Science & Technology Development Journal 2023, 26(1):2681-2691

Open Access Full Text Article

Engineering and Technology: Research Article

Toward Deep Transfer Learning for Realistic Activity Recognition in Videos

Vo Hoai Viet*

ABSTRACT

Today, videos have become popular on the internet and specified in social network and media platforms such as Youbue, Ticktok, and Vimeo. Video understanding has attracted much attention in the research community in recent years. Automatically recognizing human activity in wild videos is a trending research topic with a wide range of applications in advertising, smarthome, and surveillance camera systems. Deep convolutional neural networks have become a new de facto visual recognition problem. It achieved much success in image recognition problems that are leveraged from the ImageNet dataset. Many researchers have applied CNNs to the video domain, but the results in realistic video still have many challenges, and the recognition rate is not as expected. Because the realistic activity in video is extremely small, we cannot train a large deep convolutional network to achieve good performance. In this work, we answer the question ``***how could we transfer the visual features from the ImageNet dataset into video for activity recognition tasks***?''. We propose an approach based on the pretrained models from ImageNet for activity recognition in video that is based on transforming video into an image map for temporal information and a frame-based description for spatial information. We test our proposed approach on three datasets, UCF11, UCF50 and UCF101, and it achieves an accuracy of over 95% for the backbone of ResNet and over 88% for the backbone of MobileNet. The experimental results show that our proposed method is robust and efficient in wild video datasets.

Key words: Activity Recognition, ConvetNets, Transfer Learning, UCF Dataset

INTRODUCTION

Today, video is ubiquitous thanks to the development and popularity of video recording devices and media platforms such as YouTube, TikTok, and Vimeo. It opens new opportunities for researching and developing video understanding algorithms. Specific to the human activity domain, traditional human activities such as walking, mixing, playing guitar, etc., are visualized from video frames in Figure 1. Human activity understanding in video comprises localizing, recognizing, and predicting human behaviors. The work to identify the label of human activity in a video is called activity recognition. Several previous studies concentrate on recognition problems in very specific domains, such as human activity recognition [1–5].

These approaches for activity recognition can be categorized into two categories. The first category is based on hand-crafted features [1–6] and machine learning algorithms such as KNN, ANN and SVM for classification. The second category applies deep convolutional networks (ConvetNets) to learn activity representation from raw video data such as RGB images or optical flows from motion detection algorithms and trains the system in an end-to-end approach [7,8].

Deep convolutional networks are very successful in image classification problems [9]. The deep learning-based CNN architecture outperformed the shadow networks and hand-crafted features by a huge margin in error rate. However, the ConvetNets applied to video classification do not have a significant improvement in accuracy compared to traditional methods in wild video datasets.

The second reason is that the dataset of activity videos is much smaller than the ImageNet dataset. For instance, the UCF101 dataset [3] only contains 13,320 videos for 101 activities. However, these ConvetNets must have a large dataset with a large number of labeled samples for training and tuning the weight of the network to achieve good performance. According to the book in deep learning [10], we should have minimum samples of 5,000 for each class. The mean we should have 505,000 samples in the case of the UCF101 dataset. A huge amount of labeled samples is a major problem in costing for a real application or startup project. Therefore, using pretrained models such as ImageNet will have a large impact on visual recognition tasks.

To overcome the above issues, we propose a novel approach based on transfer learning approaches from

University Of Science, VNU-HCMC

Correspondence

Vo Hoai Viet, University Of Science, VNU-HCMC

Email: vhviet@fit.hcmus.edu.vn

History

- Received: 2022-11-13
- Accepted: 2023-04-08
- Published: 2023-04-30

DOI :

https://doi.org/10.32508/stdj.v26i1.4017

Cite this article : Viet V H. **Toward Deep Transfer Learning for Realistic Activity Recognition in Videos**. *Sci. Tech. Dev. J.; 2023,* 26(1):2681-2691.

Figure 1: Some sample frames from daily activities from the UCF dataset

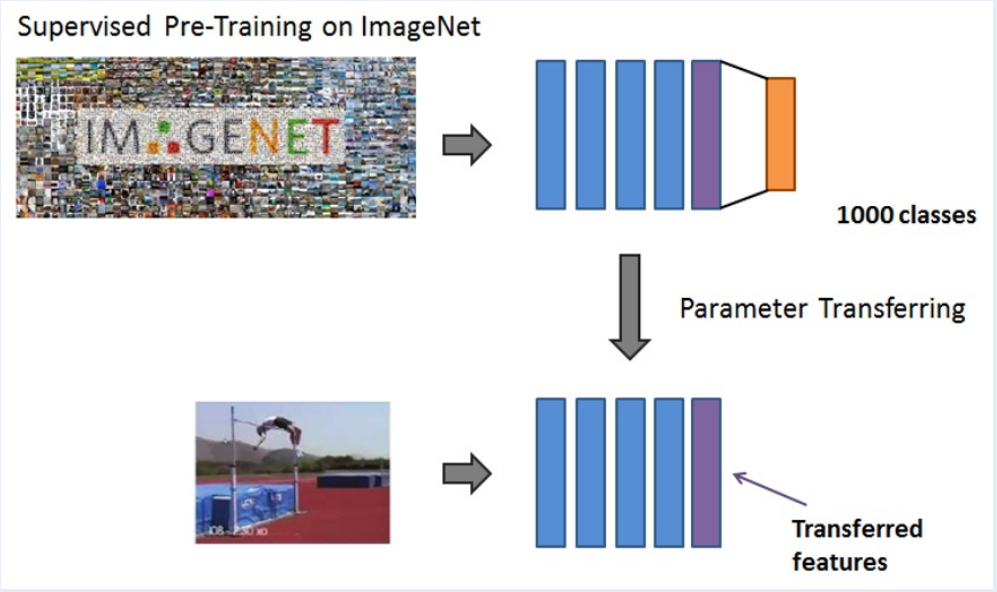

Figure 2: Overview of transfer learning from ImageNet visual features to frame-based representation

the ImageNet dataset [11] and key ideas from visual features transferred in deep convolutional neural networks [12]. The basic concept of transferring visual features in our approaches is represented in Figure 2.

Our contributions are summarized as follows: i) a transfer learning approach from image classification to activity recognition in video; ii) application of motion detection and description for temporal information in video and training a motion network for activity representation; iii) five deep network architectures are reviewed and applied to transfer visual features to the video domain; iv) experiment and evaluation of many different datasets and network architectures to demonstrate the generalization of our methodology.

LITERATURE REVIEW

Transfer learning

Transfer learning is not a new strategy, and there are many researchers applying this strategy when data are outdated or labeled data that do not have the same distribution over time [12,13]. In short, transfer learning focuses on using the knowledge that is learned from one or more previous source tasks and applying it to a new target task. With deep learning, the image classification task in the ImageNet Challenge has been very successful in closing the error rate, which may be lower than that of humans [9]. A summary of the evolution of the error rate can be seen in Figure 3. Many network architectures, such as AlexNet [14], VGG [15], GoogleNet [16], and ResNet [9], have proven robust and efficient in image classification in over 1000 classes in ImageNet. There is much knowledge or visual features learned from these architectures [12,17]. Specifically, transfer learning is also shown in the VGG architecture when training a deeper network based on

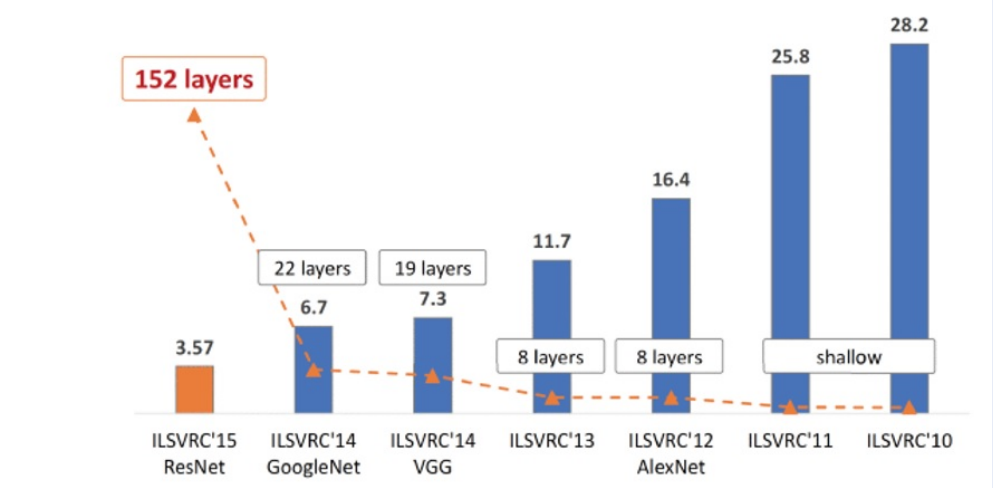

Figure 3: The evolution of the error rate on ImageNet Challenge Recognition from 2010 to 2015 [9] decreased from 28.2% to 3.57%

shallower network architectures.

Activity Recognition in video

One of the earliest studies on human activity recognition was introduced by Yamato et al. in [18], using simple shapes that were trained into a set of HMMs, and each action label was trained by one HMM. More sophisticated approaches based on motions are David [6] by transforming a video into image templates for each action. Template matching algorithms are used to identify the label of each video. After approximately 20 years, many researchers published promising results [1–5], and many applications were also considered for video understanding. Researchers have confirmed the important findings that the content of an image can be described by the spatial relationship between pixels. The video is an extended version of the image by fusion of frames in the temporal relationship. To perform well in video representation tasks, a system must extract and represent both spatial and temporal information in video. Since the success of CNNs in the image domain, many methods have been proposed to extend CNNs from images to videos for activity recognition. Two-stream networks [8] were proposed by training frame-based and motion-based networks in the separation of networks and fusion to recognize activity. AN is an extended version of CNN that applies 3D convolutional kernels such as I3D [7], SlowFast [19], and Nonlocal [20] to capture temporal information from video. These approaches require computing power and a large amount of labeled data for video training. Extending from [8], temporal segment networks [21] proposed CNN based on video segmentation that focused on reducing computational cost to scale for large-scale datasets. This research has motivated our research focus on novel methodology that could be applied in real world applications.

METHODOLOGY

In the previous section, the literature review is presented, and deep learning approaches become mainstream to improve the performance of the recognition system. The researchers also confirm that we need to have two kinds of data for activity recognition systems: spatial and temporal features. However, the deep convolutional neural network is very difficult to apply in practical applications due to the need for a huge number of video labeled datasets as well as large computing costs for large network architectures. To overcome these problems, we perform transfer learning from labeled images in Image to video datasets. The image dataset helps to learn information visual features from large semantic images, and these patterns can be applied to key frames in video that do not need training or only tuning the model for best fit to images from the video dataset. Therefore, it can be easy for practical applications, where video has limitations of labeled datasets. In this research, we propose a novel framework for activity recognition in video that uses a backbone for image description or motion template description based on these models learned from ImageNet. Our framework is illustrated in Figure 4.

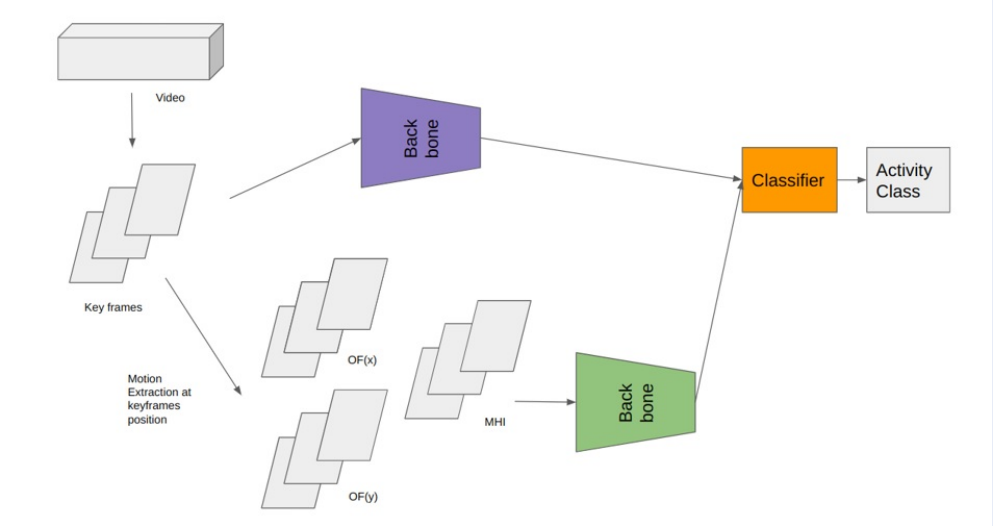

Figure 4: The flowchart for our proposed framework with keyframes and motion extraction for a deep learning backbone

Video Preprocessing

Keyframe selection. We also recognize an activity easily based on the context information from frame features. For instance, when we see the water in frame, it is easy to predict activities such as swimming and diving that cannot play violin or skiing. These characteristics are very important for wild video description. Instead of using all frames in video, these key frame selections are based on uniform sampling from the video timeline. It ensures that we capture the full content of the video in time. In this research, we adopted 5 key frames for each video, the same as the experiment in [2]. The visual features are extracted from these keyframes containing the main information for shape and context.

Image-Based Motion Transformation. The temporal relationship between frames when humans perform an activity is a key factor in distinguishing between activities of the same shape or nature, such as walking and running. We will fail when distinguishing between them that only use shape or edge features. In this research, we apply motion features from transforming video/clip into motion images in a time range of key frames selected from the previous step. The most important information that describes the content of video effectively is the temporal relationship between frames. In this research, we use the motion for x-direction OF(x) and y-direction OF(y) for frame-based motion. The optical flows are extracted by using Farnebäck [22]. We also use a motion history image (MHI) as proposed in [6] to capture how the motion is moving in the clip or how the activity motion is performing. We use these motion image templates to feed into the deep learning backbone to capture motion descriptors for activity video. The information extracted from these motion images will capture movement and the way that activity is performing at the moment time.

Network Architectures

The architecture of the network is the most important factor in the deep learning of network design. In the past, several network structures were proposed for image classification, such as AlexNet [14], VGGNet [15], Google Lenet [16], and Restnet [9]. However, their impact on activity recognition has not been fully considered in the video domain. In this section, we will briefly describe the main original architectures of AlexNet Net, Google Letnet, RestNet and MobileNet that were tested in our proposed approach. These deep neural networks are used as a backbone for learning features and transferring features in our system.

AlexNet was proposed by Alex Krizhevsky in the ImageNet competition in 2012. The general architecture is quite similar to LeNet-5 [23], although this model is considerably larger (see Figure 5). The success of this model (which took first place in the 2012 ImageNet competition) convinced much of the computer vision community to take a serious look at deep learning for computer vision tasks. The primary result in this network. To transfer visual features from image classification to video description, we adapted to use the F6

or F7 layer of this network. The details of this network can be found in [14].

VGGNet. The network opens a novel for a small kernel size in convolution and builds a deeper network based on the shallower architectures. The architecture used 3x3 for convolution operation, 1x1 for convolutional stride, and 2x2 for pooling windows. The authors systematically investigated the performance of depths from 11 to 19 layers in network architectures in image classification by training a deeper neural network from a shallower network with the same architectures. Finally, the network with 16 and 19 layers has the best performance and is often called VGG-16 and VGG-19, the architecture configuration in Figure 6. The details of this network can be found in [15].

Google LeNet. The primary idea of this network is the Hebbian principle and multiscale processing in the receptive field, called ***Interception*** with 22 layers [16], and the overall architecture is shown in Figure 7. An essential component in the network is the ***Inception module***, as shown in Figure 8. It consists of multiple convolutional filters of different sizes alongside each other. However, when this network goes deeper, changes in vanishing problems can occur. To avoid this problem, the authors proposed two auxiliary classifiers. An ***auxiliary module***, see in Figure 9, is the same as the subnetwork with full layers (convolution, pooling, and fully connected) that can learn the visual features in shadow layers and increase visual abstract features in deeper layers. Another problem of deep learning is the computational cost for computing, and there is also concern with using a 1x1 convolution operation for dimensionality reduction. The details of this network can be found in [16].

ResNet. VGG and GoogleNet notify a message in deep learning that is "go deeper, better performance". However, a deeper network will be very difficult to train. The researchers have systematically investigated deep networks with depths of 18, 34, 50, 101, and 152 layers. The authors proposed an extremely deep neural network called ResNet that is based on a novel architecture with residual learning (Figure 10) and batch normalization. The residual learning is based on the desired mapping as H(x): F(x) := H(x) - x, and the original mapping is F(x) + x. The batch normalization is a layer that will standardize and normalize the input from the previous layer into output. It will help speed up the training time and reduce internal covariate shift [24]. The details of this network can be found in [9].

MobileNet [25]. The big problem in deep learning models such AlexNet, VGG, GoogleNet or RestNet is large in computation and storage. Table 1 shows the sizel and the number of parameters of different models. The largest model is VGG with approximately 217 MB and 139 M parameters. This shows that the computational cost and memory when using the network inference stage is huge. There is a limitation of deep learning when applied in real-world applications with small memory or computing. In particular, mobile devices are currently dominant or have low hardware power in startup or manufacturing. Many studies have been proposed to resolve this problem. In this research, we briefly describe the MobileNet architecture and apply it to our architecture for human activity recognition in video. The architecture is based on depthwise separable convolutions that are factorized from a standard convolution and a 1x1 convolution, as shown in Figure 11. This factorization has reduced the computational cost for training and the size of the model. The size of the model is approximately 17 MB, and the number of parameters is 4.2 M parameters, a large reduction in both model size and parameters in deep convolutional neural networks. This model will be more practical when applied to real applications with limited computing power.

Classification

After extracting features, we have represented each video as a set of descriptors from key frames. The next problem is the identification of activity labels for each video; this step is called recognition. In this phase, we can use KNN or SVM algorithms to classify the descriptors into semantic labels. In this research, we used SVM to train a classifier for activity recognition.

EXPERIMENTS & DISCUSSION

In this section, we verify our transfer learning method from ImageNet into video datasets in activity recognition. The entire workflow of activity recognition in video is as follows: i) extract visual transfer features, ii) extract motion map and visual features, and iii) build a pool of SVM classifiers to classify each keyframe descriptor. The classification results for all keyframes and image maps in a video are aggregated using average pooling to identify the label of activity in video.

Datasets. In this work, we evaluate our approaches on the mainstream evolution of UCF datasets [1–3]. These datasets were published by the Center for Research in Computer Vision, University of Central Florida, from 2009 to 2012 with 3 versions of increasing the number of classes and complexity in activity. We visualize several frames from three versions in Figure 12. This will test our approaches in scalable and robust

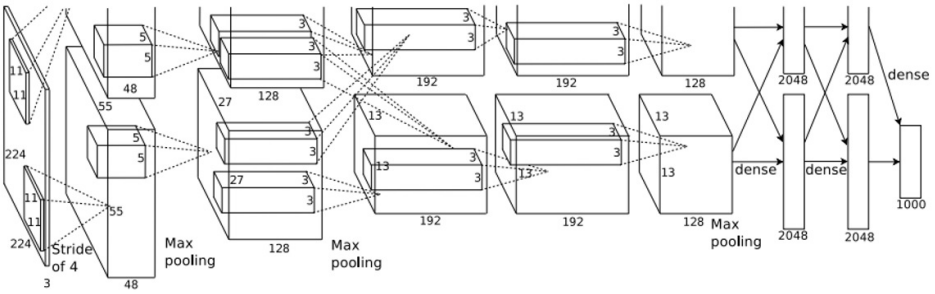

Figure 5: The network architecture of AlexNet [14] based on LeNet-5 [23]

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 6: The difference in network architecture of VGG [15], the number of layers increased from 11 to 19

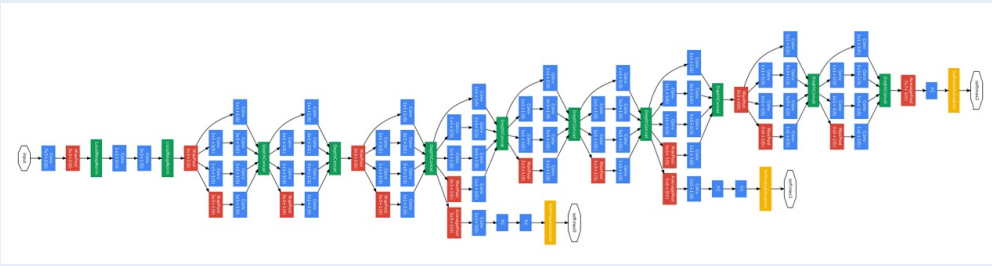

Figure 7: The overall architecture of Google Net [16]

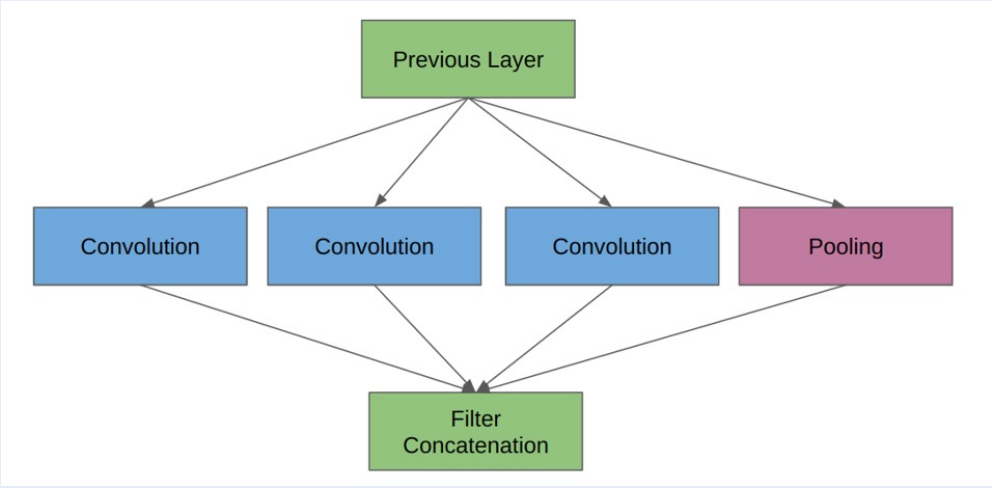

Figure 8: The architecture of the interception module

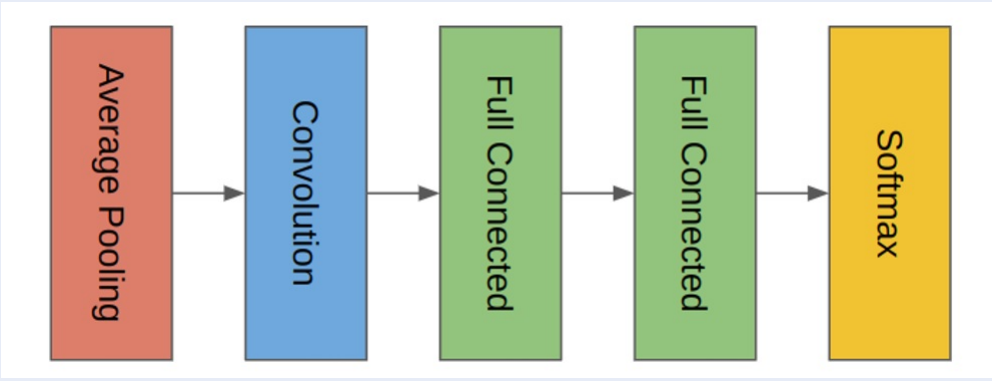

Figure 9: The architecture of the auxiliary classifier module

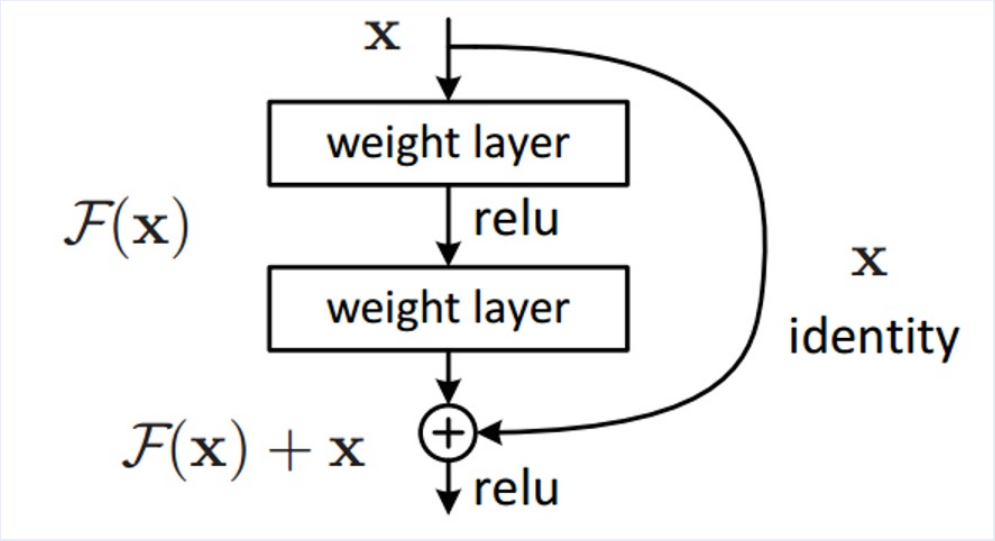

Figure 10: The residual block in RestNet[9] for residual learning

Table 1: Comparison of size and parameters of different network architectures

Model	Size (MB)	Parameters
AlexNet	217	62,970,000
VGG	553	138,360,000
GoogleLenet	163	42,710,000
RestNet-50	98	25,560,000
MobileNet	17	4,200,000

capacity in recognition rate when the datasets change in complexity and size of classes. The UCF YouTube (UCF 11)[1] contains 11 activity categories. UCF101[3] is an extended version of UCF 50[2]. It contains 13,320 videos spread over 101 categories of human activities. This challenging dataset contains a wide range of human activities that can be divided into 5 groups: i) Human-Object Interaction, ii) Body-Motion Only, iii) Human-Human Interaction, iv) Playing Musical Instruments v) Sports. All three datasets are collected from Youbute with wild human activity. Moreover, the videos in each dataset are divided into 25 folders, and each folder has more than 4 videos. Video from the same group can share the same common characteristics, such as the same person, similar background or viewpoint. The settings of the dataset for training and testing are the same as those used in original studies for UCF11[1], UCF[2], and UCF101[3].

Results: We show the experiment's results on three datasets and five backbones in Tables 2 and 3. In Table 2, we show the recognition rate from backbones transferred from different models trained from ImageNet without tuning. The results are over 93% for all datasets, and ResNet has proven its power in transferring the knowledge from ImageNet into the video domain by using key frame approaches. With a small dataset such as UCF1, all networks are robust and efficient in accuracy, and MobileNet also has over 93% accuracy when the size of the model is much smaller than the others. They confirm that we can transfer ImageNet models into the video domain in novel approaches that still have high performance. However, when the number of classes increases 5 or 10 times, such as 50 and 101 classes, the performance of the system decreases by approximately 10% for the best architecture ResNet. The accuracy is still over 86% for 101 classes, which shows the robustness of ResNet in the wild dataset. The smallest network is MobileNet, which still has a promising recognition rate of 76.14%. From these results, we have turned the parameters of these networks based on images and motion images from target datasets. The performance of the system

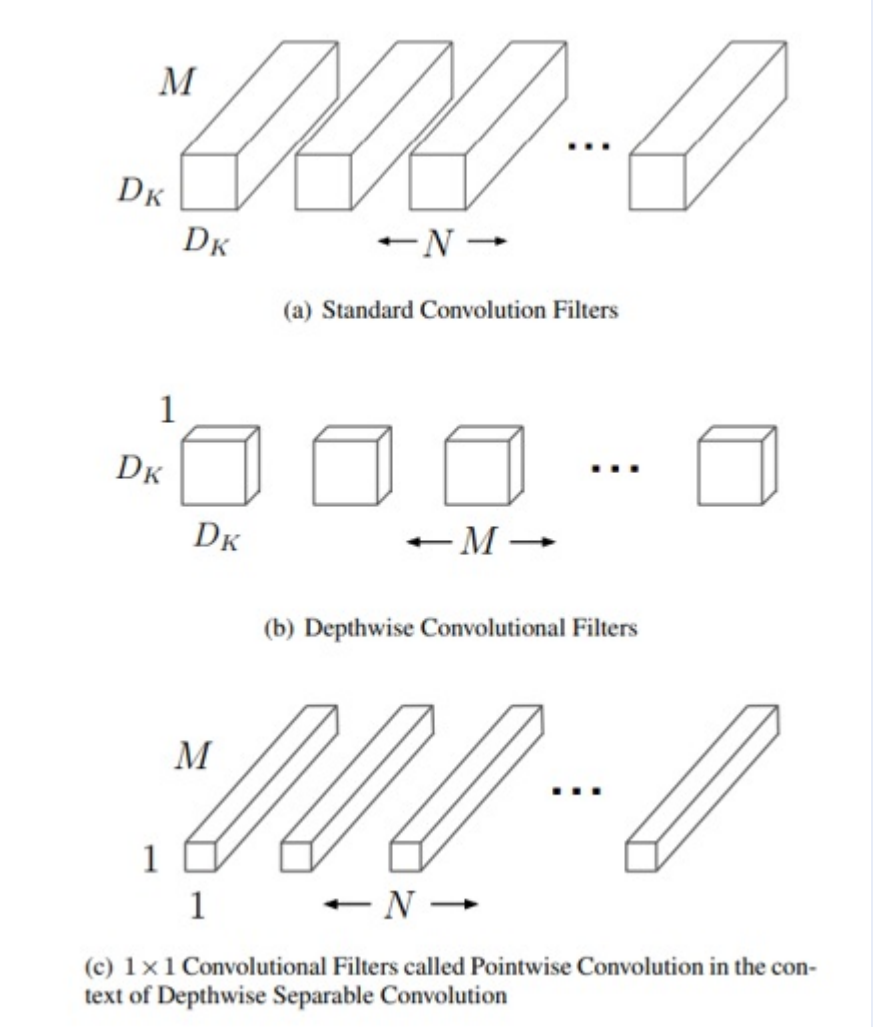

Figure 11: Illustration of convolution factorization in pointwise convolution [25] with standard, depthwise, and 1x1 convolution filters

has a significant improvement of over 10-12% for all architectures. ResNet is the most robust and efficient in performance, and MobileNet is the lowest. The results confirm that we can transfer visual features from image to video w.r.t. turning and video transformation that convert video data into image format to still capture visual information. We show a practical approach in which we can build a video understanding system based on knowledge that is trained from large-scale datasets such as ImageNet.

Comparison: We compare the recognition rate of these different network architectures on several datasets in Tables 2 and 3. These deep convoluted networks that have nearly the same depth in architecture, such as AlexNet [14], VGG [15] and GoogleNet [16], have similar performance. The deeper ResNet [9] still powers the video domain the same as the image domain when we try to convert video data into image format for transfer learning. These results show that MobileNet has a performance lower than that of ResNet [9] by approximately 7%. Meanwhile, the size of MobileNet [25] is less than 5 times that of ResNet. This means we can think that applying MobileNet in applications will be more practical than big networks. Furthermore, Tables 1 and 2 show that we will increase from 2% for a small dataset and 11% for a large dataset when tuning the networks to fit with the dataset. This proves that the hypothesis of applying a deep learning model from image to video is feasible and highly effective when we use digital image and video processing techniques in representing video content based on spatial-temporal mapping images.

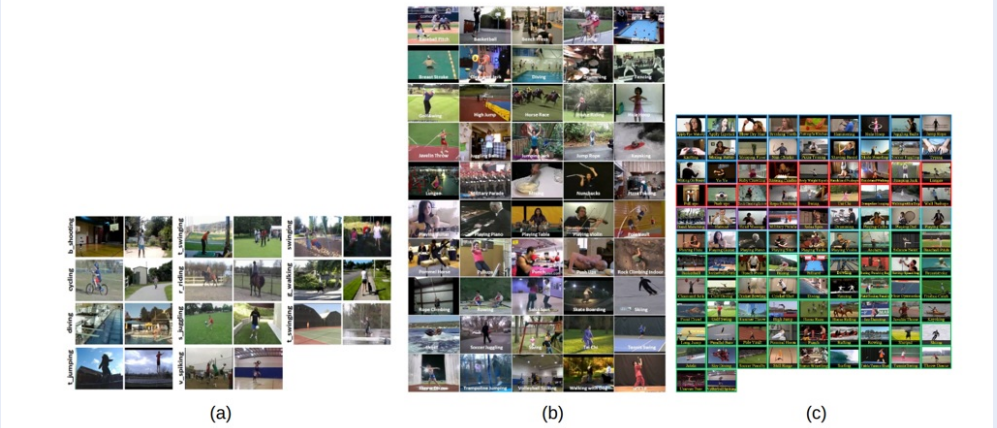

Figure 12: Sample activity frames: (a) UCF11, (b) UCF50, and (c) UCF101

Table 2: The accuracy of three datasets with transferring visual features

Networks	Accuracy (%)		
	UCF11	UCF50	UCF101
AlexNet	95.08	83.25	79.25
VGG-16	95.38	85.69	81.26
GoogleLenet	95.82	87.30	82.4
RestNet-50	97.24	92.71	86.34
MobileNet	93.59	79.40	76.14

Table 3: The accuracy of three datasets that apply tuning parameters of the network for transferred visual features

Networks	Accuracy (%)		
	UCF11	UCF50	UCF101
AlexNet	97.01	94.73	92.48
VGG	97.76	95.31	92.56
GoogleLenet	97.46	95.67	94.34
RestNet-50	98.58	97.11	95.63
MobileNet	97.31	93.86	88.14

CONCLUSION

In this article, we proposed a deep transfer learning approach from the image classification problem to realistic activity recognition in video. Because the realistic activity in video is extremely small, we cannot train a large deep convolutional network to have a good performance that does not use augmentation techniques to generate a huge amount of data frames. However, these approaches are easy for image problems, and data generation for video is a major problem due to spatial-temporal relationships in video content. To overcome this obstacle, we propose a robust approach based on transfer learning from a huge network that is trained from the ImageNet dataset. We transfer video data into image data format to apply transfer learning from convolutional neural networks that are trained from ImageNet. To keep two important pieces of information in video, shape and motion features, we trained two networks for each piece of information. The overall recognition accuracy is higher

than 88% on several datasets (UCF11, UCF50 and UCF101). Based on the experimental results from several datasets, our proposed methods are efficient and robust in realistic activity recognition in video.

In future work, we will apply this research to more complex datasets for crowd video surveillance and train motion networks that can be transferred to multiple tasks. A limitation of current research is the use of two networks for shape and motion description in separation models. This problem can be resolved by using multitask learning, which will be addressed in future research.

ACKNOWLEDGMENT

This research is supported by the University of Science, Vietnam National University — Ho Chi Minh City under grant number T2020-01.

REFERENCES

1. Liu J, Luo Jiebo, Shah M. Recognizing Realistic Actions from Videos "in the Wild", IEEE International Conference on Computer Vision and Pattern Recognition (CVPR); 2009;PMID: 20352980. Available from: https://doi.org/10.1109/CVPR.2009.5206744.
2. Reddy KK, Shah M. Recognizing 50 human action categories of web videos, machine vision and applications [journal] (MVAP). September 2012;Available from: https://doi.org/10.1007/s00138-012-0450-4.
3. Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human action classes from videos in the Wild, CRCV-TR-12-01; November 2012;.
4. Wang H, Schmid C. Action recognition with improved trajectories. In: The IEEE International Conference on Computer Vision (ICCV); 2013;Available from: https://doi.org/10.1109/ICCV.2013.441.
5. Peng X, Zou C, Qiao Y, Peng Q. Action recognition with stacked fisher vectors. In: The European Conference on Computer Vision (ECCV); 2014;Available from: https://doi.org/10.1007/978-3-319-10602-1_38.
6. Bobick AF, Davis JW. The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell. 2001;23(3):257-67;Available from: https://doi.org/10.1109/34.910878.
7. Carreira Joao, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017;Available from: https://doi.org/10.1109/CVPR.2017.502.
8. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Nips; 2014. p. 568-76;.
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conf. comput vis. Pattern Recognit (CVPR). June 2016:770-8;PMID: 26180094. Available from: https://doi.org/10.1109/CVPR.2016.90.
10. Deep learning, Ian Goodfellow and Yoshua Bengio and Aaron Courville. MIT Press; 2016;.
11. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115(3):211-52;Available from: https://doi.org/10.1007/s11263-015-0816-y.
12. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Proceedings of the 28th annual conference on neural information processing systems, Montreal; December 2014. p. 3320-8;.
13. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. October 2010;22(10):1345-59;Available from: https://doi.org/10.1109/TKDE.2009.191.
14. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks NIPS. 2012:1106-14;.
15. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Clin Orthop Relat Res [abs]/1409.1556. 2014;.
16. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D et al. Going deeper with convolutions. Clin Orthop Relat Res [abs]/1409.4842. 2014:1-9;Available from: https://doi.org/10.1109/CVPR.2015.7298594.
17. Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE CVPR; May 2014. p. 512-9;Available from: https://doi.org/10.1109/CVPRW.2014.131.
18. Yamato J, Ohya J, Ishii K. Recognizing human action in time-sequential images using a hidden Markov model. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition; 1992. p. 379-85;.
19. Feichtenhofer C, Fan Haoqi, Malik J, He K. SlowFast networks for video recognition. In: The IEEE International Conference on Computer Vision (ICCV); 2019;Available from: https://doi.org/10.1109/ICCV.2019.00630.
20. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018;Available from: https://doi.org/10.1109/CVPR.2018.00813.
21. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang Xiaoou et al. Temporal segment networks: towards good practices for deep action recognition. In: The European Conference on Computer Vision (ECCV); 2016;Available from: https://doi.org/10.1007/978-3-319-46484-8_2.
22. Farnebäck G. Two-frame motion estimation based on polynomial expansion. In: Bigun J, Gustavsson T, editors. SCIA 2003. LNCS. Vol. 2749. Heidelberg: Springer; 2003. p. 363-70;Available from: https://doi.org/10.1007/3-540-45103-X_50.
23. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278-324;Available from: https://doi.org/10.1109/5.726791.
24. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML; 2015;.
25. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T et al. Mobilenets: efficient convolutional neural networks for mobile vision applications. Clin Orthop Relat Res [abs]/1704.04861. 2017;.