Analyzing Multi-Channel Networks for Gesture Recognition

Pradyumna Narayana
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
prady@cs.colostate.edu

J. Ross Beveridge
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
ross@cs.colostate.edu

Bruce A. Draper
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
draper@cs.colostate.edu

***Abstract*—Multi-channel architectures are becoming increasingly common, setting the state-of-the-art for performance in gesture recognition challenges. Unfortunately, we lack a clear explanation of why multi-channel architectures outperform single channel ones. This paper considers two hypotheses. The *Bagging* hypothesis says that multi-channel architectures succeed because they average the result of multiple unbiased weak estimators in the form of different channels. The *Society of Experts* (SoE) hypothesis suggests that multi-channel architectures succeed because the channels differentiate themselves, developing expertise with regard to different aspects of the data.**

To distinguish between these hypotheses, this paper reports on two experiments. The first measures the drop in individual channel performance when the input is degraded by removing high frequency, color, or motion information. The second looks at fusion weights relative to gesture properties. Both experiments support the SoE hypothesis, suggesting multi-channel architectures succeed because of channel specialization.

I. Introduction

Gestures are a common form of human communication and important for human computer interfaces [37], [38], [9], [29]. There is a strong trend in gesture recognition research toward multi-channel architectures, where two or more convolutional networks process different versions of the same video in parallel [19], [20], [21], [18]. Karpathy *et al*.were the first to use a two-channel architecture [7]. Their system had a global channel that processed low-resolution versions of the whole image, and a focused channel that processed higher-resolution versions of the middle of the image. Simonyan and Zisserman were the first to divide processing by modality, with one channel processing RGB images and the other flow field [32]. They initially fused the results at the softmax layer, although later versions experimented with other fusion schemes [4], [2].

The justification for multi-channel architectures is not immediately obvious. It is easy to encode RGB data, depth data, flow field vectors and more as bands of input images, allowing all the same information to be fed to a single channel, where the data can be combined without restrictions. Yet somehow dividing processing into separate channels improves performance. Pigou *et al*.use multiple channels for sign language recognition [27], while Molchanov *et al*.use them for recognizing gestures in cars [15], [16]. Neverova *et al*.use multiple channels for different temporal scales [22]. More directly relevant to this paper, all the entries in the ChaLearn ICCV 2017 gesture recognition challenge used multi-channel architectures [34], [13], [36], [40]. Miao *et al*.won the competition with a three-channel architecture, with channels for RGB, depth, and flow field data [13]. Subsequently, Narayana *et al*.established a new state-of-the-art (SOA) for performance on ChaLearn using 12 channels, one for every combination of modality (RGB, depth, RGB flow field, or depth flow field) and focus target (right hand, left hand, and whole image) [19]. But while multi-channel approaches outperform single channel methods, there is no clear explanation as to why.

This paper considers two possible explanations. One is what we call the *Bagging* hypothesis. It is motivated by the well-known result that strong classifiers can be created by averaging (a.k.a. bagging) the results of many weak classifiers, as long as the weak classifiers are unbiased [1]. The hypothesis here is that every channel is a weak classifier, and the performance of the multi-channel system is improved by the bagging effect. In this model, every channel solves essentially the same problem, and the role of the different modalities and/or attention targets is to keep the channels from being too strongly correlated. The implications of the Bagging hypothesis are that (1) more channels will always be better than fewer, and (2) channel outputs can be combined by averaging, as in [16], [40], [36]. The alternative explanation is what we call the *Society of Experts* (SoE) hypothesis, named in the spirit of Marvin Minsky [14]. The SoE hypothesis asserts that channels specialize to the modalities and/or attention targets they are assigned to. Each channel specializes in particular sources of information and therefore become better at recognizing some gesture classes at the expense of being worse at others. The implications of the SoE hypothesis are that (1) adding a channel is only beneficial if it adds a new source of information, and (2) the outputs of channels should be combined selectively, usually through trained fusion networks as in [4], [19], [13], [7].

The Bagging and SoE hypotheses are actually two points along a spectrum of possible models. In the bagging hypothesis, all channels share similar expertise and the difference between them is estimation variance. In the SoE hypothesis, channels have unique expertise. In between are models with varying degrees of overlapping expertise among channels. Most real systems probably lie somewhere between the two extremes, particularly systems with cross-stream fusion [4],

[2]. Nonetheless, it is important to know whether bagging or SoE is the closer model in practice because of the architectural implications. To the extent that systems are bagging, we should add as many channels as possible. We could even have multiple channels processing the same data stream, as long as some other technique is used to make sure the channels are not strictly redundant such as data partitioning. If systems are societies of experts, on the other hand, we should only add channels for strong new sources of information, and we should train fusion networks instead of just averaging results.

We explore the Bagging and SoE hypotheses in the context of the 12-channel FOANet architecture introduced by Narayana *et al*.[19]. FOANet is representative of multi-channel architectures in that the internal structures of the individual channels are quite standard. At the same time, FOANet has SOA performance on the ChaLearn IsoGD dataset [35] and NVIDIA dataset [16] because it has more channels than any other architecture we know of, and it divides channels by both modality and focus of attention target. This paper first looks at how channels respond to properties of their input data. For example, it measures how the performance of a channel degrades when high frequency information is removed by pre-processing the input videos with a low-pass filter. It also looks at how channel performance degrades when color or motion information is removed. The Bagging hypothesis suggests that channels should exploit all the data available to them, and therefore the performance of all channels should degrade similarly when information is removed. The SoE hypothesis, on the other hand, suggests that channels specialize and that each channel exploits limited aspects of the input signal. Therefore some channels should degrade much more severely than others when a specific kind or source of information is removed.

This paper also measures specialization with regard to gesture properties. Some gestures, for example, are one-handed, while others are two-handed. Similarly some gestures involve large arm motions, while others involve localized hand or finger motions and still others are defined by static poses. The SoE hypothesis predicts that channels specialize and should therefore be more useful for gestures with some properties than others. The Bagging hypothesis, on the other hand, predicts that channels are generalists and their performance should not be as strongly tied to specific gesture properties.

In both sets of experiments, we find evidence that FOANet channels are highly specialized, and that they often specialize in explainable and interesting ways. For example, RGB channels specialize in a way that is analogous to differences between human foveal and peripheral vision (see Section III-B). Both sets of experiments can therefore be interpreted as preferring the SoE hypothesis, at least in the context of gesture recognition.

II. Related Work

Deep learning is a field where the practice is ahead of the theory. Efforts to explain the workings of convolutional networks have focused on feature visualization [26], [25], [31], [24], [17], [23], attribution [31], [39], [33], [30], [5], [8], or dimensionality reduction [12]. Most of this work has concentrated on single channel architectures. Recently, Feichtenhofer *et al*.visualized the spatiotemporal representations learned by deep two-stream networks for action recognition [3]. Their work focuses on cross-stream fusion between two channels and shows that cross-stream fusion enables the learning of spatiotemporal features. However, most multi-channel networks are trained independently without cross-channel interactions and with fusion performed at the FC or softmax levels. Our work focuses on the information learned by different channels of multi-channel networks that are trained independently.

III. Relating Channels to Data Properties

In order to determine if the Bagging hypothesis or Society of Experts hypothesis is a better model of multi-channel system behavior, we first investigate the extent to which channels specialize in the information they extract from videos. The methodology can be thought of as a series of data ablation studies. FOANet channels are trained on RGBD videos from the ChaLearn IsoGD dataset. In each experiment we remove a source of information from the videos and measure the drop in performance for each channel. For example, in one experiment we remove high-frequency information, while in another we remove color. The Bagging hypothesis suggests that channels consume all the information available to them, so the relative performance drop when a source of information is removed should be similar. The SoE hypothesis, on the other hand, predicts that channels specialize in the information they exploit, so some channels should degrade much more than others when a source of information is removed.

A. FOANet Channels

FOANet channels are similar to the channels in the seminal two-channel architecture paper by Simonyan and Zisserman [32]. The input to a channel is a stream of video, the type of which depends on the channel (e.g. RGB data, depth data, etc.). At each time step, the channel processes a sliding window of ten frames. This window of data is passed to a deep convolutional neural network, which produces a vector of label probabilities. The major difference between FOANet channels and the channels in [32] is that the SOA in CNNs has improved. FOANet channels use the ResNet-50 architecture [6] for their CNN.

FOANet has 12 independent channels, one for every combination of modality and attention focus targets. Specifically, the 4 modalities of RGB, depth, RGB flow and depth flow are crossed with 3 focus of attention selections: attention to the entire video frame (global), attention focused on just the right hand, and attention focused on just the left hand. The modality of a channel determines the type of input. RGB and depth channels consume 3 band RGB and depth images respectively. Flow field channels also consume 3 band images, but the bands are *dx, dy,* and *magnitude*. The difference between RGB flow

and depth flow is whether the flow fields are computed from RGB images or depth images.

The focus of attention targets in FOANet are less standard. Whereas Karpathy *et al*.positioned their attention window in the middle of the scene, FOANet actively detects and tracks the positions of the left and right hands throughout the video. The left and right hand channels (collectively, the focused channels) are created by selecting an image window that bounds the estimated position of the hand [10], [11]. The image window is extracted from the RGB, depth and flow field images and re-scaled to 128×128 pixels.

Every channel in FOANet has access to motion information, since they analyze 10-frame temporal sliding windows stacked together as 30 bands, as mentioned above. Flow field channels therefore have potential access to acceleration data, i.e. changes in motion over time.

B. Methodology

To understand how performance degrades when a source of information is removed, we look at a channel's response to every 10-frame temporal window. The ChaLearn dataset has 249 gesture classes, so channels produce a 249 element softmax vector in response to every input window. To establish a baseline for each channel, we measure the accuracy of the argmax of this vector on the ChaLearn IsoGD validation and test videos. We then alter the validation and test videos to remove a source of information and measure the relative drop in performance. A total of three experiments are run, removing three different sources of information: motion, high spatial frequencies, and color. The experiment using color is restricted to RGB channels.

1) Removing Motion: FOANet channels classify windows of 10 consecutive frames stacked together as a 30 band image. These windows capture motion information over a 10-frame interval. We remove motion information by taking the middle frame of the window and replicating it 10 times.

2) Removing High Frequency Information: High frequency information is removed from RGB and depth images by convolving them with a Gaussian mask of sigma 3. RGB flow and depth flow images are computed from the smoothed RGB and depth images. Right-hand and left-hand channels are extracted and rescaled from the smoothed images or from the flow fields extracted from smoothed images.

3) Removing Color: Color information is removed from the inputs to RGB channels by transforming RGB images to gray-scale images. As the other modalities (depth, flow fields) do not have color bands, color information is only removed for RGB channels.

C. Results

Figure 1 shows the results of data ablation on the RGB global and RGB right hand channels. The vertical axis is channel accuracy, measured relative to baseline performance on unaltered data. The gray bar, which represents baseline performance, is therefore always 100% for all channels. The blue bars show the performance of the RGB global channel

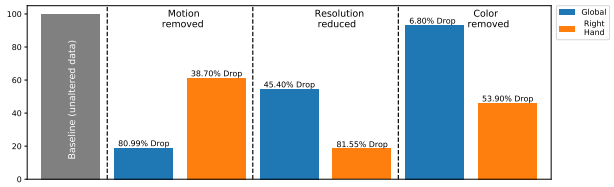

Fig. 1. Bar chart showing the result of data ablation on RGB modality of ChaLearn IsoGD dataset. The vertical axis is relative accuracy, compared to baseline performance on unaltered data. Blue and orange bars show the accuracy of the global and right-hand channels respectively. The pairs of blue and orange bars show relative accuracy after motion information is removed (left), high frequency information is removed (middle), and color information is removed (right).

when motion (left), high frequency (middle) and color (right) data is removed. The orange bars show the same for the RGB right hand channel.

In Figures 1 through 3, data is shown for the global and right hand channels, but not the left hand channel. This is to simplify the figures, because the relative performance of the left hand channel mimics the relative performance of the right hand channel (left hand channel results are given paranthetically in the text). In Section IV, the performance of the left and right hand channels will be broken out graphically.

The most striking result in Figure 1 is shown on the right of the figure. When color information is removed, the accuracy of the RGB right-hand channel is cut roughly in half. Its performance drops by 53.9%, while the performance of the left hand channel (not shown) drops by 54.01%. The RGB global channel, on the other hand, is barely changed. It drops by less than 7%. This suggests that the right and left hand channels depend on color information, but the global channel does not.

The middle of Figure 1 shows the relative performance of global and focused channels when high frequency information is removed. This time, the performance of both the global and the focused channels drop significantly, suggesting that both rely on high frequency information. They do not rely on it to the same extent, however. The performance of the global channel drops by roughly half (45.4%), but the right hand channel is more strongly effected. Its performance drops by 80.99% (the left hand drops by 81.45%). Thus while all channels depend on high frequency information, the right- and left-hand channels rely on it more than the global channel does.

The left most comparison between global and right-hand channels in Figure 1 reveals a dependence upon motion that is the opposite of that seen for color and resolution. When motion information is removed the performance of all channels drops, but this time the right and left-hand channels drop by less than half (38.7% and 40.63% respectively), while the performance of the global channel drops by over 80%. This suggests that the global channel is more concerned with motion than the focused channels are.

There is an interesting analogy here to foveal and peripheral vision. The right and left hand channels are both focused

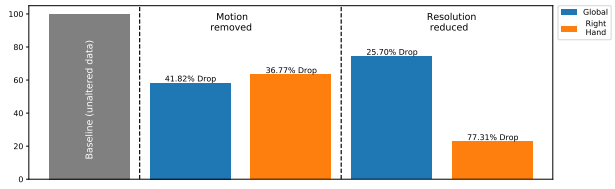

Fig. 2. Bar chart showing data ablation results on the depth modality of ChaLearn IsoGD dataset with motion removed and resolution reduced. The figure formatting otherwise matches that used in Figure 1.

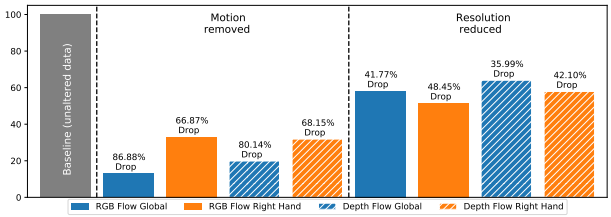

Fig. 3. Bar chart showing the results of data ablation on flow field channels of ChaLearn IsoGD dataset. Solid bars represent flow field channels computed from RGB images, while dashed bars represent flow fields computed from depth images. The first four bars show the result of removing motion information (i.e. using only a single flow field, not a window of 10). The next four show the result of removing high frequency information.

on a target, and they learn to rely on color and high frequency information, while making less use of motion, much like human foveal vision. The global channel looks at the whole scene and primarily exploits motion information, while making less use of color and high-frequency information, much like human peripheral vision. Obviously, human visual specialization is dictated by the anatomy of the eye [28]. The fovea contains densely-packed cones providing high-frequency color information to ganglion cells that largely feed the slow but color-sensitive α-channel in LGN. The peripheral retina contains mostly loosely-packed rods connected to ganglion cells that feed the faster-responding β-channel. In contrast, all channels in FOANet receive the same types of information. It is interesting to speculate, however, whether the eye evolved the way it did because focused vision demands high-frequency color information, while peripheral vision concentrates on motion information.

Figure 2 shows the results of data ablation on depth channels. When motion information is removed, the relative performance of the global depth channel drops by 41.82%, while the drop in performance for the right-hand depth channel is similar at 36.77% (The left hand drops by 38.21%). This suggests that all depth channels are sensitive to motion information, but none rely on it as their primary source of information. When high frequency information is removed, on the other hand, the relative performance of the global depth channel drops by only 25.70%, whereas the relative performance of the right hand depth channels drops by 77.31% (left: 80.66%). The focused depth channels rely on high frequency information, whereas the depth global channel can make do with low frequency information.

Figure 3 shows the results of data ablation on flow field channels. There are two types of flow fields: those computed on RGB images (shown in solid colors) and those computed from depth images (shown with cross-hatching). When only a single flow field is provided (instead of a window of 10), global flow field channel performance drops by 80.14% for depth flow and 86.88% for RGB flow. The right hand flow channels performance drops by 66.87% and 68.15%. This is somewhat counterintuitive, since even one frame of a flow field represents motion information. Apparently, these channels rely on multiple frames of flow field data, either to smooth noise or to calculate accelerations. When high frequency information is removed, on the other hand, global channel performance drops by 35.99% and 41.77%, while right-hand channel performance drops by 42.10% and 48.45%. Flow channels are not as sensitive to high frequency information as they are to multi-frame motion.

TABLE I
INFORMATION FAVORED BY DIFFERENT CHANNELS OF FOANET

	RGB	**Depth**	**Flow**
Global	Motion	Low Frequency	Motion
Focus	High Frequency & Color	High Frequency	Motion

Table I summarizes the results of Figures 1 through 3. Global RGB channels favor motion information whereas RGB focus channels (i.e. right hand and left hand channels) favor high frequency and color information. Global depth channels prefer low frequency information and focused depth channels prefer high frequency information. Both global and focus channels of flow fields capture motion information. We can see clear patterns of degradation that correspond to modality and attention target. This suggests that channels specialize and capture different information, leading us to prefer the SoE hypothesis over the Bagging hypothesis as an explanation of multi-channel behavior.

IV. Relating Channels to Gesture Properties

Section III-B shows that channels specialize in the sources of information they exploit, but do they specialize in terms of the types of gestures they recognize? To answer this question empirically, we look gesture properties and information fusion. The idea is that if FOANet's information fusion network learns how much attention to pay to each channel relative to each gesture class. If it learns to pay attention to one set of channels for gestures with a property and another set of channels for gestures without it, then the channels have specialized with regard that property. On the other hand, if there is no correlation between a gesture property and the relative importance of channels, then the channels have not specialized. A full explanation, however, requires a quick review of information fusion in FOANet and a description of gesture properties.

A. Sparse Network Fusion

The biggest difference between FOANet and previous multi-channel architectures is the number of channels. Most earlier

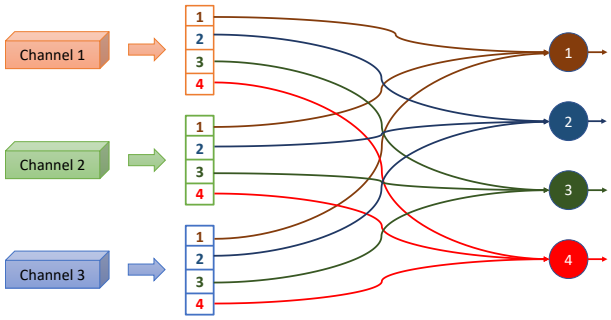

Fig. 4. Sparse network fusion architecture for a multi-channel network with 3 channels and 4 classes. Channels 1, 2 and 3 outputs softmax vectors of length 4 each. The softmax scores of a particular class from all channels are connected to the same output class.

systems had between 2 and 4 channels, whereas FOANet has 12. And the more channels there are, the more important information fusion across channels becomes. Most previous systems fused information in one of two ways: they either averaged the predictions of the channels, or they concatenated the channel feature vectors and trained a fully connected layer to generate label predictions. Unfortunately, the first approach is fairly weak whereas the second approach is prone to overfitting and doesn't scale with the number of channels. See [19] for a longer discussion.

Instead, FOANet introduced a technique called sparse network fusion. In this approach, every channel produces a label probability vector of length C, where C is the number of classes. These vectors are concatenated and a layer is trained to fuse information, but the layer is *not* fully connected. Each output unit corresponds to a gesture label, at it is only connected to the weights for that label coming from each channel. As a result, each output unit only has one weight per channel (12 in FOANet). A simplified version of this scheme with four output units and three channels is shown in Figure 4. For the purposes of this paper, the key observation is that every output unit is a gesture class, and it has 12 weights which dictate the relative importance of each channel to that gesture class.

B. Gesture Properties

We investigated three gesture properties looking for evidence of specialization. The first describes the type of motion in a gesture. Some gestures involve large arm motions, some involve smaller motions of the hands, and some are defined by static poses. The second distinguishes between one-handed and two-handed gestures. The third distinguishes between motions in the frontal plane of the body versus motion along the axis away from the torso (and roughly toward or away from the camera). Volunteer raters assign a binary (yes/no) label to all 249 ChaLearn IsoGD gestures classes for each property examined.

C. ChaLearn IsoGD Results

Table II shows the mean weights by channel for gestures with and without significant arm motions, with and without significant motion of the hand or fingers, and with and without a significant static characteristic pose. Within each of these tables are twelve columns, one for each channel, labeled by modality and attention target. In each column, the first row is the average weight for gestures with arm motions, and the second row is the average weight for those without. The third row is the signed difference between the means, while the fourth row is the p-value from Student's t-test, indicating whether the difference is statistically significant.

Without exception, flow fields channels are more important to gestures with arm motions than to those without. This confirms our intuition that flow field channels should specialize in motion information. At the same time, the lack of a significant difference for the other channels suggests that although RGB and depth channels are given access to 10-frame image windows their performance in not impacted (negatively or positively) by large motions.

Smaller hand and finger motions are slightly different. Flow fields computed from RGB images are important for detecting hand motions, but flow fields computed from depth images are not. We presume the depth sensor lacks the resolution to resolve such fine motions. Static poses, on the other hand, create an opposite scenario from arm motions. Flow field channels, whether from RGB or depth images, have significantly lower weights for gestures with static poses, meaning that flow field channels are less important to gestures with static poses than for those with movement. Once again, the RGB and depth channels are unaffected.

Table III is similar in layout to Table II and presents weights for the two other properties. Table IIIa shows the difference in fusion weights between one handed and two handed gestures. Since most ChaLearn participants are right-handed, one-handed gestures are usually performed with the right hand only. Examining the difference in weights in Table IIIa reveals that left-handed channels are discounted for one-handed gestures. This makes sense, since the left hand is not used in these cases. More interestingly, the RGB global and RGB right-hand channels are more important for one-handed gestures than two-handed gestures. The same is true for the right-hand depth channel. Our guess is that fine hand details are more important in one-handed gestures than two-handed gestures.

Table IIIb shows the weights for gestures with or without motion toward or away from the camera. It shows that the right-hand and left-hand depth channels and the global RGB flow field channel are important for motion along the optical axis. It surprises us that the focused RGB flow field channels are not significant for this task, and none of the depth flow field channels are significant. We note, however, that unlike in previous cases the p values are close to our significance cutoff threshold of 0.05, suggesting that this result is not as strong as the others.

Section III-B suggested that channels specialize in terms of the type of information they focus on. Collectively, the results in this section suggest that input specialization in turn leads to specialization in the types of gestures individual channels are

TABLE II

FUSION NETWORK WEIGHTS OF CHALEARN ISOGD DATASET COMPARED FOR GESTURES WITH AND WITHOUT SPECIFIC PROPERTIES. SIGNIFICANTLY LARGER WEIGHTS FOR A MODALITY ATTENTION COMBINATION SUGGEST THE COMBINATION IS RELIED UPON FOR RECOGNIZING GESTURES WITH THE ASSOCIATED ATTRIBUTE.

Weights comparison summary with and without Global Motion, i.e. arm motions												
Modality →	**RGB**			**Depth**			**RGB Flow**			**Depth Flow**		
Attention →	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**
Arm Motion	5.361	5.420	1.575	4.795	8.249	1.459	6.951	8.038	0.043	5.178	7.591	2.536
No Arm Motion	4.886	5.356	0.779	4.564	7.430	-0.630	5.360	6.412	-4.251	3.522	5.006	-1.434
Difference	0.475	0.064	0.796	0.232	0.819	2.089	1.591	1.626	4.294	1.656	2.525	3.970
P Value	0.385	0.911	0.561	0.709	0.117	0.116	0.002	0.006	0.004	0.008	0.000	0.004

(a)

Weights comparison summary with and without Local Motion, i.e. hand and finger motions												
Modality →	**RGB**			**Depth**			**RGB Flow**			**Depth Flow**		
Attention →	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**
Hand Motion	4.842	5.454	-0.792	4.401	7.791	0.720	6.999	9.214	3.364	4.351	6.815	2.064
No Hand Motion	5.136	5.361	-3.084	4.725	7.737	0.032	6.915	6.744	0.444	4.115	5.833	-0.440
Difference	-0.293	0.093	2.292	-0.324	0.054	0.688	1.308	2.470	2.920	0.236	0.981	2.505
P Value	0.649	0.890	0.267	0.657	0.930	0.660	0.032	0.000	0.039	0.751	0.129	0.127

(b)

Weights comparison summary with and without Static Poses, i.e. fix pose briefly held fixed												
Modality →	**RGB**			**Depth**			**RGB Flow**			**Depth Flow**		
Attention →	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**
Static Pose	5.051	5.171	-4.135	4.728	7.477	-0.454	5.026	5.896	0.258	3.415	4.794	-1.386
No Static Pose	5.090	5.587	-1.059	4.582	8.014	0.807	6.910	8.650	1.901	4.902	7.276	1.576
Difference	-0.039	-0.416	-3.076	0.146	-0.537	-1.261	-1.884	-2.754	-1.643	-1.487	-2.482	-2.962
P Value	0.943	0.455	0.072	0.810	0.293	0.331	0.000	0.000	0.018	0.015	0.000	0.029

(c)

TABLE III

FUSION NETWORK WEIGHTS COMPARED FOR GESTURES: A) USING ONLY ONE HAND VERSUS TWO HANDS, AND B) WITH SIGNIFICANT MOVEMENTS ALONG THE OPTICAL AXIS.

Weights comparison summary with only one hand versus two hands												
Modality →	**RGB**			**Depth**			**RGB Flow**			**Depth Flow**		
Attention →	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**
One Hand	5.757	6.152	-7.752	5.054	8.615	-3.933	6.285	7.642	-2.676	4.281	6.200	-3.957
Two Hand	4.114	4.307	4.636	4.096	6.541	5.924	5.554	6.798	6.338	4.009	5.841	5.786
Difference	1.643	1.844	-12.387	0.958	2.075	-9.857	0.732	0.844	-9.014	0.272	0.359	-9.743
P Value	0.002	0.001	0.000	0.118	0.000	0.000	0.156	0.121	0.000	0.663	0.510	0.000

(a)

Weights comparison summary with gestures with and without significant motion toward/away from the camera												
Modality →	**RGB**			**Depth**			**RGB Flow**			**Depth Flow**		
Attention →	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**	**Global**	**Right**	**Left**
Depth - Yes	6.217	6.545	-1.216	5.664	9.078	-2.550	7.314	7.911	2.967	3.645	6.778	3.550
Depth - No	4.908	5.216	-2.772	4.510	7.560	0.573	5.790	7.201	1.666	4.242	5.946	0.918
Difference	1.309	1.330	1.555	1.154	1.518	-3.123	1.524	0.710	1.301	-0.597	0.831	2.632
P Value	0.105	0.114	0.549	0.208	0.049	0.044	0.047	0.384	0.212	0.522	0.307	0.256

(b)

good at recognizing. This is further evidence that the Society of Experts (SOE) hypothesis is a better model of multi-channel processing that the Bagging hypothesis.

V. EXPERIMENTS ON NVIDIA DATASET

Sections III-B and III-C both report result on the ChaLearn IsoGD dataset. To see if the SoE hypothesis also holds on a second dataset, we ran the same two experiments on NVIDIA data [16]. The NVIDIA data set is smaller than IsoGD (just 25 gesture classes) and focused on a single domain: drivers. Since all the gestures are performed by right hand alone and the left hand isn't visible, the FOANet architecture is reduced to 8 channels (i.e. left hand channels are removed).

Once again, we perform a data ablation study by removing source of information and measuring drops in performance, as discussed in Section III-B. Figures 5 through 7 shows the results of the data ablation study on NVIDIA data. The clear patterns of degradation corresponding to modality and attention target that were visible in the IsoGD dataset are also visible here. Figure 5 shows that global RGB channels favor

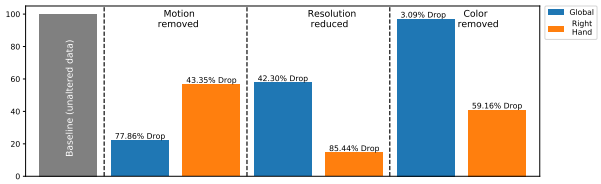

Fig. 5. Bar chart showing the result of data ablation on RGB modality of NVIDIA dataset.The figure formatting matches that used in Figure 1.

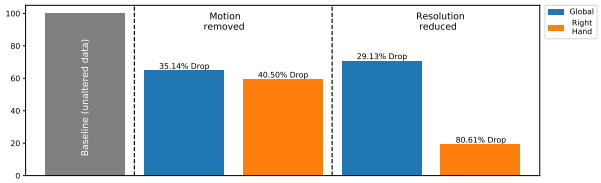

Fig. 6. Bar chart showing the result of data ablation on depth modality of NVIDIA dataset.The figure formatting matches that used in Figure 2.

motion information whereas right hand RGB channels favor high frequency and color information. Global depth channels prefer low frequency information and right hand depth channels prefer high frequency information 6. All flow channels capture motion information a shown in Figure 7. Similar to IsoGD dataset, channels specialize and capture different information on NVIDIA dataset, leading us to prefer the SoE hypothesis over the Bagging hypothesis as an explanation of multi-channel behavior.

We also repeat the experiment correlating fusion weights to gesture properties in order to see if channels are better at recognizing some types of gestures than others. We employ the methodology introduced in Section IV of comparing the fusion layer weights for the gestures with the property versus gestures without the property to see whether a channel is tuned to a gesture property. Table IV shows the mean weights by channel for gestures with and without significant arm motions (Table IVa), with and without significant motion of the hand or fingers (Table IVb), with and without a significant static characteristic pose (Table IVc), and with and without a movement towards/away from the camera (Table IVd). We can see that RGB flow (global and right hand) channels and depth flow global channels are better at recognizing gestures with significant arm motions. Right hand channels of depth, RGB flow and depth flow modalities are good at recognizing

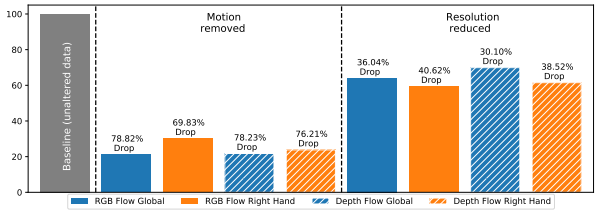

Fig. 7. Bar chart showing the result of data ablation on flow modalities of NVIDIA dataset. The figure formatting matches that used in Figure 3.

gestures with motion of the hand or fingers. Similarly, all right hand channels are good at recognizing static poses and depth modality channels are good at recognizing gestures that involve motion towards/away from the camera. These results are consistent with the results on the IsoGD data set, and suggest that channels specialize to gesture classes as predicted by the SoE hypothesis.

TABLE IV
FUSION NETWORK WEIGHTS OF NVIDIA DATASET COMPARED FOR GESTURES WITH AND WITHOUT SPECIFIC PROPERTIES. SIGNIFICANTLY LARGER WEIGHTS FOR A MODALITY ATTENTION COMBINATION SUGGEST THE COMBINATION IS RELIED UPON FOR RECOGNIZING GESTURES WITH THE ASSOCIATED ATTRIBUTE.

Weights comparison summary with and without Global Motion

Modality →	**RGB**		**Depth**		**RGB Flow**		**Depth Flow**	
Attention →	**Global**	**Right**	**Global**	**Right**	**Global**	**Right**	**Global**	**Right**
Arm Motion	1.143	2.719	2.866	4.500	6.661	4.759	6.357	2.909
No Arm Motion	0.680	2.640	2.248	3.895	4.164	2.664	3.166	4.562
Difference	0.463	0.079	0.618	0.605	2.497	2.095	3.191	-1.653
P Value	0.661	0.947	0.458	0.593	0.025	0.045	0.025	0.324

(a)

Weights comparison summary with and without Local Motion

Modality →	**RGB**		**Depth**		**RGB Flow**		**Depth Flow**	
Attention →	**Global**	**Right**	**Global**	**Right**	**Global**	**Right**	**Global**	**Right**
Hand Motion	0.097	2.183	2.138	5.877	5.041	6.253	3.223	6.042
No Hand Motion	1.502	3.111	2.730	3.547	4.892	4.829	2.932	2.179
Difference	-1.405	-0.929	-0.592	2.330	0.149	1.424	0.291	3.863
P Value	0.416	0.400	0.447	0.020	0.892	0.038	0.766	0.008

(b)

Weights comparison summary with and without Static Poses

Modality →	**RGB**		**Depth**		**RGB Flow**		**Depth Flow**	
Attention →	**Global**	**Right**	**Global**	**Right**	**Global**	**Right**	**Global**	**Right**
Static Pose	2.738	2.098	2.512	6.337	6.578	6.291	2.913	4.789
No Static Pose	1.397	0.505	2.429	3.526	7.146	3.632	3.111	1.010
Difference	1.341	1.592	0.083	2.811	-0.569	2.659	-0.198	3.778
P Value	0.329	0.038	0.933	0.025	0.569	0.018	0.871	0.046

(c)

Weights comparison of gestures with and without motion toward/away from the camera

Modality →	**RGB**		**Depth**		**RGB Flow**		**Depth Flow**	
Attention →	**Global**	**Right**	**Global**	**Right**	**Global**	**Right**	**Global**	**Right**
Depth - Yes	1.485	2.837	5.959	5.835	3.480	8.121	1.627	1.645
Depth - No	0.771	2.651	3.095	1.651	5.092	6.938	3.197	4.241
Difference	0.714	0.186	2.864	4.184	-1.613	1.184	-1.570	-2.596
P Value	0.694	0.928	0.049	0.005	0.422	0.420	0.379	0.369

(d)

VI. CONCLUSION

The literature suggests that multi-channel architectures outperform single channel ones, but does not explain why. After all, the same information can be packed into a single multi-band input, allowing a single-channel architecture to combine information across modalities without restrictions. This paper compares two explanations for the success of multi-channel systems. The Bagging hypothesis suggests that the benefit is from averaging the results of unbiased classifiers. In this model, every channel is general, and the goal of differentiating the channel inputs is to prevent redundancy. The Society of Experts (SoE) hypothesis suggests that channels specialize, with each channel becoming expert at a different aspects of the data. In this model, the benefit comes from selectively fusing information across experts. We explore these hypotheses in the context of FOANet, a 12-channel architecture with SOA performance on the ChaLearn IsoGD and NVIDIA datasets.

We find evidence that the SoE hypothesis is a better description of FOANet than the Bagging hypothesis. By altering the input streams, we show that channels specialize in the sources of information they pay attention to. For example, the global RGB channel favors motion information, while the left hand and right hand RGB channels prefer high-frequency color information. By analyzing fusion weights, we show that channels specialize to gesture properties. For example, one-handed gestures discount left-hand channels entirely while emphasizing RGB data and focused depth data over global depth data or flow field data. Collectively, our experiments suggest that channels specialize to specific sources of information, making them more or less relevant to specific types of gestures.

REFERENCES

[1] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[2] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in neural information processing systems*, pages 3468–3476, 2016.

[3] C. Feichtenhofer, A. Pinz, R. P. Wildes, and A. Zisserman. What have we learned from deep representations for action recognition? *arXiv preprint arXiv:1801.01415*, 2018.

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. 2016.

[5] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. 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.

[8] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, and S. Dähne. Patternnet and patternlrp–improving the interpretability of neural networks. *arXiv preprint arXiv:1705.05598*, 2017.

[9] N. Krishnaswamy, P. Narayana, I. Wang, K. Rim, R. Bangar, D. Patil, G. Mulay, R. Beveridge, J. Ruiz, B. Draper, et al. Communicating and acting: Understanding gesture in simulation semantics. In *IWCS 201712th International Conference on Computational SemanticsShort papers*, 2017.

[10] Z. Liu. Chalearn2017_isolated_gesture. https://github.com/ZhipengLiu6/Chalearn2017_isolated_gesture, 2017.

[11] Z. Liu, X. Chai, Z. Liu, and X. Chen. Continuous gesture recognition with hand-oriented spatiotemporal feature. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3056–3064, 2017.

[12] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[13] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, X. Cao, Z. Liu, X. Chai, Z. Liu, et al. Multimodal gesture recognition based on the resc3d network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3047–3055, 2017.

[14] M. L. Minsky. *The Society of Minds*. Simon Schuster, 1986.

[15] P. Molchanov, S. Gupta, K. Kim, and J. Kautz. Hand gesture recognition with 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015.

[16] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4207–4215, 2016.

[17] A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*, 20(14):5, 2015.

[18] P. Narayana. *Improving Gesture Recognition through Spatial Focus of Attention*. PhD thesis, Colorado State University, 8 2018.

[19] P. Narayana, R. Beveridge, and B. Draper. Gesture recognition: Focus on the hands. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] P. Narayana, R. Beveridge, and B. Draper. Continuous gesture recognition through selective temporal fusionn. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.

[21] P. Narayana, N. Krishnaswamy, I. Wang, R. Bangar, D. Patil, G. Mulay, K. Rim, R. Beveridge, J. Ruiz, J. Pustejovsky, and B. Draper. Cooperating with avatars through gesture, speech and action. In *Proceedings of the IEEE Intelligent Systems Conference (IntelliSys)*, 2018.

[22] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

[23] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. *arXiv preprint*, 2017.

[24] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[25] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[26] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

[27] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen. Sign language recognition using convolutional neural networks. In *Workshop at the European Conference on Computer Vision*, pages 572–578. Springer, 2014.

[28] S. E. Plamer. *Vision Science: Photons to Phenomenology*. MIT Press, 1999.

[29] J. Pustejovsky, N. Krishnaswamy, B. Draper, P. Narayana, and R. Bangar. Creating common ground through multimodal simulations. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*, 2017.

[30] R. Ramprasaath, D. Abhishek, V. Ramakrishna, C. Michael, P. Devi, and B. Dhruv. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CVPR 2016*, 2016.

[31] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[32] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[33] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[34] J. Wan, S. Escalera, A. Gholamreza, H. J. Escalante, X. Baró, I. Guyon, M. Madadi, A. Juri, G. Jelena, L. Chi, et al. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV*, volume 4, 2017.

[35] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016.

[36] H. Wang, P. Wang, Z. Song, and W. Li. Large-scale multimodal gesture recognition using heterogeneous networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[37] I. Wang, M. B. Fraj, P. Narayana, D. Patil, G. Mulay, R. Bangar, J. R. Beveridge, B. A. Draper, and J. Ruiz. Eggnog: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 414–421. IEEE, 2017.

[38] I. Wang, P. Narayana, D. Patil, G. Mulay, R. Bangar, B. Draper, R. Beveridge, and J. Ruiz. Exploring the use of gesture in collaborative tasks. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2990–2997. ACM, 2017.

[39] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[40] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3120–3128, 2017.