**MOOCREC: LEARNING STYLES-ORIENTED MOOCS RECOMMENDER AND SEARCH ENGINE**

Saugat Aryal

IT 14 146602

Bachelor of Science Special (Honors) in Information Technology Specializing in Software Engineering

Department of Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

September 2018

**MOOCREC: LEARNING STYLES-ORIENTED MOOCS RECOMMENDER AND SEARCH ENGINE**

Saugat Aryal

IT 14 146602

Dissertation submitted in partial fulfillment of the requirements for the degree Bachelor of Science Special (Honors) in Information Technology Specializing in Software Engineering

Department of Software Engineering

Sri Lanka Institute of Information Technology

September 2018

**DECLARATION**

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name:Saugat Aryal Date:2018-09-03

The above candidate has carried out research for the B.Sc Dissertation under my supervision.

Supervisor: Mr. Nuwan Kodagoda

Co-Supervisor: Ms. Kushnara Suriyawansa

Signature:Date:2018-09-03

**ABSTRACT**

Massive Open Online Courses (MOOCs) have emerged as new learning phenomenon in the field of e-learning. Over the recent years, it has attracted significant number of learners as well as researchers. Wide range of researches are being carried out across multiple aspects of MOOCs. Video lectures are the most fundamental component in a MOOC. There are standard video styles that are normally used across several MOOC platforms, such as, talking head, demonstration, slides, animation etc. This paper presents an Image-Based classification approach of the video styles where a single video is split into multiple image frames, and then each frame is classified into one of the video style-category. Different classifier models built on top of each state-of-the-art deep neural architectures, including, AlexNet, VGGNet, GoogleNet (Inception) and ResNet are evaluated and the comparison of results is shown. Furthermore, the paper also discusses a numeric method to calculate the composition level of a single video style in multi-style filed videos based on the classification results.

**ACKNOWLEDGEMENT**

The work described in this research paper was carried out as our 4th year research project for the subject Comprehensive Design Analysis Project. The completed final project is the result of combining all the hard work of the group members and the encouragement, support and guidance given by many others. Therefore, it is our duty to express our gratitude to all who gave us the support to complete this major task.

We are deeply indebted to our supervisor Mr. Nuwan Kodagoda and our co-supervisor Ms. Kushnara Suriyawansa, lecturers of Sri Lanka Institute of Information Technology whose suggestions, constant encouragement and support in the development of this research, particularly for the many stimulating and instructive discussions. We are also extremely grateful to Mr. Jayantha Amararachchi, Senior Lecturer/ Head-SLIIT Centre for Research who gave and confirmed the permission to carry out this research and for all the encouragement and guidance given.

We also wish to thank all our colleagues and friends for all their help, support, interest and valuable advices. Finally, we would like to thank all others whose names are not listed particularly but have given their support in many ways and encouraged us to make this a success.

Table of Contents

[1 INTRODUCTION 8](#_Toc523775905)

[1.1 Background Context 8](#_Toc523775906)

[1.2 Research Gap 8](#_Toc523775907)

[1.3 Research Questions 9](#_Toc523775908)

[2 Methodology 10](#_Toc523775909)

[2.1 Video Styles Category 10](#_Toc523775910)

[2.2 Dataset Collection 11](#_Toc523775911)

[2.3 Classification using Pre-Trained Architectures 12](#_Toc523775912)

[2.3.1 VGGNet 12](#_Toc523775913)

[2.3.2 GoogleNet 12](#_Toc523775914)

[2.3.3 ResNet 12](#_Toc523775915)

[2.4 Composition of Video Styles 13](#_Toc523775916)

[2.5 Tools and Technologies 14](#_Toc523775917)

[2.6 Research Findings 14](#_Toc523775918)

[3 Results and Discussion 14](#_Toc523775919)

[3.1 Evidence 14](#_Toc523775920)

[3.2 Discussion 15](#_Toc523775921)

[4 Conclusion 15](#_Toc523775922)

[5 References 16](#_Toc523775923)

[6 Appendices 17](#_Toc523775924)

**LIST OF FIGURES**

[Figure 1: Image-based Video Styles Classification 10](file:///F:\Year%204\Research\Final%20Report.docx#_Toc523775671)

[Figure 2: Talking Head 11](file:///F:\Year%204\Research\Final%20Report.docx#_Toc523775672)

[Figure 3: Slide 11](file:///F:\Year%204\Research\Final%20Report.docx#_Toc523775673)

[Figure 4: Discussion 11](file:///F:\Year%204\Research\Final%20Report.docx#_Toc523775674)

[Figure 5: Code 11](file:///F:\Year%204\Research\Final%20Report.docx#_Toc523775675)

[Figure 6: Class Diagram 17](#_Toc523775676)

[Figure 7: Activity Diagram 18](file:///F:\Year%204\Research\Final%20Report.docx#_Toc523775677)

[Figure 8: Use-Case Diagram 18](file:///F:\Year%204\Research\Final%20Report.docx#_Toc523775678)

**LIST OF TABLES**

[Table 1: Video Styles category and description 11](#_Toc523775690)

[Table 2: Use Case Scenario 1 20](#_Toc523775691)

[Table 3: Use Case Scenario 2 20](#_Toc523775692)

[Table 4: Use Case Scenario 3 21](#_Toc523775693)

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| MOOCs | Massive Open Online Courses |
| CNN | Convolutional Neural Network |
| RAM | Random Access Memory |
| CPU | Central Processing Unit |

# INTRODUCTION

## Background Context

Video lectures are the fundamental and significant component of MOOCs. They are produced in various standard styles that are used across different MOOCs platforms (edX, Coursera and Futurelearn). Usually, a single video is the composition of these different styles. Guo et al.[1] mentions 6 different types of production styles: Slides (PowerPoint presentation with voice-over), Code (video screencast of writing code), Khan-style (full-screen video of instructor drawing free-hand), Classroom (video captured in live classroom), Studio (recording in a studio with no audience) and Office Desk (close-up shots of instructor at an office desk). Similarly, Hansch et al. [2] also presents similar video styles using different names. Other additional styles like, Animation, Conversation, Text-Overlay, Picture-in-Picture are also described.

The production cost of the video styles is quite expensive. It is necessary to consider about factors such as, cost, application and effectiveness while developing a video. Automatically classifying the video styles in a single video and calculating their net composition in a measurable unit (%) can provide valuable insights to researchers, learners and other parties who are involved in MOOCs domain.

The video styles classification process is a sub-component of the proposed “MoocRec: Learning Styles-Oriented MOOCs Recommender and Search Engine” system. The application will work as a platform to provide the most suitable, relevant and optimal learning resources to the learners based on their preferred style of learning. The above-mentioned component deals with two sub-tasks: identifying various video lecture styles contained in a MOOC video and their level of composition and finally calculating the aggregate for the overall course. The main objective of classifying and determining the aggregate of different video styles is such that it can provide essential information when mapping MOOCs with the learning style.

## Research Gap

The popularity of videos is increasing exponentially because of its usage in multiple platforms. Also, with the advancement in deep learning, CNN (Convolutional Neural Networks) and computational power, research in video classification techniques and algorithms has spiked up. Initially, the standard approach to video classification included features extraction and passing it to an SVM classifier to distinguish the classes. [3] proposed CNNs as powerful class of models for video classification. The video is treated as a bag of short, fixed-sized clips. The authors also describe a multiresolution for addressing the computational efficiency. Comparison of various video classification techniques is presented in [4] and a new approach recognized as ‘Long Short Term memory Recurrent Neural Network’ is proposed to increase the performance of video classification with better time and space complexity. Two different methods: Feature Pooling and LSTM, capable of aggregating frame-level CNN outputs in video-level predictions is proposed in [5].

A similar work of classifying an educational video frame into a particular category is presented in [6]. The proposed system is composed of two components: Feature extractor and Feature processor. Feature extraction is carried out using AlexNet and the feature processing/classification is performed through SVM classifier. A given video was not classified into any category rather than being incorrectly classified was the most common failure. Another limitation identified was that the system classifies based on individual video frames whereas taking few adjacent video frames would be easier and efficient.

Although there is significant amount of research being carried out regarding MOOCs and video classification methods, study about video styles used in MOOCs is relatively low. With the increasing popularity of deep neural networks, there is still a huge research gap for their applications in the field of MOOCs and their components, specifically, videos.

## Research Questions

Video styles play an important role while interacting with the learners. The different type of video styles can be directly mapped with the various learning styles of learners. The content analysis of MOOC videos by classifying them into standard styles and evaluating their composition level can be helpful while recommending MOOCs to different learners with different learning characteristics and requirements. However, selecting the appropriate deep neural architecture to perform the classification is the major question that needs to be addressed.

# Methodology

There are two major tasks that are to be carried out concerning with the video styles classification component: automatically classifying the video styles present in a MOOC video and finally, calculating the net composition of each style in a single video and overall course. Image-based classification approach is adopted to classify the video styles where a single video is split into multiple image frames and each frame is classified into one of the video style classes as shown in Figure 1. Different classifier models built on top of each state-of-the-art deep neural architectures, including, VGGNet, GoogleNet and ResNet are evaluated and the most optimal architecture is selected to perform the classification. Based on the classification, the composition of a particular video style is calculated for a single video and subsequently, using average mean, the net value of each style is generated for all the videos in a course.

A screenshot of a cell phone

Description generated with high confidence

Figure 1: Image-based Video Styles Classification

## Video Styles Category

Among the various video styles defined in the literature, we will be using only four different styles that are more relevant to computer programming related courses which is the focus of this study. A full list of video styles along with the description is shown in Table 1.

Table 1: Video Styles category and description

|  |  |
| --- | --- |
| Video Style Category | Description |
| Talking Head | Shows the instructor’s head |
| Slide | PowerPoint slide presentation with educational content |
| Code | A software demonstration or tutorial or full-screen code-writing |
| Discussion | Multiple people in a single frame |

A person wearing glasses and smiling at the camera

Description generated with very high confidence

Figure 2: Talking Head

A screenshot of a cell phone

Description generated with very high confidence

Figure 3: Slide

*A person sitting at a table in front of a window

Description generated with very high confidence*A screenshot of a cell phone

Description generated with very high confidence

Figure 4: Discussion

Figure 5: Code

The image labels were chosen in an attempt to minimize ambiguity. For example, a frame that contained a full-frame picture of a person was classified as talking head rather than slide. Other variations of each styles are also used in MOOC platforms. However, only the standard types were considered.

## Dataset Collection

The image dataset for the MOOC video styles category was not publicly available. Hence, the dataset was generated using a Python script and OpenCV [11] library, where the videos are split into multiple image frames and labelled accordingly. Since, we will be using leveraging the pre-trained architectures, small dataset consisting of 500 image frames as training set and 100 images as testing set for each video style category.

## Classification using Pre-Trained Architectures

Convolutional Neural Networks, inspired from the working of human brain is a milestone in the field of computer vision tasks, like image recognition. Various state-of-the-art CNN-based deep models pre-trained on ImageNet [7] are widely used for feature extraction by instantiating only the convolutional part of the models until the fully-connected layer. Finally, a fully-connected layer based on our custom class labels can be trained on top of the extracted features. As a complement to feature extraction, we can fine-tune the top layers of a pre-trained network to achieve a better accuracy.

### VGGNet

VGGNet [8] has two versions, VGG16 and VGG19 which contains 16 and 19 layers respectively. It pushed the depth of CNN architecture from 8 layers as in AlexNet to 16–19 layers, which greatly improves the discriminative power. In addition, by using very small (3 × 3) convolutional filters, VGGNet is capable of capturing details in the input images.

### GoogleNet

GoogleNet [9] is inspired by the Hebbian principle with multi-scale processing and it contains 22 layers. A novel CNN architecture commonly referred to as Inception is proposed to increase both the depth and the width of CNN while maintaining an affordable computational cost. There are several extensions upon this work, including BN-Inception-V2, Inception-V3, and Inception-V4.

### ResNet

ResNet [10] is one of the latest deep architectures which has remarkably increased the depth of CNN to 152 layers using deep residual layers with skip connections. It has been recently updated to 1000 layers on the CIFAR-10 dataset.

Different classifier models built on top of each of the pre-trained architectures and the results were evaluated. The model which performed better with the higher accuracy rate was selected.

## Composition of Video Styles

After the video styles are classified with the most optimal architecture, we then calculate the composition of each video style type for a single video and as well as for the overall course. While all the image frames can be extracted from a video, one frame in every two minutes is considered. This is because adjacent frames in high-frame-rate videos do not change significantly.

If *t* is the time period of a single MOOC video in seconds, the number of frames *n* considered for classification is given by:

*n = t / 120* (1)

If *h*, *s,* and *c* is the number of frames predicted as talking head, slide and code by the model, then the composition of each style in percentage (%) in a single video is given by:

*head = (h / n) \* 100*  (2)

*slide = (s / n) \* 100*  (3)

*code = (c / n) \* 100* (4)

Finally, if there are total of *v* number of videos in a MOOC course, where *h1*, *h2*, *h3*…*hv* represent the amount of talking head, *s1*, *s2*, *s3*…*sv* represent the amount of slide and *c1*, *c2*, *c3*…*cv* are the amount of code for each video, then the average composition of each style in percentage(%) for the overall course is given by:

*headaverage = (h1 + h2 + h3 +…+hv / v) \* 100* (5)

*slideaverage = (s1 + s2 + s3 +…+sv / v) \* 100* (6)

*codeaverage = (c1 + c2 + c3 +…+cv / v) \* 100* (7)

These average values calculated for each video styles are used while mapping a MOOC with the learning styles.

## Tools and Technologies

**Tools**

* Anaconda IDE
* Google Colaboratory

**Technologies**

* Python
* Keras Library
* OpenCV
* Tensorflow Backend
* Matplotlib

## Research Findings

The system is still in the implementation phase and only three different video styles (talking head, slide and code) were classified using the classifier built on top of VGG16 architecture. The classifier model was trained on a laptop with 6 GB RAM and Intel Core i5 2.20 GHz CPU processor, where it took around 120 minutes. The test set accuracy obtained is 90 percent classification accuracy for the three classes.

The accuracies for different architectures will be evaluated and comparison results will be shown for all five different categories for three different pre-trained architectures mentioned above.

# Results and Discussion

## Evidence

Since, the overall implementation is yet to be completed. The evidence will be provided in the near future after the completion of the whole project.

## Discussion

The video classification was carried out only for four different classes due to the computer programming domain. However, it can be expanded to cover other subject areas in other MOOC platforms. The findings of this research can provide valuable information regarding domain of MOOCs to various interested parties. Since, the dataset was not publicly available and was manually generated, the results and performance of the models were relatively considerate due to the usage of transfer learning and fine-tuning.

The time-taken for training the dataset was comparatively faster even on CPU. Using a GPU would give more faster results. Furthermore, if the hyperparameters of the networks are varied and tested with different values, it would give different result.

# Conclusion

Classification of video styles used in MOOCs using deep architectures is a noble approach in the field of MOOCs and computer vision. Using the concept of transfer learning, to extract bottleneck features from the video image frame and fine-tuning the model on top of that, provides a fast and easy way to classify the videos. The three main design goals of speed, accuracy and trainability are met. Because the dataset available was limited, the models built are flexible and easily re-trainable with more data.

Educational videos are an increasing presence on the Internet and this research can provide useful information about the kinds of videos that are suitable for a learner with specific learning styles. Two use cases of classifying video production style and calculating the composition level of each style have shown that this study is a useful tool in educational video analysis across all platforms.

# References

[1] P. J. Guo, J. Kim, and R. Rubin, “How video production affects student engagement,” in *Proceedings of the first ACM conference on Learning @ scale conference - L@S ’14*, 2014, pp. 41–50.

[2] A. Hansch, L. Hillers, K. McConachie, C. Newman, T. Schildhauer, and P. Schmidt, “Video and Online Learning: Critical Reflections and Findings from the Field,” *SSRN Electron. J.*, Mar. 2015.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[4] Lokhande B Gharde S, “A Review on Large-scale Video Classification with Recurrent Neural Network (RNN),” *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1774–1778, 2015.

[5] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4694–4702.

[6] M. Pratusevich, “EdVidParse: Detecting People and Content in Educational Videos,” *Massachusetts Inst. Technol.*, 2013.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[8] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014.

[9] C. Szegedy *et al.*, “Going Deeper with Convolutions,” Sep. 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015.

[11] “OpenCV library.” [Online]. Available: https://opencv.org/. [Accessed: 13-Aug-2018].

# Appendices

**A screenshot of a video game

Description generated with high confidence**

Figure 6: Class Diagram

A screenshot of a cell phone

Description generated with very high confidence

Figure 7: Activity Diagram

![A screenshot of a cell phone

Description generated with high confidence]()

Figure 8: Use-Case Diagram

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | |  |
| Use case Name |  |  | Split Video into Image Frames | |  |
| Pre –Condition |  |  | Videos are downloaded from the platforms | |  |
| Post-Condition |  |  | Collection of Image Frames are available | |  |
| Actor |  |  | Backend System | |  |
|  |  |  | 1. | Select the video from the stored location. |  |
| Main Success Scenarios |  |  | 2. Decompose into set of frames.  3. The use case ends with successfully splitting the video file into set of image frames. | |  |
| Extension |  |  | 1a. Invalid video source file is selected.  1b. Select a valid source file and proceed. | |  |

Table 2: Use Case Scenario 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | |  |
| Use case Name |  | Classification of Image Frame into Video Style | |  |
| Pre –Condition |  | Image Frames are transformed into lower dimension | |  |
| Post-Condition |  | Categorization of Image Frame | |  |
| Actor |  | Backend System | |  |
|  |  | 1. | The image frames are passed through trained CNN. |  |
| Main Success Scenarios |  | 2. The neural network classifies the image frame into one of the video style class.  3. The use case ends with, the model predicting the video style of the image frame. | |  |

Table 3: Use Case Scenario 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | |  |
| Use case Name |  | Calculate the Composition of Video Styles | |  |
| Pre –Condition |  | Image Frames are successfully classified by the model | |  |
| Post-Condition |  | The amount of each video styles in the course | |  |
| Actor |  | Backend System | |  |
|  |  | 1. | Group the similarly classified video styles together. |  |
| Main Success Scenarios |  | 2. Calculate the composition level of video styles for a single video in percentage.  3. Using Mean function or other aggregate function, calculate the composition of video styles for the overall course.  4. The use case ends with successfully storing the calculated details in the database. | |  |

Table 4: Use Case Scenario 3