STP-Net: Spatio-Temporal Polarization Network for action recognition using polarimetric videos

R. Krishna Kanth
TCS Research
Bangalore, India
rokkamkrishna.kanth@tcs.com

Akshaya Ramaswamy
TCS Research
Chennai, India
akshaya.ramaswamy@tcs.com

A. Anil Kumar
TCS Research
Bangalore, India
achannaanil.kumar@tcs.com

Jayavardhana Gubbi
TCS Research
Bangalore, India
j.gubbi@tcs.com

Balamuralidhar P
TCS Research
Bangalore, India
balamurali.p@tcs.com

Abstract

Deep learning has brought tremendous progress in computer vision and natural language processing, and is used in multiple non-critical applications. A major bottleneck for its use in many other areas is the black box nature of these algorithms, resulting in a lack of explainability in their decisions. One of the key problems identified is the confounding effect, which causes confusion between the desired causes and other irrelevant factors affecting an outcome. This is more pronounced in the spatio-temporal case, such as the bias on the static background in the classification of a video. A way to handle this is by making use of sensors that capture additional scene properties, to mitigate spurious associations. In this work, we integrate the polarimetric videos with deep learning and evaluate it on the popular action recognition problem. We construct a dataset of polarimetric videos for fine-grained actions and study the effect of various parameters, extracted from the polarimetric video frames, as inputs to a deep network. Using these observations, we design a spatio-temporal polarization network (STP-Net) to effectively extract polarimetric features. This is evaluated on the recent HumanAct12 dataset for human activity recognition. Extensive evaluation clearly shows that the polarimetric modality is able to localize the correct action regions, leading to better generalizability.

1. Introduction

Deep learning has resulted in significant improvements in multiple computer vision tasks. In some cases, the accuracies have surpassed humans, resulting in industry adoption. The number of such applications are sparse. The major reason being the black box nature of these algorithms, that affects the trustworthiness and makes them vulnerable to adversarial attacks. In case of spatio-temporal data, these issues get amplified significantly. For instance, in [15], a study is performed to understand what the network has learnt for recognising actions. It is observed that around 40% of classes in UCF101 [23] and 35% of the classes in Kinetics [5] do not require any motion information for action recognition. Further, a significant number of actions can be detected using the background information in the videos. This is largely due to deep neural networks being unable to determine the confounding factors and this is being discussed widely today [9]. While neurosymbolic artificial intelligence is directed towards solving the learning problems of the machine, there is a way to address this by providing enriched data from the sensors.

It is well known that when an unpolarized light is incident on an surface, the reflected light will be partially polarized [17]. The characteristics of this reflected polarized light shall depend upon both the material as well as the shape of the reflecting object [24], which can succinctly be represented using the Stokes vector [7]. Three important parameters namely, the intensity, angle of polarization (AOP) and degree of polarization (DOP) can be derived from the Stokes vector. While the intensity is akin to information captured using a standard RGB camera, the other two are complementary information which contains rich geometrical cues. Hence, this imaging referred as *polarimetric imaging* [7] finds applications in several niche areas like

transparent object segmentation [18], imaging in harsh environments [4], dense shape reconstruction particularly on textureless surfaces [17, 8] *etc.*, which are otherwise highly challenging.

Polarimetric imaging requires images of the same scene to be captured at multiple polarizing angles and hence was limited to only static imaging scenarios. Recently, with the arrival of integrated camera capable of simultaneously providing images at multiple polarizing angles, this is increasingly finding applicability in many more scenarios. In this paper, we consider one such important application of action recognition which requires spatio-temporal polarimetric video processing.

Numerous classical and deep learning methods to perform action recognition exist in literature [27, 16, 6, 22, 25]. As mentioned earlier, these methods are data biased and often learn irrelevant features. If standard cameras are replaced with polarimetric cameras, the algorithm will have additional shape information to learn from. As with any new modality, polarimetric videos offer extra information in the form of angles of polarization and hence the input to the network will change substantially. This paper provides an in-depth study on the use of various deep learning architectures on spatio-temporal polarimetric videos. The variations are caused due to the nature of input such as the raw images captured at different polarizing angles and the derived information like the Stokes parameters or AOP and DOP. One additional challenge is the availability of datasets. Unlike standard RGB based action recognition, where exhaustive open datasets like UCF101 [23] and ActivityNet [10] are available, in this case availability of such open datasets is limited. Hence, we begin with creation of our own dataset by collecting polarimetric videos of actions using the integrated polarization camera, ensuring the background information as a constant. We then consider the HumanAct12 dataset [12] (which is created for the purpose of shape extraction) for our studies and comparisons. The contribution of this work include:

1. a detailed study on the input and multiple architectures for polarimetric video analysis
2. design of a spatio-temporal deep network for reusable polarimetric video feature extraction
3. evaluation of the proposed network on action recognition and further analysis on the use of polarimetric data.

2. Related work

Polarimetric imaging has been extensively applied for medical imaging and remote sensing applications [26, 28, 21, 2]. More recently, it is being explored specifically in applications where RGB does not give the desired properties. Polarimetric images capture photometric information and contain rich geometric cues, making it suitable in challenging environments containing reflective surfaces [8] or transparent objects [18]. They are also useful in unfavourable weather conditions such as scene analysis in the presence of rain or fog [4].

Deep learning methods have been frequently applied on polarimetric data to capture meaningful patterns and learn the mapping to downstream tasks. The usage of convolutional neural networks (CNN) with polar images was attempted by Kalra *et al.* [18] for transparent object segmentation. Multiple parameters are computed from the polar images and are used as separate inputs to parallel branches in a Mask-RCNN framework [13]. Segmentation and detection with polarized inputs was found to perform better than that using RGB intensities. The main drawback is the presence of multiple parallel branches, which increases the network complexity.

Polarization was also used by Blin *et al.* [4] to enhance object detection in adverse weather conditions. The road scene dataset consisting of both polarization and RGB data for the same scene was introduced. The polarimetric video data is recorded and one frame for every two seconds is retained. The data is collected in three different weather conditions– foggy, sunny and cloudy– and the object bounding boxes are annotated for four classes - car, person, bike and motorbike. RetinaNet [19], with ResNet50 [14] backbone and pretrained on MS-COCO dataset [20], is fine-tuned for five different combinations of polarization inputs. A comparison resulted in three out of the five polarization input combinations performing better than RGB. Here, the video frames retained in the dataset are at a very low frame rate, therefore it does not leverage the temporal information for segmentation. The common step with these approaches is the use of pre-trained models. Models trained on RGB modality may not be directly suitable on polarization data, since the information captured by both these modalities are very different and possibly complementary.

Polarimetric data is more complex and is characterised by geometric features and physical properties of objects in the scene. This makes it suitable for 3D reconstruction [31], surface normal estimation and shape recovery scenarios [3]. In [31], 3D human shape is estimated with a single polarization image input by first estimating the surface normals, and then utilizing this to reconstruct the 3D shape. In [3], the problem of shape from polarization has been attempted using deep learning. Here, the physical principles are incorporated as priors in a deep network architecture. This hybrid method achieves state-of-the-art performance on a challenging range of textures and lighting conditions. Polarization inherently suffers from physics-based ambiguities such as azimuth ambiguity due to the presence of diffuse or specular reflections. Use of data driven approaches in combination with other modalities or in a stereo setting gives advantages in many 3D applications such as in depth estimation [29] and 3D reconstruction [17].

Polarimetric images in combination with deep learning has seen a lot of attention recently but very limited work has used polarimetric videos. A recent work [30] introduced a human shape and pose estimation dataset that consists of videos of human actions captured using a four camera setup, with three color cameras and one polar camera. The RGB videos have been used for action synthesis [12] and 3D reconstruction applications [31], but to the best of our knowledge, the polarization videos have not been used for any application. A detailed study on the architecture suitable for polarimetric video analysis and the corresponding applicability is clearly missing.

3. Proposed Approach

To design a robust architecture and input configuration for the extraction of features from the polarimetric video, we start with the creation of a polarimetric video dataset for action recognition in a controlled environment. We further design a baseline network and perform an exhaustive study with different input combinations for the task of fine-grained action recognition. Based on the observations made, we propose a spatio-temporal architecture and evaluate it on a public dataset for human activity recognition.

3.1. Preliminaries

Let V be the video captured by a polarization camera. Considering that the polarizer is rotated to four angles $[0^o, 45^o, 90^o, 135^o]$, each frame in V consists of four channels $F = [F_0, F_{45}, F_{90}, F_{135}]^T$, where each channel corresponds to the intensity measured at that polarizer angle. V captures the surface characteristics and physical properties of the objects in the scene, and can be processed further to compute other polarization parameters namely Stokes vector S, Angle of Polarization AoP and Degree of Polarization DoP. For each frame F, S can be given by the following equations [7]:

$$S = [S_0, S_1, S_2]^T \tag{1}$$

$$S_0 = F_0 + F_{90} \tag{2}$$

$$S_1 = F_0 - F_{90} \tag{3}$$

$$S_2 = F_{45} - F_{135} \tag{4}$$

where, S is a vector consisting of three components as given by equations 2-4. Since we assume only linear polarization, S_3 is neglected in this paper.

Further, the AoP and DoP can be determined from S using the following equations:

$$AoP = \frac{1}{2} tan^{-1} \left[\frac{S_2}{S_1} \right] \tag{5}$$

$$DoP = \frac{\sqrt{{S_1}^2 + {S_2}^2}}{S_0} \tag{6}$$

The above equations describe parameters that encode the properties captured by polarized light. Another important parameter that can be extracted from each set of frames is the unpolarized intensity F_{un}, given by the following equation:

$$F_{un} = S_0 = F_0 + F_{90} \tag{7}$$

3.2. Polarimetric data acquisition

In order to apply spatio-temporal deep learning on polarization input, we design an experimental setup for the acquisition of polarization videos of Fine-Grained Actions (FGA). The video data is collected using a Matrix Vision grayscale polarization camera [1] in a controlled indoor environment. The camera provides four registered frames corresponding to four polarization angles $[0^o, 45^o, 90^o, 135^o]$, and is fixed on a table with a camera stand at a height of two feet. It is kept facing downwards, and all the actions are performed on the table. The focus of the camera is fixed at this height, we ensure proper illumination is present. The data is transferred from the camera to a local machine over GigE Ethernet. All the action videos are collected in an egocentric manner. Each video is captured at twenty frames per second, with spatial resolution of (1232×1028). The duration of each video ranges from six seconds to fifteen seconds depending on the complexity of the action. Videos are collected for seven fine-grained actions, with 100 videos for each action. The fine-grained actions include: 1. Moving something from left to right; 2. Moving something from right to left; 3. Moving something up; 4. Moving something away from camera; 5. Moving something towards camera; 6. Placing something; and 7. Removing something. The action classes are inspired from the Something-Something V2 (SSV2) [11] dataset of RGB videos consisting of such fine-grained actions. The FGA dataset is quite homogenous in terms of the way the actions are performed and the illumination variations. The purpose of the data collection under a constrained environment is to evaluate the classification performance with polarized input and compare with that of using unpolarized intensity input.

3.3. Spatio-temporal baseline

We implement a baseline network to evaluate on the captured polarimetric videos. Since polarimetric data captures additional surface properties and the information is complementary to the details captured by the RGB modality, we do not make use of any model pre-trained on RGB.

3.3.1 Architecture

The spatial network proposed in [3] for deep shape from polarization is utilized for this purpose. The network is an encoder-decoder architecture, and takes two inputs: the physics based priors, and the four-channel raw image

corresponding to four polarized angles. The two inputs are concatenated and processed further using a fully convolutional encoder, that encodes the high level polarization features. These are then passed through a decoder with skip connections, to estimate the surface normal. We make two modifications to this network to design our spatial baseline N_s. Firstly, we take only the four-channel image input, and do not include physics based priors. Secondly, we take only the encoder network and add dense layers followed by softmax activation, to train for classification tasks. This gives us the spatial baseline N_s. The spatio-temporal baseline N_{st} is formed by extending N_s to video inputs and converting the two-dimensional layers to three-dimensional equivalents. As stated previously, effective measures can be computed from the raw polarimetric images captured by the camera at four angles. We employ these to come up with unique input configurations to evaluate the baseline. In order to understand the effect of polarization information as input to a network, we experiment with two broad categories of inputs: one is polarized and the other unpolarized. For polarized input, we consider three types:
1) V_{raw}, which is the set of four-channel raw frames of the input video V with dimension $[N, w, h, 4]$. N indicates the number of frames, and each frame has a size of $(h \times w)$.
2) V_{stokes}, given by the three-channel Stokes vector S for a set of frames in V. This is given in equations 1-4 for each frame. V_{stokes} has a dimension of $[N, w, h, 3]$.
3) V_{pol}, defined as a three-channel vector $[AoP, DoP, F_{un}]^T$ as shown in equations 5-7, for a set of frames in V, and having a dimension of $[N, w, h, 3]$.

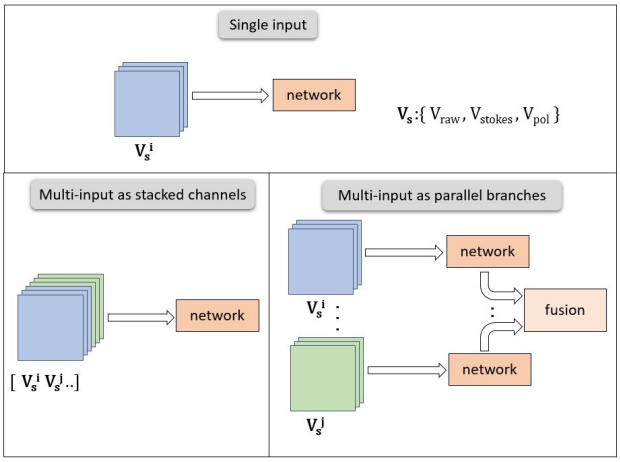

Figure 1. Three input configurations considered

These are compared with the unpolarized input V_{un} defined in equation 7 for one frame. V_{un} is a single channel output for each frame, and has size $[N, w, h, 1]$. With these inputs, we formulate three input configurations as visualized in Figure 1 and described below:

1) In single-input case, one of the above mentioned vectors is given as input to the network. This independently evaluates the contribution of each input for a classification task, and the type of input representation that is most suitable.
2) For multi-input as channels, two or more of the polarized inputs are stacked along the input channel dimension. This arrangement assumes that the different polarimetric measures are correlated, and maximizes the information captured by each, by ordering along channel dimension.
3) Multi-inputs as parallel network has each polarized input fed to an independent branch in the network. This arrangement has a distinct network for each input to learn independent features.

The purpose of experimenting with these configurations is for two reasons: i) each parameter captures a different structure in the scene, and different combinations of these are evaluated to understand the importance of each; and ii) the relation between the parameters, for example between V_{stokes} and V_{pol}, is highly complex and non-linear, therefore combining them in different ways can give insights into how successful the network is in taking advantage of the information in each of them.

3.3.2 Comprehensive study and observations

We evaluate the spatial baseline N_s and the spatio-temporal baseline N_{st}, with the designed input configurations on the acquired FGA dataset. The network is trained using a softmax cross entropy loss function and a stochastic gradient descent optimizer. An initial learning rate of 0.0001 is used, and the network is trained for 25 epochs. We compute three performance metrics: Top-1 accuracy, Macro accuracy and F1-score. The Top-K accuracy metric indicates the percentage of correctly detected action classes among the top K classes of the network detection. Macro-accuracy is the average class-wise accuracy, and gives equal importance to all the classes.

The accuracies achieved using the complete network, with the same number of layers as in [3], were above 95% for most of the inputs. We deduce that the network complexity is too high for the dataset considered, leading to overfitting. We therefore remove the last three layers and perform the evaluation. Table 1 shows the performance comparison for the different input configurations.

The first row indicates a spatial input with a single frame F_{raw} as input. All other rows are for a video input of 32 frames evenly sampled from the entire duration of a video. The row in blue indicates unpolarized input V_{un}. The following observations were made:
(a) In the single-input scenario, V_{raw} gives the best classification performance with an improvement of more than 10% as compared to the unpolarized input V_{un}. V_{pol} results in a sub-optimal performance. This can be attributed to the

Table 1. Quantitative evaluation on FGA dataset: Comparison of various input configurations to spatio-temporal baseline N_{st}

Input Configuration	Input type	No. of parameters	Top-1 Accuracy	Macro-accuracy	F1-score
Single frame input to N_s	F_{raw}	0.59 M	0.72	0.72	0.71
Single video input to N_{st}	V_{un}	1.75 M	0.75	0.75	0.75
	V_{raw}		**0.86**	**0.87**	**0.87**
	V_{stokes}		0.8	0.81	0.81
	V_{pol}		0.72	0.72	0.72
Multi-input video as stacked channels to N_{st}	V_{raw} + V_{pol}	1.75 M	0.73	0.73	0.72
	V_{raw} + V_{stokes}		*0.88*	*0.89*	*0.89*
	V_{raw} + V_{pol} + V_{stokes}		0.82	0.82	0.8
Multi-input video to parallel N_{st} branches	$V_{raw}\|V_{pol}$	3.5 M	0.96	0.96	0.96
	$V_{raw}\|V_{stokes}$		0.95	0.95	0.95
	$V_{raw}\|V_{stokes}\|V_{pol}$	5.25 M	0.99	0.99	0.99

structure of V_{pol} wherein AoP, DoP and V_{un} are stacked to form a three channel input. This arrangement does not exploit the advantages offered by each input type, and hinders the capture of useful features; (b) In the channel-wise stacking of multiple parameters, the combination of V_{raw} and V_{stokes} gives the best results with a Top-1 accuracy of 88% and an F1-score of 87%.

Again, the presence of V_{pol} could be the cause of lower performance in the other two input cases; and (c) In the parallel input configuration, the first thing to note is that the number of parameters increase considerably due to the design of separate branches for each input. All three input combinations give very high accuracies.

We apply the learning from this to develop a spatio-temporal architecture for polarimetric feature extraction.

3.4. Proposed Architecture: STP-Net

We develop a new spatio-temporal polarization network (STP-Net) to extract meaningful features from the polarimetric videos. Based on the evaluation of the baseline on the FGA dataset, we derive insights into the optimal input configuration and also identify drawbacks in the baseline architecture. Deep shape from polarization proposed by *Ba et al.* [3] is a network built for purpose of polarization images. This is used only as a baseline network N_{st} as it lacks spatial attention and the ability to capture spatio-temporal features due to 2D convolutions. We address these issues to design a new spatio-temporal polarization architecture as illustrated in figure 2. Beginning from the input, it is observed previously that the use of raw frames gives comparable performance with that using Stokes vector input and does not require additional computation. We also avoid use of parallel networks which significantly increase computational complexity and does not yield much dividends on accuracy.

Next, we focus on network-level modifications. The first layer in N_{st} makes use of $(1 \times 1 \times 1)$ convolution filters. This outputs linear combinations of the input channels, however does not make use of the spatial coherence. Moreover, placing this right at the beginning of the network leads to loss of information. The rationale behind having four filters with (1×1) kernel size in N_s could be to internally compute outputs of the form similar to Stokes vector. We replace this by directly connecting to 32 filters and kernel size $(3 \times 3 \times 3)$. The larger kernel size makes use of the spatio-temporal consistency to capture meaningful features. Instance normalization is used in between the convolution layers in N_{st}. While this may be suited for reconstruction and frame-level estimation tasks, they do not provide an advantage for classification tasks. Therefore, we replace them with batch normalization layers. We introduce regularization in the network, by adding dropout layers in between the convolution blocks and adding kernel regularizers to the convolution layers. This is to enable the network to learn robust models that can generalize well. We make minor modifications to the convolution blocks in the network. We retain the same set of filter sizes, but add more convolution layers, as shown in the Figure 2. Each strided convolution layer is replaced with a stride-1 convolution layer and a maxpool layer. A key component in the new architecture is the spatial attention module. The captured polarimetric frames are likely to contain noise and in order to focus on the correct regions in the frames, we apply a spatial attention module. This is especially essential in discrimination tasks where certain objects or actions in the scene contribute to the correct classification. After the first four convolution blocks are applied, the resulting feature vector X_{fea} is used to compute the attention weights. It is passed through another convolution layer to output attention weights W_{fea} for each spatial location. This is multiplied with all the channels in X_{fea} and the product is added to the original features according to the following equations:

$$W_{fea} = conv3D(X_{fea})|_{filters=1} \tag{8}$$

$$A_{fea} = X_{fea} + (X_{fea} * W_{fea}) \tag{9}$$

The attention weighted feature A_{fea} is passed through a global average pooling layer, and then through a set of three dense layers with filters 1024, 1024, K respectively, where K denotes the number of action classes. Softmax activation is finally applied to map the output to one of the

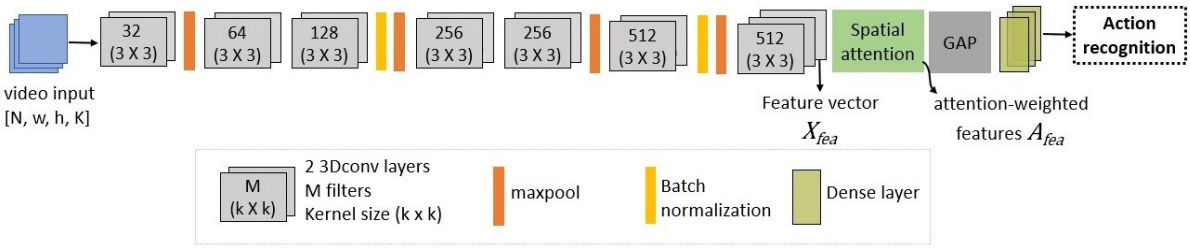

Figure 2. Block diagram of the proposed spatio-temporal polarization architecture STP-Net

K classes. The network is trained to minimize the softmax cross entropy loss. A stochastic gradient descent optimizer is used with a learning rate of 0.0001, clip value 0.5 and momentum of 0.9.

4. Experiments and Results

In this section, we cover the experimental setup and qualitative and quantitative evaluation on the HumanAct12 dataset. This is followed by the ablation study and visualization of class activation maps. The implementation is done using Tensorflow libraries, and the training is performed on an Nvidia Tesla V100 GPU.

4.1. Evaluation on HumanAct12 dataset

The HumanAct12 dataset [12] consists of a subset of videos cropped from the Polarization Human Shape and Pose Dataset (PHSPD). The data is collected synchronously with one polarization camera and three Kinect-V2 cameras. Data from the polarization camera is considered here for the experimentation. The dataset consists of 1191 video clips from twelve subjects, categorized into 12 coarse actions, and 34 fine-grained sub-actions. The coarse actions include daily activities like walk, run and also actions such as warm-up and boxing. Gray-scale videos for each scene are captured at four polarizing angles at about 13 frames per second (fps), with spatial resolution of 1224×1024. The video length ranges from 1 second to 36 seconds. Some videos in the dataset were wrongly labelled. These were corrected and videos where the action or actor mask was unclear were discarded. After these changes, the reduced dataset consists of 1157 videos.

4.1.1 Experimental setup

The dataset is split subject-wise into 739 training videos from eight subjects, 253 validation videos from two subjects and 165 testing videos from two subjects. 80% of the videos have duration less than or equal to eight seconds, so the network input length is set to eight seconds. For videos smaller than this, the frames are repeated cyclically. For longer videos, a center crop is considered. The frames are sampled at 4.5 fps resulting in a set of 35 frames spanning 8 seconds. We compute the same set of metrics for comparison: Top-1 accuracy, Macro accuracy and F1-score. The frame size is reduced to one-third of the original resolution. All the training experiments were run for 500 epochs.

We compare the proposed network STP-Net with N_{st} and with ResNet pre-trained model. With ResNet training, the input needs to be a single three-channel input. Therefore for polarized input, we consider only first three polarization angles, since the fourth frame F_{135} can be deduced from other three angles [4]. For unpolarized input, we convert the single channel intensity frames to three channels. We use the ImageNet model weights on each frame and combine it with global average pooling. We train only the dense layers, added after pooling, for classification.

4.1.2 Results

Table 2 presents the results on the HumanAct12 dataset for coarse action recognition. To improve the action recognition performance, we also experiment with masking the background in the input frames., using the ground truth annotations provided with the dataset. This can be replaced with a standard human detection or segmentation model. For both masked and unmasked scenarios, we evaluate the performance using polarized input and unpolarized input, and compare STP-Net with N_{st} and $ResNet$.

We make the following inferences:

1. An overall comparison of the three approaches shows that the proposed approach performs far better with approximately 30% average improvement over the baseline, and a larger 45% boost over ResNet pre-trained model. This clearly indicates two key points. One is that the polarization modality captures information of a scene very different from that contained in the RGB modality. Secondly, it contains rich cues about the object properties useful for classification.
2. For both the masked input as well as unmasked input, the proposed network achieves a 4% improvement with polarized input over that using unpolarized input.
3. With masked and polarized inputs to the proposed network, the macro-accuracy, which is the average class-wise accuracy, is 12% higher as compared to that using unpo-

Table 2. Quantitative evaluation on HumanAct12 dataset

Input masking	Network	Input	Top-1 Accuracy	Macro accuracy	F1-score
Unmasked	Proposed network STP-Net	V_{pol}	**0.70**	**0.69**	**0.69**
		V_{un}	0.66	0.68	0.66
	Spatio-temporal baseline N_{st}	V_{pol}	0.35	0.34	0.31
		V_{un}	0.38	0.39	0.35
	Pre-trained model $ResNet$	V_{pol}	0.21	0.08	0.07
		V_{un}	0.18	0.08	0.06
Masked	Proposed network STP-Net	V_{pol}	**0.75**	**0.77**	**0.75**
		V_{un}	0.71	0.65	0.71
	Spatio-temporal baseline N_{st}	V_{pol}	0.37	0.40	0.35
		V_{un}	0.46	0.41	0.44
	Pre-trained model $ResNet$	V_{pol}	0.21	0.15	0.08
		V_{un}	0.13	0.12	0.06

larized input. This shows better generalization capability across the action classes.

4.2. Ablation study

We perform an ablation study to validate the different network components and find the optimal parameter settings. This is done on the HumanAct12 dataset, which poses multiple challenges including class imbalance, action variability, varying video lengths and repeating nature of actions. Multiple experiments were carried out: a) to evaluate the network modifications; b) to fix the input video length and spatial resolution; and c) to handle class imbalance in the dataset. Following are the holistic inferences obtained based on the above mentioned experiments:

1. Network level: We made two modifications to N_{st} to come up with the proposed STP-Net. As ablation experiments, we train with the original settings: retaining strided convolution and removing maxpool, and retaining (1 x 1 x 1) convolution layer as the first layer; we also remove the spatial attention module and test the network. All three variations give a lower performance, clearly showing that the use of standard convolution with maxpool and the removal of the 1D convolution layer improve the action recognition. The absence of the spatial attention module results in a 5% drop in accuracy.

2. Regularization: The HumanAct12 dataset suffers from class imbalance, with two out of twelve actions containing less than 5% of the total videos. The use of regularization gives a 3% increase in the classification accuracy.

3. Input level: The optimum input dimension depends on factors such as the frame rate, frame resolution and video length. For an eight-second video length, we experiment with more number of frames and higher spatial resolution. The increase of spatial size to half of the original resolution results in larger tensor dimensions and more than 15% drop in accuracy. However, sampling more frames at 6.5 fps results in a 3% improvement but at the cost of more computations and slower training.

Table 3. Ablation study: Varying different parameters in network

Parameter	Variation	Top-1 Accuracy
Network	replace maxpool with strided conv	0.44
	adding (1 x 1 x 1) conv as first layer	0.65
	without spatial attention	0.65
Regularization	dropout	0.67
	dropout + class weighted loss	0.67
	dropout + class weighted loss + kernel regularizer	0.70
Input	Spatial resolution (612 x 512)	0.55
	52 frames spanning 8 seconds	0.73

4.3. Visualization of class activation maps

We generate the gradient weighted class activation maps for STP-Net and compare the maps for polarized inputs and unpolarized inputs. Figure 3 shows the activation maps for *eat* and *jump* actions from the HumanAct12 dataset, and *Moving object away from camera* and *Right to left* actions from the FGA dataset. These are the observations:

1. In the first two actions *eat* and *jump*, the polarized input network can be clearly seen to learn the correct action regions. In the first row, the entire region of movement from hand to mouth across frames get highlighte, and in the second rows the hand movements are localized clearly. In contrast, with unpolarized inputs, the same regions do not get focused. 2. On the FGA dataset, a distinct observation is that unpolarized input is unable to localize the correct movement regions for both actions. This is found to be the case in all the actions. With polarized input, the action areas get activated well, revealing valuable surface properties and goemetric cues captured by polarized video frames.

We visualize the class-wise recall plots for both datasets in Figure 4 and Figure 5. A key observation is that the true positive rate of each class remains above a certain level with polarization information and the network is able to generalize well for all the action classes. This is especially true for actions with fewer training samples. This also highlights that even with few samples, polarimetric data is able to generalize across different classes in both the datasets. In the absence of polarization information, while the network performs better for some actions but generalization across actions is poor.

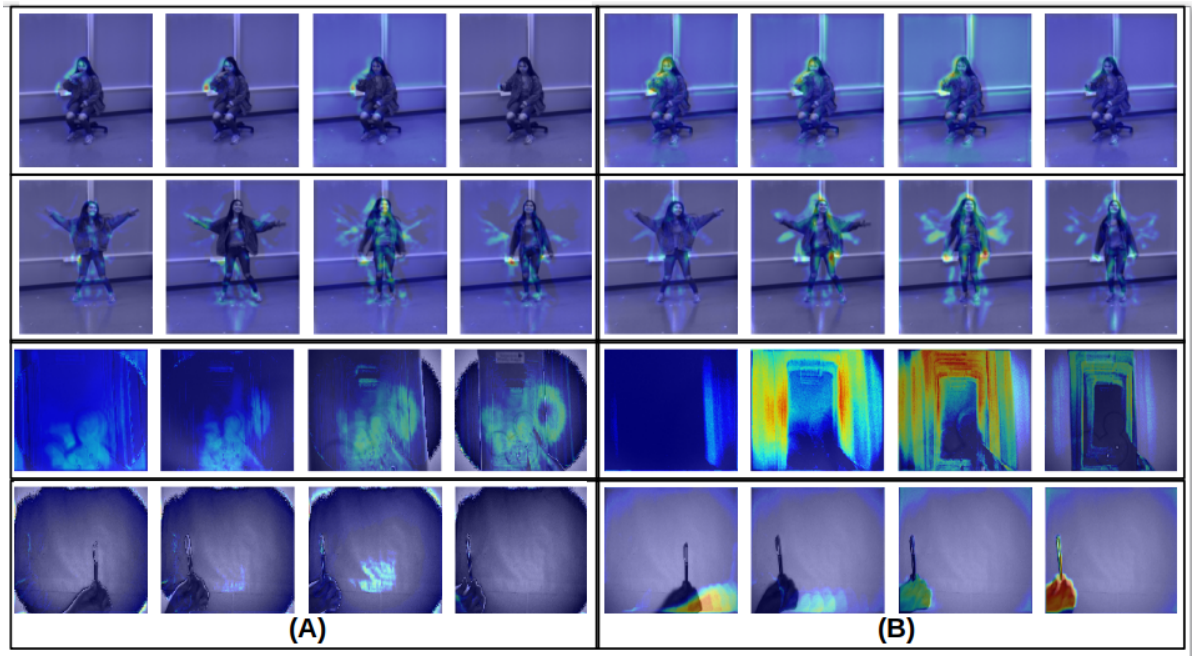

Figure 3. Class activation maps for actions from the HumanAct12 dataset: *eat* and *jump*; From the FGA dataset: *Moving away from camera* and *Right to left*, using (A) unpolarized input, and using (B) polarized input

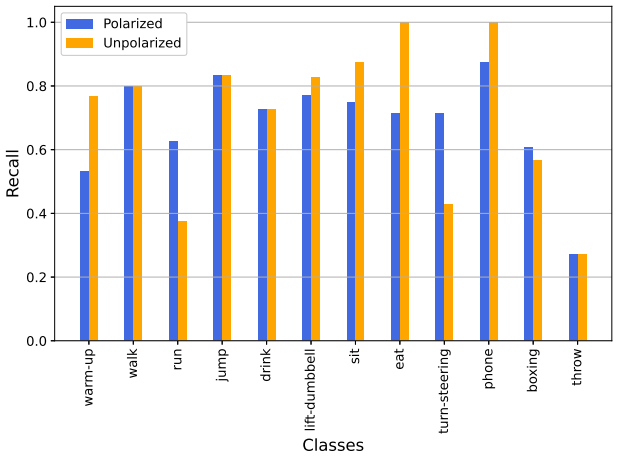

Figure 4. Class-wise recall on HumanAct12 dataset

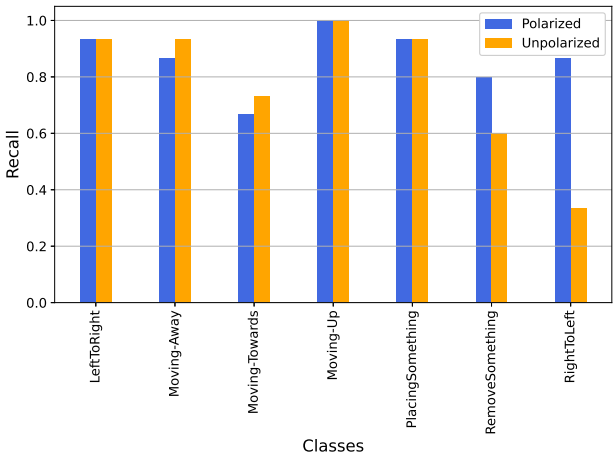

Figure 5. Class-wise recall on FGA dataset

4.4. Conclusion

This work presents an in-depth study and evaluation of spatio-temporal deep networks for action recognition using polarimetric videos. A polarimetric video dataset of fine-grained actions is constructed and an extensive analysis is conducted to understand the effect of different polarimetric input configurations. Based on the learning, STP-Net, a spatio-temporal polarization network, is introduced for the extraction of reusable features from polarimetric videos, and evaluated on the HumanAct12 dataset for human action recognition. The qualitative and quantitative results show that with additional geometric and surface cues captured by polarimetric inputs, the network is able to learn discriminative features and outperform intensity only inputs. The ablation study confirms that the polarimetric data helps in the cases of high class imbalance, which are vast in natural setting.

References

[1] Matrix vision - mvbluecougar-x.

[2] Sanaz Alali and I. Alex Vitkin. Polarized light imaging in biomedicine: emerging Mueller matrix methodologies for bulk tissue assessment. *Journal of Biomedical Optics*, 20(6):1 – 9, 2015.

[3] Yunhao Ba, Alex Ross Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. *arXiv preprint arXiv:1903.10210*, 2019.

[4] Rachel Blin, Samia Ainouz, Stéphane Canu, and Fabrice Meriaudeau. A new multimodal rgb and polarimetric image dataset for road scenes analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 216–217, 2020.

[5] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *CoRR*, abs/1808.01340, 2018.

[6] 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 (CVPR)*, July 2017.

[7] Edward Collett. Field guide to polarization. Spie Bellingham, WA, 2005.

[8] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1558–1567, 2017.

[9] Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: the 3rd wave. *CoRR*, abs/2012.05876, 2020.

[10] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[11] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *CoRR*, abs/1706.04261, 2017.

[12] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[15] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018.

[16] 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, 2013.

[17] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3d: High-quality depth sensing with polarization cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3370–3378, 2015.

[18] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8602–8611, 2020.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[21] Ji Qi and Daniel S. Elson. Mueller polarimetric imaging for surgical and diagnostic applications: a review. *Journal of Biophotonics*, 10(8):950–982, 2017.

[22] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014.

[23] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[24] Vage Taamazyan, Achuta Kadambi, and Ramesh Raskar. Shape from mixed polarization. *arXiv preprint arXiv:1605.02066*, 2016.

[25] 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 (CVPR)*, June 2018.

[26] J. Scott Tyo, Dennis L. Goldstein, David B. Chenault, and Joseph A. Shaw. Review of passive imaging polarimetry for remote sensing applications. *Appl. Opt.*, 45(22):5453–5469, Aug 2006.

[27] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[28] Lei Yan, Yanfei Li, V Chandrasekar, Hugh Mortimer, Jouni Peltoniemi, and Yi Lin. General review of optical polarization remote sensing. *International Journal of Remote Sensing*, 41(13):4853–4864, 2020.

[29] Dizhong Zhu and William AP Smith. Depth from a polarisation+ rgb stereo pair. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7586–7595, 2019.

[30] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chuan Guo, Chi Xu, Minglun Gong, and Li Cheng. Polarization human shape and pose dataset. *arXiv preprint arXiv:2004.14899*, 2020.

[31] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 3d human shape reconstruction from a polarization image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 351–368. Springer, 2020.