

# Volterra Neural Networks (VNNs)

**Siddharth Roheda**

SID.ROHEDA@GMAIL.COM

*Electrical and Computer Engineering Department  
North Carolina State University  
Raleigh, NC 27606, USA*

**Hamid Krim**

AHK@NCSU.EDU

*Electrical and Computer Engineering Department  
North Carolina State University  
Raleigh, NC 27606, USA*

**Bo Jiang**

BJIANG8@NCSU.EDU

*Electrical and Computer Engineering Department  
North Carolina State University  
Raleigh, NC 27606, USA*

**Editor:** Honglak Lee

## Abstract

The importance of inference in Machine Learning (ML) has led to an explosive number of different proposals, particularly in Deep Learning. In an attempt to reduce the complexity of Convolutional Neural Networks, we propose a Volterra filter-inspired Network architecture. This architecture introduces controlled non-linearities in the form of interactions between the delayed input samples of data. We propose a cascaded implementation of Volterra Filtering so as to significantly reduce the number of parameters required to carry out the same classification task as that of a conventional Neural Network. We demonstrate an efficient parallel implementation of this Volterra Neural Network (VNN), along with its remarkable performance while retaining a relatively simpler and potentially more tractable structure. Furthermore, we show a rather sophisticated adaptation of this network to nonlinearly fuse the RGB (spatial) information and the Optical Flow (temporal) information of a video sequence for action recognition. The proposed approach is evaluated on UCF-101 and HMDB-51 datasets for action recognition, and is shown to outperform state of the art CNN approaches. The code-base for our paper is available on github ( <https://github.com/sid-roheda/Volterra-Neural-Networks>).

**Keywords:** Volterra Filter, Activity Recognition, Activation Free Learning

## 1. Introduction

Human action recognition is an important research topic in Computer Vision, and may be useful in surveillance, video retrieval, and man-machine interaction to name a few. The survey on Action Recognition approaches Kong and Fu (2018) provides a good progress overview. Video classification usually involves three stages (Wang et al., 2009; Liu et al., 2009; Niebles et al., 2010; Sivic and Zisserman, 2003; Karpathy et al., 2014), namely, visual feature extraction (local features like Histograms of Oriented Gradients (HoG) introduced in Dalal and Triggs (2005), or global features like Hue, Saturation, etc.), feature fusion/concatenation, and lastly classification. In Yi et al. (2011), an intrinsic stochastic modeling of human activity on a shape manifold is proposed and an accurate

analysis of the non-linear feature space of activity models is provided. The emergence of Convolutional Neural Networks (CNNs) by LeCun et al. (1998), along with the availability of large training datasets and computational resources have come a long way to obtaining the various steps by a single neural network. This approach has led to remarkable progress in action recognition in video sequences, as well as in other vision applications like object detection (Sermanet et al., 2013), scene labeling (Farabet et al., 2012), image generation (Goodfellow et al., 2014), image translation (Isola et al., 2017), information distillation (Roheda et al., 2018b; Hoffman et al., 2016), etc. In the Action Recognition domain, datasets like the UCF-101 (Soomro et al., 2012), Kinetics (Kay et al., 2017), HMDB-51 (Kuehne et al., 2011), and Sports-1M (Karpathy et al., 2014) have served as benchmarks for evaluating various solution performances. In action recognition applications the proposed CNN solutions generally align along two themes: 1. *One Stream CNN* (only use either spatial or temporal information), 2. *Two Stream CNN* (integrate both spatial and temporal information). Many implementations (Carreira and Zisserman, 2017; Diba et al., 2017b; Feichtenhofer et al., 2016; Simonyan and Zisserman, 2014) have shown that integrating both streams leads to a significant boost in recognition performance. In Deep Temporal Linear Encoding (Diba et al., 2017b), the authors propose to use 2D CNNs (pre-trained on ImageNet (Deng et al., 2009)) to extract features from RGB frames (spatial information) and the associated optical flow (temporal information). The video is first divided into smaller segments for feature extraction via 2D CNNs. The extracted features are subsequently combined into a single feature map via a bilinear model. This approach, when using both streams, is shown to achieve a 95.6 % accuracy on the UCF-101 dataset, while only achieving 86.3 % when only relying on the RGB stream. Carreira and Zisserman (2017) adopt the GoogLeNet architecture which was developed for image classification in ImageNet (Deng et al., 2009), and use 3D convolutions (instead of 2D ones) to classify videos. This implementation is referred to as the Inflated 3D CNN (I3D), and is shown to achieve a performance of 88.8 % on UCF-101 when trained from scratch, while achieving a 98.0 % accuracy when a larger dataset (Kinetics) was used for pre-training the entire network (except for the classification layer). While these approaches achieve near perfect classification, the models are extremely heavy to train, and have a tremendous number of parameters (25M in I3D, 22.6M in Deep Temporal Linear Encoding). This in addition, makes the analysis, including the necessary degree of non-linearity, difficult to understand, and the tractability elusive. In this paper we explore the idea of introducing controlled non-linearities through interactions between delayed samples of a time series. We will build on the formulations of the widely known Volterra Series (Volterra, 2005) to accomplish this task. While prior attempts to introduce non-linearity based on the Volterra Filter have been proposed (Kumar et al., 2011; Zoumpourlis et al., 2017), most have limited the development up to a quadratic form on account of the explosive number of parameters required to learn higher order complexity structure. While quadratic non-linearity is sufficient for some applications (eg. system identification), it is highly inadequate to capture all the non-linear information present in videos.

**Contributions:** In this paper, we propose a Volterra Filter (Volterra, 2005) based architecture where the non-linearities are introduced via the system response functions and hence by controlled interactions between delayed frames of the video. The overall model is updated on the basis of a cross-entropy loss of the labels resulting from a linear classifier of the generated features. An efficiently cascaded implementation of a Volterra Filter is used in order to explore higher order terms while avoiding over-parameterization. The Volterra filter principle is also exploited to combine the RGB and the Optical Flow streams for action recognition, hence yielding a performance driven non-linear fusion of the two streams. We further show that the number of parameters required to realize

such a model is significantly lower in comparison to a conventional CNN, hence leading to faster training and a significant reduction of the required resources to learn, store, or implement such a model.

## 2. Background and Related Work

### 2.1 Volterra Series

A dynamical system, viewed as a black box, is characterized by its input/output relationship  $y_t/x_t$ . If a non-linear system is time invariant, the input/output relation can be expressed in the following form (Volterra, 2005; Schetzen, 1980),

$$y_t = \sum_{\tau_1=0}^{L-1} \mathbf{W}_{\tau_1}^1 x_{t-\tau_1} + \sum_{\tau_1, \tau_2=0}^{L-1} \mathbf{W}_{\tau_1, \tau_2}^2 x_{t-\tau_1} x_{t-\tau_2} + \dots + \sum_{\tau_1, \tau_2, \dots, \tau_K=0}^{L-1} \mathbf{W}_{\tau_1, \tau_2, \dots, \tau_K}^K x_{t-\tau_1} x_{t-\tau_2} \dots x_{t-\tau_K}, \quad (1)$$

where  $x_t, y_t \in \mathbb{R}$  and  $\mathbf{W}^k \in \mathbb{R}^k$ ,  $L$  is the number of terms in the filter memory (also referred to as the filter length),  $\mathbf{W}^k$  are the weights for the  $k^{\text{th}}$  order term, and  $\mathbf{W}_{\tau_1, \tau_2, \dots, \tau_k}^k = 0$ , for any  $\tau_j < 0$ ,  $j = 1, 2, \dots, k$ ,  $\forall k = 1, 2, \dots, K$  due to causality. This functional form is due to the mathematician Vito Volterra (Volterra, 2005), and is widely referred to as the Volterra Series. The calculation

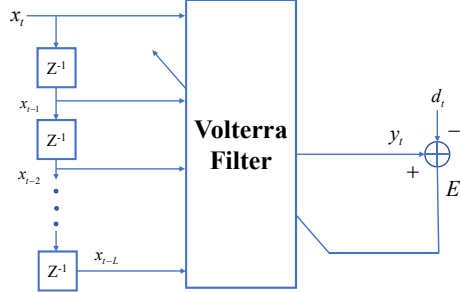


Figure 1: Adaptive Volterra Filter

of the kernels is computationally complex, and a  $K^{\text{th}}$  order filter of length  $L$ , entails solving  $L^K$  equations. The associated adaptive weights are a result of a target energy functional minimization iteratively adapting the filter taps as shown in Figure 1.

It is worth observing from Equation 1 that the linear term is actually similar to a convolutional layer in CNNs. Non-linearities in CNNs are implicitly introduced via activation functions, following the convolutional layer, while those in Equation 1 are introduced via higher order convolutions.

### 2.2 Nested Volterra Filter

A closer inspection of Equation 1 immediately suggests its somewhat simplified form by way of nesting repeated terms Osowski and Quang (1994), and yield,

$$y_t = \sum_{\tau_1=0}^{L-1} x_{t-\tau_1} \left[ \mathbf{W}_{\tau_1}^1 + \sum_{\tau_2=0}^{L-1} x_{t-\tau_2} \left[ \mathbf{W}_{\tau_1, \tau_2}^2 + \sum_{\tau_3=0}^{L-1} x_{t-\tau_3} [\mathbf{W}_{\tau_1, \tau_2, \tau_3}^3 + \dots] \right] \right]. \quad (2)$$

Each factor contained within brackets can be interpreted as the output of a linear Finite Impulse Response (FIR) filter, thus allowing a layered representation of the Filter. A nested filter implementation with  $L = 2$  and  $K = 2$  is shown in Figure 2. The length of the filter is increased by adding modules in parallel, while the order is increased by additional layers. Much like any multi-layer network, the weights of the synthesized filter are updated at successive layers according to a backpropogation scheme. The nested structure of the Volterra Filter, yields much faster learning in

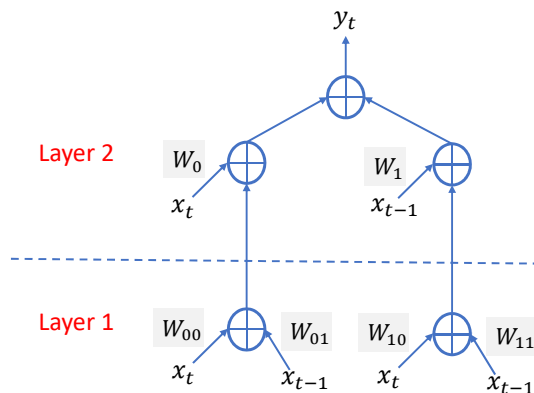


Figure 2: Nested Volterra Filter

comparison to that based on Equation 1. It, however, does not reduce the number of parameters to be learned, leading to potential over-parameterization when learning higher order filter approximations. Such a structure was used for a system identification problem in (Osowski and Quang, 1994). The mean square error between the desired signal ( $d_t$ ) and the output of the filter ( $y_t$ ) was used as the cost functional to be minimized,

$$E_t = \frac{1}{2}(d_t - y_t)^2, \quad (3)$$

and the weights for the  $k^{th}$  layer are updated per the following equations,

$$\mathbf{W}_{\tau_1, \tau_2, \dots, \tau_k}^k(t+1) = \mathbf{W}_{\tau_1, \tau_2, \dots, \tau_k}^k(t) - \eta \frac{\partial E_t}{\partial \mathbf{W}_{\tau_1, \tau_2, \dots, \tau_k}^k}, \quad (4)$$

$$\frac{\partial E_t}{\partial \mathbf{W}_{\tau_1, \tau_2, \dots, \tau_k}^k} = x_{t-\tau_k} x_{t-\tau_{k-1}} \dots x_{t-\tau_1} (y_t - d_t). \quad (5)$$

### 2.3 Relation to Bilinear Convolution Neural Networks

There has been work on introducing  $2^{nd}$  order non-linearities in the network by using a bi-linear operation on the features extracted by convolutional networks. Bilinear-CNNs (B-CNNs) were introduced in Lin et al. (2015) and were used for image classification. In B-CNNs, a bilinear mapping is applied to the final features extracted by linear convolutions leading to  $2^{nd}$  order features which are not well localized. As a result a feature extracted from the lower right corner of a frame/image in the B-CNN case, may interact with a feature from the upper right corner, and this potential non-real interaction may introduce erroneous additional characteristics, in contrast to our proposed approach which highly controls such effects. Compact Bilinear Pooling was introduced in Gao et al. (2016)

where a bilinear approach to reduce feature dimensions was introduced. This was again performed after all the features had been extracted via linear convolutions and was limited to quadratic non-linearities. In our work we will explore non-linearities of much higher orders and also account for continuity of information between video frames over a given time interval with the immediately preceding period. This effectively achieves a Recurrent Network-like property which accounts for a temporal relationship.

## 2.4 Relation to Long-Short Term Memory

Long-Short Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997) have been widely used to capture the long-term trajectory information in temporally evolving data. The sequential modeling ability of LSTMs makes them particularly appealing for capturing long-range temporal dynamics in videos. An LSTM computes a mapping from an input sequence,  $x = \{x_1, \dots, x_T\}$  to an output sequence  $h = \{h_1, \dots, h_T\}$ . The mapping of the input at time  $t$ ,  $x_t$  to the output,  $h_t$  entails the following:

### 1. Forget Gate:

$$f_t = \sigma(W_f[h_{t-1}, x_t]) = \sigma(W_f^1 \cdot h_{t-1} + W_f^2 \cdot x_t). \quad (6)$$

### 2. Input Gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t]) = \sigma(W_i^1 \cdot h_{t-1} + W_i^2 \cdot x_t), \quad (7)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t]), \quad (8)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t. \quad (9)$$

### 3. Output Gate:

$$o_t = \sigma(W_o[h_{t-1}, x_t]) = \sigma(W_o^1 \cdot h_{t-1} + W_o^2 \cdot x_t), \quad (10)$$

$$h_t = o_t \cdot \tanh(C_t). \quad (11)$$

Where  $\sigma$  is the sigmoid activation function,  $\tanh$  is the hyperbolic tangent function, and  $W_f, W_i, W_c$ , and  $W_o$  are the weight matrices characterizing the forget gate, input gate, cell state, and output gate respectively.

## 3. Problem Statement

Let a set of activities  $\mathcal{A} = \{a_1, \dots, a_I\}$ , be of interest following an observed sequence of frames  $\mathbf{X}_{T \times H \times W}$ , where  $T$  is the total number of frames, and  $H$  and  $W$  are the height and width of a frame. At time  $t$ , the features  $\mathbf{F}_t = g(\mathbf{X}_{[t-L+1:t]})$ , will be used for classification of the sequence of frames  $\mathbf{X}_{[t-L+1:t]}$  and mapped into one of the actions in  $\mathcal{A}$ , where  $L$  is the number of frames in the memory of the system/process. A linear classifier with weights  $\mathbf{W}^{cl} = \{\mathbf{w}_i^{cl}\}_{i=1, \dots, I}$ , and biases  $\mathbf{b}^{cl} = \{b_i^{cl}\}_{i=1, \dots, I}$  will then be central to determining the classification scores for each activity, followed by a soft-max function ( $\rho(\cdot)$ ) to convert the scores into a probability measure. The probability that the sequence of frames be associated to the  $i^{th}$  action class is hence the result of,

$$P_t(a_i) = \rho(\mathbf{w}_i^{clT} \cdot \mathbf{F}_t + b_i^{cl}) = \frac{\exp(\mathbf{w}_i^{clT} \cdot \mathbf{F}_t + b_i^{cl})}{\sum_{m=1}^I \exp(\mathbf{w}_m^{clT} \cdot \mathbf{F}_t + b_m^{cl})}. \quad (12)$$

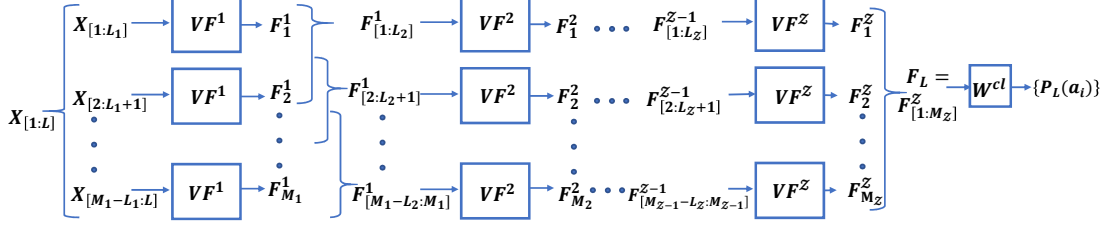


Figure 3: Block diagram for an Overlapping Volterra Neural Network

## 4. Proposed Solution

### 4.1 Volterra Filter based Classification

In our approach we propose a Volterra Filter structure to approximate a function  $g(\cdot)$ . Given that video data is of interest here, a spatio-temporal Volterra Filter must be applied. As a result, this 3D version of the Volterra Filter discussed in Section 2 is used to extract the features,

$$\begin{aligned}
 F \begin{bmatrix} t \\ s_1 \\ s_2 \end{bmatrix} &= g \left( \mathbf{X} \begin{bmatrix} t-L+1:t \\ s_1-p_1:s_1+p_1 \\ s_2-p_2:s_2+p_2 \end{bmatrix} \right) = \sum_{\tau_1, \sigma_{11}, \sigma_{21}} \mathbf{W}^1 \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} x \begin{bmatrix} t-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix} \\
 &+ \sum_{\substack{\tau_1, \sigma_{11}, \sigma_{21} \\ \tau_2, \sigma_{12}, \sigma_{22}}} \mathbf{W}^2 \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} x \begin{bmatrix} t-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix} x \begin{bmatrix} t-\tau_2 \\ s_1-\sigma_{12} \\ s_2-\sigma_{22} \end{bmatrix} + \dots
 \end{aligned} \quad (13)$$

where,  $\tau_j \in [0, L-1]$ ,  $\sigma_{1j} \in [-p_1, p_1]$ , and  $\sigma_{2j} \in [-p_2, p_2]$  respectively represent the temporal and spatial translations (horizontal and vertical directions). Following this formulation, and as discussed in Section 3, a linear classifier is used to determine the probability of each activity in  $\mathcal{A}$ . Updating the filter parameters is pursued by minimizing some measure of discrepancy relative to the ground truth and the probability determined by the model. Our adopted measure herein is the cross-entropy loss computed as,

$$E = \sum_{t,I} -d_{t_i} \log P_t(a_i), \quad (14)$$

where,  $t \in \{1, L+1, 2L+1, \dots, T\}$ ,  $i \in \{1, 2, \dots, I\}$ , and  $d_{t_i}$  is the ground truth label for  $\mathbf{X}_{[t-L+1:t]}$  belonging to the  $i^{\text{th}}$  action class. In addition to minimizing the error, we also include a weight decay in order to ensure generalizability of the model by penalizing large weights. So, the overall cost functional which serves as a target metric is written as,

$$\min_g \sum_{t,I} -d_{t_i} \log \rho(\mathbf{w}_i^{clT} \cdot g(\mathbf{X}_{[t-L+1:t]}) + b_i^{cl}) + \frac{\lambda}{2} \left[ \sum_{k=1}^K \|\mathbf{W}^k\|_2^2 + \|\mathbf{W}^{cl}\|_2^2 \right], \quad (15)$$

where  $\rho$  is the soft-max function, and  $K$  is the order of the filter.

**Proposition 1** *A VNN architecture provides any continuous function over a compact  $\Omega \subset \mathbb{R}^d$  (including so-called activation functions) an approximation up to an error margin defined by the Taylor Remainder Theorem.*



Figure 4: Approximation of ReLU and Tanh activation via the Volterra series formulation

**Proof** Based on the Weistrass Approximation Theorem (Stone, 1948) it is known that any continuous non-linear function can be approximated using a polynomial. Specifically, the Taylor expansion of the non-linear function may be used (i.e. a generic so-called activation function) for  $x \in \Omega \subset \mathbb{R}$ ,

$$\sigma(x) = c^0 + c^1x + c^2x^2 + \dots + c^kx^k + \dots + c^\infty x^\infty. \quad (16)$$

In particular, a sigmoid activation can be approximated as,

$$\sigma_{sigmoid}(x) = \frac{1}{1 + e^{-x}} = \frac{1}{2} + \frac{1}{4}x - \frac{1}{48}x^3 + \frac{1}{480}x^5 + \dots \quad (17)$$

As seen from Equation 1 the VNN formulation can specifically learn such an expansion up to a finite order,

$$\sigma_{VNN}(x) = w_0 + w_1x + w_2x^2 + \dots + w_kx^k, \quad (18)$$

which is a  $k^{th}$  order approximation of  $\sigma(x)$ . Here  $w_k$  is the  $k^{th}$  order weight and is learned during the training process. In comparison to the coefficients in Equation 17, we get an approximation of the sigmoid activation function.

A finite order polynomial expansion yields an error which can be expressed via the Taylor Remainder Theorem (Firey, 1960),

$$|\sigma(x) - \sigma_{VNN}(x)| \leq R_k = \left| \frac{\sigma^{k+1}(m)}{(k+1)!} (x-a)^{k+1} \right| \quad (19)$$

where the Taylor Expansion is centered around  $a$  and  $m$  lies between  $a$  and  $x$ . ■

Figure 4 shows the approximation of various activation functions via the Volterra series formulation. Note that the  $w'_k$ s in Equation 18 are actually learnable weights and are updated on the basis of the classification error when comparing the classifier output with the ground truth. N.B. The coefficients in Equation 17 are for a generic "sigmoid" function, and are to be distinguished from the learnable coefficients when using it as a local approximation of the data.

**Proposition 2** *A VNN approximation independently achieves optimal weighting of higher order moments and cross-moments of data, for a more adaptive representation than that possible by activation functions.*

**Proof** Consider an input  $X = [x_1, x_2]$ . A standard single layer neural network would take the dot product of  $X$  with the linear weights  $W = [w_1, w_2]$  and apply an activation function  $\sigma(\cdot)$  to get the output  $y = \sigma(W^T \cdot X) = \sigma(w_1x_1 + w_2x_2)$ .

As discussed in Proposition 1, the activation function (eg. ReLU, sigmoid, tanh)  $\sigma(\cdot)$  can be approximated by its Taylor series expansion. Considering a  $2^{nd}$  order approximation,

$$\begin{aligned} y &= c_0 + c_1(w_1x_1 + w_2x_2) + c_2(w_1x_1 + w_2x_2)^2 \\ &= c_0 + c_1w_1x_1 + c_1w_2x_2 + c_2w_1^2x_1^2 + c_2w_2^2x_2^2 + 2c_2w_1w_2x_1x_2. \end{aligned} \quad (20)$$

This can be rewritten as,

$$y = \alpha_0 + \alpha_1x_1 + \alpha_2x_2 + \alpha_3x_1^2 + \alpha_4x_2^2 + \alpha_5x_1x_2. \quad (21)$$

From Equation 20 and 21,

$$\alpha_0 = c_0; \alpha_1 = c_1w_1; \quad (22)$$

$$\alpha_2 = c_1w_2; \alpha_3 = c_2w_1^2; \quad (23)$$

$$\alpha_4 = c_2w_2^2; \alpha_5 = c_1c_2w_1w_2. \quad (24)$$

As a result,

$$y = \alpha_0 + \alpha_1x_1 + \alpha_2x_2 + \frac{c_2}{c_1}\alpha_1^2x_1^2 + \frac{1}{c_2}\alpha_2^2x_2^2 + \frac{2}{c_1}\alpha_1\alpha_2x_1x_2. \quad (25)$$

It is clear that the moments (monomials) and cross-moments (cross-products) weights are computationally tied. This unavoidable and implicit constraint among the moments and cross-moments make the adaptive approximation of such features unlikely. This clearly scales to higher order approximation highlighting the flexibility of VNN in independently adapting the associated weights.

The additional relative feature preservation limitation (at different scales) of activation functions is reflected in the allocated weighting of higher order according to the Taylor series evolution. For an  $n^{th}$  order taylor approximation we have,

$$\sigma(x) = \sum_{n=0}^N \frac{\sigma^{(n)}(a)}{n!} (x - a)^n, \quad (26)$$

where  $\sigma^{(n)}(a)$  is the  $n^{th}$  derivative of  $\sigma$  at  $a$ . As seen from this equation the  $n^{th}$  coefficient,  $c_n$  is given as  $c_n = \frac{\sigma^{(n)}(a)}{n!}$ . As a result of the  $n!$  in the denominator,  $c_{n+1} < c_n \forall n$ . This leads to decreasing importance of higher order features, and makes it difficult for the model to learn them with no regards to their potentially discriminative role in inference. ■

The Volterra series formulation, in contrast, assigns an independent learnable weight to each term in Equation 21.

## 4.2 Cascaded Volterra Filters: Managing High Order Complexity

A major challenge in learning the afore-mentioned architecture arises when higher order non-linearities are sought. The number of required parameters for a  $K^{th}$  order filter is,

$$\sum_{k=1}^K (L \cdot [2p_1 + 1] \cdot [2p_2 + 1])^k. \quad (27)$$



This complexity increases exponentially when the order is increased, thus making a higher order ( $> 3$ ) Volterra Filter implementation impractical. To alleviate this limitation, we use a cascade of  $2^{nd}$  order Volterra Filters, wherein, the second order filter is repeatedly applied until the desired order is attained.

A  $K^{th}$  order filter is realized by applying the  $2^{nd}$  order filter  $\mathcal{Z}$  times, where,  $K = 2^{2^{(\mathcal{Z}-1)}}$ . If the length of the first filter in the cascade is  $L_1$ , the input video  $\mathbf{X}_{[t-L+1:t]}$  can be viewed as a concatenation of a set of shorter videos,

$$\mathbf{X}_{[t_L:t]} = \left[ \mathbf{X}_{[t_L:t_L+L_1]} \mathbf{X}_{[t_L+L_1:t_L+2L_1]} \cdots \mathbf{X}_{[t_L+(M_1-1)L_1:t_L+M_1L_1]} \right], \quad (28)$$

where  $M_1 = \frac{L}{L_1}$ , and  $t_L = t - L + 1$ . Now, a  $2^{nd}$  order filter  $g_1(\cdot)$  applied on each of the sub-videos leads to the features,

$$\mathbf{F}_{t_{[1:M_1]}^1} = \left[ g_1(\mathbf{X}_{[t_L:t_L+L_1]}) \quad g_1(\mathbf{X}_{[t_L+L_1:t_L+2L_1]}) \cdots g_1(\mathbf{X}_{[t_L+(M_1-1)L_1:t_L+M_1L_1]}) \right]. \quad (29)$$

A second filter  $g_2(\cdot)$  of length  $L_2$  is then applied to the output of the first filter,

$$\mathbf{F}_{t_{[1:M_2]}^2} = \left[ g_2(\mathbf{F}_{t_{[1:L_2]}^1}) \quad g_2(\mathbf{F}_{t_{[L_2+1:2L_2]}^1}) \cdots g_2(\mathbf{F}_{t_{[(M_2-1)L_2+1:(M_2L_2)]}^1}) \right], \quad (30)$$

where,  $M_2 = \frac{M_1}{L_2}$ . Note that the features in the second layer are generated by taking quadratic interactions between those generated by the first layer, hence, leading to  $4^{th}$  order terms.

Finally, for a cascade of  $\mathcal{Z}$  filters, the final set of features is obtained as,

$$\mathbf{F}_{t_{[1:M_{\mathcal{Z}}]}^{\mathcal{Z}}} = \left[ g_{\mathcal{Z}}(\mathbf{F}_{t_{[1:L_{\mathcal{Z}}]}^{\mathcal{Z}-1}}) \quad g_{\mathcal{Z}}(\mathbf{F}_{t_{[L_{\mathcal{Z}+1:2L_{\mathcal{Z}}]}^{\mathcal{Z}-1}}}) \cdots g_{\mathcal{Z}}(\mathbf{F}_{t_{[(M_{\mathcal{Z}}-1)L_{\mathcal{Z}}+1:(M_{\mathcal{Z}}L_{\mathcal{Z}})]}^{\mathcal{Z}-1}}}) \right], \quad (31)$$

where,  $M_{\mathcal{Z}} = \frac{M_{\mathcal{Z}-1}}{L_{\mathcal{Z}}}$ .

Note that these filters can also be implemented in an overlapped structure yielding the following features for the  $z^{th}$  layer,  $z \in \{1, \dots, \mathcal{Z}\}$ ,

$$\mathbf{F}_{t_{[1:M_z]}^z} = \left[ g_z(\mathbf{F}_{t_{[1:L_z]}^{z-1}}) \quad g_z(\mathbf{F}_{t_{[2:L_z+1]}^{z-1}}) \cdots g_z(\mathbf{F}_{t_{[(M_z-1)L_z+1:M_z-1]}^{z-1}}) \right], \quad (32)$$

where  $M_z = M_{z-1} - L_z + 1$ . The implementation of an Overlapping Volterra Neural Network (O-VNN) to find the corresponding feature maps, for an input video is shown in Figure 3.

**Proposition 3** *If  $\mathcal{Z}$   $2^{nd}$  order filters are cascaded following the noted overlapped structure (as in Figure 3), the resulting Volterra Network has an effective order of  $K_{\mathcal{Z}} = 2^{2^{\mathcal{Z}-1}}$ .*

**Proof** Since each layer of the O-VNN is a  $2^{nd}$  order Volterra Filter, the order at the  $\mathcal{Z}^{th}$  layer can be written in terms of the order of the previous layer,

$$K_{\mathcal{Z}} = K_{\mathcal{Z}-1}^2, \quad (33)$$

where,  $K_{\mathcal{Z}-1}$  is the order of the  $(\mathcal{Z} - 1)^{th}$  layer. Since, the O-VNN only consists of  $2^{nd}$  order layers, there exists some  $p$  such that,

$$K_{\mathcal{Z}} = 2^p. \quad (34)$$

From Equations (33) and (34),

$$2^p = K_{\mathcal{Z}-1}^2 \quad (35)$$

Taking  $\log_2$  on both sides,

$$\begin{aligned} \log_2 2^p &= \log_2 K_{\mathcal{Z}-1}^2 \\ \implies p &= 2 \log_2 K_{\mathcal{Z}-1} \\ \implies p &= 2 \log_2 K_{\mathcal{Z}-2}^2 \\ \implies p &= 2^2 \log_2 K_{\mathcal{Z}-2} \\ \implies p &= 2^{(\mathcal{Z}-1)} \log_2 K_1 \end{aligned} \quad (36)$$

Since  $K_1 = 2$  and  $\log_2 2 = 1$ ,

$$p = 2^{\mathcal{Z}-1}. \quad (37)$$

Putting this in Equation (34), we get,

$$K_{\mathcal{Z}} = 2^{2^{\mathcal{Z}-1}}. \quad (38)$$

■

A natural question which arises about the resulting relative complexity of our proposed strategy, is addressed next.

**Proposition 4** *The complexity of a  $K^{\text{th}}$  order cascaded Volterra filter will consist of,*

$$\sum_{z=1}^{\mathcal{Z}} \left[ (L_z \cdot [2p_{1_z} + 1] \cdot [2p_{2_z} + 1]) + (L_z \cdot [2p_{1_z} + 1] \cdot [2p_{2_z} + 1])^2 \right] \quad (39)$$

*parameters.*

**Proof** For a  $2^{\text{nd}}$  order filter ( $K = 2$ ), the number of parameters required is  $\left[ (L \cdot [2p_1 + 1] \cdot [2p_2 + 1]) + (L \cdot [2p_1 + 1] \cdot [2p_2 + 1])^2 \right]$  (from equation 27). When such a filter is repeatedly applied  $\mathcal{Z}$  times, it will lead to  $\sum_{z=1}^{\mathcal{Z}} \left[ (L_z \cdot [2p_{1_z} + 1] \cdot [2p_{2_z} + 1]) + (L_z \cdot [2p_{1_z} + 1] \cdot [2p_{2_z} + 1])^2 \right]$  parameters with order  $K = 2^{2^{(\mathcal{Z}-1)}}$ . ■

Furthermore, if a multi-channel input/output is considered, the number of parameters is,

$$\sum_{z=1}^{\mathcal{Z}} (n_{ch}^{z-1} \cdot n_{ch}^z) \left[ (L_z \cdot [2p_{1_z} + 1] \cdot [2p_{2_z} + 1]) + (L_z \cdot [2p_{1_z} + 1] \cdot [2p_{2_z} + 1])^2 \right], \quad (40)$$

where  $n_{ch}^z$  is the number of channels in the output of the  $z^{\text{th}}$  layer.

### 4.3 System Stability and Convergence

The discussed system can be shown to be stable when the input is bounded, i.e. the system is Bounded Input Bounded Output (BIBO) stable.

**Proposition 5** *An O-VNN with  $\mathcal{Z}$  layers is BIBO stable if  $\forall z \in \{1, \dots, \mathcal{Z}\}$ ,*

$$\sum_{\tau_1, \sigma_{11}, \sigma_{21}} \left| \mathbf{W}^{z1} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \right| + \sum_{\substack{\tau_1, \sigma_{11}, \sigma_{21} \\ \tau_2, \sigma_{12}, \sigma_{22}}} \left| \mathbf{W}^{z2} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} \right| < \infty. \quad (41)$$

**Proof** Consider the  $z^{\text{th}}$  layer in the Cascaded implementation of the Volterra Filter,

$$\mathbf{F}_{[1:M_z]}^z = \left[ g_z(\mathbf{F}_{t[1:L_z]}^{z-1}) \quad g_z(\mathbf{F}_{t[2:L_z+1]}^{z-1}) \cdots g_z(\mathbf{F}_{t[(M_z-1)-L_z+1:(M_z-1)]}^{z-1}) \right], \quad (42)$$

where,  $M_z = M_{z-1} - L_z + 1$ . Then for  $m_z \in \{1, \dots, M_z\}$ ,

$$\left| \mathbf{F}_{\begin{bmatrix} s_1 \\ s_2 \end{bmatrix}}^z \right| = \left| g_z \left( \mathbf{F}_{\begin{bmatrix} m_{z-1}-L_z+1:m_z \\ s_1-p_1:s_1+p_1 \\ s_2-p_2:s_2+p_2 \end{bmatrix}}^{z-1} \right) \right| \quad (43)$$

$$\begin{aligned} &= \left| \sum_{\tau_1, \sigma_{11}, \sigma_{21}} \mathbf{W}^{z1} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} f_{\begin{bmatrix} (L_z+m_z-1)-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix}}^{z-1} \right. \\ &+ \left. \sum_{\substack{\tau_1, \sigma_{11}, \sigma_{21} \\ \tau_2, \sigma_{12}, \sigma_{22}}} \mathbf{W}^{z2} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} f_{\begin{bmatrix} (L_z+m_z-1)-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix}}^{z-1} f_{\begin{bmatrix} (L_z+m_z-1)-\tau_2 \\ s_1-\sigma_{12} \\ s_2-\sigma_{22} \end{bmatrix}}^{z-1} \right| \quad (44) \end{aligned}$$

$$\begin{aligned} &\leq \sum_{\tau_1, \sigma_{11}, \sigma_{21}} \left| \mathbf{W}^{z1} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \right| \left| f_{\begin{bmatrix} (L_z+m_z-1)-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix}}^{z-1} \right| \\ &+ \sum_{\substack{\tau_1, \tau_2 \\ \sigma_{11}, \sigma_{12} \\ \sigma_{21}, \sigma_{22}}} \left| \mathbf{W}^{z2} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} \right| \left| f_{\begin{bmatrix} (L_z+m_z-1)-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix}}^{z-1} \right| \left| f_{\begin{bmatrix} (L_z+m_z-1)-\tau_2 \\ s_1-\sigma_{12} \\ s_2-\sigma_{22} \end{bmatrix}}^{z-1} \right| \quad (45) \end{aligned}$$

$$\leq A \sum_{\tau_1, \sigma_{11}, \sigma_{21}} \left| \mathbf{W}^{z1} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \right| + A^2 \sum_{\substack{\tau_1, \sigma_{11}, \sigma_{21} \\ \tau_2, \sigma_{12}, \sigma_{22}}} \left| \mathbf{W}^{z2} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} \right|. \quad (46)$$

Note that Equation 46 states that a bounded input yields  $\sum_{\tau_1, \sigma_{11}, \sigma_{21}} \left| f_{\begin{bmatrix} (L_z+m_z-1)-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix}}^{z-1} \right| \leq A$ , for some  $A < \infty$ . Hence, the sufficient condition for the system to be BIBO stable is,

$$\sum_{\tau_1, \sigma_{11}, \sigma_{21}} \left| \mathbf{W}^{z1} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \right| + \sum_{\substack{\tau_1, \sigma_{11}, \sigma_{21} \\ \tau_2, \sigma_{12}, \sigma_{22}}} \left| \mathbf{W}^{z2} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} \right| < \infty. \quad (47)$$

If the input data (i.e. video frames) is bounded, so is the output of each layer provided that Equation 47 is satisfied  $\forall z \in \{1, \dots, Z\}$ , making the entire system BIBO stable. ■

**Proposition 6** *An O-VNN is stable and convergent under the condition  $|x_t| < \rho < \infty$ , where  $x_t$  is the input to the filter, and  $\rho$  is the radius of convergence for the proposed Volterra Filter.*

**Proof** A Volterra Filter can be viewed as a power series,

$$y_t = \sum_{k=1}^K g_k[ax_t] = \sum_{k=1}^K a^k g_k[x_t], \quad (48)$$

where  $a$  is an amplification factor and,

$$g_k[x_t] = \sum_{\tau_1, \dots, \tau_k} \mathbf{W}_{[\tau_1, \dots, \tau_k]}^k x_{t-\tau_1} x_{t-\tau_2} \dots x_{t-\tau_k}. \quad (49)$$

In general, for a power series  $\sum_{k=1}^{\infty} c_k x^k$ , converges only for  $|x| < \rho$ , where  $\rho = (\lim_{k \rightarrow \infty} \sup |c_k|^{1/k})^{-1}$  (Rudin et al., 1964). Setting  $a = 1$  in Equation 48, and replacing the coefficients  $c_k$  with the  $k^{\text{th}}$  order Volterra Kernel  $\mathbf{W}^k$ ,

$$\rho = (\lim_{k \rightarrow \infty} \sup |\mathbf{W}^k|^{1/k})^{-1}. \quad (50)$$

Furthermore, since the system must also satisfy the BIBO stability condition,

$$|x_t| < (\lim_{k \rightarrow \infty} \sup |\mathbf{W}^k|^{1/k})^{-1} < \infty. \quad (51)$$

■

#### 4.4 Synthesis and Efficient Implementation of Volterra Kernels

As noted earlier, the linear kernel (1<sup>st</sup> order) of the Volterra filter is similar to the convolutional layer in the conventional CNNs. As a result, it can easily be implemented using the 3D convolution function in Tensorflow (Abadi et al., 2016). The second order kernel may be approximated as a product of two 3-dimensional matrices (i.e. a separable operator),

$$\mathbf{W}_{L \times P_1 \times P_2 \times L \times P_1 \times P_2}^2 = \sum_{q=1}^Q \mathbf{W}_{a_{qL \times P_1 \times P_2 \times 1}}^2 \mathbf{W}_{b_{q1 \times L \times P_1 \times P_2}}^2, \quad (52)$$

where,  $P_1 = 2p_1 + 1$ , and  $P_2 = 2p_2 + 1$ , and  $p_1, p_2$  specify the spatial translations (horizontal and vertical). Accounting for Equation 13 yields,

$$g \left( \mathbf{X} \begin{bmatrix} t-L+1:t \\ s_1-p_1:s_1+p_1 \\ s_2-p_2:s_2+p_2 \end{bmatrix} \right) = \sum_{\tau_1, \sigma_{11}, \sigma_{21}} \mathbf{W}^1 \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} x \begin{bmatrix} t-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix} \\ + \sum_{\substack{\tau_1, \sigma_{11}, \sigma_{21} \\ \tau_2, \sigma_{12}, \sigma_{22}}}^Q \sum_{q=1}^Q \mathbf{W}^2_{a_q} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \mathbf{W}^2_{b_q} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} x \begin{bmatrix} t-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix} x \begin{bmatrix} t-\tau_2 \\ s_1-\sigma_{12} \\ s_2-\sigma_{22} \end{bmatrix} \quad (53)$$

$$= \sum_{\tau_1, \sigma_{11}, \sigma_{21}} \mathbf{W}^1 \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} x \begin{bmatrix} t-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix} \\ + \sum_{q=1}^Q \sum_{\substack{\tau_1, \sigma_{11}, \sigma_{21} \\ \tau_2, \sigma_{12}, \sigma_{22}}} \left( \mathbf{W}^2_{a_q} \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} x \begin{bmatrix} t-\tau_1 \\ s_1-\sigma_{11} \\ s_2-\sigma_{21} \end{bmatrix} \right) \left( \mathbf{W}^2_{b_q} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} x \begin{bmatrix} t-\tau_2 \\ s_1-\sigma_{12} \\ s_2-\sigma_{22} \end{bmatrix} \right). \quad (54)$$

A larger  $Q$  will provide a better approximation of the  $2^{nd}$  order kernel. The advantage of this class of approximation is two-fold. Firstly, the number of parameters can further be reduced, if for the  $z^{th}$  layer,  $(L_z \cdot [2p_{1z} + 1] \cdot [2p_{2z} + 1])^2 > 2Q(L_z \cdot [2p_{1z} + 1] \cdot [2p_{2z} + 1])$ . A trade-off between performance and acceptable computational complexity must be accounted for when adopting such an approximation. Additionally, this makes it easier to implement the higher order kernels in Tensorflow (Abadi et al., 2016) by using the built in convolutional operator.

The complexity of the approximate quadratic layers in the Cascaded Volterra Filter (see Figure 3) is reflected by the number of parameters as

$$\sum_{z=1}^Z \left[ (L_z \cdot [2p_{1z} + 1] \cdot [2p_{2z} + 1]) + 2Q(L_z \cdot [2p_{1z} + 1] \cdot [2p_{2z} + 1]) \right]. \quad (55)$$

**Proposition 7** *The VNN second order kernel approximation in Equation 52 is a  $Q^{th}$  rank approximation of the exact quadratic kernel  $\mathbf{W}^2$ .*

**Proof** For simplicity, consider a 1-D Volterra Filter with memory  $L$ . The quadratic weight matrix,  $\mathbf{W}^2$  in such a case is of size  $L \times L$ , and Equation (52) becomes,  $\mathbf{W}^{2(Q)}_{L \times L} = \sum_{q=1}^Q \mathbf{W}^2_{a_q L \times 1} \mathbf{W}^2_{b_q 1 \times L}$ . Consider the Singular Value Decomposition of the quadratic weight matrix,  $\mathbf{W}^2$ ,

$$\mathbf{W}^2 = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T, \quad (56)$$

where,  $\mathbf{U}$  and  $\mathbf{V}$  are  $L \times L$  matrices, and  $\mathbf{\Sigma}$  is a diagonal matrix with singular values on the diagonal. Equation (56) can be re-written as,

$$\mathbf{W}^2 = \sum_{q=1}^L u_q \sigma_q v_q^T, \quad (57)$$

where,  $u_q$  and  $v_q$  are the  $q^{th}$  column of  $\mathbf{U}$  and  $\mathbf{V}$  respectively, and  $\sigma_q$  is the  $q^{th}$  diagonal element of  $\Sigma$ . A  $Q^{th}$  rank approximation is then given as,

$$\begin{aligned} \mathbf{W}^{2(Q)} &= \sum_{q=1}^Q u_q \sigma_q v_q^T \\ &= \sum_{q=1}^Q \hat{u}_q v_q^T, \end{aligned} \quad (58)$$

where,  $\hat{u}_q = u_q \cdot \sigma_q$ . If  $\mathbf{W}_{a_q}^2 = \hat{u}_q$  and  $\mathbf{W}_{b_q}^2 = v_q^T$ ,

$$\mathbf{W}^{2(Q)} = \sum_{q=1}^Q \mathbf{W}_{a_q}^2 \mathbf{W}_{b_q}^2, \quad (59)$$

hence, confirming the approximation given in Equation 52 as a  $Q^{th}$  rank approximation of the exact quadratic kernel.  $\blacksquare$

The matrices  $\mathbf{W}_{a_q}^2$  and  $\mathbf{W}_{b_q}^2$  are unknown beforehand, and will be learned as part of the training process driven by the classification performance of the system. We will evaluate and compare both approaches when implementing the second order kernel (i.e. approximation and exact method) in Section 6. Figure 5 illustrates the implementation of a  $2^{nd}$  order filter using a  $Q^{th}$  rank approximation.

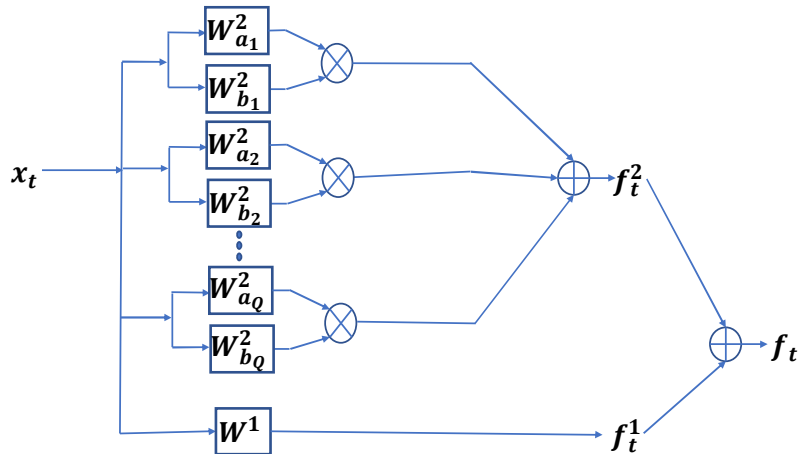


Figure 5: Implementation of a second order Volterra Filter using  $Q^{th}$  rank approximation

#### 4.5 Relation to LSTM Networks

In this section we compare the proposed Volterra Neural Networks with LSTMs, which are frequently used for processing/analyzing temporal data. An LSTM Network cell implemented as discussed in Section 2.4, can be shown to be a special case of the Volterra Filter.

**CASE-1: No Activation Functions**

**Proposition 8** *The cell state of an LSTM network at time  $t$ ,  $C_t$  is a special case of a  $2^{nd}$  order Volterra Filter, where the  $2^{nd}$  order filter is approximated using  $Q = 1$  in Equation 52, and weighed by the cell state at time  $t - 1$ ,  $C_{t-1}$ , i.e.*

$$C_t = C_{t-1} \left( \sum_{j=1}^2 \mathbf{W}_f^j s_j \right) + \sum_{j,k=1}^2 \mathbf{W}_{ic}^{jk} s_j s_k, \quad (60)$$

where,  $\mathbf{s} = [h_{t-1}, x_t]$ , and  $\mathbf{W}_{ic} = \mathbf{W}_i \cdot \mathbf{W}_c$ .

**Proof** From Equations (6), (9), (11), At time  $t=0$ ,

$$C_0 = h_0 = \mathbf{W}_0 x_0. \quad (61)$$

At time  $t=1$ ,

$$C_1 = (W_f^1 h_0 + W_f^2 x_1)(h_0) + (W_i^1 h_0 + W_i^2 x_1)(W_c^1 h_0 + W_c^2 x_1) \quad (62)$$

$$= W_f^1 h_0^2 + W_f^2 x_1 h_0 + W_i^1 h_0 \cdot W_c^1 h_0 + W_i^1 h_0 \cdot W_c^2 x_1 + W_i^2 x_1 \cdot W_c^1 h_0 + W_i^2 x_1 \cdot W_c^2 x_1. \quad (63)$$

Define the matrix  $\mathbf{W}_{ic}$  such that,  $\mathbf{W}_{ic} = \mathbf{W}_i \cdot \mathbf{W}_c$ . This leads to,

$$C_1 = W_f^1 h_0^2 + W_f^2 x_1 h_0 + W_{ic}^{11} h_0^2 + W_{ic}^{12} h_0 x_1 + W_{ic}^{21} x_1 h_0 + W_{ic}^{22} x_1^2 \quad (64)$$

$$= (W_f^1 + W_{ic}^{11}) h_0^2 + (W_f^2 + W_{ic}^{21}) x_1 h_0 + W_{ic}^{21} h_0 x_1 + W_{ic}^{22} x_1^2. \quad (65)$$

Note that the filter  $\mathbf{W}_{ic}$  is equivalent to a  $1^{st}$  rank approximation ( $Q=1$ ) of the second order kernel of a Volterra Filter ( $\mathbf{W}^2 = \sum_{q=1}^Q \mathbf{W}_{qa_L \times 1}^2 \mathbf{W}_{qb_1 \times L}^2$ ). This means that the LSTM can be considered to be a special case of the Volterra Filter where a  $1^{st}$  rank approximation of the  $2^{nd}$  order kernel is used.

At time  $t$ ,

$$C_t = W_f^1 h_{t-1} C_{t-1} + W_f^2 x_t C_{t-1} + W_{ic}^{11} h_{t-1}^2 + W_{ic}^{12} h_{t-1} x_t + W_{ic}^{21} x_t h_{t-1} + W_{ic}^{22} x_t^2 \quad (66)$$

$$= C_{t-1} \left( \sum_{j=1}^2 \mathbf{W}_f^j s_j \right) + \sum_{j,k=1}^2 \mathbf{W}_{ic}^{jk} s_j s_k, \quad (67)$$

where,  $\mathbf{s} = [h_{t-1}, x_t]$ , and  $\mathbf{W}_{ic} = \mathbf{W}_i \cdot \mathbf{W}_c$ . This formulation is equivalent to a  $2^{nd}$  order Volterra Filter applied on the input  $\mathbf{s}$ , with the linear kernel  $\mathbf{W}_f$  and quadratic kernel  $\mathbf{W}_{ic}$ . ■

**Proposition 9** *The output  $h_t$  of the LSTM network at time  $t$ , is a special case of the  $3^{rd}$  order Volterra Filter, where the  $2^{nd}$  and  $3^{rd}$  order filters are approximated using  $Q = 1$  in Equation (52), and weighed by the cell state at time  $t - 1$ ,  $C_{t-1}$ , i.e.*

$$h_t = C_{t-1} \left( \sum_{j,k=1}^2 \mathbf{W}_{of}^{jk} s_j s_k \right) + \sum_{j,k,l=1}^2 \mathbf{W}_{oic}^{jkl} s_j s_k s_l, \quad (68)$$

where,  $\mathbf{s} = [h_{t-1}, x_t]$ ,  $\mathbf{W}_{of} = \mathbf{W}_o \cdot \mathbf{W}_f$ , and  $\mathbf{W}_{oic} = \mathbf{W}_o \cdot \mathbf{W}_i \cdot \mathbf{W}_c$ .

**Proof** Based on Equation (11), the LSTM network output,  $h_t$  can be written as,

$$h_t = (W_o^1 h_{t-1} + W_o^2 x_t) C_t. \quad (69)$$

Replacing  $C_t$  by its expression in Equation (60), and using  $W_{of} = W_o \cdot W_f$  and  $W_{oic} = W_o \cdot W_i \cdot W_c$ ,

$$\begin{aligned} h_t = & W_{of}^{11} h_{t-1}^2 C_{t-1} + W_{of}^{12} h_{t-1} x_t C_{t-1} + W_{of}^{21} x_t h_{t-1} C_{t-1} + W_{of}^{22} x_t^2 C_{t-1} + W_{oic}^{111} h_{t-1}^3 + W_{oic}^{112} h_{t-1}^2 x_t \\ & + W_{oic}^{121} h_{t-1}^2 x_t + W_{oic}^{122} h_{t-1} x_t^2 + W_{oic}^{211} h_{t-1}^2 x_t + W_{oic}^{212} h_{t-1} x_t^2 + W_{oic}^{221} h_{t-1} x_t^2 + W_{oic}^{222} x_t^3. \end{aligned} \quad (70)$$

The above expression can be re-written as,

$$h_t = C_{t-1} \left( \sum_{j,k=1}^2 W_{of}^{jk} s_j s_k \right) + \sum_{j,k,l=1}^2 W_{oic}^{jkl} s_j s_k s_l, \quad (71)$$

where,  $\mathbf{s} = [h_{t-1}, x_t]$ . ■

On the other hand, for an O-VNN with  $\mathcal{Z}$  layers,

$$h_{t_{m_z}}^{\mathcal{Z}} = \sum_{j=0}^{L_{\mathcal{Z}}} W_j^{\mathcal{Z}1} h_{(m_{\mathcal{Z}}+L_{\mathcal{Z}})-j}^{\mathcal{Z}-1} + \sum_{j,k=0}^{L_{\mathcal{Z}}} W_{j,k}^{\mathcal{Z}2} h_{(m_{\mathcal{Z}}+L_{\mathcal{Z}})-j}^{\mathcal{Z}-1} h_{(m_{\mathcal{Z}}+L_{\mathcal{Z}})-k}^{\mathcal{Z}-1}, \quad (72)$$

where,  $m_{\mathcal{Z}} \in [1 : M_{\mathcal{Z}}]$ ,  $M_{\mathcal{Z}} = M_{\mathcal{Z}-1} - L_{\mathcal{Z}} + 1$ , and  $M_1 = t - L_1 + 1$ . In both scenarios, the system uses higher order relations between current and previous samples. In case of an LSTM network the cell state at  $t-1$ ,  $C_{t-1}$  is used in order to select features from previous frames that may be relevant to current frames. On the other hand the Volterra filter formulation explicitly selects the interactions between the frames and weighs them accordingly.

**CASE-2: With Activation Functions** The sigmoid activation function can be approximated as a Taylor series,

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{1}{2} + \frac{1}{4}x - \frac{1}{48}x^3 + \frac{1}{480}x^5 - \dots \quad (73)$$

This can be learned by a Volterra Network if required as it is a polynomial expression. Furthermore as long as the condition in Proposition 6 is satisfied, the series is convergent. Similarly, a tanh activation is also approximated by using a Taylor Series,

$$\tanh(x) = x - \frac{1}{3}x^3 + \frac{2}{15}x^5 - \frac{17}{315}x^7 + \dots \quad (74)$$

This means there is no longer a need to explicitly define the activation function, as the Volterra Neural Network will learn the required activation function as part of the learning/training process.

## 5. Min-Norm Solution and Risk Analysis of a $2^{nd}$ Order Filter

Given the vectorized  $i^{th}$  data sample,  $\mathbf{x}_{i_{1 \times d}} = \{x_i^1, x_i^2, \dots, x_i^d\}$ , with its associated second order terms found by  $\mathbf{x}_i^T \mathbf{x}_i$ , provide a relevant data sample rewritten as,

$$\hat{\mathbf{x}}_{i_{(d+d^2) \times 1}} = [\mathbf{x}_i, \{\text{vec}(\mathbf{x}_i^T \mathbf{x}_i)\}]. \quad (75)$$



Consider  $N$  such samples, i.e.  $i \in \{1, \dots, N\}$ , which will be used for training of the system, this leads to the data matrix,  $\hat{\mathbf{X}}_{N \times (d+d^2)} = \{\hat{\mathbf{x}}_i^T\}_{i \in \{1, \dots, N\}}$ .

Now consider,  $\hat{\mathbf{W}}_{(d+d^2) \times 1} = [\mathbf{W}_{1 \times d}^1, \mathbf{W}_{1 \times d^2}^2]^T$  where,  $\mathbf{W}^1$  and  $\mathbf{W}^2$  are the Volterra Filter weights.

If  $\mathbf{Y}_{N \times 1}$  is the ground truth, we wish to find  $\hat{\mathbf{W}}$  such that,

$$\hat{\mathbf{X}}_{N \times (d+d^2)} \hat{\mathbf{W}}_{(d+d^2) \times 1} = \mathbf{Y}_{N \times 1}. \quad (76)$$

The min-norm solution to the filter weights in this case turns out to be,

$$\hat{\mathbf{W}} = \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1} \mathbf{Y}, \quad (77)$$

where,  $\hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1}$  is the pseudo-inverse of  $\hat{\mathbf{X}}$ , and  $\hat{\mathbf{X}} \hat{\mathbf{X}}^T$  is assumed to have a full rank so that it is invertible.

**Risk Analysis** The risk of the min-norm estimate,  $\hat{\mathbf{W}}$  can be computed as discussed in Bartlett et al. (2019),

$$R(\hat{\mathbf{W}}) = \mathbb{E}_{\hat{\mathbf{x}}, y} (y - \hat{\mathbf{x}}^T \hat{\mathbf{W}})^2 - \mathbb{E} (y - \hat{\mathbf{x}}^T \hat{\mathbf{W}}^*)^2 \quad (78)$$

$$= \mathbb{E}_{\hat{\mathbf{x}}, y} (y - \hat{\mathbf{x}}^T \hat{\mathbf{W}}^* + \hat{\mathbf{x}}^T (\hat{\mathbf{W}}^* - \hat{\mathbf{W}}))^2 - \mathbb{E} (y - \hat{\mathbf{x}}^T \hat{\mathbf{W}}^*)^2 \quad (79)$$

$$= \mathbb{E}_{\hat{\mathbf{x}}} (\hat{\mathbf{x}}^T (\hat{\mathbf{W}}^* - \hat{\mathbf{W}}))^2, \quad (80)$$

where  $\hat{\mathbf{W}}^*$  is the least squares estimate, and  $\mathbb{E}$  is the expectation operator with respect to the observed data and the ground truth. Since  $\hat{\mathbf{W}} = \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1} \mathbf{Y}$ ,

$$R(\hat{\mathbf{W}}) = \mathbb{E}_{\hat{\mathbf{x}}} (\hat{\mathbf{x}}^T (\hat{\mathbf{W}}^* - \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1} \mathbf{Y})) \quad (81)$$

Furthermore, considering distortion due to noise,  $\epsilon = \mathbf{Y} - \hat{\mathbf{X}} \hat{\mathbf{W}}^*$ ,

$$R(\hat{\mathbf{W}}) = \mathbb{E}_{\hat{\mathbf{x}}} (\hat{\mathbf{x}}^T (\mathbf{I} - \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1}) \hat{\mathbf{W}}^* - \hat{\mathbf{x}}^T \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1} \epsilon)^2 \quad (82)$$

$$\leq 2 \mathbb{E}_{\hat{\mathbf{x}}} (\hat{\mathbf{x}}^T (\mathbf{I} - \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1}) \hat{\mathbf{W}}^*)^2 + 2 \mathbb{E}_{\hat{\mathbf{x}}} (\hat{\mathbf{x}}^T \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1} \epsilon)^2 \quad (83)$$

$$= 2 \hat{\mathbf{W}}^{*T} \mathbf{B} \hat{\mathbf{W}}^* + 2 \epsilon^T \mathbf{C} \epsilon, \quad (84)$$

where,  $\mathbf{B} = (\mathbf{I} - \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1}) \hat{\mathbf{\Sigma}} (\mathbf{I} - \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1})$ ,  $\hat{\mathbf{\Sigma}} = \hat{\mathbf{x}} \hat{\mathbf{x}}^T$  and  $\mathbf{C} = (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1} \hat{\mathbf{X}} \hat{\mathbf{\Sigma}} \hat{\mathbf{X}}^T (\hat{\mathbf{X}} \hat{\mathbf{X}}^T)^{-1}$ .

Equation (84) provides an upper bound for the excess risk when using the min-norm estimate  $\hat{\mathbf{W}}$ , in terms of the observed data, the optimal least-squares estimate  $\hat{\mathbf{W}}^*$  and the distortion due to noise,  $\epsilon$ , thus, showing the generalizability of this strategy.

## 5.1 Two-Stream Volterra Networks

Most previous studies in action recognition in videos have noted the importance of using both the spatial and the temporal information for an improved recognition accuracy. As a result, we also propose the use of Volterra filtering in combining the two information streams, exploring a potential non-linear relationship between them. In Section 6 we will verify that this actually boosts the performance, thereby indicating some inherent non-linear relation between the two information

streams. Independent Cascaded Volterra Filters are first used in order to extract features from each modality,

$$\mathbf{F}_{[1:M_Z]}^{\mathcal{Z}^{RGB}} = g_Z^{RGB}(\dots g_2^{RGB}(g_1^{RGB}(\mathbf{X}_{[t-L+1:t]}^{RGB}))) \quad (85)$$

$$\mathbf{F}_{[1:M_Z]}^{\mathcal{Z}^{OF}} = g_Z^{OF}(\dots g_2^{OF}(g_1^{OF}(\mathbf{X}_{[t-L+1:t]}^{OF}))). \quad (86)$$

Upon gleaning features from the two streams, an additional Volterra Filter is solely used for combining the generated feature maps from both modalities,

$$\begin{aligned} F_t^{(RGB+OF)} &= \sum_{\tau_1, \sigma_{11}, \sigma_{21}, u_1} \mathbf{W}^1 \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \\ u_1 \end{bmatrix} f \begin{bmatrix} \mathcal{Z}^{u_1} \\ M_Z - \tau_1 \\ s_1 - \sigma_{11} \\ s_2 - \sigma_{21} \end{bmatrix} \\ &+ \sum_{\substack{\tau_1, \sigma_{11}, \sigma_{21}, u_1 \\ \tau_2, \sigma_{12}, \sigma_{22}, u_2}} \mathbf{W}^2 \begin{bmatrix} \tau_1 \\ \sigma_{11} \\ \sigma_{21} \end{bmatrix} \begin{bmatrix} \tau_2 \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix} f \begin{bmatrix} \mathcal{Z}^{u_1} \\ M_Z - \tau_1 \\ s_1 - \sigma_{11} \\ s_2 - \sigma_{21} \end{bmatrix} f \begin{bmatrix} \mathcal{Z}^{u_2} \\ M_Z - \tau_2 \\ s_1 - \sigma_{12} \\ s_2 - \sigma_{22} \end{bmatrix}, \end{aligned} \quad (87)$$

where  $\tau_j \in [0, L_{Z+1}]$ ,  $\sigma_{1j} \in [-p_1, p_1]$ ,  $\sigma_{2j} \in [-p_2, p_2]$ , and  $u_j \in [RGB, OF]$ . Figure 6-(c) shows the block diagram for fusing the two information streams.

## 6. Experiments and Results

### 6.1 Action Recognition

We proceed to evaluate the performance of this approach on three action recognition datasets, namely, Kinetics-400 (Carreira and Zisserman, 2017), UCF-101 (Soomro et al., 2012) and HMDB-51 (Kuehne et al., 2011). We present two versions of the VNN, a heavier complex version: O-VNN-H and a lighter, device friendly version: O-VNN-L. The performance comparison of the results with recent state of the art implementations on Kinetics-400 is presented in Table 1, and the comparisons on UCF-101 and HMDB-51 are presented in Tables 2 and 3. Table 2 shows the comparison with techniques that only exploit the RGB stream, while Table 3 shows the comparison when both information streams are used. Note that our comparable performance to the state of the art is achieved with a significantly lower number of parameters (see Table 1). Furthermore, a significant boost in performance is achieved by allowing non-linear interaction between the two information streams. The Optical Flow is computed using the TV-L1 algorithm (Zach et al., 2007). Note that we train the network from scratch on both datasets, and do not use a larger dataset for pre-training, in contrast to some of the previous implementations. The implementations that take advantage of a different dataset for pre-training are indicated by a ‘Y’ in the pre-training column, while those that do not, are indicated by ‘N’. When training from scratch the proposed solution is able to achieve best performance for both scenarios: one stream networks (RGB frames only) and two-stream networks (RGB frames & Optical Flow). To fuse the two information streams (spatial and temporal), we evaluate the following techniques:

1. *Decision Level Fusion (Figure 6-(a))*: The decision probabilities  $P_t^{RGB}(a_i)$  and  $P_t^{OF}(a_i)$  are independently computed and are combined to determine the fused probability  $P_t^f(a_i)$  using
  - (a) *Weighted Averaging*:  $P_t^f(a_i) = \beta^{RGB} P_t^{RGB}(a_i) + \beta^{OF} P_t^{OF}(a_i)$ , where  $\beta^{RGB} + \beta^{OF} = 1$ , which control the importance/contribution of the RGB and Optical Flow streams towards making a final decision, or
  - (b) *Event Driven Fusion (Roheda et al., 2018a, 2019)*:  $P_t^f(a_i) =$

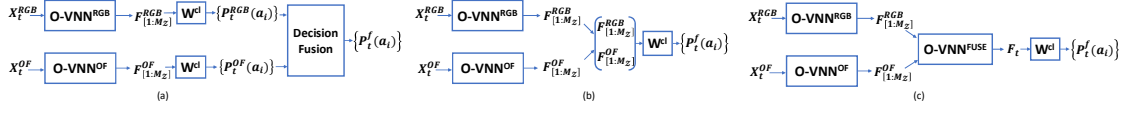


Figure 6: (a): Decision Level Fusion, (b): Feature Concatenation, (c): Two-Stream Volterra Filtering

$\gamma P_t^{\text{MAX MI}}(a_i^{\text{RGB}}, a_i^{\text{OF}}) + (1 - \gamma) P_t^{\text{MIN MI}}(a_i^{\text{RGB}}, a_i^{\text{OF}})$ , where  $\gamma$  is a pseudo measure of correlation between the two information streams,  $P_t^{\text{MAX MI}}(\cdot)$  is the joint distribution with maximal mutual information, and  $P_t^{\text{MIN MI}}(\cdot)$  is the joint distribution with minimal mutual information.

2. *Feature Level Fusion*: Features are extracted from each stream independently, and are subsequently merged before making a decision. For this level of fusion we consider a simple *Feature Concatenation* (see Figure 6-(b)), and *Two-Stream Volterra Filtering* (see Section 5.1, Figure 6-(c)).

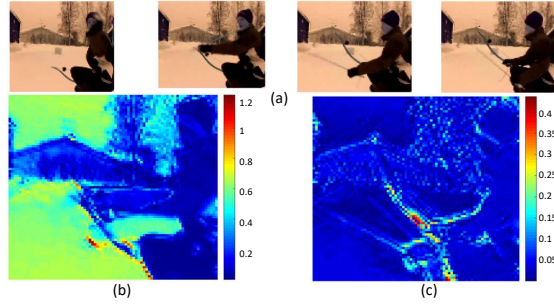


Figure 7: (a): Input Video, (b): Features extracted by only RGB stream, (c): Features extracted by Two-Stream Volterra Filtering

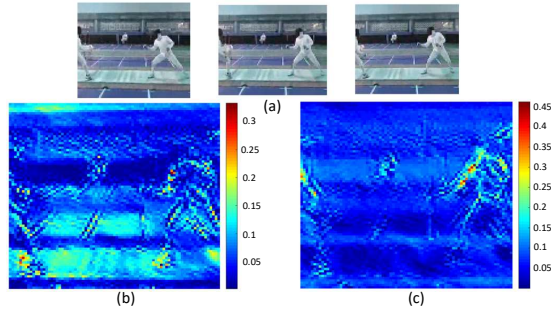


Figure 8: (a): Input Video, (b): Features extracted by only RGB stream, (c): Features extracted by Two-Stream Volterra Filtering

The techniques are summarized in Figure 6. The complex version of the proposed approach (O-VNN-H) comprises of 7  $2^{nd}$  order layers in both the RGB and the Optical stream. Each layer uses  $L_z = 2$  and  $p_{1z}, p_{2z} \in \{0, 1, 2\}$ . The outputs of the RGB filter and the Optical Flow filter

Method	Avg Accuracy Kinetics-400	Number of Parameters	GFLOPs
Temporal 3D ConvNets (Diba et al., 2017a)	69.8 %	56M	63.34
SpatioTemporal Convolution (Tran et al., 2018)	71.3 %	34M	43.72
Two-Stream Inflated 3D CNN (Carreira and Zisserman, 2017)	74.2 %	25M	37.43
Slowfast ResNet-50 (Feichtenhofer et al., 2019)	77.0 %	34M	50.58
Slowfast ResNet-101 (Feichtenhofer et al., 2019)	<b>78.5 %</b>	62M	96.79
Two Stream O-VNN-L	75.9 %	<b>10M</b>	<b>12.30</b>
<b>Two Stream O-VNN-H</b>	<b>78.8 %</b>	<b>26M</b>	<b>29.85</b>

Table 1: Performance Evaluation for Two-Stream networks (RGB &amp; Optical Flow) on Kinetics-400

Method	Pre-Training	Avg Accuracy UCF-101	Avg Accuracy HMDB-51
Slow Fusion (Karpathy et al., 2014)	Y (Sports-1M)	64.1 %	-
Deep Temporal Linear Encoding Networks (Diba et al., 2017b)	Y (Sports-1M)	86.3 %	60.3 %
Inflated 3D CNN (Carreira and Zisserman, 2017)	Y (ImageNet + Kinetics)	<b>95.1 %</b>	<b>74.3 %</b>
O-VNN-H	Y (Kinetics)	<b>95.3 %</b>	<b>75.1 %</b>
Soomro et al, 2012	N	43.9 %	-
Single Frame CNN (Karpathy et al., 2014; Krizhevsky et al., 2012)	N	36.9 %	-
Slow Fusion (Karpathy et al.; Baccouche et al.; Ji et al.)	N	41.3 %	-
3D-ConvNet (Carreira and Zisserman, 2017; Tran et al., 2015)	N	51.6 %	24.3 %
Volterra Filter	N	38.19 %	18.76 %
O-VNN-H	N	<b>58.73 %</b>	<b>29.33 %</b>
O-VNN-L	N	53.77 %	25.76 %

Table 2: Performance Evaluation for one stream networks (RGB only): The proposed algorithm achieves best performance when trained from scratch

<b>Method</b>	<b>Pre-Training</b>	<b>Avg Accuracy UCF-101</b>	<b>Avg Accuracy HMDB-51</b>
Two-Stream CNNs (Simonyan and Zisserman, 2014)	<i>Y</i> (ILSVRC-2012)	88.0 %	72.7 %
Deep Temporal Linear Encoding Networks (Diba et al., 2017b)	<i>Y</i> (BN-Inception + ImageNet)	95.6 %	71.1 %
Two Stream Inflated 3D CNN (Carreira and Zisserman, 2017)	<i>Y</i> (ImageNet + Kinetics)	98.0 %	80.9 %
Two-Stream O-VNN-H	<i>Y</i> (Kinetics)	<b>98.49 %</b>	<b>82.63 %</b>
Two Stream Inflated 3D CNN (Carreira and Zisserman, 2017)	<i>N</i>	88.8 %	62.2 %
Weighted Averaging: O-VNN-H	<i>N</i>	85.79 %	59.13 %
Weighted Averaging: O-VNN-L	<i>N</i>	81.53 %	55.67 %
Event Driven Fusion: O-VNN-H	<i>N</i>	85.21 %	60.36 %
Event Driven Fusion: O-VNN-L	<i>N</i>	80.37 %	57.89 %
Feature Concatenation: O-VNN-H	<i>N</i>	82.31 %	55.88 %
Feature Concatenation: O-VNN-L	<i>N</i>	78.79 %	51.08 %
Two-Stream O-VNN-H	<i>N</i>	<b>90.28 %</b>	<b>65.61 %</b>
Two-Stream O-VNN-L	<i>N</i>	86.16 %	62.45 %

Table 3: Performance Evaluation for two stream networks (RGB & Optical Flow): The proposed algorithm achieves best performance on both datasets.

are then fed to a fusion layer with  $L_{\text{Fuse}} = 2$  and  $p_{1\text{Fuse}}, p_{2\text{Fuse}} \in \{0, 1, 2\}$ . Similarly, a lighter version of the proposed method (O-VNN-L) is designed with 5  $2^{\text{nd}}$  order layers. It is clear from Table 3 that performing fusion using Volterra Filters significantly boosts the performance of the system. This shows that there does exist a non-linear relationship between the two modalities. This can also be confirmed from the fact that we can see significant values in the weights for the fusion layer (see Table 4). Figures 7 and 8 show one of the feature maps for an archery video and a

	$u =$ $RGB$	$u = OF$	$u = Fusion$
$\frac{1}{2} \ \mathbf{W}^u\ _2^2$	352.15	241.2	341.3

Table 4: Norm of  $\mathbf{W}^u$ , where  $u \in [RGB, OF, Fusion]$ .

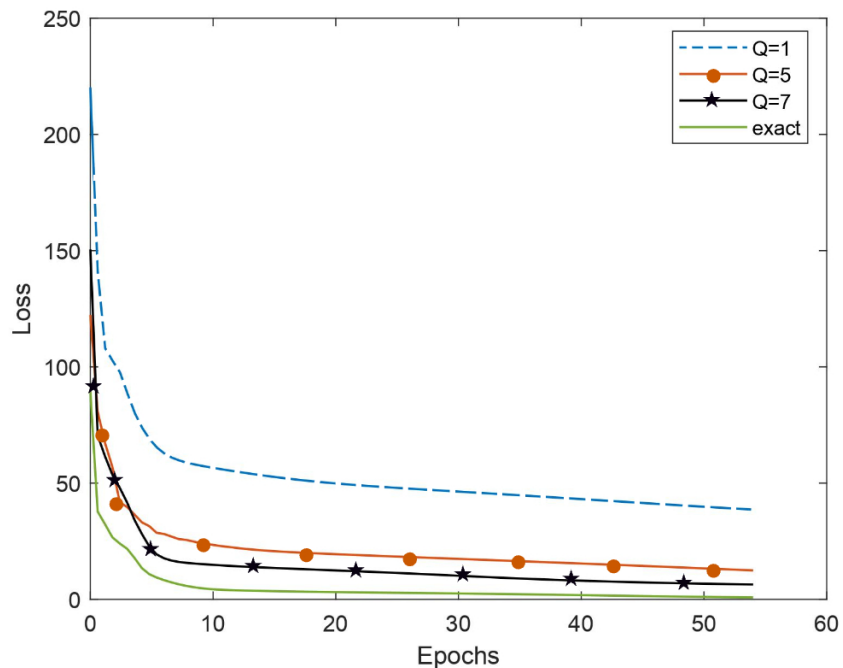


Figure 9: Epochs vs Loss for various number of multipliers for a Cascaded Volterra Filter

fencing video. From Figures 7, 8-(b),(c) it can be seen that when only the RGB stream is used, a lot of the background area has high values, while on the other hand, when both streams are jointly used, the system is able to concentrate on more relevant features. In 7-(c), the system is seen to

concentrate on the bow and arrow which are central to recognizing the action, while in 8-(c) the system is seen to concentrate on the pose of the human which is central to identifying a fencing action. Figure 9 shows the Epochs vs Loss graph for a Cascaded Volterra Filter when a different number of multipliers ( $Q$ ) are used to approximate the  $2^{nd}$  order kernel. The green plot shows the loss when the exact kernel is learned, and it can be seen that the performance comes closer to the exact kernel as  $Q$  is increased.

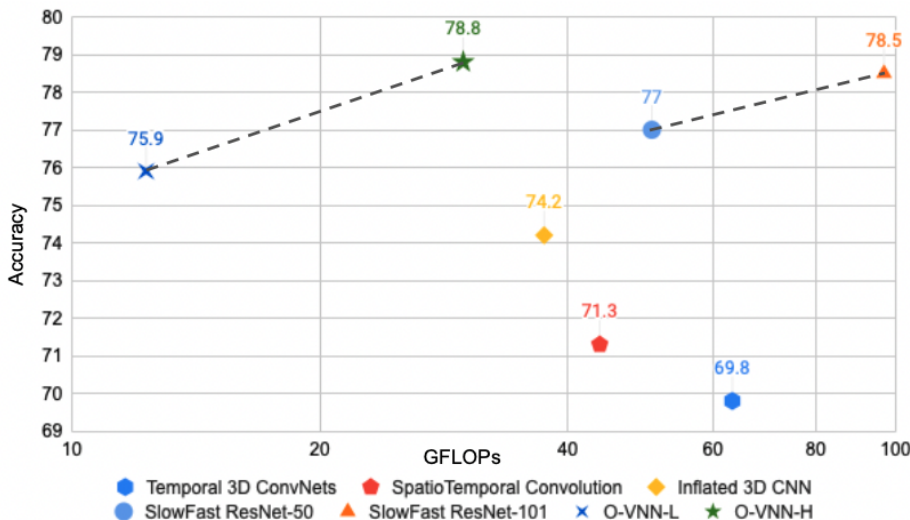


Figure 10: GFLOPs vs Accuracy for various Action Recognition methods discussed in Table 1

Figure 10 depicts the plot for GFLOPs vs Accuracy of the various models compared in Table 1 and highlights the reduction in model complexity achieved by the proposed method while achieving SOTA performance in action recognition.

Figure 11 evaluates the robustness of the proposed model (VNN) in presence of Gaussian noise and compares it with that of a CNN model (Carreira and Zisserman, 2017) for action recognition. It is observed that the VNN model is much more robust to Gaussian noise and provides up to 20 % improvement in classification accuracy. Furthermore, the VNN model experiences a graceful drop in performance compared to the CNN model which sees a severe drop at Signal to Noise Ratio (SNR) of 15 dB.

Figure 12 evaluates performance when the number of Frames Per Second (FPS) is reduced. This illustrates that the VNN is better at modelling long term relationships in the time domain as compared to a CNN model. While unlike in the case of Gaussian noise, the VNN does see a significant drop in performance, it comes at a lower frame rate of 5 fps as compared to the 10 fps in the case of CNN.

### 6.2 Image Generation

To further demonstrate the capabilities of this VNN-based architecture we proceed to evaluate its generative capacity in a Generative Adversarial Network (Goodfellow et al., 2014) using our proposed formulation. Furthermore, we also use the VNN architecture jointly with a fixed dictionary as has been recently demonstrated in Stable GANs (STGANs) (Krim et al.) to stabilize the training process.

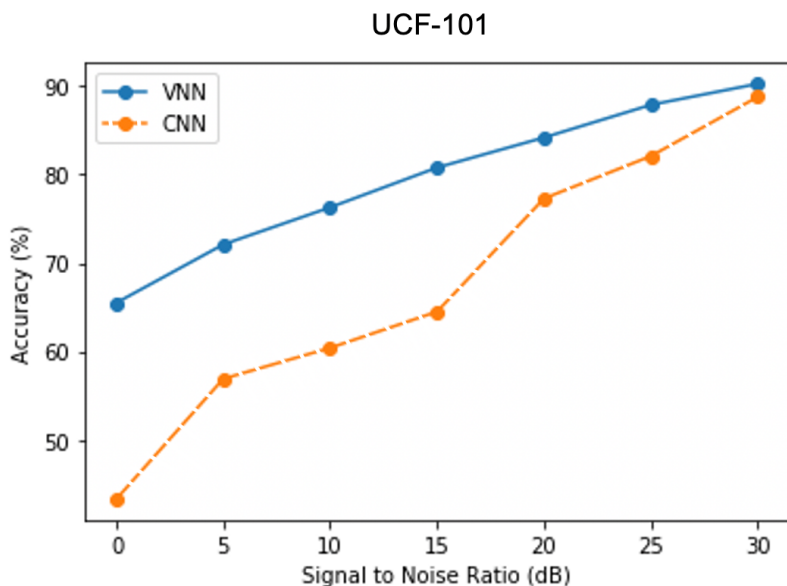


Figure 11: Performance comparison between VNN and CNN implementations when noise is added to the input videos

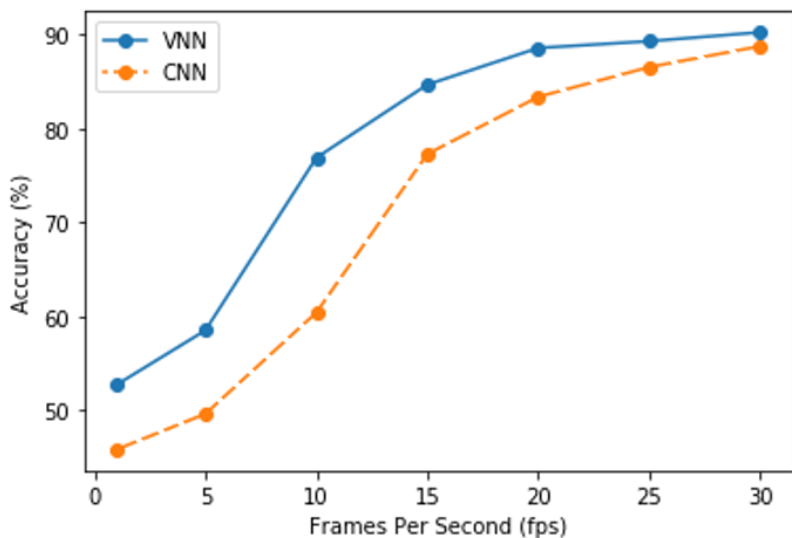


Figure 12: Performance comparison between VNN and CNN implementations for different fps

We design the experiment using the CIFAR10 dataset of 60,000  $32 \times 32$  color images of objects from 10 classes by allotting 50,000 images for training and 10,000 images for validation. To generate a  $32 \times 32$  image, the input noise vector is first transformed into a latent space using a fixed learned frame  $\Theta^*$ . The generator uses two  $2^{nd}$  order Volterra filter layers as opposed to four convolutional layers in Krim et al.. The number of channels in the final layer is such that it is equal to the number of atoms in the fixed dictionary, which in this experiment is 384. The learned fixed



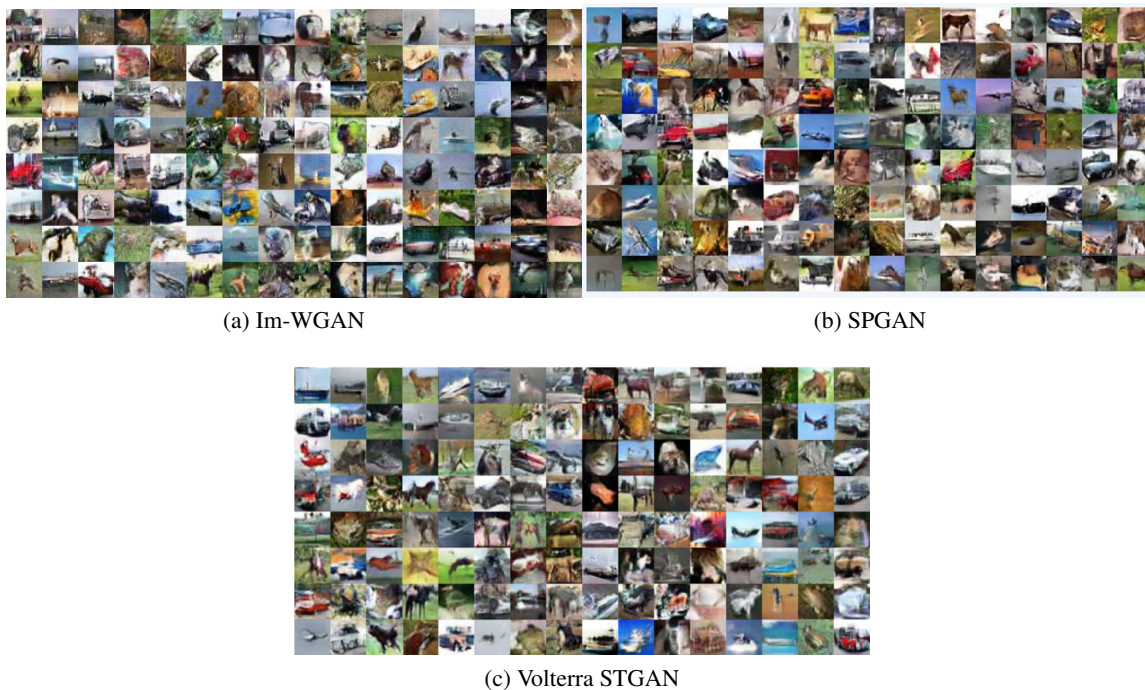


Figure 13: Generated images using Cifar10 Dataset

Method	Inception Score
WGAN	5.95
VNN-WGAN	6.20
SPGAN-recon	6.70
STGAN	6.80
Volterra STGAN	<b>7.05</b>

Table 5: Comparison of various GAN based methods with proposed VNN STGAN in terms of inception score.

dictionary  $\Theta^*$  is then multiplied by the generator output to produce generated image patches. The size of the dictionary in this experiment was  $75 \times 384$ . Figure 13 (a)-(c) shows the generated cifar10 images in this experiment and compares them with SPGAN (Mahdizadehghadam et al., 2019) and Im-WGAN (Gulrajani et al., 2017). The inception score of the proposed approach is compared with other existing approaches in Table 5 with a rather limited dictionary of Gaussians in one case (the performance would improve as the diversity of the dictionaries increases).

## 7. Conclusion

We proposed a novel network architecture for recognition of actions in videos, where the non-linearities were introduced by a Volterra Series Formulation. We propose a Cascaded Volterra Filter which leads to a significant reduction in parameters while exploring the same complexity of non-linearities in the data and attaining competitive and even better performance than SOA CNN. Such a Cascaded Volterra Filter was also shown to be a BIBO stable system. In addition, we also proposed the use of the Volterra Filter to fuse the spatial and temporal streams, hence leading to a non-linear fusion of the two streams. The combined competitive performance and a significant reduction in parameter as well as sample complexity make VNN a viable alternative to SOA CNN.

## References

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.
- Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *International workshop on human behavior understanding*, pages 29–39. Springer, 2011.
- Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.
- Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Hossein Karami, Mohammad Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv preprint arXiv:1711.08200*, 2017a.
- Ali Diba, Vivek Sharma, and Luc Van Gool. Deep temporal linear encoding networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2329–2338, 2017b.
- Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1915–1929, 2012.
- Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

- Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.
- William J Firey. Remainder formulae in taylor’s theorem. *The American Mathematical Monthly*, 67(9):903–905, 1960.
- Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.
- Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.
- Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, 2018.
- Hamid Krim, S Jaffard, Siddharth Roheda, Shahin Mahdizadehaghdam, and Ashkan Panahi. On stabilizing generative adversarial networks (stgans).
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

- Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.
- Ritwik Kumar, Arunava Banerjee, Baba C Vemuri, and Hanspeter Pfister. Trainable convolution filters and their application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1423–1436, 2011.
- Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnns for fine-grained visual recognition. *arXiv preprint arXiv:1504.07889*, 2015.
- Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos in the wild. Citeseer, 2009.
- Shahin Mahdizadehghdam, Ashkan Panahi, and Hamid Krim. Sparse generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on computer vision*, pages 392–405. Springer, 2010.
- Stanislaw Osowski and T Vu Quang. Multilayer neural network structure as volterra filter. In *Proceedings of IEEE International Symposium on Circuits and Systems-ISCAS'94*, volume 6, pages 253–256. IEEE, 1994.
- Siddharth Roheda, Hamid Krim, Zhi-Quan Luo, and Tianfu Wu. Decision level fusion: An event driven approach. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2598–2602. IEEE, 2018a.
- Siddharth Roheda, Benjamin S Riggan, Hamid Krim, and Liyi Dai. Cross-modality distillation: A case for conditional generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2926–2930. IEEE, 2018b.
- Siddharth Roheda, Hamid Krim, Zhi-Quan Luo, and Tianfu Wu. Event driven fusion. *arXiv preprint arXiv:1904.11520*, 2019.
- Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.
- Martin Schetzen. The volterra and wiener theories of nonlinear systems. 1980.
- Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.
- Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

- Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.
- Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- Marshall H Stone. The generalized weierstrass approximation theorem. *Mathematics Magazine*, 21(5):237–254, 1948.
- Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
- Vito Volterra. *Theory of functionals and of integral and integro-differential equations*. Courier Corporation, 2005.
- Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. 2009.
- Sheng Yi, Hamid Krim, and Larry K Norris. Human activity modeling as brownian motion on shape manifold. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 628–639. Springer, 2011.
- Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007.
- Georgios Zoumpourlis, Alexandros Doumanoglou, Nicholas Vretos, and Petros Daras. Non-linear convolution filters for cnn-based learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4761–4769, 2017.