CORSMAL
Collaborative object recognition, shared manipulation and learning

Audio classification of the content of food containers and drinking glasses

Santiago Donaher, Alessio Xompero, Andrea Cavallaro

European Signal Processing Conference, 23-27 August 2021

Queen Mary University of London

http://corsmal.eecs.qmul.ac.uk/audio_classification.html

CIS centre for intelligent sensing

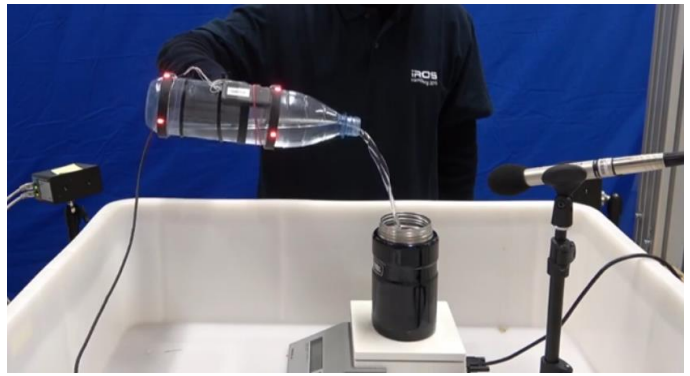
What action? What content? How much content?



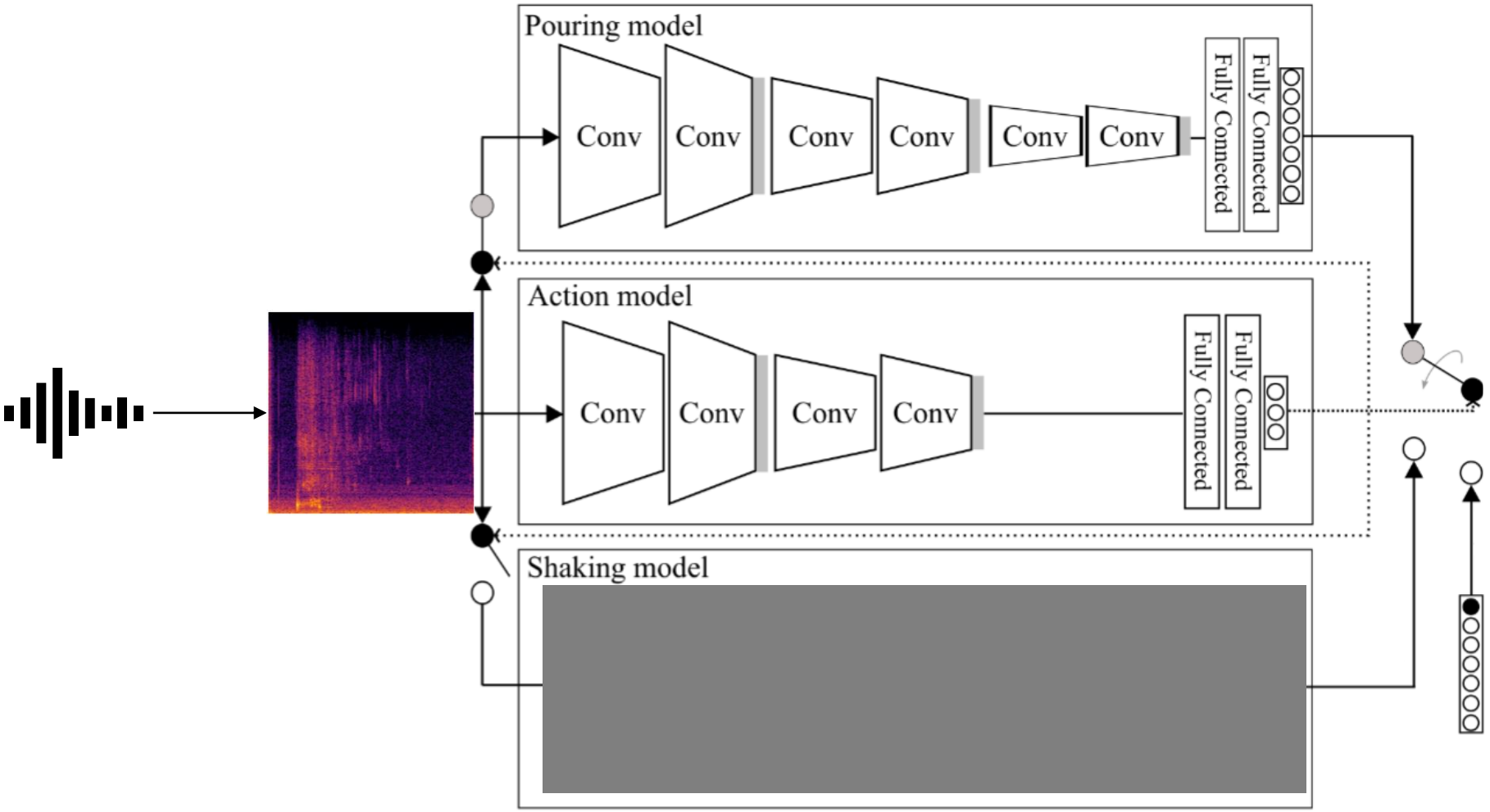
Challenges: unknown containers (shape, size, material), different contents, unknown actions (duration, varying-pose of the container), different background noises, room reverberations, varying distances to microphones.

Literature:

- No held or manipulated containers [Clarke2018]
- Only pouring of liquid for filling level estimation [Liang2019]

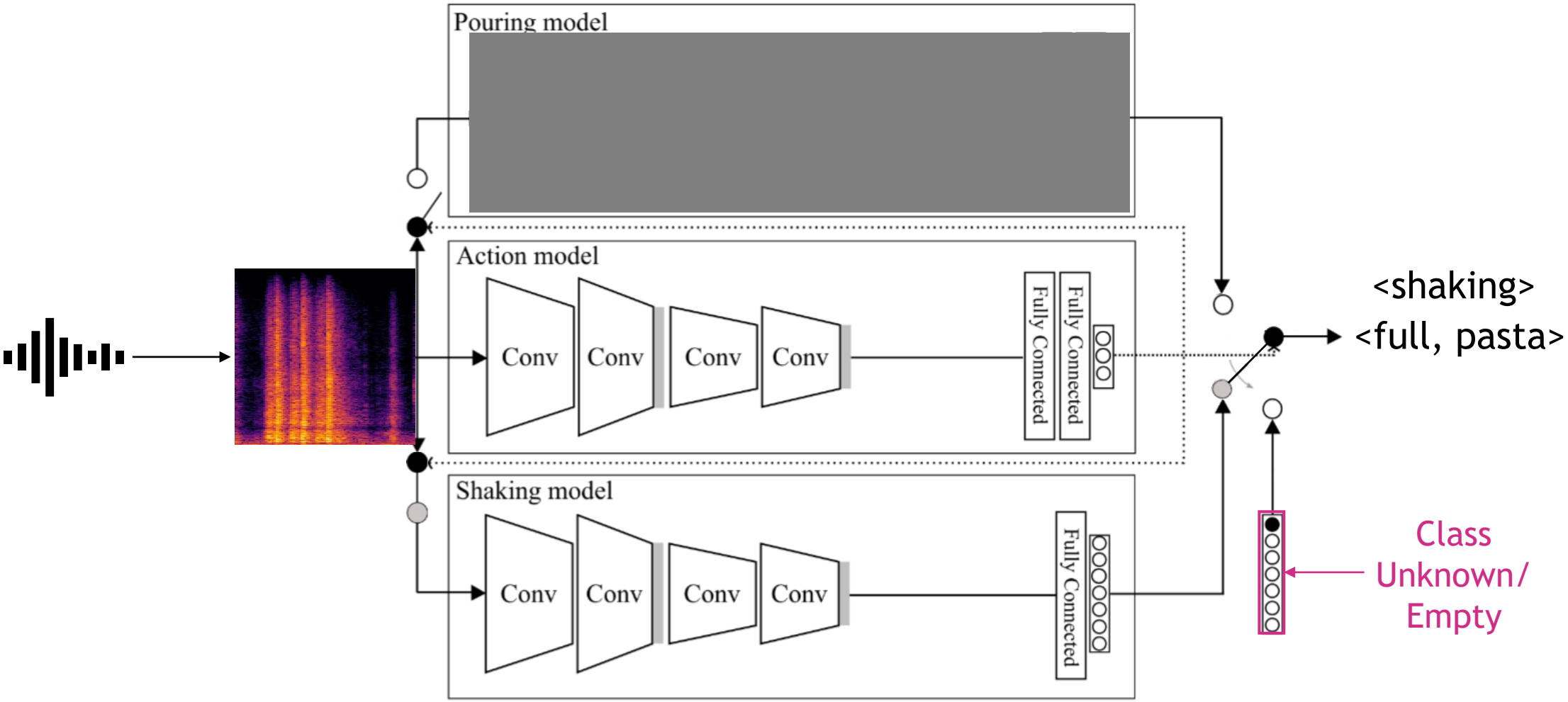


Audio Classification of Content (ACC)



<pouring>
<full, water>

Audio Classification of Content (ACC)



Datasets

CORSMAL Containers Manipulation (CCM)

1,140 audio recordings
840 pouring actions
300 shaking actions

Circular array of 8 Boya BY-M1 omnidirectional Lavelier microphones (15cm radius), at 0.5 to 3m from the action
44.1 kHz

Audio-based Containers Manipulation Setup 2

21 audio recordings
19 pouring actions
2 shaking actions

Blue Yeti Studio microphone on a table, at 12 cm for pouring actions, and 20 cm for shaking actions. 48 kHz

Different room environment and reverberation

7 filling types & level: *empty, half-full pasta, full pasta, half-full rice, full rice, half-full water, full water*

https://corsmal.eecs.qmul.ac.uk/containers_manip.html

https://zenodo.org/record/4770439#.YNpA3CAYBPZ

Experimental setup

AUDIO PRE-PROCESSING

- Down-sampling: 22,050 Hz
- Multi-to-mono channel conversion (averaging)
- Amplitude normalised to [-1, 1]
- Onset detection to identify the beginning of the action
- Spectrograms: trimmed to a fixed window (10 s)
- Resize to 96x96 input for CNNs

Weighted average F1-score

Number of classes

$$\overline{F_1} = \frac{1}{R}\sum_{n=1}^{|C|} R_n F_n$$

F1-score of class *n*

Total number of recordings

CORSMAL Containers Manipulation (CCM) splits:
648 train (9 containers) -> random split into train/val (80/20)
228 public test (3 containers)
228 private test (3 containers)

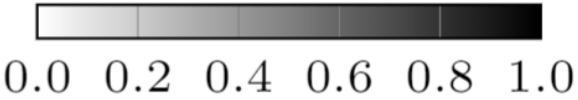
Training:
Each classifier independently, for 100 epochs
Categorical Cross-entropy loss, Adam optimiser with a Learning Rate of 0.001
Action model trained with all training recordings (518), Pouring model with pouring recordings (384), and Shaking model with shaking recordings (144)

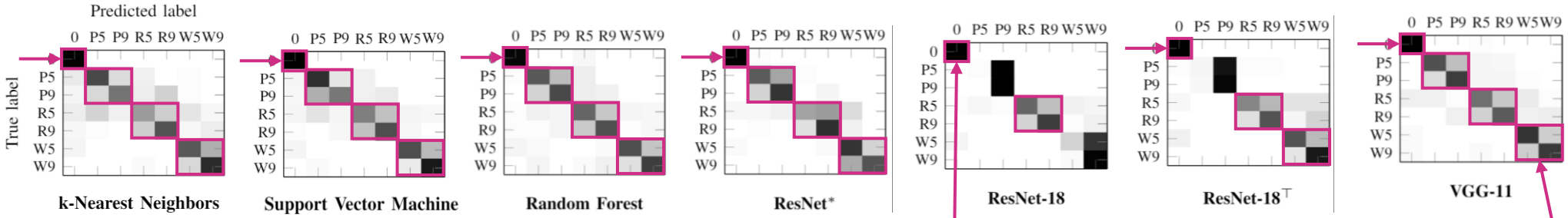
Accuracy ($\overline{F_1}$) and complexity comparison

Best score
Second best score

	Model	# Params. (x1000)	Storage (MB)	CCM Public Test	CCM Private Test	ACM2	
	Random	-	-	9.57	11.22	15.24	
Direct classification (no action)	Resized spectr. + kNN	-	63.5	70.33	62.99	19.09	
	Resized spectr. + SVM	-	34.8	69.81	73.43	26.98	
	Resized spectr. + RF	-	1.7	73.59	70.48	20.07	
	VGG-11	44,907	175.4	74.03	72.67	15.24	
	Lightweight ResNet	**179**	**0.8**	70.45	71.56	*34.95*	
	ResNet-18	11,692	45.8	62.28	55.28	30.87	
	Pre-trained ResNet-18	11,692	45.8	59.69	63.01	32.02	
Independent filling type and level classifications	Audio features + RF, Spect. + CNN + GRU, R(2+1)D (video), Fusion	-	-	75.00	77.86	-	[Iashin2020ICPRw]
	CNN + LSTM (level) CNN + Voting (type)	6,839	82.1	**82.14**	73.40	-	[Ishikawa2020ICPRw]
	ACC	16,482	64.3	*76.02*	**78.24**	**41.89**	

Confusion matrices

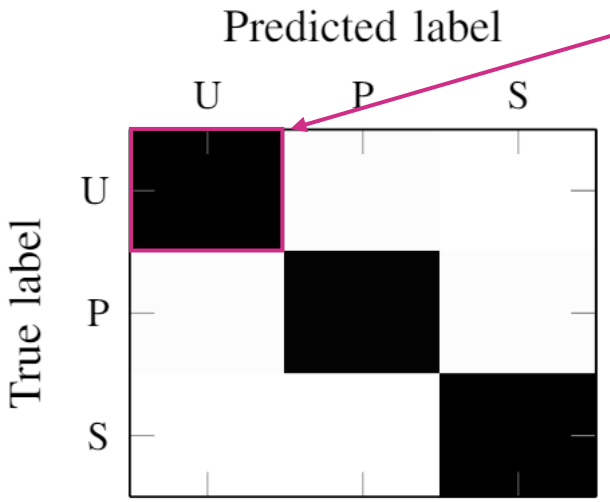




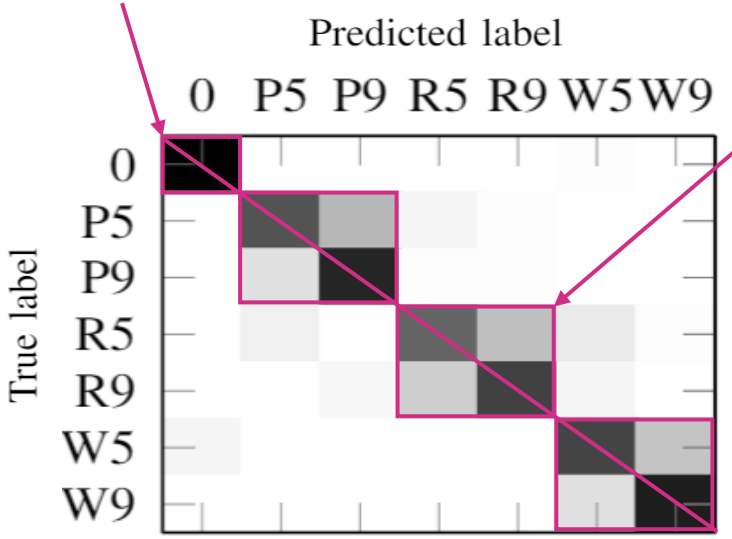
Unknown / Empty

Same type, both levels

U: Unknown action
P: Pouring action
S: Shaking action



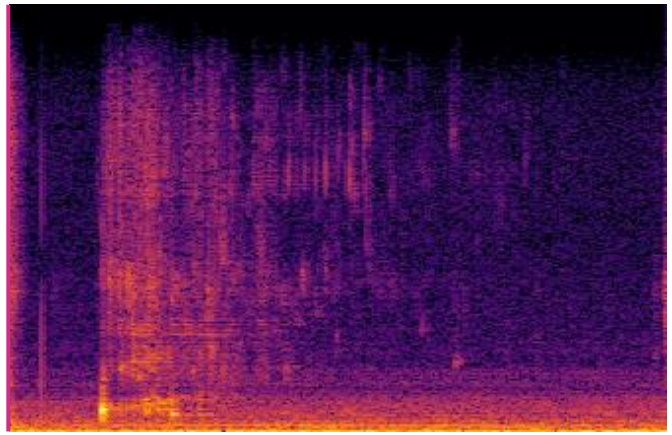
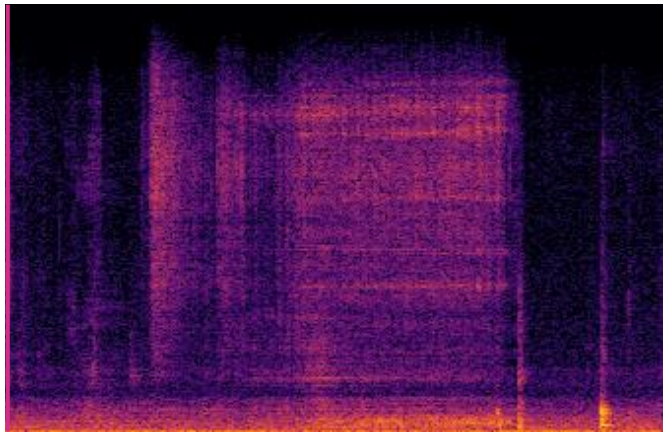
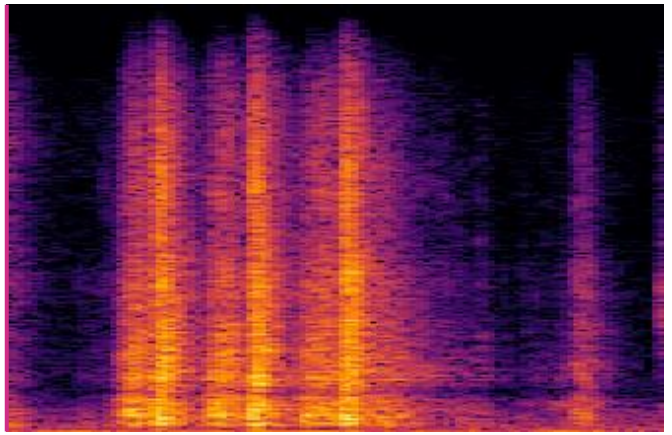
Action model



ACC

0: Empty
P5: half-full, pasta
P9: full, pasta
R5: half-full, rice
R9: full, rice
W5: half-full, water
W9: full, water

Conclusions

- Sound-based classification of content type and level in unknown food boxes and drinking glasses **handled by a person**

- Two-step model
  1. action recognition (pouring/shaking/unknown)
  2. jointly classification of content type and level with action-specific classifier

- Future work:
  generalization to other setups and environments
  multi-modal data

This work is supported by the CHIST-ERA programme through the project CORSMAL, under UK EPSRC grant EP/S031715/1