohammed Ben Abdellah University, Fez, Morocco*
[2] *Mohammed V University, Rabat, Morocco*
[3] *IMT Lille-Douai, Univ. Lille, CNRS, UMR 9189 CRIStAL, Lille, France*

Abstract

Lipreading or Visual speech recognition is the process of decoding speech from speakers 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 aim to develop a visual-only speech recognition system based only on video. Our main targeted application is in the medical field for the assistance to laryngectomized persons. To that end, we propose Hahn Convolutional Neural Network (HCNN), a novel architecture based on Hahn moments as first layer in the Convolutional neural network (CNN) architecture. We show that HCNN helps in reducing the dimensionality of video images, in gaining training time. HCNN model is trained to classify letters, digits or words given as video images. We evaluated the proposed method on three datasets, AVLetters, OuluVS2 and BBC LRW, and we show that it achieves significant results in comparison with other works in the literature.

Keywords: Visual speech recognition, Lipreading, Laryngectomy, Hahn

¶A. Mesbah and H. Hammouchi contributed equally to this work

1. Introduction

The visual speech recognition also known by lip reading is a vital task in communication for people with difficulties to interact with society. It has gained a lot of attention lately due to the need of the application of this process in many domains, especially in the medical field which is our motivation for this work. In case of laryngeal cancer, the persons affected by this disease face the loss of their natural voice after the laryngectomy surgery. Similarly, vocal cord paralysis is another disease that causes the same problems. By assuming that voice therapy can take long to recover the voice, the lipreading is an inevitable task to reconstruct the communication process and save the interaction with society. To concretize that, the advances in computer vision and image processing made this process possible to decode speech from lips movements and offer to these patients a hope to recuperate their communication functions.

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 [1]. Additionally, the differences between each speakers 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 [2], 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 [5], for example an approach using Active Shape Models (ASM) [4] and Active Appearance Models (AAM) [3] 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 [6] 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 [7], where the authors trained a model using Long-Short Term Memory (LSTM), this method achieved 58.1% accuracy. Bakry and Elgammal in 2016 [9], 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 [10], 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 [23] 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 [22] 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 [13] 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. Additionally, regarding the Lip Reading in the Wild (LRW), it was first generated by Chung and Zisserman [25] in 2016, to recognize spoken words. Authors achieved an accuracy of 61.1% using a multiple towers architecture, which uses a convolutional tower for each frame and they concatenate after the first convolution then apply a pretrained VGG-M model. Chung et. al. in [26] present a method called watch, listen, attend and spell, that aims to convert mouth motion videos to characters. They obtained on LRW an accuracy of 76.2% using both CNN and LSTM to recognize the spoken words. Also, Stafylakis and Tzimiropoulos in [28] and [27], obtained 82.97% and 88.08% respectively, using a combination of a spatiotemporal 3D CNN, ResNet and Bidirectional

LSTM.

The proposed methods above apply spatiotemporal modeling features and visual features extraction before performing classification, which make them computationally expensive and time consuming. In order to cut with these limitations we propose in this paper 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. The 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.

This paper is structured as follows. In Section 2 we will introduce Hahn moments concept and mathematical foundation. In Section 3 we will highlight our proposed architecture HCNN. Our Experiments and Results on the reported datasets are found in Section 4. And finally a conclusion to this paper will be given in Section 5.

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 [17], respect the following equation:

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

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 :

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

Hahn polynomials satisfy the orthogonal property:

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

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

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

To assure the numerical stability, the set of the weighted Hahn polynomials is defined as

$$\tilde{h}_n\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big) = h_n\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big)\sqrt{\frac{w_h(x)}{\rho_h(n)}} \tag{6}$$

The set of weighted Hahn polynomials obeys the three term recurrence relation defined as follow

$$\begin{aligned}\tilde{h}_n\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big) =& A\sqrt{\frac{\rho_h(n-1)}{\rho_h(n)}}\tilde{h}_{n-1}\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big) \\ &- B\sqrt{\frac{\rho_h(n-2)}{\rho_h(n)}}\tilde{h}_{n-2}\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big) \\ & n = 2,3,\cdots,N-1\end{aligned} \tag{7}$$

Where

$$A = 1 + B - x\frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{(n+\alpha+\beta+1)(\alpha+n+1)(N-n)} \tag{8}$$

$$B = \frac{n(n+\beta)(\alpha+\beta+n+N+1)(2n+\alpha+\beta+2)}{(2n+\alpha+\beta)(\alpha+\beta+n+1)(\alpha+n+1)(N-n)} \tag{9}$$

The initial values for the above recursion can be obtained from

$$\tilde{h}_0\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big) = \sqrt{\frac{\omega_h(x)}{\rho_h(0)}} \tag{10}$$

$$\tilde{h}_1\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big) = \left(1-\frac{x(\alpha+\beta+2)}{(\alpha+1)N}\right)\sqrt{\frac{\omega_h(x)}{\rho_h(1)}} \tag{11}$$

2.2. Weighted Hahn moments

The 2D weighted Hahn moments of order $m \times n$ of an image intensity function $f(x,y)$ are defined over the domain $[0, M-1] \times [0, N-1]$ as:

$$H_{mn} = \sum_{x=o}^{M-1}\sum_{y=0}^{N-1} \tilde{h}_m\Big(\ \alpha,\beta,M\ \Big|\ x\ \Big)\tilde{h}_n\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big) f(x,y) \tag{12}$$

where $M \times N$ is the size of the image. Due to the orthogonality property of separable Hahn polynomials allows us to reconstruct perfectly the image $f(x,y)$, if all moments are used, by utilizing the following inverse transform:

$$\tilde{f}(x,y) = \sum_{x=o}^{\hat{M}-1}\sum_{y=0}^{\hat{N}-1} \tilde{h}_m\Big(\ \alpha,\beta,M\ \Big|\ x\ \Big)\tilde{h}_n\Big(\ \alpha,\beta,N\ \Big|\ x\ \Big) H_{mn} \tag{13}$$

where $0 \le \hat{M} \le M$, $0 \le \hat{N} \le N$. The image moment set $H_{mn}(0 \le m \le M-1, 0 \le n \le N-1)$, can be speedily extracted by using the following matrix form [11]

$$H = Q_1^T A Q_2 \tag{14}$$

(a) reconstructed frame up to 12^{th} (b) reconstructed frame up to 16^{th} (c) reconstructed frame up to 20^{th}s (d) reconstructed frame up to 32^{th} (e) original frame 80×60

Figure 1: image reconstruction with Hahn moments

where H is an $m \times n$ matrix of moments $H = H_{ij}$, by
$0 \leq i \leq n,\ 0 \leq j \leq m$

$$Q_1 = \tilde{h}_m(\alpha, \beta, Mx), \text{ with } 0 \leq i \leq m \text{ and } 0 \leq x \leq M-1,$$
$$Q_2 = \tilde{h}_n(\alpha, \beta, Ny), \text{ with } 0 \leq j \leq n \text{ and } 0 \leq y \leq N-1,$$
$$\text{and } A = f(x,y),\ 0 \leq x \leq M-1, 0 \leq y \leq N-1.$$

Similarly the image can be reconstructed by

$$f = Q_1 H Q_2^T \tag{15}$$

As illustrated in fig. 1, the image can be reconstructed using eq. 15, and obviously despite the small orders Hahn moments preserve the information in image.

3. Proposed Architecture HCNN

The novel architecture HCNN as shown in fig. 2 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 [24] 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. The architecture is divided on two main phases, the Hahn moments layer and the convolutional neural network.

• **Hahn Moments layer:** the discrete orthogonal Hahn moments as the first layer of HCNN to calculate the moments of the input image and yields a matrix of moments with a size depending on the moments order value. Therefore, this layer gives an optimized representation of the image and reduce significantly the dimensionality of processing.

• **Convolutional Neural Network:** comes to further learn pattern in the data and to provide a robust classification. The CNN takes the moments matrix as input instead of the image and apply the various convolutional filters and optimization functions. The first layers of convolution, activation functions, normalization and max pooling, process the input and learn more complex patterns and structures in the data. While the classification is performed in the fully connected layers in which we apply several operations as the normalization, the activation functions and the dropout.

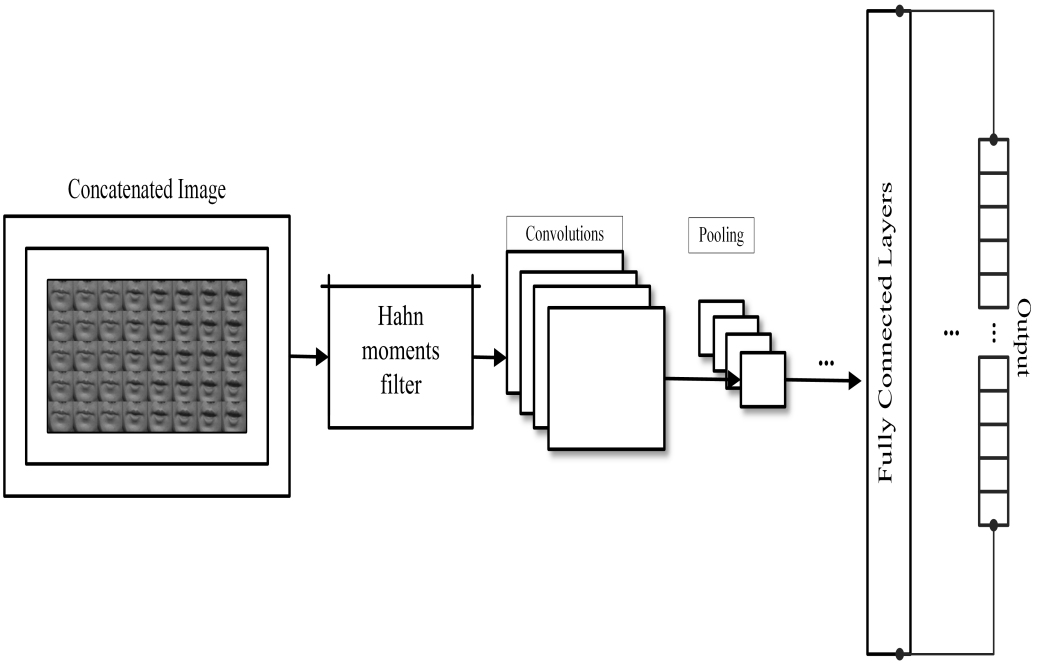

Figure 2: HCNN: the proposed architecture for lip reading which takes an input image and compute its correspond Hahn moments, then the returned representation is passed through several convolutions, normalization, max pooling, and fully connected layers

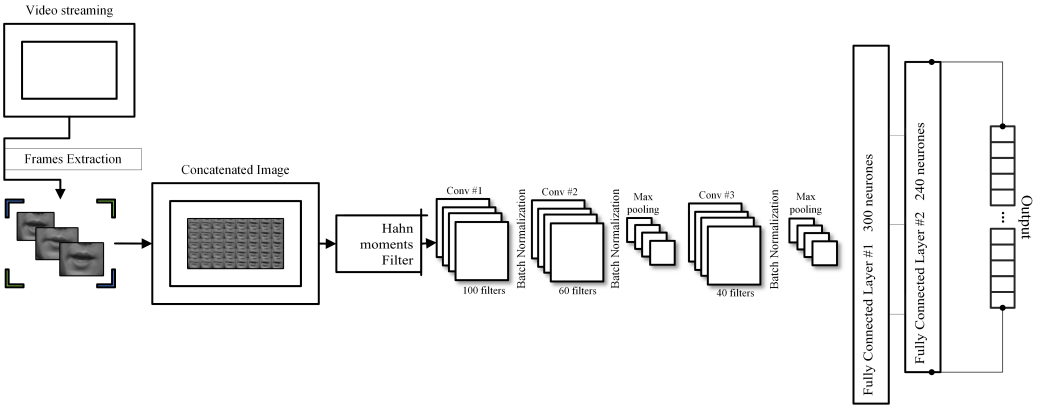

Figure 3: HCNN model parameters: Hahn moments until the given order, first convolution (kernel 3x3 and 100 filters), second convolution (kernel 3x3 and 60 filters), first max pooling (pool size 3x3). Third convolution (kernel 3x3 and 40 filters). Second max pooling (pool size 3x3). First fully connected layer (300 neurons), and a second layer (240 neurons), and finally an output layer (26 classes)

4. Experiments & Results

In this section, simulation results are carried out through a set of appropriate experiments in order to evaluate the classification performance of the proposed architecture (HCNN). Three datasets have been selected from the literature : AVLetters, OuluVS2 and BBC LRW.

4.1. datasets

• **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. The speaker start and finish with a closed mouth, and every video has a variable number of frames. The frames as shown in figure4 , are an example of images in AVLetters dataset for the mouth region with a dimension of 80x60.

• **OuluVS2:** a dataset that contains 52 speakers uttering 10 digits sequences with a repetition of 3 times each. The digits series are as follow : "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". The dataset as shown in figure 6 is provided with cropped mouth region and with multiple views, with an angle of rotation of the speaker. In our experiments we use the frontal view, with resizing the extracted images to the size of 50×50.

• **Lip Reading in the Wild:** The BBC Lip Reading in the Wild (LRW) dataset contains 500 unique words with up to 1000 utterances per word spoken by different speakers. The dataset as given provide the train, validation and test sets, as well as the metadata indicating the time where the word

appear. It is a very challenging dataset with a high number of classes and with words close in spelling which increase the chances of confusion between words.

4.2. Data Augmentation and Preprocessing

4.2.1. Data Augmentation

In order to conduct a fair comparison with other works in the literature, especially those who worked on OuluVS2 dataset, we performed a data augmentation (DA) by applying rotations with angle degrees [-15, -10, 10, 15], on each frame of each video. As for the Lip Reading in the Wild dataset, we augment by flipping horizontally each frame in the dataset. Consequently, the size of the dataset is doubled, and since the dataset is very large, augmenting it by applying other operations makes the processing and the training expensive.

4.2.2. Preprocessing

In order to prepare the datasets, we first proceed to retrieve frames from the given videos in both AVLetters and OuluVS2 digits datasets. The extracted frames appear in various numbers depending on video length, which make their manipulation unfitting. To correct this situation we supplement the last frame in each video at the end to unify the numbers of frames in all videos. We choose to add the last frame from each video because we believe that it will not affect the content of the video, since the last frame is just repetition and does not add any information. For the AVLetters we have set the number of frames to 40 frames in each video, so for each video which the number of frames is under 40 we repeat the last frame until we get 40

frames. In the case of OuluVS2 dataset the number of frames is relatively high (the sequence length ranges from 130 to 220), so we proceed differently. We process each video separately depending on its sequence length and we repeat the last frame until we get a perfect square number of frames, and to determine that number we use the following manner:
let n_f be the number of frames in a video, the number c of images to be concatenated as given as follow :

$$c = (\lceil \sqrt{n_f} \rceil)^2 \quad \text{by } \lceil x \rceil = min\{n \in \mathbb{Z} \,|\, n \geq x\}$$

A further issue to tackle is the modeling of spatiotemporal aspect in the video. For this matter we use the method of concatenated frame image (CFI) proposed by Saitoh *et al.* in [13] to represent all frames in one concatenated image. The concatenated image is constructed in a way to conserve the chronological order of appearance of frames in the video. For the AVLetters we arrange the 40 frames as one image with 8 frame in each row as illustrated in fig. 7 and fig. 8. In the case of OuluVS2 digits we constructed a square shape image with the size of $\sqrt{c}$ frames in rows and $\sqrt{c}$ in columns.
Consequently, we have augmented the OuluVS2 digits dataset to 7650 images.

As for the Lip Reading in the Wild dataset, we start by retrieving the frames from the given videos, then we use the provided meta data to cut the video and keep only the frames corresponding to the spoken word. finally we crop the region of each speaker's mouth (ROI).

Figure 4: AVLatters Frames

Figure 5: OuluVS2 Digits frames

Figure 6: Lip Reading in the Wild frames sample

Figure 7: Example of CFI for the AVLetters dataset

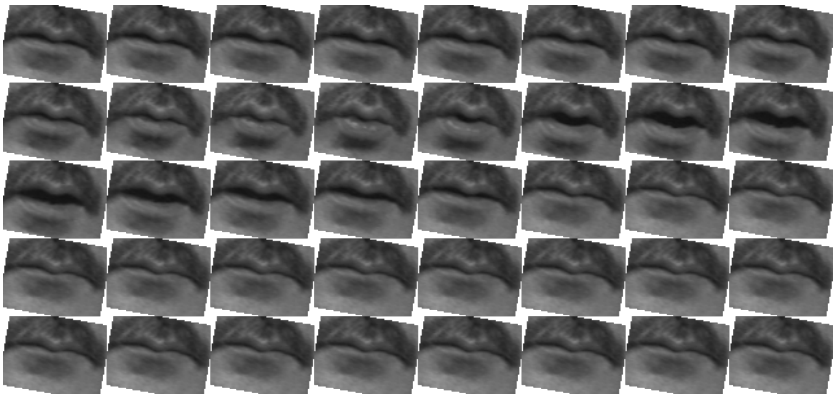

Figure 8: Concatenated Frame Image with rotation

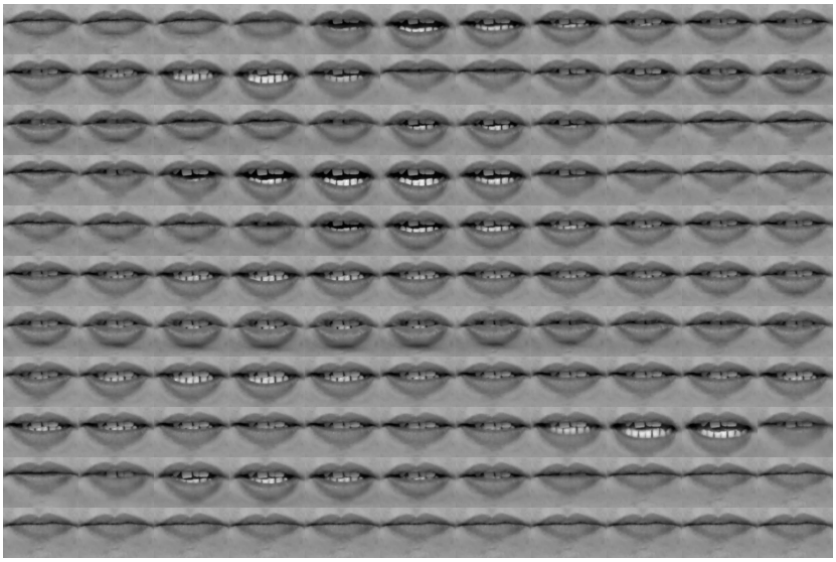

Figure 9: Concatenated Frame Image for OuluVS2 dataset

4.3. Model training parameters

In this section, we highlight the different parameters we used to train our model. After the preprocessing discussed in data augmentation and preprocessing section above, we split our dataset into train and test. For the AVLetters dataset we followed the split given in [5] by taking the two utterances of each speaker for the training and the remaining utterance for the test, which results in 520 images for training and 260 for test, and we train our model for 5000 epochs using a batch size of 520. The OuluVS2 dataset we used the speaker-independent protocol where the twelve speakers (s06, s08, s09, s15, s26, s30, s34, s43, s44, s49, s51, and s52, 10 males and 2 females) were used for testing and the remaining speakers for the training. After the split we had 1800 images for test and 5850 for train, used to train the model for 1000 iterations with a batch size of 1070. As for the Lip

Table 1: HCNN Model parameters

Layer	**Purpose**	**Filter**	**Nb of filters**	**Stride**	**Activation**	**Dropout Probability**
-	input image	-	-	-	N×N	-
1	Moment layer	-	-	-	n×n	-
2	Conv+BN+RELU	3×3	100	1	n×n×100	-
3	Conv+BN	3×3	60	1	n×n×60	
4	Maxpooling+ELU	3×3	-	2	$\frac{n}{2} \times \frac{n}{2} \times 40$	-
5	Conv+BN	3×3	40	1	$\frac{n}{2} \times \frac{n}{2} \times 40$	-
6	Maxpooling+ELU	3×3	-	2	$\frac{n}{4} \times \frac{n}{4} \times 40$	-
7	Fully connected (RELU)	-	-	-	300	0.55
8	Fully connected (RELU)	-	-	-	240	0.55
9	Softmax	-	-	-	nb of classes	-

Reading in the Wild dataset we respect the split provided by default, and we train a model for 105 epochs with a batch size of 300.

To train our model we first apply the moments filter on each CFI to generate its moments matrix at different orders, where the choice of orders is based on the concatenated image size, as for the images of AVLetters dataset we have calculated moments up to the $16^{th}, 32^{th}, 52^{th}\, and\, 60^{th}, 64^{th}\, and\, 72^{th}$ order. In the OuluVS2 as the generated concatenated images are large we have computed the moments up to the $12^{th}, 16^{th}, 32^{th}, 44^{th}, 56^{th}\, and\, 60^{th}$. The choice of moments order is based on the performance obtained on the test set. We change the order incrementally and we report the results obtained using several orders to show how the performance changes. In addition, We perform an ablation study by removing the Hahn filter and use the same architecture parameters reported in table 1. For AVLetters we resize the CFI to (150×320), and for OuluVS2 we set the CFI size to (225×225).

In the case of LRW dataset, we use the same model as above (2D for the image dimensions and the third dimension for the sequence length). As we used moments up to the 22^{th} order, the input has a size of $(sequence\, length \times 22 \times 22)$. We feed then the moments matrix to the CNN with parameters illustrated in table 1, with a slight change in the fully connected layers. We augment the two layers with 300 and 240 neurons to 2048 and 1000 respectively because of the large number of classes (500 words).

4.4. Results and Discussion

The recognition rate of our model in comparison with the previous works on AVLetters dataset are shown in table 4. 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 3 two works for comparison, SyncNet [22] and CFI-based CNN [13], 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.

Moreover, the results obtained on LRW dataset are listed in table 6. Our method achieves 56.44% *top*@1% accuracy with data augmentation, with an improvement of about 9% over HCNN without any data augmentation. Therefore, the augmentation enhances clearly the performance of the model. In comparison to other works, We perform worse, however, conducting any comparison is not fair, because our model is shallow against the proposed in [25, 26, 28, 27]. Also, because the data augmentation they applied is extensive against only flipping in our experiments, which allows the model to learn more patterns in the sequences.

In the light of the above, the combination of Hahn moments and CNN proved its effectiveness on a complex problem such as the Lip reading problem. HCNN by the component of Hahn moments extract only the significant information by capturing the global features in the image. Further, we significantly reduce the dimensionality and the complexity of the model compared to using the standard CNN architecture. Furthermore, with a small architecture such the ours we can achieve results better than other works that used large architecture such as GoogleNet in the case of OuluVS2, or large pre-trained models like VGG-M.

Table 2: 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 3: 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 4: Obtained results on AVLetters in comparison with other methods

Method	HMM [5]	LBP-SVM [6]	LSTM [7]	**HCNN**	CNN Without Hahn
Accuracy	44.6%	58.85%	58.1%	**59.23%**	39.23%

Table 5: Obtained results on OuluVS2 Digits (frontal view) in comparison with other methods

Method	Accuracy
CFI-based CNN (GoogLeNet) +DA [13]	89.4%
SyncNet (CNN+LSTM) +DA [22]	92.8%
HCNN +DA (SI)	**93.72%**
CNN Without Hahn +DA (SI)	42.27%

Table 6: Obtained results on BBC LRW words dataset whole 500 classes in comparison with other methods

Method	$Top@1\ Acc$	$Top@5\ Acc$	$Top@10\ Acc$
VGG-M [25]	61.1%	-	90.4%
Watch-Attend-Spell [26]	76.2%	-	-
ResNet-LSTM [28]	82.97%	96.3%	98.3%
Bi-LSTM & ResNet [27]	88.08%	96.28%	-
HCNN (without DA)	**55.86**%	**82.93**%	**89.95**%
HCNN (+ flip DA)	**58.02**%	**84.54**%	**90.86**%

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 and deep learning, and to extract the main and useful characteristics of the image to perform the classification with effectiveness. With a small ar-

chitecture such the HCNN, we have demonstrated the effectiveness of HCNN on three datasets, AVLetters, OuluVS2 digits and the LRW against deep architectures, despite the shortfalls of some results. These findings would assist researchers to tackle the problem of lipreading and could be a useful aid for the laryngectomized persons to decode their speech from their lips movements. Nevertheless, we believe our work could be a starting point to apply this method on real-time basis, experiment it on real patients, and also to recognize speech independently of the language.

Acknowledgment

The authors would like to thank Prof. Ismail Berrada for providing us with the Nvidia GTX 1080Ti GPU used in our experiments.

References

[1] C. G. Fisher, "Confusions Among Visually Perceived Consonants," Journal of Speech, Language, and Hearing Research, vol. 11, pp. 796-804, Dec 1968.

[2] Sarah Hilder, Richard Harvey, Barry-John Theobald, "Comparison of human and machine-based lip-reading" Auditory-Visual Speech Processing, Norwich,, p. 8689, Sept 10th-13th, 2009.

[3] T. F. Cootes, G. J. Edwards, C. J. Taylor, "Active Appearance Models" Proc. European Conf. Computer Vision, pp. 484-498, June 1998.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, "Active Shape

Models-Their Training and Application" Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38-59, Jan. 1995.

[5] I. Matthews, T.Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 198-213, FEB, 2002.

[6] Guoying Zhao, Mark Barnard and Matti Pietikinen, "Lipreading With Local Spatiotemporal Descriptors" IEEE Transactions on Multimedia, vol. 11, no. 7, pp. 1254-1265, 2009.

[7] Stavros Petridis, Maja Pantic, "Deep Complementary Bottleneck Features for Visual Speech" IEEE, pp. 2304-2308, 2016.

[8] Di Hu, Xuelong Li, Xiaoqiang Lu, "Temporal Multimodal Learning in Audiovisual Speech Recognition" In IEEE Conference on Computer Vision and Pattern Recognition, pp. 35743582, 2016.

[9] Amr Bakry, Ahmed Elgammal, "Manifold-Kernels Comparison in MK-PLS for Visual Speech Recognition" arXiv:1601.05861 [cs.CV].

[10] Chunlin Tian, Weijun Ji, "Auxiliary Multimodal LSTM for Audio-visual Speech Recognition and Lipreading" arXiv:1701.04224v2, 2017.

[11] Zhou J., Shu H., Zhu H., Toumoulin C., Luo L. (2005) Image Analysis by Discrete Orthogonal Hahn Moments. In: Kamel M., Campilho A. (eds) Image Analysis and Recognition. ICIAR 2005. Lecture Notes in Computer Science, vol 3656. Springer, Berlin, Heidelberg

[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, "Backpropagation applied to handwritten zip code recognition" Neural Comput., vol. 1, no. 4, p. 541551, 1989.

[13] Saitoh T., Zhou Z., Zhao G., Pietikinen M. (2017) , Concatenated Frame Image Based CNN for Visual Speech Recognition. In: Chen CS., Lu J., Ma KK. (eds) Computer Vision ACCV 2016 Workshops. ACCV 2016. Lecture Notes in Computer Science, vol. 10117, Springer, Cham.

[14] Stephen Cox, Richard Harvey, Yuxuan Lan, Jacob Newman, Barry-John Theobald, "The challenge of multispeaker lip-reading" International Conference on Auditory-Visual Speech Processing, Citeseer,, p. 179184, 2008.

[15] Noda, K., Yamaguchi, Y., Nakadai, K. et al. Appl Intell (2015) 42: 722. doi:10.1007/s10489-014-0629-7

[16] Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A. Bernal, Jiebo Luo, "Deep Multimodal Representation Learning from Temporal Data" arXiv:1704.03152v1 [cs.CV].

[17] Zhou J., Shu H., Zhu H., Toumoulin C., Luo L. (2005) Image Analysis by Discrete Orthogonal Hahn Moments. In: Kamel M., Campilho A. (eds) Image Analysis and Recognition. ICIAR 2005. Lecture Notes in Computer Science, vol 3656. Springer, Berlin, Heidelberg

[18] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao, Matti Pietikinen, "A Compact Representation of Visual Speech Data Using Latent Variables"

IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 1, pp. 181-187, Jan, 2014.

[19] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-Based Learning Applied to Document Recognition" Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[20] Sergey Ioffe, Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift" arXiv:1502.03167v3 [cs.LG], 2015.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting" Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, Jan. 2014.

[22] Chung J.S., Zisserman A. (2017) Out of Time: Automated Lip Sync in the Wild. In: Chen CS., Lu J., Ma KK. (eds) Computer Vision ACCV 2016 Workshops. ACCV 2016. Lecture Notes in Computer Science, vol 10117. Springer, Cham

[23] Anina I., Zhou Z., Zhao G., and Pietikinen 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 (FG15), Ljubljana, Slovenia, 1-5.

[24] A. Mesbah, A. Berrahou, M. El Mallahi, H. Qjidaa. Fast and efficient computation of three-dimensional Hahn moments. Journal of Electronic Imaging, 25 (6), doi: 10.1117/1.JEI.25.6.061621, (2016)

[25] J. S. Chung, A. Zisserman. Lip Reading in the Wild, In ACCV, (2016)

[26] J. S. Chung, A. W. Senior, O. Vinyals, A. Zisserman, Lip Reading Sentences in the Wild, In CVPR, pp.3444-3453, (2017).

[27] T. Stafylakis, G. Tzimiropoulos, Deep word embeddings for visual speech recognition, In CoRR, abs/1710.11201, (2017)

[28] T. Stafylakis, G. Tzimiropoulos, Combining Residual Networks with LSTMs for Lipreading, In Interspeech, (2017)