Extracting Production Style Features of Educational Videos with Deep Learning

Fatima Maya[1], Philipp Krieter[2], Karsten D. Wolf[3] and Andreas Breiter[4]

Abstract: Enforced by the pandemic, the production of videos in educational settings and their availability on learning platforms allow new forms of video-based learning. This has a strong benefit of covering multiple topics with different design styles and facilitating the learning experience. Consequently, research interest in video-based learning has increased remarkably, with many studies focusing on examining the diverse visual properties of videos and their impact on learner engagement and knowledge gain. However, manually analysing educational videos to collect metadata and to classify videos for quality assessment is a time-consuming activity. In this paper, we address the problem of automatic video feature extraction related to video production design. To this end, we introduce a novel use case for object detection models to recognize the human embodiment and the type of teaching media used in the video. The results obtained on a small-scale custom dataset show the potential of deep learning models for visual video analysis. This will allow for future use in developing an automatic video assessment system to reduce the workload for teachers and researchers.

Keywords: Video-based learning; MOOC; Video lecture design; Video features; Object detection; YOLOv4-algorithm; Deep learning

1 Introduction

The use of educational videos such as explanatory videos, tutorials and video lectures has increased massively over the past 20 years both in formal and informal education. YouTube has become the central distribution platform for audio-visual explanations of even complex topics such as dark matter or programming on quantum computers. Explanatory videos are used by students to review lessons they did not understand, to do homework and write term papers, to deepen school knowledge and to prepare for exams [Wo21, PGM17]. In online education programs such as MOOCs, videos also play a central

[1] University of Bremen, Faculty 3 - Mathematics and Computer Science, Am Fallturm 1, 28359 Bremen, famaya@uni-bremen.de

[2] University of Bremen, Faculty 3 - Mathematics and Computer Science, Am Fallturm 1, 28359 Bremen, pkrieter@uni-bremen.de

[3] University of Bremen, Faculty 12 - Educational Sciences, Centre for Media, Communication and Information Research (ZeMKI), Linzer Str. 4, 28359 Bremen, wolf@uni-bremen.de

[4] University of Bremen, Faculty 3 - Mathematics and Computer Science, Am Fallturm 1, 28359 Bremen, abreiter@uni-bremen.de

role for content delivery. Moreover, the creation of explanatory videos by students for educational purposes have become a fairly common exam assignment in secondary and tertiary education to foster deep learning [Ho19, KW21].

While it is advantageous to have a large and free availability of video-based learning resources, the quality of videos in terms of their didactic and media design has to be taken into consideration in order to achieve an effective learning process. Therefore, there is a growing interest in different scientific fields such as instructional psychology, didactic, film studies, communication and media science, as well as video-based learning to further understand how to design effective and motivational explanatory videos for teaching and learning.

Screening or evaluating educational videos for specific design criteria both in educational practice as well as in empirical studies has mostly been done manually [Ho22]. Human analysis of videos requires an enormous amount of time resources for even small samples. To further open up audio-visual content for large scale research as well as for assessment purposes, it would be desirable to partially automate this process.

The main contribution of this work is to demonstrate how deep learning can support the automatic classification of educational videos and the extraction of the relevant production design characteristics with a selected set of features. To address this, we utilize object detection algorithms to identify design features such as human protagonists, animated visualization or slides in educational videos. Additionally, we investigate how reliably these algorithms recognize the target elements. For this aim, we collect and present a suitable small-scale dataset for training our model and show that we can reach a mean average precision (mAP) of 62.2 % demonstrating the great potential of the presented approach.

2 Related Work

In a review on the recent advances in video-based learning research, *Navarrete et al.* [NHE21] studied 41 reviewed articles that examines specific characteristics of the educational videos. Based on this study, they proposed a taxonomy to group the video features in the following categories: (a) audio features, (b) visual features, (c) textual features, (d) production style, (e) instructor’s behaviour, (f) learner interaction with the video, (g) interactive features and (h) instructional design principles. In the context of *audio features* which include all information related to audio stream and spoken text, *Shi et al.* [Sh19] found an important positive correlation between the modulated loudness and the learner's knowledge gain and a negative correlation between average syllable duration (talking speed) and knowledge acquisition, where these 2 features were extracted automatically using an open source toolkit for audio features retrieval. The *visual features* category consists of the information that can be extracted from the video frames. In [Sh19], the authors computed the text ratio and the image ratio in order to evaluate the design quality of the slide presentation.

Any text-based information that could be extracted from the video can be grouped in the *textual features* category. A common practice in text retrieval from video is to generate speech transcripts from spoken content, which are processed and analysed using NLP models, and then used for video segmentation [GHE20]. Many features related to instructor's behaviour, as a central facilitator, have been also retrieved from videos, for instance, detecting gesture and facial emotions using deep learning models [CWJ20]. The learner's interaction with a certain educational video was examined by many studies, for example, predicting the learner's performance by analysing video clickstream data with LSTM neural network [MCA20]. The interactive features denote the functionalities that help the learner to interact with the video. In this study [El21], analysing quizzes answers was used to predict the learner's knowledge gain. The *production style* category groups the video features based on design choices. In his work, *Chorianopoulos* [Ch18] proposed a taxonomy of video production styles by considering two dimensions: human embodiment (whether the instructor is visible) and the used teaching media type (whiteboard, slides, animation film, lecture, etc.) to help classify videos in a comprehensive way for further analysis. Previous work has explored the impact of production design on the learning outcomes. According to [CW15], *Chen et al.* found that compared to voice over presentation video type, other types of video composition improve the learning performance and reduce the cognitive load: both lecture capture, where the classroom is recorded, and picture to picture videos, where the instructor overlays slides, as well as animated content, are suitable for online learning. Furthermore, students who watched videos where the instructor's face is visible reported to learn better and needed less effort compared to those who watched them without the face [KBG15].

Animation-based learning has also proven to be an effective learning medium. Compared to traditional learning environments based on verbal explanations, learning through graphic animation helped improve understanding of complex notions [Ro09]. For example, it is difficult to visualize the movement of electrons or chemical reactions using traditional teaching tools. Therefore, using this medium can help create mental representations of phenomena that cannot be visualized in the classroom or laboratory.

In contrast to the other feature categories, we observed an absence of technology components that could help to extract video properties related to the video's design. Most studies use *manual* approaches to extract these features by assigning humans to watch a large set of videos and encode their properties, which could be an exhausting task and affect the advancement of the research in this field. Therefore, there is a need to automate the feature extraction from educational videos to study larger sets of videos.

Within computer vision, deep learning (DL) approaches can provide efficient methods to analyse videos automatically. In fact, deep learning networks have been widely used for image classification, and particularly, for object detection as a technique that allows to recognize and locate objects of specific classes in digital images. These DL models are the backbone of real-time video analysis and have a wide range of applications, from pedestrian and traffic sign detection for autonomous driving [Li20], detection of fire and personal protective equipment at construction sites [Ku21] to human abnormal behaviour recognition [Liu20]. However, in the context of educational videos, the object detection models have not yet been used for the visual analysis.

3 Methodology

In this work, the task of identifying visual properties related to the production style of educational videos is formulated as an object recognition problem, where we investigate the possible implementation of the state-of-the-art YOLOv4 object detection algorithm [BWL20] in identifying the human presence (teacher, hand, no person) and type of the used teaching tool (whiteboard, pen, slides, animation). To this end, a custom dataset is collected and annotated to train the model to perform the target task.

3.1 Dataset

Due to the lack of suitable datasets that include the target objects, we collected a custom dataset from the following sources to ensure diversity and to obtain a generalizable model: *Open Images Dataset V6*[5], *Google images* using the following search keywords: {*PowerPoint Slides, Latex Slides, Slide Presentation*}, *SlideImages* [MME20] and frames taken from educational videos with different production styles. As a next step, the dataset needs to be labelled to train the object detection model. For this purpose, the open source annotation tool *LabelImg*[6] is used to manually create regions around the target objects in the images called bounding boxes. The corresponding object class and its location coordinates will be assigned to each of the boxes and saved in a *txt* file for training. Our custom dataset contains 3797 images[7] with more than 500 samples for each of the seven object classes and a total of 10 729 annotations (labelled objects) distributed as follows: man (2982), human hand (2257), woman (2206), slide (1141), pen (962), whiteboard (617), animation frames (564). It exists an uneven distribution of objects numbers among the classes. Although approximately the same number of sample images has been collected for all object classes, some images, for instance, contain more than one person which explains the dominance of the classes *Man*, *Woman* and *Hand*. For the classes *Whiteboard* and *Animation*, there is often just one object per image which explains why these classes are under-represented in the dataset.

3.2 Object Detection with YOLOv4-Algorithm

We aim to automatically extract visual features related to the instructor's presence and the used teaching tool in educational video frames. In computer vision, this problem can be solved using an object detection model which simultaneously performs object

[5] https://storage.googleapis.com/openimages/web/index.html
[6] https://github.com/heartexlabs/labelImg
[7] https://gitlab.informatik.uni-bremen.de/famaya/edu_vid_YOLOv4

classification and localization within an image. *YOU Only Look Once (YOLO)* is a state-of-the art object detection algorithm that is able to predict the bounding boxes and the corresponding object class probability at one pass. This algorithm was chosen due to its ability to run object detection in real time making it suitable for processing videos. Additionally, it has outperformed the previous object detection networks in terms of achieving a significant balance between accuracy and speed [BWL20]. YOLOv4 has a complex architecture composed of the following three main blocks: It uses the *CSPDarknet53* network (backbone) as a features extractor which has 53 convolutional layers organized in 2 consecutive modules: the CBM (Convolution, Batch Normalization and MISH) contains a convolutional layer, a batch normalization layer and a mish activation function. This module is used to extract the image features. The CSP (Center and Scale Prediction) block is deployed to improve the learning ability of the convolutional layer. It takes the extracted features from the previous CBM block as input, applies zero padding to preserve the input size and pass it to an CBM block, where the input is then divided into two parts: one half is sent to a residual unit followed by another CBM block, and the other half is routed directly to an CBM module, where at the end of the CSP block both outputs will be concatenated. This technique helps to preserve low-level features. An *SPP* block is then applied to increase the receptive field of the model by using multiple max-pooling layers. Additionally, *PANet* is incorporated in the network to enhance the process of instance segmentation by identifying objects present in the images in a pixel level. The CBL (Convolution, Batch Normalization and Leaky-ReLU) block is the main structure of the PANet. The described blocks build the neck of YOLOv4 algorithm. Finally, three *YOLO* heads with different sizes placed in different stages of the network are used to process the extracted feature maps in order to detect objects of different scales.

3.3 Model training

To set its architecture for training, YOLOv4 uses *Darknet* [Re16], an open-source neural network framework written in C and CUDA which is faster and supports CPU (Central Processing Unit) and GPU (Graphics Processing Unit) computations. *Google Colab*[8] is chosen as an efficient environment to build this project. It offers the access to *the Nvidia Tesla T4 GPU*[9], and has the necessary libraries to train the models. To train the model, the cleaned dataset with a total of 3797 images was split into training set consisting of 80% of the samples (3037 images), and the rest 20% was used for validation (759 images). The default YOLOv4 model is modified adapting it to solve the given problem: the maximum number of training iterations is set to 14000, the input images are rescaled to the size of 416 × 416, the learning rate was first initialized with a fixed value of 0.001, and after 80% and 90% of the steps it was divided by 10 to ensure that the weights are not updated aggressively by the end of the training and miss the local minimum. The number of classes in the YOLO layers (output layers) is set to 7, and the number of filters in the convolutional

[8] https://colab.research.google.com/?utm_source=scs-index

[9] https://gist.github.com/cedrickchee/273cf787f8d998f1ba61a7265b3e1957

layers before each of the 3 YOLO layers is changed to 36, obtained by the following equation: filters = (classes + 5) × 3. Training a CNN from scratch might be challenging as it requires a large-scale dataset with hyper-parameters tuning in order to achieve optimal results. To address this issue, pre-trained YOLOv4 weights on the *COCO-dataset* [Li14] with 80 object classes including persons, were used to initialize the weights. Additionally, hue, saturation and exposure data augmentation techniques were used to expand the size of the training set. A training duration of 17 hours was recorded for YOLOv4. The loss over the training set and the mean average precision (mAP) over the validation set are evaluated on a 1000 iteration interval basis. The average loss, which estimates how much the model prediction distribution deviates from the ground truth distribution, settles at a value of 1.99 after 14000 iterations. The training yields a mean average precision of 62.2% which is a popular metric for measuring object detection accuracy that compares the predicted bounding box with the ground truth box and returns the corresponding score.

3.4 Results and Discussion

Most images of the collected dataset do not come from an educational context. additionally, some objects and person classes such as male/female instructor and slides were not seen combined enough during the training. To evaluate the performance of our model, we selected a separate test set consisting of video frames from different video production designs: instructor using a whiteboard, instructor presenting with slides, hands writing on a whiteboard, and animation frames ranging from simple to sophisticated drawing. Another factor is considered when selecting the test images is to ensure that it contains the target objects with different sizes and captured from different angles as shown in Figure 1. The object detection model is evaluated with a set of 100 images, where each object appears 20 times (40 times for the class hand) within the test set. To assess the accuracy of YOLOv4, the *root-mean-square error (RMSE)* was chosen to measure the deviation between the predicted number of objects and the ground truth number expressed as follows

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - D_i)^2},$$

where N denotes the number of images for each class, A_i is the ground truth number of objects in an image and D_i represents the number of the detected objects by the model. Furthermore, the average prediction probability was also evaluated for each class.

Table 1 displays the RMSE and the average prediction score achieved for each object. The results show that the YOLOv4 has a low error value and a good prediction score in detecting all classes. The custom model works well in recognizing the classes whiteboard, slides, male instructor and animation frames. However, it has relatively lower performance when detecting female instructor, hands and pen. To gain more insight into the observed results, we analysed the test images and focused on those where the model fails to detect a certain object. As visualized in Figure 1, YOLOv4 is quite strong in locating objects of big size and with different angles within the images such as whiteboard, slides, animation and instructor. However, the model could not detect the pen, hands and far located instructors as shown in Figures (a) and (f). In the test image (g), the model successfully recognized the frame as animation and did not confuse with a slide, although it contains text. Additionally, the drawn pen and hand in figure (h) were not false detected, but in Figure (i) we observe that an animation character was detected as a real person due to the high-quality drawing.

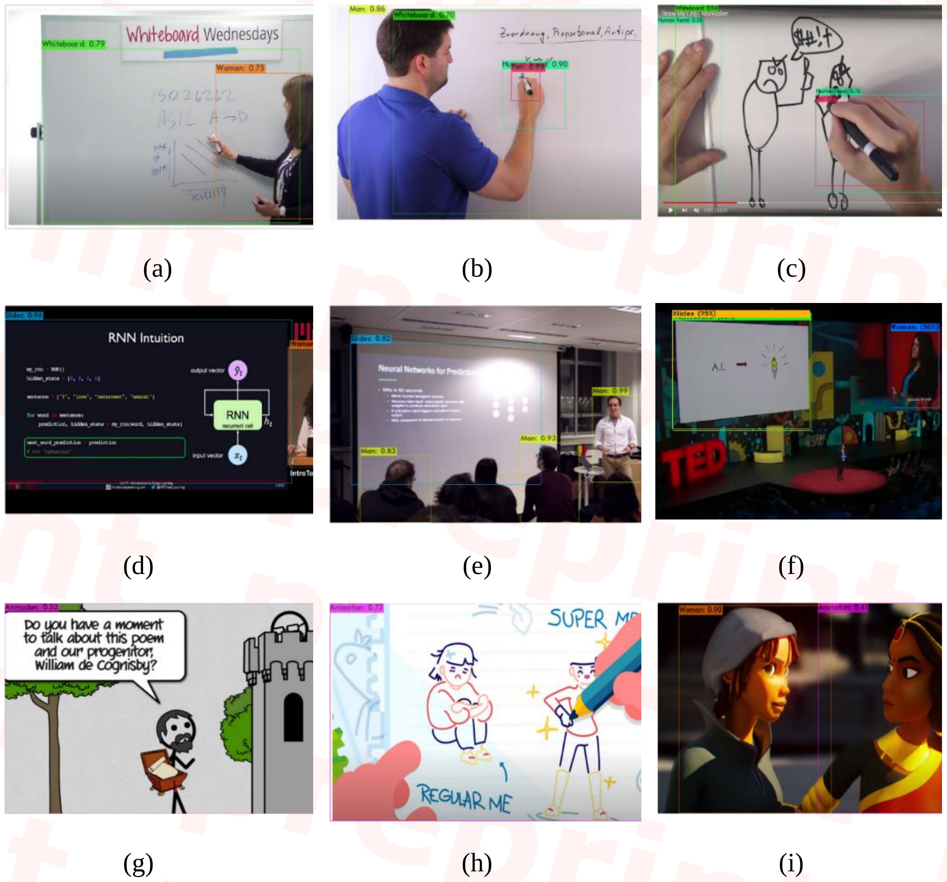
Although all the experiments have been carried on static images, the findings of this work are also applicable to video samples. Testing the custom YOLOv4 model in the settings described in section 3.3, we recorded an average detection speed of 26.7 frame per second (FPS), achieving a real-time inference.

Object Class	Detected	Not Detected	Total Objects	Images Number	RMSE	Average Prediction Score (%)
Male Instructor	17	3	20	20	0.38	84.37
Female Instructor	15	5	20	20	0.5	75
Hand	30	17	47	20	0.59	64
Whiteboard	20	0	20	20	0	87.7
Pen	12	8	20	20	0.63	70
Slide	19	1	20	20	0.22	82.37
Animation	16	4	20	20	0.44	76.9

Tab. 1: Experiments results with test images.

4 Conclusion

This paper explores the use of deep learning techniques in analysing educational videos. More specifically, it examines the suitability of using object detection algorithms to automate the recognition of relevant features related to the production design of instructional videos. To this end, the state-of-the-art algorithm YOLOv4 was trained on a custom dataset of images representing the presence of the instructor and the type of the used teaching tool. To reflect the most probable performance of the algorithm on unseen



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 1: Example output images representing the different educational scenarios.

data and in real applications, the utilized test images were selected from different educational videos based on the complexity of the represented scene. We can conclude that object detection algorithms have a remarkable potential to be used in the task of extracting design features of educational videos.

There are several starting points for future work. The custom dataset could be improved by adding more samples for each class and from different educational scenarios to improve the classification and the detection accuracy. Another promising aspect is to include more object classes such as digital whiteboard, code screencast and tablet screen.

While this work focuses only on the production style features, it would be highly relevant to develop a pipeline combining our custom model with the other algorithms mentioned in the literature review to obtain an end-to-end system for video analysis. Such system would significantly support researchers to conduct large-scale studies to efficiently evaluate the quality of each video type and develop an automatic video assessment system. In addition, such system could be also beneficial for recommendation systems, which are mostly based on the video meta data such as title and search topics but not the actual design qualities of the video.

5 Bibliography

[BWL20] Bochkovskiy, A.; Wang, C.Y.; Liao, H.M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. In: CoRR abs/2004.10934, 2020.

[Ch18] Chorianopoulos, K.: A Taxonomy of Asynchronous Instructional Video Styles. In: The International Review of Research in Open and Distributed Learning, 19(1), 2018.

[CW15] Chen, C.M.; Wu, C.H.: Effects of different video lecture types on sustained attention, emotion cognitive load and learning performance. Computers & Education, Volume 80, pp. 108-121, 2015.

[CWJ20] Chen Y.; Wang, C.; Jian, Z.: Research on Evaluation Algorithm of Teacher's Teaching Enthusiasm Based on Video. In: 2020 6th International Conference on Robotics and Artificial Intelligence (ICRAI 2020). Association for Computing Machinery, New York, NY, USA, 184–191, 2020.

[El21] El Aouifi, H. et.al.: Toward Student Classification in Educational Video Courses Using Knowledge Tracing. In: Business Intelligence. Ed. by Mohamed Fakir, Mohamed Baslam, and Rachid El Ayachi, pp. 73–82, 2021.

[GHE20] Ghauri, J. A.; Hakimov, S.; Ewerth, R.: Classification of Important Segments in Educational Videos using Multimodal Features. In: CoRR abs/2010.13626, 2020.

[Ho19] Hoogerheide, V. et.al.: Generating an instructional video as homework activity is both effective and enjoyable. In: Learning and Instruction 64. pp. 101226. ISSN: 0959-4752, 2019.

[Ho22] Honkomp-Wilkens, V. et.al.: Informelles Lernen auf YouTube: Entwicklung eines Analyseinstruments zur Untersuchung didaktischer und gestalterischer Aspekte von Erklärvideos und Tutorials. In: MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung 18.Jahrbuch Medienpädagogik, pp. 495–528, 2022.

[KBG15] Kizilcec, R.; Bailenson, J.; Gomez, C.: The Instructor’s Face in Video Instruction: Evidence from Two Large-Scale. In: Journal of Educational Psychology 107, 2015.

[Ku21] Kumar, S. et.al.: YOLOv4 algorithm for the real-time detection of fire and personal protective equipments at construction sites. In: Multimedia Tools and Applications, 2021.

[KW21] Kulgemeyer, C.; Wolf, K. D.: Lehren und Lernen mit Erklärvideos im Fachunterricht. In: Handbuch: Lernen mit digitalen Medien, 1. Auflage., G. Brägger und H.-G. Hrsg. Weinheim: Beltz, pp. 474–487, 2021.

[Li14] Lin, T. Y. et al.: Microsoft COCO: Common Objects in Context. In: CoRR abs/1405.0312, 2014.

[Li20] Li, Y. et.al.: A Deep Learning-Based Hybrid Framework for Object Detection and Recognition in Autonomous Driving. In: IEEE Access 8, pp. 194228–194239, 2020.

[Liu20] Liu, Y. et.al.: Abnormal Behavior Recognition Based on Key Points of Human Skeleton. In: IFAC-PapersOnLine, Volume 53, Issue 5, pp. 441-445, 2020.

[MCA20] Mubarak, A. A.; Cao, H. K.; Ahmed, S. A. M.: Predictive learning analytics using deep learning model in MOOCs’ courses videos. In: Education and Information Technologies 26, pp. 371–392, 2020.

[MME20] Morris, D.; Müller-Budack, E.; Ewerth, R.: SlideImages: a dataset for educational image classification. In: European Conference on Information Retrieval. Springer, Cham, 2020.

[NHE21] Navarrete, E.; Hoppe, A.; Ewerth, R.: A Review on Recent Advances in Video-based Learning Research: Video Features, Interaction, Tools, and Technologies. In: Proceedings of the CIKM 2021 Workshops co-located with 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Gold Coast, Queensland, Australia, November 1-5, 2021. Vol. 3052. CEUR Workshop Proceedings. CEUR-WS.org, 2021.

[PGM17] Pappas, I.; Giannakos, M.; Mikalef, P.: Investigating students’ use and adoption of with-video assignments: lessons learnt for video-based open educational resources. In: Journal of Computing in Higher Education 29, pp. 160-177, 2017.

[Re16] Redmon, J.: Darknet: Open Source Neural Networks in C. http://pjreddie.com/darknet/. 2013–2016.

[Ro09] Rosen, Y.: The Effects of an Animation-Based On-Line Learning Environment on Transfer of Knowledge and on Motivation for Science and Technology Learning. In: Journal of Educational Computing Research 40.4, pp. 451–467, 2009.

[Sh19] Shi, J. et.al.: Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality. In: SALMM ’19. Nice, France: Association for Computing Machinery, pp. 11–19, 2019.

[Wo21] Wolf, K. D. et.al.: Leistungsoptimierung von Schülerinnen und Schülern durc schulbezogene Erklärvideonutzung auf YouTube: Entschulungsstrategie oder Selbsthilfe?. In: MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung 42 (Optimierung): pp. 380-408, 2021.