The Importance of the Instantaneous Phase for Face Detection using Simple Convolutional Neural Networks

Luis Sanchez Tapia*, Marios S. Pattichis†, Sylvia Celedón-Pattichis‡, and Carlos LópezLeiva‡
*Department of Electrical and Computer Engineering
The University of New Mexico, Albuquerque, NM, USA. Email: luis2sancheztapia@unm.edu
†Department of Electrical and Computer Engineering, and
Center for Collaborative Research and Community Engagement, College of Education
The University of New Mexico, Albuquerque, NM, USA. Email: pattichi@unm.edu
‡Dept. of Language, Literacy, and Sociocultural Studies
The University of New Mexico, Albuquerque, NM, USA. Email: {sceledon, callopez}@unm.edu

***Abstract*—Large scale training of Deep Learning methods requires significant computational resources. The use of transfer learning methods tends to speed up learning while producing complex networks that are very hard to interpret.**

This paper investigates the use of a low-complexity image processing system to investigate the advantages of using AM-FM representations versus raw images for face detection. Thus, instead of raw images, we consider the advantages of using AM, FM, or AM-FM representations derived from a low-complexity filterbank and processed through a reduced LeNet-5.

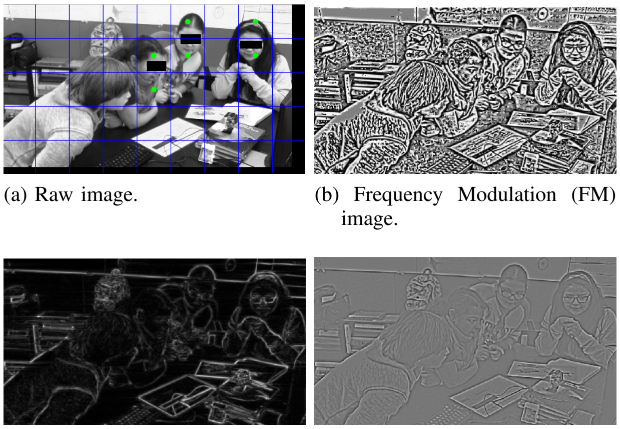
The results showed that there are significant advantages associated with the use of FM representations. FM images enabled very fast training over a few epochs while neither IA nor raw images produced any meaningful training for such low-complexity network. Furthermore, the use of FM images was 7× to 11× faster to train per epoch while using 123× less parameters than a reduced-complexity MobileNetV2, at comparable performance (AUC of 0.79 vs 0.80).

***Index Terms*—Instantaneous phase, AM-FM representations, low-complexity neural networks.**

I. Introduction

Convolutional Neural Networks dominate image and video analysis methods. Large-scale training of deep learning networks using large databases requires significant computational resources. Once trained, the resulting networks are very large and hard to interpret. As a result, we have the use of Transfer learning methods that re-train pre-trained networks on smaller problems. Yet, the result is still unsatisfactory since the resulting networks are still large and hard to interpret and still require significant computational resources.

The current paper seeks to investigate the possibility of fast training using interpretable inputs derived from low-complexity processing. Our work is motivated by earlier research that showed the importance of the phase in representing images [1]. Unfortunately, since the phase derived from the FFT is hard to understand, we instead consider the use of AM-FM representations that allow us to visualize instantaneous phase through the use of the FM components (e.g., see [2]).


(a) Raw image.

(b) Frequency Modulation (FM) image.

(c) Instantaneous Amplitude (IA) image.

(d) AM-FM image.

Fig. 1: Image representations for face detection. The input image is broken into 50×50 blocks. In (a), each block that contains a face is marked by a green marker.

For comparison, we select a classic application of face detection to examine the advantages of using FM components versus the standard use of raw images with simple image classifiers. Once again, we select the classical LeNet classifier [3] to see if it is possible to derive any advantages in using FM, AM, or AM-FM versus using raw images as inputs.

For face detection, we consider images derived from a collection of videos from the Advancing Out-of-school Learning in Mathematics and Engineering (AOLME) [4] after-school program. Our dataset is derived from an unconstrained classroom environment without any restrictions on pose or the number of faces. Our interest in developing low-complexity methods comes from the need to process thousands of hours of videos.

We introduce the basic problem in Fig 1. The image is broken into non-overlapping blocks. The problem is then to

978-1-7281-5745-0/20/$31.00 ©2020 IEEE

determine whether a face appears in each block (see Fig. 1(a)). We consider the development of low-complexity face detectors based on the raw image, the derived dominant instantaneous amplitude (AM) the derived FM, and AM-FM image.

Our paper claims two main contributions. First, for Face detection using LeNet-5 type architectures, we establish that FM representations are much more effective than raw images, IA, or even dominant-component AM-FM representations. In fact, we found that the derived FM-based face detectors can be easily trained with few epochs while other representations cannot provide low-complexity detectors that cannot match this performance. Second, we develop a low-complexity face detection system that closely approximates the performance of MobileNetV2 [5], a low-complexity neural network architecture. Towards this end, we develop a low-complexity filterbank for deriving FM representations using a directional filterbank with just sixteen filters and two scales. We show that our derived architecture is much faster to train and contains significantly fewer parameters than [5].

The rest of the paper is organized as follows. We summarize related AM-FM research in section II. We then describe our method in section III. We provide results in section IV and concluding remarks in section V.

II. Background

We break our discussion of the background into two parts. First, we describe related research on the use of AM-FM representations. Second, we provide a brief overview of low-complexity neural networks.

A. Related AM-FM Research

We focus on the use of dominant component analysis to derive an AM-FM representation based on a single filterbank (e.g., see [2]). After applying the Hilbert transform along each row, we process the image through a Gabor filterbank to extract the dominant component given by:

$$I(x,y) \approx a(x,y)\cos\varphi(x,y)$$

where $I(x,y)$ denotes the input image, $a(x,y)$ denotes the instantaneous amplitude, and $\varphi(x,y)$ denotes the instantaneous phase. The FM image is given by $\cos\varphi(x,y)$.

AM-FM methods have been applied on the AOLME dataset as described in [6] and [7]. In [6], the authors described an AM-FM method for detecting heads. The approach considered both back of the head and face detection. The approach was extended in [7] to detect dynamic group interactions. This prior research served as the motivation for the current paper. Moreover, AM-FM applications extends to other areas such as biomedical image analysis (e.g. [8]).

The AM-FM decompositions of [6] and [7] required the use of an extensive filterbank with a relatively large number of coefficients. Since we are interested in developing methods for processing thousands of hours of videos, we want to consider a low-complexity filterbank. More specifically, we reduce the number of filters from 54 to 16 and reduce the number of coefficients from 21×21 to 11×11. As a result, we estimate a tenfold reduction in the amount of time required to process each input image through the filterbank.

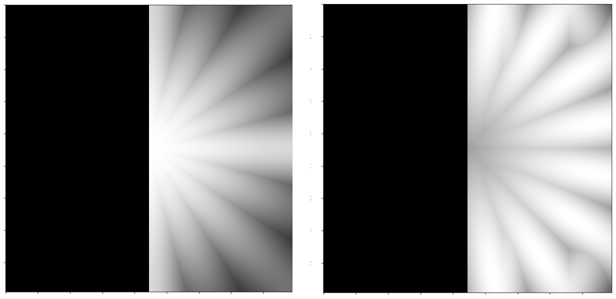

(a) Directional Gaussian centered at the DC.
(b) Directional Gabor filters centered at a radius of $\pi/2$.

Fig. 2: Frequency magnitude plots for low-complexity filterbank. Each filter uses 11×11 coefficients.

B. Low-complexity Neural Networks

We consider two different approaches. First, as a baseline low-complexity network, we consider simple variations of the original LeNet-5 architecture [3]. As we shall discuss in our methodology, in our final network, we further reduced LeNet-5 to having just one convolutional layer. Second, for comparison on the state of the art, we consider the use of MobileNetV2 [5].

MobileNetV2 had been specifically designed for low-complexity, mobile applications. MobileNetV2 consists of a total of 21 layers. The layers include two convolutional layers and 19 residual bottleneck layers. The original MobileNetV2 was designed to process images of size 224×224, requiring 300 million multiply-adds per run. MobileNetV2 uses 3.4 million parameters that were trained with ImageNet for image classification and the COCO dataset for object detection.

III. Methodology

Our general method consists of two basic steps. First, we apply dominant component analysis to estimate the AM-FM representations. Second, we develop an optimization method for developing a low-complexity classifier based on AM-FM representations for comparing against the use of raw images.

A. Gabor Filterbank

To support directional sensitivity, we consider the use of a directional, Gabor filterbank as shown in Fig. 2. The filterbank was generated using two scales with 11×11 coefficients used for each filter. The ellipsoidal support for each filter uses $\sigma_x = 1.5$ and $\sigma_y = \sigma_x/4$. The directions are generated by rotating each filter by 0.39 radians. For better frequency coverage, we also have that the second scale angles are rotated by 0.39/2 radians with respect to the first scale. In the plots of Fig. 2, the frequency components with negative horizontal frequencies are set to zero due to the application of the Hilbert transform along the rows.

TABLE I: Architecture of the Low-complexity Neural Net for face detection using the FM image. The input is an image block of 50×50 pixels.

Layer	Type	Kernel	Size	Stride	Act.
In	Input	-	50x50	-	-
C1	Conv2D	6 (5x5)	46x46	1	selu
S2	MaxPool	6 (5x5)	23x23	2	-
F3	Dense	-	40	-	selu
F4	Dense	-	24	-	selu
Out	Dense	-	1	-	sigmoid

B. Dominant Component Analysis

Following the application of the Gabor filterbank, we have complex-valued image outputs for each filter. For each filter output we estimate the instantaneous amplitude (IA) and the instantaneous phase (IP):

$$a_i(x, y) \approx \text{abs}\,(g_i(x, y))$$
$$\varphi_i(x, y) \approx \text{angle}\,(g_i(x, y))$$

where $g_i(x, y)$ denotes the i-th filtered output that is complex valued. Then, over all of the pixels, we set the dominant component estimates of the IA and the IP to the estimates of the filter that produced the largest IA.

C. Fast training and architecture optimization using a low-complexity neural network

We present our approach for computing a low-complexity model in Fig. 3. We consider LeNet-5 as the initial model and consider different reductions to it. Here, we note that the approach only worked for FM images which allowed us to train with just 5 epochs. We could not train the low-complexity model on the IA, the AM-FM images, or on the raw images. For reducing the network, we consider reducing the number of convolutional layers and max pooling sizes.

IV. Results

A. Face Detection Dataset

For comparing the different methods, we use 12 video clips extracted from 12 different sessions. The test set consisted of 6 video clips extracted from 6 different sessions. From each video clip, one frame is extracted every minute, resulting in 24 frames per video clip. Each frame was reduced by half along its rows and columns. Then, each frame was broken into 50×50 blocks and each block was marked as a face block or not. Overall, 12960 blocks were extracted for training and validation, while 6480 blocks from different frames were used for testing.

B. Face-Detection Results

As mentioned earlier, the FM images provided reasonable performance with a low-complexity network. In comparison, the IA, AM-FM, and raw images did not produce any meaningful results when used with the low-complexity network. For comparison, we also compare our results against the use of raw images with MobileNetV2.

```
function BuildModel(TSet, ValSet, TLbl, ValLbl)
  ▷ Builds a low complexity model from the TrainSet
  ▷ Inputs:
  ▷   Face detection is performed over 50x50 blocks.
  ▷   TSet, ValSet: training and validation sets.
  ▷   TLbl, ValLbl: training and validation labels.
  ▷ Outputs:
  ▷   LowCompModel: low complexity model

    Pre-compute AM-FM representations
    CandModel ← LeNet5
    ContinueConds ← True
    while ContinueConds do
        LowCompModel ← CandModel
        BestLoss ← Loss(Fit(LowCompModel))

  ▷ Consider a reduced model:
        Reduce CandModel
        TrainLoss ← Loss(CandModel(TrainOnlySet))
        ValLoss ← Loss(CandModel(ValidationSet))

  ▷ Ensure that there is no degradation in performance:
        Cond1 ← $|\text{ValLoss} - \text{TrainLoss}| < 0.1 \cdot \text{TrainLoss}$
        Cond2 ← ValLoss $\leq$ BestLoss
        ContinueConds ← Cond1 and Cond2
    end while
    return LowCompModel
end function
```

Fig. 3: Method for computing low-complexity model. Note that we applied this approach multiple times with different model reduction techniques.

We present an example in Fig. 4. In this example, we were able to detect face blocks from three different students. However, for two of the students, the upper half blocks were missed and they are thus marked with yellow dots. We also had false positives that are marked by red dots. We do not show the true negatives, that formed the majority of our blocks.

We present the final, low-computational complexity network for FM images in Table I. Without sacrificing performance, the number of convolutional layers was reduced from 2 to 1, and the max pooling was increased to 5×5.

We present ROC results in Fig. 5. Using the original LeNet-5, we note the that the raw images produced an AUC near 0.5 (0.48). Yet, the FM gave an AUC of 0.79 with a much simpler network.

C. Comparisons between low-complexity FM detector and MobileNetV2

We compare the low-complexity FM detector against a MobileNetV2 in Table II and Fig. 6. We note that the FM detector trains significantly faster per epoch (6.8× to 11×

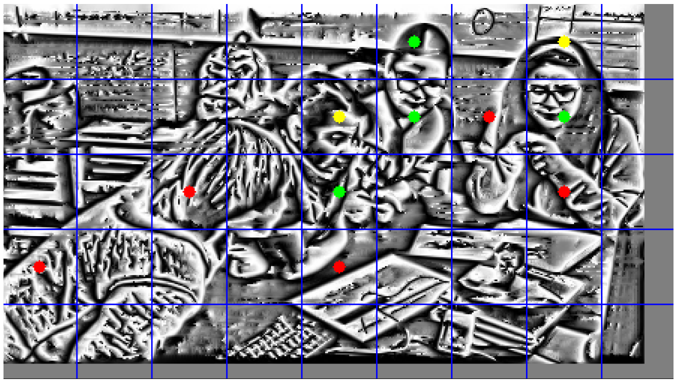


Fig. 4: Face detection using FM image. Over the FM image, we have marks for True Positives (green), False Positives (red) and False Negatives (yellow).

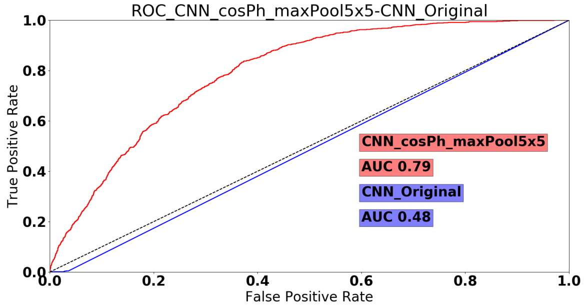


Fig. 5: ROC curves for face detection for FM component in red and raw image in blue.

faster). Furthermore, the proposed FM detector is significantly simpler than MobileNetV2 (123 times less parameters).

V. Conclusions and Future Work

The paper demonstrated the importance of the instantaneous phase in images through an application in block-based face detection. More specifically, unlike the standard use of raw images, the use of FM representations were found to lead to effective classifiers with simple neural net models. On the other hand, MobileNetV2 can match our performance while requiring training that is from 7 to 11 times slower with more than hundred more parameters to train. Future research will focus on the development of fast FM methods for large scale video databases.

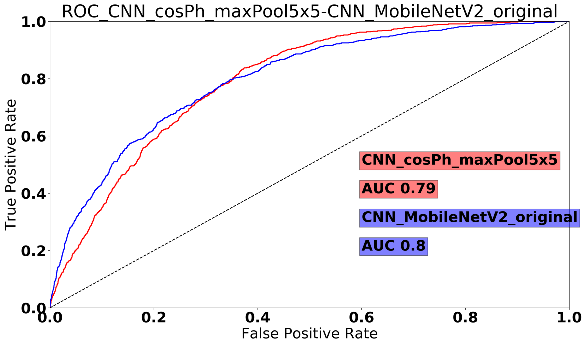


Fig. 6: ROC curves for face detection for: (i) FM representations with a low-complexity classifier in red (AUC: 0.79), and (ii) MobileNetV2 in blue (AUC: 0.80).

TABLE II: Comparison of training time for one epoch MobileNetV2 vs the Reduced LeNet-5 using GPUs. The 4GB NVIDIA GeForce GTX 1050 was installed in a laptop with an Intel core i5-7300HQ CPU @2.50GHz with 8GB of RAM. The 8GB NVIDIA GeForce GTX 1080 was installed in a desktop with an Intel Xeon CPU ES-2630 v4 @2.20GHz with 32GB of RAM.

Network	**GTX 1050** 640 cores	**GTX 1080** 2560 cores	**Num. of Params**
MobileNetV2	220 sec (11×)	68 sec (6.8×)	1205073 (123×)
Reduced LeNet-5	20 sec (1×)	10 sec (1×)	9775 (1×)

VI. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1613637 and Grant No. 1842220. Any opinions or findings of this thesis reflect the views of the author. They do not necessarily reflect the views of NSF.

References

[1] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[2] V. Murray, P. Rodriguez, and M. S. Pattichis, "Multiscale am-fm demodulation and image reconstruction methods with improved accuracy," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1138–1152, 2010.

[3] Y. LeCunn, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[4] Sylvia Celedon-Pattichis, Carlos Alfonso LopezLeiva, Marios S. Pattichis, and Daniel Llamocca, "An interdisciplinary collaboration between computer engineering and mathematics/bilingual education to develop a curriculum for underrepresented middle school students," *Cultural Studies of Science Education*, vol. 8, no. 4, pp. 873–887, Dec 2013.

[5] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation," *CoRR*, vol. abs/1801.04381, 2018.

[6] W. Shi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Robust head detection in collaborative learning environments using am-fm representations," in *2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, April 2018, pp. 1–4.

[7] W. Shi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Dynamic group interactions in collaborative learning videos," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Oct 2018, pp. 1528–1531.

[8] C. Agurto, S. Barriga, V. Murray, S. Nemeth, R. Crammer, W. Bauman, G. Zamora, M. Pattichis, and P. Soliz, "Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images," *Investigative Ophthalmology and Visual Science*, vol. 52, no. 8, pp. 5862–5871, 2011.