Transfer Learning for Classifying Single Hand Gestures on Comprehensive Bharatanatyam Mudra Dataset

Anuja P. Parameshwaran[1*], Heta P. Desai[1*], Rajshekhar Sunderraman[1], Michael Weeks[1]
[1]Georgia State University
{aparameshwaran1, hdesai1}@student.gsu.edu {raj, mweeks}@cs.gsu.edu

Abstract

For any dance form, either classical or folk, visual expressions - facial expressions and hand gestures play a key role in conveying the storyline of the accompanied music to the audience. Bharatanatyam – a classical dance form which has origins from the southern states of India, is on the verge of being completely automated partly due to an acute dearth of qualified and dedicated teachers/gurus. In an honest effort to speed up this automation process and at the same time preserve the cultural heritage, we have chosen to identify and classify the single hand gestures/mudras/hastas against their true labels by using two variations of the convolutional neural networks (CNNs) that demonstrates the exceeding effectiveness of transfer learning irrespective of the domain difference between the pre-training and the training dataset. This work is primarily aimed at 1) building a novel dataset of 2D single hand gestures belonging to 27 classes that were collected from Google search engine (Google images), YouTube videos (dynamic and with background considered) and professional artists under staged environment constraints (plain backgrounds), 2) exploring the effectiveness of Convolutional Neural Networks in identifying and classifying the single hand gestures by optimizing the hyperparameters, and 3) evaluating the impacts of transfer learning and double transfer learning, which is a novel concept explored in this paper for achieving higher classification accuracy.

1. Introduction

In recent years, there has been a lot of ongoing research being carried out in the field of sign language recognition [1-9] and general hand gesture recognition [10-14] using traditional image processing and machine learning techniques. Conventional image processing techniques mostly depend on using moments, shape descriptors as hand crafted features along with uses of classifiers like support vector machine or artificial neural network for classification tasks. Most of these works have not transcended into the domain of deep learning yet, nor have adequately explored the extent of effectiveness of a simple convolutional neural network (CNN) for solving complex multi-class classification problems.

As part of our ongoing research, multi-classification of single hand gestures of the Bharatanatyam dance form was achieved using the CNN architectures. In all, this dance form has a total of fifty-two hand gestures known as *hastas* or *mudras*. Among these, twenty-eight are single hand gestures or *asamyukta hastas* and the remaining twenty-four, double hand gestures or *samyukta hastas*. This paper focuses on 1) using preprocessing techniques to build a novel dataset 27 classes of *asamyukta hastas* collated from search engine, YouTube videos and actors in staged setting.

The works cited in [10-14] deal with: 1) traditional image processing techniques-extracting and classifying hand crafted features, 2) only a subset of the single hand mudras (out of 28), except [14], 3) very small datasets mostly collected under controlled environment settings without dynamic backgrounds rendering these methods ineffective when used in a real-scenario setting like when the shots of the performance are taken outdoors against backgrounds of nature etc., particularly true for [14] and lastly, 4) inadequate focus on deep architectures.

2. Data Acquisition, Cleansing, Pre-processing and Augmentation

The dataset is a novel dataset of single hand gestures belonging to the selected 27 categories (reason behind the 27 is because one hand gesture cannot be described by a single frame). The 2D images of every *mudra* were acquired by three different sources as mentioned earlier. The acquired data required a good amount of cleansing mainly to either remove data not related to the dataset or properly classify mislabeled data as well as data with noisy label to improve the quality of the training data. The above steps increase the classification accuracy of the model by improving the overall quality of training instances collected and at the same time minimizing the manual time required for weeding out the misfits. Towards this end, we attempted automation of the data cleansing process by treating it as an image classification problem, solving the same using a convolutional autoencoder architecture for image

[*] Denotes equal contribution.

classification. We decided on using the convolutional autoencoder instead of a CNN primarily because of limited data. The training dataset in this case comprised of images collected only from actors and the YouTube videos. The images collected from the google search engine are web labeled images rather than human labeled images. They are used because they are one of the fastest mediums to gather and construct big datasets but not necessarily accurate as the web search itself need not be accurate and there is bound to be label noise. Having human verifiers to verify the web labeled images is time consuming and expensive, hence we used a convolutional autoencoder to achieve the same.

The next step i.e. data prepping and preprocessing, involves using techniques like median filter smoothing and region of interest (ROI) extraction operations. For using a deep network or a wide network or a deep and wide network like a ResNet, it is essential to perform data augmentation - increase the data in the dataset to achieve better classification accuracy that fully utilizes the depth of the network model. We perform various data augmentation techniques using both the TensorFlow data augmentation commands as well as the Keras Image generator. After augmentation the dataset has a total of 18,992 images.

3. Architecture, Transfer Learning and Double Transfer Learning

This work uses the transfer learning concept in CNNs to achieve higher classification accuracy. The hyperparameters such as number of epochs, optimizer and batch size were optimized using grid search. Model1 is an ImageNet pre-trained VGG16 architecture [15] which has 2 additional dense layers along with the softmax layer for our 27 classes. Model1 has an overall classification accuracy of 94.56% for 70 epochs. The pre-training dataset is the ImageNet dataset [16] and the training dataset is our comprehensive dataset.

Model2 is an ImageNet pre-trained model with two additional dense layers, a Batch Normalization layer, a Dropout layer (50% dropout) and a softmax layer for 10 classes, pre-trained again using the dataset in [17], and the model saved to disk. This dataset essentially consists of a set of near infrared images acquired by the Leap Motion sensor. The above pre-training dataset has 10 classes, collected using 10 different subjects (5 men and 5 women). The saved model is then trained on our dataset by substituting the existing softmax layer with a new softmax layer for the 27 classes. Our findings show that double transfer learning utilizing two different domain datasets (with respect to object of interest and categories/classes) for pre-training the model is highly effective, yielding a high accuracy of 98.25% in just 20 epochs. Model1 and Model2 used the SGD optimizer with initial learning rate of 0.0001 with step decay after every 5 epochs.

4. Results

The overall classification accuracy achieved with Model1 is 94.56%, whereas with Model2 it is 98.25%. Model2 achieves a higher classification accuracy than existing work [9-14] in this domain (hand gestures) while making use of a more challenging dataset.

Ref. No	Classification accuracy
[10]	95.25%
[11]	80% (SIFT, SURF, LBP, Haar features + SVM)
[11]	90% (HOG features + SVM)
[12]	85.10%
[13]	85.29%
[14]	94.71% (shallow CNN)
[14]	97.1%, 98% and 96.8% (Hu, eigen vector and intersection as features + ANN)
Our Model1	94.56% (Single transfer learning)
Our Model2	98.25% (Double transfer learning)

Table 1: Comparison of model performances with other published works

5. Conclusion

Convolutional neural networks, due to its inherent ability to automatically learn features without any human supervision, are very popular deep learning architectures used in different domains including Computer Vision, Natural Language Processing etc. Transfer learning is an effective tool, especially when the dataset size is limited. A higher classification accuracy is achieved when the domains of both pre-training and training datasets are similar. We were able to achieve a high classification accuracy of 98.25% using the above Double transfer learning techniques. In short, this method is a giant step towards facile implementation of e-learning techniques for the Indian classical dance, Bharatanatyam.

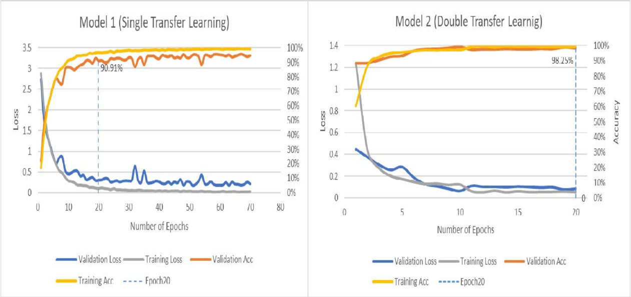

Figure 1: Single Transfer Learning and Double Transfer Learning model training loss/accuracy and validation loss accuracy plots. Model 1 achieved 90.91% while model 2 achieved 98.25% validation accuracies at epoch 20.

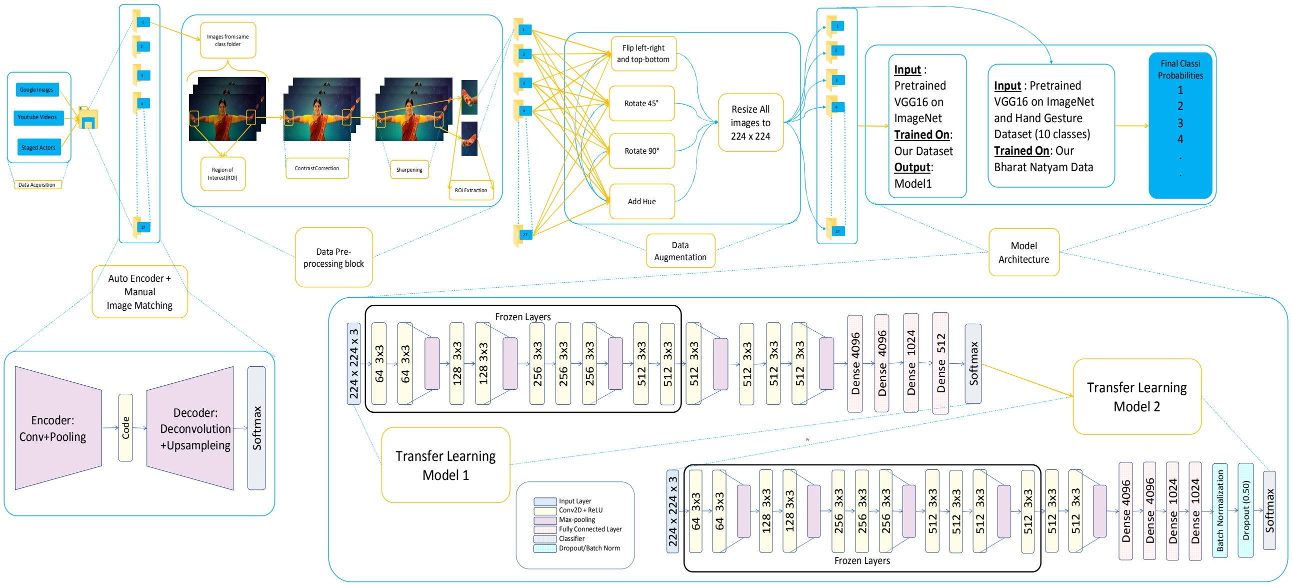

Figure 2: Overall procedures of data-transformation and model architecture

References

[1] Solís, F., Martínez, D. and Espinoza, O., 2016. Automatic mexican sign language recognition using normalized moments and artificial neural networks. *Engineering*, *8*(10), pp.733-740.

[2] Zadghorban, M. and Nahvi, M., 2018. An algorithm on sign words extraction and recognition of continuous Persian sign language based on motion and shape features of hands. *Pattern Analysis and Applications*, pp.1-13.

[3] Fagiani, M., Principi, E., Squartini, S. and Piazza, F., 2015. Signer independent isolated Italian sign recognition based on hidden Markov models. *Pattern Analysis and Applications*, *18*(2), pp.385-402.

[4] Fernando, M. and Wijjayanayake, J., 2015. Novel approach to use HU moments with image processing techniques for real time sign language communication. *Int. J. Image Process*, *9*, pp.335-345.

[5] Dixit, K. and Jalal, A.S., 2013, February. Automatic Indian sign language recognition system. In *2013 3rd IEEE International Advance Computing Conference (IACC)* (pp. 883-887). IEEE.

[6] Premaratne, P., Yang, S., Zou, Z. and Vial, P., 2013, July. Australian sign language recognition using moment invariants. In *International Conference on Intelligent Computing* (pp. 509-514). Springer, Berlin, Heidelberg.

[7] Pradhan, A., Kumar, S., Dhakal, D. and Pradhan, B., 2016. Implementation of PCA for recognition of hand gesture representing alphabets. *International Journal*, *6*(3).

[8] Adithya, V., Vinod, P.R. and Gopalakrishnan, U., 2013, April. Artificial neural network-based method for Indian sign language recognition. In *2013 IEEE Conference on Information & Communication Technologies* (pp. 1080-1085). IEEE.

[9] Singha, J. and Das, K., 2013. Indian sign language recognition using eigen value weighted Euclidean distance-based classification technique. *arXiv preprint arXiv:1303.0634*.

[10] Anami, B.S. and Bhandage, V.A., 2018. A vertical-horizontal-intersections feature based method for identification of bharatanatyam double hand mudra images. *Multimedia Tools and Applications*, *77*(23), pp.31021-31040.FirstName Alpher,, FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? Journal of Foo, 14(1):234–778, 2004.

[11] Kumar, K.V.V. and Kishore, P.V.V., 2017. Indian Classical Dance Mudra Classification Using HOG Features and SVM Classifier. *International Journal of Electrical & Computer Engineering (2088-8708)*, *7*(5). Authors. The frobnicatable foo filter, 2014. Face and Gesture 2014 submission ID 324. Supplied as additional material efg324.pdf.

[12] Saha, S., Ghosh, L., Konar, A. and Janarthanan, R., 2013, September. Fuzzy L membership function-based hand gesture recognition for Bharatanatyam dance. In *2013 5th International Conference and Computational Intelligence and Communication Networks* (pp. 331-335). IEEE.

[13] Hariharan, D., Acharya, T. and Mitra, S., 2011, June. Recognizing hand gestures of a dancer. In *International conference on Pattern Recognition and Machine Intelligence* (pp. 186-192). Springer, Berlin, Heidelberg.

[14] Anami, B.S. and Bhandage, V.A., 2018. A Comparative Study of Suitability of Certain Features in Classification of Bharatanatyam Mudra Images Using Artificial Neural Network. *Neural Processing Letters*, pp.1-29.

[15] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. CVPR*, 2009, pp. 248–255.

[17] T. Mantecón, C.R. del Blanco, F. Jaureguizar, N. García, “Hand Gesture Recognition using Infrared Imagery Provided by Leap Motion Controller”, Int. Conf. on Advanced Concepts for Intelligent Vision Systems, ACIVS 2016, Lecce, Italy, pp. 47-57, 24-27 Oct. 2016. (doi: 10.1007/978-3-319-48680-2_5)