Rethinking genre classification with fine-grained semantic experts

Edward Fish, Jon Weinbren, Andrew Gilbert

UNIVERSITY OF SURREY

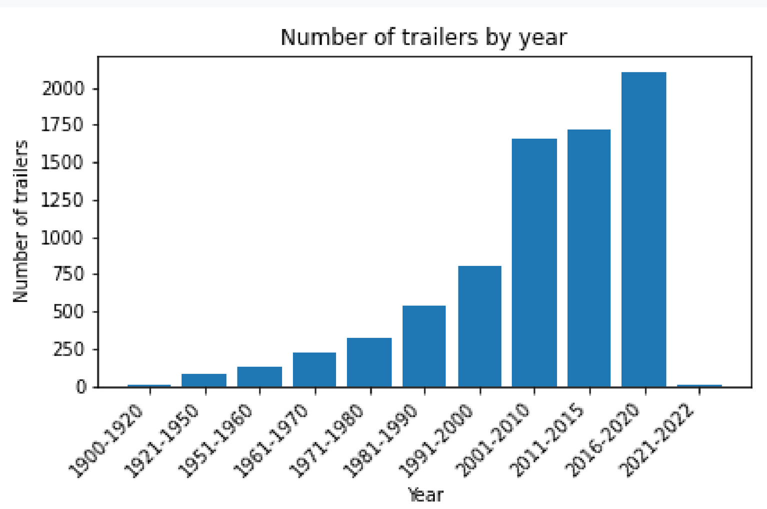
ICIP INTERNATIONAL CONFERENCE ON INFORMATION PROCESSING 2021

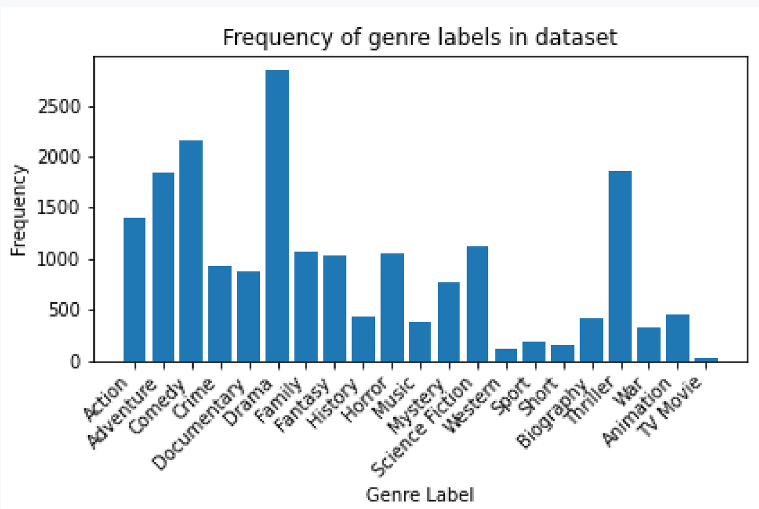
A self-supervised method for generating sub-genres from sparsely labelled movie data.

MMX-Trailer-20 Dataset

Available at ed-fish.github.io

- ~ 37 Million Frames
- ~ 8800 Movie Trailers
- Pre-computed expert embeddings
- 6 labels per sample





Dataset	Video Source	Number Trailers	Frames	Label Source	Num. Genres	Genre/ Trailer
Rasheed [34]	Apple	101	-	-	4	1
Huang [20]	Apple	223	-	IMDb	7	1
Zhou [50]	IMDb+Apple	1239	4.5M	IMDb	4	3
LMTD-9 [44]	Apple	4000	12M	IMDb	9	3
Moviescope [9]	IMDb	5000	20M	IMDb	13	3
MMX-Trailer-20	Apple+YT	8803	37M	IMDb	20	6

01 Overview

Genre labels are useful for conveying a general overview of the narrative and plot of a movie. But even with multi-label examples there can be large audio-visual semantic differences between movies with the same genre labels.

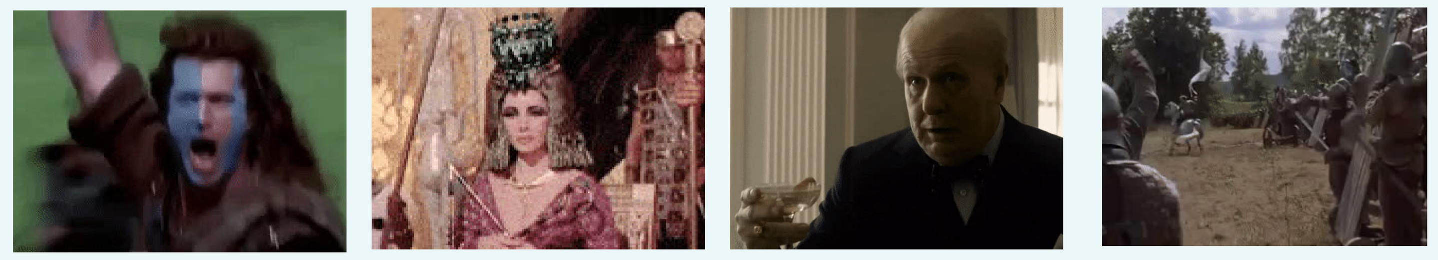


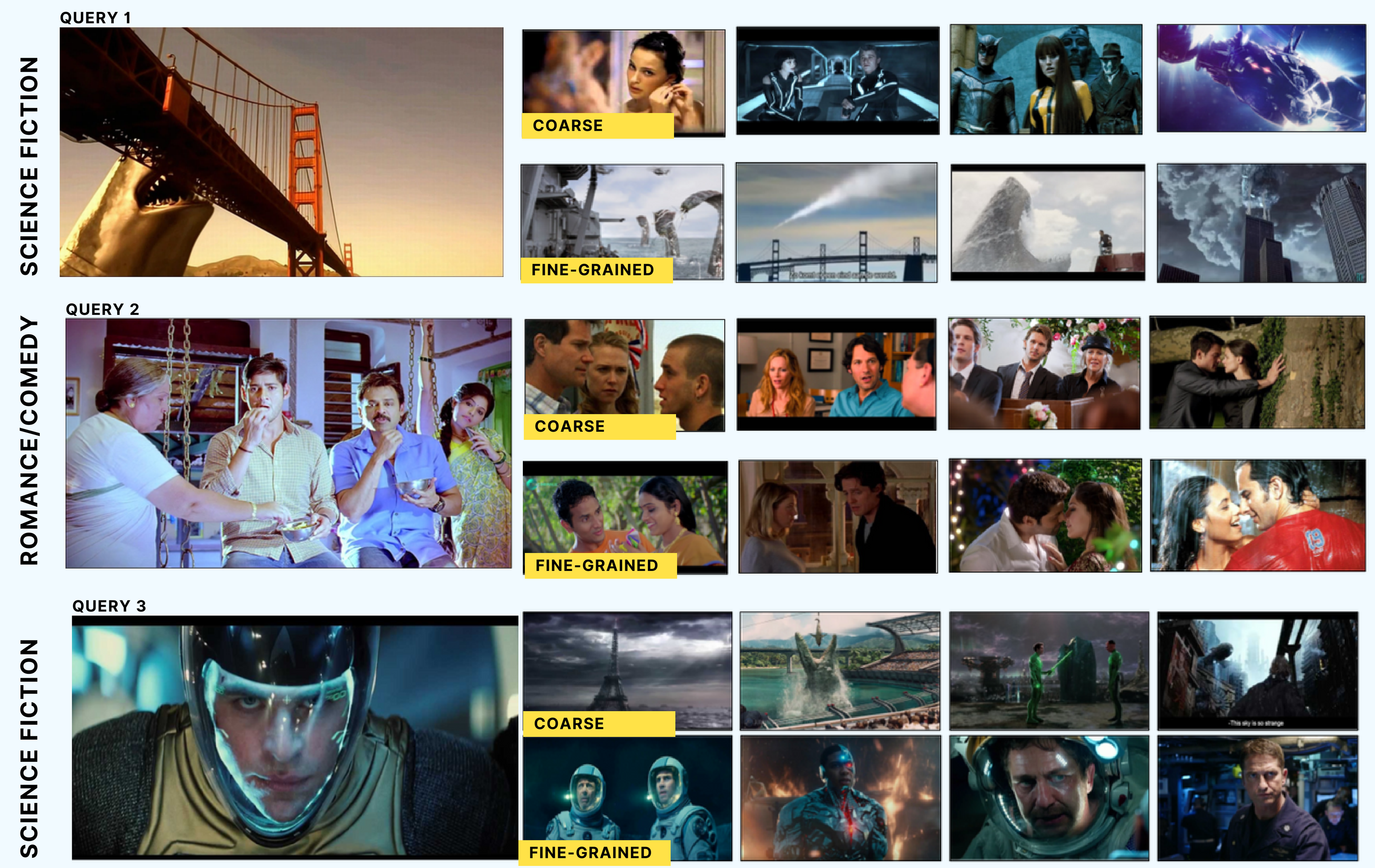
Fig 1. All the examples above share the same genre label combination **History, Biography, Drama,** but the audio-visual content differs between examples significantly.

We train a genre classification network using labels and extract the feature representations as shown below. This **coarse encoder network** groups movie trailers with the same labels closely together despite the variety of audio-visual content.

Then we continue to train the network self-supervised to find similarities in audio-visual content while still retaining some genre information. This leads to a **fine-grained embedding space** where clusters represent new sub-genres.

03 Results

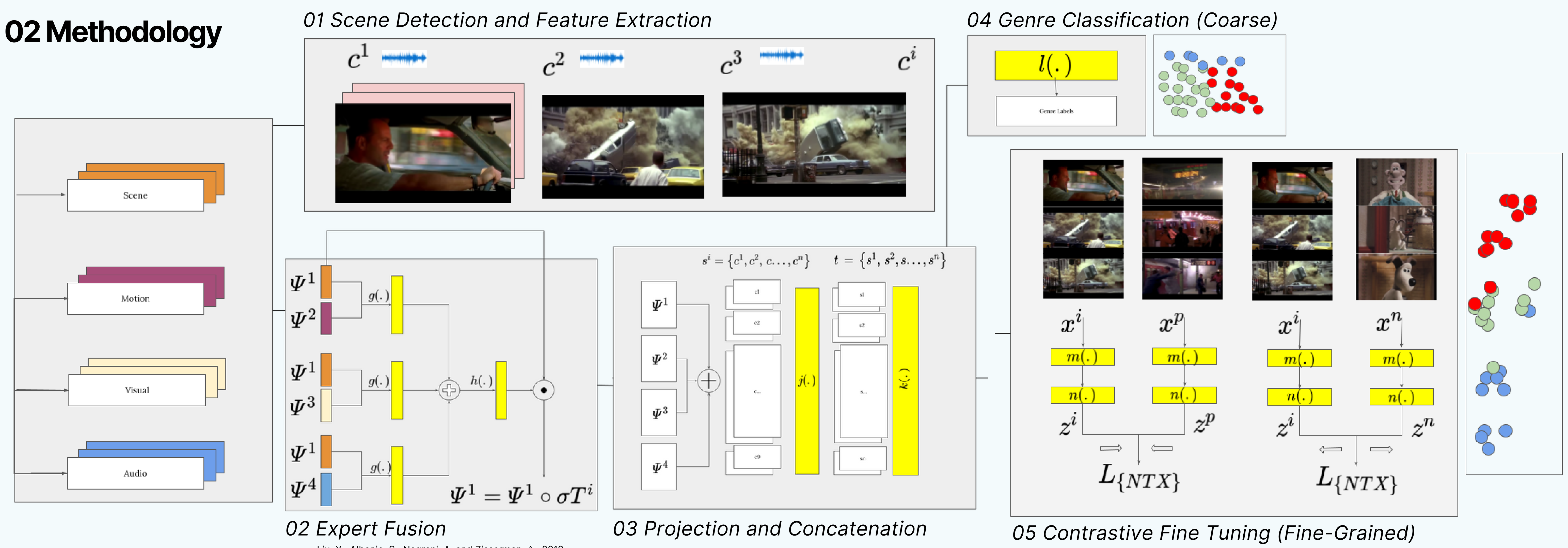
Here we show some retrieval results for target trailers. We can see how the **coarse genre encoder** retrieves trailers with the same genre label as the query despite differences in audio-visual content. Following **fine-grained self-supervised learning,** retrieval yields results much closer to the original trailer.



Model	Actn	Advnt	Animtn	Bio	Cmdy	Crme	Doc	Drma	Famly	Fntsy	Hstry	Hrror	Mystry	Music	SciFi	Wstrn	Sprt	Shrt	Thrll	War	$F1_w$	$AU(\overline{PRC})_w$	P_w	R_w
Support	130	197	46	13	224	102	87	267	117	115	44	104	41	86	107	181	30	45	12	21	-	-	-	-
Random	0.29	0.41	0.11	0.03	0.46	0.24	0.21	0.52	0.27	0.26	0.11	0.24	0.1	0.2	0.25	0.39	0.08	0.11	0.03	0.05	0.318	0.134	0.19	1
Scene [11]	0.43	0.55	0.74	0	0.49	0.38	0.63	0.55	0.51	0.28	0.24	0.42	0.3	0.28	0.41	0.51	0.22	0.19	0.11	0.33	0.434	0.489	0.437	0.48
Audio [1]	0.47	0.51	0.40	0.10	0.61	0.38	0.58	0.55	0.51	0.37	0.11	0.34	0.39	0.30	0.35	0.55	0.16	0.15	0.13	0.12	0.454	0.449	0.400	0.537
Motion [6]	0.5	0.59	0.74	0	0.62	0.33	0.63	0.56	0.55	0.36	0.2	0.38	0.45	0.24	0.37	0.57	0.23	0.14	0.10	0.13	0.463	0.487	0.448	0.494
Image [12]	0.48	0.63	0.79	0.12	0.65	0.41	0.60	0.59	0.55	0.42	0.25	0.47	0.42	0.29	0.50	0.54	0.34	0.19	0.12	0.31	0.516	0.554	0.493	0.572
Image + Audio	0.52	0.63	0.78	**0.15**	0.65	0.42	0.68	0.6	0.63	0.46	0.25	0.50	0.51	0.34	0.49	0.59	0.38	**0.28**	0.12	0.42	0.544	0.558	0.476	0.65
Image + Motion	0.59	0.64	0.78	0	0.59	0.39	0.66	0.6	0.6	0.5	0.29	0.54	0.53	0.25	**0.52**	0.57	0.4	0.2	0.24	0.12	0.535	0.553	0.511	0.583
Image + Scene	0.52	0.61	0.80	0.12	0.61	0.37	0.65	**0.62**	0.58	0.49	0.15	0.51	0.49	0.37	0.48	0.56	**0.43**	0.26	0.12	0.46	0.531	0.539	0.490	0.600
Naive Concat	0.56	0.61	0.64	0.09	0.64	0.35	0.69	0.60	0.58	0.39	0.19	0.49	0.45	0.21	0.48	0.6	0.39	0.28	0.27	0.41	0.525	0.497	0.522	0.551
MMX-Trailer-20	**0.62**	**0.69**	**0.71**	0.11	**0.71**	**0.53**	**0.73**	**0.62**	**0.64**	**0.51**	**0.34**	**0.56**	**0.60**	**0.45**	0.50	**0.64**	0.30	0.11	**0.13**	**0.55**	**0.597**	**0.583**	**0.554**	**0.697**

Fig 2. Ablation studies on the effect of individual experts on **coarse genre classification.**

02 Methodology



Liu, Y., Albanie, S., Nagrani, A. and Zisserman, A., 2019. Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487.

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

edward.fish@surrey.ac.uk