# Automatic MOOC video classification using transcript features and convolutional neural networks

Houssem Chatbri
Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
houssem.chatbri@dcu.ie

Kevin McGuinness
Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
kevin.mcguinness@dcu.ie

Suzanne Little
Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
suzanne.little@dcu.ie

Jiang Zhou
Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
jiang.zhou@insight-centre.org

Keisuke Kameyama
Faculty of Engineering, Information and Systems, University of Tsukuba
Tsukuba, Japan
keisuke.kameyama@cs.tsukuba.ac.jp

Paul Kwan
School of Science & Technology,
University of New England
Armidale, NSW, Australia
paul.kwan@une.edu.au

Noel O'Connor
Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
noel.oconnor@dcu.ie

## ABSTRACT

The amount of MOOC video materials has grown exponentially in recent years. Therefore, their storage and analysis need to be made as fully automated as possible in order to maintain their management quality. In this work, we present a method for automatic topic classification of MOOC videos using speech transcripts and convolutional neural networks (CNN). Our method works as follows: First, speech recognition is used to generate video transcripts. Then, the transcripts are converted into images using a statistical co-occurrence transformation that we designed. Finally, a CNN is used to produce video category labels for a transcript image input. For our data, we use the Khan Academy on a Stick dataset that contains 2,545 videos, where each video is labeled with one or two of 13 categories. Experiments show that our method is strongly competitive against other methods that are also based on transcript features and supervised learning.

## CCS CONCEPTS

**•Information systems → Content analysis and feature selection;**

## KEYWORDS

MOOC video classification, transcript features, convolutional neural networks (CNN)



,
© 2016 Copyright held by the owner/author(s). 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

**ACM Reference format:**
Houssem Chatbri, Kevin McGuinness, Suzanne Little, Jiang Zhou, Keisuke Kameyama, Paul Kwan, and Noel O'Connor. 2016. Automatic MOOC video classification using transcript features and convolutional neural networks. In *Proceedings of , , ,* 6 pages.
DOI: 10.475/123_4

## 1 INTRODUCTION

In recent years, a growing number of highly-regarded academic institutions adopted MOOC and started to collaborate with online platforms such as Coursera and Udemy [21]. This led to a big interest by students worldwide and has been translated into an emerging online industry that formed large communities of MOOC students and educators [6].

Due to MOOC videos being produced frequently, systems for automatic video storage, indexing, classification, and retrieval should be designed and constantly improved to maintain the sustainability of MOOC platforms. In this paper, we focus on automatic video classification which aims to classify a video into one or many of known categories that describe the video content. Automatic video classification is of important benefit for numerous applications of video analysis such as content-based retrieval and content recommendation.

We propose an approach that exploits an intrinsic characteristic of MOOC videos, which is the fact that the video topic category that can be extracted from the speech. For this purpose, our approach extracts the video transcript in a first step, and uses a convolutional neural network (CNN) for classification as a last step. In between, transcripts are transformed into images in order to leverage the proven high performances of CNNs when dealing with visual data. We do that by designing a statistical co-occurrence transform.

For our experiments, we use the Khan Academy educational platform, which has become one of the most popular MOOC platforms

due to the high quality and intuitiveness of its videos [26]. We compare our approach with several other methods and we report significantly improved performances.

The remainder of this paper is as follow: In Sec. 2, we overview the related research on MOOC, the Khan Academy data that we use, and key automatic video classification methods. Sec. 3 describes our video classification method. We report our results in Sec. 4, and finally our concluding remarks and future directions in Sec. 5.

## 2 RELATED WORK

### 2.1 Massive open online courses (MOOC)

Massive open online courses (MOOC) received a lot of interest and become a commodity for a large number of students worldwide [6]. The concept has grown from recording lectures or talks and providing them online (such as in VideoLectures.net) to producing high quality educational videos that follow a specific template.

Most MOOC videos are relatively short while examination is managed automatically or with peer-review and group collaboration [3]. MOOC platforms such as Khan Academy, Coursera, Udacity, edX and Udemy have been adopted in influential institutions of higher education [21]. Coursera and Udacity are for-profit companies while Khan Academy, edX and Udemy are non-profit organizations.

Research has been conducted to understand the characteristics of the MOOC community and the recipe of an engaging MOOC video. In [8], Christensen et al. showed that the demography of University of Pennsylvania's MOOC students is largely young educated males from developing countries whose main reasons are advancing in their job and satisfying curiosity. In [11], Guo et al. empirically showed that informal styles such as talking-head and Khan Academy's drawings are more engaging than typical long lectures and PowerPoint slides (Fig. 1). In [3], Alraimi et al. showed that perceived reputation (i.e. trustworthiness, confirmation of user expectations) and perceived openness (i.e. institution's freedom of information access, resource sharing) are the two strongest predictors to explain MOOCs continuance intention to use. Using such empirical findings, MOOC providers can both differentiate themselves from competitors and enhance an individual's intention for continued MOOCs enrollment.

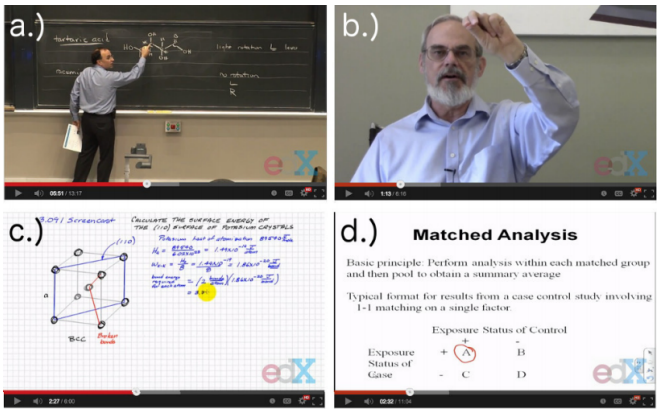


**Figure 1: MOOC video production style [11]: (a) classroom lecture, (b) "talking-head" style, (c) digital drawing tablet style of Khan Academy, (d) PowerPoint presentation style.**

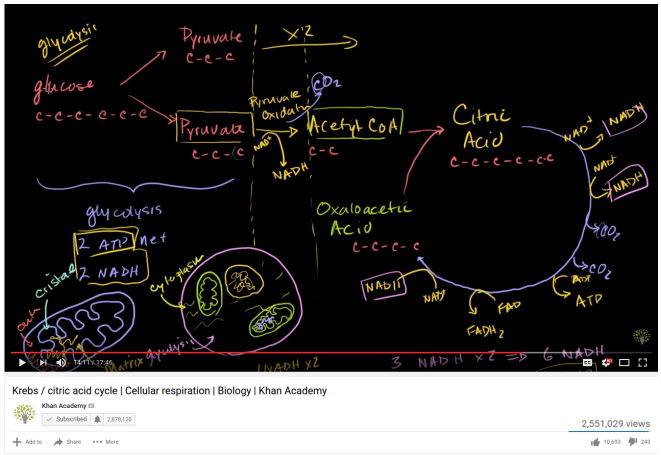


**Figure 2: A Khan Academy video with black background and colored sketched content, which is the style used in the Khan Academy on a Stick dataset.**

### 2.2 Khan Academy

Khan Academy is one of the most popular MOOC platform and it has gained large popularity among students in recent years due to the high quality of its videos and the excellent presentation skills, which led to its integration in a number of educational institutions [20, 26]. Usually, Khan Academy videos contain freehand sketched content on a digital tablet (Fig. 2).

In order to help institutions in developing countries with limited Internet access, Khan Academy put together an offline dataset called the *Khan Academy on a Stick* dataset[1]. This dataset contains 2,545 videos that were recorded between 2006 and 2013. The videos depict sketched content on a black background (Fig. 2), and they are annotated with 13 labels (Table 1). Fig. 3 shows the histogram of video labels, Fig. 4 shows the histogram of video durations, and Fig. 5 shows the histogram of video frame resolutions. It can be seen from the figures that the distributions of classes, video durations and frame resolutions are not balanced. Of the 2,545 videos, 238 have more than one label (e.g. Algebra and Trigonometry, Biology and Healthcare and Medicine).

Although Khan Academy provides a rich repository of videos for free, not much research was reported from the computational intelligence community. To the best of our knowledge, the work by Shin et al. on generating *visual transcripts* (i.e. structured visual documents) from Khan Academy videos [23] is worth noting and shares the same background with our work.

### 2.3 Video classification

The literature on automatic video classification is wide and it includes research on action recognition [13], anomaly detection [19], lecture video classification [4], etc. Video classification methods

[1] http://khan.mujica.org

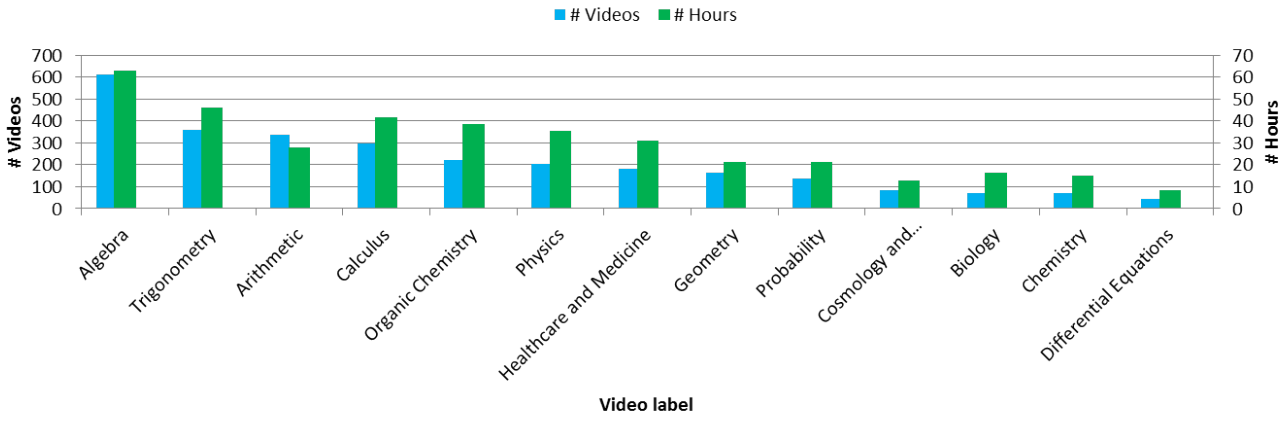


**Figure 3: Number of videos and footage duration for each video label.**

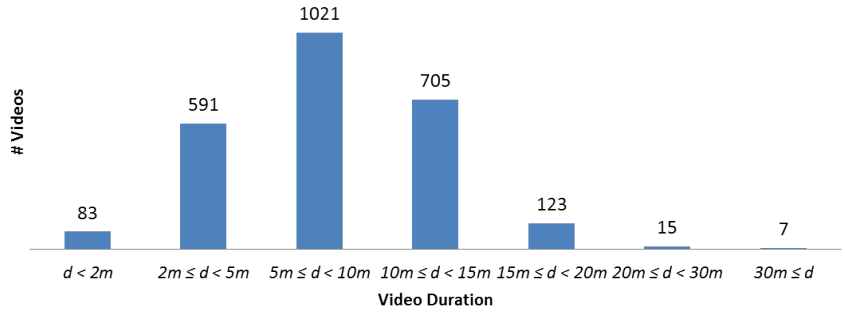


**Figure 4: Number of video for each duration interval.**

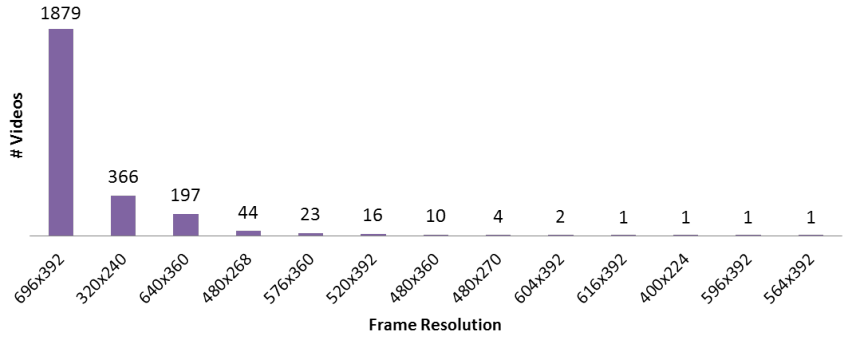


**Figure 5: Number of video of each frame resolution.**

have traditionally used shallow features [5] before deep learning based approaches started to emerge [29].

Shallow features correspond to hand-crafted visual descriptors that encode appearance and motion information. This includes the Histogram of Oriented Gradients (HOG) [9], Histogram of Optical Flow (HOF) [7], and spatio-temporal interest points [18]. Local feature descriptors are extracted using dense grids [27] or by interest point detection [17, 18], and graph structures have been used to encode spatio-temporal information [12]. On the other hand, video classification has been achieved by using text features. This includes video closed captions and viewable text (e.g. scene text, news bar), in which case OCR is used to extract text from video

**Table 1: Topics of the Khan Academy on a Stick dataset videos.**

| Category | Label |
|---|---|
| Math | Algebra, Calculus, Geometry, Trigonometry, Arithmetic, Differential Equations, Probability |
| Science | Biology, Cosmology and Astronomy, Organic Chemistry, Chemistry, Healthcare and Medicine, Physics |

frames [5], in addition to features that are extracted from video transcripts [4].

Contrary to shallow methods, deep learning techniques automatically learn discriminant features for video classification and they leverage the abundance of large amount of online videos [1]. Convolutional Neural Networks (CNNs) have been used for this purpose and they are fed with single frames or stacked frames [13]. Deep Learning approaches are suitable to data with discriminative visual information, so they perform well in datasets where object motion is an important feature such as human action datasets [24]. Research has also been conducted to deal with noisy data [2].

# 3 MOOC VIDEO CLASSIFICATION USING TRANSCRIPT FEATURES AND CNN

In this work, we rely on text features instead of frame visual features. We generate the video transcripts using a standard toolkit (Sec. 3). Then, we split them into 80% for training and 20% for testing while making sure that all video labels and label combinations are represented in that ratio.

We rely on transcripts instead of the frames' visual information for the following reasons:

- The speech contains keywords that are associated with certain topics, and they are usually easily discriminated. Contrarily, frames tend to follow the same template by the video producer, which makes them harder to discriminate video topics.
- Visually similar sketches are used in videos of different topic labels, e.g. Calculus and Arithmetic, Chemistry and Organic Chemistry, etc. This makes discrimination based on visual features more challenging.

Our approach starts by extracting the transcript of each video using the CMU Sphinx toolkit [16]. This tool is based on Hidden Markov Models (HMM) and it reaches word error rates (WER) of 26.9% and 22.7% on the VM1 and WSJ1 datasets respectively [10]. Afterwards, transcripts are converted into images using a co-occurrence transform that we designed (Sec. 3.1). Finally, the transcript image is fed to a CNN that produces the labels of the video (Sec. 3.2).

## 3.1 Transcript to image transform

The purpose of this step is to convert each transcript to an image-like representation that characterizes the transcript file content and that can be fed to a CNN classifier. To this end, we designed a statistical co-occurrence transform that works as follows (Algorithm 1): Considering a transcript $T$ as an array of characters, each five adjacent characters $C = T[j \bmod length(\mathrm{T})], j = i, ..., i+4\}$ are used to populate a $128 \times 128$ image $I$ by using the ASCII codes of the first four characters of $C$ to calculate a pixel coordinate and the ASCII code of the fifth character as an increment to the existing pixel value (128 corresponds to the total number of ASCII codes). Finally, $I$ is normalized by dividing on its maximum cell value.

The result of the proposed transform can be visualized with grayscale images (Fig. 6). Despite of their sparseness, we expect high performances by a CNN classifier due to their proven effectiveness when coupled with sparse encoded features [28].

## 3.2 CNN model

We use a Convolutional Neural Network (CNN) to produce the labels of a video transcript image from labels among the 13 video topics (Table 1). As illustrated in Fig. 7, our CNN has a Zero Padding layer to make sure that all transcript image pixels are considered, 3 convolutional layers with ReLU non-linearity, 1 fully connected layer with 128 neurons and ReLU non-linearity, and 1 output layer with 13 neurons and Softmax non-linearity. The 13 output neurons are activated correspondingly to the video labels. Dropout layers are used to prevent overfitting [25]. An Adamax optimizer [14] is used with a learning rate of 0.002, and a categorical cross entropy is used as a loss function.

**Algorithm 1** Transcript to image transform

**Precondition:** Transcript file $T$ : array of characters

**function** TranscriptToImage($T$)
  define $I$ : $128 \times 128$ matrix
  **for** $i \leftarrow 1$ to $length(T)$ **do**
    $C \leftarrow \{T[j \bmod length(\mathrm{T})], j = i, ..., i+4\}$
    $x \leftarrow$ | ASCII ( $C[1]$ ) - ASCII ( $C[0]$ ) |
    $y \leftarrow$ | ASCII ( $C[3]$ ) - ASCII ( $C[2]$ ) |
    $v \leftarrow$ ASCII ( $C[4]$ )
    $I[x, y] \leftarrow I[x, y] + v$
  **end for**
  $I \leftarrow \frac{1}{max(I)} I$
  **return** $I$
**end function**

# 4 EXPERIMENTAL RESULTS

## 4.1 Comparison baseline

We compare our model with a baseline algorithm that uses shallow features [5]. The baseline method works as follows: The video transcript is generated using the CMU Sphinx toolkit [16]. Then, a vector of word frequencies is generated for each transcript. The vector is initially 354,986 dimensional corresponding to a list of most used English words[2], then it is reduced to a 7,937 dimensional vector by storing only the words that exist more than once in the training dataset's videos. The vector is finally normalized by dividing on the total sum of frequencies.

[2] https://github.com/dwyl/english-words

**Figure 6: Examples of transcript images (pixel values are normalized by $\times 255$). The left image corresponds to a biology video, and the right one corresponds to a physics video.**

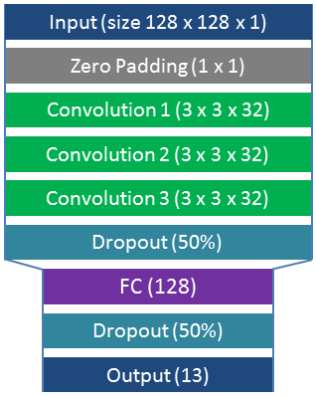


**Figure 7: Architecture of the CNN model.**

After extracting word frequencies features vectors, we experiment with different classifiers: multilayer perceptions (MLP), Decision Trees , K-Nearest Neighbors (K-NN), and Random Forests [22].

## 4.2 Evaluation metrics

Evaluation of our model against the baseline is done with two metrics: *Label Accuracy* (Eq. 1) expresses the ability of the model to correctly generate a label for a video, so it is penalized every time a single label is incorrect. The *Class Accuracy* (Eq. 2) expresses the ability of the model to correctly generate all the $N = 13$ labels to a video, which means that a classification is considered incorrect as soon as one label is incorrect.

$$Label\ Accuracy = \frac{1}{K}\sum_{k=1}^{K}(1 - \frac{|\overrightarrow{y_{test}} - \overrightarrow{y_{predicted}}|}{N}) \quad (1)$$

$$Class\ Accuracy = \frac{1}{K}\sum_{k=1}^{K} 1, \text{if } \overrightarrow{y_{test}} = \overrightarrow{y_{predicted}} \quad (2)$$

Where $K$ is the number of videos. Both metrics are in $[0, 1]$ and the larger the values the better the performances. Naturally, *Class Accuracy* would be inferior to *Label Accuracy*. For instance, if a video is classified as 0100000000000 while its ground truth is 0100000010000, *Label Accuracy* $= \frac{12}{13} = 0.92$ while *Class Accuracy* $= 0$.

## 4.3 Results

Fig. 8 shows the progress of CNN and MLP models training. The performance difference is more noticeable with the *Class Accuracy* metric than with the *Label Accuracy* metric. The proposed model outperforms the baseline using an MLP in terms of *Label Accuracy* slightly and in terms of *Class Accuracy* with more than 9%. After 50 iterations, the baseline shows overfitting and its performance decreases. Table 2 shows best performances of all models that we experimented with. Our model outperforms all the baselines, and the best baseline performances were obtained by an MLP that has 7,937 input neurons, 1 hidden layer with 1024 neurons, and 13 neurons in the output layer. A part from the CNN and MLP, other baselines using Decision Trees, K-NN and Random Forest have not given satisfactory performances. Trying different configurations of the baseline models, including the MLP (by adding hidden layers and neurons) has not led to improved performances.

It is worth noting that performances of our method are high despite of the imperfection of the speech recognition tool [16] that can cause a word error rate (WER) as high as 26.9% [10]. Given that the KAS dataset is currently single speaker, we explain our model's high performances by the fact that the transcript errors would be associated with certain video topic labels and lead to correct classifications despite of word mistakes.

We also confirmed our hypothesis of using speech transcript instead of using frames. We trained an Alexnet [15] that takes single video frames and produce topic labels, and we prepared a dataset of 1,263,227 frames by extracting equidistant keyframes in 1s intervals from the KAS videos. Frames were resized to $160 \times 112$ in order to overcome the high resolution differences (Fig. 5). This classifier gave *Class Accuracy* and *Label Accuracy* values below 0.5, which is way inferior to the models using transcripts.

# 5 CONCLUSION AND FUTURE WORK

In this paper, we presented a method for automatic classification of MOOC videos. Our method works as follows: First, it extracts video transcripts using a standard toolkit. Then, it converts the transcripts into images using a statistical co-occurrence transform. Finally, it uses a convolutional neural network (CNN) to generate the video topic labels.

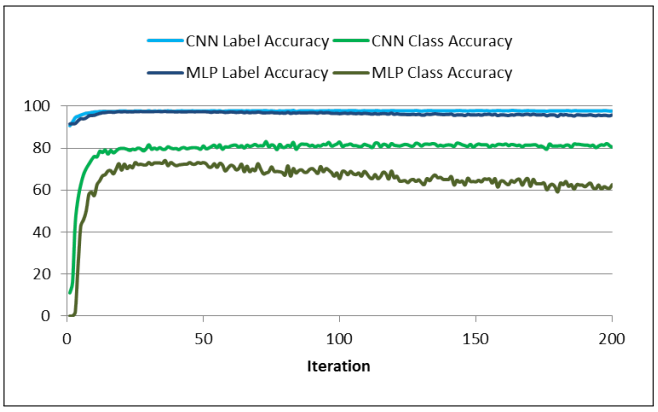


**Figure 8: Metrics values during models training.**

**Table 2: Classifier performances. Best ones are the CNN model (after 73 training iterations) and the MLP model (after 36 training iterations).**

| Model | *Label Accuracy* (100%) | *Class Accuracy* (100%) |
|---|---|---|
| CNN | 97.87% | 83.10% |
| MLP | 97.53% | 74.08% |
| Decision Trees | 90.71% | 37.42 |
| K-NN (K=3) | 88.51% | 22.64% |
| Random Forest | 91.87% | 6.33% |

To evaluate our method, we use the Khan Academy on a Stick (KAS) dataset and evaluate our model against a baseline that is based on transcript word frequency feature vectors. Results demonstrate the effectiveness of our approach compared to the baseline with significant performance improvement.

This paper reports one module of our ongoing work to involve content-based indexing in the MOOC videos in order to enable more intuitive video retrieval. As a future work, we will evaluate our method using datasets collected from different MOOC platforms, and we will work towards content-based video retrieval using sketches and audio keywords.

## REFERENCES


[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Olaoluwa Adigun and Bart Kosko. 2017. Using noise to speed up video classification with recurrent backpropagation. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 108–115.

[3] Khaled M Alraimi, Hangjung Zo, and Andrew P Ciganek. 2015. Understanding the MOOCs continuance: The role of openness and reputation. *Computers & Education* 80 (2015), 28–38.

[4] Subhasree Basu, Yi Yu, and Roger Zimmermann. 2016. Fuzzy clustering of lecture videos based on topic modeling. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–6.

[5] Darin Brezeale and Diane J Cook. 2008. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C* 38, 3 (2008), 416–430.

[6] Joseph Chapes. 2017. Online Video in Higher Education: Uses and Practices. In *EdMedia: World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education (AACE), 1133–1138.

[7] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. 2009. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1932–1939.

[8] Gayle Christensen, Andrew Steinmetz, Brandon Alcorn, Amy Bennett, Deirdre Woods, and Ezekiel J Emanuel. 2013. The MOOC phenomenon: Who takes massive open online courses and why? *University of Pennsylvania. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2350964*. (2013).

[9] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE, 886–893.

[10] Christian Gaida and others. 2014. Comparing open-source speech recognition toolkits. (2014).

[11] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 41–50.

[12] Iveel Jargalsaikhan, Suzanne Little, Remi Trichet, and Noel E O'Connor. 2015. Action recognition in video using a spatial-temporal graph-based feature representation. In *International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

[13] Andrej Karpathy and others. 2014. Large-scale video classification with convolutional neural networks. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 1725–1732.

[14] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[16] Paul Lamere and others. 2003. The CMU SPHINX-4 speech recognition system. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 1. IEEE, 2–5.

[17] Ivan Laptev. 2005. On space-time interest points. *International Journal of Computer Vision* 64, 2-3 (2005), 107–123.

[18] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1–8.

[19] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor. 2016. Holistic features for real-time crowd behaviour anomaly detection. In *International Conference on Image Processing (ICIP)*. IEEE, 918–922.

[20] Michael Noer. Accessed on 19/07/2017. One Man, One Computer, 10 Million Students: How Khan Academy Is Reinventing Education. www.forbes.com/sites/michaelnoer/2012/11/02/one-man-one-computer-10-million-students-how-khan-academy-is-reinventing-education. (Accessed on 19/07/2017).

[21] Laura Pappano. 2012. The Year of the MOOC. *The New York Times* 2, 12 (2012), 2012.

[22] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

[23] Hijung Valentina Shin, Floraine Berthouzoz, Wilmot Li, and Frédo Durand. 2015. Visual transcripts: lecture notes from blackboard-style lecture videos. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 240.

[24] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[25] Nitish Srivastava and others. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[26] Clive Thompson. 2011. How Khan Academy is changing the rules of education. *Wired Magazine* 126 (2011), 1–5.

[27] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3169–3176.

[28] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. 2015. Deep networks for image super-resolution with sparse prior. In *International Conference on Computer Vision (ICCV)*. 370–378.

[29] Joe Yue-Hei Ng and others. 2015. Beyond short snippets: Deep networks for video classification. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 4694–4702.