MultiEdTech 2017

1st International Workshop on Educational and Knowledge Technologies, ACM Multimedia Conference
October 23-27, 2017, Mountain View, CA, USA

Automatic MOOC video classification using transcript features and convolutional neural networks

Authors

Houssem Chatbri, Kevin McGuinness, Suzanne Little, Jiang Zhou, Noel E. O'Connor (Dublin City University, Ireland)

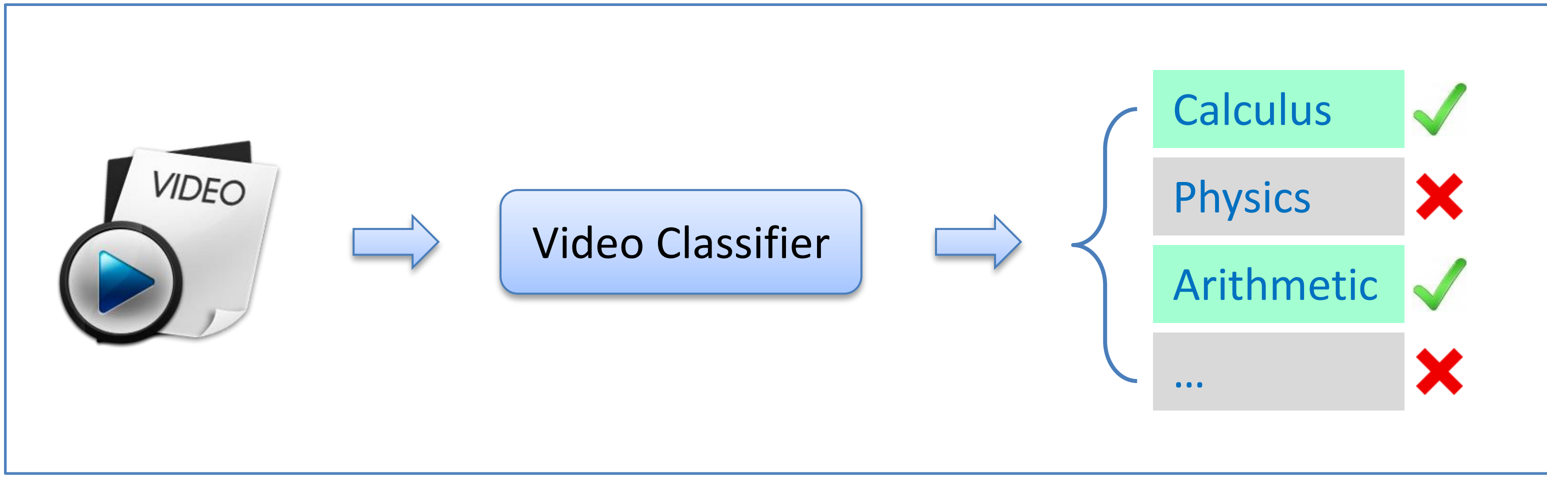
Keisuke Kameyama (Univ. Tsukuba, Japan)

Paul Kwan (University of New England, Australia)

Goal

Input: MOOC video

Output: Topic labels



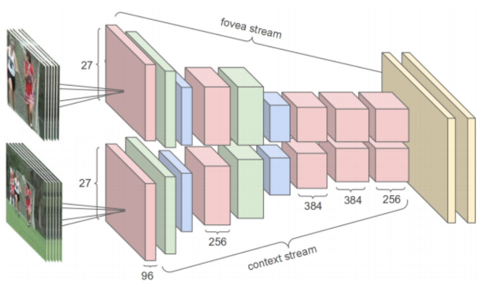
Advantages

- Automatic indexing
- Semantic retrieval

Existing methods

End to end systems

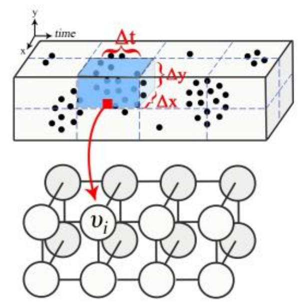
Action recognition using convolutional neural networks (CNN) [Karpathy et al. CVPR 2014]



A stack of frames is used as input to a CNN with two separate processing streams. Each stream process a different resolution.

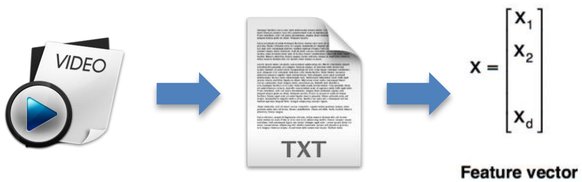
Feature extraction based systems

Action recognition using graphs of frame features [Jargalsaikhan et al. AVSS 2015]



Spatiotemporal local features are extracted to build a feature graph. Then, an SVM classifier is used.

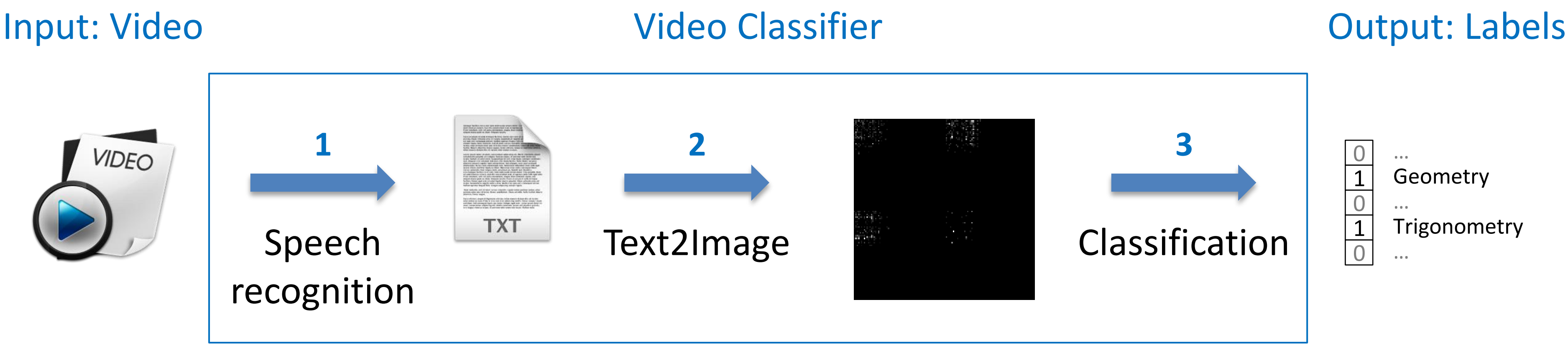
Systems that transform the problem domain



Educational video classification using transcripts [Brezeale and J Cook, IEEE Trans. SMC 2008]

Word frequencies are used to calculate feature vectors.

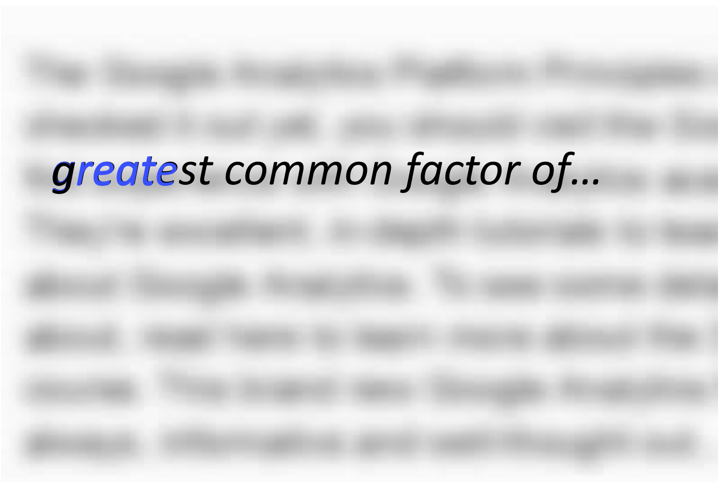
Pipeline



1. Speech recognition using the CMU Sphinx toolkit
2. Synthetic feature image (SFI) generation using a co-occurrence transform
3. Classification using a convolutional neural network (CNN)

Text2Image transform

Video transcript



Use ASCI values to fill a 2D matrix

x: ASCII(‘**e**’) - ASCII(‘**r**’)
y: ASCII(‘**t**’) - ASCII(‘**a**’)
v: ASCII(‘**e**’)

Text2Image transform

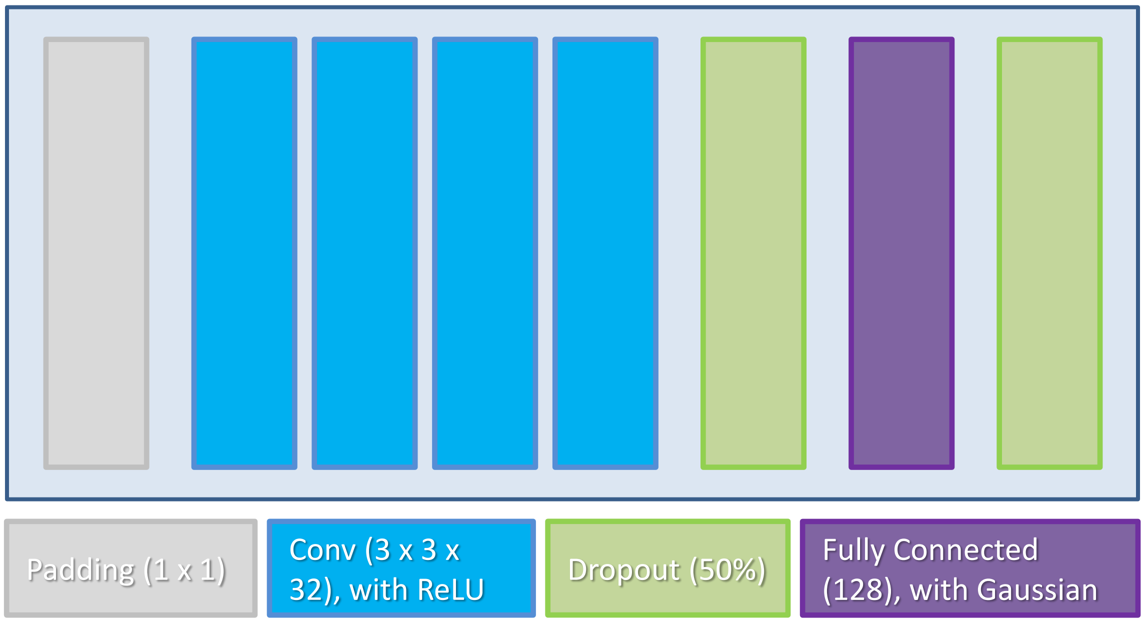
Video tran

matrix

greatest common

Classifier

Input	*Model*	*Output*
Transcript image	A Convolutional Neural Network (CNN)	Class labels



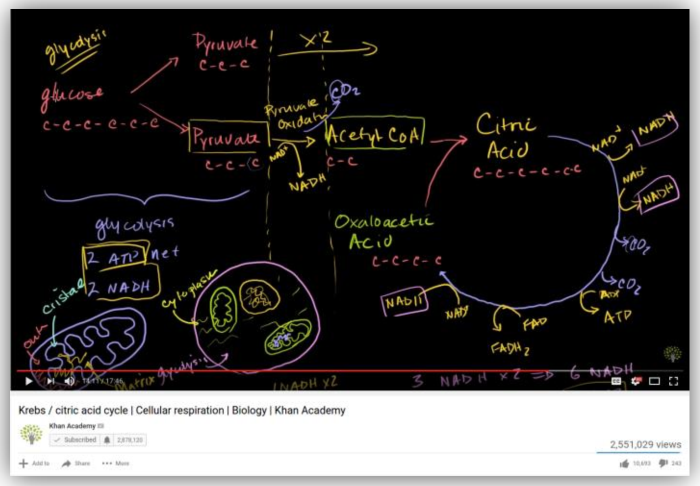
0	Algebra
0	Arithmetic
1	Geometry
0	Probability
0	Calculus
0	Differential Equations
1	Trigonometry
0	Biology
0	Cosmology and Astronomy
0	Organic Chemistry
0	Chemistry
0	Healthcare and Medicine
0	Physics

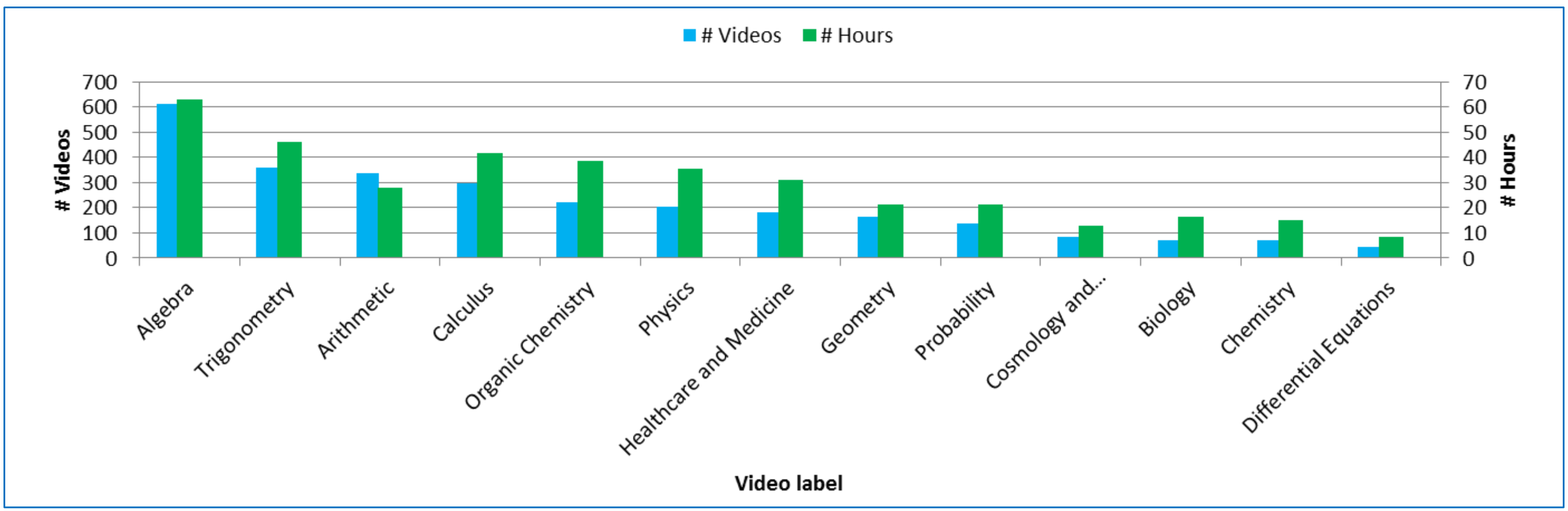
Parameters

Optimiser	Learning rate	Loss function
Adamax	0.002	Categorical Cross Entropy

Data

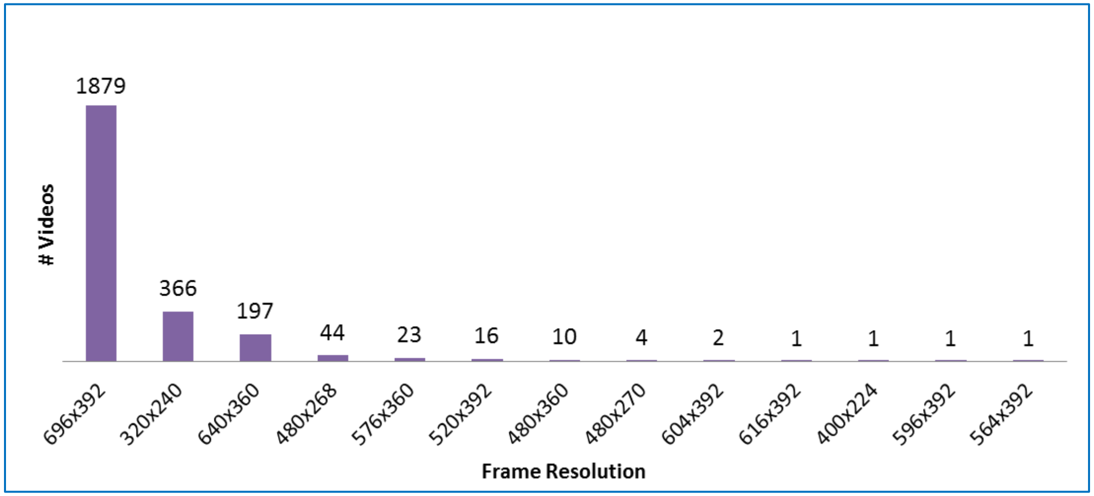
- The “Khan Academy on a Stick” public dataset
- 2,545 videos recorded from 2006 to 2013



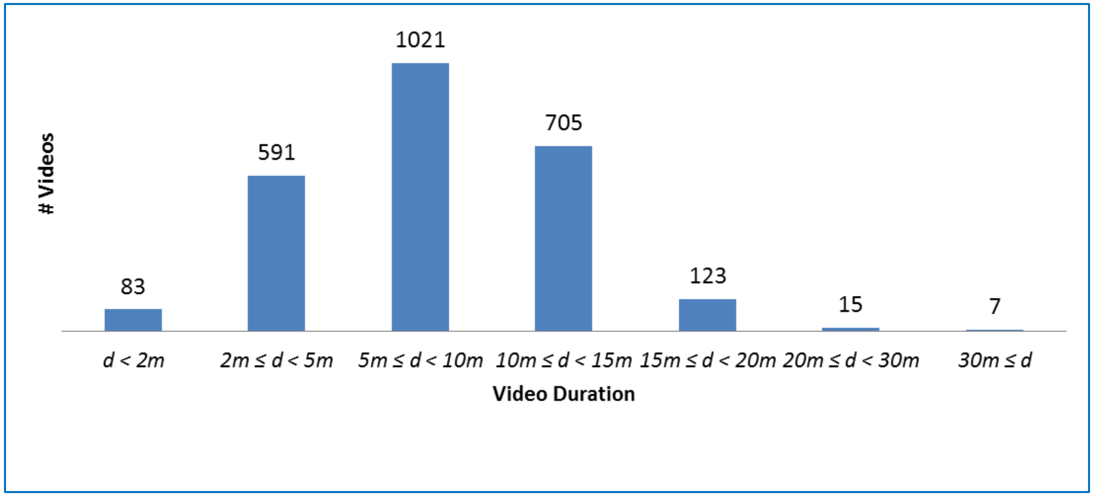


Video statistics per class. Total of 2,545 videos running 380 hours

Data



Histogram of video frame resolutions



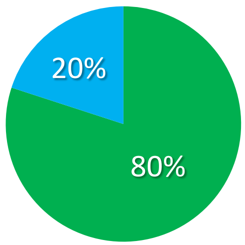
Histogram of video durations

- Variance of the frame resolution due to sketching tablet change

- Variance of video duration

Data split

Training and testing subsets of videos



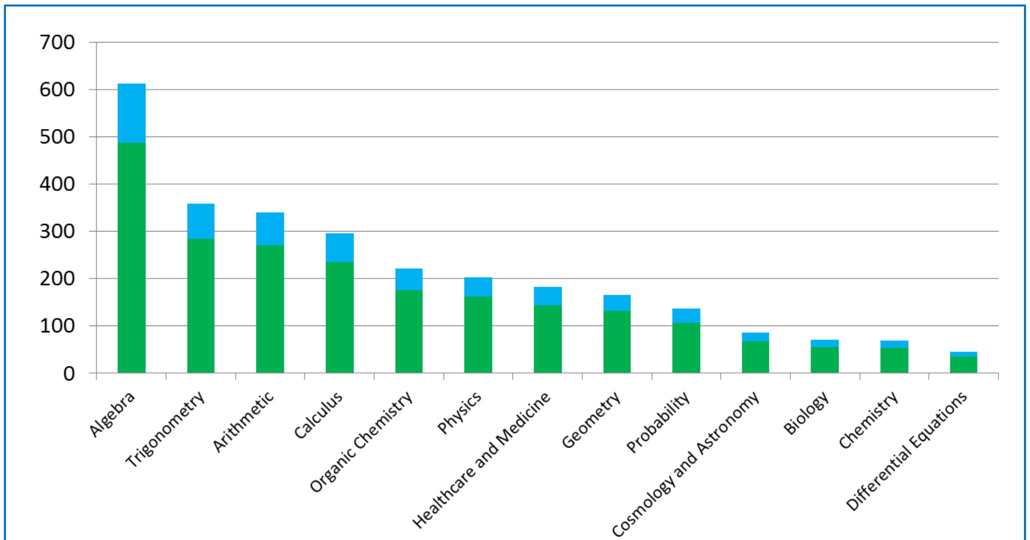
<  ,

0	…
1	Geometry
0	…
1	Trigonometry
0	…

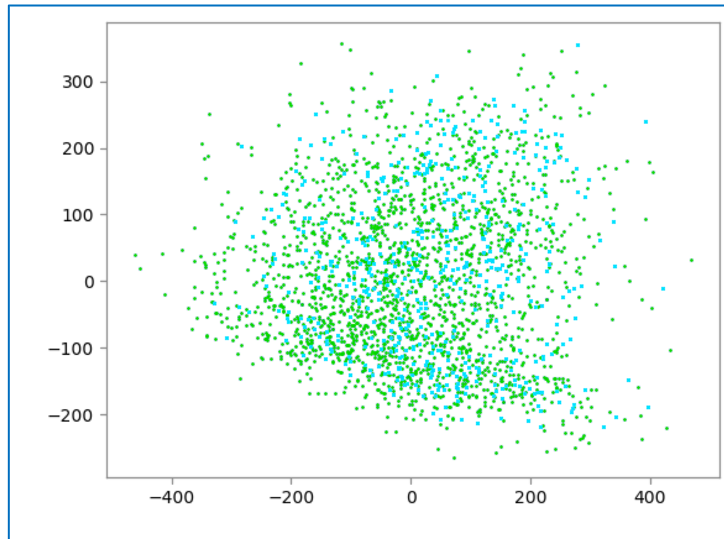
>

Input Output

Video statistics per class (2,545 videos running 380 hours)



Projection of the training and testing sets using PCA extracted from SFIs



Metrics

$$Label\ Accuracy = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{|\overrightarrow{y^k_{test}} - \overrightarrow{y^k_{predicted}}|}{N}\right)$$

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

K: Number of images, N: 13 labels, y: output

Prediction	0	0	1	0	0	0	0	0	0	0	1	0	0
Ground truth	0	0	1	0	0	0	0	0	0	0	1	0	0

Label Accuracy = 1.0

Prediction	0	0	1	0	0	0	0	0	0	0	1	0	0
Ground truth	0	0	1	0	0	0	0	0	0	0	1	0	0

Class Accuracy = 1.0

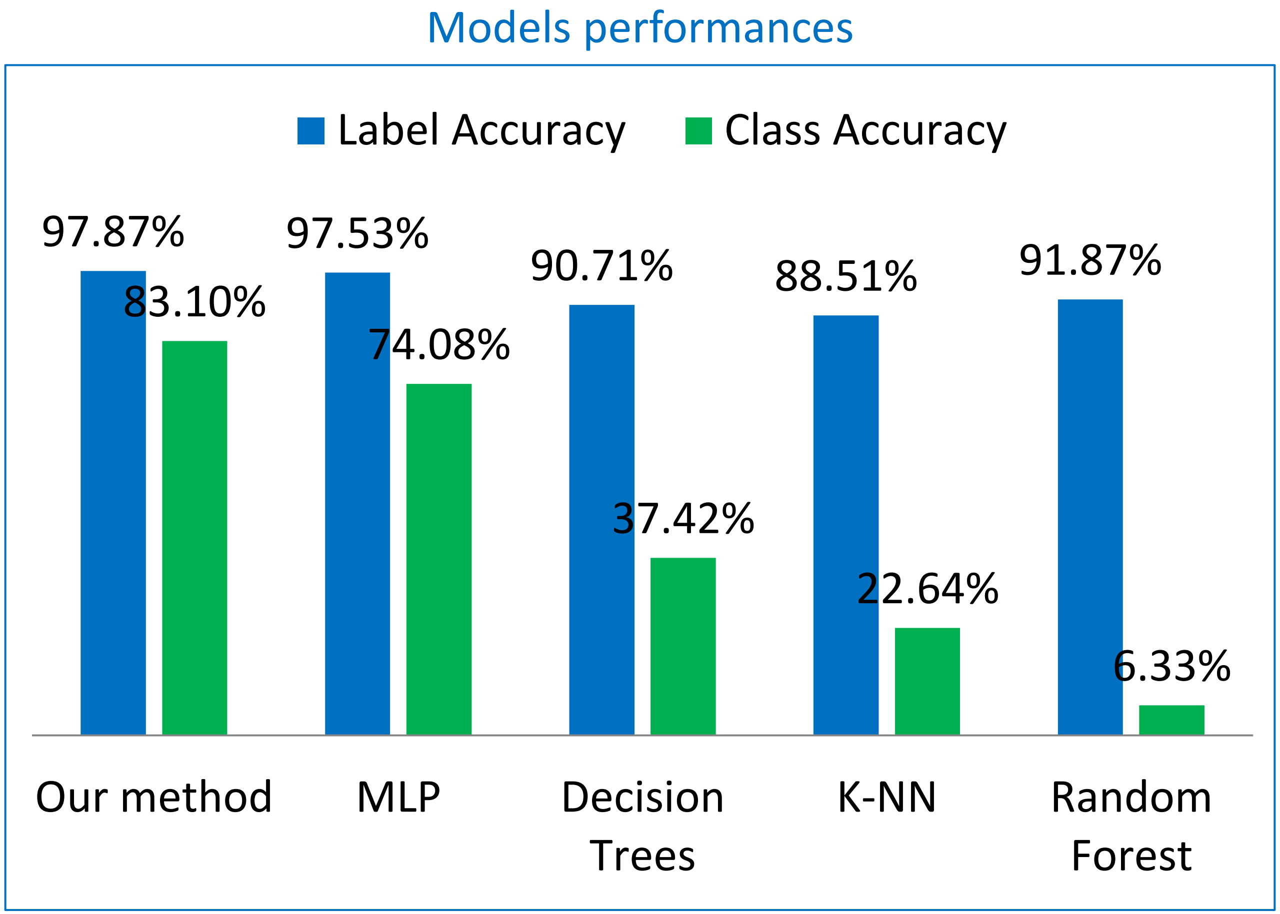
Prediction	0	0	1	0	0	1	0	0	0	0	1	0	0
Ground truth	0	0	1	0	0	0	0	0	0	0	1	0	0

Label Accuracy = 0.92

Prediction	0	0	1	0	0	1	0	0	0	0	1	0	0
Ground truth	0	0	1	0	0	0	0	0	0	0	1	0	0

Class Accuracy = 0.0

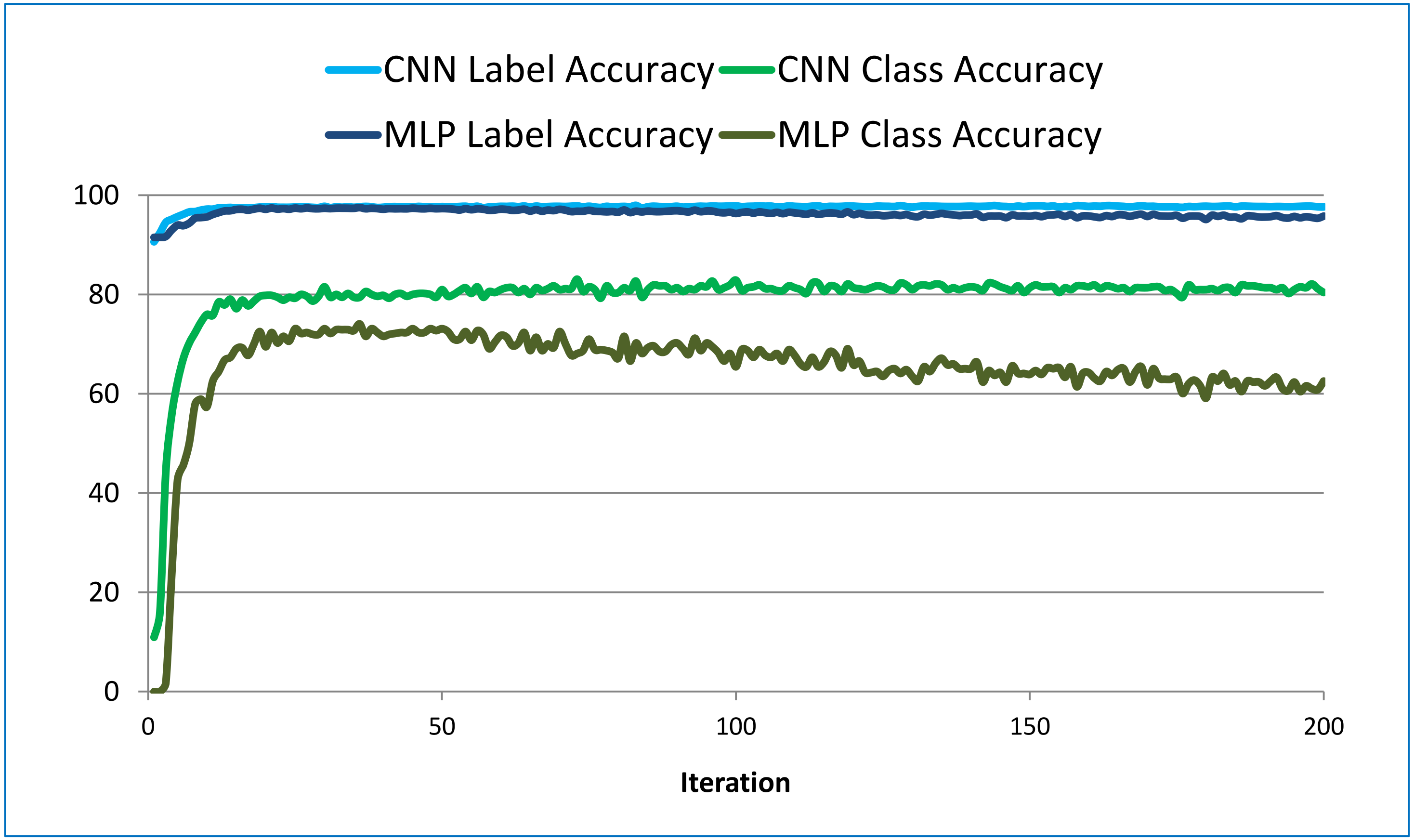
Results



- MLP, Decision Trees, K-NN, and Random Forest are fed feature vector of word frequencies [B and Cook, IEEE Trans. SMC 2008]

Results

Performances of CNN and MLP during training

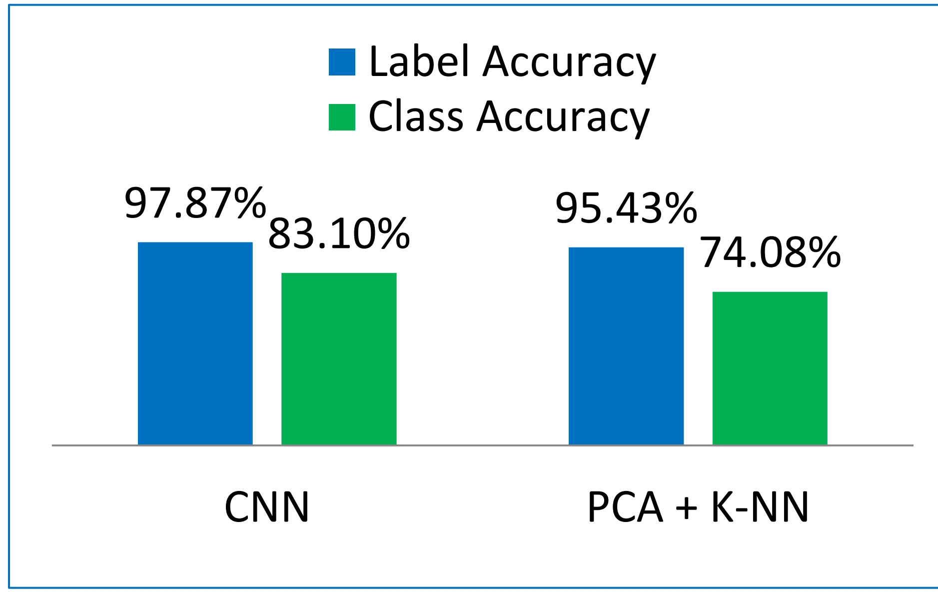


- MLP, Decision Trees, K-NN, and Random Forest are fed feature vector of word frequencies [B and Cook, IEEE Trans. SMC 2008]

Results

Why does the model work despite of sparse inputs?

Synthetic feature image



Classifier performances

- ✓ CNNs perform well on sparse inputs, supporting results by Wang et al. ICCV 2015

Results

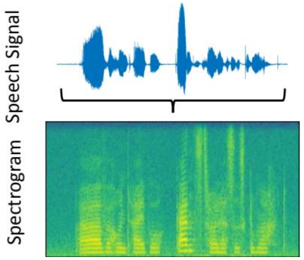
Why does the model work despite of indistinctive data?

SFI

- Synthetic feature images (SFI) from low-level text features are used to train a CNN with a high accuracy

✓ Our results support *Cummins et al.* and *Zhang et al.'s*

Speech spectrograms



- *Cummins et al. (INTERSPEECH 2017, ACM MM 2017)* have use a pre-trained CNN on image spectrums

Random image

- *Zhang et al. (ICCV 2016)* have trained a CNN with an ImageNet-size of random pixels images. The model memorised hem with high accuracy

Contribution

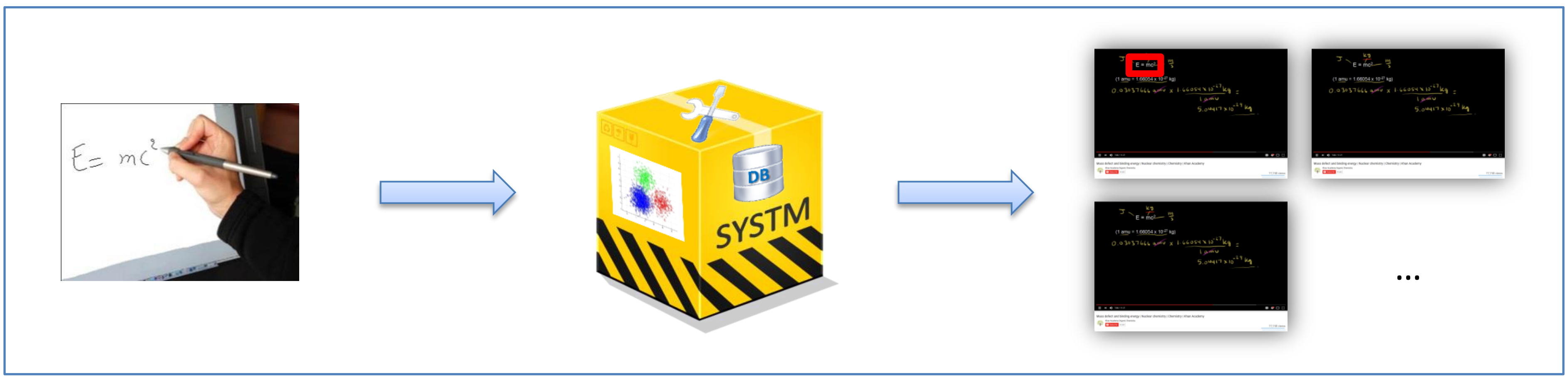
- Improving the state of the art in educational video classification
- Supporting research in deep learning using sparse and synthetic images

Future work

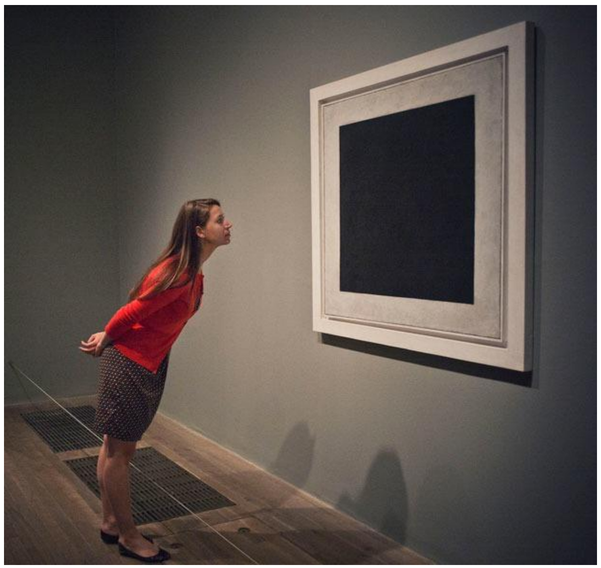
Input: Sketch query

Video retrieval system

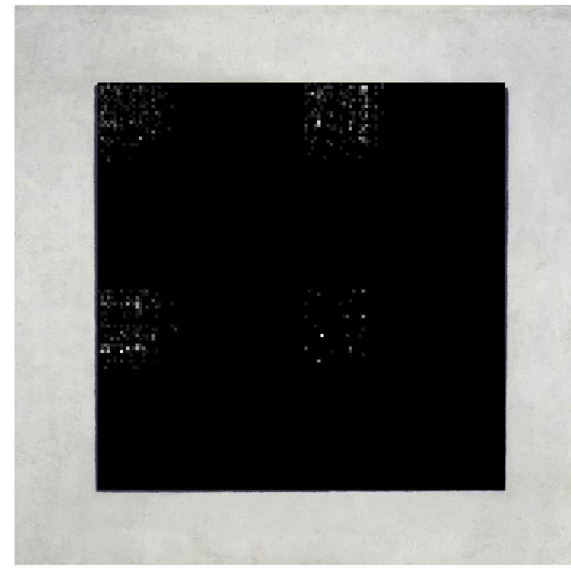
Output: Khan Academy videos



Modern Art!



Malevich Square
Kazimir Malevich (1915)



SFI Square
Dublin City University (2017)

References

Video classification methods

End-to-end methods

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. FeiFei. Large-scale video classification with convolutional neural networks. CVPR, 2014

Feature extraction based methods

S. Basu, Y. Yu, and R. Zimmermann. Fuzzy clustering of lecture videos based on topic modeling. IEEE CBMI, 2016

CNN architecture and settings

Adamax optimiser

D. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv, 2014

Reducing overfitting with dropouts

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014

Deep learning with synthetic feature images

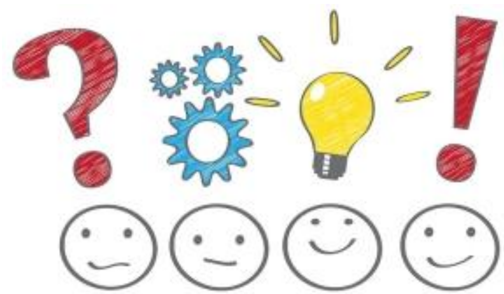
CNN trained on an ImageNet-sized dataset of random images

Hang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." *arXiv,* 2016

DNN trained with image spectrum images

Nicholas Cummins. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. ACM MM 2017

Thanks for your attention!



Acknowledgment

Insight Centre for Data Analytics

sfi Science Foundation Ireland

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn