Multi-stream Architecture with Symmetric Extended Visual Rhythms for Deep Learning Human Action Recognition

Hemerson Tacon[1], André de Souza Brito[1], Hugo de Lima Chaves[1], Marcelo Bernardes Vieira[1], Saulo Moraes Villela[1], Helena de Almeida Maia[2], Darwin Ttito Concha[2] and Helio Pedrini[2]
[1]*Department of Computer Science, Federal University of Juiz de Fora (UFJF), Juiz de Fora, MG, Brazil*
[2]*Institute of Computing, University of Campinas (UNICAMP), Campinas, SP, Brazil*

Keywords: Deep Learning, Human Action Recognition, Data Augmentation, Visual Rhythm, Video Analysis.

Abstract: Despite the significant progress of Deep Learning models on the image classification task, it still needs enhancements for the Human Action Recognition task. In this work, we propose to extract horizontal and vertical Visual Rhythms as well as their data augmentations as video features. The data augmentation is driven by crops extracted from the symmetric extension of the time dimension, preserving the video frame rate, which is essential to keep motion patterns. The crops provide a 2D representation of the video volume matching the fixed input size of a 2D Convolutional Neural Network. In addition, multiple crops with stride guarantee coverage of the entire video. We verified that the combination of horizontal and vertical directions leads do better results than previous methods. A multi-stream strategy combining RGB and Optical Flow information is modified to include the additional spatiotemporal streams: one for the horizontal Symmetrically Extended Visual Rhythm (SEVR), and another for the vertical one. Results show that our method achieves accuracy rates close to the state of the art on the challenging UCF101 and HMDB51 datasets. Furthermore, we assessed the impact of data augmentations methods for Human Action Recognition and verified an increase of 10% for the UCF101 dataset.

1 INTRODUCTION

In the last years, revolutionary advances were accomplished in the Computer Vision field. This progress is due to the development of Deep Learning (DL) methods, driven by the technological enhancements of GPU. The major DL breakthrough was the Convolutional Neural Network (CNN) architecture and many architectures for image classification were developed. All of them benefited from the emergence of large image datasets, such as ImageNet (Deng et al., 2009). A natural consequence of this success was the exploitation of these achievements in the field of video classification. In this domain, the problem of Human Action Recognition (HAR) consists in recognizing the main action being represented by a person along a video. A solution to this problem is crucial to automate many tasks and it has outstanding applications: video retrieval, intelligent surveillance and autonomous driving (Kong and Fu, 2018).

In most datasets, the actions are simplistic, lasting for only a few seconds. However, due to scene dynamics, the challenge of HAR relies on detecting the action under different viewpoints, light conditions, pose orientations and in spite of significant differences in manner and speed that a video can present.

This work presents a method for HAR taking advantage of a DL architecture for classification. We propose the usage of VR (Ngo et al., 1999; Concha et al., 2018) obtained from two directions: horizontal and vertical. The VR is a 2D video representation which combines 1D RGB information varying over time. The specific feature used in this work to classify videos is called Symmetric Extended Visual Rhythm (SEVR). We propose to use horizontal and vertical SEVRs in conjunction, instead of only using the horizontal (Tacon et al., 2019). The two VRs are combined to increase the performance. The results corroborate for the usage of the combination of the two directions instead of choosing the best direction as proposed by (Concha et al., 2018).

Furthermore, we assess the employment of conventional data augmentation for image classification in the context of HAR with VR. We show that the

Tacon, H., Brito, A., Chaves, H., Vieira, M., Villela, S., Maia, H., Concha, D. and Pedrini, H.
Multi-stream Architecture with Symmetric Extended Visual Rhythms for Deep Learning Human Action Recognition.
DOI: 10.5220/0008958003510358
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 351-358
ISBN: 978-989-758-402-2; ISSN: 2184-4321

usage of classic data augmentation methods (zoom and horizontal and vertical flips) together with specific data augmentation for video (symmetric extension) increase the UCF101 dataset accuracy by 10%. Moreover, the VRs were combined with RGB images and Optical Flow (OF) in a multi-stream architecture to achieve competitive results.

The main contributions of this work are: the extension of the concepts of the Weighted Visual Rhythm (WVR) and Symmetrically Extended Visual Rhythm (SEVR) to the vertical direction of the spatial dimension of videos, the combination of vertical and horizontal SEVR to other streams to form a four-stream architecture, and an ablation study about the impact of data augmentation methods for HAR. Experiments were performed on two well-known challenging datasets, HMDB51 and UCF101, to evaluate our method.

2 RELATED WORK

The recent approaches to learning automatic features are mostly based on DL architectures. They can be viewed as single or multi-stream models. Since the success of the AlexNet (Krizhevsky et al., 2012) in the image classification problem, CNNs have become state of the art for this task. Since the 3D counterpart of an image is a video, the emergence of methods using 3D CNNs to address the video classification problem was a natural consequence. However, the transition from 2D to 3D CNNs implies an exponential increase of parameters, making the network more prone to overfitting. A successful architecture based on a 3D-like method is the Two-Stream Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017). The Two-Stream I3D was built upon on the inflation of CNNs by the expansion of kernels to a 3D space, making the network capable of learning spatiotemporal features. The main contribution of that work was the transfer learning from pre-training on both ImageNet (Deng et al., 2009) and a larger HAR dataset named Kinetics.

Simonyan and Zisserman (Simonyan and Zisserman, 2014) proposed to exploit and merge multiple features in a multi-stream method. They used RGB and Optical Flow (OF). The RGB representation is basically the usage of one colored frame to represent the whole video sample. The OF is a method for estimating and quantifying a pixel motion between subsequent frames (Zach et al., 2007). Essentially, the OF is a 2D displacement vector of the apparent velocities of brightness patterns in an image (Horn and Schunck, 1981). As a major contribution, they proposed to stack the OF from ten consecutive frames. Since the horizontal and vertical components of the OF vector fields were computed individually, the employed CNN architecture was modified to have an input layer with 20 channels ($224 \times 224 \times 20$). The temporal stream by itself outperformed the spatial one, which conferred importance to the specific motion information.

Since the two-stream method achieved state of the art results, recent works have proposed to explore 2D representations of videos to use image-based CNNs for the HAR problem. Wang et al. (Wang et al., 2015b) used the ten frame approach to train a temporal network. They observed that the usage of the ImageNet dataset (Deng et al., 2009) to pre-train the temporal stream could increase its performance. Their improved temporal stream is used in our work.

Derived from the OF definition, the Optical Flow guided Feature (OFF), introduced by (Sun et al., 2018), aimed to represent compactly the motion for video action recognition. This method consisted of applying the OF concepts to the difference of feature maps of consecutive frames. One of the main purposes of this work was to avoid the expensive runtime in the classical OF computation. However, it only achieved state of the art comparable results when combined with a temporal stream based on the OF.

Multi-stream methods have the problem of not allowing communication between the streams. This lack of interaction hinders the models from learning spatiotemporal features (Kong and Fu, 2018). Choutas et al. (Choutas et al., 2018) proposed a representation, named Pose moTion (PoTion), to encode motion of some video key points. In every frame, heat maps for human joints were acquired by human pose estimation. The heat maps were summed to obtain the final PoTion representation, with the same dimension of a frame. This representation is the input of a shallower CNN that predicts the final class. The PoTion representation alone was not able to achieve good results. However, combined with the Two-Stream I3D ConvNet (Carreira and Zisserman, 2017), it slightly improved the state of the art accuracy on the UCF101 dataset.

In the context of multi-stream video classification, data augmentation is often applied to the spatial stream inputs. At the training phase of the two-stream method, a randomly selected frame was rescaled to have the smallest dimension equals to 256. From this frame, a sub-image matching the employed network input dimension was randomly cropped and transformed in several ways. Wang et al. (Wang et al., 2015b) also proposed a multi-scale cropping method. It consisted of resizing the frame to 256×340 and

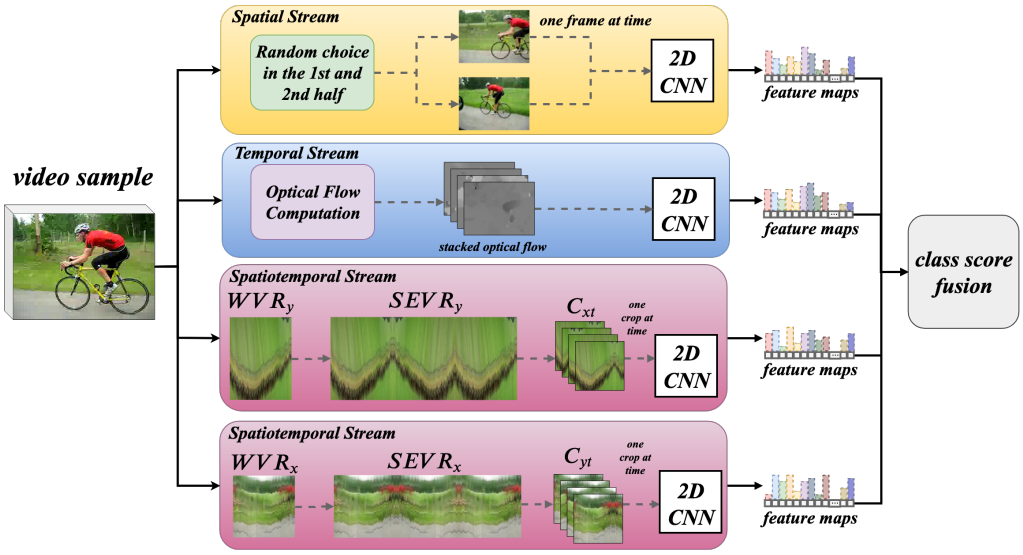

Figure 1: Final multi-stream architecture. The training of each stream is performed individually, and a weighted sum of the feature maps determines a descriptor utilized in the final classification.

randomly sampling from specific positions. After that, the crop was rescaled to match the network input dimension. These data augmentation methods showed to be helpful to avoid overfitting.

3 PROPOSED METHOD

The proposed method consists of combining two VRs, extracted from distinct directions, with other two streams that operate with RGB frames and OF. To this end, we use the WVR and its augmentation driven by the symmetric extension as new spatiotemporal streams. The streams are combined in a multi-stream architecture illustrated in Figure 1, where the WVR instances are depicted in purple.

3.1 Weighted Visual Rhythm

The horizontal Visual Rhythm WVR_y is defined as:

$$\mathrm{WVR}_y = \sum_{r=1}^{h} VR_{P_r} \cdot g(r-y,\sigma_y) \cdot \left[\sum_{r=1}^{h} g(r-y,\sigma_y)\right]^{-1}, \qquad (1)$$

where y is the reference row of the horizontal Visual Rhythm and $g(s,\sigma) = e^{-\frac{s^2}{\sigma^2}}$ is the weighting function that decays as the video embedded planes goes far from the reference row y.

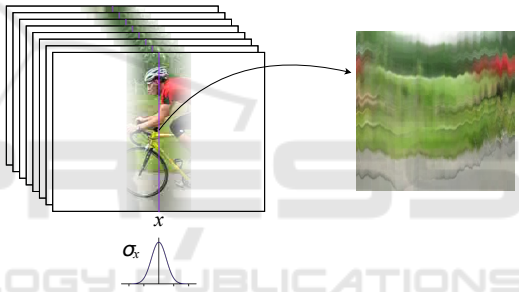

Figure 2: Vertical WVR of a sample video of the *biking* class from the UCF101 dataset: x is the middle column and σ_x is equal to 33.

The vertical Visual Rhythm WVR_x can be defined in the same way:

$$\mathrm{WVR}_x = \sum_{c=1}^{w} VR_{P_c} \cdot g(c-x,\sigma_x) \cdot \left[\sum_{c=1}^{w} g(c-x,\sigma_x)\right]^{-1}, \qquad (2)$$

where x is the reference column of the vertical VR and $g(s,\sigma) = e^{-\frac{s^2}{\sigma^2}}$ is the weighting function that in this case decays as the other VRs get farther from the reference column x. Thus, the WVRs used in the present work are defined by two parameters: the reference row y and standard deviation σ_y, for the horizontal version; and the reference column x and standard deviation σ_x, for the vertical one. In practice, some simplifications are adopted. An interval $y \pm d_y$ is defined from σ_y such that outer rows have zero weight. Furthermore, to make the parameter y invariant to video

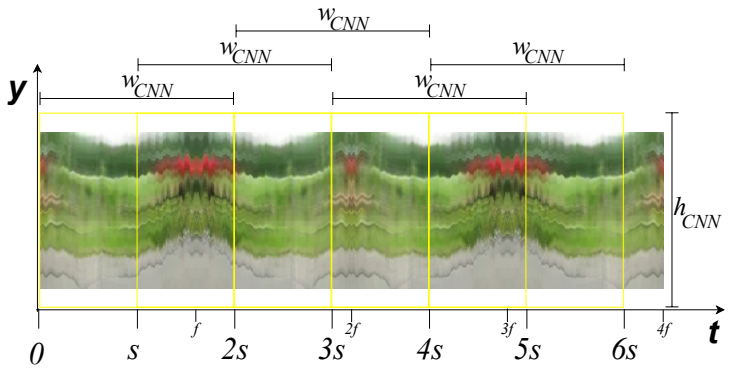

Figure 3: Extraction of five squared crops from the symmetric extensions of the vertical VR of a video of the *biking* class from UCF101 dataset. The frame width is $w = 320$ pixels, the frame height is $h = 240$ pixels, and the corresponding video length is $f = 240$ frames. The stride between crops is $s = 150$ pixels and the crop dimensions are $w_{CNN} = h_{CNN} = 299$. The central area in Y is selected thus the rhythm will be stretched in Y to cover the crop dimensions.

height h, we define a factor α_y such that $y = \alpha_y \cdot h$. Those same simplifications also apply for the vertical VRs. An interval $x \pm d_x$ is defined from σ_x such that outer columns have zero weight, and the α_x factor is defined as $x = \alpha_x \cdot w$ to make the x parameter invariant to the video width w. Figure 2 depicts a video of the *biking* class from the UCF101 dataset (240 frames with 320×240 pixels), forming a vertical WVR of 240×240 pixels.

3.2 Symmetric Extension

The symmetric extension for the WVR_y is defined as:

$$\text{SEVR}_y(i,k) = \begin{cases} \text{WVR}_y(i, f-m), \text{ for } \lfloor k/f \rfloor \text{ odd} \\ \text{WVR}_y(i, m+1), \text{ otherwise} \end{cases} \quad (3)$$

where $1 \leq i \leq w$, m is the remainder of the integer division of k by f and $k \in \mathbb{Z}$. Analogously, the symmetric extension of the WVR_x as:

$$\text{SEVR}_x(i,k) = \begin{cases} \text{WVR}_x(i, f-m), \text{ for } \lfloor k/f \rfloor \text{ odd} \\ \text{WVR}_x(i, m+1), \text{ otherwise} \end{cases} \quad (4)$$

where $1 \leq i \leq h$, m is the remainder of the integer division of k by f and $k \in \mathbb{Z}$. Thus, the SEVR is composed of several copies of the VR concatenated several times along the temporal dimension with the even occurrences being horizontally flipped. Figure 3 shows the vertical WVR of a video of the *biking* class of UCF101 extended three times. The premise is as follows: the majority of actions are constituted of symmetrical gestures, e.g. *typing*, *brushing teeth*, *drumming*, *pull ups*, *playing guitar*, etc. Thus, the action performed backward in time also represents the class and can be used to reinforce the CNN training. The symmetric extension circumvents the temporal limitation of videos and turns feasible the application of some data augmentation methods.

3.2.1 Symmetric Extension with Fixed Stride Crops

A crop C_{xt} from the SEVR_y with lower left coordinates x and t is defined as:

$$\text{C}_{xt}(a,b) = \text{SEVR}_y(x+a, t+b), \quad (5)$$

with $x \leq a < x + h_{CNN}$ and $t \leq b < t + w_{CNN}$. And a crop C_{yt} from the vertical SEVR with lower left coordinates y and t is defined as:

$$\text{C}_{yt}(a,b) = \text{SEVR}_x(y+a, t+b), \quad (6)$$

with $y \leq a < y + h_{CNN}$ and $t \leq b < t + w_{CNN}$. The VR is extended symmetrically until n_c crops are extracted using a stride s, i.e., the first crop is taken at $t = 0$ and all subsequent $n_c - 1$ crops are taken s frames ahead the previous one. The resulting set of crops for a fixed column y is $\{C_{yt} \mid t = js\}$, for $j \in \{0, 1, ..., n_c - 1\}$.

There is no guarantee that a complete cycle of the action is portrayed in a single crop since we do not have any information about its commencement and conclusion in a video sample. The probability of getting at least one complete cycle of action increases proportionally with the number of excerpted crops. Besides that, the stride is helpful for this purpose. This parameter is an attempt to adjust the temporal limits of the crop, aiming to comprehend a full action cycle.

3.3 Spatiotemporal Stream Classification Protocol

For the spatiotemporal streams, depicted in purple in Figure 1, a WVR is computed for each video, and its data augmentation is driven by symmetric extension. Multiple crops with fixed stride are extracted from the symmetric extension. At the inference time

and for video classification, all the augmented crops are individually applied to the CNN, and their last layer feature maps are extracted (just before softmax classification) and averaged. We observed that fusing the feature maps before the softmax normalization, as performed by (Diba et al., 2017) and (Zhu et al., 2018) achieves better results. A softmax classification layer is applied to this average feature map. The final class prediction is the averaged prediction of all crops.

We argue that this process might yield better class predictions based on the assumption that multiple crops taken at different time positions are representative of a distinct portion of the underlying action in the video. In the training stage, however, each crop is processed as a distinct sample and separately classified, i.e., the average is not taken into account.

3.4 Multi-stream Classification Protocol

Two instances of the spatiotemporal stream, one operating on horizontal WVRs and another operating on vertical WVRs, are used together with spatial and temporal streams to form a multi-stream architecture. In the overview of the proposed multi-stream architecture, depicted in Figure 1, the spatiotemporal streams are represented in purple, and the spatial and temporal streams are represented in orange and blue, respectively.

Each stream is trained individually, and all of them use a version of the InceptionV3 network pre-trained with ImageNet. The following details concern to both UCF101 and HMDB51 datasets. We adopt the improved spatial stream described by Concha et al. (Concha et al., 2018) as well as the training and testing protocols used in that work for both spatial and temporal streams.

A fusion of the feature maps of each stream produces the final classification for the multi-stream architecture. This fusion is a weighted sum of the feature maps. A grid search strategy is used to find out the best weights. The feature maps used for this are also extracted before the application of the softmax.

4 EXPERIMENTAL RESULTS

In this section, we evaluate the proposed methods. To make our results more competitive and to show the complementarity of the spatiotemporal streams, we explore the multi-stream classification using every possible combination between the employed streams: horizontal WVR, vertical WVR, static RGB frames, and Optical Flow. The proposed method was evaluated through experiments performed on two challenging video action datasets: UCF101 and HMDB51.

The evaluation protocol used for both datasets is the same. The average accuracy of the three training/test splits available for both datasets is reported as the final result.

4.1 Data Augmentation Ablation Study

In this section, we present an experiment to evaluate the real contribution of our data augmentation method apart from Keras data augmentation methods. To this end, we executed a baseline method using the resized WVR and without Keras data augmentation (horizontal flip, vertical flip, and zoom in the range of 0.8 to 1.2). Therefore, this is the scenario without any other data augmentation methods. Table 1 shows the results of these experiments. All results were obtained using the horizontal version of the WVR. The SEVR scenarios used 4 crops with a fixed stride of 299.

Table 1: Comparison of accuracy rates (%) for UCF101 and HMDB51 with (w/) and without (w/o) data augmentation methods.

Scenarios	UCF101 (%)		HMDB51 (%)	
	w/ Keras DA	w/o Keras DA	w/ Keras DA	w/o Keras DA
Baseline	65.19	57.64	34.46	28.93
SEVR	67.70	60.55	34.99	28.80

As expected, the usage of Keras data augmentation is significant, with an improvement of 7.35% and 5.86%, in UCF101 and HMDB51, respectively. The performance increase with our data augmentation methods (SEVR with crops) on UCF101 was also relevant. In this dataset, the mean increase was 2.71%. However, for HMDB51, our approach showed a slight decrease without using Keras data augmentation, 0.13%, and a small increase using both data augmentation methods, 0.53%. Thus, our data augmentation is at least complementary to the basic ones for the more complex dataset. Overall, the usage of both data augmentation methods for both datasets leads to the performance increase, highlighting the performance on UCF101, which increased over 10%.

4.2 Multi-stream Classification using Visual Rhythms

The goal in this section is to show that our spatiotemporal streams can complement a multi-stream architecture to get more competitive accuracy rates. The results of individual streams are shown in Table

2. The first five approaches, OF, RGB*, Horizontal-mean, Vertical-mean, and Adaptive Visual Rhythm (AVR), are results from the work of Concha et al. (Concha et al., 2018). It is worth remembering that the use of OF as a temporal stream is not a contribution of Concha et al. (Concha et al., 2018), but of Simonyan and Zisserman (Simonyan and Zisserman, 2014). However, the result of such work is shown since it comes from the use of the InceptionV3 network in the temporal stream, which is not performed in the original two-stream work (Simonyan and Zisserman, 2014). For the WVR_x, the best obtained accuracy for UCF101 was obtained with $\sigma_x = 33$, middle column ($\alpha_x = 0.5$), 4 crops and stride 299, and for HMDB51 was using $\sigma_x = 65$, middle column ($\alpha_x = 0.5$), 4 crops and stride 299. Similar to other multi-stream networks (Simonyan and Zisserman, 2014; Wang et al., 2015b), the OF performs better on both datasets. It is possible to notice that the horizontal SEVR presented superior performance if compared with the vertical one, independent of the dataset, due to the prevalence of horizontal motion in the videos. The same outcome appeared in the mean rhythm results. Excepting the $SEVR_y$ in the HMDB51, the SEVR was superior to the mean rhythm approach. Concerning the comparison between the $SEVR_y$ with the AVR, the results are divided. Even using only horizontal motion information, the $SEVR_y$ is better than AVR in the UCF101 scenario. However, the lack of vertical information may be one of the factors that led to a worse result in the HMDB51 dataset.

Table 2: Results of single-stream approaches.

Single streams	UCF101 (%)	HMDB51 (%)
Optical Flow (Concha et al., 2018)	**86.95**	**59.91**
RGB* images (Concha et al., 2018)	86.61	51.77
Horizontal-mean (Concha et al., 2018)	62.37	35.57
Vertical-mean (Concha et al., 2018)	53.87	30.12
AVR (Concha et al., 2018)	64.74	39.63
WVR_y	65.19	34.46
WVR_x	60.41	32.37
$SEVR_y$	68.01	35.29
$SEVR_x$	63.50	32.11

In order to our approach achieve more competitive results, we proposed a final multi-stream architecture merging the $SEVR_y$ and $SEVR_x$ best setups, with the RGB* and the OF streams. This final combination is not enough to assess the complementarity of the streams. To this end, we also conducted experiments incrementally fusing the streams. The strategy used for merging the streams was the same described by Concha et al. (Concha et al., 2018). Concerning the combination of all streams, the best combination found for UCF101 was 9.0, 7.0, 1.5 and 1.5, respectively for OF, RGB*, $SEVR_y$ and $SEVR_x$. And the best combination found for HMDB51 was 7.5, 3.5, 1.0 and 0.5, respectively for OF, RGB*, $SEVR_y$ and $SEVR_x$. We obtained 94.06% for UCF101 and 67.73% for HMDB51.

Table 3: Results of the combination of the streams.

Streams	UCF101 (%)	HMDB51 (%)
OF + RGB*	**93.21**	**66.43**
OF + $SEVR_y$	89.07	62.85
OF + $SEVR_x$	88.50	61.68
RGB* + $SEVR_y$	89.87	56.49
RGB* + $SEVR_x$	88.83	55.75
$SEVR_y$ + $SEVR_x$	75.90	41.85
OF + RGB* + $SEVR_y$	**93.70**	**67.15**
OF + RGB* + $SEVR_x$	93.53	66.91
OF + $SEVR_y$ + $SEVR_x$	89.72	63.20
RGB* + $SEVR_y$ + $SEVR_x$	90.76	58.43
All streams	**94.06**	**67.73**
Simple mean of OF + RGB*	**92.06**	**65.03**
Simple mean of OF + RGB* + $SEVR_y$	91.01	60.98
Simple mean of all streams	90.17	58.45

Table 3 shows the results of the incremental experiments. It contains the $\binom{4}{2}$ and $\binom{4}{3}$ combinations besides the combination of all streams. The combination of the best single-streams (OF and RGB*) generated the best two-stream combination, and the merging of the best two-stream formed the best three-stream combination with the third best single-stream result ($SEVR_y$). However, there is no guarantee that the best results are also the most complementary among each other. This is verified, for the UCF101, by the best two-streams containing $SEVR_y$ and $SEVR_x$ separately, which are formed with RGB* instead of OF. So, the other streams are crucial to complement the OF and to improve accuracy when combined.

Table 3 also shows the simple mean for the best combinations in each modality. In this case, we used the feature maps after the softmax normalization because of the magnitude difference between streams. The simple mean is the merge strategy adopted by some works in the literature (Simonyan and Zisserman, 2014; Wang et al., 2015b). Although the impact on two-stream is not harmful, the results tend to be more negatively affected by the streams increase, which is possibly a consequence of the accuracy gap

Table 4: Comparison of accuracy rates (%) for UCF101 and HMDB51 datasets.

Method	Pre-training Dataset	UCF101 (%)	HMDB51 (%)
Two-Stream I3D (Carreira and Zisserman, 2017)	ImageNet + Kinetics	98.0	80.7
I3D + PoTion (Choutas et al., 2018)	ImageNet + Kinetics	**98.2**	**80.9**
TDD+IDT (Wang et al., 2015a)	ImageNet	**91.5**	**65.9**
OFF (Sun et al., 2018)	—	**96.0**	**74.2**
Two-Stream (Simonyan and Zisserman, 2014)	ImageNet	88.0	59.4
Three-Stream TSN (Wang et al., 2016)	ImageNet	94.2	69.4
Three-Stream (Wang et al., 2017)	ImageNet	94.1	**70.4**
Multi-Stream + ResNet152 (Concha et al., 2018)	ImageNet	**94.3**	68.3
Multi-Stream + InceptionV3 (Concha et al., 2018)	ImageNet	93.7	69.9
Our method	ImageNet	94.1	67.7

among them. The superior results of the weighted sum in all cases demonstrate that the methods that treat each stream differently lead to better results in multi-stream architectures.

Furthermore, the two-stream combination of our both spatiotemporal stream ($SEVR_y$ + $SEVR_x$ line in Table 3) surpassed the AVR (Table 2). Both methods combined data about horizontal and vertical motion. While this comparison is not fair, because the two-stream had two types of data per sample and the AVR counts with only one type per sample, it gives a clue that combining vertical and horizontal motion is more advantageous than using the information from the most prominent movement direction.

Notice that the use of all four streams gives the best result. This means that $SEVR_y$ and $SEVR_x$ provide worthy complementary information to reduce classification confusion. A deeper study is needed to find out which types of motion benefit from our proposals.

Table 4 presents a comparison of our method combining all stream features through multi-stream late fusion and the other methods in the literature. Although the $SEVR_y$ and $SEVR_x$ streams do not achieve accuracy rates comparable to the state of the art individually (Table 4), the improved multi-stream method produced fairly competitive accuracy rates.

However, our method is overcome by some works. The works pre-trained with the Kinetics dataset have access to a volume of information that is crucial to achieving higher accuracy on UCF101 and HMDB51. However, a substantial amount of computational power required for pre-training with Kinetics makes its use impractical in most cases. Thus, we do not consider a direct comparison with these approaches. The merging with IDT features (Wang and Schmid, 2013) is another way to increase performance. DL methods often benefit from exploring this specific hand-crafted feature. In future works, we plan to verify the complementarity of this feature with our SEVR. Furthermore, the OFF approach stands out by being a method that does not use Kinetics pre-training and still achieves very close results to those that explore it.

Considering the VR approaches for the UCF101 dataset, our method outperforms the proposal of (Concha et al., 2018), using the InceptionV3. Our approach is not better than the ResNet152 result for the UCF101. The ResNet152 is deeper than the InceptionV3, and this may be the reason for the difference between outcomes. Considering that our approach used four streams, the change to a deeper model certainly would increase significantly the computational time required for training and testing. Further investigation is needed to evaluate the deep-accuracy tradeoff on multi-stream architectures.

5 CONCLUSIONS

In this present work, we propose an approach to deal with HAR in videos. It consists of the improvement of a two-stream method reasoned on the inclusion of complementary information through two spatiotemporal streams. The proposed spatiotemporal streams are 2D CNNs operating on the SEVR. We also evaluated the influence of the conventional data augmentation for image classification in the Visual Rhythms. It was verified that these data augmentation methods are very relevant for HAR using Visual Rhythms and that the data augmentation provided by the SEVR with fixed stride crops is also appropriate, providing complementary information to improve the classification. Concerning our multi-stream architecture, the results endorsed the complementarity between the spatial and temporal streams with our spatiotemporal streams. However, more experiments need to be carried out to evaluate the statistical signif-

icance of our approach, including a per class analysis. Our approach did not surpass some state-of-the-art methods, mainly due to restricted information of the used datasets. However, our results showed that our data augmentation might improve HAR accuracy. To achieve more competitive results, in future works, we intend to explore the complementarity of our multi-stream architecture with other features, such as IDT (Wang et al., 2013) and I3D (Carreira and Zisserman, 2017). In addition, the SEVR principles could also be employed to 3D CNNs for video classification problems.

ACKNOWLEDGEMENTS

Authors thank CAPES, FAPEMIG (grant CEX-APQ-01744-15), FAPESP (grants #2017/09160-1 and #2017/12646-3), CNPq (grant #305169/2015-7) for the financial support, and NVIDIA Corporation for the donation of two Titan Xp (GPU Grant Program).

REFERENCES

Carreira, J. and Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733.

Choutas, V., Weinzaepfel, P., Revaud, J., and Schmid, C. (2018). Potion: Pose motion representation for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Concha, D. T., Maia, H. D. A., Pedrini, H., Tacon, H., Brito, A. D. S., Chaves, H. D. L., and Vieira, M. B. (2018). Multi-stream convolutional neural networks for action recognition in video sequences based on adaptive visual rhythms. In *IEEE International Conference on Machine Learning and Applications*, pages 473 – 480.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Diba, A., Sharma, V., and Van Gool, L. (2017). Deep temporal linear encoding networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2329–2338.

Horn, B. K. and Schunck, B. G. (1981). Determining Optical Flow. *Artificial intelligence*, 17(1-3):185–203.

Kong, Y. and Fu, Y. (2018). Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Ngo, C.-W., Pong, T.-C., and Chin, R. T. (1999). Detection of Gradual Transitions through Temporal Slice Analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 36–41.

Simonyan, K. and Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems*, pages 568–576.

Sun, S., Kuang, Z., Sheng, L., Ouyang, W., and Zhang, W. (2018). Optical flow guided feature: a fast and robust motion representation for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1390–1399.

Tacon, H., Brito, A. S., Chaves, H. L., Vieira, M. B., Villela, S. M., de Almeida Maia, H., Concha, D. T., and Pedrini, H. (2019). Human action recognition using convolutional neural networks with symmetric time extension of visual rhythms. In *International Conference on Computational Science and Its Applications*, pages 351–366. Springer.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79.

Wang, H. and Schmid, C. (2013). Action Recognition with Improved Trajectories. In *IEEE International Conference on Computer Vision*, pages 3551–3558.

Wang, H., Yang, Y., Yang, E., and Deng, C. (2017). Exploring Hybrid Spatio-Temporal Convolutional Networks for Human Action Recognition. *Multimedia Tools and Applications*, 76(13):15065–15081.

Wang, L., Qiao, Y., and Tang, X. (2015a). Action recognition with trajectory-pooled deep-convolutional descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4305–4314.

Wang, L., Xiong, Y., Wang, Z., and Qiao, Y. (2015b). Towards Good Practices for very Deep Two-Stream Convnets. *arXiv preprint arXiv:1507.02159*.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016). Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision*, pages 20–36. Springer.

Zach, C., Pock, T., and Bischof, H. (2007). A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer.

Zhu, J., Zhu, Z., and Zou, W. (2018). End-to-end video-level representation learning for action recognition. In *24th International Conference on Pattern Recognition*, pages 645–650.