Multimodal Dance Recognition

Monika Wysoczańska[1,3] and Tomasz Trzciński[1,2]
[1] *Warsaw University of Technology, Poland*
[2] *Tooploox, Poland*
[3] *Sport Algorithmics and Gaming, Poland*

Keywords: Multimodal Learning, Activity Recognition, Music Genre Classification, Multimodal Fusion.

Abstract: Video content analysis is still an emerging technology, and the majority of work in this area extends from the still image domain. Dance videos are especially difficult to analyse and recognise as the performed human actions are highly dynamic. In this work, we introduce a multimodal approach for dance video recognition. Our proposed method combines visual and audio information, by fusing their representations, to improve classification accuracy. For the visual part, we focus on motion representation, as it is the key factor in distinguishing dance styles. For audio representation, we put the emphasis on capturing long-term dependencies, such as tempo, which is a crucial dance discriminator. Finally, we fuse two distinct modalities using a late fusion approach. We compare our model with corresponding unimodal approaches, by giving exhaustive evaluation on the *Let's Dance* dataset. Our method yields significantly better results than each single-modality approach. Results presented in this work not only demonstrate the strength of integrating complementary sources of information in the recognition task, but also indicate the potential of applying multimodal approaches within specific research areas.

1 INTRODUCTION

Dance being one of the factors in shaping cultural identities around the world has its many different forms (Sachs and Schönberg, 1963). It emerges from humans' innate reaction to hearing rhythmical music (Madison et al., 2011). Therefore, dancing creation process is highly multimodal. This observation inspired research community to analyse how humans match body movements to music (Ofli et al., 2012), (Kleiman and Cohen-Or, 2018). In (Lee et al., 2019) authors use a generative approach for music-to-dance synthesis and developed a method which generates realistic and style-consistent dances given a piece of music. In this work, we also take the leverage of the above observations and prove that multimodal approach can be highly beneficial for the dance recognition task.

The diversity of existing dance styles impacts the level of difficulty of our task. For instance, it takes a lot of domain knowledge to properly discriminate between different latin dances. Dancers perform in similar conditions, mostly competition venues, and wear the same costumes when performing all of the them. The same applies to standard ballroom dances, where a scene context is not as important as exact dancers' movements. Therefore, effective motion representation seems to be crucial for dance recognition task. Figure 1 shows exemplary frames taken from videos of different ballroom dances. Although each example represents different dance style, all of them present highly overlapping scenes.

Musically-wise the problem of dance classification is also non-trivial. For the above mentioned ballroom dances there are actually well-defined international standards which describe important audio characteristics, such as tempo. Nevertheless, when it comes to other dance genres, there are no such requirements or constraints. In fact, dance as a constantly evolving art form has no limits when it comes to music interpretation.

In this work we explicitly tackle dance videos classification problem as we believe current methods, based on visual information, are insufficient for this complex task. We define our problem as a multimodal learning challenge of integrating information coming from distinct signals. In case of our problem, the considered modalities are vision and sound. We propose an audio-visual solution consisting of two modules. For the visual part, we propose an improved architec-

Wysoczańska, M. and Trzciński, T.
Multimodal Dance Recognition.
DOI: 10.5220/0009326005580565
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 558-565
ISBN: 978-989-758-402-2; ISSN: 2184-4321

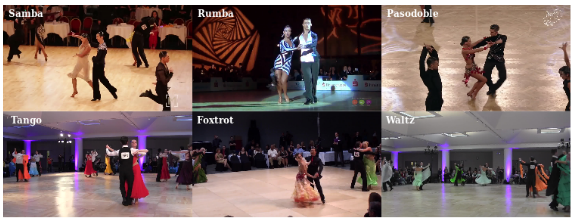

Figure 1: Exemplary images illustrating the extent of visual overlap between different ballroom dances.

ture of the state-of-the-art model for activity recognition task, that captures temporal dependencies more efficiently. Regarding audio representation, we adapt the state-of-the-art method for music genre recognition and implement it within our pipeline. Finally, we examine fusion methods for our both unimodal representations and propose a late fusion approach.

2 RELATED WORK

Activity Recognition. In computer vision the problem of video classification with a focus on dance domain falls into human activity recognition, where the most important problem has always been how to efficiently extract and represent humans' movement. Some of the earlier methodologies focused on assigning a specific part of a video clip to predefined gestures, similar to words in dictionary in natural language processing techniques. In (Samanta et al., 2012) a task is to recognise one of the three Indian classical dances, namely: *Bharatnatyam*, *Kathak* and *Odissi*. Each video frame is represented by a motion encoder, and a sequence of consecutive frames creates one sample called a 'visual word' - movement.

In (Anbarsanti and Prihatmanto, 2014) the aim is to properly assign given gesture pattern from *Likok Pulo* dance. This is the example of a trajectory-based approach with the use of depth information for more accurate human body representation, which in the case of subtle gestures from this traditional Indonesian aceh dance is crucial.

Authors of (Castro et al., 2018) address the dance classification problem based on visual information using deep learning methods. They adapt the concept of two-stream networks for action recognition, introduced in (Simonyan and Zisserman, 2014), and give the exhaustive evaluation of its different variations. The proposed model, namely Three-stream Temporal CNN, is the extension of the idea of concurrent RGB frames and optical flow processing in a stream manner to incorporate motion information. The additional stream analyses the exact humans' poses based on visualised skeletons, which are previously computed using some pose estimation technique. Additionally, authors introduce a first publicly available dataset specifically targeted for dance video classification task, named *Let's dance*.

Music Genre Recognition. In audio analysis the problem of audio classification for dance videos falls into music genre recognition. There are many benchmark datasets dedicated for this task, such as *GTZAN* and *Extended Ballroom* (Marchand and Peeters, 2016), with the latter actually focusing on the dance domain. It consists of over thousand, 30-second long pieces of music that meet international standards' criteria for 13 ballroom dances.

In this research area, similarly to computer vision, deep neural networks dominated currently explored methods. Unlike the time-consuming and oftentimes ineffective manual feature engineering, they allow for the automatic feature learning. The efficient algorithms for audio classification analyse signals in both, time and frequency domain, thus today's music genre recognition research focuses on the design of robust deep architectures, capable of capturing complex sequential dependencies in music. Currently the most effective methods base on convolutional neural networks applied on visual audio representations, usually spectrograms in mel-frequency domain. In (Hershey et al., 2016) authors experiment with popular image classification architectures, and prove that this approach achieves state-of-the-art performance.

In (Liu et al., 2019) authors discuss shortcomings of using traditional CNN architectures designed for image processing tasks. They propose a novel architecture, namely Bottom-up Broadcast Neural Network (BBNN) which fully exploits and preserves the low-level information on the higher layers of deep network. BBNN achieves state-of-the-art performance on *GTZAN* as well as *Extended Ballroom* datasets.

3 PROPOSED METHOD

We propose a mutlimodal pipeline that leverages both sources of information, vision and audio, as we believe they are both equally important in recognising dance genres. The pipeline consists of two modules, for visual and audio representations respectively, as shown in Figure 2. Given a video clip, visual and acoustic signals are first processed separately. In the visual module, we focused on the effective motion representation, while still preserving contextual cues. Based on the concept of Three-Stream Temporal CNN, we propose its modified architecture and implement it within our pipeline. For audio representation we adapt Bottom-up Broadcast Neural Network as it achieves the highest classification accuracy on the benchmark datasets for music genre recognition. Both modalities are first trained separately. We then infer corresponding models to obtain the respective representations and combine them to give a final decision for a whole video clip.

Motion Representation. We adapt the idea of Three-Stream Temporal CNN and propose our modified architecture. Similarly to the original model, our method uses visualised optical flow, as well as visualised skeletal data to incorporate motion information. We preserve contextual information by using extracted RGB frames in a third processing stream of the model. Each stream is processed separately through a couple of convolutional layers. Instead of originally proposed AlexNet for a single-stream extractor, we use the state-of-the-art architecture for image classification - Inception v3. proposed in (Szegedy et al., 2015). Each stream's outputs are then concatenated and passed through the additional fully-connected layer to train a joint representation. At the end, the softmax classifier is applied giving the final decision for each input as shown in Figure 3.

To obtain a final decision for a given video clip we follow the Three-Stream Temporal original paper practices. We implement the averaging method based on the final softmax outputs. For each video fragment a visual input chunk consisting of RGB, optical flow and skeleton representations is fed into Three-Stream Temporal Inception core CNN model. We average the consecutive output vectors to finally obtain one motion representation vector for a whole video fragment.

Audio Representation. Regarding audio representation, our main indicator was the performance on the *Extended Ballroom* dataset, which perfectly corresponds to our problem. We use BBNN architecture applied on mel-spectrograms of 128 mel-bands, as proposed originally in the paper. We process one audio input for a whole video clip, as our experiments showed that shorter audio fragments, precisely 0.5-second, give much worse results. We concluded that 0.5-second input is simply too short to capture all the necessary audio characteristics, especially rhythm.

Late Fusion. Having computed the vector representations of both modalities, corresponding to final unimodal decisions per video fragment, the two following methods for the late fusion are considered:

1. softmax scores averaging across modalities,
2. SVM classifier trained on the concatenation of both modalities' vector representations.

Table 1: Hyper-parameter tuning configurations for SVM randomised grid search.

kernel type	C	gamma
RBF	[0.01,10]	[0.001, 10]
Linear	[0.01,10]	-
Polynomial	[0.01,10]	[0.001, 10]

To perform the SVM-based fusion method, we obtain last-layer vector representations of each modality. Such a multimodal vector forms an input to SVM classifier. We then use *Randomised Grid Search* method, implemented in Scikit-Learn (Pedregosa et al., 2011) for hyper-parameter tuning. In our experiments we run 10-fold cross validation for 500 iterations. The parameters grid is presented in Table 1.

With our design choices, the proposed pipeline comes with its advantages but also with some shortcomings. The late fusion approach results in high flexibility of the proposed solution. With the separate unimodal building blocks, we can conduct an in-depth analysis of the respective representations. This allows us to identify potential pipeline's weak points and easily interpret obtained results.

On the other hand, the flexibility of our solution comes at a price of higher complexity. As opposed to an end-to-end architecture, our pipeline is harder to train, since each of the components needs to be trained separately. This also affects the inference process, meaning additional steps are required in order to automatically run our pipeline.

4 EXPERIMENTAL RESULTS

Evaluation Dataset. We evaluate our method on the *Let's dance dataset*. The dataset consists of 10-second videos taken from Youtube at a quality of 720p. It includes both, dancing performances and plain-clothes practising. In this work we use the extended version of the original dataset consisting of video recordings presenting 16 different dances.

According to international standards, some of those 16 classes have the same origin, thus are usually grouped together. Figure 4 depicts the taxonomy of dances included in *Let's dance* dataset. We denote the originally named *latin* class as *salsa*, since it refers to the general family of dance styles, thus not properly assigned. The video examples indeed show different variations of *salsa* dance, such as *bachata*.

The examples of the *Let's dance* dataset are provided in three different representations, respectively

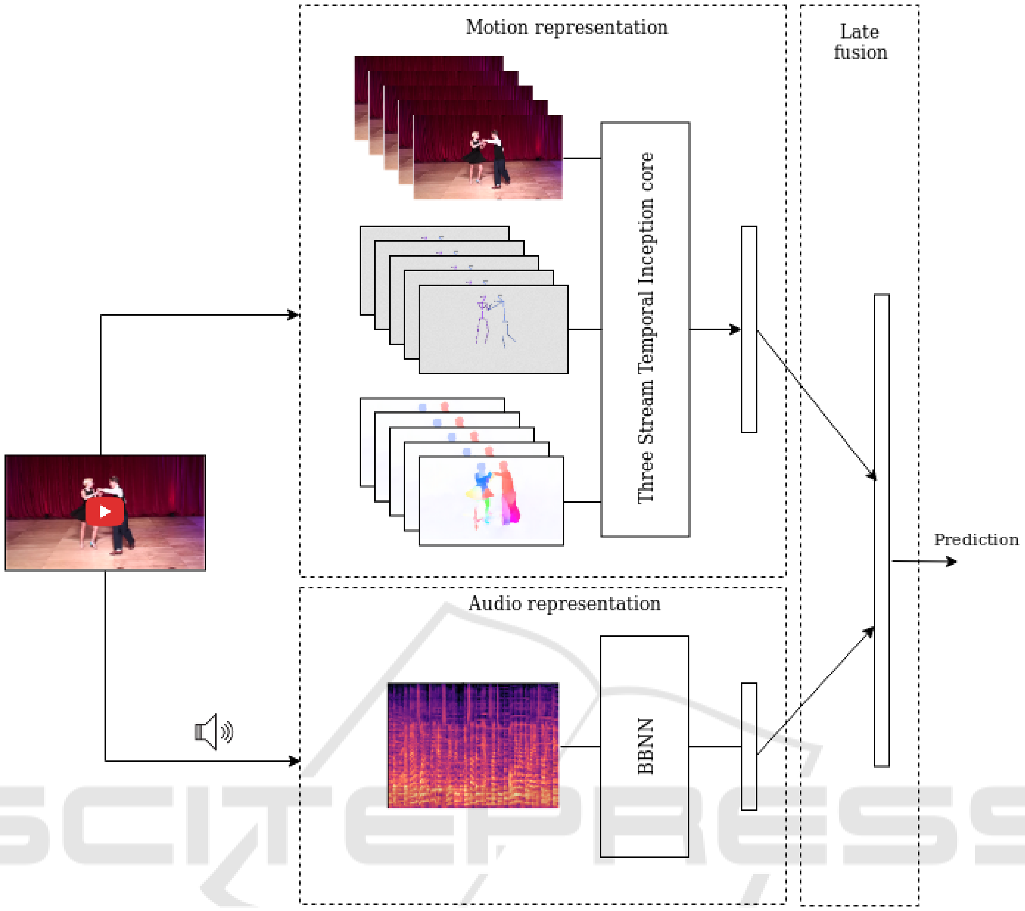

Figure 2: High-level architecture of our proposed method for dance videos classification. We use a late fusion approach, where visual and audio signals are first processed separately by respective deep architectures. We then combine those representations to give a final decision for a given video clip.

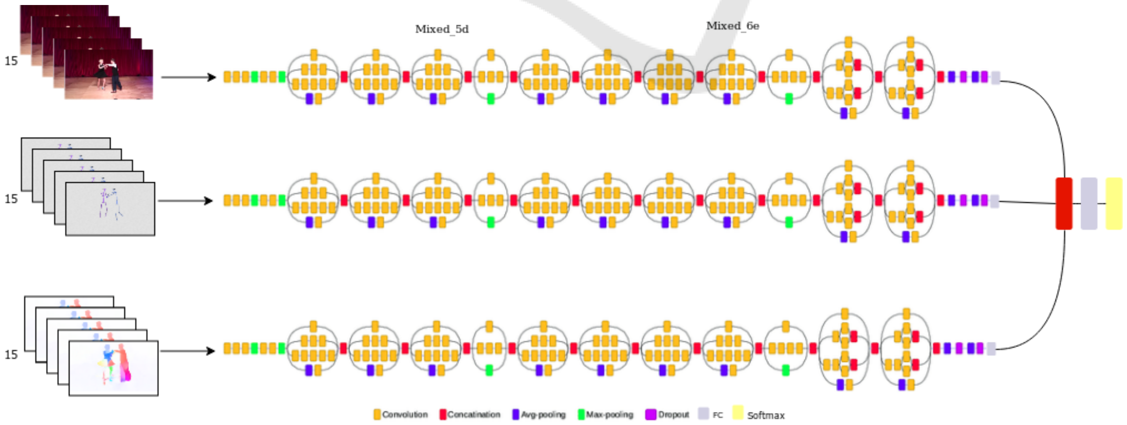

Figure 3: The high level architecture diagram of Three-stream Temporal Inception core CNN.

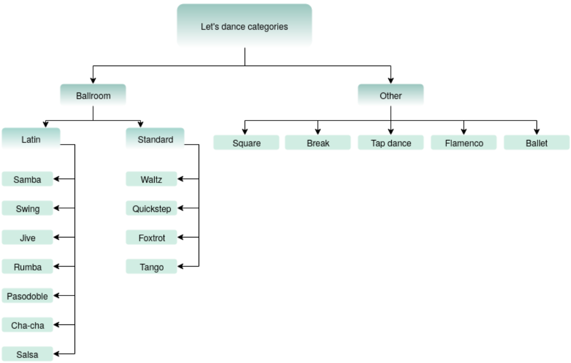

Figure 4: *Let's dance* dataset classes taxonomy. This is our own elaboration based on international standards.

to the corresponding model presented in the original work:

- extracted RGB frames,
- rendered optical flow visualisation images - obtained using FlowNet2.0. presented in (Ilg et al., 2016)
- rendered images of visualised skeletons - using Densepose presented in (Güler et al., 2018)

The examples of each visual input representation are shown in Figure 5.

In order to built the audio set we used *youtube-dl* command line tool. Each audio sample was trimmed to match the exact video fragment from the visual set. Some of the videos have already been taken down from the platform, so the final dataset used in our experiments counts 1381 video clips in total.

After the initial pre-processing we generated mel-spectrograms for each audio sample. Since the duration of different video samples in the dataset varies between 8 to 10 seconds, we fixed the length of each audio input to 8 seconds for simplicity. The mel-spectrograms of 128 mel-bands were extracted using Librosa (Brian McFee et al., 2015). The frame length was set to 2048 and the hop size to 1024, to finally obtain an image of size 176x128.

Experimental Setup. For all of the experiments described in the following sections, we randomly split the dataset into 3 subsets, taking into account its slight imbalance:

- Training set - 80% of the dataset, used for the networks training as well as SVM fusion classifier.
- Test set - 10% of the dataset, used for the evaluation at the networks training time.
- Validation set - 10% of the dataset, used only as a final evaluation set.

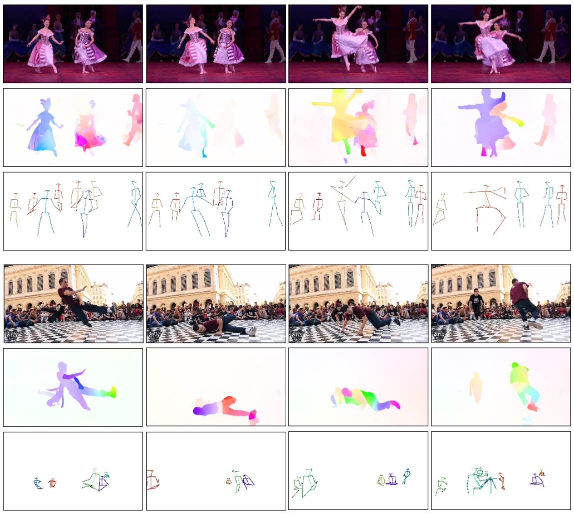

Figure 5: Exemplary visual representations of *Let's dance* dataset videos.

Evaluation Metrics. We conduct the qualitative evaluation using following metrics:

- accuracy score,
- precision score (macro-averaged),
- recall score (macro-averaged),
- f1-score (calculated and analysed per class).

Evaluation Results. We implemented our models, Three-Stream Temporal Inception core CNN as well as BBNN in PyTorch (Paszke et al., 2017). Both networks are evaluated on the test set only, which is representative enough for this task.

We initialise Three-Stream Temporal Inception core CNN with the weights pretrained on the ImageNet dataset (Deng et al., 2009), following the original paper implementation. Our experiments showed that the model achieves the best results when only fine-tuned from *Mixed 5d layer*. We tested other transfer learning scenarios such as: whole network fine-tuning, and alternatively, freezing layers up to *Mixed 6e layer*, but obtained significantly worse results. We train the network with Stochastic Gradient Descent optimiser with momentum. We input 15-consecutive frames in a stack manner and train the network with the batch size of 32.

Regarding BBNN, we train the network using ADAM optimiser (Kingma and Ba, 2014), starting with the learning rate equal to 0.005. Finally, we remove softmax layers at the end of each network and infer the models to obtain final vector representations. For the SVM fusion method we concatenate both vectors and train the classifier on the training set.

Table 2: Qualitative comparison of the unimodal and multimodal approaches on the final validation set of the *Let's dance* dataset.

Modalities	Fusion method	Accuracy	Precision	Recall
Vision	-	58.3 %	63.5%	57.0%
Audio	-	69.1%	71.3%	70.1%
V+A	Average	74.1%	75.4%	74.5%
V+A	**SVM (Linear, C=6.0)**	**77.0%**	**78.0%**	**78.6%**

Table 3: The per-class f1-scores for our best model on the final validation set of the *Let's dance* dataset.

dance	F1-score
ballet	0.84
break	0.96
cha	0.75
flamenco	0.53
foxtrot	0.86
jive	0.8
salsa	0.86
pasodoble	0.74
quickstep	0.6
rumba	0.86
samba	0.71
square	0.91
swing	0.67
tango	0.8
tap	0.57
waltz	0.89

The SVM-based multimodal late fusion approach yields the best results, achieving 77% accuracy on the final validation set. It slightly improves the results obtained with simple averaging method, and outperforms both single-modality approaches significantly. Visual-based model achieves only 58.3%, whereas audio-based approach scored 69.1% accuracy. The results of per-class F1-scores, presented in Table 3 indicate that our method achieves the worst overall performance on *flamenco* and *tap* classes. In fact, from the musical standpoint, these genres may be considered very similar. Both dances involve rhythmical tapping and clapping, oftentimes without any instrumental background. Visually, they can also be difficult to discriminate, as those dances base on small hands' and feet movements and are usually performed by soloists. On the other hand, we found the best performance of our final model for *break* class. This may indeed be considered a very different dance form, originating from the family of street styles. Visually, it involves a lot of tricks and floor work and musically-wise *break* is usually performed to hip-hop music.

To get a better understanding of the results of our proposed approach, we visualised multimodal vectors on 2D surface using Scikit-learn implementation of t-SNE method (van der Maaten and Hinton, 2008).

Considering the taxonomy presented in Figure 4, we observe similar grouping with our multimodal representations, as can be seen in Figure6. Standard ballroom dances (blue ones) are clearly separated from latin dances (red ones). They are also quite well-separated within their groups. In terms of other dance genres, especially *tap*, *flamenco* and *ballet* there is no such clear separation, as already indicated by the F1-scores analysis. Moreover, *break* appropriately lies far from other dances, especially *ballet* and standard ballroom dances, being indeed a completely different genre, taking into account both modalities.

We should also mention the fact that YouTube is a common video-sharing platform, open to everyone. The examples from the dataset used in our experiments include amateur videos recorded in very different circumstances and environments. It influences the visual as well as sound quality, resulting in the noisy data our models struggled with.

5 CONCLUSIONS

In this work we propose a novel approach for dance videos classification. We define dance classification problem as a multimodal learning task and solve it with an audio-visual approach. Inspired by human perception, our solution demonstrates the strength of combining complementary signals for the recognition of complex actions, such as different dance styles.

Our method uses deep neural network based approaches for the unimodal representations. Although the models achieve state-of-the art performance in their particular fields, they are insufficient for the demanding task of dance videos classification separately. We proposed two multimodal late fusion methods, simple averaging and a model-based approach using SVM classifier. Both of the proposed fusions demonstrate a significant classification accuracy improvement, with the latter achieving 77% accuracy on the *Let's dance* dataset being our best proposed method.

In the future work, we plan to evaluate our approach on different datasets, such as a subset of 'dancing' classes from the *Kinetics 700* (Carreira et al., 2019). To the best of our knowledge, there are no

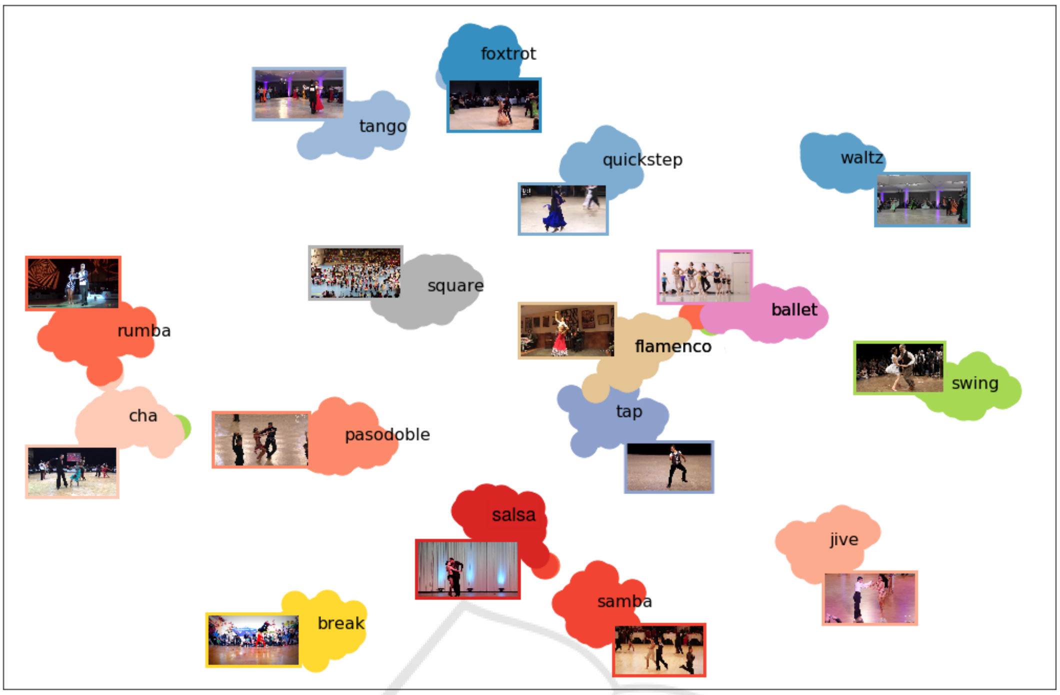

Figure 6: Multimodal feature vectors' visualisation on the Let's dance dataset. We used t-SNE method with perplexity=25 to visualise high-dimensional vectors on the 2D surface. The results correspond with the taxonomy presented in Figure 4.

other multimodal methods specifically targeted for the dance recognition task. Therefore, we would like to compare our late fusion approach with existing audio-visual methods, such as an early-fusion approach presented in (Owens and Efros, 2018). Finally, with the extended version of the *Let's dance* we would like to explore the remaining multimodal challenges within this area, such as co-learning, and check to what extent the considered modalities could provide supplementary information.

ACKNOWLEDGEMENTS

This research was partially supported by the Polish National Science Centre grant no. UMO-2016/21/D/ST6/01946.

REFERENCES

Anbarsanti, N. and Prihatmanto, A. S. (2014). Dance modelling, learning and recognition system of aceh traditional dance based on hidden markov model. In *2014 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 86–92.

Bahuleyan, H. (2018). Music genre classification using machine learning techniques. *CoRR*, abs/1804.01149.

Baltrusaitis, T., Ahuja, C., and Morency, L. (2017). Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto (2015). librosa: Audio and Music Signal Analysis in Python. In Kathryn Huff and James Bergstra, editors, *Proceedings of the 14th Python in Science Conference*, pages 18 – 24.

Carreira, J., Noland, E., Hillier, C., and Zisserman, A. (2019). A short note on the kinetics-700 human action dataset. *CoRR*, abs/1907.06987.

Castro, D., Hickson, S., Sangkloy, P., Mittal, B., Dai, S., Hays, J., and Essa, I. A. (2018). Let's dance: Learning from online dance videos. *CoRR*, abs/1801.07388.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2014). Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, LNCS 2749, pages 363–370, Gothenburg, Sweden.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. *CoRR*, abs/1604.06573.

Feng, L., Liu, S., and Yao, J. (2017). Music genre classification with paralleling recurrent convolutional neural network. *CoRR*, abs/1712.08370.

Güler, R. A., Neverova, N., and Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. W. (2016). CNN architectures for large-scale audio classification. *CoRR*, abs/1609.09430.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2016). Flownet 2.0: Evolution of optical flow estimation with deep networks. *CoRR*, abs/1612.01925.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kleiman, Y. and Cohen-Or, D. (2018). Dance to the beat : Enhancing dancing performance in video.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA. Curran Associates Inc.

Lee, H.-Y., Yang, X., Liu, M.-Y., Wang, T.-C., Lu, Y.-D., Yang, M.-H., and Kautz, J. (2019). Dancing to music. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 3581–3591. Curran Associates, Inc.

Liu, C., Feng, L., Liu, G., Wang, H., and Liu, S. (2019). Bottom-up broadcast neural network for music genre classification. *CoRR*, abs/1901.08928.

Madison, G., Gouyon, F., Ullén, F., and Hörnström, K. (2011). Modeling the tendency for music to induce movement in humans: first correlations with low-level audio descriptors across music genres. *Journal of experimental psychology. Human perception and performance*, 37 5:1578–94.

Marchand, U. and Peeters, G. (2016). The Extended Ballroom Dataset. Late-Breaking Demo Session of the 17th International Society for Music Information Retrieval Conf.. 2016.

Moon, S., Kim, S., and Wang, H. (2014). Multimodal transfer deep learning for audio visual recognition. *CoRR*, abs/1412.3121.

Ofli, F., Erzin, E., Yemez, Y., and Tekalp, A. (2012). Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE TRANSACTIONS ON MULTIMEDIA*, 14:747–759.

Owens, A. and Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *The European Conference on Computer Vision (ECCV)*.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS-W*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. (2014). Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575.

Sachs, C. and Schönberg, B. (1963). *World History of the Dance*. The Norton Library. Allen & Unwin.

Samanta, S., Purkait, P., and Chanda, B. (2012). Indian classical dance classification by learning dance pose bases. In *Proceedings of the 2012 IEEE Workshop on the Applications of Computer Vision*, WACV '12, pages 265–270, Washington, DC, USA. IEEE Computer Society.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc.

Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. *CoRR*, abs/1409.4842.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition*.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Wang, L., Xiong, Y., Wang, Z., and Qiao, Y. (2015). Towards good practices for very deep two-stream convnets. *CoRR*, abs/1507.02159.

Zhang, H.-B., Zhang, Y.-X., Zhong, B., Lei, Q., Yang, L., Du, J.-X., and Chen, D.-S. (2019). A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19:1005.