Comparison of deep learning techniques for video-based automatic recognition of Greek folk dances

Georgios Loupas, Theodora Pistola, Sotiris Diplaris, Konstantinos Ioannidis, Stefanos Vrochidis, and Ioannis Kompatsiaris

Information Technologies Institute - CERTH, Thessaloniki, Greece
{loupgeor, tpistola, diplaris, kioannid, stefanos, ikom}@iti.gr

Abstract. Folk dances consist an important part of the Intangible Cultural Heritage (ICH) of each place. Nowadays, there is a great amount of videos related to folk dances. An automatic dance recognition algorithm can ease the management of this content and enforce the promotion of folk dances to the younger generations. Automatic dance recognition is still an open research area that belongs to the more general field of human activity recognition. Our work focuses on the exploration of existing deep neural network architectures for automatic recognition of Greek folk dances depicted in standard videos, as well as the experimentation with different representations of input. For our experiments, we have collected YouTube videos of Greek folk dances from north-eastern Greece. Specifically, we have validated three different deep neural network architectures using raw RGB and grayscale video frames, optical flow, as well as "visualised" multi-person 2D poses. In this paper, we describe our experiments, and, finally, we present the results and findings of the conducted research.

Keywords: Dance recognition · Deep learning · Folk dances · Intangible cultural heritage.

1 Introduction

Folk dances comprise an important part of the Intangible Cultural Heritage (ICH) of each place. They are associated with folk music and social events (e.g., celebrations, customs) and help people to maintain their cultural identity. In the past, their preservation was only possible through their transmission from the old generations, which often led to variations of the original dances. In modern days, video recording has been the main mean of digitising and preserving the different types of dances, either folk or contemporary. This fact led to a great amount of related data, which can be exploited by machine learning technologies for better understanding and automatic recognition of different types of dances from computers.

Automatic dance recognition is a subdomain of the wider research area of human activity recognition. This technology aims at the automatic categorisation of dances depicted in videos into specific categories based on the extracted features. Due to a multitude of constraints, dance recognition is a challenging task for computer vision. Video resolution, frame rate, poor lighting, blurring, complex backgrounds, and occlusions of dancers are just some noteworthy limitations that should be considered. In addition, the usage of dance videos for the wider problem of video classification is a more challenging task, as dance videos belong to the category of highly dynamic videos and there is a great variety of dynamics for the different dance types. Despite the challenges that automatic dance recognition may face, such a system can be useful for a variety of applications in the fields of Cultural Heritage (CH) and education. For example, an automatic system for the recognition of Greek folk dances can enhance dance videos with metadata describing the type of the depicted dance. As a result, these videos could be easily retrieved from big databases based on the recognised type of dance facilitating the preservation and management of relevant data to this kind of ICH. Moreover, automatic dance recognition can assist the learning of different dances. For instance, an idea is to have a mobile application that enables the user to capture a video sample of people dancing, return an analysis result based on an integrated automatic dance recognition model and be able to provide more information about this type of dance (e.g., steps, history, information about the music etc.). Another suggestion could be to exploit automatic dance recognition to find the parts of a video that contain a specific dance.

In this work, we focus on the recognition of Greek folk dances from a specific area of north-eastern Greece, namely Thrace. Specifically, we experiment with well known deep learning network architectures designed for general video classification, in order to create a system for automatic recognition of four selected traditional Greek dances (*Karsilamas*, *Hasapikos*, *Gikna* and *Baintouska*). In the context of our research, we have created a new dataset of YouTube videos that depict the aforementioned Greek folk dances. In addition, we explored different types of input representations (raw RGB and grayscale frames, optical flow and multi-person 2D poses of dancers). Our main contribution is the evaluation and comparison of common deep learning architectures used for the more general tasks of video classification and human activity recognition towards automatic dance recognition from videos. The challenge of this task is that the networks should discriminate the motion features of each dance and not rely on background features for the classification, as backgrounds are either the same or similar for different dances. To the best of our knowledge, the selected architectures have not been used for the task of automatic dance recognition from videos before.

The rest of this paper is organised as follows. In section 2, we briefly present the most prominent related works in the general field of human activity recognition and its sub-field of automatic dance recognition. Section 3 describes the dataset we used for our experiments. In section 4, a quick presentation of the network architectures is made, while training details and results of the conducted experiments are provided. Finally, we conclude our paper with section 5, where the main outcomes and future steps are outlined.

2 Related Work

Automatic dance recognition belongs to the wider domain of human activity recognition. In this section, we provide a summary of the most dominant works related to human activity recognition and automatic dance recognition through videos.

Human Activity Recognition: Early video classification and activity recognition methods, like [25], were based on hand-crafted motion features extracted from the pixels of the video frames that were exploited by machine learning approaches, such as Bag-of-Words model [9]. In the last decade, the great performance of deep learning in image classification led to its application also for human activity recognition from video, finding ways to integrate the temporal information too. Initial works developed fusion techniques to exploit temporal information, like in [16], while others created two-stream Convolutional Neural Networks (CNNs) [21], where one stream is fed with spatial information (e.g., video frame/image) and the second with the temporal information in the form of stacked optical flow vectors. The combination of CNNs with Long Short-Term Memory Networks (LSTMs) [6] was also tested as a solution. In addition, CNNs with 3D convolutions were developed [14] to take into account the temporal aspect, with promising results, such as C3D [24], I3D [4] and SlowFast [8]. 3D Residual Networks (3D ResNets) [11] were also proposed for the task of human activity recognition through video. In order to decrease the complexity of the training procedure for the 3D networks, the idea of 3D factorisation was introduced in P3D [20]. Temporal Segment Networks (TSN) [26] is another deep learning architecture utilised for video-based human activity recognition. Skeleton-based methods have also gained a lot of attention recently. The human skeletons in a video are mainly represented like a sequence of coordinates that are extracted from 2D pose estimation algorithms [3]. The results of these methods are unaffected by background variations and changes in illumination conditions because only human skeletons are used for activity recognition. Graph Convolutional Networks (GCNs) [28] is another famous method for human activity recognition. Finally, networks with attention, such as transformers including TimesFormer [2], ViVit [1] and Video Masked AutoEncoders (VideoMAE) [23], are explored in this direction too.

Automatic Dance Recognition: Three different methods for the extraction of handcrafted features were explored in [15], along with a Bag-of-Words approach, for the recognition of five Greek folk dances. The authors of [10] explored the problem of distinguishing Greek folk dances from other kinds of activities, as well as from other dance genres, using video recordings. To achieve this, they adopted dense trajectories descriptors along with Bag-of-Words (BoW) model to present the motion depicted in the videos. For the classification step they have used a Support Vector Machine (SVM). In [17] the authors present a system for the classification of Indian classical dance actions from videos using CNNs that were previously used for action recognition. Similarly, in [13] Indian classical dance classification is achieved through the use of a new deep convolutional neural network (DCNN) that is based on ResNet50. The above methods, though, do not exploit temporal information, but only sets of specific poses depicted in video frames. In [5] different approaches for automatic dance recognition from videos are proposed, considering the temporal information. More specifically, the authors present a comparison of numerous state-of-the-art techniques on the "Let's Dance" dataset using three different representations (video, optical flow and multi-person 2D pose data). In [19] a differential evolutionary convolutional neural network model is applied for the classification of dance art videos. The dance videos were collected from large video websites and consist of seven dance categories: classical dance, ballet, folk dance, modern dance, tap dance, jazz dance and Latin dance. Automatic dance recognition approaches that combine visual data with other types of data, like audio [27], were also proposed in the literature. However, in the context of this paper we focus only on visual-based approaches.

3 Dataset Description

Each country or either region has its own particular dancing style, which evolved from its culture and history. In some dances, specific parts of the body are more dominant than others, like legs and arms. Moreover, there are dances, where a specific sequence of movements must be followed, while others allow the dancer to create their own choreography.

Greek folk dances are mostly danced in a circle, where the dancers are connected through their hands in different manners depending on the dance. There are also dances that are danced in couples or more freely, where specific steps are followed. In this work, we focus on the automatic recognition of four characteristic dances of the region of Thrace in north-eastern Greece. These dances are: i) *Karsilamas*, ii) *Hasapikos*, iii) *Gikna* and iv) *Baintouska.* Karsilamas is danced in couples, where one dancer faces the other, without joined hands. Hasapikos, Gikna and Baintouska are danced in a circle, where the hands of the dancers are connected in different ways depending on the dance. You can find example snapshots of these dances in *Figure 1*.

Our new dataset consists of videos that we crawled from YouTube. These videos contain the four selected types of Greek folk dances performed at Greek folk dance festivals in most cases. See *Table 1* for more details. The dataset cannot be public at the moment due to copyright issues.

[1] https://www.youtube.com/watch?v=YIcJa0Te6z0

[2] https://www.youtube.com/watch?v=LRqXRLQ4hQ0

[3] https://www.youtube.com/watch?v=f61-KQeXtlo

[4] https://www.youtube.com/watch?v=bjx8Vv5H7kst=62s

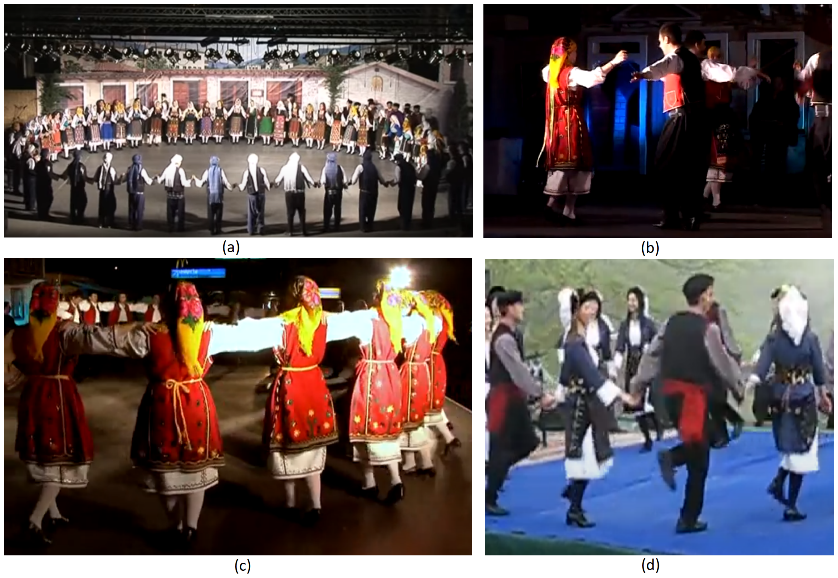

Fig. 1. Examples of each dance category from our dataset: a) Gikna [1], b) Karsilamas [2], c) Hasapikos [3], and d) Baintouska [4].

Dance	Karsilamas	Hasapikos	Gikna	Baintouska
Number of Videos	15	15	15	15
Total Duration	40.02 min	37.35 min	31.92 min	28.02 min

Table 1: Number of videos used in our experiments and total duration per dance category.

4 Experiments and Results

In this section, we briefly describe the deep neural networks that we utilised for our experiments and the preprocessing of the input videos. We also provide details about the training procedure. Finally, we present our experimental results.

4.1 Network Architectures and Input Representations

In the context of our research, we experimented with three deep neural networks that are widely used in the field of video classification, namely the *C3D* [24], *3D ResNet* [11] and *SlowFast* [8] architectures that are briefly described bellow. We choose SlowFast as it is a state-of-the-art network that has shown strong performance on the problem of video classification and activity recognition and its comparison to the other two baseline networks would give us interesting outcomes.

C3D Architecture: *C3D* [24] comprises a 3D Convolutional Network (3D ConvNet), which aims to address the problem of learning spatiotemporal features from videos. This architecture consists of 8 convolution layers, 5 pooling layers, followed by two fully connected layers, and a softmax output layer.

3D ResNet Architecture: *Residual Networks (ResNets)* [12] introduced the concept of the Residual Blocks. The core of the Residual Blocks is the Skip Connection, which is a direct connection that skips some layers of the network. These connections pass through the gradient flows of the network from later layers to early layers, and facilitate the training of very deep networks. The need for the Residual Block was the limitation of the number of stacked layers that one can use to build a Deep CNN, as it has been observed that the conventional CNNs have a maximum depth threshold. The difference of *3D ResNets* compared to the original ResNets is that they perform 3D convolution and 3D pooling.

SlowFast Architecture: The *SlowFast* network [8] was originally proposed for video recognition. Its architecture consists of two pathways, namely the *Slow* and the *Fast* pathways. The *Slow* pathway gathers spatial and semantic information from images or sparse frames, and it works at low frame rates and slow refreshing speed. On the other hand, the *Fast* pathway captures rapidly changing motion, as it works at high refreshing speed and temporal resolution. These two pathways are finally fused by lateral connections. In this way, this network takes into account both the static and dynamic content of a video.

Moreover, we experimented with different representations of input to feed the above networks (See Figure 2). We tested the networks using the raw RGB and grayscale frames. This means that the whole information depicted in the video clips passed through the networks that automatically extracted the features that can differentiate one dance type from the others. We also experimented with optical flow data that can help us track how each pixel changes from one frame to the next. Optical flow is independent of the visual information in the original frames and emphasizes only motion information, which helps the network focus on motion properties. We use the RAFT algorithm [22] for the extraction of the optical flow in our experiments.

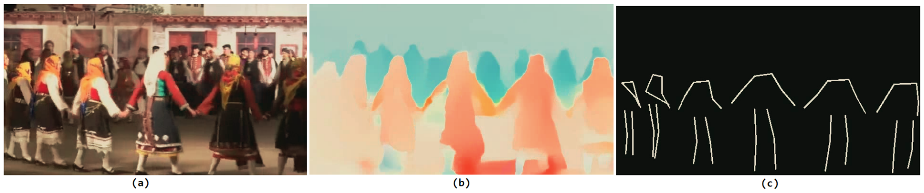

Fig. 2. Different input representations are used to feed the networks: (a) RGB video frame, (b) RAFT optical flow and (c) multi-person 2D poses. Video source [5]

For the extraction of the multi-person 2D poses, we utilised the AplhaPose algorithm [7], which also supports pose tracking and, based on experiments, has better results compared to OpenPose [3] and OpenPifPaf [18]. We have to note here that we trained the networks using the “visualized” pose data that practically are the grayscale images of the extracted pose skeletons. This repre-

[5] https://www.youtube.com/watch?v=YIcJa0Te6z0

sentation is also independent of the visual information in the original frames, as well the body shape of each subject, and encodes only the body position, which helps the network to focus on the spatio-temporal information of the body pose.

4.2 Training Procedure and Experimental Results

For the training procedure, a total duration of 20 minutes was used for each class that correspond to a different number of videos per class, because the videos had different durations. For validation and testing, 4 minutes per class were used, respectively, that correspond to 3 videos per class. There was no overlapping between the videos that were used for testing and those exploited in the training procedure. This leads to a split of 80% for training, 15% for validation and 15% for testing (consistent across experiments). Also, for every video of the final dataset, the extraction of RAFT optical flow and multi-person 2D poses using the AlphaPose algorithm followed (see *Figure 1*). Afterwards, each video with the specific representation according to the experiment, was splitted into clips of a constant number of frames. In this work, 30 frames per clip were used that correspond to a duration of 2.36 seconds. This is due to the fact that we sample every other frame and all videos were at 25 frames per second (fps). 30 frames per clip showed better results as opposed to 15 frames (1.16 seconds) and this make sense, if we consider that dances like Gikna and Baintouska have a similar form, so a larger duration of a clip can capture more temporal information leading to better discrimination. Each clip was then fed to the deep neural network, along with its label to contribute to the training of each model. We trained all the architectures from scratch using our dataset.

Each network received the aforementioned representations of the input, except for the Slowfast, where the optical flow representation was omitted due to the inherent ability of this network to capture rapidly changing motion through the fast pathway. All the videos were spatially transformed according to the input requirements of each network. The size of each sample is n channels × 30 frames × Height × Width, where n=3 for RGB and RAFT representations and n=1 for grayscale, Gray-RAFT and multi-person 2D poses. The batch size is 20 clips and for all of our experiments we used an Adam optimizer and an early stopping technique. Specifically, in the case where the validation loss was saturated for 4 epochs, the learning rate was divided by 10. In case there was no progress on validation loss for 20 epochs, the training was forced to stop.

For our experiments, we used a NVIDIA GeForce RTX 3090 GPU. This fact restricted our ability to choose bigger batch sizes. We could increase the batch size and spatial resolutions with the use of more than one GPU, which would lead to improved classification results.

Below, we provide detailed information for the training of each architecture, along with the corresponding results in terms of F1-measure (F1) and accuracy (Acc).

C3D parameters and results: In the context of this paper, the *C3D* network was trained from scratch. We spatially resize each sample at 112 × 112 pixels according to the input requirements of the network. In the training process we use Adam optimizer with a starting learning rate of 0.0001 for RAFT, Gray-RAFT and Skeleton input representations, 0.001 for grayscale and 0.00001 for RGB.

Architecture	**F1 (clips)**	**Acc (clips)**	**F1 (videos)**	**Acc (videos)**
C3D-RGB	32.65%	35.33%	27.50%	41.66%
C3D-Gray	41.24%	50.92%	48.93%	58.33%
C3D-RAFT	53.20%	52.65%	66.43%	66.66%
C3D-Gray-RAFT	43.32%	42.90%	**74.16%**	**75.00%**
C3D-Poses	**56.84%**	**57.82%**	72.91%	75.00%

Table 2: Experimental results for the C3D architectures.

We find that C3D works best for the "visualised" multi-person 2D poses input representation with RAFT to follow, having quite a difference from the rest of the representations. This is something we would expect as these 2 representations are independent of visual information and enable the network to focus on features of body posture and limb movement, i.e. features related to dance.

For the categorization of an entire video, it is important that the model correctly predicts the majority of the video clips that make it up. We notice that for the C3D network a representation independent of visual information in the original frames can do this with a notable difference.

3D ResNet parameters and results: In the context of this paper, we have experimented with 3D ResNet-50 and 3D ResNet-101 testing the aforementioned different input representations, while training the networks from scratch. We spatially resized each sample at 224 × 224 pixels according to the input requirements of the network. For the training of these networks on our dataset, we have used Adam optimiser. The starting learning rate for 3D-ResNet50 was 0.0001 for all representations, while for 3D-ResNet101 was 0.001 for RGB and Gray-RAFT input, 0.0001 for grayscale, RAFT and "visualized" multi-person 2D poses input.

Architecture	**F1 (clips)**	**Acc (clips)**	**F1 (videos)**	**Acc (videos)**
3D ResNet-50-RGB	49.10%	55.07%	58.93%	66.66%
3D ResNet-101-RGB	63.09%	62.23%	**74.16%**	**75.00%**
3D ResNet-50-Gray	48.12%	51.21%	47.62%	50.00%
3D ResNet-101-Gray	54.78%	62.94%	56.84%	66.66%
3D ResNet-50-RAFT	29.65%	29.55%	30.83%	33.33%
3D ResNet-101-RAFT	35.02%	36.30%	35.00%	41.66%
3D ResNet-50-Gray-RAFT	28.49%	28.55%	29.92%	33.33%
3D ResNet-101-Gray-RAFT	25.56%	34.72%	23.21%	33.33%
3D ResNet-50-Poses	45.51%	43.58%	52.14%	50.00%
3D ResNet-101-Poses	**63.28%**	**65.44%**	72.68%	**75.00%**

Table 3: Experimental results for the 3D ResNet architectures.

3D-ResNet50 has a similar performance for RGB and gray input, which outperforms the rest of the input representations. The performance decreases noticeably for RAFT and Gray-RAFT representations, while for the "visualized" multi-person 2D poses it is better than these two. The poor performance on representations that are independent of visual information seems to be contrary to what we might expect and perhaps finer tuning of the hyperparameters is required to increase performance.

Similar results are given by 3D-ResNet101 which, however, achieves higher performance overall than 3D-ResNet50 and gives the best results for the "visualized" multi-person 2D poses, in line with our intuition. However, the results for optical flow representation are also worse than the original frames, which indicates that perhaps a better setting of the hyperparameters is required.

We notice that 3D-ResNet50 correctly predicts the majority of video clips, thus categorizing more accurately the videos for RGB input relative to the rest representations, while 3D-ResNet101 performs better for RGB and "visualized" multi-person 2D poses.

SlowFast parameters and results: As part of our work, we have trained the *SlowFast* architecture with a ResNet-50 and ResNet-101[12] backbone, speed ratio $\alpha = 8$, channel ratio $\beta = 1/8$ and $\tau = 16$ on our dataset. We spatially resized each sample at 224 × 224 pixels according to the input requirements of the network. In the training process we used Adam optimizer and the starting learning rate for SlowFast-ResNet50 was 0.001 for multi-person 2D poses input, 0.00001 for RGB and grayscale input, while for SlowFast-ResNet101 was 0.001 for RGB input, 0.001 for grayscale and multi-person 2D poses input.

Architecture	**F1 (clips)**	**Acc (clips)**	**F1 (videos)**	**Acc (videos)**
SlowFast-ResNet50-RGB	36.47%	43.92%	45.00%	50.00%
SlowFast-ResNet101-RGB	45.88%	53.93%	50.00%	58.33%
SlowFast-ResNet50-Gray	48.56%	56.65%	45.89%	50.00%
SlowFast-ResNet101-Gray	56.33%	65.38%	60.00%	66.66%
SlowFast-ResNet50-Poses	**69.67%**	**70.52%**	**81.25%**	**83.33%**
SlowFast-ResNet101-Poses	46.37%	48.37%	57.50%	58.33%

Table 4: Experimental results for the SlowFast architectures.

For the SlowFast-ResNet50 we notice that the simpler the input becomes from the point of view of visual information, the network gets better performances with the best performance being for the "visualized" multi-person 2D poses, again following our intuition and what we would expect as the network does not focus on information related to the space or the costumes of the dancers.

On the contrary, SlowFast-ResNet101 seems to be able to generalize better when also given the visual information in the original frames comparatively with the "visualized" multi-person 2D poses, something that may indicate a tendency for overfitting, as the model is more complex in relation to SlowFast-ResNet50 and the "visualized" multi-person 2D poses representation is simpler in relation to the original frames.

For whole video classification, SlowFast-ResNet50 correctly predicts the majority of video clips, thus categorizing more accurately the videos for 2D skeleton input representation relative to original frames, while SlowFast-ResNet101 performs better for grayscale input representation, results that makes sense according to the results per video clip.

5 Conclusions

In this paper, we focus on the task of automatic dance recognition, specifically for traditional Greek dances from the region of Thrace in north-eastern Greece. In the context of our research, we created a new dataset by crawling videos that contain four different kinds of traditional Greek Thracian dances from YouTube. The specific dance types that consist our dataset are *Karsilamas*, *Hasapikos*, *Gikna* and *Baintouska*. We used this dataset in order to evaluate and compare existing architectures that were previously used in the wider field of video recognition and activity recognition for the more specific task of automatic dance recognition. Towards this goal, we trained the C3D, 3D ResNet and SlowFast architectures on our dataset, while also tested different representations of the input (raw RGB and grayscale frames, optical flow and multi-person 2D poses). Through our results, we observe that the representation of the input plays a crucial role in the efficiency of the networks on the problem of dance recognition from video and representations based on motion and multi-person 2D poses could help the networks to learn essential information about the motion of the body increasing the performance on the task. Moreover, we conclude that the "visualized" multi-person 2D poses are giving the best performance in most of the cases, in line with our intuition, with SlowFast network to be the most effective, confirming its power to handle challenging activity recognition tasks.

One of the next steps for our research on the automatic recognition of Greek folk dances is to extend our dataset by adding more videos per dance type or experimenting with more dance types. Moreover, we consider some video preprocessing techniques in order to remove non-relevant to dance parts of the videos (e.g., titles, other scenes). As a future work, other architectures except for 3D CNNs will also be examined, like networks with attention, such as transformers (e.g TimesFormer [2], ViVit [1]) and video masked autoencoders (VideoMAE [23]), in order to pretrain them in a large population of dance videos and then fine-tune them for the dances of our interest to increase their performance.

Acknowledgments

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T2EDK-01856).

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846 (2021)
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML (2021)
3. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
5. Castro, D., Hickson, S., Sangkloy, P., Mittal, B., Dai, S., Hays, J., Essa, I.: Let's dance: Learning from online dance videos. arXiv preprint arXiv:1801.07388 (2018)
6. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
7. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2334–2343 (2017)
8. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
9. Foggia, P., Percannella, G., Saggese, A., Vento, M.: Recognizing human actions by a bag of visual words. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics. pp. 2910–2915. IEEE (2013)
10. Fotiadou, E., Kapsouras, I., Nikolaidis, N., Tefas, A.: A bag of words approach for recognition of greek folk dances. In: Proceedings of the 9th Hellenic Conference on Artificial Intelligence. pp. 1–4 (2016)
11. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Jain, N., Bansal, V., Virmani, D., Gupta, V., Salas-Morera, L., Garcia-Hernandez, L.: An enhanced deep convolutional neural network for classifying indian classical dance forms. Applied Sciences **11**(14), 6253 (2021)
14. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35**(1), 221–231 (2012)
15. Kapsouras, I., Karanikolos, S., Nikolaidis, N., Tefas, A.: Feature comparison and feature fusion for traditional dances recognition. In: International Conference on Engineering Applications of Neural Networks. pp. 172–181. Springer (2013)
16. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)

17. Kishore, P., Kumar, K., Kiran Kumar, E., Sastry, A., Teja Kiran, M., Anil Kumar, D., Prasad, M.: Indian classical dance action identification and classification with convolutional neural networks. Advances in Multimedia **2018** (2018)
18. Kreiss, S., Bertoni, L., Alahi, A.: Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. IEEE Transactions on Intelligent Transportation Systems (2021)
19. Li, L.: Dance art scene classification based on convolutional neural networks. Scientific Programming **2022** (2022)
20. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE International Conference on Computer Vision. pp. 5533–5541 (2017)
21. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems **27** (2014)
22. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
23. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. ArXiv **abs/2203.12602** (2022)
24. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
25. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision **103**(1), 60–79 (2013)
26. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2740–2755 (2018)
27. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020)
28. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)