COSC 6342 - Machine Learning

TStreamPS: A Two-Stream deep learning method for recognizing human interactions with musical instruments using Pose information in Still images

Christos Smailis (1478522) Michail Koumpanakis (2001935)
Theodoros Tavoulareas (1901216)

Abstract

Many approaches for action recognition focus on general actions, such as "running" or "walking". This work introduces TStreamPS, a method for recognizing interactions between humans and musical instruments in still images, by utilizing a two-stream CNN Deep Learning architecture. The first CNN stream of our method consists of a CNN model that takes as input raw image image pixel information. The second CNN stream of our architecture takes as input pose information extracted from a pose estimation tool, in the form of images. The features extracted from the two CNN streams are then fused to perform the classification process. The experimental results on the PPMI+ dataset, demonstrate that our method boosted accuracy by 2.33% when compared to a baseline single single stream CNN image classification architecture, that utilized only appearance information. Also TStreamPS surpassed the best state-of-the-art method that was evaluated using the PPMI+ dataset by boosting accuracy by 5.13%.

Introduction

The problem of single image action recognition has attracted significant interest from the research community in the past. Current single image action recognition approaches focus on recognizing general actions, such as "running" or "walking" in images captured from the visible spectrum. Not many of them however give a straightforward way to combine pose information with pixel information in a deep learning framework. To address this gap we propose a two-stream CNN architecture that can take advantage of pose information to perform action recognition in images that contain interactions of humans with musical instruments. The contributions of this project report are:

i) Developed and implemented a Deep Learning method for recognizing actions in static images images.

ii) The method takes into account pose information to assist the task of action recognition in still images of people interacting with musical instruments.

iii) Explored and evaluated several backbone CNN architectures for our method.

iv) Explored and evaluated two strategies for fusing the pixel level information with pose information.

Related Work

This section is intended to provide the reader with an overview of the recent trends in the problem of action recognition for static images. The related work section has three categorizations of works, namely (i) Standard image classification methods, (ii) General methods for action recognition in still images and (iii) Action recognition methods in still images that depict interactions with musical instruments. Each section is structured in such a way so as to guide the reader from the more general version of the problem to its specializations. Since the action recognition literature draws heavily from the more general problem of deep learning methods for image classification, it was deemed necessary to also include a synopsis of the latter in this section.

Standard image classification methods

Convolutional neural networks (CNNs) are the pillars of modern computer vision systems. In the last few years, significant progress has been made in the image classification and computer vision fields by state-of-the-art convolutional neural networks architectures. Simple CNN architectures such as VGG16 [16] were introduced first as a solution to image classification problems. The model achieved outstanding performance at the ImageNet dataset [15] and as a result was considered state-of-the art back in 2014. Throughout the years, more convolutional models were introduced to the public that consisted of more complex and deep architectures. ResNet [5] and its variants (Resnet50, Resnet101, Resnet151) outperformed previous CNN models on the ImageNet and CIFAR-10 [9] datasets and also achieved higher training efficiency by introducing layers that learned residual functions. At the same time, InceptionV3 [18] was introduced to the public. Just like ResNet, the model's selling point was the optimization of crucial parameters that improved its computational efficiency. Specifically, its improvements were label smoothing, factorized 7 x 7 convolutions, the use of an auxiliary classifier to propagate label information lower down the network and finally the use of batch normalization. Later on, Xception [1] was introduced as an improved version of InceptionV3 that utilizes a modified depth-wise separable convolution plus residual connections.

General methods for action recognition in still images

Most recent approaches for action recognition in still images, make use of deep learning based architectures that take into account different types of cues to assist the action recognition process. Thus the methods can be categorized based on what types of cues they use:

Contextual region based methods: Inspired by the success of the R-CNN object detector [3], Gkioxari *et al.* [4] proposed a similar action recognition method namely R*CNN static images utilizing the contextual regions produced by the selective search algorithm based on their importance to assist and performs score level of the most informative contextual region with the full person crop to infer the action label.

Attention based methods: Other methods, such as the one introduced by Diba *et*

al.[2], emphasize visual attention mechanisms within deep learning models to learn mid-level representations of actions.

Person body parts based methods: Another category of methods gives emphasis person body parts in order to perform action classification for images. In Gkioxari *et al.* [4], the authors attempt to perform classification of actions using only features extracted only from a person's body part regions in the image such as the hands, the feed and the head. In Zhao *et al.* [24], the authors attempt to fuse features from body parts with the full crop of the person to infer the action label for an image.

Boxless based methods: In Zhang *et al.* [23], authors attempt to perform action classification in full images depicting scenes, without prior knowledge for bounding boxes or the crops (boxless) of the persons involved involved in these actions.

Learning using Privileged Information Based (LUPI) Methods: Leveraging additional information available only while training image classification models is a concept that has been addressed in many different contexts in the literature. In one of their demonstrated uses of LUPI, Vapnik *et al.* [19], leveraged textual descriptions of hand-drawn digit images as privileged information, to further assist the recognition of handwritten characters. Currently only one method exists for applying LUPI in the context of action recognition from still images, namely RECASPIA, introduced by Smailis *et al.* [17]. The authors of this work treated pose annotations as well as person attributes annotations as privileged information, available only during training time in order to formulate a LUPI method that can perform action classification in still images for recognizing carrying actions.

Pose based methods: Only one work in the past literature introduces a method utilizing pose cues to model actions for the problem of action recognition in still images. More specifically in Wang *et al.* [20], the authors proposed a new body descriptor, named limb angle descriptor (LAD), which uses the relative angles between the limbs in 2D skeletal information. The information from the pose vector is then incorporated to a deep learning architecture based on R*CNN by passing it through a stream of fully connected layers and performing score level fusion.

Table 1: Overview of the general methods for action recognition in still images.

Methods	Category					
	Part	**Context**	**Visual Attention**	**Boxless**	**LUPI**	**Pose**
Gkioxari et al.	-	Y	-	-	-	-
Zhao et al.	Y	-	-	-	-	-
Gkioxari et al.	Y	-	-	-	-	-
Diba et al.	-	-	Y	-	-	-
Zhang et al.	-	-	-	Y	-	-
Smailis et al.	-	-	-	-	Y	-
Wang et al.	-	-	-	-	-	Y

Action recognition methods in still images that depict interactions with musical instruments

In this part of our related work section, the most recent existing studies that deal specifically with the problem of action recognition methods in still images that depict interactions with musical instruments are presented. The most well known dataset in the literature, containing still images with people and a variety of musical instruments is named People Playing Musical Instruments (PPMI) [22]. We thus adopt it in this project. The following works also use PPMI for their evaluation. Li *et al.* [12] proposed a deep selective feature learning network on the PPMI+ dataset, which can automatically learn the feature maps with both fine-grained and global information. Li *et al.* [11] modeled human-object interaction (HOI) by using some recent pre-trained deep networks as feature extractors to create a hierarchical representation of the images, which was evaluated on a smaller version of the PPMI+ dataset that contains only 7 musical instruments: bassoon, erhu, flute, French horn, guitar, saxophone, and violin. Qi *et al.* [14] used a convolutional neural network (CNN) originally learned for object recognition (VGG16) as a base network and then transferred it to action recognition by training the base network jointly with an inference of poses obtained using the "poselet activation vector" (PAV) [13]. Finally, Yan *et al.* [21] introduced a method that combines a CNN with the Vector of Locally Aggregated Descriptors (VLAD) encoding scheme [7], in order to better capture global contextual information. It should be taken into consideration that

the last two studies evaluated their approaches on the 24-category classification task (both PPMI+ and PPMI- datasets).

Methodology

Problem Statement: Given an image $\boldsymbol{I}$ depicting human actions, the goal of this work is to predict the action labels $y \in \mathcal{Y}$, performed by the central person within image $\boldsymbol{I}$, where $\mathcal{Y} = \{y_1, ...y_N\}$ is a set of N actions.

Method Overview: The overview of our method can be seen in Fig.1. Our method is based on previous architectures for CNN based image classification. Pose information in the form of skeleton images, can be a very strong indicator of what type of action is performed in an image. To this end our method consists of two information streams representing the appearance of a human, as well as pose information from his skeleton. These two streams are further discussed in the following sections:

Appearance Stream: The purpose of the appearance stream is to extract features from RGB pixel values from the original image $\boldsymbol{I}$ that displays a human interacting with a music instrument.

Pose Stream: The purpose of the pose stream is to extract features from an image that contains pose information that corresponds to each image $\boldsymbol{I}$. Pose information in our method are represented as an image depicting the skeleton of the human depicted in image $\boldsymbol{I}$. To produce the pose information needed by the Pose Stream of our method we employed the AlphaPose Pose Estimation Method [10].

Fusion: Features from both the appearance and pose streams are extracted from the fully connected layer of each of the streams. The two feature vectors are then concatenated into a single vector that is then passed to the softmax classifier to infer the class of the action depicted in the original image $\boldsymbol{I}$.

Implementation Details: The adopted CNN architecture, was based on Xception [6] and was trained from scratch with random initial weights but with a specific random state that reproduces our results. Training was performed through the Adam optimizer [8]. Since this is a multi-label classification problem, the Softmax categorical cross-entropy loss was adopted. The batch size was set to 32 samples with a learning rate of 0.001, β_1 and β_2 values of 0.9 and 0.999 respectively. We

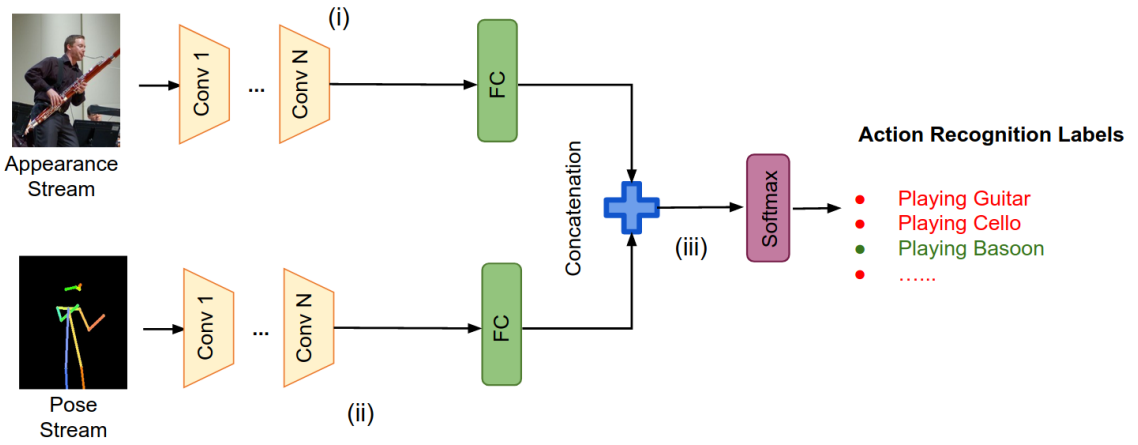

Figure 1: Overview of the TStreamPS Method: (i) The appearance stream consists of CNN architecture that extracts features from a raw image. (ii) The pose stream consists of CNN network that takes as input the result of a pose estimation method from the raw image. (iii) The features from the fully connected layers at the end of each stream are concatenated to a single feature vector and passed to a softmax classifier to obtain the final class prediction.

performed random shuffling with a specific random seed for both skeleton pose images and appearance images (sorted data with one-to-one correspondence between the raw images and thier pose information). Appearance images and skeleton pose images were scaled to 224×224 pixels for each of the two information streams of the architecture.

Evaluation

In this section, we assess the performance of the TStreamPS method using the PPMI+ dataset. To better understand TStreamPS's characteristics, it is compared against another baseline for image action recognition. We also perform a baseline study to evaluate the performance of each individual information stream of the TStreamPS architecture. To establish the best way to perform fusion we conduct experiments comparing the performance of the TStreamPS architecture using two fusion methods (namely feature concatenation and score level fusion). Finally, we compare the performance obtained by TStreamPS against other action recognition methods from the literature that have been evaluated with the PPMI+ dataset.

Figure 2: Sample images from the PPMI+ setting of the PPMI dataset: each image corresponds to each one of the 12 categories with the action of "playing the instrument".

PPMI Dataset

The People Playing Musical Instrument (PPMI) dataset is introduced by Yao and Fei-Fei [22] and contains images of humans interacting with 12 different musical instruments, namely: bassoon, cello, clarinet, erhu, flute, french horn, guitar, harp, recorder, saxophone, trumpet, and violin. The dataset makes the following distinctions on images based on whether, for each musical instrument, people are "playing the instrument" "holding the instrument". Thus two settings are formed were images of people only playing the instruments are used (PPMI+ setting) or either holding or playing the instruments (PPMI- setting). Hence, there are 24 categories to classify. The PPMI dataset provides cropped images that contain the person of interest at the center. Some images from the dataset are shown in Fig. 2. We evaluated the TStreamPS method using the PPMI+ setting and followed the evaluation protocol defined by the PPMI authors. More specifically, for each playing instrument class, 100 cropped images are used for training (20% validation) and 100 cropped images are used for testing.

Experiments

We performed four types of using the PPMI dataset. Each of the experiments have a different goal:

- **Experiment 1:** The first experiment attempts to figure out which general im-

Table 2: Results from Experiment 1: Comparison of the general image classification methods using appearance information. The best performing architecture is the Xception with an accuracy of 85% on the PPMI+ dataset.

General Image Classification Methods	**Xception**	**Resnet50**	**Resnet101**	**VGG16**	**VGG19**
Test Set - Accuracy % (Appearance Information)	**85.00**	72.92	76.00	65.00	63.00

age classification CNN architecture would be the most successful in classifying the PPMI+ dataset using appearance information. We adopt the most successful architecture as the backbone of the appearance stream of the TStreamPS method. From Table 2 that contains the results of the experiment we infer that the best performing general image classification method using appearance information, is the Xception architecture with an accuracy of 85%. We thus opted to adopt the Xception architecture as the backbone architecture for the appearance stream of the TStreamPS method.

- **Experiment 2:** The second experiment attempts to assess the performance of each of the pose stream of the TStreamPS model when attempting to use it individually for performing action recognition in the PPMI+ dataset. This way we can get an intuition over how well it can contribute in the final result and which general image architecture would be the ideal for this type of data. As we can see in Table 3. the Xception architecture achieves the best results again. Thus we also adopt it as the backbone architecture of the pose stream in the TStreamPS method.

- **Experiment 3:** In the third experiment we attempt to figure out what is the optimal way to fuse information from the two streams of the TStreamPS method. We thus compare the performance of two deep learning based fusion methods, namely, feature concatenation and score level fusion. Using feature concatenation between appearance and pose features the TStreamPS method achieves the best accuracy score which is equal to **87.33%**. Using score based fusion over separately trained appearance and motion streams and by averaging the scores obtained by two softmax classifiers attached at the end of each

Table 3: Results from Experiment 2: Comparison of the general image classification methods using pose information. The best performing architecture is the Xception with an accuracy of 57.75% on the PPMI+ dataset.

General Image Classification Methods	**Xception**	**Resnet50**	**Resnet101**	**VGG16**	**VGG19**
Test Set- Accuracy % (Pose Information)	**57.75**	56.00	55.00	50.00	52.00

Table 4: Results from Experiment 4: Comparison of the TStreamPS method with other state of the art methods evaluated under the same settings of the PPMI+ dataset. The best performing method is TStreamPS. Part of the performance boosting is attributed to Xception being used as a backbone architecture for the two streams.

Methods	**TStreamPS (Ours)**	**DSFNet + Resnet34[12]**	**VGG16 + VLAD SP**[21]	**VGG16 + PAV**[14]
Test Set - Accuracy %	**87.33**	72.36	81.30	82.20

of the streams we obtain an accuracy score 85.08%. We thus opted for using concatenation based fusion for the two streams in the TStreamPS method.

- **Experiment 4:** In the fourth experiment we compare the TStreamPS method against other state-of-the-art deep learning approaches that have been evaluated under the same settings of the PPMI+ dataset, in order to figure out if it surpasses them or not. As we can see in Table 4, TStreamPS surpassed the best state-of-the-art method that was evaluated using boosting accuracy by 5.13%. However, the part of the performance increase is related to the Xception network being used as the backbone of the TStreamPS information streams.

Conclusion

In this project report, we presented a deep learning method named TStreamPS for performing action recognition in still images containing persons interacting with musical instruments. Our method utilizes two streams for using appearance and pose information from the images. We carefully examined and reviewed related

past works from the literature and we performed several experiments to justify the decision choices for the design of our method, by assessing each individual component of our method. Finally, we evaluated and compared the performance of our method against other state of the art works from the recent literature.

References

[1] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. CoRR (2016)

[2] Diba, A., Pazandeh, A.M., Pirsiavash, H., Gool, L.V.: Deepcamp: Deep convolutional action attribute mid-level patterns. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3557–3565. Las Vegas, NV (2016)

[3] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

[4] Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with R*CNN. In: Proc. IEEE International Conference on Computer Vision, pp. 1080–1088. Santiago, Chile (2015)

[5] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR (2015)

[6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. Las Vegas, NV (2016)

[7] Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3304–3311 (2010)

[8] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)

[9] Krizhevsky, A., Nair, V., Hinton, G.: Cifar-10 (canadian institute for advanced research) URL http://www.cs.toronto.edu/ kriz/cifar.html

[10] Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. arXiv:1812.00324 [cs] (2019)

[11] Li, R., Liu, Z., Tan, J.: Reassessing Hierarchical Representation for Action Recognition in Still Images. IEEE Access **6**, 61386–61400 (2018)

[12] Li, Z., Ge, Y., Feng, J., Qin, X., Yu, J., Yu, H.: Deep Selective Feature Learning for Action Recognition. IEEE (2020)

[13] Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR 2011, pp. 3177–3184 (2011)

[14] Qi, T., Xu, Y., Quan, Y., Wang, Y., Ling, H.: Image-based action recognition using hint-enhanced deep neural networks. Neurocomputing **267**, 475–488 (2017)

[15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)

[16] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[17] Smailis, C., Vrigkas, M., Kakadiaris, I.A.: Recaspia: Recognizing carrying actions in single images using privileged information. In: IEEE International Conference on Image Processing (ICIP), pp. 26–30 (2019)

[18] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR (2015)

[19] Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22**(5-6), 544–557 (2009)

[20] Wang, X., Li, K., Li, Y.: A deep model combining structural features and context cues for action recognition in static images. In: Proc. International Conference on Neural Information Processing, pp. 622–632. Guangzhou, China (2017)

[21] Yan, S., Smith, J.S., Zhang, B.: Action Recognition from Still Images Based on Deep VLAD Spatial Pyramids. Signal Processing: Image Communication **54**, 118–129 (2017)

[22] Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 9–16 (2010)

[23] Zhang, Y., Cheng, L., Wu, J., Cai, J., Do, M.N., Lu, J.: Action recognition in still images with minimum annotation efforts. IEEE Transactions on Image Processing **25**(11), 5479–5490 (2016)

[24] Zhao, Z., Ma, H., You, S.: Single image action recognition using semantic body part actions. In: Proc. IEEE International Conference on Computer Vision, pp. 3411–3419. Venice, Italy (2017)