Rhythmic Activity Detection in Footwork in Kathak Dance using Machine Learning

Project-I (MA57011) report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Master of Science

in

Mathematics and Computing

by

Akhil Raj

(15MA20057)

Under the supervision of

Professor Debjani Chakraborty, Department of Mathematics
and
Professor Partha Pratim Das, Department of Computer Science and Engineering

Department of Mathematics

Indian Institute of Technology Kharagpur

Autumn Semester, 2019-20

November 11, 2019

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisors.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: November 11, 2019 (Akhil Raj)

Place: IIT Kharagpur (15MA20057)

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Rhythmic Activity Detection in Footwork in Kathak Dance using Machine Learning**" submitted by **Akhil Raj** (Roll No. 15MA20057) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Master of Science in Mathematics and Computing is a record of bona fide work carried out by him under my supervision and guidance during Autumn Semester, 2019-20.

Professor Debjani Chakraborty
Department of Mathematics

Professor Partha Pratim Das
Department of Computer Science and Engineering

Date: November 11, 2019
Place: IIT Kharagpur

Abstract

Name of the student: **Akhil Raj** Roll No: **15MA20057**

Degree for which submitted: **Master of Science**

Department: **Department of Mathematics**

Thesis title: **Rhythmic Activity Detection in Footwork in Kathak Dance using Machine Learning**

Thesis supervisors: **Professor Debjani Chakraborty and Professor Partha Pratim Das**

Month and year of thesis submission: **November 11, 2019**

This thesis presents a detailed analysis of various approaches used to solve the problem of identifying rhythmic pattern occurring at beat intervals, from a sequence of frames, in which a dancer is performing rhythmic dance steps of Kathak.

We have implemented various techniques obtained from a thorough study on works done in the past to achieve the purpose of the project.

We have implemented video segmentation algorithms to segment the audio-less video dataset that we have, frame by frame. Several algorithms are being used to achieve that. Statistical models, graph cut techniques are employed. In some cases, where the segmentation was not up to the mark, we have made necessary changes in the dataset to rectify the errors.

Acknowledgements

I would like to thank Professor Debjani Chakraborty for being my supervisor in this project. Her invaluable support has immensely helped me in writing of this paper.

I would also like to thanks to Professor Partha Pratim Das for allowing me to work in his project. He provided the much needed guidance and inspiration for the thesis.

Contents

List of Figures

Chapter 1

Introduction

In computer science, digital image processing is the use of computer algorithms to perform image processing on digital images. As a subcategory or field of digital signal processing, digital image processing has many advantages over analog image processing. It allows a much wider range of algorithms to be applied to the input data and can avoid problems such as the build-up of noise and signal distortion during processing. Since images are defined over two dimensions (perhaps more) digital image processing may be modeled in the form of multidimensional systems. A crucial part of image processing is computer vision.

Computer vision tasks include methods for acquiring, processing, analyzing and understanding digital images, and extraction of high-dimensional data from the real world in order to produce numerical or symbolic information, e.g. in the forms of decisions. Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that can interface with other thought processes and elicit appropriate action. This image understanding can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.

My thesis work is heavily based on computer vision. Broadly, from a sequence of frames, performing rhythmic dance steps of Kathak, the objective is to identify rhythmic pattern occurring at beat intervals. From Kathak footwork, our aim is to identify the frame instances within the start and end of each beat. It involves video segmentation.

Chapter 2

Motivation

Kathak is one of the main genres of ancient Indian classical dance and is traditionally regarded to have originated from the travelling bards of North India referred as Kathakars or storytellers. These Kathakars wandered around and communicated legendary stories via music, dance and songs quite like the early Greek theatre. The genre developed during the Bhakti movement, the trend of theistic devotion which evolved in medieval Hinduism. The Kathakars communicate stories through rhythmic foot movements, hand gestures, facial expressions and eye work. This performing art that incorporates legends from ancient mythology and great Indian epics, especially from the life of Lord Krishna became quite popular in the courts of North Indian kingdoms.

Despite of utmost importance of this dance, there has not been much research in applying image processing tools in analysing it. We have tried to explore the possibility by analysis an audio-less video footage of a kathak dance and then using computer vision to analyse it's beat timings.

Chapter 3

Literature Survey

In this part of the thesis, I would like to highlight the past research work that we have consulted in order to get a deep insight on how to solve the problem at hand. Research papers are described(in brief) below:

Monocular video fore-ground/background segmentation by tracking spatial-color gaussian mixture models

[1] Yu et al. presents a new approach to segmenting monocular videos captured by static or hand-held cameras filming large moving non-rigid foreground objects.

Foreground/background segmentation is of great interest in many applications. It can be used in Video Conferencing for replacing the background of the speaker with a better environment, hence preventing the privacy of the speaker. It is used for tracking objects, motion detection, and various other image processing tasks.

The paper presents a new approach to segment foreground/background objects in a video. Instead of a stereo camera pair, this paper's method can be applied to videos captured just by monocular videos, though videos are assumed to be captured by static or hand-held monocular cameras filming large moving non-rigid foreground objects. The method is flexible to some extent since cameras are allowed to shake to somewhat extent and background objects can move too. To solve the segmentation problem, first foreground and background are modelled separately using Spatio-color Gaussian Mixture models(SCGMM), which are built into Random Markov

field energy function. This energy function is minimized by graph cut algorithm which leads to binary segmentation of the video frames.

An energy-based objective function can be formulated over the unknown labeling variables of every pixel, fi , i = 1, . . . , N , in the form of a first-order Markov random field (MRF) energy function:

$$\begin{aligned} E(f) &= E_{\text{data}}(f) + \lambda E_{\text{smooth}}(f) \\ &= \sum_{p \in \mathcal{P}} D_p\left(f_p\right) + \lambda \sum_{\{p,q\} \in \mathcal{N}} V_{p,q}\left(f_p, f_q\right) \end{aligned} \tag{3.1}$$

where N denotes the set of 8-connected pair-wise neighboring pixels, P is the set of pixels in each image. The role of lambda is to balance the data Dp(fp) and smooth cost Vp,q (fp , fq). The above energy function can be efficiently minimized by a two-way graph cut algorithm [3], where the two terminal nodes represent foreground and background labels respectively.

We model the pair-wise smoothness energy term Esmooth (f) as:

$$\begin{aligned} E_{\text{smooth}}(f) &= \sum_{\{p,q\} \in \mathcal{N}} V_{p,q}\left(f_p, f_q\right) \\ &= \sum_{\{p,q\} \in \mathcal{N}} \frac{1}{d(p,q)} e^{-\frac{(t_p - t_q)^2}{2\sigma^2}} \end{aligned} \tag{3.2}$$

where Ip denotes the intensity of pixel p, sigma is the average intensity difference between neighboring pixels in the image, and d(p, q) is the distance between two pixels p and q.

A five dimensional SCGMM model is obtained for each video frame. The likelihood of a pixel belonging to the foreground or background can be written as:

$$p(\mathbf{z}|l) = \sum_{k=1}^{K_l} p_{l,k} G\left(\mathbf{z}; \mu_{l,k}, \mathbf{\Sigma}_{l,k}\right) \tag{3.3}$$

where l fg, bg, representing foreground or background; pl,k is the prior of the kth Gaussian component in the mixture model, and G is the kth Gaussian component as:

$$G\left(\mathbf{z};\mu_{l,k},\Sigma_{l,k}\right)=\frac{1}{(2\pi)^{\frac{d}{2}}\left|\mathbf{\Sigma}_{l,k}\right|^{\frac{1}{2}}}e^{-\frac{\left(\mathbf{z}-\mu_{l,k}\right)^{T}\mathbf{\Sigma}_{l,k}^{-1}\left(\mathbf{z}-\mu_{l,k}\right)}{2}} \tag{3.4}$$

Given the SCGMM models, the data cost Edata(f)defined as:

$$E_{\text{data}}(f)=\sum_{p\in\mathcal{P}}D_{p}\left(f_{p}\right)=\sum_{p\in\mathcal{P}}-\log p\left(\mathbf{z}_{p}|f_{p}\right) \tag{3.5}$$

where p(zp —fp) is computed using Equation 3.

The major problem in this process was that if the objects are moving rapidly across the frames or there are occlusions, then SCGMM models learns from previous frames are not usable for segmenting the current frames. To solve this problem, this paper introduces a 'foreground/background SCGMM joint tracking algorithm', which can propogate SCGMM models over the frames. The drawback of the approach is that in the present of occlusions or rapid motions, segmentation of objects is poor for subsequent some frames, though it gets better with time due to the joint tracking algorithm

Probabilistic Motion Diffusion of Labeling Priors for Coherent Video Segmentation

[2] Wang et al. presents a robust algorithm for temporally coherent video segmentation.

There has been several improvements in Video Segmentation but temporally coherent segmentation remains challenging for real-world video of even moderate complexity. This is because inter-frame motion estimation is often inaccurate. This paper introduces a novel 'motion diffusion model' to produce a probabilistic motion estimate modelling the distribution of motion vectors in each pixel. This information is propagated from previous frames and used in the segmentation in the current frame. Along with temporal coherence, this paper also uses spatial coherence by imposing

labelling consistency obtained through unsupervised segmentation. Several inaccuracies in boundaries, region over-segmentation problems were greatly alleviated with the help of above mentioned coherences.

Fast Approximate Energy Minimization via Graph Cuts

[3] Boykov et al. addresses the problem of minimizing a large class of energy functions that occur in early vision.

The major restriction is that the energy function's smoothness term must only involve pairs of pixels. Authors propose two algorithms that use graph cuts to compute a local minimum even when very large moves are allowed.

Many early vision problems require estimating some spatially varying quantity (such as intensity or disparity) from noisy measurements. Such quantities tend to be piecewise smooth; they vary smoothly at most points, but change dramatically at object boundaries.

These vision problems can be naturally formulated in terms of energy minimization.

$$E(f) = E_{\text{smooth}}(f) + E_{\text{data}}(f) \tag{3.6}$$

Here Esmooth measures the extent to which f is not piecewise smooth, while Edata measures the disagreement between f and the observed data.

The energy functions that we consider in the paper arise in a variety of different contexts, including the Bayesian labeling of MRF's. We allow Dp to be arbitrary, and consider smoothing terms of the form

$$E_{\text{smooth}} = \sum_{\{p,q\} \in \mathcal{N}} V_{\{p,q\}} \left(f_p, f_q\right) \tag{3.7}$$

where N is the set of pairs of adjacent pixels.

V is called a semi-metric on the space of labels L if for any pair of labels alpha, beta L it satisfies two properties: V (alpha, beta) = V (beta, alpha) ¿= 0 and V (alpha, beta) = 0 iff alpha = beta. If V also satisfies the triangle inequality

$$V(\alpha, \beta) \leq V(\alpha, \gamma) + V(\gamma, \beta) \tag{3.8}$$

for any alpha, beta, gamma in L then V is called a metric.

The drawback of this approach is that it only works for metric V p,q 's (i.e., the additional triangle inequality constraint is required)

Exact Maximum A Posteriori Estimation for Binary Images

[4] Greig et al. shows how the image with maximum a posteriori (MAP) probability, the MAP estimate, can be evaluated exactly using efficient variants of the Ford-Fulkerson algorithm for finding the maximum flow in a certain capacitated network.

With xi denoting the category or value of pixel i in the image x = (x;, ... , xn), a Bayesian formulation specifies an a priori distribution p(x) over all allowable images. Usually,p(x) is taken to be a locally dependent Markov random field (MRF), a convenient model for quantifying the belief that the unknown true scene x* consists of, for example, large homogeneous patches, or smoothly varying grey levels which occasionally change level discontinuously. With y = (y1, ••• , y11) denoting the observed records of x*, the likelihood l(ylx) of any image xis combined with p(x), in accordance with Bayes's theorem, to form an a posteriori distribution p(xl y) o: l(ylx)p(x). The MAP estimate of x* is that image x which maximizes p(xl y).

In the research paper, MAP estimation for binary images is reformulated as a minimum cut problem in a certain capacitated network, and the classic Ford-Fulkerson algorithm can then be used to find x exactly.

3.1 Research gaps

We observed several research gaps while reading and implementing the research papers. They are:

1) In the paper of video segmentation by Yu et al.[1], the graph cut algorithm is a very computational task and hence the method is not real time runnable. For our data set, it took about 30 seconds for each frame to be processed.

Also, one of it's assumption is that they are using Markov Random field energy function which assumes the local dependency upto only 1 distance of every pixel.

2) In the paper of object tracking by Wang et al.[5], main drawback is that it is not completely unsupervised but rather it requires an initial rectangular box to begin the process of object tracking. This can be cumbersome if there are large number of videos.

Another problem with the approach is that it is inaccurate and unstable in detecting boundaries of the objects, especially at legs unlike the apprach by Yu et al.[1], which is mostly accurate if the camouflaging parts are manually coloured.

Moreover, the two scenarios in which there is motion blur or "non-object" instance make the model performs badly. Despite being different in nature, these two cases arguably arise from the complete lack of similar training samples in a training sets, which are focused on objects

Chapter 4

Objectives

The objectives of our project are as follows:

1) Perform the Semantic segmentation of the dancer's video dataset.

2) From a sequence of frames, performing rhythmic dance steps of Kathak, our objective is to identify rhythmic pattern occurring at beat intervals.

3)From Kathak footwork, our aim is to identify the frame instances within the start and end of each beat.

Our focus till now have been on the first objective since it is the step which is prior to other two objectives. We have used video segmentation[1] and SiamMask[5] to achieve the segmentation.

We are also working on the second and third objective. For that, the approach of Temporal Cycle-Consistency Learning[6] is being used and currently yet to be implemented

How each of the objective are accomplished is more elaborately explained in Methodology section and further sections.

Chapter 5

Methodology

I will briefly introduce the methodology that we have followed to obtain our objective. It can be broken down into three broad steps.

1) Performing video segmentation on the video
2) Matching the frames of the video
3) Determining the rhythmic patterns from the results of the matching

Till now, our work has been more or less involved the first two steps of the pipeline.

The first part is about video segmentation of the video given. There are many existing video segmentation algorithms available and we have thoroughly studied the work done in the past and implemented them for our project. More information about it is in the Literature Survey and results and discussion sections.

We have also worked on the second part of the pipeline. More details about it can be found in the literature survey and results and discussion sections.

Third part of the pipeline has not been explored till now and we shall implement and research on it in the next phase of the project in the next semester.

Chapter 6

Results and Discussion

Various Research papers models were applied on our Kathak Dataset to obtaind results. The results are mentioned below :

1) On applying video segmentation from [1] initially, we obtained a very clear segmentation of the video, but the toes of the dancer were incorrectly segmented as shown below. To train the image, a ground truth data of black and white images was required. We obtained it using Mask RCNN instance segmentation model.

(a) Frame1 (b) Frame2 (c) Frame3

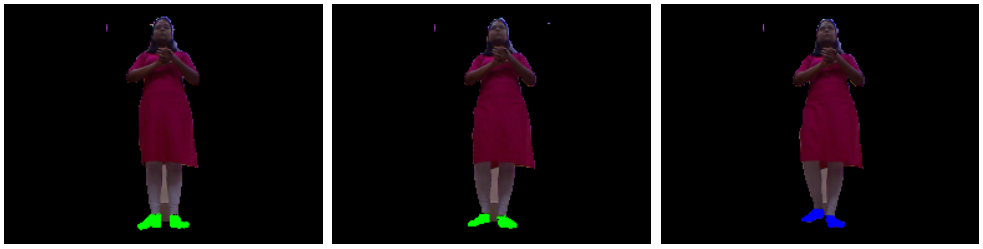
FIGURE 6.1: Original Images

In order to improve upon the toes segmentation, we manually coloured the toes in the original images and then got the results as shown below :

2) On applying SiamMask[5] on our dataset, we segmented the video. Few frames of the segmented video are shown below:

(a) Frame1 (b) Frame2 (c) Frame3

FIGURE 6.2: Model's Output



(a) Frame1 (b) Frame2 (c) Frame3

FIGURE 6.3: Model's Output on Edited Input

(a) Frame1

FIGURE 6.4: Initialization

(a) Frame1 (b) Frame2

FIGURE 6.5: SiamMask's Results

Chapter 7

Future Work

We have implemented various video segmentation algorithms for monocular video foreground/background segmentation. We have shown that these algorithms improves the segmentation results significantly on a number of challenging sequences.

In the next phase of the project, we will implement the second and third part of the pipeline which is matching of various frames and using it to determine the rhythmic patterns from the results of the matching.

Bibliography

[1] Yu, T., Zhang, C., Cohen, M., Rui, Y. and Wu, Y., 2007, February. Monocular video foreground/background segmentation by tracking spatial-color gaussian mixture models. In 2007 IEEE Workshop on Motion and Video Computing (WMVC'07) (pp. 5-5). IEEE.

[2] Wang, T. and Collomosse, J., 2011. Probabilistic motion diffusion of labeling priors for coherent video segmentation. IEEE Transactions on Multimedia, 14(2), pp.389-400.

[3] Boykov, Y., Veksler, O. and Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Transactions on pattern analysis and machine intelligence, 23(11), pp.1222-1239.

[4] Greig, D.M., Porteous, B.T. and Seheult, A.H., 1989. Exact maximum a posteriori estimation for binary images. Journal of the Royal Statistical Society: Series B (Methodological), 51(2), pp.271-279.

[5] Wang, Q., Zhang, L., Bertinetto, L., Hu, W. and Torr, P.H., 2019. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1328-1338).

[6] Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P. and Zisserman, A., 2019. Temporal Cycle-Consistency Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1801-1810).