

Video Understanding using Multimodal Deep Learning

Arsha Nagrani



Wolfson College

University of Oxford

A thesis presented for the degree of
Doctor of Philosophy

Trinity 2020

To Saleena and Vineet Nagrani, my role models forever.

The artist brain is the sensory brain: sight and sound, smell and taste, touch. These are the elements of magic.

— Julia Cameron

Acknowledgements

None of this would have been possible without the work of so many.

First and foremost, my supervisor Andrew Zisserman to whose consistent and hands-on guidance I will always be indebted. Thank you for encouraging me to be creative, thoughtful and hardworking. I am extremely fortunate and grateful to have worked with you so closely for the last four years. You have moulded me into the researcher I am today.

To Joon Son Chung, for helping me write my first paper and teaching me that no project is too ambitious. To Samuel Albanie, for showing me not only how to define a problem, but how to investigate and write about it. And to all the fine members of VGG, thank you for creating an environment so conducive to collaborative research. It has been a privilege to share an office with such talented people from all over the world. In particular to my collaborators - Andrea Vedaldi, Weidi Xie, Max Bain, Yang Liu and Andrew Brown, thank you for generously sharing your research outputs, and teaching me that research is best done with others.

To all the researchers in Oxford Anthropology, thank you for showing me the impact that computer vision has on other disciplines. And to Rahul Sukthankar and others at Google, thank you for generously funding this research, and making my time in California so memorable. You have encouraged me to expand my horizons and I am excited for what the future holds.

To all the committees and teams I have worked with over the years: Women in Computer Vision, Black in AI, and numerous others – I am inspired by your commitment to meaningful causes that benefit others. I am also grateful to the wider research community, the dreamers, thinkers and visionaries of today and of yesterday. They are the giants whose shoulders I firmly stand on today.

To my large and warm family all over the world - which has too many members to all name here. In particular, to my father Vineet Nagrani, for being my beacon of support. Thank you for always encouraging me to pursue my dreams, even when they take me far away from home. To my mother Saleena Nagrani for being my guide and sounding board every step of the way. Your never failing compassion and patient listening (to constant weather complaints) has navigated me through this journey, and I owe

all my successes to you. To my fun-loving brother Ishan Nagrani, whose sense of humour and perpetual optimism has gotten me through the darkest of times. To Alpha, whose unwavering loyalty and wagging tail brightens my day. To my grandparents, aunts and uncles, for being proud of me always. And to Deepa and Ajit Bhushan, for making my life in the U.K. just that much more comfortable and enjoyable.

I am also grateful to my dear friends in Cambridge, Oxford and London, who have been the sunlight I needed on the many dark and rainy days. Thank you for being a crucial part of my home away from home, in this beautiful and historic corner of the world. Through every new person and every new conversation I have had, I am grateful for the impact on my ever-expanding world view. I cherish having met and getting to know each one of you.

And finally, to Mihir Bhushan, my greatest cheerleader – who has faith in me and my research when I scarcely have any myself. Thank you for being my pillar of support, my confidante, advisor, proof-reader, therapist and role model all in one. Of all the things the U.K. has given me, you are by far the best one.

Abstract

Our experience of the world is multimodal, however deep learning networks have been traditionally designed for and trained on unimodal inputs such as images, audio segments or text. In this thesis we develop strategies to exploit multimodal information (in the form of vision, text, speech and non-speech audio) for the automatic understanding of human-centric videos. The key ideas developed in this thesis are (i) Cross-modal Supervision, (ii) Self-supervised Representation Learning, and (iii) Modality Fusion.

In *cross-modal supervision*, data labels from a supervision-rich modality are used to learn representations in another, supervision-starved target modality, eschewing the need for costly manual annotation in the target modality domain. This effectively exploits the redundant, or overlapping information between modalities. We demonstrate the utility of this technique for three different tasks; First we use face recognition and visual active speaker detection to curate a large scale audio-visual dataset of human speech called VoxCeleb, training on which yields state of the art models for speaker recognition; Second we train a text-based model to predict action labels from transcribed speech *alone*, and transfer these labels to accompanying videos. Training with these labels allows us to outperform fully supervised action recognition models trained with costly manual supervision; Third, we distill the information from a face model trained for emotion recognition to the speech domain, where manual emotion annotation is expensive.

The second key idea explored in this thesis is the use of modality redundancy for *self-supervised representation learning*. Here we learn audio-visual representations without any manual supervision in either modality, specifically for human faces and voices. Unlike existing representations, our joint representations enable cross-modal retrieval from audio to vision and vice-versa. We then extend this work to explicitly remove learnt biases, enabling greater generalisation.

Finally, we effectively combine the complementary information in different modalities through the development of new *modality fusion* architectures. By distilling the information from multiple modalities in a video to a single, compact video representation, we achieve robustness to unimodal inputs which can be missing, corrupted, occluded, or have various levels of background noise. With these models we achieve state of the art results in both action recognition and video-text retrieval.

Keywords— multimodal, deep learning, cross-modal, multisensory, neural networks

Table of Contents

1	Introduction and Background	5
1.1	Motivation	6
1.2	Key Ideas	9
1.2.1	Cross-Modal Supervision	9
1.2.2	Self-Supervised Representation Learning	9
1.2.3	Multimodal Fusion	10
1.3	Thesis Outline and Contributions	12
1.3.1	Publications	14
I	Cross-Modal Supervision	16
2	VoxCeleb: Large-scale Speaker Verification in the Wild	18
2.1	Introduction	19
2.2	Related Works	21
2.3	The VoxCeleb Dataset	23
2.4	Dataset Collection Pipeline	25
2.5	VGGVox	30
2.6	Experiments	34
2.7	Results	39
2.8	Conclusion	43
3	<i>Speech2Action</i>: Cross-modal Supervision for Action Recognition	45
3.1	Introduction	45
3.2	Related Works	48
3.3	<i>Speech2Action</i> Model	50
3.4	Mining Videos for Action Recognition	54
3.5	Action Classification	58
3.6	Conclusion	63

4	Emotion Recognition in Speech using Cross-Modal Transfer in the Wild	65
4.1	Introduction	66
4.2	Related Work	68
4.3	Cross Modal Transfer	72
4.4	EMOVOXCELEB Dataset	77
4.5	Experiments	80
4.6	Conclusions	88
II	Self-Supervised Representation Learning	89
5	<i>Seeing Voices and Hearing Faces: Cross-modal biometric matching</i>	91
5.1	Introduction	92
5.2	Related Work	94
5.3	Cross-Modal Models	97
5.4	Datasets and Training	101
5.5	Experiments	103
5.6	Results and Discussion	106
5.7	Ablation Analysis	109
5.8	Conclusion	111
6	<i>Learnable PINs: Cross-Modal Embeddings for Person Identity</i>	113
6.1	Introduction	114
6.2	Related Work	115
6.3	Learning Joint Embeddings	117
6.4	The Importance of Curriculum-based Mining	119
6.5	Dataset	122
6.6	Experiments	123
6.7	Evaluation	124
6.8	One-Shot Learning for TV Show Character Retrieval	129
6.9	Conclusion	132

7	Disentangled speech embeddings using cross-modal self-supervision	134
7.1	Introduction	135
7.2	Related Work	137
7.3	Model	138
7.4	Experiments	142
7.5	Conclusion	146
III	Multimodal Fusion	147
8	EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition	149
8.1	Introduction	150
8.2	Related Work	152
8.3	The Temporal Binding Network	155
8.4	Experiments	161
8.5	Conclusion	171
9	Use What You Have: Video Retrieval Using Representations From Collaborative Experts	172
9.1	Introduction	173
9.2	Related Work	175
9.3	Collaborative Experts	177
9.4	Experiments	181
9.5	Conclusion	188
10	Discussion	190
10.1	Achievements and Impact	190
10.2	What Comes Next?	194
10.3	Conclusion	197
Appendix A	Feature Aggregation	235
A.1	Spatiotemporal Aggregation	235
A.2	Multimodal Aggregation	236

Appendix B	Objective Functions	238
Appendix C	Statements of Authorship	240

1 | Introduction and Background

Our experience of the world is multimodal – we see objects, hear sounds, read words, feel textures and taste flavors. Humans interact with their environment through these multiple sensory streams, combining information and forming associations between senses [Edelman, 1987; Bahrick & Lickliter, 2000; L. Smith & Gasser, 2005]. When a baby eats an apple, for instance, taste is not her only experience – she will also hear the apple crunch, see its shiny skin, and feel its smooth surface on her skin [L. Smith & Gasser, 2005]. Psychologists suggest that these multiple overlapping and time-locked sensations are an crucial enabler of human perceptual learning about the world [Bahrick & Lickliter, 2000; L. Smith & Gasser, 2005].

Artificial machine learning, in contrast, has traditionally focused on solving tasks within single modality domains (e.g. in computer vision, speech recognition or natural language processing). Research communities in these domains have made valuable contributions by developing and learning from large, *unimodal* corpora containing millions of data samples, such as images (ImageNet [J. Deng et al., 2009]), words (BooksCorpus [Zhu et al., 2015]), or sounds (AudioSet [Gemmeke et al., 2017]). A particular successful technique for exploiting these large datasets is *deep learning* [LeCun et al., 2015], which learns representations that map raw data formats to easily ingestible and compact embedding vectors. This is often done through versatile, multi-step function approximators, or deep neural networks [LeCun et al., 1998, 2015] which learn increasingly semantic, hierarchical representations by minimising a suitable loss function on input-output data pairs. This paradigm has seen the development of new neural network architectures that are particularly effective in distilling images [Krizhevsky et al., 2012], audio [Dahl et al., 2011], and text [Sutskever et al., 2014] into compact, meaningful representations.

Given the explosion of multimodal content, however, learning from unimodal data in isolation is becoming an increasingly unnatural and artificially contrived scenario. Online videos are naturally multimodal, often containing an audio track accompanying visual content. Images uploaded to social media are frequently accompanied by contextual text in the form of captions or hashtags, while news broadcasts often show text on screen. Movies, TV shows and user-uploaded videos are likely to be

accompanied by music streams, subtitle tracks and other metadata. Autonomous vehicles will not be equipped only with vision cameras, but also with active sensors providing depth information (LiDAR), Radio Detection and Ranging (RADAR), and Global Navigation Satellite Systems (GNSS) [Otaegui, 2018]. A major challenge that the machine learning community faces today is, what is the best way to exploit these multiple, heterogenous streams of information?

In this thesis, we show that multimodal data can provide us with two key benefits: (1) *Redundancy* – the common, or overlapping information between modalities; and (2) *Complementarity*, the information obtained from combining multiple modalities that is unattainable from a single modality. We explore training and developing multimodal neural networks to exploit both these two benefits, the former as a source of *multimodal supervision*, and the latter to guide *fusion architectures* that outperform unimodal architectures. These two key ideas are described in detail in the next few sections.

We apply these ideas to enable the understanding of *human-centric* videos. Learning and understanding human stories and interactions in videos involves recognising human identities, emotions, actions and situations. Our goal is to use deep learning to map high dimensional sensory inputs into meaningful embedding vectors that can be used to accurately understand and represent these human behaviours in video. We focus primarily on four sensory inputs: visual signals which are derived from a single or multiple frames of a video; speech including para-verbal information such as prosody and vocal expressions; non-speech audio signals including environmental sounds; and written natural language (text).

1.1 Motivation

Redundancy for Supervision: While deep learning has allowed ground-breaking results across many tasks [Krizhevsky et al., 2012; Girshick, 2015; Dahl et al., 2011; Y. Wu et al., 2016], in the absence of explicit feature engineering, these methods are data hungry – depending on millions of manually annotated training examples to discover and distil patterns. Obtaining this annotation is expensive and time consuming, and does not scale to the growing amount of unlabelled data available.

A major goal of the research community has been to remove the need for this manual supervision -

creating systems that, instead, teach themselves by analyzing unlabeled *unimodal* data. These tasks are designed in such a way that they force a deep learning model to capture underlying semantics for a single modality. For example, exploiting the spatial correspondences between image patches by solving a Jigsaw puzzle proxy task [M. Noroozi & Favaro, 2016], or using a video frame order verification task to leverage visual temporal correspondences in video [D. Xu et al., 2019].

Learning from a single modality, however is contrary to human experience [Barlow, 1989]. Why do humans interact with the world through so many sensory streams— vision, audition, touch, smell, proprioception and balance? As early as the 1980s, psychologists proposed that the answer lies in the concept of redundancy, or as referred to in psychology – *degeneracy* [Edelman, 1987]. Redundancy allows a system to function even with the loss of one component. For example, our experience of space is not limited to sight alone, but is present in sound, movement, touch, and even smell, and hence the lack of a single modality does not interfere with our concept of space [Edelman, 1987]. This is evident in studies of blind children [Edelman, 1987], where comparable spatial concepts are developed through different clusters of modalities. Redundancy also means that sensory systems can educate each other, without an external teacher, providing a form of *self-supervision* [de Sa, 1994]. Experimental observations of infants have discovered that they spend literally hours gazing at their own actions [Piaget, 1936; Bushnell, 1994; Borjon et al., 2018] – holding their hands in front of their faces, watching as they turn them back and forth, and focusing visual attention on objects based on the objects’ feel and weight [Yuan et al., 2019].

Another form of evidence for sensory redundancy is *re-entry* [Edelman, 1987], the explicit inter-relating of multiple simultaneous representations across modalities, where the sensation of one modality can immediately trigger the memory of another. For example, psychologists note that humans can visualise the face of a person after simply being exposed to their voice [Kamachi et al., 2003; H. M. Smith et al., 2016b], without having seen or heard the person before. This phenomenon can be explained by a highly-influential cognitive model [Bruce & Young, 1986], which proposed that ‘person identity nodes’ or ‘PINs’ are a portion of associative memory holding identity-specific codes that are entirely abstracted from the input modality. This suggests that humans are able to effortlessly link high-level semantics across different sensory input modalities.

In this work, we argue that similar to human perceptual learning, modality redundancy can help train artificial neural network architectures as well. We exploit cross-modal redundancy in two ways; first as a source of cross-modal supervision (Part I) – where labels from a supervision-rich modality are used to learn representations in another, supervision-starved modality, and secondly in a fully self-supervised way (Part II), where redundant information across modalities is distilled into representations with no manual supervision in either modality. These concepts are described in more detail later in this chapter.

Complementarity for Fusion: Another challenge for deep learning architectures is the heterogeneity of input data. In ‘real world’ applications of deep learning, unimodal inputs can be missing, corrupted, occluded, or have various levels of background noise, making it difficult to disambiguate different semantic concepts. Even small perturbations due to degradation in visual inputs can significantly distort the feature embeddings and output of a neural network [Zheng et al., 2016]. Achieving robustness to such effects is imperative for deploying deep learning models ‘in the wild’.

One method used to obtain perceptual robustness in humans is the fusion of information from multiple modalities. For example, vision and hearing are closely intertwined in the perceptual system – the motion of a speaker’s lips, for instance, can profoundly change our perception of hearing. This is known as the McGurk effect [McGurk & MacDonald, 1976], where the visual modality provides *complementary* information on the place of articulation and muscle movements [Summerfield, 1992], and hence helps to disambiguate between speech with similar acoustics (e.g., the unvoiced consonants /p/ and /k/). Another beautiful study of audio-visual fusion for *disambiguation* [Sekuler et al., 1997], shows participants two identical objects moving towards one another, coinciding and then moving apart (but with the same linear path and speed). With the visual display alone, participants guessed a collision and bouncing 20% of the time, the rest of the time reporting that the objects just crossed and continued in their original directions. Presenting a loud click at the time of coincidence increased the frequency of bouncing reports to 60%. In the first case, vision helps to disambiguate the sound signal, whereas in the second, sound helps to disambiguate a vision signal.

In this thesis we develop multimodal neural network architectures, that combine and distil the information from multiple modalities into a single representation. We show that such architectures outperform single modality architectures. Note that here the supervision is provided manually.

1.2 Key Ideas

The work in this thesis is organised into three key ideas or themes for multimodal deep learning; (i) Cross-Modal Supervision, (ii) Self-Supervised Representation Learning and (iii) Modality Fusion. While the first two exploit modality redundancy, the third exploits complementarity of modalities.

1.2.1 Cross-Modal Supervision

Labelled data in one modality (the source) can be used to aid learning in another modality (the target modality). This is particularly useful in cases where there are large labelled datasets in one modality (e.g. face recognition [Parkhi et al., 2015; Cao et al., 2018]), but it is more challenging to obtain labelled data for the same semantic task in another modality (e.g. identity recognition from speech). Since identity information is present in both face images and speech segments, we use face recognition in unlabelled videos to obtain identity supervision for corresponding speech segments (Chapter 2), thereby creating a dataset for the task of speaker recognition. We also explore the transfer of supervision from speech to vision (Chapter 3), for the task of human action recognition. In both cases, the cross-modal redundancy present in paired data is used to directly transfer data labels from the source to the target modality (Chapters 2 and 3).

Another way to exploit labels in the source modality is to guide representation learning in the target modality without the explicit transfer of labels. We use this concept to distill information from a strong emotion face recognition model, to a speech emotion recognition model using unlabelled audio-visual data as a bridge (Chapter 4). In this setting, data from multiple modalities is available only during feature learning; during the testing phase, only data from the target modality is provided (Chapter 4).

1.2.2 Self-Supervised Representation Learning

Cross-modal redundancy can also be exploited when labelled data is unavailable in either modality. In this case, the synchrony or co-occurrence of modalities can be exploited as a source of *self-supervision*, to learn shared representations. This has been exploited by a number of audio-visual methods [Arand-

jelović & Zisserman, 2017; Owens et al., 2016], where the objective is to match visual and audio components extracted from the same video. Unlike such works that are interested in objects (such as instruments) and other items in general, we focus particularly on humans. As mentioned earlier, a major aspect of human perception is the theory of ‘person identity nodes’ or ‘PINs’ [Bruce & Young, 1986] – portions of associative memory holding *modality free* codes, e.g. identity-specific semantic codes that can be accessed via the face, the voice, or other modalities [Bruce & Young, 1986], and hence are entirely abstracted from the input modality. With a deep neural network, we represent the concept of PINs with a joint representation space, where the network maps inputs from different modalities into the same embedding space. We learn this joint representation space by training visual and audio networks *simultaneously and from scratch* to predict whether visual information is semantically related to audio information (for example, whether a face image and voice segment share the same identity). These joint representations can then be evaluated individually in each modality, as well as evaluated for correlations across different modalities via *cross-modal retrieval* (Chapters 5, 6 and 7).

1.2.3 Multimodal Fusion

While the first two ideas exploit multimodal redundancy, we also explore the usefulness of multimodal complementarity. Here we distill the information from different modalities in a video into a single representation.

While training fusion models, it is worthwhile to note that the information coming from different modalities may have varying predictive power, noise topology and representational formats. For example, language is symbolic and can be represented with a fixed vocabulary, while audio and visual modalities are represented as continuous signals. Sound is a single dimensional waveform in time, where confounding object/event noise is always additive. In contrast, images are 2D (both dimensions spatial), and confounding elements of a scene can be occlusive – making it difficult to recover relevant content. Hence input representations, and consequently neural network architectures tend to vary wildly for different modalities.

To deal with this, until recently, multimodal fusion in machine learning was mostly restricted to *late fusion*, in which modalities are treated largely independently right until the end, where scores from each

unimodal system are then combined for the final output. This combination can be a simple [Simonyan & Zisserman, 2014] or weighted [Natarajan et al., 2012] score average, a bilinear product [Ben-Younes et al., 2017], or a more robust combination such as rank minimization [Ye et al., 2012]. This also conveniently allows unimodal systems to deal with short-term *temporal* ranges *within* different modalities independently. To account for varying representation sizes and frame-rates, most multi-modal architectures apply temporal aggregation functions to each modality in the form of average pooling or other temporal pooling functions (e.g. maximum or NetVLAD [Arandjelović et al., 2016]), before attempting multimodal fusion.

However, fusing modalities at their respective deepest features is not necessarily optimal. In this thesis, we instead focus on *mid-fusion*, where features are merged earlier on. This is an open research area, with some recent works even applying artificial architecture search [Pérez-Rúa et al., 2019] to automatically discover optimal mid-fusion architectures. We first focus on developing novel mid fusion architectures for combining audio and visual information for the task of egocentric action recognition [Damen et al., 2018] (Chapter 8). The egocentric domain in particular offers rich sounds resulting from the interactions between hands and objects, as well as the close proximity of the wearable microphone to the undergoing action. Audio can also capture actions that are out of the wearable camera’s field of view, but audible (e.g. ‘eat’ can be heard but not seen).

While mid-fusion is an effective technique for learning the optimal way to combine different modality inputs, the task of learning from video content is made extremely challenging by the high dimensionality of the sensory data contained in a single video. To train deep networks *end-to-end* with discriminative training would require prohibitively expensive and detailed annotation of a vast number of videos, to exhaustively cover the entire annotation space. Hence in this thesis, we also explore fusion architectures for pre-extracted feature embeddings (Chapter 9), which are semantic representations of the video data learnt by individual experts (in audio, scenes, actions, etc). In essence, this approximation enables us to exploit knowledge from existing individual sources where the cost of annotation is significantly reduced (e.g. classification labels for objects and scenes in images, labels for actions in videos etc.) and where consequently, there exist very large-scale labelled datasets.

1.3 Thesis Outline and Contributions

In this section we summarise the contributions of this thesis, and provide an outline of the chapters. The thesis is divided into three parts – (i) Cross-Modal Supervision, (ii) Self-Supervised Representation Learning, and (iii) Multimodal Fusion.

While the text for each chapter contains a section on related work specific to that chapter, we also provide some broad background material in the Appendix, knowledge of which is assumed for most of the chapters in the thesis. The key stages for most of the deep learning based systems for multimodal and crossmodal representation learning in this thesis are: (i) frame level feature extraction using deep neural networks or similar feature extractors; (ii) aggregation of frame level features (this can be temporal, spatial, or across modalities); and (iii) optimisation of an objective loss function. We provide a review of the latter two components in the context of multimodal and cross-modal deep learning: aggregation architectures in Appendix A and objective functions or learning constraints in Appendix B.

For Chapters 2 to 9, we summarise the main contribution below. Finally, Chapter 10 discusses the impact of this work and avenues for future exploration.

Part I: Cross-Modal Supervision

We first introduce a *fully automated pipeline* based on face recognition and visual active speaker detection to collect a very large-scale *speaker recognition* dataset from open source media (YouTube videos), called VoxCeleb (Chapter 2). VoxCeleb is the largest publicly available dataset of human speech ‘in the wild’. This multimodal dataset consists of interviews of celebrities from unconstrained YouTube videos. Inspired by computer vision architectures, we also develop and train new convolutional neural network architectures for speaker recognition, achieving state of the art results on standard benchmarks. The VoxCeleb datasets are also used as a primary data source for Chapters 4 to 7.

While in Chapter 2 we show how vision can be used to automatically create a large dataset of speech, in Chapter 3 we show how transcribed speech can be used to obtain labelled data for vision, i.e. specifically for the task of human action recognition. We train a model on text material to predict action labels from transcribed speech segments, and using the predictions of this model, obtain weak

action labels for over 800K video clips automatically. By training on these video clips, we demonstrate superior action recognition performance on standard action recognition benchmarks.

In [Chapter 4](#), we do not explicitly transfer labels from one modality to another, but instead use face models trained for emotion recognition as a ‘teacher’, to distill knowledge into a speech emotion recognition model (the ‘student’). Labelling emotion in speech manually is an extremely challenging task, and our method eschews the need for this costly manual annotation. The representations learnt outperform other models trained for speech emotion recognition on standard benchmarks, even approaching fully supervised performance.

Part II: Self-Supervised Representation Learning

In [Chapters 5](#) and [6](#) we use cross-modal self-supervision to learn representations for identity recognition from face and voice inputs. Our representations are learnt without any manual supervision in either modality. Such representations enable cross-modal retrieval from voice to face and from face to voice. In [Chapter 7](#), we extend this work to explicitly remove content information from these identity embeddings, enabling greater generalisation.

Part III: Multimodal Fusion

In [Chapter 8](#), we propose a novel mid fusion architecture for combining inputs from different modalities, including RGB, optical flow, and audio, focusing on the task of egocentric action recognition [[Damen et al., 2018](#)]. Our proposed architecture is trained end-to-end, outperforming individual modalities as well as late-fusion of modalities.

In [Chapter 9](#), we focus on the task of fusing pre-extracted embeddings from different modalities into a single video embedding for the task of video-text retrieval. The embeddings include ‘general’ features such as motion, appearance, and scene features from visual content, as well as more ‘specific’ cues from ASR and OCR which are intermittently available for videos. We propose a *collaborative experts* model to aggregate information from these different pre-trained experts and achieve state of the art results on five retrieval benchmarks: MSR-VTT [[J. Xu et al., 2016](#)], LSMDC [[Rohrbach et al., 2015](#)], MSVD [[D. L. Chen & Dolan, 2011](#)], DiDeMo [[Anne Hendricks et al., 2017](#)] and ActivityNet-captions [[Krishna et al., 2017](#)], covering a challenging set of domains which include videos from YouTube, personal collections and movies.

1.3.1 Publications

Chapters 2 to 9 each contain a paper which has been peer-reviewed and accepted for publication in a conference or journal. The papers have been left unmodified from their published forms, with the exception of formatting changes. For each publication, we also provide a statement of authorship in Appendix C. The publications included in this thesis are:

Chapter 3: “Voxceleb: Large-scale speaker verification in the wild.”

Arsha Nagrani*, Joon Son Chung*, Weidi Xie, and Andrew Zisserman. In *Computer Speech Language*, 60, p.101027 2020.

Chapter 3: “Speech2Action: Cross-modal Supervision for Action Recognition”.

Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, Andrew Zisserman In *Conference on Computer Vision and Pattern Recognition (CVPR)* 2020.

Chapter 4: “Emotion Recognition in Speech using Cross-Modal Transfer in the Wild”

Samuel Albanie*, **Arsha Nagrani***, Andrea Vedaldi, Andrew Zisserman, In *ACM Multimedia*, 2018

Chapter 5: “Seeing Voices and Hearing Faces: Cross-modal biometric matching”

Arsha Nagrani, Samuel Albanie, Andrew Zisserman In *Conference on Computer Vision and Pattern Recognition (CVPR)* 2018.

Chapter 6: “Learnable PINs: Cross-Modal Embeddings for Person Identity”

Arsha Nagrani*, Samuel Albanie*, Andrew Zisserman In *European Conference on Computer Vision (ECCV)* 2018.

Chapter 7: “Disentangled Speech Embeddings using Cross-Modal Self-Supervision”.

Arsha Nagrani*, Joon Son Chung*, Samuel Albanie*, Andrew Zisserman In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2020.

Chapter 8: “EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition”

Evangelos Kazakos, **Arsha Nagrani**, Andrew Zisserman, Dima Damen. In *International Conference on Computer Vision (ICCV)*, 2019

Chapter 9: “Use What You Have: Video Retrieval Using Representations From Collaborative Experts”:

Yang Liu*, Samuel Albanie*, **Arsha Nagrani***, Andrew Zisserman, In *British Machine Vision Conference (BMVC)*, 2019

Publications not included:

“VoxCeleb: a large-scale speaker identification dataset”[†]

Arsha Nagrani*, Joon Son Chung*, Andrew Zisserman. In *INTERSPEECH*, 2017 (Oral Presentation, Best Student Paper Award)

“VoxCeleb2: Deep Speaker Recognition”[†]

Joon Son Chung*, **Arsha Nagrani***, Andrew Zisserman. In *INTERSPEECH*, 2018

“Utterance-level Aggregation For Speaker Recognition In The Wild”[†]

Weidi Xie, **Arsha Nagrani**, Joon Son Chung, Andrew Zisserman. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019 (Oral Presentation)

“From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script”[‡]

Arsha Nagrani, Andrew Zisserman. In *British Machine Vision Conference (BMVC)*, 2017 (Oral Presentation)

“Chimpanzee face recognition from videos in the wild using deep learning”[‡]

Daniel Schofield*, **Arsha Nagrani***, Andrew Zisserman, Misato Hayashi, Tetsuro Matsuzawa, Dora Biro, Susana Carvalho. In *Science Advances*, 2019

“Count, Crop and Recognise: Fine-Grained Recognition in the Wild”[‡]

Max Bain, **Arsha Nagrani**, Daniel Schofield, Andrew Zisserman. In *ICCV Workshops*, 2019 (Oral Presentation)

[†] **Chapter 2** collates and extends, and therefore supersedes these publications.

[‡] These papers are only loosely related to the topic of this thesis and hence are excluded.

Part I

Cross-Modal Supervision

Humans use but a tiny percentage of their senses.
They barely look, they rarely listen, they never
smell, and they think that they can only experience
feelings through their skin. But they talk, oh, do
they talk.

— Michael Scott

2 | VoxCeleb: Large-scale Speaker Verification in the Wild

Arsha Nagrani^{1*} Joon Son Chung^{1,2*} Weidi Xie¹ Andrew Zisserman¹

Visual Geometry Group, Oxford¹ Naver Corporation²

(* Equal Contribution)

Abstract

The objective of this work is speaker recognition under noisy and unconstrained conditions. We make two key contributions. First, we introduce a very large-scale *audio-visual* dataset collected from open source media using a *fully automated pipeline*. Most existing datasets for speaker identification contain samples obtained under quite constrained conditions, and usually require manual annotations, hence are limited in size. We propose a pipeline based on computer vision techniques to create the dataset from open-source media. Our pipeline involves obtaining videos from YouTube; performing active speaker verification using a two-stream synchronization Convolutional Neural Network (CNN), and confirming the identity of the speaker using CNN based facial recognition. We use this pipeline to curate VoxCeleb which contains over a million ‘real-world’ utterances from over 6,000 speakers. This is several times larger than any publicly available speaker recognition dataset. Second, we develop and compare different CNN architectures with various aggregation methods and training loss functions that can effectively recognise identities from voice under various conditions. The models trained on our dataset surpass the performance of previous works by a significant margin.

Published in the [Computer Speech and Language Journal](#). This paper consolidates three separate conference papers [[Nagrani et al., 2017](#); [J. S. Chung et al., 2018](#); [W. Xie et al., 2019](#)] accepted to Interspeech 2017, Interspeech 2018 and ICASSP 2019 respectively.

2.1 Introduction

Speaker recognition under noisy and unconstrained conditions is an extremely challenging task. Applications of speaker recognition vary from authentication in high-security systems and forensic tests, to searching for persons in large corpora of speech data. All such tasks require high speaker recognition performance under ‘real-world’ conditions. This is extremely difficult due to both extrinsic and intrinsic variations; extrinsic variations include background chatter and music, laughter, reverberation, channel and microphone effects; while intrinsic variations are factors inherent to the speakers themselves such as age, accent, emotion, intonation and manner of speaking, amongst others [Stoll, 2011].

Deep Convolutional Neural Networks (henceforth, CNNs) have given rise to substantial improvements in speech recognition, computer vision and related fields due to their ability to deal with real-world, noisy datasets without the need for handcrafted features [Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; K. He et al., 2016]. One of the most important ingredients for the success of such methods, however, is the availability of large training datasets.

Unfortunately, large-scale public datasets in the field of speaker identification with unconstrained speech samples are lacking. While large-scale evaluations are held regularly by the National Institute of Standards in Technology (NIST), these datasets are not freely available to the research community. The only freely available dataset curated from multimedia is the Speakers in the Wild (SITW) dataset [McLaren et al., 2016], which contains speech samples of 299 speakers across unconstrained or ‘wild’ conditions. This is a valuable dataset, as the speech samples have been manually annotated, however, scaling it further, for example to thousands of speakers across tens of thousands of utterances, would require the use of a service such as Amazon Mechanical Turk (AMT). In the computer vision community AMT like services have been used to produce very large-scale datasets, such as ImageNet [Russakovsky et al., 2015].

We make two contributions towards the goal of speaker recognition under noisy and unconstrained conditions. The first contribution is to propose a fully automated and scalable pipeline for creating a large-scale ‘real-world’ speaker identification dataset. Benefiting from the success of face recognition research in computer vision, our method circumvents the need for human annotation completely. We

use this method to curate VoxCeleb¹, a large-scale dataset with over a million utterances for over seven thousand speakers. Since the dataset is collected ‘in the wild’, the speech segments are corrupted with real-world noise including laughter, cross-talk, channel effects, music and other sounds. The dataset is also multilingual, with speech from speakers of 145 different nationalities, covering a wide range of accents, ages, ethnicities and languages. The dataset is audio-visual, so is also useful for a number of other applications, for example, visual speech synthesis [J. S. Chung, Jamaludin, & Zisserman, 2017; Karras et al., 2017], speech separation [Afouras et al., 2018; Ephrat et al., 2018], cross-modal transfer from face to voice or vice versa [Nagrani et al., 2018b,a], emotion recognition [Albanie et al., 2018] and training face recognition from video to complement existing face recognition datasets [Cao et al., 2018; Kemelmacher-Shlizerman et al., 2016; Guo et al., 2016]. Since its official release in 2017, the dataset has already been downloaded over 3,000 times.

The second contribution is a deep CNN based neural speaker verification system, named VGGVox, which is trained to map voice spectrograms to a compact embedding space. We then use the cosine distance between vectors in this embedding space to measure the similarity between speakers. Besides speaker recognition and verification, clustering and novel speaker discovery can be straightforwardly implemented using standard techniques, with our embeddings as features. In developing VGGVox we investigate current popular CNN architectures, *e.g.* variants of VGG-M [Chatfield et al., 2014] and ResNet [K. He et al., 2016], different aggregation strategies, *e.g.* global average pooling, NetVLAD [Arandjelović et al., 2016], GhostVLAD [Zhong et al., 2018], and different loss functions for training the model, *e.g.* standard softmax classification, large-margin softmax and the contrastive loss. Our methods achieve state-of-the-art performance on the VoxCeleb1 speaker verification task, outperforming all other traditional methods and recent deep learning methods.

This paper consolidates three separate conference papers [J. S. Chung et al., 2018; Nagrani et al., 2017; W. Xie et al., 2019] into a single coherent document. In addition, we have added new results based on a new relation module, and added a more detailed comparison to previous work and the discussion of results.

¹The dataset can be downloaded from <http://www.robots.ox.ac.uk/~vgg/data/voxceleb>.

2.2 Related Works

Traditional methods. For a long time, speaker identification was dominated by Gaussian Mixture Models (GMMs) trained on low dimensional feature vectors [Reynolds et al., 2000; Reynolds & Rose, 1995]. The state of the art in more recent times involves both the use of joint factor analysis (JFA) based methods which model speaker and channel subspaces separately [Kenny, 2005], and i-vectors which attempt to model both subspaces into a single compact, low-dimensional space [Dehak et al., 2011]. These methods rely on a low dimensional representation of the audio input, such as Mel Frequency Cepstrum Coefficients (MFCCs). However, not only does the performance of MFCCs degrade rapidly in real-world noise [Yapanel et al., 2002; Hansen et al., 2001], but by focusing only on the overall spectral envelope of short frames, MFCCs may be lacking in speaker-discriminating features (such as pitch information). An in-depth review of these traditional methods is given in [Hansen & Hasan, 2015].

Deep learning methods. Deep neural networks (DNN) have been used successfully as feature extractors to learn discriminative embeddings in both computer vision and speech. Such methods [Variani et al., 2014; Lei et al., 2014; Snyder et al., 2017; Ghalehjegh & Rose, 2015; Snyder et al., 2018] are often combined with classifiers, both being trained independently. While such fusion methods are highly effective, since they are not trained end to end they still require hand-crafted engineering. In contrast, CNN architectures can be applied directly to raw spectrograms and trained in an end-to-end manner. End-to-end deep learning based systems for speaker recognition usually follow a similar three-stage pipeline: (i) frame level feature extraction using a DNN; (ii) temporal aggregation of frame level features; and (iii) optimisation of a classification loss. In the following, we review the three components in turn.

The trunk DNN architecture used is often either a 2D CNN with convolutions in both the time and frequency domain [Nagrani et al., 2017; Bhattacharya et al., 2017; J. S. Chung et al., 2018; Cai, Chen, & Li, 2018b; Hajibabaei & Dai, 2018; Cai, Chen, & Li, 2018a], or a 1D CNN with convolutions applied only to the time domain [Snyder et al., 2017; Shon et al., 2018; Okabe et al., 2018; Snyder et al., 2018]. A number of papers [Wan et al., 2018; Chowdhury et al., 2017] have also used LSTM-based front-

end architectures, including the work by Heigold et al. [Heigold et al., 2016], which unlike our work focused on *text-dependant* speaker verification.

The output from the feature extractor is a variable length feature vector, dependant on the length of the input utterance. Average pooling layers have been used in [Nagrani et al., 2017; J. S. Chung et al., 2018; Wan et al., 2018] to aggregate frame-level feature vectors to obtain a fixed length utterance-level embedding. [Snyder et al., 2017] introduces an extension of the method in which the standard deviation is used as well as the mean – this method is termed *statistical pooling*, and used by [Shon et al., 2018; Snyder et al., 2018]. Unlike these methods which ingest information from all frames with equal weighting, [Bhattacharya et al., 2017; Chowdhury et al., 2017] have employed attention models to assign weight to the more discriminative frames. [Okabe et al., 2018] combines the attention models and the statistical model to propose *attentive statistics pooling*. The final pooling strategy of interest is the Learnable Dictionary Encoding (LDE) proposed by [Cai, Cai, et al., 2018; Cai, Chen, & Li, 2018b]. This method is closely based on the NetVLAD layer [Arandjelović et al., 2016] designed for image retrieval.

Typically, such systems are trained end-to-end for classification with a softmax loss [Okabe et al., 2018] or one of its variants, such as the angular softmax [Cai, Chen, & Li, 2018b]. In some cases, the network is further trained for verification using the contrastive loss [Nagrani et al., 2017; J. S. Chung et al., 2018; D. Chen et al., 2011] or other metric learning losses such as the triplet loss [C. Li et al., 2017]. Similarity metrics like the cosine similarity or PLDA are often adopted to generate a final pairwise score.

Datasets. Many existing datasets are obtained under controlled conditions, for example: forensic data intercepted by police officials [van der Vloed et al., 2014], data from telephone calls [Hennebert et al., 2000], speech recorded live in high quality environments such as acoustic laboratories [Millar et al., 1994; Garofolo et al., 1993], or speech recorded from mobile devices [McCool & Marcel, 2009; Woo et al., 2006]. [Morrison et al., 2015] consists of more natural speech but has been manually processed to remove extraneous noises and crosstalk. All the above datasets are also obtained from single-speaker environments, and are free from audience noise and overlapping speech.

Datasets obtained from multi-speaker environments include those from recorded meeting data [Janin et

Name	Cond.	Free	# of Speakers	# of Utter.
ELSDSR [Feng & Hansen, 2005]	Clean Speech	✓	22	198
MIT Mobile [Woo et al., 2006]	Mobile Devices	-	88	7,884
SWB [Godfrey et al., 1992]	Telephony	-	3,114	33,039
POLYCOST [Hennebert et al., 2000]	Telephony	-	133	1,285‡
ICSI Meeting Corpus [Janin et al., 2003]	Meetings	-	53	922
Forensic Comparison [Morrison et al., 2015]	Telephony	✓	552	1,264
ANDOSL [Millar et al., 1994]	Clean speech	-	204	33,900
TIMIT [Fisher et al., 1986]†	Clean speech	-	630	6,300
MGB Challenge Dataset [Bell et al., 2015]	Broadcast Data	**	Unknown	1,600 hours
SITW [McLaren et al., 2016]	Multi-media	✓	299	2,800
NIST SRE [Greenberg, 2012]	Clean speech	-	2,000+	*
VoxCeleb1	Multi-media	✓	1,251	153,516
VoxCeleb2	Multi-media	✓	6,112	1,128,246

Table 2.1: Comparison of existing speaker identification datasets. **Cond.:** Acoustic conditions; **Utter.:** Approximate number of utterances. †And its derivatives. ‡Number of telephone calls. * varies by year. ** Only available to participants of the challenge. This dataset was mainly used for speech recognition (ASR).

al., 2003; McCowan et al., 2005], or from audio broadcasts [Bell et al., 2015]. These datasets usually contain audio samples under less controlled conditions. Some datasets contain artificial degradation in an attempt to mimic real-world noise, such as those developed using the TIMIT dataset [Garofolo et al., 1993]: NTIMIT, (transmitting TIMIT recordings through a telephone handset) and CTIMIT, (passing TIMIT files through cellular telephone circuits).

Table 2.1 summarises existing speaker identification datasets. Besides lacking real-world conditions, to the best of our knowledge, most of these datasets have been collected with great manual effort, other than [Bell et al., 2015] which was obtained by mapping subtitles and transcripts to broadcast data.

2.3 The VoxCeleb Dataset

We released the dataset in two stages, as VoxCeleb1 and VoxCeleb2. VoxCeleb1 contains over 100,000 utterances for 1,251 celebrities, while VoxCeleb2 contains over 1 million utterances for over 6,000 celebrities extracted from videos uploaded to YouTube. We attempt to minimise gender imbalance (VoxCeleb1 – 55% male, VoxCeleb2 – 61% male). The speakers span a wide range of different ethnicities, accents, professions and ages. Videos included in the dataset are shot in a large number of challenging visual and auditory environments. These include interviews from red carpets, outdoor stadiums and indoor studios, speeches given to large audiences, excerpts from professionally

shot multimedia, and even crude videos shot on hand-held devices. Crucially, all are degraded with real-world noise, consisting of background chatter, laughter, overlapping speech, room acoustics, and there is a range in the quality of recording equipment and channel noise. We also provide face detections and face-tracks for the speakers in the dataset, and the face images are similarly ‘in the wild’, with variations in pose (including profiles), lighting, image quality and motion blur. Table 2.2 gives the general statistics, and Figure 2.1 shows examples of cropped faces as well as utterance length, gender and nationality distributions.

Dataset	VoxCeleb1	VoxCeleb2
# of speakers	1,251	6,112
# of male speakers	690	3,761
# of videos	22,496	150,480
# of hours	352	2,442
# of utterances	153,516	1,128,246
Avg # of videos per speaker	18	25
Avg # of utterances per speaker	116	185
Avg length of utterances (s)	8.2	7.8

Table 2.2: Dataset statistics for both VoxCeleb1 and VoxCeleb2. Note VoxCeleb2 is more than 5 times larger than VoxCeleb1.

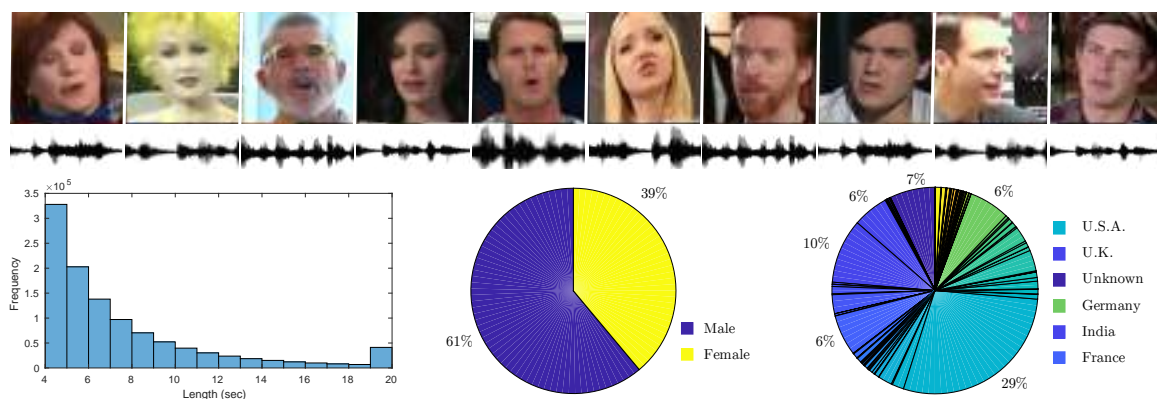


Figure 2.1: **Top row:** Examples from the VoxCeleb2 dataset. We show cropped faces of some of the speakers in the dataset. Both audio and face detections are provided. **Bottom row:** (left) distribution of utterance lengths in the dataset – lengths shorter than 20s are binned in 1s intervals and all utterances of 20s+ are binned together; (middle) gender distribution and (right) nationality distribution of speakers. For readability, the percentage frequencies of only the top-5 nationalities are shown. Best viewed zoomed in and in colour.

Both datasets contain development and test sets, with disjoint speakers. The development set of

Dataset	V1 Dev	V1 Test	V1 All	V2 Dev	V2 Test	V2 All
# of speakers	1,211	40	1,251	5,994	118	6,112
# of videos	21,819	677	22,496	145,569	4,911	150,480
# of utterances	148,642	4,874	153,516	1,092,009	36,237	1,128,246

Table 2.3: Development and test set splits for VoxCeleb1 and VoxCeleb2.

	Vox1 Train	Vox1 Test	Vox2 Train	Vox2 Test	SITW
Vox1 Train	Y	N	N	Y	Y
Vox1 Test	N	Y	N	Y	Y
Vox2 Train	N	N	Y	N	N
Vox2 Test	Y	Y	N	Y	Y
SITW	Y	Y	N	Y	Y

Table 2.4: Overlap between development and test sets for VoxCeleb1, VoxCeleb2 and SITW. N refers to there definitely being no overlap, Y refers to the possibility of overlap between the sets.

VoxCeleb2 has no overlap with the identities in the VoxCeleb1 or SITW datasets. Since we have created a number of evaluation benchmarks using the VoxCeleb1 dataset for testing (Sec. 2.6.1), we encourage others to use the *development* set of VoxCeleb2 *only* to train models for the speaker recognition task so that they can evaluate their methods fairly on VoxCeleb1. The VoxCeleb2 *test* set should prove useful for other applications of audio-visual learning for which the dataset might be used. The statistics for all the dev/test splits are given in Table 2.3. For clarity, we also provide a summary of the possible overlap between the development and test sets of VoxCeleb1, VoxCeleb2 and SITW in Table 2.4. This is useful for researchers wishing to train on one of these datasets and test on another.

2.4 Dataset Collection Pipeline

This section describes our multi-stage approach for collecting a large speaker recognition dataset, starting from YouTube videos. Our pipeline involves obtaining videos from YouTube; performing active speaker verification using a two-stream synchronization Convolutional Neural Network (CNN), and confirming the identity of the speaker using CNN based facial recognition. Using this fully automated pipeline, we have obtained over a million utterances for thousands of different speakers. The overall pipeline is the same for both VoxCeleb1 and VoxCeleb2, but the methods used in key stages differ since we selected the state-of-the-art face recognition systems at the time of dataset curation. The

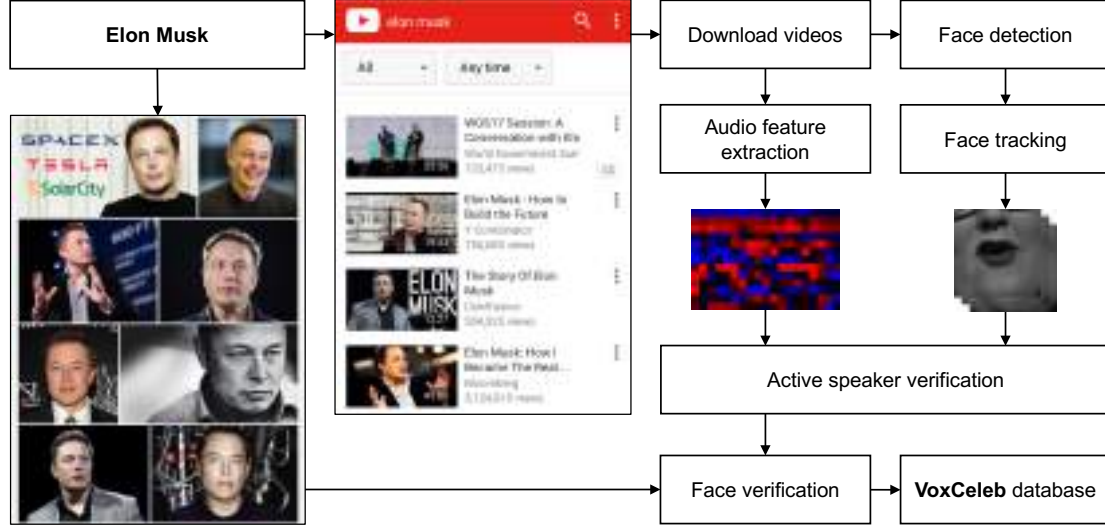


Figure 2.2: The multi-stage automatic pipeline used to create the VoxCeleb dataset automatically from YouTube videos. Our pipeline involves obtaining videos from YouTube; performing active speaker verification using a two-stream synchronization Convolutional Neural Network (CNN), and confirming the identity of the speaker using CNN based facial recognition.

pipeline is summarised in Figure 2.2, and key stages are discussed in the following subsections:

2.4.1 Candidate list of speakers.

The first stage is to obtain a list of speakers.

VoxCeleb1. We start from the list of people that appear in the VGGFace1 dataset [Parkhi et al., 2015], which is based on an intersection of the most searched names in the Freebase knowledge graph, and the Internet Movie Data Base (IMDB). This list contains 2,622 identities, ranging from actors and sportspeople to entrepreneurs, of which approximately half are male and the other half female.

VoxCeleb2. The list of candidate names are drawn from the VGGFace2 dataset [Cao et al., 2018], which has greater ethnic diversity compared to VGGFace1. This list contains over 9,000 identities, ranging from actors and sportspeople to politicians. There are a number of overlapping identities between VGGFace1 and VGGFace2 – these are excluded from the development set of VoxCeleb2, so that any models trained on VoxCeleb2 can be tested on VoxCeleb1.

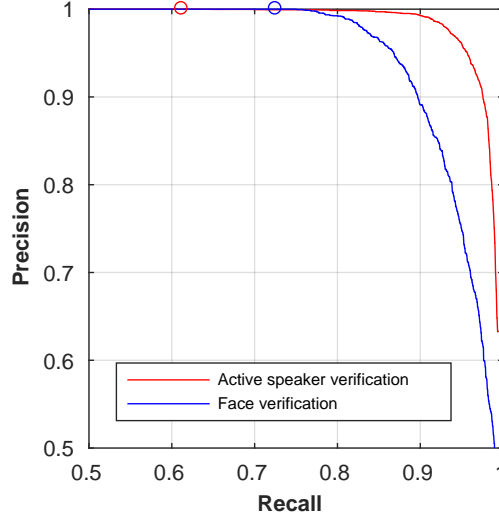


Figure 2.3: Precision-recall curves for the active speaker verification (using a 25-frame window) and the face verification steps, tested on standard benchmark datasets [Parkhi et al., 2015; Chakravarty & Tuytelaars, 2016]. Operating points are shown in circles.

2.4.2 Downloading videos from YouTube.

The top 50 or 100 videos for each of the speakers are automatically downloaded using YouTube search for VoxCeleb1 and VoxCeleb2, respectively. The word ‘interview’ is appended to the name of the speaker in search queries to increase the likelihood that the videos contain an instance of the speaker speaking, and to filter out sports or music videos. No other filtering is done at this stage.

2.4.3 Face tracking.

VoxCeleb1. The HOG-based face detector [King, 2009] is used to detect the faces in every frame of the video. Facial landmark positions are detected for each face detection using the regression tree based method of [Kazemi & Sullivan, 2014].

VoxCeleb2. The CNN face detector based on the Single Shot MultiBox Detector (SSD) [W. Liu et al., 2016] is used to detect face appearances on every frame of the video. This detector is a distinct improvement over [King, 2009], allowing the detection of faces in profile and extreme poses.

For both datasets, the shot boundaries are detected by comparing colour histograms across consecu-

tive frames. Within each detected shot, face detections are grouped together into face tracks using a position-based tracker. This stage is closely related to the tracking pipeline of [J. S. Chung & Zisserman, 2016a; Everingham et al., 2009], but optimised to reduce run-time given the very large number of videos to process.

2.4.4 Active speaker verification.

The goal of this stage is to determine the audio-video synchronisation between mouth motion and speech in a video in order to determine which (if any) visible face is the speaker. This is done by using ‘SyncNet’, a two-stream CNN described in [J. S. Chung & Zisserman, 2016b] which estimates the correlation between the audio track and the mouth motion of the video. For VoxCeleb2, the SyncNet model is replaced with a multi-view variant [J. S. Chung & Zisserman, 2017], so that talking faces can be detected even when the face is off-frontal. This method is able to reject the clips that contain dubbing or voice-over.

2.4.5 Face verification.

Active speaker face tracks are then classified into whether they are of the speaker or not using the VGGFace and VGGFace2 CNNs for VoxCeleb1 and VoxCeleb2 respectively. Verification is done by directly comparing the cosine similarity of the face embedding from the pretrained networks – the face classification networks have been trained on images of the same set of speakers (the VGGFace CNN is trained on the VGGFace image dataset, and VoxCeleb1 starts from the same list of speakers, similarly for VGGFace2).

2.4.6 Duplicate removal.

A caveat of using YouTube as a source for videos is that often the same video (or a section of a video) can be uploaded twice, albeit with different URLs. Duplicates are identified and removed as follows: each speech segment is represented by a 1024D vector using the model in [Nagrani et al., 2017] as a feature extractor. The Euclidean distance is computed between all pairs of features from the same

speaker. If any two speech segments have a distance smaller than a very conservative threshold (of 0.1), then the the speech segments are deemed to be identical, and one is removed. This method will certainly identify all exact duplicates, and in practice we find that it also succeeds in identifying near-duplicates, e.g. speech segments of the same source that are differently trimmed.

2.4.7 Manual filtering.

Since VoxCeleb1 is intended to be used as a test set for speaker verification, the data is checked manually for any errors. This is done using a simple web-based tool that shows all video segments for each identity. In order to highlight the segments which are more likely to contain errors, face and voice embeddings are generated from SphereFace [W. Liu et al., 2017] and our own model trained on VoxCeleb2 respectively, and those with lower confidence are highlighted with a different colour. By running this check, we discovered around 300 label errors, which account for around 0.2% of the VoxCeleb1 data.

2.4.8 Obtaining nationality labels.

Nationality labels are crawled from Wikipedia for all the celebrities in the dataset. We crawl for country of *citizenship*, and not *ethnicity*, as this is often more indicative of accent. In total, nationality labels are obtained for all but 428 speakers, who were labelled as unknown. Speakers in the dataset were found to hail from 36 nationalities for VoxCeleb1 and 145 for VoxCeleb2. The VoxCeleb2 is a far more ethnically diverse dataset, with a smaller percentage of U.S. speakers (29% in VoxCeleb2 compared to 64% in VoxCeleb1).

2.4.9 Discussion.

In order to ensure that our system is extremely confident that a person is speaking (Section 2.4.4), and that they have been correctly identified (Section 2.4.5) without any manual interference, we set very conservative thresholds in order to minimise the number of false positives. This conservative threshold allows us to operate in a high precision low recall regime. The large number of videos downloaded

initially allows us to discard many, and only keep the ones with extremely high confidence. Precision-recall curves for both tasks on their respective benchmark datasets [Parkhi et al., 2015; Chakravarty & Tuytelaars, 2016] are shown in Figure 2.3, and the values at the operating point are given in Table 2.5. Employing these thresholds ensures that although we discard a lot of the downloaded videos, we can be reasonably certain that the dataset has few labelling errors. Since VoxCeleb2 is designed primarily as a training-only dataset, the thresholds are less strict compared to those used to compile VoxCeleb1, so that fewer videos are discarded.

This ensures an automatic pipeline that can be scaled up to any number of speakers and utterances (if available) as required.

Task	Dataset	Precision	Recall
Active speaker verification	[Chakravarty & Tuytelaars, 2016]	1.000	0.613
Face verification	[Parkhi et al., 2015]	1.000	0.726

Table 2.5: Precision-recall values at the chosen operating points for VoxCeleb1.

2.5 VGGVox

In this section we describe our neural embedding system, called VGGVox. Our aim is to move from techniques that require traditional hand-crafted features, to a CNN architecture that can train end-to-end for the task of speaker recognition. The system is trained on short-term magnitude spectrograms extracted directly from raw audio segments, with no other pre-processing. A deep neural network trunk architecture is used to extract frame level features, and the features are aggregated to obtain utterance-level speaker embeddings. The entire model is then trained end-to-end.

We use 2D CNNs as feature extractors and treat 2D spectrograms as single-channel images. It is perhaps unnatural to treat spectrograms in this manner where the same convolution is used at every point since, unlike in a visual image where an object may appear at any location, a pattern can appear at any point on the time axis but we would not expect patterns to also be frequency independent. However, deep networks can potentially learn frequency-specific filters if they are needed for solving a downstream task; for instance, some filters can only fire on specific patterns existing in the low frequency region, whilst fully connected layers can be position dependent. Therefore, even if a 2D

CNN uses shared filters on the spectrogram, the model has the capability to divide the filters into low/high frequency groups.

We experiment with different trunk architectures, aggregation strategies as well as training losses. We describe the trunk and aggregation architectures here, and the losses in section 2.6.2.

2.5.1 Input features.

We use short-term magnitude spectrograms as input to our deep CNN architecture. Mean and variance normalisation is performed on every frequency bin of the spectrum. No other speech-specific preprocessing (e.g. silence removal, voice activity detection, or removal of unvoiced speech) is used. Precise implementation details are provided in section 2.6.3.

2.5.2 Trunk Architecture

We experiment with both VGG [Chatfield et al., 2014] and ResNet style CNN architectures.

VGG-M: The baseline trunk architecture is the CNN introduced in [Nagrani et al., 2017]. This architecture is a modification of the VGG-M [Chatfield et al., 2014] CNN, known for high efficiency and good performance on image classification. The modification concerns the addition of an aggregation layer, and is described below. The complete CNN architecture is specified in Table 2.6.

ResNets: The residual-network (ResNet) architecture [K. He et al., 2016] is similar to a standard multi-layer CNN, but with added skip connections such that the layers add residuals to an identity mapping on the channel outputs. In this paper, we experiment with three variants of ResNets, e.g. ResNet-34, ResNet-50 and a Thin-ResNet which contains fewer parameters. We modify the layers to adapt to the spectrogram input. The architectures are specified in Table 2.7.

2.5.3 Aggregation Strategies

Features produced by the trunk CNN architecture are then aggregated in time to produce a single fixed length representation for each audio input. We experiment with two aggregation strategies: simple non-trainable average pooling, as well as a trainable aggregation layer based on the NetVLAD layer.

Layer	Support	Filt dim.	# filts.	Stride	Data size
conv1	7×7	1	96	2×2	254×148
mpool1	3×3	-	-	2×2	126×73
conv2	5×5	96	256	2×2	62×36
mpool2	3×3	-	-	2×2	30×17
conv3	3×3	256	384	1×1	30×17
conv4	3×3	384	256	1×1	30×17
conv5	3×3	256	256	1×1	30×17
mpool5	5×3	-	-	3×2	9×8
fc6	9×1	256	4096	1×1	1×8
apool6	$1 \times n$	-	-	1×1	1×1
fc7	1×1	4096	1024	1×1	1×1
fc8	1×1	1024	1251	1×1	1×1

Table 2.6: VGG style architecture. The data size on the right is the *output* data size for each layer. Here we assume input spectrograms of size 512×300 , and up to *fc6* the sizes have been calculated for an input with a temporal dimension of 300, but the network is able to accept inputs of variable lengths. Note that the first layer also has zero padding.

Here we provide a brief overview of both the average pooling aggregation layer, and also the NetVLAD (for full details please refer to [Arandjelović et al., 2016]).

Average pooling aggregation The fully connected *fc6* layer from the original VGG-M is replaced by two layers – a fully connected layer of 9×1 (support in the frequency domain), and an aggregation layer – global average pooling along the temporal axis. The benefit of this modification is that the network becomes invariant to temporal position but *not* frequency, which is desirable for speech, but not for images. It also helps to keep the output dimensions the same as those of the original fully connected layer, and reduces the number of network parameters for our given input size this reduction is fivefold, i.e. from 319M in VGG-M to 67M in our network) which helps avoid overfitting.

NetVLAD aggregation The CNN trunk architecture maps the input spectrogram to frame-level descriptors, as described in the Thin-ResNet shown in Tabel 2.7, the output feature is downsampled by a factor of 32. The NetVLAD layer then takes dense descriptors as input and produces a single $K \times D$ matrix V , where K refers to the number of chosen cluster, and D refers to the dimensionality of each

layer name	ResNet34	ResNet50	Thin-ResNet
conv1	$7 \times 7, 64$, stride 2	$7 \times 7, 64$, stride 2	$7 \times 7, 64$, stride 1
pool1	3×3 , max pool stride 2	3×3 , max pool stride 2	3×3 , max pool stride 2
conv2_x	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 48 \\ 3 \times 3, 48 \\ 1 \times 1, 96 \end{bmatrix} \times 2$
conv3_x	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 96 \\ 3 \times 3, 96 \\ 1 \times 1, 128 \end{bmatrix} \times 3$
conv4_x	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv5_x	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 512 \end{bmatrix} \times 3$
fc1	$9 \times 1, 512$, stride 1	$9 \times 1, 2048$, stride 1	$7 \times 1, 512$, stride 1
pool_time	$1 \times N$, avg pool stride 1	$1 \times N$, avg pool stride 1	$3 \times N$, max pool stride 2

Table 2.7: Modified ResNet34, ResNet50 and Thin-ResNet architectures with average pool layer at the end. Batch normalisation is used before the rectified linear unit (ReLU) activations. Each row specifies the number of convolutional filters and their sizes as **size** \times **size**, # **filters**. Square brackets indicate blocks over which there are residual connections.

cluster. Concretely, the matrix of descriptors V is computed using the following equation:

$$V(k, j) = \sum_{t=1}^{T/32} \frac{e^{w_k^T x_t + b_k}}{\sum_{k'=1}^K e^{w_k^T x_t + b_{k'}}} (x_t(j) - c_k(j)) \quad (2.1)$$

where $\{w_k\}$, $\{b_k\}$ and $\{c_k\}$ are trainable parameters, with $k \in [1, 2, \dots, K]$. The first term corresponds to the soft-assignment weight of the input vector x_i for cluster k , while the second term computes the residual between the vector and the cluster centre. Each row in V , i.e. the residual from each cluster is then L2 normalized. The final output is then obtained by flattening this matrix into a long vector, i.e. row vectors are concatenated. To keep computational and memory requirements low, dimensionality reduction is performed via a Fully Connected (FC) layer, where we pick the output dimensionality to be 512. We also experiment with the recently proposed **GhostVLAD** [Zhong et al., 2018] layer, where some of the clusters are not included in the final concatenation, and so do not contribute to the final representation, these are referred to as ‘ghost clusters’ (we used *two* in our implementation). Therefore,

Dataset	# of speakers	# of utterances	# of pairs
VoxCeleb1	40	4,715	37,720
VoxCeleb1 (cleaned)	40	4,708	37,611
VoxCeleb1-E	1,251	145,375	581,480
VoxCeleb1-E (cleaned)	1,251	145,160	579,818
VoxCeleb1-H	1,190	138,137	552,536
VoxCeleb1-H (cleaned)	1,190	137,924	550,894

Table 2.8: VoxCeleb test sets.

while aggregating the frame-level features, the contribution of the noisy and undesirable sections of a speech segment to normal VLAD clusters is effectively down-weighted, as most of their weights have been assigned to the ‘ghost cluster’. For further details, please refer to [Zhong et al., 2018].

2.6 Experiments

This section describes our experimental setup for speaker verification, loss functions, baselines, and implementation details. Along with releasing the VoxCeleb dataset, we also release a number of different evaluation benchmarks for testing speaker verification. These have been used extensively by the speech community to compare methods. In particular, we provide both easy pairs and hard pairs for testing; for the hard pairs, speakers with the same nationality and gender are chosen which makes distinguishing between them more challenging. This is described in more detail in the next section.

2.6.1 Evaluation Splits and Metrics

The methods are evaluated on a number of different test sets. These are described below and summarised in Table 2.8. All test set lists can be found on the VoxCeleb website².

Original VoxCeleb1 test set. The original verification test set from VoxCeleb1 consists of 40 speakers. All speakers with names starting with ‘E’ are reserved for testing, since this gives a good balance of male and female speakers.

Extended VoxCeleb1-E test set – using the entire dataset. Since the above test set is limited in the

²<http://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html>

number of speakers, there is a danger that models achieving high performance on this test set might not generalise to other sets of speakers. Hence we also propose a larger test set of 581,480 random pairs sampled from the entire VoxCeleb1 dataset, covering 1,251 speakers.

Hard VoxCeleb1-H test set – within the same nationality and gender. This is a ‘hard’ evaluation set consisting of 552,536 pairs with the same nationality and gender, sampled from the entire VoxCeleb1 dataset. There are 18 nationality-gender combinations each with at least 5 individuals, of which ‘USA-Male’ is the most common.

Evaluation Metric. We evaluate the models with Equal Error Rate (EER) and the minimum detection cost function (minDCF). EER measures the value at which the false-reject (miss) rate equals the false-accept (false-alarm) rate, and minDCF is defined as a weighted sum of false-reject and false-accept error probabilities. These are common metrics used by existing datasets and challenges, such as NIST SRE12 [Greenberg, 2012] and SITW [McLaren et al., 2016].

2.6.2 Training Loss

We experiment with a number of different training losses.

Softmax + Contrastive Loss. We employ a contrastive loss [Chopra et al., 2005; Hadsell et al., 2006] on paired embeddings, which seeks to minimise the distance between the embeddings of positive pairs and penalises the negative pair distances for being smaller than a margin parameter α . Pair-wise losses such as the contrastive loss are notoriously difficult to train [Hermans et al., 2017], and hence to avoid suboptimal local minima early on in training, we proceed in two stages: first, pre-training for identification using a softmax loss, then, second, fine-tuning with the contrastive loss (described in more detail below).

Additive Margin Softmax. Besides the standard softmax loss, we also experiment with the additive margin softmax (AM-Softmax) classification loss [F. Wang et al., 2018] during training. This loss is designed explicitly for improving verification performance by introducing a margin in the angular space, meaning that we do not need to train with the contrastive loss after. The loss is given by the following equation:

$$L_i = -\log \frac{e^{s(\cos \theta_{y_i} - m)}}{e^{s(\cos \theta_{y_i} - m)} + \sum_{j \neq y_i} e^{s \cos(\theta_j)}} \quad (2.2)$$

where L_i refers to cost of assigning the sample to the correct class, $\theta_y = \arccos(w^T x)$ refers to the angle between sample features (x) and the decision hyperplane (w), as both vectors have been L2 normalised. The goal is therefore to minimise this angle by making $\cos(\theta_{y_i}) - m$ as large as possible, where m refers to the angular margin. The hyper-parameter s controls the “temperature” of the softmax loss, producing higher gradients to the well-separated samples (and further shrinking the intra-class variance). We used the default values $m = 0.4$ and $s = 30$ [F. Wang et al., 2018].

Relation Loss.

In this work, as another contribution, we introduce a novel relation module as a scoring mechanism. It is similar to a contrastive loss function, but uses a simple binary classifier rather than Euclidean distance. The relation module is shown in Figure 2.4. It is inspired by the relation networks and their use in face comparisons [Santoro et al., 2017a; W. Xie et al., 2018].

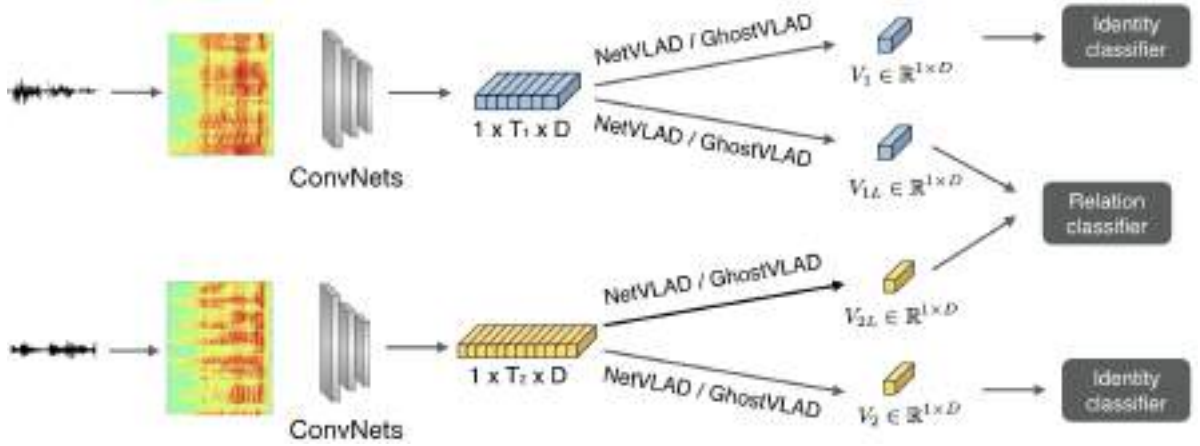


Figure 2.4: Relation module. The relation module is added to a Siamese network in training and inference. During training, two sets of aggregation modules are used: V_1 and V_2 are computed from a shared NetVLAD/GhostVLAD and trained for identity classification; V_{1L} and V_{2L} are computed from a shared NetVLAD/GhostVLAD and used to train the relation classifier for identity matching.

A Siamese network is constructed from two standard classification models, *i.e.* two branches share

the same network and parameters ThinResNet is fixed until conv4_x, refer to Table 2.7 that have been pretrained for speaker classification based on standard softmax, and the small relation module is then trained to distinguish if two voice samples are from same identity or not (binary classifier, implemented as a softmax with two classes). A separate NetVLAD/GhostVLAD aggregator is incorporated for the classification and relation network paths. As most of the feature extractor are fixed, the relation module only costs a very limited additional computation. During inference the output scores of a voice pair is computed as the average of the cosine similarity (between feature embeddings) and the classification score (from the small relation module).

2.6.3 Implementation details and training

During training, we randomly sample segments from each utterance. For the VGG based model, we use 3-second long segments with a 1024 FFT giving us spectrograms of size 512×300 , and for the Thin-Resnet model we use 512 point FFTs giving us spectrograms of size 257×250 (frequency \times temporal). Earlier models (ResNet34 and ResNet50) are trained using the deep learning toolbox MatConvNet [Vedaldi & Lenc, 2014], and the latest models (Thin-ResNet) are in Keras (tensorflow).³ The models and training code are publically available⁴. The model is trained using a fixed size spectrogram corresponding to a 2.5 second interval. All audio is first converted to single-channel, 16-bit streams at a 16kHz sampling rate for consistency. Spectrograms are then generated in a sliding window fashion using a hamming window of width 25ms and step 10ms. We normalise the spectrogram by subtracting the mean and dividing by the standard deviation of all frequency components in a single time step. No voice activity detection (VAD), or automatic silence removal is applied. We use the Adam optimizer with an initial learning rate of 0.001, and decrease the learning rate by a factor of 10 after every 36 epochs until convergence.

Pre-training for contrastive loss Our first strategy is to use softmax pre-training to initialise the weights of the network. The cross entropy loss produces more stable convergence than the contrastive

³For our earlier spectrogram generation code, we discard the DC component, i.e. a 1024 pt FFT gives us spectrograms with $1024/2$ frequency channels, but in later models we added the DC component to the spectrogram, hence for a 512 pt FFT we get spectrograms with $512/2 + 1$ frequency channels.

⁴<http://www.robots.ox.ac.uk/~vgg/research/speakerID/> [W. Xie et al., 2019]

loss, possibly because softmax training is not impacted by the difficulty of pairs when using the contrastive loss. To evaluate the identification performance, we create a held-out validation test which consists of all the speech segments from a single video for each identity.

We take the model pre-trained on the identification task, and replace the classification layer with a fully connected layer of output dimension 512. This network is then trained with the contrastive loss.

Mining hard examples A key challenge associated with learning embeddings via the contrastive loss is that as the dataset gets larger, the number of possible pairs grows quadratically. In such a scenario, the network rapidly learns to correctly map the easy examples, and hard negative mining is often required to improve performance to provide the network with a more useful learning signal. We use an offline hard negative mining strategy, which allows us to select harder negatives (*e.g.* top 1-percent of randomly generated pairs) than is possible with online (in-batch) hard negative mining methods [Sung, 1996; Hermans et al., 2017; H. O. Song et al., 2016] limited by the batch size. We do not mine hard positives, since false positive pairs are much more likely to occur than false negative pairs in a random sample (due to possible label noise on the face verification), and these label errors will lead to poor learning dynamics.

While training the relation module, a similar strategy is applied, we pre-compute the feature embeddings for all the voice samples in the entire VoxCeleb2 dataset. In addition to negative pairs, we mine both hard positive and negative pairs for training relation modules.

2.6.4 Non deep learning based baselines

For the sake of comparison, we also implement some traditional non-CNN methods and train them on the VoxCeleb1 dev set.

GMM-UBM. The GMM-UBM system uses MFCCs of dimension 13 as input. Cepstral mean and variance normalisation (CMVN) is applied on the features. Using the conventional GMM-UBM framework, a single speaker-independent universal background model (UBM) of 1024 mixture components is trained for 10 iterations from the training data.

I-vectors/PLDA. Gender independent i-vector extractors [Dehak et al., 2011] are trained on the VoxCeleb1

dataset to produce 400-dimensional i-vectors. Probabilistic LDA (PLDA) [Ioffe, 2006] is then used to reduce the dimension of the i-vectors to 200.

Inference. For identification, a one-vs-rest binary SVM classifier is trained for each speaker m ($m \in 1 \dots K$). All feature inputs to the SVM are L2 normalised and a held out validation set is used to determine the C parameter (determines trade off between maximising the margin and penalising training errors). Classification during test time is done by choosing the speaker corresponding to the highest SVM score. The PLDA scoring function [Ioffe, 2006] is used for verification.

2.7 Results

In this section, we show all the evaluation results on three publicly available test sets created from VoxCeleb1, i.e. VoxCeleb1 test-set, VoxCeleb1-E, VoxCeleb1-H. Discussions of our main observations from these experiments are included, e.g. benefits from the end-to-end trained CNN, size of training data, network architecture, different loss functions, and choice of aggregation strategy. We also compare performance of the CNN architectures to a number of other deep learning methods and more traditional state of the art methods.

2.7.1 Results on VoxCeleb1

2.7.1.1 Comparison to Non-CNN Methods

Comparing with the baseline methods that are based on traditional methods, e.g. GMM-UBM, I-vectors+PLDA, achieving 15.0% EER and 8.8% EER on the standard VoxCeleb1 testing set respectively, most of the Neural Networks (NN) based methods have shown clear advantages, for instance, one of our earliest VGG-M models [Nagrani et al., 2017] trained with Softmax and Contrastive has outperformed the traditional methods (obtaining 7.8% EER).

VoxCeleb1 test set						
	Front-end model	Loss	Dims	Aggregation	Training set	EER (%)
INTERSPEECH17 [Nagrani et al., 2017]	GMM-UBM	–	–	–	VoxCeleb1	15.0
INTERSPEECH17 [Nagrani et al., 2017]	I-vectors+PLDA	–	–	–	VoxCeleb1	8.8
INTERSPEECH17 [Nagrani et al., 2017]	VGG-M	Softmax	1024	TAP	VoxCeleb1	10.2
INTERSPEECH17 [Nagrani et al., 2017]	VGG-M	Softmax+Contr.	1024	TAP	VoxCeleb1	7.8
INTERSPEECH18 [J. S. Chung et al., 2018]	VGG-M	Softmax+Contr.	1024	TAP	VoxCeleb2	5.94
INTERSPEECH18 [J. S. Chung et al., 2018]	ResNet-34	Softmax+Contr.	512	TAP	VoxCeleb2	5.04
INTERSPEECH18 [J. S. Chung et al., 2018] TTA-2	ResNet-34	Softmax+Contr.	512	TAP	VoxCeleb2	5.11
INTERSPEECH18 [J. S. Chung et al., 2018] TTA-3	ResNet-34	Softmax+Contr.	512	TAP	VoxCeleb2	4.83
INTERSPEECH18 [J. S. Chung et al., 2018]	ResNet-50	Softmax+Contr.	512	TAP	VoxCeleb2	4.19
INTERSPEECH18 [J. S. Chung et al., 2018] TTA-2	ResNet-50	Softmax+Contr.	512	TAP	VoxCeleb2	4.43
INTERSPEECH18 [J. S. Chung et al., 2018] TTA-3	ResNet-50	Softmax+Contr.	512	TAP	VoxCeleb2	3.95
ICASSP19 [W. Xie et al., 2019]	Thin-ResNet-34	Softmax	512	TAP	VoxCeleb2	10.48
ICASSP19 [W. Xie et al., 2019]	Thin-ResNet-34	Softmax	512	NetVLAD	VoxCeleb2	3.57
ICASSP19 [W. Xie et al., 2019]	Thin-ResNet-34	AM-Softmax	512	NetVLAD	VoxCeleb2	3.32
ICASSP19 [W. Xie et al., 2019]	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	3.22
ICASSP19 [W. Xie et al., 2019]	Thin-ResNet-34	AM-Softmax	512	GhostVLAD	VoxCeleb2	3.23
ICASSP19 (cleaned †) [W. Xie et al., 2019]	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	3.24
Ours + Relation Module	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	2.87

VoxCeleb1-E						
INTERSPEECH18 [J. S. Chung et al., 2018]	ResNet-50	Softmax+Contr.	512	TAP	VoxCeleb2	4.42
ICASSP19	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	3.24
ICASSP19 (cleaned †)	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	3.13
Ours + Relation Module †	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	2.95

VoxCeleb1-H						
INTERSPEECH18 [J. S. Chung et al., 2018]	ResNet-50	Softmax+Contr.	512	TAP	VoxCeleb2	7.33
ICASSP19 [W. Xie et al., 2019]	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	5.17
ICASSP19 (cleaned †) [W. Xie et al., 2019]	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	5.06
Ours + Relation Module †	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	4.93

Table 2.9: Comparison of our different models for verification on the original VoxCeleb1 test set [Nagrani et al., 2017] and the extended and hard test sets (VoxCeleb-E and VoxCeleb-H) [J. S. Chung et al., 2018]. TAP: Temporal Average Pooling. TTA: Test Time Augmentation. Contr.: Contrastive Loss † Cleaned up versions of the test lists have been released publicly. We encourage other researchers to evaluate on these lists.

2.7.1.2 Size of Training Data

Deep neural networks are well-known for their capability to process large amount of data, in this section, we focus on exploring the benefits from a large dataset. In our experiments, we keep all settings unchanged, and only varying the datasets from VoxCeleb1 to VoxCeleb2, e.g. VGG-M model trained with Softmax and Contrastive [Nagrani et al., 2017; J. S. Chung et al., 2018]. When testing on the standard VoxCeleb1 test set, larger training set (VoxCeleb2) leads to better performance (5.94 % EER vs. 7.8 % EER). This is due to the fact that we expect a larger dataset to naturally provide more diversity and variation, and therefore lead to better generalization.

2.7.1.3 Effect of CNN Architecture

Following the continuous development of new architectures in computer vision, we also experiment with different trunk architectures, ranging from VGG to ResNet. In this comparison, we fix all the experimental settings and only vary the network architecture, i.e. we compare three models (VGG-M, ResNet-34, ResNet-50) trained with temporal average pooling (TAP), and Softmax+Contrastive loss on VoxCeleb2. Evaluation is done on the standard VoxCeleb1 test set without any test-time augmentation. Similar to the observations found in computer vision research, deeper networks lead to better generalization, therefore, ResNet-50 (4.19 % EER) outperforms the ResNet-34 (5.04 % EER) and VGG-M (5.94 % EER), despite the fact that the VGG-M model uses a higher dimensional embedding (1024D).

2.7.1.4 Aggregation Strategy and Training Loss

We next explore different aggregation methods and loss functions in this section. Once again, other experimental settings are fixed, for instance, we train the same Thin-ResNet on VoxCeleb2 and only vary the aggregation strategy (TAP, NetVLAD, GhostVLAD) and training loss (Softmax vs. AMSoftmax). As shown in Table 2.9, the Thin-ResNet trained with standard softmax loss and NetVLAD aggregation layer outperforms the previous model [J. S. Chung et al., 2018] by a significant margin (EER of 3.57% vs 4.19%). The fact that the Thin-ResNet is actually shallower than the ResNet-50 (Table 2.7), and contain fewer number of parameters, further illustrates the benefits of the NetVLAD aggregation layer. By replacing the standard softmax with the additive margin softmax (AM-Softmax), a further performance gain is achieved (3.32% EER). The GhostVLAD layer, which excludes irrelevant information from the aggregation, additionally makes a modest contribution to performance (3.22% EER).

On the challenging VoxCeleb1-H test set, we outperform the previous best architecture [J. S. Chung et al., 2018] (EER of 5.17% vs 7.33%), which is by a larger margin than on the original VoxCeleb1 test set. We note that training a softmax loss based on features from temporal average pooling (TAP) yields extremely poor results (EER of 10.48%). We conjecture that the features trained using a softmax loss are typically good at separating different speakers), but not good at reducing the intra-class vari-

ation (i.e. making features of the same speaker compact). Therefore, contrastive loss with online hard sample mining leads to a significant performance boost, as demonstrated in [J. S. Chung et al., 2018] for TAP.

2.7.1.5 Test Time Augmentation

Here, we experiment with different augmentation protocols for evaluating the performance at test time. We propose three methods:

Baseline: Here we use variable average pooling where we evaluate the entire test utterance at once, by changing the size of an average pooling layer during test time according to the length of the test sample;

(TTA-2) Here we sample ten 3-second temporal crops from each test segment, and take the mean of the final embeddings;

(TTA-3) Here we sample ten 3-second temporal crops from each test segment, compute the distances between the every possible pair of crops ($10 \times 10 = 100$) from the two speech segments, and use the mean of the 100 distances. This final method results in a marginal improvement in performance, as shown in Table 2.9.

2.7.1.6 Comparison with State-of-the-art Models

In Table 2.10, we compare with the recent state-of-the-art models based on TDNN and x-vectors [Snyder et al., 2018] on the standard VoxCeleb1 test set. While only training on Voxceleb2, our Thin-ResNet-34 [W. Xie et al., 2019] achieves comparable performance to the model based on x-vectors trained on Voxceleb1&2 (EER 3.2 vs. EER 3.1).

As a fair comparison to our original model [W. Xie et al., 2019], we only train the relation module on the Voxceleb2 dataset. However, as explored in previous sections, we expect that incorporating both Voxceleb1 and Voxceleb2 can further boost the performance of all of our models. Overall, our Thin-ResNet with a GhostVLAD layer and a relation module currently holds the state-of-the-art result on the VoxCeleb1 dataset (Table 2.10)

VoxCeleb1 test set							
	Front-end model	Loss	Dims	Aggregation	Training set	EER (%)	minDCF†
Cai et al. [Cai, Chen, & Li, 2018b]	ResNet-34	A-Softmax+PLDA	128	TAP	VoxCeleb1	4.46	-
Cai et al. [Cai, Chen, & Li, 2018b]	ResNet-34	A-Softmax+PLDA	128	SAP	VoxCeleb1	4.40	-
Cai et al. [Cai, Chen, & Li, 2018b]	ResNet-34	A-Softmax+PLDA	128	LDE	VoxCeleb1	4.48	-
Okabe et al. [Okabe et al., 2018]	TDNN (x-vector)	Softmax	1500	TAP	VoxCeleb1	4.70	-
Okabe et al. [Okabe et al., 2018]	TDNN (x-vector)	Softmax	1500	SAP	VoxCeleb1	4.19	-
Okabe et al. [Okabe et al., 2018]	TDNN (x-vector)	Softmax	1500	ASP	VoxCeleb1	3.85	-
Hajibabaei et al. [Hajibabaei & Dai, 2018]	ResNet20	A-Softmax	128	TAP	VoxCeleb1	4.40	-
Hajibabaei et al. [Hajibabaei & Dai, 2018]	ResNet20	AM-Softmax	128	TAP	VoxCeleb1	4.30	-
Synder et al. [Synder et al., 2018]	TDNN (x-vector)	Softmax	1500	SP	VoxCeleb1	3.10	0.33
					VoxCeleb2		
					MUSAN RIR_NOISES		
Ours (ICASSP19 [W. Xie et al., 2019])	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	3.22	0.35
Ours + Relation Module	Thin-ResNet-34	Softmax	512	GhostVLAD	VoxCeleb2	2.87	0.31

Table 2.10: Comparison of our best performing model to the state-of-the-art on the VoxCeleb1 original test set. TAP: Temporal Average Pooling. SAP: Self-attentive Pooling Layer [Cai, Chen, & Li, 2018b]. SP: Statistical Pooling. TTA: Test Time Augmentation. † We calculate minDCF at (0.01) using the standard parameters used in the NIST SRE 18.^a

^ahttps://www.nist.gov/sites/default/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf

2.8 Conclusion

In this paper we have introduced a scalable method to automatically generate a speaker recognition dataset, and used it to produce the VoxCeleb1 and VoxCeleb2 datasets, which are several times larger than any other speaker recognition dataset. These datasets have become a standard for the speech community to train and evaluate speaker recognition performance on. They have also played a large part in the recent NIST-SRE challenge in 2018. As mentioned by [K. A. Lee et al., 2019], introducing the ‘VAST partition’ in SRE18, comprising the VoxCeleb and SITW datasets, represents a ‘new initiative towards speaker recognition in the wild’, since ‘a signature feature of the VAST partition is multi-speaker conversation with considerable background noise.’ The VoxCeleb datasets are also the subject of the first VoxSRC challenge to be held at Interspeech 2019. We believe that the use of these datasets in challenges has allowed a paradigm shift in speaker recognition efforts in the community, encouraging the development of systems under noisy and ‘in-the-wild’ conditions.

We have also introduced new architectures and training strategies for the task of speaker verification. Our learnt identity embeddings are compact (512D) and hence easy to store and useful for other tasks such as diarisation and retrieval.

The relation module, also introduced in this paper, has been shown to outperform all previous models

by a significant margin on the VoxCeleb1 dataset.

Whilst our models are based on 2D convolutions applied to spectrogram inputs, further work will involve investigating alternatives that may be more efficient, such as 1D time convolutions with the frequencies of the spectrogram arranged as input channels, or 1D convolutions applied to raw waveforms directly.

We have publicly released all code, models and data.

Acknowledgements

Funding for this research is provided by the EPSRC Programme Grant Seebibyte EP/M013774/1. q aA

Statement of Authorship

A statement of authorship for this work can be found in Appendix [C](#).

3 | *Speech2Action*: Cross-modal Supervision for Action Recognition

Arsha Nagrani¹ Chen Sun² David Ross²
Rahul Sukthankar² Cordelia Schmid² Andrew Zisserman^{1,3}

¹VGG, Oxford ²Google Research ³DeepMind

Abstract

Is it possible to guess human action from dialogue alone? In this work we investigate the link between spoken words and actions in movies. We note that movie screenplays describe actions, as well as contain the speech of characters and hence can be used to learn this correlation with no additional supervision. We train a BERT-based *Speech2Action* classifier on over a thousand movie screenplays, to predict action labels from transcribed speech segments. We then apply this model to the speech segments of a large unlabelled movie corpus (188M speech segments from 288K movies). Using the predictions of this model, we obtain weak action labels for over 800K video clips. By training on these video clips, we demonstrate superior action recognition performance on standard action recognition benchmarks, without using a single manually labelled action example.

Published in the Proceedings of the [Computer Vision and Pattern Recognition Conference 2020](#).

3.1 Introduction

Often, you can get a sense of human activity in a movie by listening to the dialogue alone. For example, the sentence ‘*Look at that spot over there*’, is an indication that somebody is

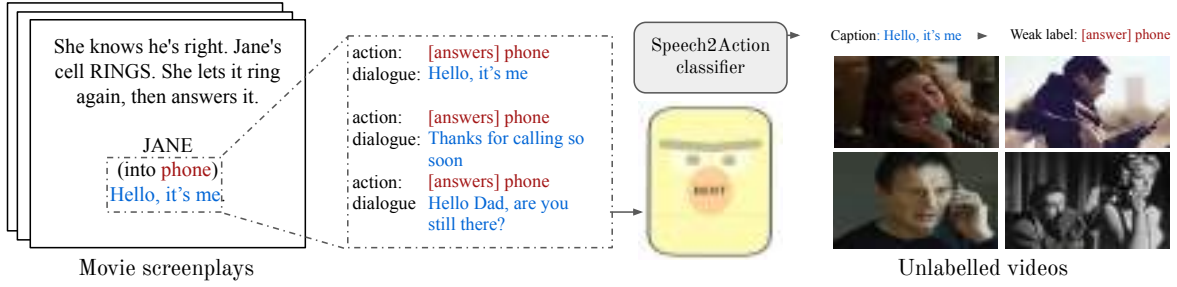


Figure 3.1: **Weakly Supervised Learning of Actions from Speech Alone:** The co-occurrence of speech and scene descriptions in movie screenplays (text) is used to learn a Speech2Action model that predicts actions from transcribed speech *alone*. Weak labels for visual actions can then be obtained by applying this model to the speech in a large *unlabelled* set of movies.

pointing at something. Similarly, the words ‘Hello, thanks for calling’, is a good indication that somebody is speaking on the phone. Could this be a valuable source of information for learning good action recognition models?

Obtaining large scale human labelled video datasets to train models for visual action recognition is a notoriously challenging task. While large datasets, such as Kinetics [Kay et al., 2017] or Moments in Time [Monfort et al., 2019] consisting of individual short clips (e.g. 10s) are now available, these datasets come at formidable human cost and effort. Furthermore, many such datasets suffer from heavily skewed distributions with long tails – i.e. it is difficult to obtain manual labels for *rare* or *infrequent* actions [Gu et al., 2018].

Recently, a number of works have creatively identified *certain domains* of videos, such as *narrated instructional videos* [Miech, Zhukov, et al., 2019; Tang et al., 2019; L. Zhou et al., 2018] and *lifestyle vlogs* [Ignat et al., 2019; Fouhey et al., 2018] that are available in huge numbers (e.g. on YouTube) and often contain narration with the explicit intention of explaining the visual content on screen. In these video domains, there is a direct link between the action being performed, and the speech accompanying the video – though this link, and the visual supervision it provides, can be quite weak and ‘noisy’ as the speech may refer to previous or forthcoming visual events, or be about something else entirely [Miech, Zhukov, et al., 2019].

In this paper we explore a complementary link between speech and actions in the more general domain of movies and TV shows (not restricted to instructional videos and vlogs). We ask: is it possible given only a speech sentence to predict whether an action is happening, and, if so, what the action is? While it appears that in some cases the speech is correlated with action – ‘*Raise your glasses to ...*’, in the more general domain of movies and TV shows it is *more* likely that the speech is completely uncorrelated with the action – ‘*How is your day going?*’. Hence in this work, we *explicitly* learn to identify when the speech is discriminative. While the supervision we obtain from the speech–action correlation is still noisy, we show that at scale it can provide sufficient weak supervision to train visual classifiers (see Fig. 3.1).

Luckily, we have a large amount of literary content at our disposal to learn this correlation between speech and actions. Screenplays can be found for hundreds of movies and TV shows and contain rich descriptions of the identities of people, their actions and interactions with one another and their dialogue. Early work has attempted to *align* these screenplays to the videos themselves, and use that as a source of weak supervision [Bojanowski et al., 2013; Duchenne et al., 2009; Laptev et al., 2008; Marszałek et al., 2009]. However, this is challenging due to the lack of explicit correspondence between scene elements in video and their textual descriptions in screenplays [Bojanowski et al., 2013], and notwithstanding alignment quality, is also fundamentally limited in scale to the amount of aligned movie screenplays available. Instead we learn from *unaligned* movie screenplays. We *first* learn the correlation between speech and actions from written material *alone* and use this to train a `Speech2Action` classifier. This classifier is then applied to the speech in an unlabelled, unaligned set of videos to obtain visual samples corresponding to the actions confidently predicted from the speech (Fig. 9.1). In this manner, the correlations can provide us with an effectively infinite source of weak training data, since the audio is freely available with movies.

Concretely, we make the following four contributions: (i) We train a `Speech2Action` model from literary screenplays, and show that it is possible to predict certain actions from transcribed speech *alone* without the need for any manual labelling; (ii) We apply the

Speech2Action model to a large unlabelled corpus of videos to obtain weak labels for video clips from the speech alone; (iii) We demonstrate that an action classifier trained with these weak labels achieves state of the art results for action classification when fine-tuned on standard benchmarks compared to other weakly supervised/domain transfer methods; (iv) Finally, and more interestingly, we evaluate the action classifier trained only on these weak labels with *no* fine-tuning on the mid and tail classes from the AVA dataset [Gu et al., 2018] in the zero-shot and few-shot setting, and show a large boost over fully supervised performance for some classes without using a *single* manually labelled example.

3.2 Related Works

Aligning Screenplays to Movies: A number of works have explored the use of screenplays to learn and automatically annotate character identity in TV series [Everingham et al., 2006; Naim et al., 2016; Cour et al., 2009; Sivic et al., 2009; Tapaswi et al., 2012]. Learning human actions from screenplays has also been attempted [Bojanowski et al., 2013; Duchenne et al., 2009; Laptev et al., 2008; Marszałek et al., 2009; Miech, Alayrac, et al., 2017]. Crucially, however, all these works rely on aligning these screenplays to the actual videos themselves, often using the speech (as subtitles) to provide correspondences. However, as noted by [Bojanowski et al., 2013], obtaining supervision for actions in this manner is challenging due to the lack of explicit correspondence between scene elements in video and their textual descriptions in screenplays.

Apart from the imprecise temporal localization inferred from subtitles correspondences, a major limitation is that this method is not scalable to all movies and TV shows, since screenplays with stage directions are simply not available at the same order of magnitude. Hence previous works have been limited to a small scale, no more than *tens* of movies or a season of a TV series [Bojanowski et al., 2013; Duchenne et al., 2009; Laptev et al., 2008; Marszałek et al., 2009; Miech, Alayrac, et al., 2017]. A similar argument can be applied to works that align

books to movies [Zhu et al., 2015; Tapaswi et al., 2015]. In contrast, we propose a method that can exploit the richness of information in a modest number of screenplays, and then be applied to a virtually limitless set of edited video material with no alignment or manual annotation required.

Supervision for Action Recognition: The benefits of learning from large scale supervised video datasets for the task of action recognition are well known, with the introduction of datasets like Kinetics [Kay et al., 2017] spurring the development of new network architectures yielding impressive performance gains, e.g. [Carreira & Zisserman, 2017; S. Xie et al., 2018; Tran et al., 2018; X. Wang et al., 2018; L. Wang, Xiong, et al., 2016; Feichtenhofer et al., 2019]. However, as described in the introduction, such datasets come with an exorbitant labelling cost. Some work has attempted to reduce this labeling effort through heuristics [H. Zhao et al., 2017] (although a human annotator is required to clean up the final labels) or by procuring weak labels in the form of accompanying meta data such as hashtags [Ghadiyaram et al., 2019].

There has *also* been a recent growing interest in using *cross-modal supervision* from the audio streams readily available with videos [Owens et al., 2016; H. Zhao et al., 2018; Arandjelović & Zisserman, 2017; Owens & Efros, 2018; Korbar et al., 2018]. Such methods, however, focus on *non-speech* audio, e.g. ‘guitar playing’, the ‘thud’ of a bouncing ball or the ‘crash’ of waves at the seaside, rather the transcribed speech. As discussed in the introduction, transcribed speech is used only in certain narrow domains, e.g. instruction videos [Miech, Zhukov, et al., 2019; Tang et al., 2019; L. Zhou et al., 2018] and lifestyle vlogs [Ignat et al., 2019; Fouhey et al., 2018], while in contrast to these works, we focus on the domain of movies and TV shows (where the link between speech and actions is less explicit). Further, such methods use most or *all* the speech accompanying a video to learn a better overall visual embedding, whereas we note that often the speech is completely uninformative of the action. Hence we *first* learn the correlation between speech and actions from written material, and then apply

this knowledge to an unlabelled set of videos to obtain video clips that can be used directly for training.

3.3 Speech2Action Model

In this section we describe the steps in data preparation, data mining and learning, required to train the `Speech2Action` classifier from a large scale dataset of screenplays. We then assess its performance in predicting visual actions from transcribed speech segments.

# movies	# scene desc.	# speech seg.	# sentences	# words	# unique words	# genres
1,070	539,827	595,227	2,570,993	21,364,357	590,959	22

Table 3.1: **Statistics of the IMSDb dataset of movie screenplays.** This dataset is used to learn the correlation between speech and verbs. We use 850 screenplays for training and 220 for validation. Statistics for sentences and words are from the entire text of the screenplays. **scene desc.:** scene descriptions, **speech seg.:** speech segments

3.3.1 The IMSDb Dataset

Movie screenplays are a rich source of data that contain both stage directions (*‘Andrew walked over to open the door’*) and the dialogues spoken by the characters (*‘Please come in’*). Since stage directions often contain described actions, we use the co-occurrence of dialogue and stage directions in screenplays to learn the relationship between ‘actions’ and dialogue (see Fig. 3.1). In this work, we use a corpus of screenplays extracted from IMSDb (www.imsdb.com). In order to get a wide variety of different actions (*‘push’* and *‘kick’* as well as *‘kiss’* and *‘hug’*) we use screenplays covering a range of different genres¹. In total our dataset consists of 1,070 movie screenplays (statistics of the dataset can be seen in Table 3.1). We henceforth refer to this dataset as the IMSDb dataset.

¹Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, Film-Noir, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Short, Sport, Thriller, War, Western

PHONE	KISS	DRINK
Hello, it's me. May I have the number for Dr George Honey I asked you not to call unless hey, it's me Hello, it's me. Hello?	One more kiss Give me a kiss Good night my darling I love you my darling Noone had ever kissed me there before Goodnight angel my sweet boy	To us Raise your glasses to Charlie Heres a toast You want some water Drink deep and live Drink up its party time
DANCE	DRIVE	POINT
Shes a beautiful dancer Waddaya say you wanna dance Come on Ill take a break and well all dance Ladies and Gentlemen the first dance Excuse me would you care for this dance Hattie do you still dance	So well drop Rudy off at the bus Ill drive her just parking it out of the way all you have to do is drop me off at the bank Wait down the road He drove around for a long long time driving	Officer Van Dorn is right down that hall OK Print that one the Met Museum of Art is right there Over there And her The one with the black spot

Table 3.2: **Examples of the top ranked speech samples for six verb categories.** Each block shows the action verb on the left, and the speech samples on the right. All speech segments are from the validation set of the IMSDb dataset of movie screenplays. Best viewed zoomed in.

Screenplay Parsing: While screenplays (generally) follow a standardized format for their parts (e.g., stage direction, dialogue, location, timing information etc.), they can be challenging to parse due to discrepancies in layout and format. We follow the grammar created by Winer et al. [Winer & Young, 2017] which is based on ‘The Hollywood Standard’ [Riley, 2009], to parse the scripts and separate out various screenplay elements. The grammar provided by [Winer & Young, 2017] parses scripts into the following four different elements, (1) Shot Headings, (2) Stage Directions (which contain mention of actions), (3) Dialogue and (4) Transitions. More details are provided in the longer version of this paper ².

In this work we extract only (2) Stage Directions and (3) Dialogue. We extract over 500K stage directions and over 500K dialogue utterances (see Table 3.1). It is important to note that since screenplay parsing is done using an automatic method, and sometimes hand-typed screenplays follow completely non-standard formats, this extraction is not perfect. A quick manual inspection of 100 randomly extracted dialogues shows that around 85% of these are actually dialogue, with the rest being stage directions that have been wrongly labelled as

²A longer ArXiv paper for this chapter can be found at <http://www.robots.ox.ac.uk/~vgg/publications/2020/Nagrani20/nagrani20.pdf>

dialogue.

Verb Mining the Stage Directions: Not all actions will be correlated with speech – e.g. actions like ‘sitting’ and ‘standing’ are difficult to distinguish based on speech alone, since they occur commonly with all types of speech. Hence our first endeavour is to automatically determine verbs rendered ‘discriminative’ by speech alone. For this we use the IMSDb dataset described above. We first take all the stage directions in the dataset, and break up each sentence into clean word tokens (devoid of punctuation). We then determine the part of speech (PoS) tag for each word using the NLTK toolkit [Loper & Bird, 2002] and obtain a list of all the verbs present. Verbs occurring fewer than 50 times (includes many spelling mistakes) or those occurring too frequently, i.e. the top 100 most frequent verbs (these are stop words like ‘be’ etc.) are removed. For each verb, we then group together all the conjugations and word forms for a particular word stem (e.g. the stem *run* can appear in many different forms – running, ran, runs etc.), using the manually created verb conjugations list from the UPenn XTag project³. All such verb classes are then used in training a BERT-based speech to action classifier, described next.

3.3.2 BERT-based Speech Classifier

Each stage direction is then parsed for verbs belonging to the verb classes identified above. We obtain *paired* speech-action data using proximity in the movie screenplays as a clue. Hence, the nearest speech segment to the stage direction (as illustrated in Fig. 3.1) is assigned a label for every verb in the stage direction. This gives us a dataset of speech sentences matched to verb labels. As expected, this is a very noisy dataset. Often, the speech has no correlation with the verb class it is assigned to, and the same speech segment can be assigned to many different verb classes. To learn the correlation between speech and action, we train a classifier with 850 movies and use the remaining ones for validation. The classifier used is a pretrained BERT [Devlin et al., 2018] model with an additional classification layer, finetuned on the

³<http://www.cis.upenn.edu/~xtag/>

dataset of speech paired with weak ‘action’ labels. Exact model details are described below.

Implementation Details: The model used is BERT-Large Cased with Whole-Word Masking (L=24, H=1024, A=16, Total Parameters=340M) [Devlin et al., 2018] pretrained only on English data (BooksCorpus (800M words, [Zhu et al., 2015]) and the Wikipedia corpus (2,500M words)), since the IMsDb dataset consists only of movie screenplays in English⁴. We use WordPiece embeddings [Y. Wu et al., 2016] with a 30,000 token vocabulary. The first token of every sequence is always a special classification token ([CLS]). We use the final hidden vector $C \in \mathbb{R}^H$ corresponding to the first input token ([CLS]) as the aggregate representation. The only new parameters introduced during fine-tuning are classification layer weights $W \in \mathbb{R}^{K \times H}$ where K is the number of classes. We use the standard cross-entropy loss with C and W , i.e., $\log(\text{softmax}(W^T C))$. We use a batch size of 32 and finetune the model end-to-end on the IMsDb dataset for 100,000 iterations using the Adam solver with a learning rate of 5×10^{-5} .

Results: We evaluate the performance of our model on the 220 movie screenplays in the val set. We plot the precision-recall curves using the softmax scores obtained from the `Speech2Action` model. Only those verbs that achieve an average precision (AP) higher than 0.01 are inferred to be correlated with speech. The highest performing verb classes are ‘phone’, ‘open’ and ‘run’, whereas verb classes like ‘fishing’ and ‘dig’ achieve a very low average precision. We finally conclude that there is a strong correlation for 18 verb classes.⁵ Qualitative examples of the most confident predictions (using softmax score as a measure of confidence) for 6 verb classes can be seen in Fig. 3.2. We note here that we have learnt the correlation between action verb and speech from the movie screenplays using a purely data-driven method. The key assumption is that if there is a *consistent* trend of a verb appearing in the screenplays before or after a speech segment, and our model is able to exploit this trend to minimise a classification objective, we infer that the speech is correlated with the action verb.

⁴The model can be found here: <https://github.com/google-research/bert>

⁵The verb classes are: ‘open’, ‘phone’, ‘kiss’, ‘hug’, ‘push’, ‘point’, ‘dance’, ‘drink’, ‘run’, ‘count’, ‘cook’, ‘shoot’, ‘drive’, ‘enter’, ‘fall’, ‘follow’, ‘hit’, ‘eat’.

Because the evaluation is performed purely on the basis of the proximity of speech to verb class in the stage direction of the movie screenplay, it is *not* a perfect ground truth indication of whether an action will actually be performed in a *video* (which is impossible to say only from the movie scripts). We use the stage directions in this case as *pseudo* ground truth, i.e. if the stage direction contains an action and the actor then says a particular sentence, we infer that these two must be related. As a sanity check, we also manually annotate some videos in order to better assess the performance of the `Speech2Action` model. This is described in Sec. 3.4.2.3.

3.4 Mining Videos for Action Recognition

Now that we have learned the `Speech2Action` model to map from transcribed speech to actions (from *text* alone), in this section we demonstrate how this can be applied to video. We use the model to automatically mine video examples from large, unlabelled corpora (the corpus is described in Sec. 3.4.1), and assign them with weak labels from the `Speech2Action` model prediction. Armed with this weakly labelled data, we then train models directly for the downstream task of visual action recognition. Detailed training and evaluation protocols for the mining are described in the following sections.

3.4.1 Unlabelled Data

In this work, we apply the `Speech2Action` model to a large internal corpus of movies and TV shows. The corpus consists of 222,855 movies and TV show episodes. For these videos, we use the closed captions (note that this can be obtained from the audio track directly using automatic speech recognition). The total number of closed captions for this corpus is 188,210,008, which after dividing into sentences gives us a total of 390,791,653 (almost 400M) sentences. While we use this corpus in our work, we would like to stress here that there is no correlation between the text data used to train the `Speech2Action` model and

this unlabelled corpus (other than both belonging to the movie domain), and such a model can be applied to any other corpus of unlabelled, edited film material.

3.4.2 Obtaining Weak Labels

In this section, we describe how we obtain weak action labels for short clips from the speech alone. We do this in two ways, (i) using the `Speech2Action` model, and (ii) using a simple keyword spotting baseline described below.

3.4.2.1 Using `Speech2Action`

The `Speech2Action` model is applied to a single sentence of speech, and the prediction is used as a weak label if the confidence (softmax score) is above a certain threshold. The threshold is obtained by taking the confidence value at a precision of 0.3 on the IMSDb validation set, with some manual adjustments for the classes of ‘phone’, ‘run’ and ‘open’ (since these classes have a much higher recall, we increase the threshold in order to prevent a huge imbalance of retrieved samples). More details are provided in the Appendix, Sec. A. We then extract the visual frames for a 10 second clip centered around the midpoint of the timeframe spanned by the caption, and assign the `Speech2Action` label as the weak label for the clip. Ultimately, we successfully end up mining 837,334 video clips for 18 action classes. While this is a low yield, we still end up with a large number of mined clips, greater than the manually labelled Kinetics dataset [Kay et al., 2017] (600K).

We also discover that the verb classes that have high correlation with speech in the IMSDb dataset tend to be *infrequent* or *rare* actions in other datasets [Gu et al., 2018] – as shown in Fig. 3.2, we obtain two orders of magnitude more data for certain classes in the AVA training set [Gu et al., 2018]. Qualitative examples of mined video clips with action labels can be seen in Fig. 3.3. Note how we are able to retrieve clips with a wide variety in background and actor, simply from the speech alone. Refer to Fig. 10 in the online version of this chapter for more

examples showing diversity in objects and viewpoint ⁶.



Figure 3.2: **Distribution of training clips mined using Speech2Action.** We compare the distribution of mined clips to the number of samples in the AVA training set. Although the mined clips are noisy, we are able to obtain far more, in some cases up to *two* orders of magnitude more training data (note the **log scale** in the x-axis).

3.4.2.2 Using a Keyword Spotting Baseline

In order to validate the efficacy of our `Speech2Action` model trained on movie screenplays, we also compare to a simple keyword spotting baseline. This involves searching for the action verb in the speech directly – a speech segment like ‘*Will you eat now?*’ is directly assigned the label ‘eat’. This itself is a very powerful baseline, e.g. speech segments such as ‘*Will you dance with me*’, are strongly indicative of the action ‘dance’. To implement this baseline, we search for the presence of the action verb (or its conjugations) in the speech segment directly, and if the verb is present in the speech, we assign the action label to the video clip directly. The fallacy of this method is that there is no distinction between the different semantic meanings of a verb, e.g. the speech segment ‘*You’ve missed the point entirely*’ will be weakly labelled with the verb ‘point’ using this baseline, which is indicative of a different semantic meaning to the physical action of ‘pointing’. Hence as we show in the results,

⁶A longer ArXiv paper for this chapter can be found at <http://www.robots.ox.ac.uk/~vgg/publications/2020/Nagrani20/nagrani20.pdf>



Figure 3.3: **Examples of clips mined automatically using the `Speech2Action` model applied to *speech alone* for 8 AVA classes.** We show only a single frame from each video. Note the diversity in background, actor and view point. We show false positives for eat, phone and dance (last in each row, enclosed in a red box). Expletives are censored. More examples are provided in the Appendix.

this baseline performs poorly compared to our `Speech2Action` mining method (Tables 3.5 and 3.4).

3.4.2.3 Manual Evaluation of `Speech2Action`

We now assess the performance of `Speech2Action` applied to videos. Given a speech segment, we check whether a prediction made by the model on the speech translates to the action being performed visually in the frames aligned to the speech. To assess this, we do a manual inspection of a random set of 100 retrieved video clips for 10 of the verb classes, and report the true positive rate (number of clips for which the action is visible) in Table 3.3. We find that a

dance	phone	kiss	drive	eat	drink	run	point	hit	shoot
42	68	18	41	27	51	83	52	18	27

Table 3.3: **Number of true positives for 100 randomly retrieved samples for 10 classes.** These estimates are obtained through manual inspection of video clips that are labelled with `Speech2Action`. While the true positive rate for some classes is low, the other samples still contain valuable information for the classifier. For example, although there are only 18 true samples of ‘kiss’, many of the other videos have two people with their lips very close together, or even if they are not ‘eating’ strictly, many times they are holding food in their hands.

surprising number of samples actually contain the action during the time frame of 10 seconds, with some classes noisier than others. The high purity of the classes ‘run’ and ‘phone’ can be explained by the higher thresholds used for mining, as explained in Sec. 3.4.2.1. Common sources of false positives are actions performed off screen, or actions performed at a temporal offset (either much before or much after) the speech segment. We note that at no point do we ever actually use any of the manual labels for training, these are purely for evaluation and as a sanity check.

3.5 Action Classification

Now that we have described our method to obtain weakly labelled training data, we train a video classifier with the S3D-G [S. Xie et al., 2018] backbone on these noisy samples for the task of action recognition. We first detail the training and testing protocols, and then describe the datasets used in this work.

3.5.1 Evaluation Protocol

We evaluate our video classifier for the task of action classification in the following two ways: First, we follow the typical procedure adopted in the video understanding literature [Carreira & Zisserman, 2017]: pre-training on a large corpus of videos weakly labelled using our `Speech2Action` model, followed by fine-tuning on the training split of a labeled target

dataset (‘test bed’). After training, we evaluate the performance on the test set of the target dataset. In this work we use HMDB-51 [Kuehne et al., 2011], and compare to other state of the art methods on this dataset.

Second, and perhaps more interestingly, we apply our method by training a video classifier on the mined video clips for some action classes, and evaluating it *directly* on the test samples of *rare* action classes in the target dataset (in this case we use the AVA dataset [Gu et al., 2018]).

Note: At this point we also manually verified that there is no overlap between the movies in the IMSDb dataset and the AVA dataset (not surprising since AVA movies are older and more obscure – these are movies that are freely available on YouTube). Here not a single manually labelled training example is used, since there is no finetuning (we henceforth refer to this as zero-shot⁷). We also report performance for the few-shot learning scenario, where we fine-tune our model on a *small* number of labelled examples. We note that in this case, we can only evaluate on the classes that directly overlap with the verb classes in the IMSDb dataset.

3.5.2 Datasets and Experimental Details

HMDB51: HMDB51 [Kuehne et al., 2011] contains 6766 realistic and varied video clips from 51 action classes. Evaluation is performed using average classification accuracy over three train/test splits from [Idrees et al., 2017], each with 3570 train and 1530 test videos.

AVA: The AVA dataset [Gu et al., 2018] is collected by exhaustively manually annotating videos and exhibits a strong imbalance in the number of examples between the common and rare classes. Eg. a common action, like ‘stand’, has 160k training and 43k test examples, compared to ‘drive’ (1.18K train and 561 test) and ‘point’ (only 96 train and 32 test). As a result, methods relying on full supervision struggle on the categories in the middle and the end of the tail. We evaluate on the 14 AVA classes that overlap with the classes present in the IMDSDB dataset (all from the middle and tail). While the dataset is originally a detection

⁷In order to avoid confusion with the strict meaning of this term, we clarify that in this work we use it to refer to the case where not a single *manually labelled* example is available for a particular class. We do however train on multiple weakly labelled examples.

dataset, we repurpose it simply for the task of action classification, by assigning each frame the union of labels from all bounding box annotations. We then train and test on samples from these 14 action classes, reporting per-class average precision (AP).

Implementation Details: We train the S3D with gating (S3D-G) [S. Xie et al., 2018] model as our visual classifier. Following [S. Xie et al., 2018], we densely sample 64 frames from a video, resize input frames to 256×256 and then take random crops of size 224×224 during training. During evaluation, we use all frames and take 224×224 center crops from the resized frames. Our models are implemented with TensorFlow and optimized with a vanilla synchronous SGD algorithm with momentum of 0.9. For models trained from scratch, we train for 150K iterations with a learning rate schedule of 10^2 , 10^3 and 10^4 dropping after 80K and 100K iterations, and for finetuning we train for 60K iterations using a learning rate of 10^2 .

Loss functions for training: We try both the softmax cross-entropy and per-class sigmoid loss, and find that the performance was relatively stable with both choices.

3.5.3 Results

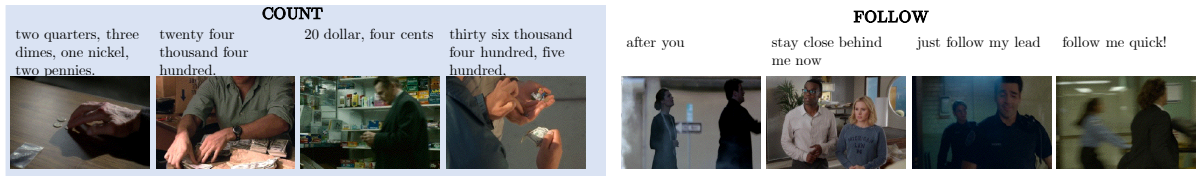


Figure 3.4: **Examples of clips mined for more abstract actions.** These are actions that are not present in standard datasets like HMDB51 or AVA, but are quite well correlated with speech. Our method is able to automatically mine clips weakly labelled with these actions from unlabelled data.

HMDB51: The results on HMDB51 can be seen in Table 3.4. Training on videos labelled with Speech2Actions leads to a significant 17% improvement over from-scratch training. For reference, we also compare to other self-supervised and weakly supervised works (note that these methods differ both in architecture and training objective). We show a 14% improvement over previous self-supervised works that use *only* video frames (no other modalities).

Method	Architecture	Pre-training	Acc.
Shuffle&Learn [Misra et al., 2016]★	S3D-G (RGB)	UCF101† [Soomro et al., 2012]	35.8
OPN [H.-Y. Lee et al., 2017]	VGG-M-2048	UCF101† [Soomro et al., 2012]	23.8
ClipOrder [D. Xu et al., 2019]	R(2+1)D	UCF101† [Soomro et al., 2012]	30.9
Wang et al. [J. Wang et al., 2019]	C3D	Kinetics† [Soomro et al., 2012]	33.4
3DRotNet [Jing & Tian, 2018]★	S3D-G (RGB)	Kinetics†	40.0
DPC [T. Han et al., 2019]	3DResNet18	Kinetics†	35.7
CBT [C. Sun, Baradel, et al., 2019]	S3D-G (RGB)	Kinetics†	44.6
DisInit (RGB) [Girdhar, Tran, et al., 2019]	R(2+1)D18 [Tran et al., 2018]	Kinetics**	54.8
Korbar et al [Korbar et al., 2018]	I3D (RGB)	Kinetics†	53.0
-	S3D-G (RGB)	Scratch	41.2
Ours	S3D-G (RGB)	KSB-mined	46.0
Ours	S3D-G (RGB)	S2A-mined	58.1
Supervised pretraining	S3D-G (RGB)	ImageNet	54.7
Supervised pretraining	S3D-G (RGB)	Kinetics	72.3

Table 3.4: **Action classification results on HMDB51.** Pre-training on videos labelled with Speech2Action leads to a 17% improvement over training from scratch and also outperforms previous self-supervised and weakly supervised works. **KSB-mined:** video clips mined using the keyword spotting baseline. **S2A-mined:** video clips mined using the Speech2Action model. †videos without labels. **videos with labels distilled from ImageNet. When comparing to [Korbar et al., 2018], we report the number achieved by their I3D (RGB only) model which is the closest to our architecture. For ★, we report the reimplementations by [C. Sun, Baradel, et al., 2019] using the S3D-G model (same as ours). For the rest, we report performance directly from the original papers.

We also compare to Korbar *et al.* [Korbar et al., 2018] who pretrain using audio and video synchronisation on AudioSet, DisInit [Girdhar, Tran, et al., 2019], which distills knowledge from ImageNet into Kinetics videos, and simply pretraining on ImageNet and then inflating 2D convolutions to our S3D-G model [Kay et al., 2017]. We improve over these works by 3-4% – which is impressive given that the latter two methods rely on access to a large-scale manually labelled image dataset [J. Deng et al., 2009], whereas ours relies only on 1000 unlabelled movie scripts. Another point of interest (and perhaps an unavoidable side-effect of this stream of self- and weak-supervision) is that while all these previous methods do not use labels, they still pretrain on the Kinetics data, which has been carefully curated to cover a wide diversity of over 600 different actions. In contrast, we mine our training data directly from movies, without the need for any manual labelling or careful curation, and our pretraining data was mined for only 18 classes.

AVA-scratch: The results on AVA for models trained from scratch with *no* pretraining, can

Data	Per-Class AP													
	drive	phone	kiss	dance	eat	drink	run	point	open	hit	shoot	push	hug	enter
AVA (fully supervised)	0.63	0.54	0.22	0.46	0.67	0.27	0.66	0.02	0.49	0.62	0.08	0.09	0.29	0.14
KS-baseline †	0.67	0.20	0.12	0.53	0.67	0.18	0.37	0.00	0.33	0.47	0.05	0.03	0.10	0.02
S2A-mined (zero-shot)	0.83	0.79	0.13	0.55	0.68	0.30	0.63	0.04	0.52	0.54	0.18	0.04	0.07	0.04
S2A-mined + AVA	0.84	0.83	0.18	0.56	0.75	0.40	0.74	0.05	0.56	0.64	0.23	0.07	0.17	0.04
AVA (few-shot)-20	0.82	0.83	0.22	0.55	0.69	0.33	0.64	0.04	0.51	0.59	0.20	0.06	0.19	0.13
AVA (few-shot)-50	0.82	0.85	0.26	0.56	0.70	0.37	0.69	0.04	0.52	0.65	0.21	0.06	0.19	0.15
AVA (few-shot)-100	0.84	0.86	0.30	0.58	0.71	0.39	0.75	0.05	0.58	0.73	0.25	0.13	0.27	0.15
AVA (all)	0.86	0.89	0.34	0.58	0.78	0.42	0.75	0.03	0.65	0.72	0.26	0.13	0.36	0.16

Table 3.5: **Per-class average precision for 14 AVA mid and tail classes.** These actions occur *rarely*, and hence are harder to get manual supervision for. For 8 of the 14 classes, we exceed fully supervised performance without a single manually labelled training example (highlighted in pink, best viewed in colour). S2A-mined: Video clips mined using Speech2Action. † Keyword spotting baseline. First 4 rows: models are trained from scratch. Last 4 rows: we pre-train on video clips mined using Speech2Action.

be seen in Table 3.5 (top 4 rows). We compare the following: training with the AVA training examples (Table 3.5, top row), training only with our mined examples, and training jointly with both. For 8 out of 14 classes, we exceed fully supervised performance without a single AVA training example, in some cases (‘drive’ and ‘phone’) almost by 20%.

AVA-finetuned: We also show results for pre-training on Speech2Action mined clips first, and then fine-tuning on a gradually increasing number of AVA labelled training samples per class (Table 3.5, bottom 4 rows). Here we keep all the weights from the fine-tuning, including the classification layer weights, for initialisation, and fine-tune only for a single epoch. With 50 training samples per class, we exceed fully supervised performance for all classes (except for ‘hug’ and ‘push’) compared to training from scratch. The worst performance is for the class ‘hug’ – ‘hug’ and ‘kiss’ are often confused, as the speech in both cases tends to be similar – ‘I love you’. A quick manual inspection shows that most of the clips are wrongly labelled as ‘kiss’, which is why we are only able to mine very few video clips for this class. For completeness, we also pretrain a model with the S2A mined clips (only 14 classes) and then finetune on AVA for *all* 60 classes used for evaluation, and get a 40% overall classification acc. vs 38% with training on AVA alone.

Mining Technique: We also train on clips mined using the keyword spotting baseline (Table 3.5). For some classes, this baseline itself exceeds fully supervised performance. Our `Speech2Action` labelling beats this baseline for all classes, indeed the baseline does poorly for classes like ‘point’ and ‘open’ – verbs which have many semantic meanings, demonstrating that the semantic information learnt from the IMSTDb dataset is valuable. However we note here that it is difficult to measure performance quantitatively for the class ‘point’ due to idiosyncrasies in the AVA test set (wrong ground truth labels for very few test samples) and hence we show qualitative examples of mined clips in Fig. 3.3. We note that the baseline comes very close for ‘dance’ and ‘eat’, demonstrating that simple keyword matching on speech can retrieve good training data for these actions.

Abstract Actions: By gathering data directly from the stage directions in movie screenplays, our action labels are post-defined (as in [Fouhey et al., 2018]). This is unlike the majority of the existing human action datasets that use pre-defined labels [Caba Heilbron et al., 2015; Sigurdsson et al., 2016; Gu et al., 2018; Monfort et al., 2019]. Hence we also manage to mine examples for some unusual or *abstract* actions which are quite well correlated with speech, such as ‘count’ and ‘follow’. While these are not present in standard action recognition datasets such as HMDB51 or AVA, and hence cannot be evaluated numerically, we show some qualitative examples of these mined videos in Fig. 3.4.

3.6 Conclusion

We provide a new data-driven approach to obtain weak labels for action recognition, using speech alone. With only a thousand unaligned screenplays as a starting point, we obtain weak labels automatically for a number of rare action classes. However, there is a plethora of literary material available online, including plays and books, and exploiting these sources of text may allow us to extend our method to predict other action classes, including composite actions of ‘verb’ and ‘object’. We also note that *besides* actions, people talk about physical objects,

events and scenes – descriptions of which are also present in screenplays and books. Hence the same principle used here could be applied to mine videos for more general visual content.

Acknowledgments: Arsha is supported by a Google PhD Fellowship. We are grateful to Carl Vondrick for early discussions.

Appendices

Appendices and further visualisations for this chapter can be accessed online.⁸

Statement of Authorship

A statement of authorship for this work can be found in Appendix C.

⁸<https://www.robots.ox.ac.uk/~vgg/research/speech2action/>

4 | Emotion Recognition in Speech using Cross-Modal Transfer in the Wild

Samuel Albanie* Arsha Nagrani* Andrea Vedaldi Andrew Zisserman

VGG, Oxford

*Equal Contribution

Abstract

Obtaining large, human labelled speech datasets to train models for emotion recognition is a notoriously challenging task, hindered by annotation cost and label ambiguity. In this work, we consider the task of learning embeddings for speech classification without access to any form of labelled audio. We base our approach on a simple hypothesis: that the emotional content of speech correlates with the facial expression of the speaker. By exploiting this relationship, we show that annotations of expression can be transferred from the visual domain (faces) to the speech domain (voices) through *cross-modal distillation*. We make the following contributions: (i) we develop a strong teacher network for facial emotion recognition that achieves the state of the art on a standard benchmark; (ii) we use the teacher to train a student, *tabula rasa*, to learn representations (embeddings) for speech emotion recognition *without access to labelled audio data*; and (iii) we show that the speech emotion embedding can be used for speech emotion recognition on external benchmark datasets. Code, models and data are available¹.

Published in the Proceedings of the [ACM Multimedia Conference, 2018](#).

¹<http://www.robots.ox.ac.uk/~vgg/research/cross-modal-emotions>

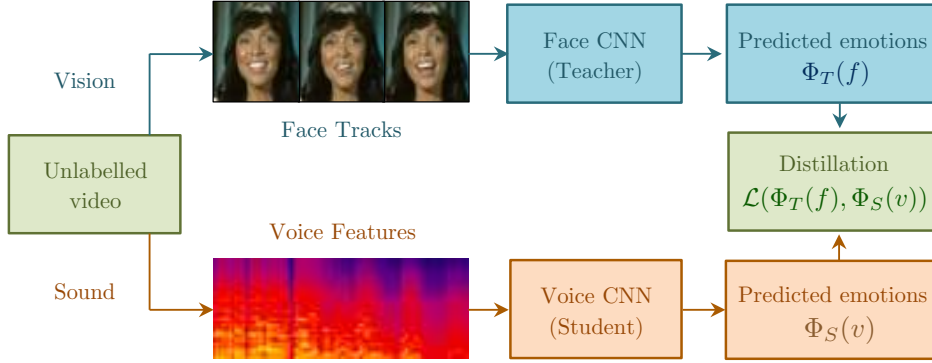


Figure 4.1: Cross-modal transfer: A CNN for speech emotion recognition (the student, Φ_S) is trained by distilling the knowledge of a pretrained facial emotion recognition network (the teacher, Φ_T) across unlabelled video. The student aims to exploit redundancy between the audio and visual signals v and f to learn embeddings, reducing dependence on labelled speech.

4.1 Introduction

Despite recent advances in the field of speech emotion recognition, learning representations for *natural* speech segments that can be used efficiently under noisy and unconstrained conditions still represents a significant challenge. Obtaining large, labelled human emotion datasets ‘in the wild’ is hindered by a number of difficulties. First, since labelling naturalistic speech segments is extremely expensive, most datasets consist of elicited or acted speech. Second, as a consequence of the subjective nature of emotions, labelled datasets often suffer from low human annotator agreement, as well as the use of varied labelling schemes (i.e., dimensional or categorical) which can require careful alignment [Mariooryad & Busso, 2013]. Finally, cost and time prohibitions often result in datasets with low speaker diversity, making it difficult to avoid speaker adaptation. Fully supervised techniques trained on such datasets hence often demonstrate high accuracy for only intra-corpus data, with a natural propensity to overfit [Latif et al., 2018].

In light of these challenges, we pose the following question: is it possible to learn a represen-

tation for emotional speech content for natural speech, from *unlabelled* audio-visual speech data, simply by transferring knowledge from the facial expression of the speaker?

Given the recent emergence of large-scale video datasets of human speech, it is possible to obtain examples of unlabelled human emotional speech at massive scales. Moreover, although it is challenging to assess the accuracy of emotion recognition models precisely, recent progress in computer vision has nevertheless enabled deep networks to learn to map faces to emotional labels in a manner that consistently matches a pool of human annotators [Albanie & Vedaldi, 2016]. We show how to transfer this discriminative visual knowledge into an audio network using unlabelled video data as a bridge. Our method is based on a simple hypothesis: that the emotional content of speech correlates with the facial expression of the speaker.

Our work is motivated by the following four factors. First, we would like to learn from a large, *unlabelled* collection of ‘talking faces’ in videos as a source of free supervision, without the need for any manual annotation. Second, evidence suggests that this is a possible source of supervision that infants use as their visual and audio capabilities develop [Grossmann, 2010]. Newborns look longer at face-like stimuli and track them farther than non-face-like stimuli [Goren et al., 1975; Johnson et al., 1991], and combining these facial stimuli together with voices, detect information that later may allow for the discrimination and recognition of emotional expressions. Our third motivation is that we would like to be able to handle ambiguous emotions gracefully. To this end, we seek to depart from annotation that relies on a single categorical label per segment, but instead incorporate a measure of uncertainty into the labelling scheme, building on prior work by [S. Zhao et al., 2017] and [J. Han et al., 2017]. Finally, accepting that the relationship between facial and vocal emotion will be a noisy one, we would like to make use of the remarkable ability of CNNs to learn effectively in the presence of label noise when provided with large volumes of training data [Rolnick et al., 2017; Mahajan et al., 2018].

We make the following contributions: (i) we develop a strong model for facial expression

emotion recognition, achieving state of the art performance on the FERPlus benchmark (section 4.3.1), (ii) we use this computer vision model to label face emotions in the VoxCeleb [Nagrani et al., 2017] video dataset – this is a large-scale dataset of emotion-unlabelled speaking face-tracks obtained in the wild (section 4.4); (iii) we transfer supervision *across modalities* from faces to a speech, and then train a speech emotion recognition model using speaking face-tracks (section 4.5); and, (iv) we demonstrate that the resulting speech model is capable of classifying emotion on two external datasets (section 4.5.2). A by-product of our method is that we obtain emotion annotation for videos in the VoxCeleb dataset automatically using the facial expression model, which we release as the EMOVOXCELEB dataset.

4.2 Related Work

Teacher-student methods. Teaching one model with another was popularised by [Bucilua et al., 2006] who trained a single model to match the performance of an ensemble, in the context of model compression. Effective supervision can be provided by the “teacher” in multiple ways: by training the “student” model to regress the pre-softmax logits [Ba & Caruana, 2014], or by minimising cross entropy between both models’ probabilistic outputs [J. Li et al., 2014], often through a high-temperature softmax that softens the predictions of each model [Hinton et al., 2015; Crowley et al., 2017]. In contrast to these methods which transfer supervision within the same modality, *cross-modal* distillation obtains supervision in one modality and transfers it to another. This approach was proposed for RGB and depth paired data, and for RGB and flow paired data by [Gupta et al., 2016]. More recent work [Aytar et al., 2016; Arandjelović & Zisserman, 2017; Aytar, Vondrick, & Torralba, 2017; Owens et al., 2016] has explored this concept by exploiting the correspondence between synchronous audio and visual data in teacher-student style architectures [Aytar et al., 2016; Aytar, Vondrick, & Torralba, 2017], or as a form of self-supervision [Arandjelović & Zisserman, 2017] where networks for both modalities are learnt from scratch. Some works have also examined cross-modal relationships

between faces and voices in order to learn identity representations [Nagrani et al., 2018b,a; C. Kim et al., 2018]. Differently from these works, our approach places an explicit reliance on the correspondence between the facial and vocal *emotions* emitted by a speaker during speech, discussed next.

Links between facial and vocal emotion. Our goal is to learn a representation that is aware of the emotional content in speech *prosody*, where prosody refers to the extra-linguistic variations in speech (e.g. changes in pitch, tempo, loudness, or intonation), by transferring such emotional knowledge from face images extracted synchronously. For this to be possible, the emotional content of speech must correlate with the facial expression of the speaker. Thus in contrast to multimodal emotion recognition systems which seek to make use of the complementary components of the signal between facial expression and speech [Busso et al., 2004], our goal is to perform cross-modal learning by exploiting the redundancy of the signal that is common to both modalities. Fortunately, given their joint relevance to communication, person perception, and behaviour more generally, interactions between speech prosody and facial cues have been intensively studied [Cvejic et al., 2010; Pell, 2005; Swerts & Krahmer, 2008]. The broad consensus of these works is that during conversations, speech prosody is typically associated with other social cues like facial expressions or body movements, with facial expression being the most ‘privileged’ or informative stimulus [Rigoulot & Pell, 2014].

Deep learning for speech emotion recognition. Deep networks for emotional speech recognition either operate on hand-crafted acoustic features known to have a significant effect on speech prosody, (e.g. MFCCs, pitch, energy, ZCR, ...), or operate on raw audio with little processing, e.g. only the application of Fourier transforms [Cummins et al., 2017]. Those that use handcrafted features focus on global suprasegmental/prosodic features for emotion recognition, in which utterance level statistics are calculated. The main limitation of such global-level acoustic features is that they cannot describe the dynamic variation along an utterance [Aldeneh & Provost, 2017]. Vocal emotional expression is shaped to some extent by differences in the temporal structure of language and emotional cues are not equally salient

throughout the speech signal [Rigoulot & Pell, 2014; Y. Kim & Provost, 2016]. In particular, there is a well-documented propensity for speakers to elongate syllables located in word- or phrase-final positions [Oller, 1973; Pell, 2001], and evidence that speakers vary their pitch in final positions to encode gradient acoustic cues that refer directly to their emotional state [Pell, 2001]. We therefore opt for the second strategy, using minimally processed audio represented by short term magnitude spectrograms, directly as inputs to the network. Operating on these features can potentially improve performance “in the wild” where the encountered input can be unpredictable and diverse [J. Kim et al., 2017]. By using CNNs with max pooling on spectrograms, we encourage the network to determine the emotionally salient regions of an utterance.

Corpus	Speakers	Naturalness	Labelling method	Audio-visual
AIBO★ [Batliner et al., 2004]	51	Natural	Manual	Audio only
EMODB [Burkhardt et al., 2005]	10	Acted	Manual	Audio only
ENTERFACE [Martin et al., 2006]	43	Acted	Manual	✓
LDC [Lieberman et al., 2002]	7	Acted	Manual	Audio only
IEMOCAP [Busso et al., 2008]	10	Both†	Manual	✓
AFEW 6.0♠ [Dhall et al., 2012]	unknown ⁺	Acted	Subtitle Analysis	✓
RML	8	Acted	Manual	✓
EMOVoxCELEB	1,251	Natural	Expression Analysis	✓

Table 4.1: Comparison to existing public domain speech emotion datasets. † contains both improvised and scripted speech. ★ contains only emotional speech of children. ♠ has not been commonly used for audio only classification, but is popular for audio-visual fusion methods. ⁺ identity labels are not provided.

Existing speech emotion datasets. Fully supervised deep learning techniques rely heavily on large-scale labelled datasets, which are tricky to obtain for emotional speech. Many methods rely on using actors [Burkhardt et al., 2005; Martin et al., 2006; Liberman et al., 2002; Busso et al., 2008] (described below), and automated methods are few. Some video datasets are created using subtitle analysis [Dhall et al., 2012]. In the facial expression domain, labels can be generated through reference events [Albanie & Vedaldi, 2016], however this is challenging to imitate for speech. A summary of popular existing datasets is given in Table 4.1. We

highlight some common disadvantages of these datasets below, and contrast these with the VoxCeleb dataset that is used in this paper:

(1) Most speech emotion datasets consist of elicited or acted speech, typically created in a recording studio, where actors read from written text. However, as [Douglas-Cowie et al., 2000] points out, full-blown emotions very rarely appear in the real world and models trained on acted speech rarely generalise to natural speech. Furthermore there are physical emotional cues that are difficult to consciously mimic, and only occur in natural speech. In contrast, VoxCeleb consists of interview videos from YouTube, and so is more naturalistic.

(2) Studio recordings are also often extremely clean and do not suffer from ‘real world’ noise artefacts. In contrast, videos in the VoxCeleb dataset are degraded with real world noise, consisting of background chatter, laughter, overlapping speech and room acoustics. The videos also exhibit considerable variance in the quality of recording equipment and channel noise.

(3) For many existing datasets, cost and time prohibitions result in low speaker diversity, making it difficult to avoid speaker adaptation. Since our method does not require any emotion labels, we can train on VoxCeleb which is two orders of magnitude larger than existing public speech emotion datasets in the number of speakers.

Note that for any machine learning system that aims to perform emotion recognition using vision or speech, the ground truth emotional state of the speaker is typically unavailable. To train and assess the performance of models, we must ultimately rely on the judgement of human annotators as a reasonable proxy for the true emotional state of a speaker. Throughout this work we use the term emotion recognition to mean accurate prediction of this proxy.

4.3 Cross Modal Transfer

The objective of this work is to learn useful representations for emotion speech recognition, without access to labelled speech data. Our approach, inspired by the method of cross modal distillation [Gupta et al., 2016], is to tackle this problem by exploiting readily available annotated data in the visual domain.

Under the formulation introduced in [Gupta et al., 2016], a student model operating on one input modality learns to reproduce the features of a teacher model, which has been trained for a given task while operating on a different input modality (for which labels are available). The key idea is that by using a sufficiently large dataset of modality paired inputs, the teacher can transfer task supervision to the student without the need for labelled data in the student’s modality. Importantly, it is assumed that the paired inputs possess the same attributes with respect to the task of interest.

In this work, we propose to use the correspondence between the emotion expressed by the facial expression of a speaker and the emotion of the speech utterance produced synchronously. Our approach relies on the assumption that there is some redundancy in the emotional content of the signal communicated through the concurrent expression and speech of a speaker. To apply our method, we therefore require a large number of *speaking face-tracks*, in which we have a known correspondence between the speech audio and the face depicted. Fortunately, this can be acquired, automatically and at scale using the recently developed SyncNet [J. S. Chung & Zisserman, 2016b]. This method was used to generate the large-scale VoxCeleb dataset [Nagrani et al., 2017] for speaking face-tracks, which forms the basis of our study.

As discussed in Sec. 4.2, there are several ways to distill the knowledge of the teacher to the student. While [Gupta et al., 2016] trained the student by regressing the intermediate representations at multiple layers in the teacher model, we found in practice that the approach introduced in [Hinton et al., 2015] was most effective for our task. Specifically, we used a cross

Method	Accuracy (PrivateTest)
PLD [Barsoum et al., 2016]	85.1 \pm 0.5%
CEL [Barsoum et al., 2016]	84.6 \pm 0.4%
ResNet+VGG [†] [Huang, 2017]	87.4
SENet Teacher (Ours)	88.8 \pm 0.3%

Table 4.2: Comparison on the FERplus facial expression benchmark. [†] denotes performance of model ensemble. Where available, the mean and std. is reported over three repeats. The SENet Teacher model is described in Sec. 4.3.1.

entropy loss between the outputs of the networks after passing both both sets of predictions through a softmax function with temperature T to produce a distribution of predictions:

$$p_i = \frac{\exp(x_i/T)}{\sum_j \exp(x_j/T)}, \quad (4.1)$$

where x_i denotes the logit associated with class i and p_i denotes the corresponding normalised prediction. A higher temperature softmax produces a softer distribution over predictions. We experimented with several values of T to facilitate training and found, similarly to [Hinton et al., 2015], that a temperature of 2 was most effective. We therefore use this temperature value in all reported experiments.

4.3.1 The Teacher

This section describes how we obtain the teacher model which is responsible for classifying facial emotion in videos.

Frame-level Emotion Classifier. To construct a strong teacher network (which is tasked with performing emotion recognition from *face images*), training is performed in multiple stages. We base our teacher model on the recently introduced Squeeze-and-Excitation architecture [J. Hu et al., 2019] (the ResNet-50 variant). The network is first pretrained on the large-scale VGG-Face2 dataset [Cao et al., 2018] (\approx 3.3 million faces) for the task of identity verifica-

tion. The resulting model is then finetuned on the *FERplus* dataset [Barsoum et al., 2016] for emotion recognition. This dataset comprises the images from the original FER dataset ($\approx 35k$ images) [Goodfellow et al., 2013] together with a more extensive set of annotations (10 human annotators per image). The emotions labelled in the dataset are: *neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, *fear* and *contempt*. Rather than training the teacher to predict a single correct emotion for each face, we instead require it to match the *distribution* of annotator labels. Specifically, we train the network to match the distribution of annotator responses with a cross entropy loss:

$$\mathcal{L} = - \sum_n \sum_i p_i^{(n)} \log q_i^{(n)}, \quad (4.2)$$

where $p_i^{(n)}$ represents the probability of annotation n taking emotion label i , averaged over annotators, and $q_i^{(n)}$ denotes the corresponding network prediction.

During training, we follow the data augmentation scheme comprising affine distortions of the input images introduced in [Z. Yu & Zhang, 2015] to encourage robustness to variations in pose. To verify the utility of the resulting model, we evaluate on the FERPlus benchmark, following the test protocol defined in [Barsoum et al., 2016], and report the results in Table 4.2. To the best of our knowledge, our model represents the current state of the art on this benchmark.

From Frames to Face-tracks. Since a single speech segment typically spans many frames, we require labels at a face-track level in order to transfer knowledge from the face domain to the speech domain. To address the fact that our classifier has been trained on individual images, not with *face-tracks*, we take the simplest approach of considering a single face-track as a set of individual frames. A natural consequence of using still frames extracted from video, however, is that the emotion of the speaker is not captured with equal intensity in every frame. Even in the context of a highly emotional speech segment, many of the frames that correspond to transitions between utterances exhibit a less pronounced facial expression, and are therefore

often labelled as ‘neutral’ (see Figure 4.2 for an example track). One approach that has been proposed to address this issue is to utilise a single frame or a subset of frames known as *peak frames*, which best represent the emotional content of the face-track [Zhalehpour et al., 2016; Poria et al., 2015]. The goal of this approach is to select the frames for which the dominant emotional expression is at its apex. It is difficult to determine which frames are the key frames, however, while [Poria et al., 2015] select these frames manually, [Zhalehpour et al., 2016] add an extra training step which measures the ‘distance’ of the expressive face from the subspace of neutral facial expressions. This method also relies on the implicit assumption that all facial parts reach the peak point at the same time.

We adopt a simple approximation to peak frame selection by representing each track by the maximum response of each emotion across the frames in the track, an approach that we found to work well in practice. We note that prior work has also found simple average pooling strategies over frame-level predictions [Bargal et al., 2016; P. Hu et al., 2017] to be effective (we found average pooling to be slightly inferior, though not dramatically different in performance). To verify that max-pooling represents a reasonable temporal aggregation strategy, we applied the trained SENet Teacher network to the individual frames of the AFEW 6.0 dataset, which formed the basis of the 2016 Emotion Recognition in the Wild (EmotiW) competition [Dhall et al., 2016]. Since our objective here is not to achieve the best performance by specialising for this particular dataset (but rather to validate the aggregation strategy for predicting tracks), we did not fine-tune the parameters of the teacher network for this task. Instead, we applied our network directly to the default face crops provided by the challenge organisers and aggregated the emotional responses over each video clip using max pooling. We then treat the predictions as 8-dimensional embeddings and use the AFEW training set to fit a single affine transformation (linear transformation plus bias), followed by a softmax, allowing us to account for the slightly different emotion categorisation (AFEW does not include a *contempt* label). By evaluating the resulting re-weighted predictions on the validation set we obtained an accuracy of 49.3% for the 7-way classification task, strongly outperforming the



Figure 4.2: An example set of frames accompanying a single speech segment in the VoxCeleb dataset illustrating the *neutral transition-face* phenomenon exhibited by many face tracks: the facial expression of the speaker, as predicted by the static image-based face classifier often takes a ‘neutral’ label while transitioning between certain phonemes.

baseline of 38.81% released by the challenge organisers.

4.3.2 The Student

The student model, which is tasked with performing emotion recognition *from voices*, is based on the VGG-M architecture [Chatfield et al., 2014] (with the addition of batch normalization). This model has proven effective for speech classification tasks in prior work [Nagrani et al., 2017], and provides a good trade-off between computational cost and performance. The architectural details of the model are described in section 4.5.1.

4.3.3 Time-scale of transfer

The time-scale of transfer determines the length of the audio segments that are fed into the student network for transferring the logits from face to voice. Determining the optimal length of audio segment for which emotion is discernable is still an open question. Ideally, we would like to learn only features related to speech *prosody* and not the lexical content of speech, and hence we do not want to feed in audio segments that contain entire sentences to the student network. We also do not want segments that are too short, as this creates the risk of capturing largely neutral audio segments.

Rigoulot and Pell [Rigoulot & Pell, 2014] studied the time course for recognising vocally



Figure 4.3: Examples of emotions in the EMOVOXCELEB dataset. We rely on the facial expression of the speaker to provide clues about the emotional content of their speech.

expressed emotions on human participants, and found that while some emotions were more quickly recognised than others (fear as opposed to happiness or disgust), after four seconds of speech emotions were usually classified correctly. We therefore opt for a four second speech segment input. Where the entire utterance is shorter than four seconds, we use zero padding to obtain an input of the required length.

4.4 EMOVOXCELEB Dataset

We apply our teacher-student framework on the VoxCeleb [Nagrani et al., 2017] dataset, a collection of *speaking face-tracks*, or contiguous groupings of talking face detections from video. The videos in the VoxCeleb dataset are interview videos of 1,251 celebrities uploaded to YouTube, with over 100,000 utterances (speech segments). The speakers span a wide range of different ages, nationalities, professions and accents. The dataset is roughly gender balanced. The audio segments also contain speech in different languages. While the identi-

	Train	Heard-Val	Unheard-Val
# speaking face-tracks	118.5k	4.5k	30.5k

Table 4.3: The distribution of speaking face-tracks in the EMOVOXCELEB dataset. The Heard-Val set contains identities that are present in Train, while the identities in Unheard-Val are disjoint from Train.

ties of the speakers are available, the dataset has *no emotion labels*, and the student model must therefore learn to reason about emotions entirely by transferring knowledge from the face network. The identity labels allow us to partition the dataset into three splits: Train, Heard-Val and Unheard-Val. The Heard-Val split contains held out speech segments from the same identities in the training set, while the Unheard-Val split contains identities that are disjoint from the other splits². Validating on unheard identities allows us to ascertain whether the student model is exploiting identity as a bias to better match the predictions of the teacher model. The identity labels may also prove useful for researchers tackling other tasks, for example evaluating the effect of emotional speech on speaker verification, as done by [Parthasarathy et al., 2017]. The total size of each partition is given in Table 4.3.

By applying the teacher model to the frames of the VoxCeleb dataset as described in section 4.3.1, we automatically obtain emotion labels for the face-tracks and the speech segments. These labels take the form of a predicted distribution over eight emotional states that were used to train the teacher model: *neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, *fear* and *contempt*. These frame-level predictions can then be directly mapped to synchronous speech segments by aggregating the individual prediction distributions into a single eight-dimensional vector for each speech segment. For all experiments we perform this aggregation by max-pooling across frames. However, since the best way to perform this aggregation remains an open topic of research, we release the frame level predictions of the model as part of the dataset annotation. The result is a large-scale audio-visual dataset of human emotion,

²The Unheard-Val split directly corresponds to the Test (US-UH) set defined in [Nagrani et al., 2018a].

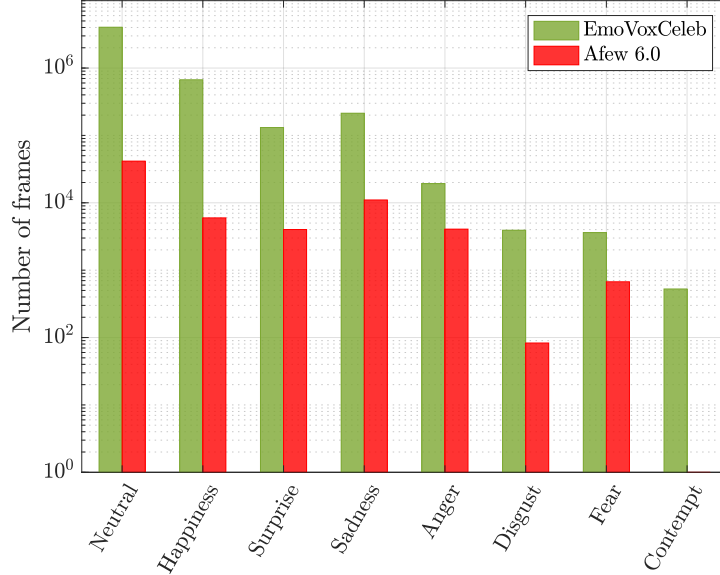


Figure 4.4: Distribution of frame-level emotions predicted by the SENet Teacher model for EMOVOXCELEB (note that the y-axis uses a log-scale). For comparison, the distribution of predictions are also shown for the Afew 6.0 dataset.

which we call the EMOVOXCELEB dataset. As a consequence of the automated labelling technique, it is reasonable to expect that the noise associated with the labelling will be higher than for a manually annotated dataset. We validate our labelling approach by demonstrating quantitatively that the labels can be used to learn useful speech emotion recognition models (Sec. 4.5.2). Face-track visualisations can be seen in Figure 4.3, and audio examples are available online³.

Distribution of emotions. As noted above, each frame of the dataset is annotated with a distribution of predictions. To gain an estimate of the distribution of emotional content in EMOVOXCELEB, we plot a histogram of the *dominant* emotion (the label with the strongest prediction score by the teacher model) for each extracted frame of the dataset, shown in Figure 4.4. While we see that the dataset is heavily skewed towards a small number of emotions (particularly neutral, as discussed in Sec. 4.3), we note that it still contains some diversity

³<http://www.robots.ox.ac.uk/~vgg/research/cross-modal-emotions>

of emotion. For comparison, we also illustrate the distribution of emotional responses of the teacher model on ‘Afew 6.0’ [Dhall et al., 2012], an emotion recognition benchmark. The Afew dataset was collected by selecting scenes in movies for which the subtitles contain highly emotive content. We see the distribution of labels is significantly more balanced but still exhibits a similar overall trend to EMOVOXCELEB. Since this dataset has been actively sampled to contain good diversity of emotion, we conclude that the coverage of emotions in EMOVOXCELEB may still prove useful, given that no such active sampling was performed. We note that Afew does not contain segments directly labelled with the *contempt* emotion, so we would therefore not expect there to be frames for which this is the predicted emotion. It is also worth noting that certain emotions are rare in our dataset. Disgust, fear and contempt are not commonly exhibited during natural speech, particularly in interviews and are therefore rare in the predicted distribution.

Data Format. As mentioned above, we provide logits (the pre-softmax predictions of the teacher network) at a frame level which can be used to directly produce labels at an utterance level (using max-pooling as aggregation). The frames are extracted from the face tracks at an interval of 0.24 seconds, resulting in a total of approximately 5 million annotated individual frames.

4.5 Experiments

To investigate the central hypothesis of this paper, namely that it is possible to supervise a speech emotion recognition model with a model trained to detect emotion in faces, we proceed in two stages. First, as discussed in Sec. 4.4, we compute the predictions of the SENet Teacher model on the frames extracted from the VoxCeleb dataset. The process of distillation is then performed by randomly sampling segments of speech, each four seconds in duration, from the training partition of this dataset. While a fixed segment duration is not required by our method (the student architecture can process variable-length clips by dynamically modi-

fying its pooling layer), it leads to substantial gains in efficiency by allowing us to batch clips together. We experimented with sampling speech segments in a manner that balanced the number of utterance level emotions seen by the student during training. However, in practice, we found that it did not have a significant effect on the quality of the learned student network and therefore, for simplicity, we train the student without biasing the segment sampling procedure.

For each segment, we require the student to match the response of the teacher network on the facial expressions of the speaker that occurred *during the speech segment*. In more detail, the responses of the teacher on each frame are aggregated through max-pooling to produce a single 8-dimensional vector per segment. As discussed in Section 4.3, both the teacher and student predictions are passed through a softmax layer before computing a cross entropy loss. Similarly to [Hinton et al., 2015], we set the temperature of both the teacher and student softmax layers to 2 to better capture the confidences of the teacher’s predictions. We also experimented with regressing the pre-softmax logits of the teacher directly with an Euclidean loss (as done in [Ba & Caruana, 2014]), however, in practice this approach did not perform as well, so we use cross entropy for all experiments. As with the predictions made by the teacher, the distribution of predictions made by the student are dominated by the neutral class so the useful signal is primarily encoded through the relative soft weightings of each emotion that was learned during the distillation process. The student achieves a mean ROC AUC of 0.69 over the teacher-predicted emotions present in the unheard identities (these include all emotions except disgust, fear and contempt) and a mean ROC AUC of 0.71 on validation set of heard identities on the same emotions.

4.5.1 Implementation Details

The student network is based on the VGGVox network architecture described in [Nagrani et al., 2017], which has been shown to work well on spectrograms, albeit for the task of speaker verification. The model is based on the lightweight VGG-M architecture, however the fully

connected *fc6* layer of dimension $9 \times n$ (support in both dimensions) is replaced by two layers – a fully connected layer of 9×1 (support in the frequency domain) and an average pool layer with support $1 \times n$, where n depends on the length of the input speech segment (for example for a 4 second segment, $n = 11$). This allows the network to achieve some temporal invariance, and at the same time keeps the output dimensions the same as those of the original fully connected layer. The input to the teacher image is an RGB image, cropped from the

Layer	Support	Filt dim.	# filts.	Stride	Data size
conv1	7×7	1	96	2×2	254×198
mpool1	3×3	-	-	2×2	126×99
conv2	5×5	96	256	2×2	62×49
mpool2	3×3	-	-	2×2	30×24
conv3	3×3	256	256	1×1	30×24
conv4	3×3	256	256	1×1	30×24
conv5	3×3	256	256	1×1	30×24
mpool5	5×3	-	-	3×2	9×11
fc6	9×1	256	4096	1×1	1×11
apool6	$1 \times n$	-	-	1×1	1×1
fc7	1×1	4096	1024	1×1	1×1
fc8	1×1	1024	1251	1×1	1×1

Table 4.4: The CNN architecture for the student network. The data size up until *fc6* is depicted for a 4-second input, but the network is able to accept inputs of variable lengths. Batchnorm layers are present after every conv layer.

source frame to include only the face region (we use the face detections provided by the VoxCeleb dataset) resized to 224×224 , followed by mean subtraction. The input to the student network is a short-term amplitude spectrogram, extracted from four seconds of raw audio using a Hamming window of width 25ms and step (hop) 10ms, giving spectrograms of size 512×400 . At train-time, the four second segment of audio is chosen randomly from the entire speaking face-track, providing an effective form of data augmentation. Besides performing mean and variance normalisation on every frequency bin of the spectrogram, no other speech-specific processing is performed, e.g. silence removal, noise filtering, etc. (following the approach outlined in [Nagrani et al., 2017]). While randomly changing the speed of audio segments can be useful as a form of augmentation for speaker verification [Nagrani et al., 2017], we do no such augmentation here since changes in pitch may have a significant impact

on the perceived emotional content of the speech.

Training Details. The network is trained for 50 epochs (one epoch corresponds to approximately one full pass over the training data where a speech segment has been sampled from each video) using SGD with momentum (set to 0.9) and weight decay (set to 0.0005). The learning rate is initially set to $1E-4$, and decays logarithmically to $1E-5$ over the full learning schedule. The student model is trained from scratch, using Gaussian-initialised weights. We monitor progress on the validation set of unheard identities, and select the final model to be the one that minimises our learning objective on this validation set.

4.5.2 Results on external datasets

To evaluate the quality of the audio features learned by the student model, we perform experiments on two benchmark speech emotion datasets.

RML: The RML emotion dataset is an acted dataset containing 720 audiovisual emotional expression samples with categorical labels: *anger, disgust, fear, happiness, sadness and surprise*. This database is language and cultural background independent. The video samples were collected from eight human subjects, speaking six different languages (English, Mandarin, Urdu, Punjabi, Persian, Italian). To further increase diversity, different accents of English and Chinese were also included.

eNTERFACE [Martin et al., 2006]: The eNTERFACE dataset is an acted dataset (in English) recorded in a studio. Forty-two subjects of fourteen nationalities were asked to listen to six successive short stories, each of which was designed to elicit a particular emotion. The emotions present are identical to those found in the RML dataset.

Both external datasets consist of acted speech, and are labelled by human annotators. Since the external datasets are obtained in a single recording studio, they are also relatively clean, in contrast to the noisy segments in EMOVOXCELEB. We choose the RML dataset for evaluation specifically to assess whether our embeddings can generalise to multilingual speech. Both

datasets are class-balanced.

Method	RML		eNTERFACE	
	Modality	Acc.	Modality	Acc.
Random	A	16.7	A	16.7
Student	A	49.7 ± 5.4	A	34.3 ± 4.0
Teacher	V	72.6 ± 3.9	V	48.3 ± 4.9
Noroozi et al. [F. Noroozi et al., 2017]	A	65.3	A	47.1

Table 4.5: Comparison of method accuracy on RML and eNTERFACE using the evaluation protocol of [F. Noroozi et al., 2017]. Where available, the mean \pm std. is reported.

We do not evaluate the predictions of the student directly, for two reasons: first, the set of emotions used to train the student differ from those of the evaluation test set, and second, while the predictions of the student carry useful signal, they skew towards neutral as a result of the training distribution. We therefore treat the predictions as 8-dimensional embeddings and adopt the strategy introduced in Sec. 4.3.1 of learning a map from the set of embeddings to the set of target emotions, allowing the classifier to re-weight each emotion prediction using the class confidences produced by the student. In more detail, for each dataset, we evaluate the quality of the student model embeddings by learning a single affine transformation (comprising a matrix multiply and a bias) followed by a softmax to map the 8 predicted student emotions to the target labels of each dataset. Although our model has been trained using segments of four seconds in length, its dynamic pooling layer allows it to process variable length segments. We therefore use the full speech segment for evaluation.

To assess the student model, we compare against the following baselines: the expected performance at chance level by a random classifier; and the performance of the teacher network, operating on the faces modality. We also compare with the recent work of [F. Noroozi et al., 2017], whose strongest speech classifier consisted of a random forest using a combination of 88 audio features inc. MFCCs, Zero Crossings Density (ZCD), filter-bank energies (FBE) and other pitch/intensity-related components. We report performance using 10-fold cross validation (to allow comparison with [F. Noroozi et al., 2017]) in Table 4.5. While it falls short of the performance of the teacher, we see that the student model performs significantly better than

chance. These results indicate that, while challenging, transferring supervision from the facial domain to the speech domain is indeed possible. Moreover, we note that the conditions of the evaluation datasets differ significantly from those on which the student network was trained. We discuss this domain transfer problem for emotional speech in the following section.

Anger	0.61	0.07	0.17	0.03	0.04	0.07
Disgust	0.07	0.85	0.05	0.03	0.01	0.00
Fear	0.14	0.02	0.44	0.01	0.29	0.10
Happiness	0.03	0.02	0.01	0.93	0.00	0.02
Sadness	0.01	0.05	0.23	0.01	0.68	0.03
Surprise	0.03	0.03	0.07	0.01	0.03	0.84
	Anger	Disgust	Fear	Happiness	Sadness	Surprise

Anger	0.72	0.07	0.01	0.03	0.00	0.18
Disgust	0.06	0.67	0.05	0.10	0.07	0.06
Fear	0.06	0.17	0.35	0.17	0.19	0.07
Happiness	0.10	0.17	0.13	0.35	0.07	0.18
Sadness	0.00	0.08	0.15	0.05	0.71	0.01
Surprise	0.28	0.09	0.04	0.22	0.02	0.36
	Anger	Disgust	Fear	Happiness	Sadness	Surprise

Figure 4.5: Normalised confusion matrices for the teacher model (left) and the student model (right) on the RML dataset (ground truth labels as rows, predictions as columns).

4.5.3 Discussion

Evaluation on external corpora: Due to large variations in speech emotion corpora, speech emotion models work best if they are applied under circumstances that are similar to the ones they were trained on [Schuller et al., 2010]. For cross-corporal evaluation, most methods rely heavily on domain transfer learning or other adaptation methods [Z. Zhang et al., 2011; J. Deng, Zhang, Eyben, & Schuller, 2014; J. Deng, Zhang, & Schuller, 2014]. These works generally agree that cross-corpus evaluation works to a certain degree only if corpora have similar contexts. We show in this work that the embeddings learnt on the EMOVOXCELEB dataset can generalise to different corpora, even with differences in nature of the dataset (natural versus acted) and labelling scheme. While the performance of our student model falls short of the teacher model that was used to supervise it, we believe this represents a useful step towards the goal of learning useful speech emotion embeddings that work on multiple

corpora without requiring speech annotation.

Challenges associated with emotion distillation: One of the key challenges associated with the proposed method is to achieve a consistent, high quality supervisory signal by the teacher network during the distillation process. Despite reaching state-of-the-art performance on the FERplus benchmark, we observe that the teacher is far from perfect on both the RML and eNTERFACE benchmarks. In this work, we make two assumptions: the first is that distillation ensures that even when the teacher makes mistakes, the student can still benefit, provided that there is signal in the uncertainty of the predictions. The second is a broader assumption, namely that deep CNNs are highly effective at training on large, noisy datasets (this was recently explored in [Rolnick et al., 2017; Mahajan et al., 2018], who showed that despite the presence of high label noise, very strong features can be learned on large datasets). To better understand how the knowledge of the teacher is propagated to the student, we provide confusion matrices for both models on the RML dataset in Figure 4.5. We observe that the student exhibits reasonable performance, but makes more mistakes than the teacher for every emotion except sadness and anger. There may be several reasons for this. First, EMOVOXCELEB used to perform the distillation may lack the distribution of emotions required for the student to fully capture the knowledge of the teacher. Second, it has been observed that certain emotions are easier to detect from speech than faces, and vice versa [Busso et al., 2004], suggesting that the degree to which there is a redundant emotional signal across modalities may differ across emotions.

Limitations of using interview data: Speech as a medium is intrinsically oriented towards another person, and the natural contexts in which to study it are interpersonal. Interviews capture these interpersonal interactions well, and the videos we use exhibit real world noise. However, while the interviewees are not asked to act a specific emotion, i.e. it is a ‘natural’ dataset, it is likely that celebrities do not act entirely naturally in interviews. Another drawback is the heavily unbalanced nature of the dataset where some emotions such as contempt and fear occur rarely. This is an unavoidable artefact of using real data. Several works have

shown that the interpretation of certain emotions from facial expressions can be influenced to some extent by contextual clues such as body language [Aviezer et al., 2009; Hassin et al., 2013]. Due to the talking-heads nature of the data, this kind of signal is typically not present in interview data, but could be incorporated as clues into the teacher network.

Student Shortcuts: The high capacity of neural networks can sometimes allow them to solve tasks by taking shortcuts by exploiting biases in the dataset [Doersch et al., 2015]. One potential for such a bias in EMOVOXCELEB is that interviewees may often exhibit consistent emotions which might allow the student to match the teacher’s predictions by learning to recognise the identity, rather than the emotion of the speaker. As mentioned in Sec. 4.5, the performance of the student on the `heardVal` and `unheardVal` splits is similar (0.71 vs 0.69 mean ROC AUC on a common set of emotions), providing some confidence that the student is not making significant use of identity as a shortcut signal.

Extensions/Future Work: First, we note that our method can be applied as is to other mediums of unlabelled speech, such as films or TV shows. We hope to explore unlabelled videos with a greater range of emotional diversity, which may help to improve the quality of distillation and address some of the challenges discussed above. Second, since the act of speaking may also exert some influence on the facial expression of the speaker (for example, the utterance of an o sound could be mistaken for surprise) we would also like to explore the use of proximal *non-speech* facial expressions as a supervisory signal in future work. Proximal supervision could also address the problem noted in Section 4.3, that speaking expressions can tend towards neutral. Finally, facial expressions in video can be learnt using self-supervision [Wiles et al., 2018], and this offers an alternative to the strong supervision used for the teacher in this paper.

4.6 Conclusions

We have demonstrated the value of using a large dataset of emotion unlabelled video for cross-modal transfer of emotions from faces to speech. The benefit is evident in the results – the speech emotion model learned in this manner achieves reasonable classification performance on standard benchmarks, with results far above random. We also achieve state of the art performance on facial emotion recognition on the FERPlus benchmark (supervised) and set benchmarks for cross-modal distillation methods for speech emotion recognition on two standard datasets, RML and eNTERFACE.

The great advantage of this approach is that video data is almost limitless, being freely available from YouTube and other sources. Future work can now consider scaling up to larger unlabelled datasets, where a fuller range of emotions should be available.

Acknowledgements. The authors would like to thank the anonymous reviewers, Almut Sophia Koepke and Judith Albanie for useful suggestions. We gratefully acknowledge the support of EPSRC CDT AIMS grant EP/L015897/1, and the Programme Grant Seebibyte EP/M013774/1.

Appendices

Appendices, data, code and further visualisations for this chapter can be accessed online.⁴

Statement of Authorship

A statement of authorship for this work can be found in Appendix C.

⁴<http://www.robots.ox.ac.uk/~vgg/research/cross-modal-emotions/>

Part II

Self-Supervised Representation Learning

And then, when it was made simple, distilled,
counted in bits, information was found to be
everywhere.

— James Gleick

5 | *Seeing Voices and Hearing Faces: Cross-modal biometric matching*

Arsha Nagrani Samuel Albanie Andrew Zisserman

VGG, Oxford

Abstract

We introduce a seemingly impossible task: given only an audio clip of someone speaking, decide which of two face images is the speaker. In this paper we study this, and a number of related cross-modal tasks, aimed at answering the question: how much can we infer from the voice about the face and vice versa?

We study this task “in the wild”, employing the datasets that are now publicly available for face recognition from static images (VGGFace) and speaker identification from audio (VoxCeleb). These provide training and testing scenarios for both static and dynamic testing of cross-modal matching. We make the following contributions: (i) we introduce CNN architectures for both binary and multi-way cross-modal face and audio matching; (ii) we compare dynamic testing (where video information is available, but the audio is not from the same video) with static testing (where only a single still image is available); and (iii) we use human testing as a baseline to calibrate the difficulty of the task. We show that a CNN can indeed be trained to solve this task in both the static and dynamic scenarios, and is even well above chance on 10-way classification of the face given the voice. The CNN matches human performance on easy examples (e.g. different gender across faces) but exceeds human performance on more challenging examples (e.g. faces with the same gender, age and nationality).

Published in the Proceedings of the [Conference on Computer Vision and Pattern Recognition, 2018](#).

5.1 Introduction

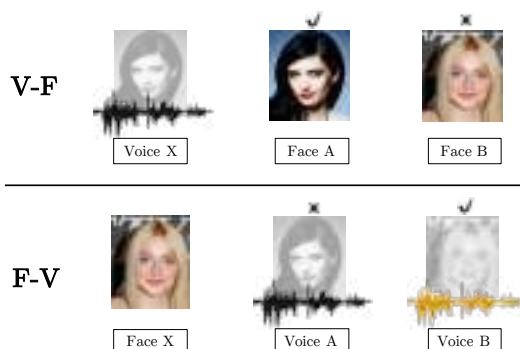


Figure 5.1: We introduce the task of cross-modal biometric matching, and consider two specific formulations of the problem: (*Top*) *V-F*: given an audio clip of a voice and two or more face images/videos, select the face image/video that corresponds to the voice. (*Bottom*) *F-V*: given an image or video of a face, determine the corresponding voice.

Can you recognise someone’s face if you have only heard their voice? Or recognise their voice if you have only seen their face? As humans, we may ‘*see voices*’ or ‘*hear faces*’ by forming mental pictures of what a person looks like after only hearing their voice, or vice versa. This phenomenon has been investigated in a number of studies on human perception and neurology [Kamachi et al., 2003; H. M. Smith et al., 2016b], where participants completed a sequential two-alternative forced choice matching task. They were asked to listen to a human voice (Voice X), and then pick the face corresponding to the same identity, between two still/dynamic face images. It is also a familiar trope in Hollywood films that someone can be recognised after only hearing their voice – for example in the film “Die Hard”, where the Bruce Willis character (John Mclane) emerges from the building towards the end of the film and is instantly able to recognise the cop (Sgt. Al Powell) who he has only spoken to by radio throughout the film, but never seen. This task of *cross-modal* (face and voice) recognition, or ‘cross-modal biometric matching’, is the objective of this paper. As illustrated in figure 5.1, there are two related tasks: first, given an image or video of a face, determine which of two or more voices it corresponds to; second, and conversely, given an audio clip of a voice, deter-

mine which of two or more face images or videos it corresponds to. Note, the voice and face video are not acquired simultaneously, so methods of active speaker detection that may rely on synchronisation of the audio and lip motion, e.g. [J. S. Chung & Zisserman, 2016b] cannot be employed here.

That this task might be possible at all is due to the existence of factors that are common to both modalities; in particular, specific latent properties (like age, gender, ethnicity/accent) influence both the facial appearance and voice. Besides these, there exist other, more subtle cross-modal biometrics. Studies in biology and evolutionary perception [Wells et al., 2013] show that hormone levels during puberty affect both face morphology and voice pitch. In males, higher testosterone-oestrogen ratios lead to a prominent eyebrow ridge, broad chin, small eyes, and thin lips [Thornhill & Møller, 1997], while vocal folds situated in the larynx also increase in size, thus leading to a lower voice pitch [Hollien & Moore, 1960]. Similarly for females, higher oestrogen levels cause large eyes and full lips [Thornhill & Møller, 1997] and prevent the vocal folds from enlargement, leading to higher voice pitch [Hollien & Moore, 1960]. Besides the above static properties, given a video stream, we expect there to exist more (dynamic) cross-modal biometrics. For example, the ‘manner of speaking’ can be an important cross-modal biometric. Sheffert and Olson [Sheffert & Olson, 2004] suggested that visual information about a person’s particular idiosyncratic speaking style is related to the speaker’s auditory attributes. The origins of this link lie in the mechanics of speech production, which, when shaping the vocal tract, determines both facial motion as well as the sound of the voice [Kamachi et al., 2003].

Apart from establishing that it is indeed possible to solve cross-modal biometric matching, which is an interesting scientific result on its own, there are also practical applications of the technology – not least in surveillance. Imagine the following scenario: the only information we have about a person is a handful of speaking (audio) samples, because the data was recorded from telephone conversations. We then want to identify the individual from a video stream, for example from CCTV. A more benign application is automatically labelling char-

acters in TV and film material where characters may be heard but not seen at the same time, and so cross-modal matching can be used to infer the labels.

In this paper we approach the problem using the tools of deep learning trained on large-scale datasets. We make the following contributions: first, we introduce a CNN architecture that ingests face images and voice spectrograms, and is able to infer the correspondence between them. The network is trained on a large-scale dataset of voices (VoxCeleb [Nagrani et al., 2017]) and faces (VGGFace [Parkhi et al., 2015]) from the same identities. Second, we investigate the performance of the network using still, dynamic images, or both. We show, in contrast to the findings in the perception literature, that the task can be solved far better than chance using static images alone, and that the performance improves further using dynamic images. We also carry out our own study of human performance using AMT. Finally, we generalise the two-alternative forced choice architecture to multi-way classification and report results for this more challenging task.

5.2 Related Work

Human Perception Studies: The broad consensus among studies exploring cross-modal matching of faces and voices using human participants, is that matching is only possible when dynamic visual information about articulatory patterns is available [Kamachi et al., 2003; Lachs & Pisoni, 2004; Rosenblum et al., 2006]. In particular, works have demonstrated coupling between an individual’s idiosyncratic speaking style, the sound of their voice and the manner in which their face moves [Lander et al., 2007; Yehia et al., 1998; Cvejic et al., 2012], suggesting the presence of dynamic information which can be exploited to solve the matching task. Although these studies demonstrate that static face voice matching performance lies at chance level [Kamachi et al., 2003; Lachs & Pisoni, 2004], we note that there has been research which challenges this perspective [H. M. Smith et al., 2016a; Krauss et al., 2002]. However, while Krauss et al. [Krauss et al., 2002] showed that people could match a voice to

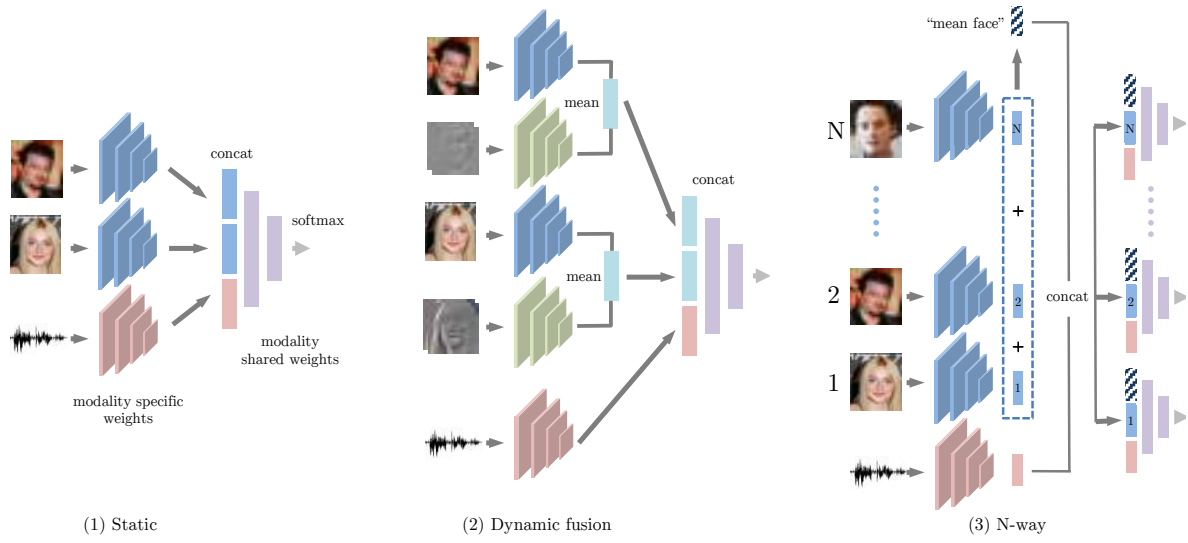


Figure 5.2: The three main networks architectures used in this paper. From left to right: (1) The *static* 3-stream CNN architecture consisting of two face sub-networks and one voice network, (2) a 5-stream *dynamic-fusion* architecture with two extra streams as dynamic feature sub-networks, and finally (3) the *N-way classification* architecture which can deal with any number of face inputs at test time due to the concept of query pooling (see Sec. 5.3.2). Voice-only weights shown in red, face-only weights in blue/green (static/dynamic), and modality-shared weights in purple (best viewed in colour). Modality-specific weights of the same colour are shared amongst different inputs.

one of two static images with above-random accuracy, the stimuli were *full-body* images rather than images of faces, which may have provided additional cues to inform accurate matching (see [H. M. Smith et al., 2016b] for a detailed discussion of these contradictory results). It is worth noting that the difficulty of the task is highly dependent on the specific stimuli sets provided—as we show in this work, some face-voice combinations are more distinctive than others.

Face Recognition and Speaker Identification: The tasks of face recognition and speaker identification are longstanding problems in the vision and speech research communities, and consequently an in-depth review of these topics is beyond the scope of this work. However, we note that the recent advent of deep CNNs with large datasets has considerably advanced the state-of-the-art in both face recognition [Taigman et al., 2014a; Parkhi et al., 2015; Y. Sun et al., 2014; Kemelmacher-Shlizerman et al., 2016] and speaker recognition [Dehak et al., 2011; Nagrani et al., 2017; Saon et al., 2013; Snyder et al., 2017]. Unfortunately, while these recognition models have proven remarkably effective at representation learning from a single modality, the alignment of learned representations across the modalities is less developed. In this work we address this issue through the development of a multimodal architecture that directly ingests data from both faces and voices and learns a correspondence between them.

Cross-modal Matching: Cross-modal matching has received considerable attention using visual data and text (natural language). Methods have been developed to establish mappings from images [Frome et al., 2013; G. Kulkarni et al., 2013; Karpathy & Fei-Fei, 2015; Kiros et al., 2014; Vinyals et al., 2015] and videos [Venugopalan et al., 2014] to textual descriptions (e.g. captioning), generating visual models from text [J. Wang et al., 2009; Zitnick et al., 2013] and solving visual question answering problems [Antol et al., 2015; X. Lin & Parikh, 2015; Malinowski & Fritz, 2014]. In cross-modal matching between video and *audio* however, work is limited, particularly in the field of biometrics (person or speaker recognition). Recent work has begun to explore the tasks of audio-visual matching for scenes and objects [Aytar et al., 2016; Aytar, Castrejon, et al., 2017; Arandjelović & Zisserman, 2017; Owens et al.,

2016] and audio-visual speech recognition (lip reading [J. S. Chung, Senior, et al., 2017], lip sync [J. S. Chung & Zisserman, 2016b] etc). In biometrics, there has also been work that uses both modalities to improve performance [Brunelli & Falavigna, 1995; Khoury et al., 2014] but not one to recognise the other. Le and Odobez [Le & Odobez, 2017] use transfer learning from face embeddings to try and improve speaker diarisation results. The only attempt we can find to solve a similar task to the one proposed here (but only for videos, and not still face images) is by [Roy & Marcel, 2010]. This work seeks to map a statistical model of the features in one modality to a statistical model of the features in another modality. It is evaluated on the M2VTS audio-visual database for 25 male subjects who count from zero to nine. In contrast, we aim to solve this task at large-scale, ‘in the wild’, and with longer more natural speech segments from unconstrained interview videos.

5.3 Cross-Modal Models

For the task of forced matching between two faces and voice input (V-F formulation), our objective is to identify which of a pair of given faces possesses the same identity as the voice. Since this problem admits a natural symmetry with the F-V formulation (matching between two voices and a face), each component of our method can be readily adapted to address either task. For notational clarity, we focus our description on the V-F formulation. The forced matching task can be defined as follows; let $x = \{v, f_1, f_2\}$ denote a set consisting of an anchor voice segment v and two face images f_1 and f_2 . Each input set x contains one positive and one negative face, where face f_i is defined as positive if it possesses the same identity as the anchor voice, and negative otherwise. We pose the matching task as a binary classification problem, in which the objective is to predict the position $y \in \{1, 2\}$ of the positive face. Given images and voices of known identity, we can construct a dataset of training examples $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$ by simply randomising the position of the positive face in each face pair. The learning problem corresponds to maximising likelihood: $\theta =_{\theta} L(g_{\theta}; \mathcal{D})$,

where g_θ is the parameterised model to be learned. The loss to be minimised can then be framed as a cross-entropy loss on target label positions.

We instantiate g_θ as a three-stream convolutional neural network, taking inspiration from the *odd-one-out* network architecture proposed in [Fernando et al., 2016]. Our forced matching task, however, is unique in the sense that we would like to perform it across two different *modalities*. Our model design consists of three modality specific sub-networks (or streams); two parameter-sharing face sub-networks that ingest image data and a voice sub-network which ingests spectrograms. The three streams are then combined through a fusion layer (via feature concatenation) and fed into modality-shared fully connected layers on top. The fusion layer is required to enable the network to establish a correspondence between faces and voices.

Our model hence has two kinds of layers, modality specific (face and voice) layers and higher-level layers which are shared between both modalities. Similar to the motivation in [Aytar, Castrejon, et al., 2017], the rationale behind this architecture is to force early layers to specialise to modality specific features (such as edges in face images and spectral patterns in audio segments), while allowing later layers to capture higher-level latent cross-modal variables (such as gender, age, ethnicity and identity). For the sake of clarity, we state here the three main tasks that we solve in this paper: 1) Static matching, which uses only still face images, 2) Dynamic matching, which involves videos of faces during speech, and 3) N-way classification, which is an extension of the matching task to any number of faces (greater than two). These tasks are described in more detail in section 5.5. In order to capture dynamic facial appearance, we introduce an additional sub-network which ingests dynamic features extracted from videos. To motivate the design of each of these sub-networks, we next discuss the input representations upon which they will operate.

5.3.1 Input Representations

Voices: The input to the voice stream is a short term magnitude spectrogram extracted directly from raw audio. The audio stream is extracted from the video and converted to single-channel, 16-bit streams at a 16kHz sampling rate for consistency. Spectrograms are then generated in a similar manner to that in [Nagrani et al., 2017], giving spectrograms of size 512×300 for three seconds of speech. We perform mean and variance normalisation on every frequency bin of the spectrum, but apply no further speech-specific preprocessing (e.g. silence removal, voice activity detection, or background noise suppression).

Static Faces: Each input to the face stream consists of an RGB image, which has been cropped from a source image to contain only the region of the image surrounding a face. The locations of these crops are provided by the datasets used in our experiments (discussed further in Sec. 5.4)¹. The resulting region is then resized to a fixed 224×224 input.

Dynamic Faces: The annotated face regions contained in video data are processed as *face-tracks*, defined to be contiguous sequences of frames possessing the same identity. To exploit dynamic cross-modal information from idiosyncratic speaking styles, an estimate of motion is required. Previous work that seeks to perform speaker recognition solely with visual information [G. Zhao & Pietikäinen, 2013; Çetingül et al., 2006; Cetingul et al., 2005; Ouyang & Lee, 2006] tends to focus primarily on the lip region. While useful biometric information is concentrated around the lips, (e.g. when uttering the same phoneme or word, different speakers have different mouth shapes [Ouyang & Lee, 2006]), we hypothesise that the motion of other facial features, e.g. eyes or eyebrows, or even the entire motion of the head during speech, could be useful biometric cues for identification. We would therefore like to work with a representation of this data that is capable of extracting temporal information from each full face-track.

A wide range of approaches have been proposed to enable CNNs to exploit temporal infor-

¹Since in both datasets, the specified face regions yield a tight face crop, we expand all crops by a factor of $\times 1.6$ to incorporate additional context into the face region.

mation from video, including 3D convolutions [Ji et al., 2013], optical flow [Simonyan & Zisserman, 2014] and dynamic images [Bilen et al., 2016] which have proven to be particularly effective in the context of human action recognition. In this work, we employ the dynamic image representation, which computes a fixed size representation of a video sequence by learning a ranking machine on the raw pixel input across a given sequence of frames. See section 5.7 for variant implementation details.

5.3.2 Architectures

(1) Static Architecture: Our base architecture comprises two face sub-networks and one voice sub-network. Both the face and voice streams use the VGG-M architecture [Chatfield et al., 2011], which achieves a good trade-off between efficiency and performance. The features from each stream are fused through concatenation to form a 3072-dimensional feature², which is then processed by three fully connected layers with hidden units of dimensionality 1024, 512 and 2 respectively. Further details of each sub-network can be found in the appendix.

(2) Dynamic-Fusion Architecture: Motivated by the effectiveness of dual stream architectures that combine RGB images with temporal features in action recognition [Simonyan & Zisserman, 2014; Bilen et al., 2016], we also explore a variant of the base architecture which includes an additional dynamic image stream for each input face. The features computed for each face (RGB + dynamic) are combined after the final fully connected layer in each stream through summation. In more detail, given an augmented input set $x = \{v, f_1, f_2, d_1, d_2\}$, where d_1 and d_2 are dynamic face inputs, we compute the representation

$$\phi(x) = \phi_2 \circ \left[(\phi_f(f_1) + \phi_d(d_1)) \oplus (\phi_f(f_2) + \phi_d(d_2)) \oplus \phi_a(v) \right]$$

²Each of the three sub-networks produces a 1024-dimensional vector.

where \oplus denotes concatenation, ϕ_f represents the RGB face sub-network, ϕ_d the dynamic image face sub-network, ϕ_a the operations of the audio sub-network, and ϕ_2 the modality-shared fully connected layers on top. The two static face streams and two dynamic face streams share separate weights, allowing the different types of face input to be treated accordingly.

(3) N-way Classification Architecture: We further extend the architecture to deal with the more challenging task of developing a general cross-modal biometric system that is capable of solving an $N : 1$ identification problem. The input to this network consists of an anchor voice segment v , 1 positive face and $N - 1$ negative faces. As before, the target label $y \in \{1, 2, \dots, N\}$ denotes the position of the positive face, resulting in an N -way classification problem.

As a consequence of using concatenation as a fusion layer in our base architecture, the number of face streams cannot be adjusted during inference. This shortcoming is common to many CNN architectures, where it is difficult to change the number of inputs at test time. One approach to resolving this issue would be to concatenate the voice to each face stream separately, however in this scenario each face stream would be unaware of the presence of the other streams. In order to avoid this problem, we add a mean pooling layer to each face stream which calculates the ‘mean face’ of all the faces in a particular query, thereby making each stream *context aware*. We refer to this simple concept as ‘query pooling’.

5.4 Datasets and Training

Due to the novel nature of the task explored in this work, no large-scale public benchmarks exist for evaluating our approach. We therefore construct a new dataset to train and evaluate our method by combining two available datasets with overlapping identities:

VGGFace [Parkhi et al., 2015]: VGGFace is a large-scale dataset of still face images collected from search engines. We use the ‘curated’ version of this dataset.

	Train	Val	Test	Total
# of identities	942	116	189	1,247
<i>VGGFace Dataset</i>				
# of face images	873,382	47,759	74,564	995,705
<i>VoxCeleb Dataset</i>				
# of speech segments	116,480	14,630	22,376	153,486
# of videos	16,820	2,044	3,425	22,295

Table 5.1: Statistics for the VGGFace and VoxCeleb datasets. The numbers shown here are only for the overlapping identities in the two datasets.

VoxCeleb [Nagrani et al., 2017]: VoxCeleb is a large-scale audio-visual dataset of human speech collected from YouTube videos. Since this dataset is collected ‘in the wild’, it covers a wide range of different recording environments and background noise levels. This dataset contains both video and audio.

For the purposes of this task, we use only the data for the 1,247 identities that overlap between the two datasets.

Train/Test Split: Identities in the train and test datasets do not overlap. All speakers whose names start with ‘A’ or ‘B’ are reserved for validation, while speakers with names starting with ‘C’, ‘D’, ‘E’ are reserved for testing. This yields a good balance of male and female speakers (dataset statistics are given in table 5.1).

Gender, Nationality and Age (GNA) Variation: To enable a more thorough analysis of our method, gender and nationality labels for speakers in the dataset were obtained by crawling Wikipedia. Note that we crawl for *nationality*, and not *ethnicity*, since this is a variable typically more informative of accent. We use these labels to construct a more challenging test set, wherein each triplet contains speakers of the same gender, broad age bracket (speakers between the ages of 30-50 years old were selected manually), and nationality (we restrict to U.S nationals). We note that these conditions are similar to those established during the human perception studies discussed previously [Kamachi et al., 2003; Lachs & Pisoni, 2004; Rosenblum et al., 2006].

In the sections that follow, we refer to each input x containing two face (either from stills or video) and one voice representation as a *triplet*.

5.4.1 Training Protocol

All networks are trained end-to-end using stochastic gradient descent with batch normalisation. We use a minibatch size of 64, momentum (0.9), weight decay ($5E - 4$) and a logarithmically decaying learning rate (initialised to 10^{-2} and decaying to 10^{-8}). The face and voice sub-networks are initialised using the pre-trained weights from the VGGFace and VoxCeleb models trained for face and speaker identification respectively, while the modality shared weights are initialised from a Gaussian distribution. When processing face images, we apply the data augmentation techniques used on the ImageNet classification task by [Krizhevsky et al., 2012] (i.e. random cropping, flipping, colour shift). For the audio segments, we change the speed of each segment by choosing a random speed ratio between 0.95 to 1.05. We then extract a random 3s segment from the audio clip at train time. Training uses $1.2M$ triplets that are selected at random (and the choice is then fixed). Networks are trained for 10 epochs, or until validation error stops decreasing, whichever is sooner.

5.5 Experiments

5.5.1 Tasks

Static Matching: Under the static evaluation, each test sample consists of two *static* face images and single speech segment. To construct the test set for this benchmark, we use audio segments from VoxCeleb [Nagrani et al., 2017] and face images from VGGFace [Parkhi et al., 2015]. We make use of both still images from VGGFace and frames extracted from the videos in the VoxCeleb dataset during training. When processing frames extracted from the VoxCeleb videos, we ensure that the audio segments and frames in a single triplet are not

sourced from the same video.

Dynamic Matching: The dynamic evaluation assesses performance on videos of human speech. In addition to static cross-modal biometrics, in this setting a person’s ‘manner of speaking’ may also provide important source of identity information. The dataset for this benchmark consists of videos and audio both extracted from VoxCeleb [Nagrani et al., 2017]. A triplet in this case consists of two face-tracks and one audio segment.

For the purposes of this task, it is important to minimise any correlation or mutual information based on audio-visual synchrony which could arise if the audio and visual data were extracted at the same time (for example: mouth motion based on the exact lexical content of the sentence, the emotional state of the speaker etc.). While interesting in their own right, these factors do not constitute cross modal biometrics for person verification, and therefore exploiting their presence to solve the matching task is not the objective of this work; we wish to be sensitive *only* to identity. Hence we ensure that the audio segments and face-tracks in a single triplet are not sourced from the same video. While we experiment with different methods for extracting dynamic information from a face-track (described in detail in section 5.7), the best results were obtained using dense sampling in order to obtain multiple aligned RGB and dynamic images from each face-track. These inputs are then fed into the fusion architecture.

At test time, we adopt a simple strategy to combine predictions from the densely extracted frames. The predictions from each frame are averaged to give a single prediction per triplet. Since we may have two face-tracks of differing lengths in each triplet, the frames from the longer face-track are selected using a stride s , where $s = \left\lfloor \frac{\text{length}(F_1)}{\text{length}(F_2)-1} \right\rfloor$ and F_1 and F_2 are the two face-tracks.

N-way Classification: We also extend the V-F task to one of $1 : N$ classification. It is important to note that such a task is extremely challenging, particularly since as N increases, the likelihood of solving the problem using isolated variables such as age, gender or ethnicity in isolation (or in combination) diminishes. We use the N-way classification architecture

(Figure 5.2, right) to tackle this task. This architecture allows us to train with any number of face images N_{Tr} , and then test with any number of test images N_{Te} , where N_{Tr} does not have to be equal to N_{Te} . We experimented with different values of $N_{Tr} = 2, 3, 5$, however we found that changing the number of faces at train time did not significantly improve results. We therefore report results of accuracy A_I vs N_{Te} trained using $N_{Tr} = 2$ (figure 5.3).

Evaluation Protocol: The static and dynamic cases are evaluated on 10,000 triplets randomly chosen from the test set. This gives a good balance of easy and difficult triplets. The N-way case is evaluated on 10,000 inputs, again chosen randomly. For all three of the above cases, we use the entire audio segment at test time with standard average pooling, following the exact procedure used in [Nagrani et al., 2017].

5.5.2 AMT Human Baselines:

Since there are no prior baselines to compare to, it is useful to have a measure of how well humans are able to perform forced matching between faces and voices. Direct comparison with studies on human perception [Kamachi et al., 2003; Lachs & Pisoni, 2004; Rosenblum et al., 2006] is infeasible given the likely difference in distributions of datasets. We therefore calibrate the difficulty of our dataset by performing our own human study on Amazon Mechanical Turk (AMT). For this study, a set of 500 triplets were randomly sampled from the static test set. Each sample was shown to 20 different workers on AMT in batches of five triplets. An in-depth description of the study can be found in the appendices provided online³, and the results are shown in table 5.2.

³<http://www.robots.ox.ac.uk/vgg/publications/2018/Nagrani18a/nagrani18a.pdf>

	Static Test		Dynamic Test	
	$A_I(\text{Total})$	$A_I(\text{GNA-var removed})$	$A_I(\text{Total})$	$A_I(\text{GNA-var removed})$
V-F	81.0	63.9	84.3	67.4
F-V	79.5	63.4	82.9	65.6
Human Baseline (V-F)	81.3	57.1	-	-

Table 5.2: Results are reported using % Identification Accuracy A_I which is calculated using 10,000 test triplets. Since this is a 2-way forced matching task, chance is 50%.

5.5.3 Evaluation Measures

We define two metrics to evaluate performance; *Identification Accuracy* and *Marginal Accuracy*. Following the notation introduced in Sec. 5.3, let $D = \{x_n, y_n\}_{n=1}^N$ denote a set of labelled examples where each input triplet takes the form $x_n^{(i,j)} = \{v^{(i)}, f^{(i)}, f^{(j)}\}, i, j \in \mathcal{I}$ (here \mathcal{I} denotes the set of identities). We define the identification accuracy of a predictive model g as follows:

$$A_I = \frac{1}{N} \sum_n |g(x_n^{(i,j)}) = y_n|$$

We further define the marginal accuracy of a predictive model g as:

$$m_A(s) = \frac{1}{\mathcal{N}_s} \sum_{(i=s \vee j=s)} |g(x_n^{(i,j)}) = y_n|$$

where $\mathcal{N}_s := |\{x_n^{(i,j)} : i = s \vee j = s\}|$ represents the number of triplets containing the speaker s . Identification accuracy provides a measure of performance on the entire test set, while marginal accuracy enables us to determine speaker-specific performance.

5.6 Results and Discussion

Static and Dynamic Matching: We report the results of both the F-V and the V-F formulations for the static and dynamic cases in Table 5.2. The results for the dynamic task are better

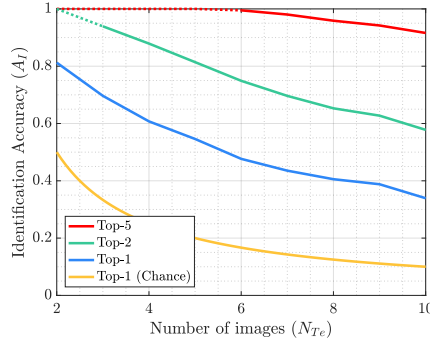


Figure 5.3: Top-1, top-2 and top-5 identif. acc. (A_I) vs the number of face images N_{Te} at test time. As can be seen from the graph, the model performs well above chance for all values of N_{Te} .

than those for the static task (by more than 3% for the V-F case). Since the identities in the two datasets are exactly the same, we infer that this increase in accuracy may be due to the presence of visually dynamic information from articulatory patterns.

N-way classification: The accuracy A_I vs N_{Te} for a model trained using $N_{Tr} = 2$ is shown in Figure 5.3. As observed from Figure 5.3, although A_I reduces as the number of faces at test time N_{Te} increases, the ratio A_I/A_R indicating the relative improvement of the proposed system compared to chance A_R remains relatively stable, validating the efficacy of our approach.

V-F vs F-V cases: The similar accuracies suggest that the task is highly symmetric in nature, aligning closely with the outcome of a prior study in human perception [Kamachi et al., 2003].

Comparison to the Human Benchmark: The AMT studies show good human performance on the static test set without GNA-variation removal. As can be seen in Table 5.2, the model is comparable to human performance on this task. On the more challenging test set with GNA-variation removed, however, human performance is significantly lower. Interestingly, on this setting the model manages to exceed human performance, which may suggest the presence of *subtle* cross-modal biometrics that are difficult for un-trained humans to identify. We note however, that it is difficult to eliminate all biases that may be exploitable by humans/algorithms performing this task. For instance, since the images in our dataset are

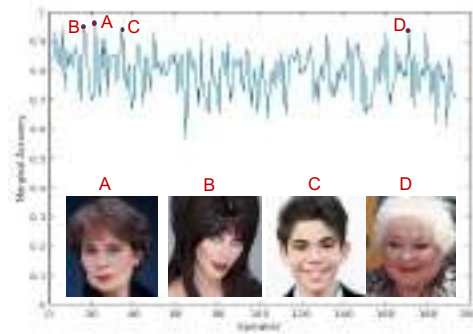


Figure 5.4: Marginal Accuracies: Marginal accuracies are computed for all speakers in the test set for the static V-F task. The highest marginal accuracies are for A. Celia Imrie, B. Cassandra Peterson, C. Cameron Boyce and D. Estelle Harris.

sourced from the identities of celebrities (which may occasionally be recognised by the workers), there is likely to be a degree of positive bias in the results from the human study. Moreover, as a consequence of the large-scale nature of the data, it may be possible for the model to learn to use other correlated factors which would be difficult to detect without extensive annotation, e.g. speakers of certain professions, such as sportspeople, may be more likely to be interviewed outside.

Marginal Accuracies: An examination of the marginal accuracies of our model shows that some face-voice combinations are significantly more discriminative than others (Figure 5.4). Three speakers who appear to be particularly distinctive, (all with marginal accuracies above 90%), are ‘Cameron Boyce’, who is a child actor, ‘Estelle Harris’, who (quoting directly from her Wikipedia page) is ‘easily recognised by her distinctive, high-pitched voice’ and ‘Cassandra Peterson’ who portrays the horror hostess character Elvira, Mistress of the Dark. Both Estelle and Cassandra have unconventional hairstyles and make up, as well as highly distinctive manners of speaking. While the task is made easier with particularly distinctive identities such as those mentioned above, we observe that accuracies are above random regardless of the identity of the speaker in the test set, suggesting that the trained model should generalise reasonably well to other speakers.



Figure 5.5: Examples of dynamic images for two identities represented using a single image per face-track (leftmost image in each row) and multiple images per track (5 images on the right). Note how lip motion has clearly been encoded in some of the frames.

5.7 Ablation Analysis

What is the best way to capture articulatory patterns? We experiment with three different methods of incorporating dynamic features in our architecture: The first computes a single dynamic image per face-track via approximate rank pooling. Since a single face track can contain a long sequence of facial dynamics which can be challenging to capture compactly, we also experiment with the Multiple Dynamic Image (MDI) formulation proposed in [Bilen et al., 2016] in which a sequence of k dynamic images are computed from sets of m contiguous frames at uniformly sampled locations. Each dynamic image is processed independently by the early part of the network and then fused later in the architecture through temporal pooling. In our experiments we take both k and m to be 10, which was found to be most effective in [Bilen et al., 2016]. Differently from the task of action recognition, where the entire video may be needed to inform some actions (e.g. “backhand flick”), we require only local descriptors of motion in order to effectively capture the ‘manner of speaking’ of a speaker. We therefore also experiment with a third “dense” dynamic image formulation sampled at a fixed stride (thus m remains fixed at 10 but k depends on the length of the face-track).

Temporal Pooling: To enable the network to ingest multiple dynamic images (MDI), we adopt a simple temporal pooling scheme: for a sequence of input frames $\mathcal{X} = x_1, \dots, x_T$, we compute a representation $\phi(\mathcal{X}) = \phi_2 \circ \text{pool}(\phi_1(x_1), \dots, \phi_1(x_T))$, where ϕ_1 comprises the set of operations up to `pool5` in the face sub-network, followed by a max-pooling over the time

dimension before the fully connected layers ϕ_2 of the network are applied.

Experiment 1: Video as a collection of static frames: Our first dynamic experiment simply treats each video as a bag of independent static frames. Since each face-track is a representation of the person speaking, some limited information, including mouth *shape* and facial contortion during speech, can be learnt from individual frames. The static architecture (Figure 5.2, left) can be used in this case, on RGB frames extracted from the video, with frames extracted in a dense manner with a stride of 6.

Experiment 2: SDI (Single Dynamic Image) per face-track: In this experiment a single dynamic image is computed to represent the entire face-track. Since there is a single image per face input, we can once again make direct use of the static architecture (Figure 5.2, left).

Experiment 3: MDI (Multiple Dynamic Images) with temporal pooling: We also experiment with multiple dynamic images in order to capture local variations in motion. In this experiment 10 uniformly sampled dynamic images are extracted per face-track and the static architecture (Figure 5.2, left) with temporal pooling is used.

Experiment 4: RGB + SDI fusion: In this experiment we use the dynamic fusion architecture (Figure 5.2, middle). A single RGB frame and dynamic image per track are fed into the network along with the audio segment.

Experiment 5: RGB + MDI fusion: We also use dense sampling in order to obtain multiple aligned RGB and dynamic images from each face-track. These inputs are then fed into the fusion architecture, and results are ensembled at test time (this is done in a similar manner to experiment 1, but over the dynamic images as well).

The results on the dynamic test set are given in table 5.3. To provide a measure of variance, we sampled 10,000 triplets from the set of all possible triplets on the test identities ten times (with replacement). We report means and std devs for the % accuracy. As expected, using multiple dynamic images instead of a single dynamic image per track leads to a substantial increase in classification accuracy, possibly due to the ability to capture local variations in motion. In order to determine whether the network is learning the face motion of the speaker (and not

Dynamic Formulation	$A_T(\text{Total})$
1. RGB	79.2 ± 0.1
2. Dynamic (SDI)	76.9 ± 0.6
3. Dynamic (MDI)	79.9 ± 0.2
4. RGB + SDI Fusion	82.4 ± 0.3
5. RGB + MDI Fusion	84.3 ± 0.2

Table 5.3: Results using different dynamic formulations for the dynamic matching F-V task; **SDI**: Single dynamic Image; **MDI**: **Multiple Dynamic Images**; The best performance is achieved using both RGB and MDI fusion.

simply relying on the RGB frame input) we also trained the network with dynamic images only (experiments 2 & 3). As seen from Figure 5.5, it is harder to discern latent variables like age, gender, ethnicity in these images, while mouth motion is clearly encoded. Using these dynamic images alone, we still achieve an accuracy of 77%, suggesting that the network may be able to exploit dynamic cross-modal biometrics.

5.8 Conclusion

In this paper, we have introduced the novel task of cross-modal matching between faces and voices, and proposed a corresponding CNN architecture to address it. Under a binary forced matching constraint, the model is able to match human performance on easy faces and exceed human performance under the more challenging setting in which the speaker pair possesses the same gender, age and nationality. The results of the experiments strongly suggest the existence of *cross-modal* biometric information, leading to the conclusion that perhaps our faces are more similar to our voices than we think.

Acknowledgements: The authors gratefully acknowledge the support of EPSRC CDT AIMS EP/L015897/1 and the Programme Grant Seebibyte EP/M013774/1. The authors would also like to thank Erika Lu for help with the AMT study, Hakan Bilen and Joe Levy for useful discussions, and Joon Son Chung for being a living legend.

Appendices

Data, models and appendices can be accessed online.⁴

Statement of Authorship

A statement of authorship for this work can be found in Appendix C.

⁴<http://www.robots.ox.ac.uk/~vgg/research/CMBiometrics>

6 | *Learnable PINs*: Cross-Modal Embeddings for Person Identity

Arsha Nagrani* Samuel Albanie* Andrew Zisserman

VGG, Oxford

*Equal Contribution

Abstract

We propose and investigate an identity sensitive joint embedding of face and voice. Such an embedding enables cross-modal retrieval from voice to face and from face to voice.

We make the following four contributions: first, we show that the embedding can be learnt from videos of talking faces, without requiring any identity labels, using a form of cross-modal self-supervision; second, we develop a curriculum learning schedule for hard negative mining targeted to this task, that is essential for learning to proceed successfully; third, we demonstrate and evaluate cross-modal retrieval for identities unseen and unheard during training over a number of scenarios and establish a benchmark for this novel task; finally, we show an application of using the joint embedding for automatically retrieving and labelling characters in TV dramas.

Published in the Proceedings of the [European Conference on Computer Vision, 2018](#).

6.1 Introduction

Face and voice recognition, both non-invasive and easily accessible biometrics, are the tools of choice for a variety of tasks. State of the art methods for face recognition use face embeddings generated by a deep convolutional neural network [Schroff et al., 2015; Taigman et al., 2014b; Parkhi et al., 2015] trained on a large-scale dataset of labelled faces [Cao et al., 2018; Kemelmacher-Shlizerman et al., 2016; Guo et al., 2016]. A similar path for generating a voice embedding is followed in the audio community for speaker recognition [Nagrani et al., 2017; McLaren et al., 2016; J. S. Chung et al., 2018; C. Zhang et al., 2018]. However, even though a person can be identified by their face or their voice, these two ‘modes’ have been treated quite independently – could they not be considered jointly?

To that end, the objective of this paper is to learn a *joint* embedding of faces and voices, and to do so using a virtually free and limitless source of unlabelled training data – videos of human speech or ‘talking faces’ – in an application of cross-modal self-supervision. The key idea is that a subnetwork for faces and a subnetwork for voice segments can be trained jointly to predict whether a face corresponds to a voice or not, and that training data for this task is freely available: the positives are faces and voice segments acquired from the same ‘talking face’ in a video, the negatives are a face and voice segment from different videos.

What is the motivation for learning such a joint embedding? First, a joint embedding of the modalities enables cross-modal retrieval – a person’s face can retrieve face-less voice segments, and their voice can retrieve still photos and speech-less video segments. Second, this may in fact be how humans internalise identity. A highly-influential cognitive model due to the psychologists Bruce and Young [Bruce & Young, 1986] proposed that ‘person identity nodes’ or ‘PINs’ are a portion of associative memory holding identity-specific semantic codes that can be accessed via the face, the voice, or other modalities: and hence are entirely abstracted from the input modality.

It is worth first considering if a joint embedding is even possible. Certainly, if we task a network with learning a joint embedding then it is likely to succeed on the training data – since arbitrary associations can be learnt even from unrelated data [C. Zhang et al., 2016]. However, if the relationship between face and voice is completely arbitrary, and the network has ‘memorised’ the training data then we would expect chance behaviour for cross-modal retrieval of identities that were *unseen and unheard* during training. It is unlikely that the relationship between face and voice is completely arbitrary, because we would expect some dependence between gender and the face/voice, and age and the face/voice [Nagrani et al., 2018b]. Somewhat surprisingly, the experiments show that employing cross-modal retrieval on the joint embeddings for unseen-unheard identities achieves matches that go beyond gender and age.

In this paper we make the following four contributions. First, in Sec. 6.3, we propose a network architecture for jointly embedding face and voice, and a training loss for learning from unlabelled videos from YouTube. Second, in Sec. 6.4, we develop a method for curriculum learning that uses a single parameter to control the difficulty of the within-batch hard negatives. Scheduling the difficulty of the negatives turns out to be a crucial factor for learning the joint embedding in an unsupervised manner. Third, in Sec. 6.7, we evaluate the learnt embedding for unseen-unheard identities over a number of scenarios. These include using the face and voice embedding for cross-modal verification, and ‘1 in N’ cross-modal retrieval where we beat the current state of the art [Nagrani et al., 2018b]. Finally, in Sec. 6.8, we show an application of the learnt embedding to one-shot learning of identities for character labelling in a TV drama. This again evaluates the embeddings on unseen-unheard identities.

6.2 Related Work

Cross-modal embeddings: The relationship between visual content and audio has been researched in several different contexts, with common applications being generation, matching

and retrieval [D. Li et al., 2003; Kidron et al., 2005; Lampert & Krömer, 2010]. The primary focus of this work, however, is to construct a shared representation, or joint embedding of the two modalities. While joint embeddings have been researched intensively for images and text, [Barnard et al., 2003; Duygulu et al., 2002; Kiros et al., 2014; L. Wang, Li, & Lazebnik, 2016; Gordo & Larlus, 2017], they have also started to gain traction for audio and vision [Ngiam et al., 2011; Srivastava & Salakhutdinov, 2012; Aytar et al., 2016; Albanie et al., 2018]. There are several ways in which this embedding may be learned—we take inspiration from a series of works that exploit audio-visual correspondence as a form of self-supervised learning [Arandjelović & Zisserman, 2017; Owens et al., 2016]. It is also possible to learn the embedding via cross-modal distillation [Aytar et al., 2016; Albanie et al., 2018; Harwath et al., 2016] in which a trained model (the teacher) transfers its knowledge in one modality to a second model (the student) in another to produce aligned representations.

Of particular relevance is recent work [Arandjelović & Zisserman, 2018] that learns a joint embedding between visual frames and sound segments for musical instruments, singing and tools. Our problem differs from theirs in that ours is one of fine grained recognition: we must learn the subtle differences between pairs of faces or pairs of voices; whereas [Arandjelović & Zisserman, 2018] must learn to distinguish between different types of instruments by their appearance and sound. We also note a further challenge; human speech exhibits considerable variability that results not only from *extrinsic* factors such as background chatter, music and reverberation, but also from *intrinsic* factors, which are variations in speech from the same speaker such as the lexical content of speech (the exact words being spoken), emotion and intonation [Nagrani et al., 2017]. A person identity-sensitive embedding must achieve invariance to both sets of factors.

Cross-modal learning with faces and voices: In biometrics, an active research area is the development of multimodal recognition systems which seek to make use of the *complementary* signal components of facial images and speech [Brunelli & Falavigna, 1995; Khoury et al., 2014], in order to achieve better performance than systems using a single modality, typically

through the use of feature fusion. In contrast to these, our goal is to exploit the *redundancy* of the signal that is common to both modalities, to facilitate the task of cross-modal retrieval. Le and Odobez [Le & Odobez, 2017] try to instill knowledge from face embeddings to improve speaker diarisation results, however their focus is only to achieve better audio embeddings.

In our earlier work [Nagrani et al., 2018b] we established, by using a forced matching task, that strong correlations exist between faces and voices belonging to the same identity. These occur as a consequence of cross-modal biometrics such as gender, age, nationality and others, which affect both facial appearance and the sound of the voice. This paper differs from [Nagrani et al., 2018b] in two key aspects. First, while [Nagrani et al., 2018b] used identity labels to train a discriminative model for matching, we approach the problem in an *unsupervised* manner, learning directly from videos without labels. Second, rather than training a model restricted to the task of matching, we instead learn a *joint* embedding between faces and voices. Unlike [Nagrani et al., 2018b], our learnt representation is no longer limited to forced matching, but can instead be used for other tasks such as cross-modal verification and retrieval.

6.3 Learning Joint Embeddings

Our objective is to learn functions $f_\theta(x_f) : \mathbb{R}^F \rightarrow \mathbb{R}^E$ and $g_\phi(x_v) : \mathbb{R}^V \rightarrow \mathbb{R}^E$ which map faces and voices of the same identity in \mathbb{R}^F and \mathbb{R}^V respectively onto nearby points in a shared coordinate space \mathbb{R}^E . To this end, we instantiate $f_\theta(x_f)$ and $g_\phi(x_v)$ as convolutional neural networks and combine them to form a two-stream architecture comprising a face subnetwork and a voice subnetwork (see Fig. 6.1). To learn the parameters of f_θ and g_ϕ , we sample a set \mathcal{P} of training pairs $\{x_f, x_v\}$, each consisting of a face image x_f and a speech segment x_v and attach to each pair an associated label $y \in \{0, 1\}$, where $y = 0$ if x_f and x_v belong to different identities (henceforth a negative pair) and $y = 1$ if both belong to the same identity (a positive pair). We employ a contrastive loss [Chopra et al., 2005; Hadsell et al., 2006] on the paired data $\{(x_{f_i}, x_{v_j}, y_{i,j})\}$, which seeks to optimise f_θ and g_ϕ to minimise the distance between the

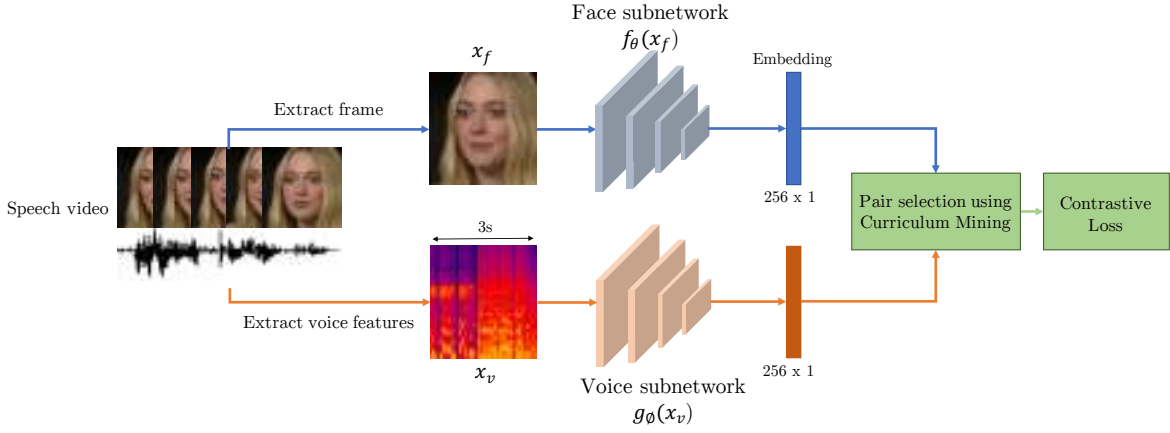


Figure 6.1: Learning a joint embedding between faces and voices. Positive face-voice pairs are extracted from speech videos and fed into a two-stream architecture with a face subnetwork $f_\theta(x_f)$ and a voice subnetwork $g_\phi(x_v)$, each producing 256-D embeddings. A curriculum-based mining schedule is used to select appropriate negative pairs which are then trained using a contrastive loss.

embeddings of positive pairs and penalises the negative pair distances for being smaller than a margin parameter α . Concretely, the cost function is defined as:

$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} y_{i,j} D_{i,j}^2 + (1 - y_{i,j}) \max\{0, \alpha - D_{i,j}\}^2 \quad (6.1)$$

where $(i, j) \in \mathcal{P}$ is used to indicate $(x_{f_i}, x_{v_j}, y_{i,j}) \in \mathcal{P}$ and $D_{i,j}$ denotes the Euclidean distance between normalised embeddings, $D_{i,j} = \left\| \frac{f_\theta(x_{f_i})}{\|f_\theta(x_{f_i})\|_2} - \frac{g_\phi(x_{v_j})}{\|g_\phi(x_{v_j})\|_2} \right\|_2$. Details of the architectures for each subnetwork are provided in Sec. 6.6.1.

6.3.1 Generating face-voice pairs

Obtaining speaking face tracks: In contrast to previous audio-visual self-supervised works that seek to exploit naturally synchronised data [Arandjelović & Zisserman, 2017; Aytar et al., 2016], simply extracting audio and video frames at the same time is not sufficient to obtain pairs of faces and voice samples (of the same identity) required to train the contrastive loss described in Eqn. 6.1. Even for a given video tagged as content that may contain a

talking human, a short sample from the associated audio may not contain any speech, and in cases when speech is present, there is no guarantee that the speaker of the audio is visible in the frame (e.g. in the case of ‘reaction shots’, flashbacks and dubbing of videos [Nagrani & Zisserman, 2017]). Furthermore, even when the face of the speaker is present there may be more than one face occupying the frame.

We address these issues by using SyncNet [J. S. Chung & Zisserman, 2016b], an unsupervised method that obtains speaking face-tracks from video automatically. SyncNet consists of a two-stream convolutional neural network which estimates the correlation between the audio track and the mouth motion of the video. This allows the video to be accurately segmented into *speaking face-tracks*—contiguous groupings of face detections from the video of the *speaker* (described in more detail in Sec. 6.5).

Selecting face-voice pairs: Given a collection of speaking face-tracks, we can then construct a collection of labelled training pairs with the following simple labelling algorithm. We define face and voice segments extracted from the *same* face-track as *positive pairs* and define face and voice segments extracted from *different* face-tracks as *negative pairs* (this approach was also taken for single modality in [Cinbis et al., 2011]).

Since our objective is to learn embeddings that place identities together, rather than capturing synchronous, intrinsic factors (such as emotion expressions, or lexical content), we do not constrain the face associated with a positive pair to be temporally aligned with the audio. Instead it is sampled uniformly from the speaking face-track, preventing the model from learning to use synchronous clues to align the embeddings (see Fig. 6.2). We next describe the procedure for pair selection during training.

6.4 The Importance of Curriculum-based Mining

One of the key challenges associated with learning embeddings via contrastive losses is that as the dataset gets larger the number of possible pairs grows quadratically. In such a scenario,

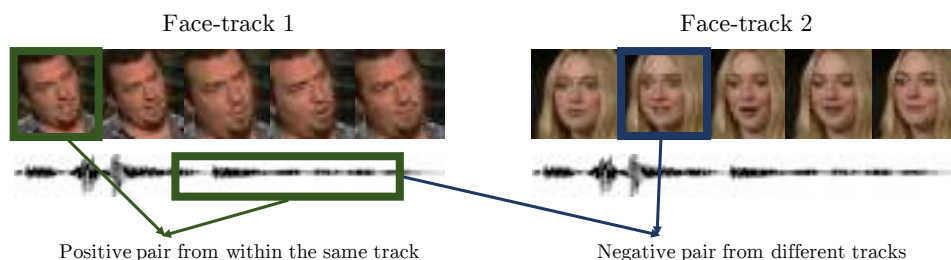


Figure 6.2: Generating positive and negative face/voice pairs (Sec. 6.3.1). To prevent the embeddings from learning to encode synchronous nuisance factors, the frame for the positive face is not temporally aligned with the sequence for the voice.

the network rapidly learns to correctly map the easy examples, but hard positive and negative mining [Sung, 1996; Hermans et al., 2017; X. Wang & Gupta, 2015; J. S. Chung & Zisserman, 2016b; H. O. Song et al., 2016] is often required to improve performance further. In the context of our task, a neural network of sufficient capacity quickly learns to embed faces and voices of differing genders far apart—samples from different genders then become “easy” negative pairs. Since gender forms only one of the many components that make up identity, we would like to ensure that the embeddings also learn to encode other factors. However, as we do not know the identities of the speaker face-tracks a priori, we cannot enforce sampling of gender-matched negative pairs. We tackle this issue with a hard negative mining approach that does not require knowledge of the identities during training.

When used in the unsupervised setting, hard negative selection is a somewhat delicate process, particularly when networks are trained from scratch. If the negative samples are too hard, the network will focus disproportionately on outliers, and may struggle to learn a meaningful embedding. In our setting, the hardest negatives are particularly dangerous, since they may in fact correspond to false negative labels (in which a voice and a face of the *same* identity has been sampled by chance from different speaking face-tracks)¹.

¹For a given face image and voice sampled from different speaking face-tracks, the false negative rate of the labelling diminishes as the number of identities represented in the videos grows.

6.4.1 Controlling the difficulty of mined negatives

Standard online hard example mining (OHEM) techniques [Shrivastava et al., 2016; Hermans et al., 2017] sample the hardest positive and negative pairs within a minibatch. However, in our setting hard positive mining may be of limited value since we do not expect the video data to exhibit significant variability within speaking face-tracks. If the hardest negative example within each mini-batch is selected, training with large batches leads to an increased risk of outliers or false negatives (i.e. pairs labelled as negatives which are actually positives), both of which will lead to poor learning dynamics. We therefore devise a simple curriculum-based mining system, which we describe next. Each mini-batch comprises K randomly sampled face-tracks. For each face-track we construct a positive pair by uniformly sampling a single frame x_f , and uniformly sampling a three second audio segment x_v . This sampling procedure can be viewed as a form of simple data augmentation and makes good use of the available data, producing a set of K positive face-voice pairs. Next, we treat each face input x_f among the pairs as an *anchor face* and select an *appropriately hard* negative sample from within the mini-batch. This is achieved by computing the distances between its corresponding face embedding and all voice embeddings with the exception of its directly paired voice, leading to a total of $K - 1$ potential negatives. The potential negatives are then ranked in descending order based on their distance to the anchor face (with the last element being the hardest negative in the batch), and the appropriate negative is chosen according to a ‘negative difficulty parameter’ τ . This parameter simply corresponds to the percentile of the ranked negatives: $\tau = 1$ is the hardest negative, $\tau = 0.5$ the median, and $\tau = 0$ the easiest. This parameter τ can be tuned just like a learning rate. In practice, we found that a schedule that selects easier negatives during early epochs of training, and harder negatives for later epochs to be particularly effective². While

²It is difficult to tune this parameter based on the loss alone, since a stagnating loss curve is not necessarily indicative of a lack of progress. As the network improves its performance at a certain difficulty, it will be presented with more difficult pairs and continue to incur a high loss. Hence we observe the mean distance between positive pairs in a batch, mean distance between negative pairs in the batch, and mean distance between *active* pairs (those that contribute to the loss term) in the batch, and found that it was effective to increase τ by 10 percent every two epochs, starting from 30% up until 80%, and keeping it constant thereafter.

selecting the appropriate negative, we also ensure that the distance between the anchor face to the threshold negative is larger than the distance between the anchor face and the positive face, (following the semi-hard negative mining procedure outlined in [Schroff et al., 2015]). Pseudocode for the mining procedure and an ablation analysis on the curriculum mining is provided in the longer version of this paper (accessible online³).

6.5 Dataset

We learn the joint face-voice embeddings on VoxCeleb [Nagrani et al., 2017], a large-scale dataset of audio-visual human speech video extracted ‘in the wild’ from YouTube. The dataset contains over 100,000 segmented *speaking face-tracks* obtained using SyncNet [J. S. Chung & Zisserman, 2016b] from over 20k challenging videos. The speech audio is naturally degraded with background noise, laughter, and varying room acoustics, while the face images span a range of lighting conditions, image quality and pose variations (see Fig. 6.5 for examples of face images present in the dataset). VoxCeleb also contains labels for the identities of the celebrities, which, we stress, are not used while learning the joint embeddings. We make use of the labels only for the purposes of analysing the learned representations – they allow us to evaluate their properties numerically and visualise their structure (e.g. Fig. 6.4). We use two train/test splits for the purpose of this task. The first split is provided with the dataset, and consists of disjoint videos from the same set of speakers. This can be used to evaluate data from identities seen and heard during training. We also create a second split which consists of 100 randomly selected disjoint identities for validation, and 250 disjoint identities for testing. We train the model using the intersection of the two training sets, allowing us to evaluate on both test sets, the first one for seen-heard identities, and the second for unseen-unheard identities. The statistics of the dataset are given in Table 6.1.

³<http://www.robots.ox.ac.uk/~vgg/research/LearnablePins/>

	Train	Test(S-H)	Val(US-UH)	Test(US-UH)
# speaking face-tracks	105,751	4,505	12,734	30,496
# identities	901	901	100	250

Table 6.1: Dataset statistics. Note the identity labels are not used at any point during training. SH: Seen-heard. US-UH: Unseen-unheard. The identities in the unseen-unheard test set are disjoint from those in the train set.

6.6 Experiments

We experiment with two initialisation techniques, training from scratch (where the parameters for both subnetworks are initialised randomly) and using pretrained subnetworks. In the latter formulation, both the subnetworks are initialised using weights trained for identification within a single modality. We also experiment with a teacher-student style architecture, where the face subnetwork is initialised with pretrained weights which are frozen during training (teacher) and the voice subnetwork is trained from scratch (student), however we found that this leads to a drop in performance. We use weights pretrained for identity on the VGG-face dataset for the face subnetwork, and weights pretrained for speaker identification on the VoxCeleb dataset for the voice subnetwork.

6.6.1 Network architectures and implementation details

Face subnetwork: The face subnetwork is implemented using the VGG-M [Chatfield et al., 2011] architecture, with batch norm layers [Ioffe & Szegedy, 2015] added after every convolutional layer. The input to the face subnetwork is an RGB image, cropped from the source frame to include only the face region and resized to 224×224 . The images are augmented using random horizontal flipping, brightness and saturation jittering, but we do not extract random crops from within the face region. The final fully connected layer of the VGG-M architecture is reduced to produce a single 256-D embedding for every face input. The embeddings are then L2-normalised before being passed into the pair selection layer for negative

mining (Sec. 6.4).

Voice subnetwork: The audio subnetwork is implemented using the VGG-Vox architecture [Nagrani et al., 2017], which is a modified version of VGG-M suitable for speaker recognition, also incorporating batch norm. The input is a short-term amplitude spectrogram, extracted from three seconds of raw audio using a 512-point FFT (following the approach in [Nagrani et al., 2017]), giving spectrograms of size 512×300 . At train-time, the three second segment of audio is chosen randomly from the entire audio segment. Mean and variance normalisation is performed on every frequency bin of the spectrogram. Similarly to the face subnetwork, the dimensionality of the final fully connected layer is reduced to 256, and the 256-D voice embeddings are L2-normalised. At test time, the entire audio segment is evaluated using average pooling in an identical manner to [Nagrani et al., 2017].

The lightweight VGG-M inspired architectures described above have the benefit of computational efficiency and in practice we found that they performed reasonably well for our task. We note that either subnetwork could be replaced with a more computationally intensive trunk architecture without modification to our method.

Training procedure: The networks are trained on three Titan X GPUs for 50 epochs using a batch-size of 256. We use SGD with momentum (0.9), weight decay ($5E - 4$) and a logarithmically decaying learning rate (initialised to 10^{-2} and decaying to 10^{-8}). We experimented with different values of the margin for the contrastive loss (0.2,0.4,0.6,0.8) and found that a margin of 0.6 was optimal.

6.7 Evaluation

6.7.1 Cross-modal Verification

We evaluate our network on the task of *cross-modal verification*, the objective of which is to determine whether two inputs from different modalities are semantically aligned. More specifically, given a face input and a speech segment, the goal is to determine if they belong to the

	AUC %	EER %
Seen-Heard		
Random	50.3	49.8
Scratch	73.8	34.1
Pretrained	87.0	21.4
Unseen-Unheard		
Random	50.1	49.9
Scratch	63.5	39.2
Pretrained	78.5	29.6

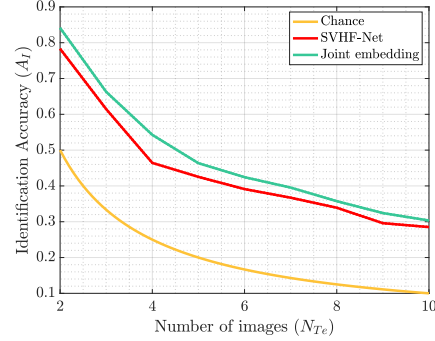


Table 6.2: **Cross-modal Verification:** Results are reported for an untrained model (with random weights), as well as for the two initialisations described in Sec. 6.6.

Figure 6.3: **N-way forced matching:** We compare our joint embedding to SVHF-Net [Nagrani et al., 2018b]. Our method comfortably beats the current state of the art for all values of N .

same identity. Since there are no available benchmarks for this task, we create two evaluation protocols for the VoxCeleb dataset, one for *seen-heard* identities and one for *unseen-unheard* identities. For each evaluation benchmark test pairs are randomly sampled, 30,496 pairs from unseen-unheard identities and 18,020 pairs from seen-heard identities (a description of the evaluation protocol is in Appendix C) using the identity labels provided by VoxCeleb: positives are faces and voices of the same identity, and negative pairs are from differing identities.

The results for cross-modal verification are reported in Table 6.2. We use standard metrics for verification, i.e area under the ROC curve (AUC) and equal error rate (EER). As can be seen from the table, the model learned from scratch performs significantly above random, even for unseen-unheard identities, providing evidence to support the hypothesis that it is, in fact, possible to learn a joint embedding for faces and voices with no explicit identity supervision. A visualisation of the embeddings is provided in Fig. 6.4, where we observe that the embeddings form loose groups of clusters based on identity. Initialising the model with two pretrained subnetworks brings expected performance gains and also performs surprisingly

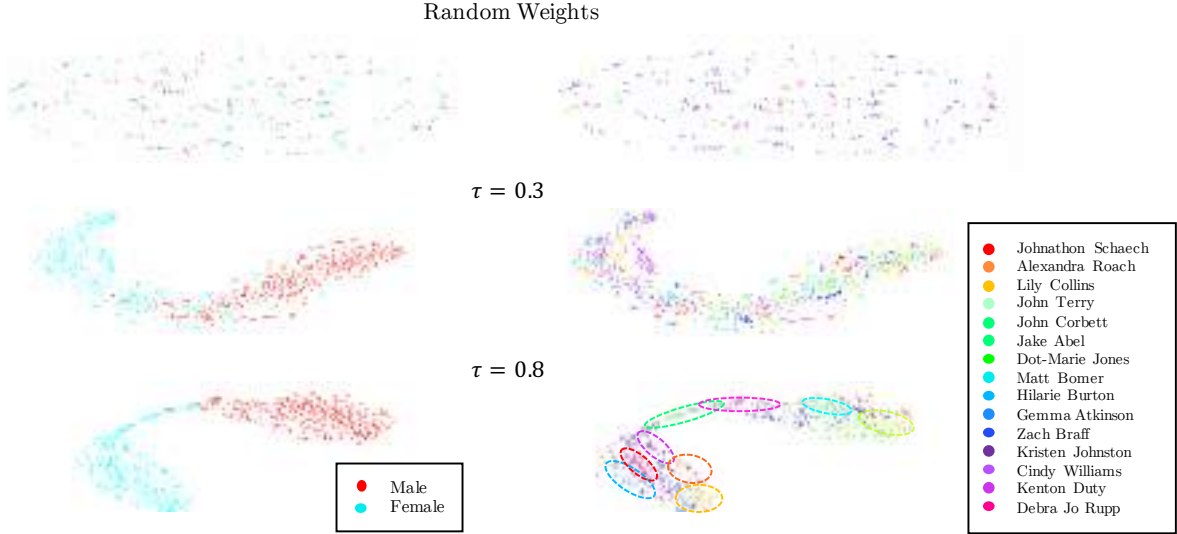


Figure 6.4: t-SNE [Maaten & Hinton, 2008] visualisation of learnt embeddings for *faces only* from 15 identities from the VoxCeleb test set. The model is trained entirely from scratch. For visualisation purposes, embeddings are coloured with (left) gender labels and (right) identity labels (no labels were used during training). The embeddings are shown for three stages, from top to bottom; a non-trained network (random weights), a model trained with $\tau = 0.3$ and the final model trained using our curriculum learning schedule, with τ increasing from 0.3 till 0.8. Best viewed in colour.

well for unseen-unheard identities, a task that is even difficult for humans to perform. Previous work has shown that on the less challenging forced matching task (selecting from two faces given a voice), human performance is around 80% [Nagrani et al., 2018b].

Effect of cross-modal biometrics: In this section we examine the effect of specific latent properties (age, gender and nationality) which influence both face and voice. We evaluate the model by sampling negative test pairs while holding constant each of the following demographic criteria: gender (G), nationality (N) and age (A). Gender and nationality labels are obtained from Wikipedia. Since the age of a speaker could vary over different videos, we apply an age classifier [Rothe et al., 2018] to the face frames (extracted at 1fps) and average the age predictions over each video.

We find that gender is the most influential demographic factor. Studies in biology and evolutionary perception [Wells et al., 2013; Thornhill & Møller, 1997] also show that other more

demographic criteria	random	G	N	A	GNA
unseen-unheard (AUC %)	78.5	61.1	77.2	74.9	58.8
seen-heard (AUC %)	87.0	74.2	85.9	86.6	74.0

Table 6.3: **Analysis of cross-modal biometrics under varying demographics:** Results are reported for both seen-heard and unseen-unheard identities using AUC: Area Under Curve. Chance performance is 50%.

subtle factors such as hormone levels during puberty affect both face morphology and voice pitch, eg. lower voice pitch correlating with a stronger jawline. However since these factors are harder to quantify, we leave this analysis for future work.

Searching for shortcuts (bias): As a consequence of their high modelling capacity, CNNs are notorious for learning to exploit biases that enables them to minimise the learning objective with trivial solutions (see [Doersch et al., 2015] for an interesting discussion in the context of unsupervised learning). While we are careful to avoid correlations due to lexical content and emotion, there may be other low level correlations in the audio and video data that the network has learned to exploit. To probe the learned models for bias, we construct two additional evaluation sets. In both sets, negative pairs are selected following the same strategy as for the original evaluation set (they are faces and voices of different identities). However, we now sample positives pairs for the bias evaluation test sets as follows. For the first test set we sample positive pairs from the *same speaking face-track*, as opposed to sampling pairs from the same identity across all videos and speaking face-tracks (as done in our original evaluation set), and for the second test set we sample positive pairs from the *same video*. We then evaluate the performance of the model trained from scratch on the task of cross-modal verification. We obtain results that are slightly better when positive pairs are always from the same video (AUC: 74.5, EER: 33.8) vs (AUC:73.8, EER: 34.1, Table 6.2) on the original test set, but with minimal further improvement when they are constrained to belong to the same track (AUC: 74.6, EER: 33.6). This suggests that audio and faces taken from the same video have small additional correlations beyond possessing the same identity which the network has learned to exploit. For example, it is likely that blurry low quality videos are often accompanied by low

quality audio, and that faces from professionally shot studio interviews often occur with high quality audio. While these signals are unavoidable artefacts of working with datasets collected ‘in the wild’, the difference in performance is slight, providing some measure of confidence that the network is relying primarily on identity to solve the task.

6.7.2 Cross-modal Retrieval with varying gallery size

The learned joint embedding also enables cross-modal retrieval. Given a single query from one modality, the goal is to retrieve all semantically matching templates from another modality (here the set of all possible templates is referred to as the *gallery set*). This can be done for both the F-V formulation (using a face to retrieve voices of the same identity) and the V-F formulation (using a voice segment to retrieve matching faces). Since there are limited baselines available for this task, we instead perform a variant of cross-modal retrieval to allow us to compare with previous work [Nagrani et al., 2018b] (which we refer to as SVHF-Net), which represents the current state of the art for matching faces and voices. In [Nagrani et al., 2018b], a forced matching task is used to select the *single* semantically matching template from N options in another modality, and the SVHF-Net is trained directly to perform this task. Unlike this work where we learn a joint embedding, SVHF-Net consists of a concatenation layer which allows comparison of the two modalities, i.e. learnt representations in each modality are not aligned. In order to compare our method to SVHF-Net, a query set is made using all the available test samples in a particular modality. For example for the V-F formulation (used in [Nagrani et al., 2018b]), we use all the voice segments in our unseen-unheard test set. A gallery of size N is then created for each query – a gallery consists of a single positive face and $N-1$ negative faces from different identities. We adopt a simple method to perform the task: the query embedding is compared directly to the embeddings of all the faces in the gallery using the Euclidean distance, and the closest embedding is chosen as the retrieved result. We compare to SVHF-Net directly on our test set, for values $N = 2$ to 10 . A comparison of the results is given in Fig. 6.3.



Figure 6.5: Qualitative results for cross-modal *forced matching* (selecting the matching template from N samples). We show results for $N = 10$. A query sample from one modality is shown on the left, and 10 templates from the other modality are shown on the right. For each formulation, we show four successful predictions, with the matching template highlighted in green (top four rows in each set) and one failure case (bottom row in each set) with the ground truth highlighted in green and the model prediction in red. Best viewed zoomed in and in colour.

We observe that learning a joint embedding and using this embedding directly to match faces and voices, outperforms previous work [Nagrani et al., 2018b] for all values of N . In addition, note that in contrast to the SVHF-Net [Nagrani et al., 2018b] which cannot be used if there is more than one matching sample in the gallery set, our joint embedding can be used directly to provide a ranking. In addition to the numerical results for the V-F formulation (this is the formulation used by [Nagrani et al., 2018b]) we present qualitative results for both the V-F and face to voice (F-V) formulations in Fig. 6.5.

6.8 One-Shot Learning for TV Show Character Retrieval

One shot retrieval in TV shows is the extremely challenging task of recognising all appearances of a character in a TV show or feature film, with only a single face image as a query. This is difficult because of the significant visual variation of character appearances in a TV show caused by pose, illumination, size, expression and occlusion, which can often exceed those due to identity. Recently there has been a growing interest in the use of the audio-

track to aid identification [Nagrani & Zisserman, 2017; Budnik et al., 2014; Tapaswi et al., 2012] which comes for free with multimedia videos. However, because face and voice representations are usually not aligned, in prior work the query face cannot be directly compared to the audio track, necessitating the use of complex fusion systems to combine information from both modalities. For example, [Budnik et al., 2014] use clustering on face-tracks and diarised speaker segments after a round of human annotation for both, [Nagrani & Zisserman, 2017] use confidence labels from one modality to provide supervision for the other modality, and [Tapaswi et al., 2012] fuse the outputs of a face recognition model, and a clothing model, with a GMM-based speaker model. With a joint embedding, however, the query face image can be compared directly to the audio track, leading to an extremely simple solution which we describe below.

Method: For this evaluation, we use the tracks and labels provided by [Nagrani & Zisserman, 2017] for episode 1 of the TV series ‘Sherlock’. In order to demonstrate the effectiveness of using voice information as well, we use only the 336 speaking face-tracks from the episode, which are often the most difficult to classify visually due to large variations in head pose (it is extremely rare for the speaker to look directly at the camera during a conversation). We demonstrate our method on the retrieval of the two most frequently appearing characters, Sherlock and John, from among all the other 17 classes in the episode (16 principal characters and a single class for all the background characters).

A single query face is selected randomly for Sherlock and for John, and an embedding computed for the query using our face representation. Each face-track from the set of total tracks is then split into frames, and embeddings for each face detection are computed using our learned face representation, giving a 256-D vector for each face. The vectors are then averaged over all frames, leading to a single 256-D embedding for every track. Audio segments are also extracted for each track, and an embedding computed using our learned voice representation, giving a 256-D vector for each track in a similar fashion.

Because our representations are aligned, for each track, we can compare both the visual track

	Sherlock (AUC %)	John (AUC %)
Face only	35.0	44.6
Voice only	28.7	37.2
Max Fusion	37.5	45.4

Table 6.4: **One-shot retrieval results:** Retrieval from amongst 17 categories, 16 principal characters and 1 class for all the background characters. A higher AUC is better.

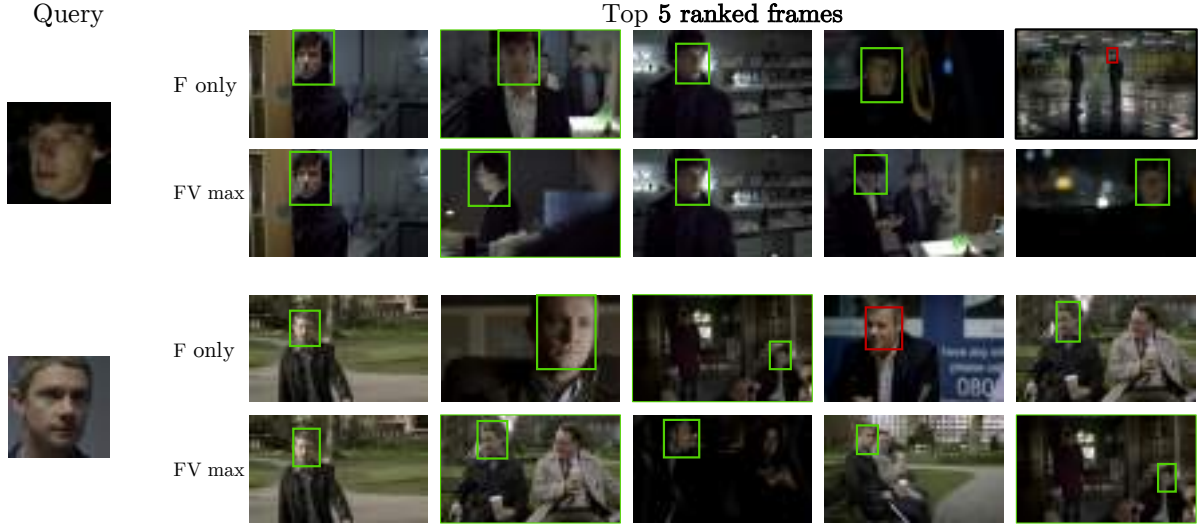


Figure 6.6: Results of one-shot retrieval for speaking face-tracks from the TV series ‘Sherlock’. A single query image and the top 5 retrieved results are shown. For each query we show tracks retrieved using only the face embeddings of the tracks (F only), and using both the face and voice embeddings (FV max). The middle frame of each retrieved track is shown. Note how FV fusion allows more profile faces to be retrieved – row 2, second and fourth frames, and row 4, third ranked frame. Face detections are green for correctly retrieved faces and red otherwise. Best viewed in colour.

and the audio track embeddings directly to the features of the query image, using L2 Euclidean distance. The tracks are then ranked according to this final score. We report results for 3 cases, retrieval using visual embeddings alone, retrieval using audio embeddings alone, and a simple fusion method where we take the maximum score out of the two (i.e. we pick the score of the modality that is closest in distance to the query image). Note, none of the identities in the episode are in the VoxCeleb training set, this test is for unseen-unheard identities. As can be seen from Table 6.4, using information from both modalities provides a slight improvement over using face or speech alone. Such a fusion method is useful for cases when one modality

is a far stronger cue, e.g. when the face is too small or dark, or for extreme poses where the voice can still be clear [Nagrani & Zisserman, 2017]. On the other hand facial appearance scores can be higher when voice segments are corrupted with crosstalk, background effects, music, laughter, or other noise. We note that a superior fusion strategy could be applied in order to better exploit this complementary information from both modalities (e.g. an attention based strategy) and we leave this for future work.

6.9 Conclusion

We have demonstrated the somewhat counter-intuitive result – that face and voice can be jointly embedded and enable cross-modal retrieval for unseen and unheard identities. We have also shown an application of this joint embedding to character retrieval in TV shows. Other possible applications include biometric security, for example a face in video footage can be directly compared to an existing dataset which is in another modality, e.g. a scenario where only voice data is stored because it was obtained from telephone conversations. The joint embedding could also be used to check whether the face in a video actually matches the voice, as part of a system to detect tampering (e.g. detecting ‘Deepfakes’ [H. Kim et al., 2018]).

Identity is more than just the face. Besides voice, identity is also in a person’s gait, the way the face moves when speaking (a preliminary exploration is provided in the online Appendix), the way expressions form, etc. So, this work can be extended to include more cues – in accord with the original abstraction of a PIN.

Acknowledgements. The authors gratefully acknowledge the support of EPSRC CDT AIMS grant EP/L015897/1 and the Programme Grant Seebibyte EP/M013774/1. The authors would also like to thank Judith Albanie for helpful suggestions.

Appendices

Appendices can be accessed online.⁴

Statement of Authorship

A statement of authorship for this work can be found in Appendix [C](#).

⁴[urlhttp://www.robots.ox.ac.uk/vgg/research/LearnablePins/](http://www.robots.ox.ac.uk/vgg/research/LearnablePins/)

7 | Disentangled speech embeddings using cross-modal self-supervision

Arsha Nagrani^{1*}, Joon Son Chung^{1,2*}, Samuel Albanie^{1*}, Andrew Zisserman¹

¹VGG, Oxford, ²Naver Corporation

*Equal Contribution

Abstract

The objective of this paper is to learn representations of speaker identity without access to manually annotated data. To do so, we develop a self-supervised learning objective that exploits the natural cross-modal synchrony between faces and audio in video. The key idea behind our approach is to tease apart—without annotation—the representations of linguistic content and speaker identity. We construct a two-stream architecture which: (1) shares low-level features common to both representations; and (2) provides a natural mechanism for explicitly disentangling these factors, offering the potential for greater generalisation to novel combinations of content and identity and ultimately producing speaker identity representations that are more robust.

Published in the Proceedings of the [International Conference on Acoustics Speech and Signal Processing, 2020](#).

7.1 Introduction

The coupling of deep neural networks with large-scale labelled training datasets has produced a number of notable successes, yielding improved performance in speech related tasks such as ASR [Chiu et al., 2018] and speaker verification [Torfi et al., 2018; W. Xie et al., 2019]. However, the considerable cost of manually producing such labels ultimately limits the potential of fully supervised approaches. By contrast, methods which are able to learn effective representations from data with few labelled examples can in principle benefit from the ever-increasing quantity of existing unlabelled speech data.

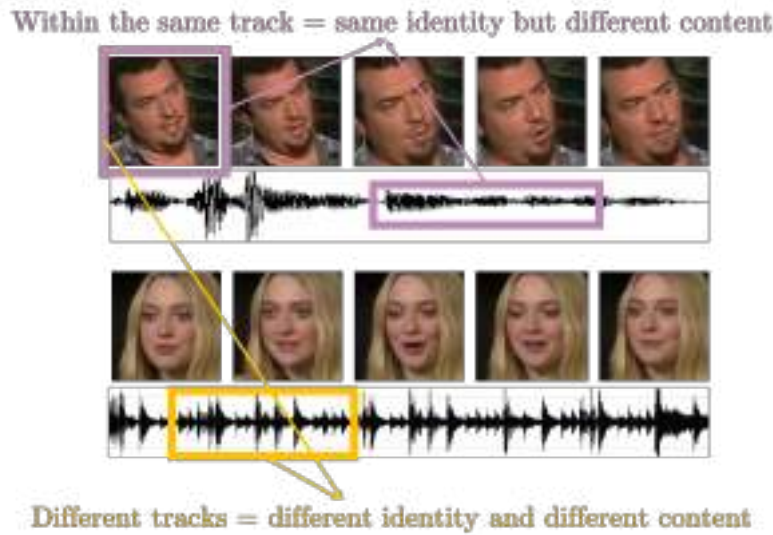


Figure 7.1: To learn representations for speaker recognition without labels, our method relies on two hypotheses: (1) face and voice samples within a single face-track are likely to share a common identity, but different linguistic content across time; (2) face and voice samples from different face-tracks are likely to have both different speaker identities and different linguistic content.

The objective of this paper is to develop one such method for learning compact and robust representations of speaker identity without supervision. Ultimately, these representations can then be used for a number of downstream tasks such as speaker recognition, clustering and diarisation etc. To achieve this goal, we propose to exploit the natural synchrony between

faces and audio in audio-visual video data as a supervisory signal, removing the need for speaker annotation. To facilitate our method, we assume access to a large-scale collection of unlabelled speaking face-tracks [Nagrani et al., 2017], which can be readily obtained through self-supervised techniques for active speaker detection [J. S. Chung & Zisserman, 2016b]. Beyond access to this data, our approach makes use of two weak statistical cues to define a self-supervised learning objective (Fig. 7.1): we assume that faces and voices extracted within a face-track at small offsets are likely to have the same speaker identity but different linguistic content, while faces and voices from different face-tracks are likely to differ in both content and speaker identity. As we show in Sec. 7.3, these cues can be combined to learn representations of speaker identity which minimise their dependence on speaker content. The motivation for doing so is simple: unlike earlier datasets such as TIMIT [Garofolo et al., 1993] that are carefully balanced for phonetic and dialectal coverage, more modern (and larger) datasets created from uncontrolled speech ‘in the wild’ are likely to contain a strong correlation between identity and linguistic content. For example, VoxCeleb2 [J. S. Chung et al., 2018] consists of interviews of famous celebrities from a wide variety of professions, whose speech can be closely tied to their occupation—the cricketer Adam Gilchrist says the word ‘*cricket*’ 17 times and ‘*president*’ 0 times; whereas the politician Nancy Pelosi says the word ‘*president*’ 88 times, ‘*Democrats*’ 19 times and ‘*cricket*’ 0 times. Consequently, a model trained to represent identity may be incentivised to use linguistic content as a discriminative cue. While some coupling between content and identity is natural, over-reliance on content can prevent generalisation to new settings, harming robustness. More broadly, disentangled representations can, in principle, achieve an exponential improvement in generalisation efficiency over their entangled counterparts, because they are able to represent novel combinations of factors that were encountered separately (but never in combination) during training. In this work, we make the following contributions: (1) We propose a novel framework for learning speech representations capturing information at different time scales in the speech signal, including in particular the identity of the speaker; (2) we show that we can learn these representations

from a large, unlabelled collection of talking faces in videos as a source of free supervision, without the need for any manual annotation; (3) we show that sharing a trunk architecture for two different tasks (content and speaker identity) and adding an explicit disentanglement objective between the two improves performance; and, (4) we evaluate the performance of our self-supervised embeddings on the popular VoxCeleb1 speaker recognition benchmark and compare to fully supervised methods. All data and models will be released.

7.2 Related Work

Representation Learning. The ability to represent variable-length high-dimensional audio segments using compact, fixed-length representations has proven useful for many speech applications such as speaker verification [W. Xie et al., 2019; J. S. Chung et al., 2018], audio emotion classification [Albanie et al., 2018], and spoken term detection (STD) [Miller et al., 2007], where the representation can be coupled with a standard classifier. The use of fixed-length representations also enables efficient storage and retrieval when paired with an inverted index. These representations can either be hand-crafted, such as MFCCs or learned from data - such as i-vectors and deep neural networks. While the former may fail to capture the correct underlying factors for the task, the latter require large amounts of expensively labeled training data to be effective. As a consequence, there has recently been renewed interest in learning unsupervised audio representations [Y.-A. Chung et al., 2016].

Disentangled Representation Learning. Motivated by their attractive compositional properties and theoretical ability to generalise efficiently, a number of models that seek to learn disentangled representations in a weakly supervised or self-supervised manner have been proposed, such as DC-IGN [T. D. Kulkarni et al., 2015], InfoGAN [X. Chen et al., 2016] and VQ-VAE [van den Oord et al., 2017]. Due to the proliferation of video data, there has also been a renewed interest in learning representations from sequential data [Fabius & van Amersfoort, 2014; J. Chung et al., 2015; Fraccaro et al., 2016; J. Chung et al., 2016]. These self-

supervised works focus on predicting future, missing or contextual information, all within the same modality. However to the best of our knowledge, no prior method has sought to learn disentangled representations through cross-modal self-supervision.

Audio-Visual Self Supervision. A number of recent works [Aytar et al., 2016; Arandjelović & Zisserman, 2017; Aytar, Vondrick, & Torralba, 2017; Owens et al., 2016; Korbar et al., 2018] have explored the concept of exploiting the correspondence between synchronous audio and visual data in teacher-student style architectures (where the ‘teacher’ is represented by a pretrained network) [Aytar, Vondrick, & Torralba, 2017], or two-stream networks where both networks are trained from scratch [Arandjelović & Zisserman, 2017; J. S. Chung & Zisserman, 2016b]. Additional work has examined cross-modal relationships between faces and voices specifically in order to learn identity [Nagrani et al., 2018b; C. Kim et al., 2018; Nagrani et al., 2018a] or emotion [Albanie et al., 2018] representations. In contrast to these works, we aim to learn representations of both content and identity with a view to explicitly disentangling separate factors—we compare our approach with theirs in Sec. 7.4.

7.3 Model

Speech, like many sequential natural signals, can be decomposed into the interaction of several largely-independent causal factors which express themselves over different time scales. The central observation that underpins our approach is that the speaker identity affects fundamental frequency, pitch and volume at the utterance level while linguistic content affects spectral contour and duration of formants more locally.

Without labels, we have no way to directly separate these factors. Instead, we can impose our prior knowledge as to how such representations should behave. Intuitively, representations of identity should change *slowly* over time (remaining constant for a given speaker), whereas representations of content should change *quickly*, capturing the local variation in the speech

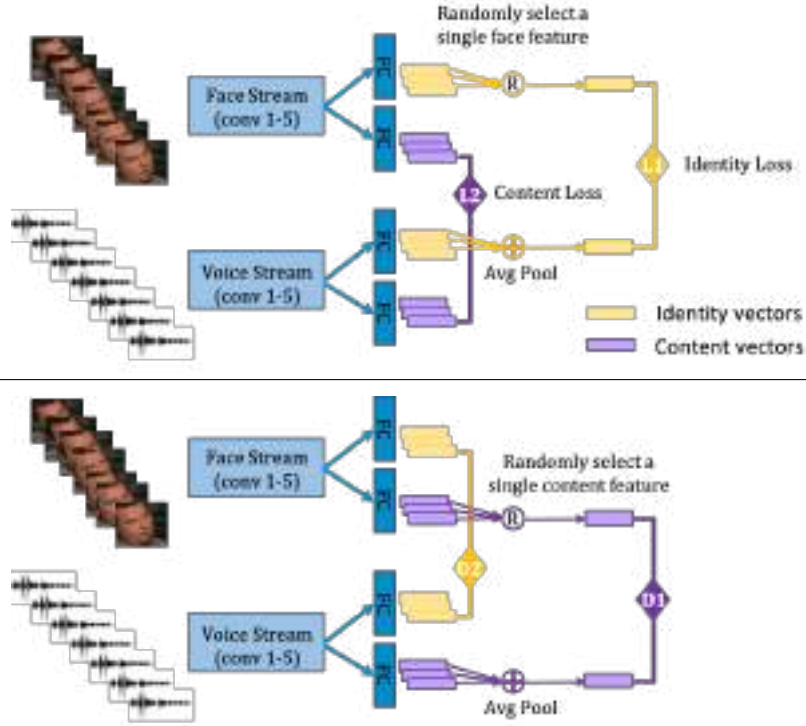


Figure 7.2: The framework for learning representations of speaker identity. We aim to explicitly disentangle speech representations into content and identity embeddings. Under the notation of Sec. 7.3, we show $B = 1$ facetracks in the input and $N = 3$ samples per track. Top diagram: the identity (L1) and content classification losses (L2); bottom diagram: the disentanglement losses (D1 and D2).

signal. Concretely, we enforce these properties by exploiting the known correspondence between a speech signal and the face of its speaker within a facetrack to impose three constraints on the representations for content and identity:

Content constraints. Within a given speaking facetrack, speech and face signals extracted concurrently contain redundant (or overlapping) linguistic content (while this information is trivially available in the speech signal, it is perhaps less obvious that it is also present in the face—in fact, it is this signal that enables lipreading). By contrast, the face signal at a small temporal offset from the speech signal is likely to convey *different* linguistic content. These cues provide a natural source of paired data (positive and negative examples) that we can use to learn a self-supervised representation of linguistic content from a speech signal [J. S. Chung

& Zisserman, 2016b].

Identity constraints. By considering instead face and voice signals across face tracks, we can obtain a different form of constraint: signals from the same face-track should come from the same speaker, while those from different face-tracks are likely to come from different speakers. This idea was demonstrated in [Nagrani et al., 2018a].

Disentangling constraint. Although representations that have been trained to satisfy the intra-track and inter-track constraints may capture a measure of both linguistic content and speaker identity, there is no guarantee that both factors will be disentangled (represented independently of one-another). To achieve this last goal, we employ a further constraint on the speech representations themselves, requiring that variation within one factor cannot be predicted from the other to enforce their independence.

Learning framework. In this work, we train a single model in an end-to-end self-supervised manner to satisfy the constraints described above (the framework is depicted in Fig. 7.2). In the next section, we describe the architecture used for representation learning and the losses that are used to implement these constraints. All losses are across modalities.

7.3.1 Network Architecture

Our architecture consists of two sub-networks, one sub-network that ingests five cropped faces as input, and another sub-network that takes in short-term magnitude spectrograms of 0.2-second speech segments. Each sub-network contains a block of five convolutional layers as the basic feature extraction trunk (these are shared for both content and identity, as it has been speculated that lower level features, e.g. edges for images and formants for speech, are likely to be common [Shinohara, 2016] for different high level tasks). Both sub-networks are based on the VGG-M architecture [Chatfield et al., 2014] which strikes a good trade-off between efficiency and performance. See [S.-W. Chung et al., 2019] for the exact filter sizes. After this, each block branches into two separate fully connected layers, one that produces

identity embeddings and one that produces content embeddings, both of dimension 1024. For $N+4$ input frames, N identity and content embeddings are produced for each modality stream (Fig. 7.2), since both sub-networks have temporal receptive fields of 5 frames (0.2 second) and strides of 1 frame (0.04 second). During training, the identity vectors from the audio stream are then averaged into a single vector, while a single identity vector is selected from the face stream at random. To understand this choice, note that if we were to also average the face embeddings, then the task of matching identity representations would simply become one of lip reading, i.e. matching the linguistic content of the audio and visual signals. Hence we pick a single random face vector and make the assumption that a face from a single frame is insufficient to encode linguistic content.

Self-Supervised Paired Data Inputs. In a single minibatch, we take B face-tracks, each of 1.2 seconds. Within a face-track, we sample $N + 4$ consecutive face images and $N + 4$ temporally aligned speech segments from the 1.2-second speech segment. Hence the total number of input samples per batch is $(N + 4) \times B$ face images and $(N + 4) \times B$ speech segments.

7.3.2 Loss Functions

A *content loss (CL)* is used to implement the content constraint via a multi-way matching task, as described in [S.-W. Chung et al., 2019]. The loss takes one input feature from the visual stream and N features from the audio stream. Since only one of these audio features is a positive sample (i.e. in sync with the visual stream), this can be set up as any (N) -way feature matching task. Euclidean distances between the audio and video features are computed, resulting in N distances. A cross-entropy loss is applied on the inverse of this distance after passing through a softmax, encouraging the similarity between matching pairs to exceed that of non-matching pairs.

An *identity loss (IL)* is used to implement the identity constraint. It is similar in form to the

content loss, but the negative samples are now obtained from different tracks, as opposed to *within* a track. The task becomes one of selecting the correct track averaged identity speech representation for a single face representation from all the B tracks in a batch, i.e. this is a B-way classification task.

Disentanglement losses (DL) are used to encourage explicit separation of representations—for this we use the confusion loss implemented by [Alvi et al., 2018] (inspired by [Tzeng et al., 2015]). This loss is used to assess the amount of spurious variation information left in either feature representation and then remove it (for the identity representation, content information is a spurious variation and vice versa). Minimizing this loss seeks to change the feature representation, such that it becomes invariant to the spurious variations. To remove identity from content, we perform the B-way identity matching task *across* facetracks using the content vectors as input instead (D1 in Fig. 7.2). We then minimise the cross-entropy between the output predicted from the model and a uniform distribution. Similarly, we apply the N-way content classification loss to the identity vectors and minimise the cross-entropy with the output to a uniform distribution (D2 in Fig. 7.2). See [Alvi et al., 2018], Equations 1–3 for exact details.

7.4 Experiments

We train our model using the following loss combinations: (1) Only the content loss: in this case the identity streams are not present in the network; (2) Using only the identity loss: in this case the content streams are not present in the network; (3) Joint training with both the content and the identity loss; (4) Joint training with the content, identity and disentanglement losses. In all cases the model uses the same trunk architecture and training hyperparameters.

Implementation Details. The model is implemented using PyTorch. It is trained end-to-end with batch size $B = 30$ and $N = 30$ samples per face-track using SGD (initial learning rate

of $1e-2$ which decays by 0.95 per epoch).

7.4.1 Dataset

We train our model on VoxCeleb2 [J. S. Chung et al., 2018], a large-scale audio-visual dataset of interviews obtained from unedited YouTube videos. The dataset consists of over a million utterances for 6,112 identities. No identity labels are used during training. To reduce computational cost, we sample only 20% of the speech per speaker for training from the VoxCeleb *dev* set, and validate performance of the self-supervised learning objectives on 120 speakers from the VoxCeleb2 test set. The statistics of the dataset can be seen in Table 7.1.

	# face-tracks	# identities
Training set	218,340	5,994
Test set	36,600	120

Table 7.1: Dataset Statistics. Although we report the no. of identities in the dataset, the identities are *not used at any point during training*.

7.4.2 Evaluation

We first evaluate the performance of our model on the two self-supervised learning objectives that it was trained for, and then evaluate the learned representations on the downstream task of speaker recognition on the standard VoxCeleb1 speaker recognition benchmark.

Learning Objective. We evaluate the self-supervised learning objectives on 120 speakers from the VoxCeleb2 test set (Table 7.1), and the results can be seen in Table 7.2. We evaluate the learned identity representations on the N-way classification task within a facetrack (content task), as well as evaluating it on the identity B-way classification task. From Table 7.2, it is clear that training both self-supervised objectives jointly improves performance on the identity classification task over training for identity alone (48.2 % vs 44.3 %) and training with the disentanglement losses provides a further improvement (49.6 %). In order to further probe

	Content Task	Identity Task	
	N -way cls.	B -way cls.	EER
Random	3.3%	3.3%	50.0%
Content loss only	49.0%	–	–
Identity loss only	–	44.3%	24.8%
<i>Content Embeddings</i>			
Con. and Id. Loss	46.7%	8.5%	45.7%
Con., Id. and Dis. Loss	49.0%	10.5%	45.2%
<i>Identity Embeddings</i>			
Con. and Id. Loss	19.3%	48.2%	23.1%
Con., Id. and Dis. Loss	12.0%	49.6%	18.9%

Table 7.2: Results on the self-supervised training objectives. The content task is N -way classification (N = number of samples per face-track), and the Identity task is B -way classification (B = number of face-tracks per minibatch). With $N = B = 30$, random performance is 3.3%. Lower EER, higher cls. accuracy is better. We want good performance of identity embeddings on the identity task, and low performance on the content task.

the effect of the disentanglement losses, however, we look at the performance of the identity embeddings on the content classification task (which ideally it should perform poorly on). From Table 7.2, it can be seen that disentanglement helps remove content information from the identity embedding – the accuracy drops from 19.3 % to 12.0 %, on the N -way content classification task.

As an aside, we also report performance of the content embeddings in the middle two rows of Table 7.2 (although learning content representations for their own sake is not the objective of this work) and note that joint training actually harms the performance compared to training with the content loss alone (from 49.0% to 46.7%) on the content classification task, however this performance is recovered by adding in the disentanglement losses. This is to be expected, as it is very difficult for identity information to leak into the content representation when it is trained for content alone (the content objective is trained with a large number of *negative* pairs within the same face-track, discouraging the embedding from learning identity).

Speaker Recognition. We then extract identity embeddings for the data in the VoxCeleb1 *test* set (VoxCeleb1-O, 40 speakers) [Nagrani et al., 2017]. We first evaluate using the self-supervised embeddings directly (i.e. without *any* speaker identity labels at all), and report results in Table 7.3. The negative cosine distance between embeddings is calculated directly and used as the similarity score between verification pairs. Once again we see a similar trend in the results, both joint training and disentanglement show cumulative gains in performance. We then compare our method to fully supervised performance, by freezing the layers of our network and then finetuning a single fully connected layer on the embedding network with n-pair loss, using labels from the VoxCeleb1 *dev* set. We do this for various subsets of the VoxCeleb1 *dev* set, and demonstrate in Table 7.4 that up until 500 speakers, our self-supervised method (even with only the identity loss, and with gains using the other two losses) outperforms full supervision. The fully supervised baseline is trained end-to-end, and for a fair comparison, has the exact same architecture as the audio stream of the cross-modal model.

Method	EER
Identity loss only	23.15%
Identity loss + Content loss	22.59%
Identity loss + Content loss + Dis. loss	22.09%

Table 7.3: Speaker verification results on the VoxCeleb1 test set. Lower is better. EER: Equal Error Rate.

# speakers	100	250	500	1,211
# utterances	1,228	6,019	12,146	ALL
Id. loss only	15.05%	13.00%	11.16%	9.85%
Id.+Cont.+Dis. loss	14.33%	12.69%	10.94%	9.43%
Fully supervised	19.84%	13.60%	11.35%	7.28%

Table 7.4: Comparison to fully supervised performance on the VoxCeleb1 test set measured in EER. For the first two rows, a single fully connected layer is trained on the self-supervised embeddings. The fully supervised model is trained end-to-end with labels. Lower is better.

7.5 Conclusion

In this work we develop a self-supervised method that learns speaker recognition embeddings from speech without access to any training labels, simply by using the co-occurrence of faces in video. By explicitly disentangling factors of variation such as content and identity, and training for both objectives with a common trunk architecture, we show improvements in generalisation to unseen speakers, and in the case of small amounts of training data, even outperform fully supervised methods.

Acknowledgements This work is funded by the EPSRC Programme Grant Seebibyte EP/M013774/1 and ExTol EP/R03298X/1. Arsha is funded by a Google PhD Fellowship.

Statement of Authorship

A statement of authorship for this work can be found in [Appendix C](#).

Part III

Multimodal Fusion

The whole is greater than the sum of its parts.

— Aristotle

8 | EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition

Evangelos Kazakos¹ Arsha Nagrani² Andrew Zisserman² Dima Damen¹

¹Visual Information Lab, Bristol ²VGG, Oxford

Abstract

We focus on multi-modal fusion for egocentric action recognition, and propose a novel architecture for multi-modal temporal-binding, i.e. the combination of modalities within a range of temporal offsets. We train the architecture with three modalities – RGB, Flow and Audio – and combine them with mid-level fusion alongside sparse temporal sampling of fused representations. In contrast with previous works, modalities are fused before temporal aggregation, with shared modality and fusion weights over time. Our proposed architecture is trained end-to-end, outperforming individual modalities as well as late-fusion of modalities.

We demonstrate the importance of audio in egocentric vision, on per-class basis, for identifying actions as well as interacting objects. Our method achieves state of the art results on both the seen and unseen test sets of the largest egocentric dataset: EPIC-Kitchens, on all metrics using the public leaderboard.

Published in the Proceedings of the [International Conference on Computer Vision, 2019](#).

8.1 Introduction

With the availability of multi-sensor wearable devices (e.g. GoPro, Google Glass, Microsoft Hololens, MagicLeap), egocentric audio-video recordings have become popular in many areas such as extreme sports, health monitoring, life logging, and home automation. As a result, there has been a renewed interest from the computer vision community on collecting large-scale datasets [Damen et al., 2018; Sigurdsson et al., 2018] as well as developing new or adapting existing methods to the first-person point-of-view scenario [Y. Zhou & Berg, 2015; Pirsiavash & Ramanan, 2012; Y. J. Lee et al., 2012; Ma et al., 2016; Damen et al., 2014; Yonetani et al., 2016].

In this work, we explore audio as a prime modality to provide complementary information to visual modalities (appearance and motion) in egocentric action recognition. While audio has been explored in video understanding in general [Arandjelović & Zisserman, 2018; Aytar et al., 2016; Arandjelović & Zisserman, 2017; Owens et al., 2016; Owens & Efros, 2018; Aytar, Vondrick, & Torralba, 2017; Nagrani et al., 2018b; Miech et al., 2018; Senocak et al., 2018; Gao & Grauman, 2019] the egocentric domain in particular offers rich sounds resulting from the interactions between hands and objects, as well as the close proximity of the wearable microphone to the undergoing action. Audio is a prime discriminator for some actions (e.g. ‘wash’, ‘fry’) as well as objects within actions (e.g. ‘put plate’ vs ‘put bag’). At times, the temporal progression (or change) of sounds can separate visually ambiguous actions (e.g. ‘open tap’ vs ‘close tap’). Audio can also capture actions that are out of the wearable camera’s field of view, but audible (e.g. ‘eat’ can be heard but not seen). Conversely, other actions are *sound-less* (e.g. ‘wipe hands’) and the wearable sensor might capture irrelevant sounds, such as talking or music playing in the background. The opportunities and challenges of incorporating audio in egocentric action recognition allow us to explore new multi-sensory fusion approaches, particularly related to the potential *temporal asynchrony* between the action’s appearance and the discriminative audio signal – the main focus of our work.

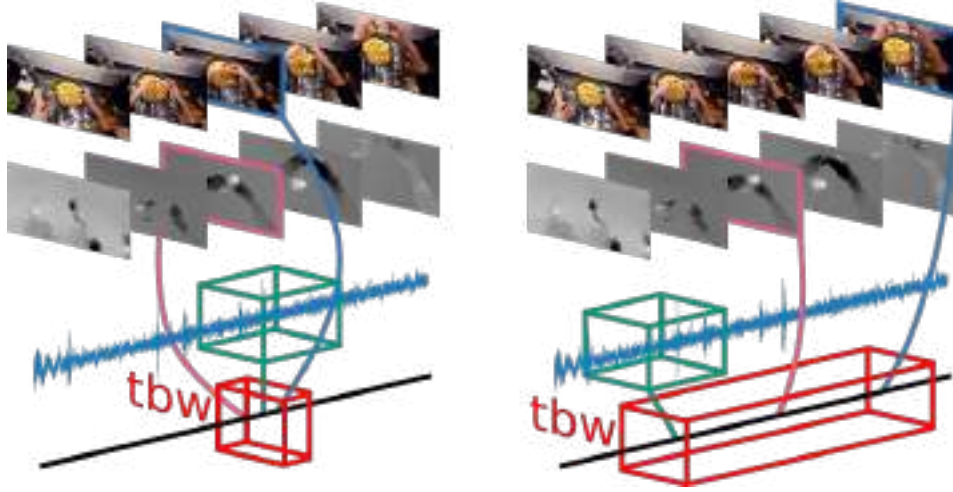


Figure 8.1: As the width of the temporal binding window increases (left to right), modalities (appearance, motion and audio) are fused with varying temporal shifts.

While several multi-modal fusion architectures exist for action recognition, current approaches perform temporal aggregation *within* each modality *before* modalities are fused [L. Wang, Xiong, et al., 2016; Miech, Laptev, & Sivic, 2017] or embedded [Miech et al., 2018]. Works that do fuse inputs before temporal aggregation, e.g. [Feichtenhofer et al., 2016], do so with inputs synchronised across modalities. In Fig. 8.1, we show an example of ‘breaking an egg into a pan’ from the EPIC-Kitchens dataset. The distinct sound of cracking the egg, the motion of separating the egg and the change in appearance of the egg occur at different frames/temporal positions within the video. Approaches that fuse modalities with synchronised input would thus be limited in their ability to learn such actions. In this work, we explore fusing inputs within a Temporal Binding Window (TBW) (Fig 8.1), allowing the model to train using asynchronous inputs from the various modalities. Evidence in neuroscience and behavioural sciences points at the presence of such a TBW in humans [Parise et al., 2012; Wallace & Stevenson, 2014]. The TBW offers a “range of temporal offsets within which an individual is able to perceptually bind inputs across sensory modalities” [Stevenson et al., 2013]. This is triggered by the gap in the biophysical time to process different senses [Magavand et al., 2013]. Interestingly, the width of the TBW in humans is heavily task-dependant, shorter for

simple stimuli such as flashes and beeps and intermediate for complex stimuli such as a hammer hitting a nail [Wallace & Stevenson, 2014].

Combining our explorations into audio for egocentric action recognition, and using a TBW for asynchronous modality fusion, our contributions are summarised as follows. First, an end-to-end trainable mid-level fusion Temporal Binding Network (TBN) is proposed¹. Second, we present the first audio-visual fusion attempt in egocentric action recognition. Third, we achieve state-of-the-art results on the EPIC-Kitchens public leaderboards on both seen and unseen test sets. Our results show (i) the efficacy of audio for egocentric action recognition, (ii) the advantage of mid-level fusion within a TBW over late fusion, and (iii) the robustness of our model to background or irrelevant sounds.

8.2 Related Work

We divide the related works into three groups: works that fuse visual modalities (RGB and Flow) for action recognition (AR), works that fuse modalities for egocentric AR in particular, and finally works from the recent surge in interest of audio-visual correspondence and source separation.

Visual Fusion for AR: By observing the importance of spatial and temporal features for AR, two-stream (appearance and motion) fusion has become a standard technique [Simonyan & Zisserman, 2014; Feichtenhofer et al., 2016; L. Wang, Xiong, et al., 2016]. *Late fusion*, first proposed by Simonyan and Zisserman [Simonyan & Zisserman, 2014], combines the streams’ independent predictions. Feichtenhofer et al. [Feichtenhofer et al., 2016] proposed *mid-level fusion* of the spatial and temporal streams, showing optimal results by combining the streams after the last convolutional layer. In [Carreira & Zisserman, 2017], 3D convolution for spatial and motion streams was proposed, followed by late fusion of modalities. All these approaches do not model the temporal progression of actions, a problem addressed by [L. Wang, Xiong,

¹Code at: <http://github.com/ekazakos/temporal-binding-network>

et al., 2016]. Temporal Segment Networks (TSN) [L. Wang, Xiong, et al., 2016] perform sparse temporal sampling followed by temporal aggregation (averaging) of softmax scores across samples. Each modality is trained independently, with late fusion of modalities by averaging their predictions. Follow-up works focus on pooling for temporal aggregation, still training modalities independently [B. Zhou et al., 2018; Girdhar et al., 2017]. Modality fusion before temporal aggregation was proposed in [W. Lin et al., 2018], where the appearance of the current frame is fused with 5 uniformly sampled motion frames, and vice versa, using two temporal models (LSTM). While their motivation is similar to ours, their approach focuses on using predefined asynchrony offsets between two modalities. In contrast, we relax this constraint and allow fusion from any random offset within a temporal window, which is more suitable for scaling up to many modalities.

Fusion in Egocentric AR: Late fusion of appearance and motion has been frequently used in egocentric AR [Damen et al., 2018; S. Song et al., 2016; Moltisanti et al., 2017; Sudhakaran & Lanz, 2018], as well as extended to additional streams aimed at capturing egocentric cues [Ma et al., 2016; S. Singh et al., 2016; S. Song et al., 2016]. In [Ma et al., 2016], the spatial stream segments hands and detects objects. The streams are trained jointly with a triplet loss on objects, actions and activities, and fused through concatenation. [S. Singh et al., 2016] uses head motion features, hand masks, and saliency maps, which are stacked and fed to both a 2D and a 3D ConvNet, and combined by late fusion. All previous approaches have relied on small-scale egocentric datasets, and none utilised audio for egocentric AR.

Audio-Visual Learning: Over the last three years, significant attention has been paid in computer vision to an underutilised and readily available source of information existing in video: the audio stream [Arandjelović & Zisserman, 2018; Aytar et al., 2016; Arandjelović & Zisserman, 2017; Owens et al., 2016; Owens & Efros, 2018; Aytar, Vondrick, & Torralba, 2017; Nagrani et al., 2018b; Miech et al., 2018; Senocak et al., 2018; Gao & Grauman, 2019]. These fall in one of four categories: i) *audio-visual representation learning* [Arandjelović & Zisserman, 2017; Aytar et al., 2016; Aytar, Vondrick, & Torralba, 2017; Miech et al., 2018; Owens

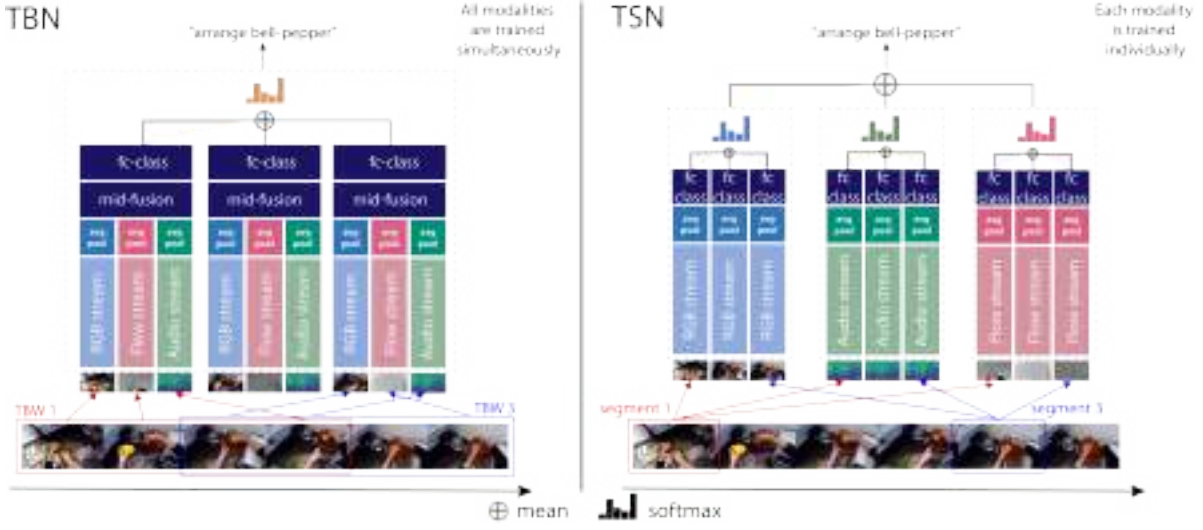


Figure 8.2: **Left:** our proposed Temporal Binding Network (TBN). Modalities are sampled within a TBW, and modality-specific weights (same colour) are shared amongst different inputs. Modalities are fused with mid-level fusion and trained jointly. Predictions from multiple TBWs, possibly overlapping, are averaged. **Right:** TSN [L. Wang, Xiong, et al., 2016] with an additional audio stream performing *late* fusion. Modalities are trained independently. Note that while in TSN a prediction is made for each modality, TBN produces a single prediction per TBW after fusing all modality representations. Best viewed in colour.

& Efros, 2018; Owens et al., 2016], ii) *sound-source localisation* [Arandjelović & Zisserman, 2018; Owens & Efros, 2018; Senocak et al., 2018], iii) *audio-visual source separation* [Owens & Efros, 2018; Gao & Grauman, 2019] and (iv) *visual-question answering* [Alamri et al., 2018]. These approaches attempt fusion [Arandjelović & Zisserman, 2017; Owens & Efros, 2018] or embedding into a common space [Arandjelović & Zisserman, 2018; Aytar, Vondrick, & Torralba, 2017; Nagrani et al., 2018a]. Several works sample the two modalities with temporal shifts, for learning better synchronous representations [Owens & Efros, 2018; Korbar et al., 2018]. Others sample within a 1s temporal window, to learn a correspondence between the modalities, e.g. [Arandjelović & Zisserman, 2017, 2018]. Of these works, [Owens & Efros, 2018; Korbar et al., 2018] note this audio-visual representation learning could be used for AR, by pretraining on the self-supervised task and then fine-tuning for AR.

Fusion for AR using three modalities (appearance, motion and audio) has been explored in [Z. Wu et al., 2016], employing late-fusion of predictions, and [Long, Gan, de Melo, et al., 2018; Long, Gan, Melo, et al., 2018] using attention to integrate local features into a global representation. Tested on UCF101, [Z. Wu et al., 2016] shows audio to be the least informative modality for third person action recognition (16% accuracy for audio compared to 80% and 78% for spatial and motion). A similar conclusion was made for other third-person datasets (AVA [Girdhar et al., 2018] and Kinetics [Long, Gan, de Melo, et al., 2018]).

In this work, we show audio to be a competitive modality for egocentric AR on EPIC-Kitchens, achieving comparable performance to appearance. We also demonstrate that audio-visual modality fusion in egocentric videos improves the recognition performance of both the action and the accompanying object.

8.3 The Temporal Binding Network

Our goal is to find the optimal way to fuse multiple modality inputs while modelling temporal progression through sampling. We first explain the general notion of temporal binding of multiple modalities in Sec 8.3.1, then detail our architecture in Sec 8.3.2.

8.3.1 Multimodal Temporal Binding

Consider a sequence of samples from one modality in a video stream, $m_i = (m_{i1}, m_{i2}, \dots, m_{iT/r_i})$ where T is the video’s duration and r_i is the modality’s framerate (or frequency of sampling). Input samples are first passed through unimodal feature extraction functions f_i . To account for varying representation sizes and frame-rates, most multi-modal architectures apply pooling functions G to each modality in the form of average pooling or other temporal pooling functions (e.g. maximum or VLAD [Arandjelović et al., 2016]), before attempting multimodal fusion.

Given a pair of modalities m_1 and m_2 , the final class predictions for a video are hence obtained

as follows:

$$y = h(G(f_1(m_1)), G(f_2(m_2))) \quad (8.1)$$

where f_1 and f_2 are unimodal feature extraction functions, G is a temporal aggregation function, h is the multimodal fusion function and y is the output label for the video. In such architectures (e.g. TSN [L. Wang, Xiong, et al., 2016]), modalities are temporally aggregated for a prediction before different modalities are fused; this is typically referred to as ‘late fusion’.

Conversely, multimodal fusion can be performed at *each* time step as in [Feichtenhofer et al., 2016]. One way to do this would be to synchronise modalities and perform a prediction at *each* time-step. For modalities with matching frame rates, synchronised multi-modal samples can be selected as (m_{1j}, m_{2j}) , and fused according to the following equation:

$$y = h(G(f_{sync}(m_{1j}, m_{2j}))) \quad (8.2)$$

where f_{sync} is a multimodal feature extractor that produces a representation for each time step j , and G then performs temporal aggregation over all time steps. When frame rates vary, and more importantly so do representation sizes, only approximate synchronisation can be attempted,

$$y = h(G(f_{sync}(m_{1j}, m_{2k}))) \quad : k = \lceil \frac{jr_2}{r_1} \rceil \quad (8.3)$$

We refer to this approach as ‘synchronous fusion’ where synchronisation is achieved or approximated.

In this work, however, we propose fusing modalities within temporal windows. Here modalities are fused within a range of temporal offsets, with all offsets constrained to lie within a finite time window, which we henceforth refer to as a temporal binding window (TBW). Formally,

$$y = h(G(f_{tbw}(m_{1j}, m_{2k}))) \quad : k \in [\lceil \frac{jr_2}{r_1} - b \rceil, \lceil \frac{jr_2}{r_1} + b \rceil] \quad (8.4)$$

where f_{tbw} is a multimodal feature extractor that combines inputs within a binding window of width $\pm b$. Interestingly, as the number of modalities increases, say from two to three modalities, the TBW representation allows fusion of modalities each with different temporal offsets, yet within the same binding window $\pm b$:

$$y = h\left(G(f_{tbw}(m_{1j}, m_{2k}, m_{3l}))\right) : \begin{aligned} k &\in \left[\lceil \frac{jr_2}{r_1} - b \rceil, \lceil \frac{jr_2}{r_1} + b \rceil\right] \\ l &\in \left[\lceil \frac{jr_3}{r_1} - b \rceil, \lceil \frac{jr_3}{r_1} + b \rceil\right] \end{aligned} \quad (8.5)$$

This formulation hence allows a large number of different inputs combinations to be fused. This is different from proposals that fuse inputs over predefined temporal differences (e.g. [W. Lin et al., 2018]). Sampling within a temporal window allows fusing modalities with various temporal shifts, *up to* the temporal window width $\pm b$. This: 1) enables straightforward scaling to multiple modalities with different frame rates, 2) allows training with a variety of temporal shifts, accommodating, say, different speeds of action performance and 3) provides a natural form of data augmentation.

With the basic concept of a TBW in place, we now describe our proposed audio-visual fusion model, TBN.

8.3.2 TBN with Sparse Temporal Sampling

Our proposed TBN architecture is shown in Fig 8.2 (left). First, the action video is divided into K segments of equal width. Within each segment, we select a random sample of the first modality $\forall k \in K : m_{1k}$. This ensures the temporal progression of the action is captured by sparse temporal sampling of this modality, as with previous works [L. Wang, Xiong, et al., 2016; B. Zhou et al., 2018], while random sampling within the segment offers further data for training. The sampled m_{1k} is then used as the centre of a TBW of width $\pm b$. The other

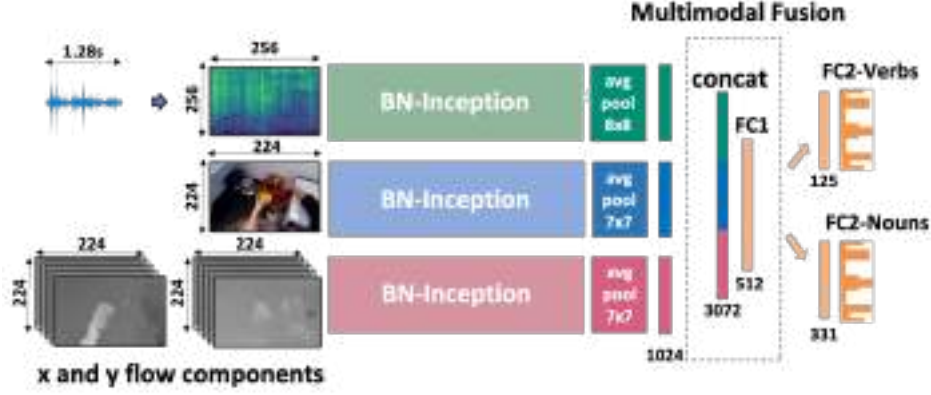


Figure 8.3: A single TBN block showing architectural details and feature sizes. Outputs from multiple TBN blocks are averaged as shown in Fig. 8.2. We model the problem of learning both verbs and nouns as a multi-task learning problem, by adding two output FC layers, one that predicts verbs and the other nouns (as in [Damen et al., 2018]). Best viewed in colour.

modalities are selected randomly from within each TBW (Eq. 8.5). In total, the input to our architecture in both training and testing is $K \times M$ samples from M modalities.

Within each of the K TBWs, we argue that the complementary information in audio and vision can be better exploited by combining the internal representations of each modality before temporal aggregation, and hence we propose a *mid-level* fusion. A ConvNet (per modality) extracts *mid-level* features, which are then fused through *concatenating* the modality features and feeding them to a fully-connected layer, making multi-modal predictions per TBW. We backpropagate all the way to the inputs of the ConvNets. Fig 8.3 details the proposed TBN block. The predictions, for each of these unified multimodal representations, are then aggregated for video-level predictions. In the proposed architecture, we train all modalities simultaneously. The convolutional weights for each modality are shared over the K segments. Additionally, mid-level fusion weights and class prediction weights are also shared across the segments.

To avoid biasing the fusion towards longer or shorter action lengths, we calculate the window width b relative to the action video length. Our TBW is thus of variable width, where the

width is a function of the length of the action. We note again that b can be set independently of the number of segments K , allowing the temporal windows to overlap. This is detailed in Sec. 8.4.1.

Relation to TSN. In Fig 8.2, we contrast the TBN architecture (left) to an extended version of the TSN architecture (right). The extension is to include the audio modality, since the original TSN only utilises appearance and motion streams. There are two key differences: first, in TSN each modality is temporally aggregated independently (across segments), and the modalities are only combined by late fusion (e.g. the RGB scores of each segment are temporally aggregated, and the flow scores of each segment are temporally aggregated, individually). Hence, it is not possible to benefit from combining modalities *within* a segment which is the case for TBN. Second, in TSN, each modality is trained independently first after which predictions are combined in inference. In the TBN model instead, all modalities are trained simultaneously, and their combination is also learnt.

8.3.3 Different mid-fusion strategies

As Fig 8.2 indicates, our TBW framework performs mid-level fusion on the modalities within the binding window. In this section we discuss a number of multimodal fusion strategies, for which results are provided in section 8.4.2.

(i) Multi-modal fusion with concatenation (*mid-level fusion*) has been extensively used in the literature [Feichtenhofer et al., 2016; Arandjelović & Zisserman, 2017; Nagrani et al., 2018b], including for egocentric AR [Ma et al., 2016], where the feature maps of each modality are concatenated, and a fully-connected layer is used to model the cross-modal relations, where each unit is a weighted combination of the units across all modalities.

$$f_{tbw}^{concat} = \phi(W[m_{1j}, m_{2k}, m_{3l}] + b) \quad (8.6)$$

where ϕ is a non-linear activation function. When used within TBWs, shared weights f_{tbw} are to be learnt between modalities with a range of temporal shifts (Fig 8.3).

(ii) Context gating was used in [Miech, Laptev, & Sivic, 2017], with the aim to introduce non-linear interactions along the dimensions of a feature representation, and to recalibrate the strength of the activations of different units with a self-gating mechanism:

$$h = \phi(W[m_{1j}, m_{2k}, m_{3l}] + b_h) \quad (8.7)$$

$$f_{tbw}^{context} = \sigma(Wh + b_z) \circ h \quad (8.8)$$

where \circ is element-wise multiplication. We apply context gating on top of our multi-modal fusion with concatenation, where h in (8.16) is the output of the mid-level fusion FC layer.

(iii) Multi-modal gating fusion was introduced in [Arevalo et al., 2017], where a gate neuron takes as input the features from all modalities to learn the importance of one modality w.r.t. its relations to all modalities. It is described in [Arevalo et al., 2017] for three modalities $i \in \{1, 2, 3\}$ as:

$$h_i = \phi(W_i m_{ij} + b_i) \quad \forall i \quad (8.9)$$

$$z_i = \sigma(W_{zi}[m_{1j}, m_{2k}, m_{3l}] + b_{zi}) \quad \forall i \quad (8.10)$$

$$f_{tbw}^{gating} = z_1 \circ h_1 + z_2 \circ h_2 + z_3 \circ h_3, \quad (8.11)$$

In Eq. (8.19), the gate neurons are unconstrained, in that they are not tied to trade-off between modalities as mentioned in [Arevalo et al., 2017]. We propose an alternative formulation of f_{tbw}^{gating} , keeping Eq. (8.17) the same, by binding pairs of modalities together against the third one:

$$z_i = \sigma(W_{zi}[m_{ij}, \sum_{n \neq i} m_{nk}] + b_{z_i}) \quad (8.12)$$

$$\begin{aligned} f_{tbw}^{gating} &= z_1 \circ h_1 + (1 - z_1) \circ (h_2 + h_3) \\ &+ z_2 \circ h_2 + (1 - z_2) \circ (h_1 + h_3) \\ &+ z_3 \circ h_3 + (1 - z_3) \circ (h_1 + h_2). \end{aligned} \quad (8.13)$$

8.4 Experiments

Dataset: We evaluate the TBN architecture on the largest dataset in egocentric vision: EPIC-Kitchens [Damen et al., 2018], which contains 39,596 action segments recorded by 32 participants performing non-scripted daily activities in their native kitchen environments. In EPIC-Kitchens, an action is defined as a combination of a *verb* and a *noun*, e.g. ‘cut cheese’. There are in total 125 verb classes and 331 noun classes, though these are heavily-imbalanced. The test set is divided in two splits: Seen Kitchens (S1) where sequences from the same environment are in both training, and Unseen Kitchens (S2) where the complete sequences for 4 participants are held out for testing. Importantly, EPIC-Kitchens sequences have been captured using a head-mounted Go-Pro with the audio released as part of the dataset. No previous baseline on using audio for this dataset is available.

8.4.1 Implementation Details

RGB and Flow: We use the publicly available RGB and computed optical flow with the dataset [Damen et al., 2018].

Audio Processing: We extract 1.28s of audio, convert it to single-channel, and resample it to 24kHz. We then convert it to a log-spectrogram representation using an STFT of window length 10ms, hop length 5ms and 256 frequency bands. This results in a 2D spectrogram

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
s_1	RGB	45.68	36.80	19.86	85.56	64.19	41.89	61.64	34.32	09.96	23.81	31.62	08.81
	FLOW	55.65	31.17	20.10	85.99	56.00	39.30	48.83	26.84	09.02	27.58	24.15	07.89
	Audio	43.56	22.35	14.21	79.66	43.68	27.82	32.28	19.10	07.27	25.33	18.16	06.17
	RGB-Flow (late)	55.32	39.16	24.75	86.81	64.66	44.59	52.20	37.72	13.56	28.47	33.13	11.68
	RGB-Audio (late)	49.54	32.82	20.33	86.23	62.30	40.26	40.98	27.28	11.29	28.59	26.95	09.72
	Flow-Audio (late)	52.21	28.69	19.12	86.73	56.73	37.94	41.34	22.96	10.13	29.52	22.64	08.45
	All (late)	55.49	36.27	23.95	87.04	64.17	44.26	53.85	30.94	13.55	30.60	29.82	11.11
	All (mid)	64.75	46.03	34.80	90.70	71.34	56.65	55.67	43.65	22.07	45.55	42.30	21.31
s_2	RGB	34.89	21.82	10.11	74.56	45.34	25.33	19.48	14.67	04.77	11.22	17.24	05.67
	FLOW	48.21	22.98	14.48	77.85	45.55	29.33	23.00	13.29	05.63	19.61	16.09	07.61
	Audio	35.43	11.98	06.45	69.20	29.49	16.18	22.46	09.41	04.59	18.02	09.79	04.19
	RGB-Flow (late)	45.79	24.93	14.99	77.34	49.37	29.06	23.92	15.60	07.46	17.51	19.37	09.38
	RGB-Audio (late)	39.23	18.00	09.26	76.07	44.76	24.00	24.78	13.75	05.86	18.09	13.84	05.31
	Flow-Audio (late)	45.24	18.10	10.73	78.19	44.80	26.23	30.17	12.02	06.17	21.97	13.64	06.15
	All (late)	46.61	22.50	13.05	78.19	48.59	29.13	28.92	15.48	06.47	21.58	16.61	07.55
	All (mid)	52.69	27.86	19.06	79.93	53.78	36.54	31.44	21.48	12.00	28.21	23.53	12.69

Table 8.1: Baseline results

matrix of size 256×256 , after which we compute the logarithm. Since many egocentric actions are very short (< 1.28 s), we extract 1.28s of audio from the untrimmed video, allowing the audio segment to extend beyond the action boundaries.

Training details: We implement our model in PyTorch [Paszke et al., 2017]. We use Inception with Batch Normalisation (BN-Inception) [Ioffe & Szegedy, 2015] as a base architecture, and fuse the modalities after the average pooling layer. We chose BN-Inception as it offers a good compromise between performance and model-size, critical for our proposed TBN that trains all modalities simultaneously, and hence is memory-intensive. Compared to TSN, the three modalities have 10.78M, 10.4M and 10.4M parameters, with only one modality in memory during training. In contrast, TBN has 32.64M paramaters.

We train using SGD with momentum [Qian, 1999], a batch size of 128, a dropout of 0.5, a momentum of 0.9, and a learning rate of 0.01. Networks are trained for 80 epochs, and the learning rate is decayed by a factor of 10 at epoch 60. We initialise the RGB and the Audio streams from ImageNet. While for the Flow stream, we use stacks of 10 interleaved horizontal and vertical optical flow frames, and use the pre-trained Kinetics [Carreira & Zisserman, 2017] model, provided by the authors of [L. Wang, Xiong, et al., 2016].

Note that our network is trained end-to-end for all modalities and TBWs. We train with $K = 3$

segments over the $M = 3$ modalities, with $b = T$, allowing the temporal window to be as large as the action segment. We test using 25 evenly spaced samples for each modality, as with the TSN basecode for direct comparison.

Pretraining: We want each stream to encapsulate prior discriminative knowledge on the modality that is trained on. To this end, we initialise each stream’s weights from a pretrained model, relevant to each modality. For the RGB stream, we use ImageNet pretrained weights. For the optical flow stream, we use the pretrained Kinetics [Carreira & Zisserman, 2017] model, provided by the authors of [L. Wang, Xiong, et al., 2016]. We pretrain the Audio stream using AudioSet [Gemmeke et al., 2017], the largest annotated audio dataset, containing 2,084,320 million sound clips extracted from YouTube videos. AudioSet consists of an ontology of sound events organised in a hierarchical structure, where each audio segment may have more than one labels. The ontology contains 527 audio event classes, and covers a wide range of sounds, such as, human sounds, animal sounds, natural sounds, music, etc. We use the balanced training split that contains 22,176 audio segments, and we hold out 10% for validation. We model the problem in a multi-task learning setting, where the labels are encoded with a 527-d vector, with one(s) at the location(s) of the label(s) and zeros everywhere else. We use the *multi-label soft margin loss*, defined as:

$$\begin{aligned} loss(pred, gt) = & - \sum_i gt(i) \log \left((1 + e^{-pred(i)})^{-1} \right) \\ & + (1 - gt(i)) \log \left(\frac{e^{-pred(i)}}{1 + e^{-pred(i)}} \right) \end{aligned} \quad (8.14)$$

8.4.2 Results

This section is organised as follows. First, we show and discuss the performance of single modalities, and compare them with our proposed TBN, with a special focus on the efficacy of

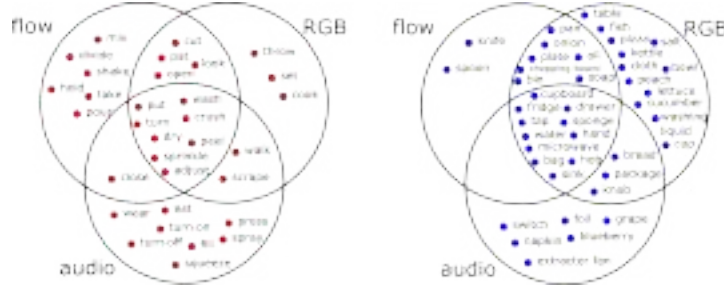


Figure 8.4: Verb (left) and noun (right) classes’ performances using single modalities for top-performing 32 verb and 41 noun classes, using single modality accuracy. For each, we consider whether the accuracy is high for Flow, Audio or RGB, or for two or all of these modalities. It can be clearly seen that noun classes can be predicted with high accuracy using RGB alone, whereas for many verbs, Flow and Audio are also important modalities.

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	RGB	45.68	36.80	19.86	85.56	64.19	41.89	61.64	34.32	09.96	23.81	31.62	08.81
	Flow	55.65	31.17	20.10	85.99	56.00	39.30	48.83	26.84	09.02	27.58	24.15	07.89
	Audio	43.56	22.35	14.21	79.66	43.68	27.82	32.28	19.10	07.27	25.33	18.16	06.17
	TBN (RGB+Flow)	60.87	42.93	30.31	89.68	68.63	51.81	61.93	39.68	18.11	39.99	38.37	16.90
	TBN (All)	64.75	46.03	34.80	90.70	71.34	56.65	55.67	43.65	22.07	45.55	42.30	21.31
S2	RGB	34.89	21.82	10.11	74.56	45.34	25.33	19.48	14.67	04.77	11.22	17.24	05.67
	Flow	48.21	22.98	14.48	77.85	45.55	29.33	23.00	13.29	05.63	19.61	16.09	07.61
	Audio	35.43	11.98	06.45	69.20	29.49	16.18	22.46	09.41	04.59	18.02	09.79	04.19
	TBN (RGB+Flow)	49.61	25.68	16.80	78.36	50.94	32.61	30.54	20.56	09.89	21.90	20.62	11.21
	TBN (All)	52.69	27.86	19.06	79.93	53.78	36.54	31.44	21.48	12.00	28.21	23.53	12.69

Table 8.2: Comparison of our fusion method to single modality performance. For both splits, the fusion outperforms single modalities. For the seen split, the RGB and Flow modalities perform comparatively, whereas for the unseen split the Flow modality outperforms RGB by a large margin. Audio is comparable to RGB on top-1 verb accuracy for both splits.

the audio stream. Second, we compare different mid-level fusion techniques. And finally, we investigate the effect of the TBW width on both training and testing.

Single-modality vs multimodal fusion performance: We examine the overall performance of each modality individually in Table 8.2. Although it is clear that RGB and optical flow are stronger modalities than audio, an interesting find is that audio performs comparably to RGB on some of the metrics (e.g. top-1 verb accuracy), signifying the relevance of audio on recognising egocentric actions. While as expected optical flow outperforms RGB in S2, interestingly for S1, the RGB and Flow modalities perform comparatively, and in some cases RGB performs better. This matches the expectation that Flow is more invariant to the environment.

To obtain a better analysis of how these modalities perform, we examine the accuracy of *individual* verb and noun classes on **S1**, using single modalities. Fig 8.4 plots top-performing verb and noun classes, into a Venn diagram. For each class, we consider the accuracy of individual modalities. If all modalities perform comparably (within 0.15), we plot that class in the intersection of the three circles. On the other hand, if one modality is clearly better than the others (more than 0.15), we plot the class in the outer part of the modality’s circle. For example, for the verb ‘close’, we have per-class accuracy of 0.23, 0.47 and 0.42 for RGB, Flow and Audio respectively. We thus note that this class performs best for two modalities: Flow and Audio, and plot it in the intersection of these two circles.

From this plot, many verb and noun classes perform comparably for all modalities (e.g. ‘wash’, ‘peel’ and ‘fridge’, ‘sponge’). This suggests all three modalities contain useful information for these tasks. A distinctive difference, however, is observed in the importance of individual modalities for verbs and nouns. Verb classes are strongly related to the temporal progression of actions, making Flow more important for verbs than nouns. Conversely, noun classes can be predicted with high accuracy using RGB alone. Audio, on the other hand, is important for both nouns and verbs, particularly for some verbs such as ‘turn-on’, and ‘spray’. For nouns, Audio tends to perform better for objects with distinctive sounds (e.g. ‘switch’, ‘extractor fan’) and materials that sound when manipulated (e.g. ‘foil’).

In Table. 8.2, we compare single modality performance to the performance over the three modalities. Single modalities are trained as in TSN, as TBN is designed to bind multiple modalities. We find that the fusion method outperforms single modalities, and that audio is a significantly informative modality across the board. Per-class accuracies, for individual modalities as well as for TBN trained on all three modalities, can be seen in Figure 8.6. The advantage of the fusion method is more pronounced for verbs (where we expect motion and audio to be more informative) than nouns, and more for particular noun classes than others, such as ‘pot’, ‘kettle’, ‘microwave’, and particular verb classes eg. ‘spray’ (fusion 0.54, RGB 0.09, Flow 0, Audio 0.3). This suggests that the mixture of complementary and redundant

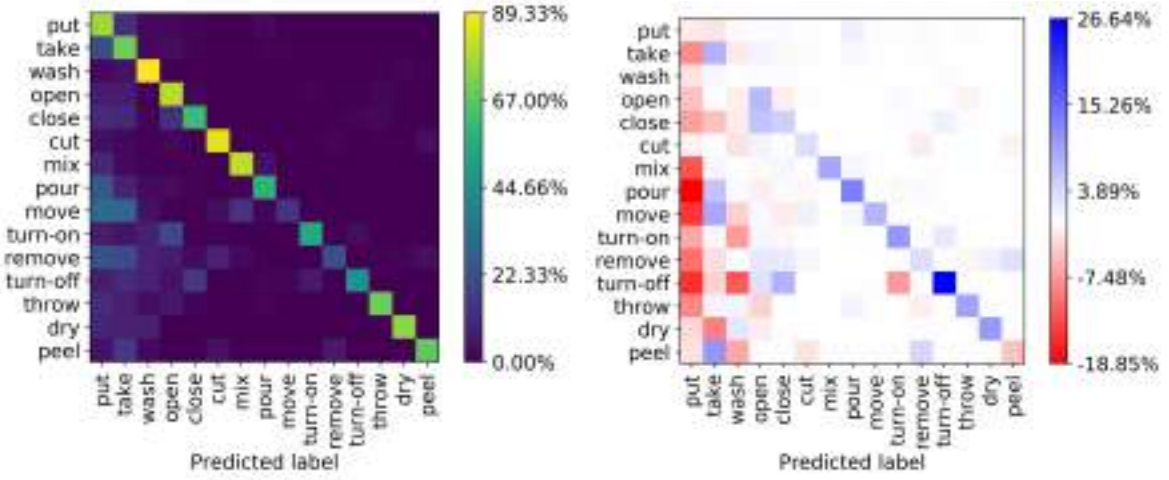


Figure 8.5: Confusion matrix for the largest-15 verb classes, with audio (left), as well as the difference to the confusion matrix without audio (right).

		All			RGB+Flow		
	TBN	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	irrelevant	61.37	46.46	32.63	57.28	42.55	27.73
	rest	65.28	45.97	35.14	61.44	42.99	30.72
S2	irrelevant	47.32	23.36	15.30	44.41	20.45	12.39
	rest	57.21	31.66	22.22	54.00	30.09	20.52

Table 8.3: Comparing top-1 accuracy of All modalities (left) to RGB+Flow (right). Actions are split in segments with ‘irrelevant’ background sounds, and the ‘rest’ of the test set.

information captured in a video is highly dependant on the action itself, yielding the fusion method to be more useful for some classes than for others. We also note that the fusion method helps to significantly boost the performance of the tail classes (Fig. 8.6, right), where individual modality performance tends to suffer.

Efficacy of audio: We train TBN only with the visual modalities (RGB+Flow) and the results can be seen in Table 8.2. An increase of 5% (S1) and 4% (S2) in top-5 action recognition accuracy with the addition of audio demonstrates the importance of audio for egocentric action recognition. Fig 8.5 shows the confusion matrix with the utilisation of audio for the largest-15 verb classes (in S1). Studying the difference (Fig 8.5 right) clearly demonstrates an increase (blue) in confidence along the diagonal, and a decrease (red) in confusion elsewhere.

Audio with irrelevant sounds: In the recorded videos for EPIC-Kitchens, background sounds

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	Concatenation	64.75	46.03	34.80	90.70	71.34	56.65	55.67	43.65	22.07	45.55	42.30	21.31
	Context gating [Miech, Laptev, & Sivic, 2017]	63.77	44.33	33.47	90.04	69.09	54.10	57.31	42.20	21.72	45.63	41.53	20.20
	Gating fusion [Arevalo et al., 2017]	61.52	43.54	31.61	89.54	68.42	52.57	52.07	39.62	18.39	42.55	39.77	18.66
S2	Concatenation	52.69	27.86	19.06	79.93	53.78	36.54	31.44	21.48	12.00	28.21	23.53	12.69
	Context gating [Miech, Laptev, & Sivic, 2017]	52.65	27.35	19.16	79.25	52.00	36.40	30.82	23.16	11.72	23.39	25.03	12.58
	Gating fusion [Arevalo et al., 2017]	50.16	27.25	18.41	78.80	50.84	34.04	28.42	22.42	12.34	23.92	24.15	13.14

Table 8.4: Comparison of mid-level fusion techniques for the TBN architecture.

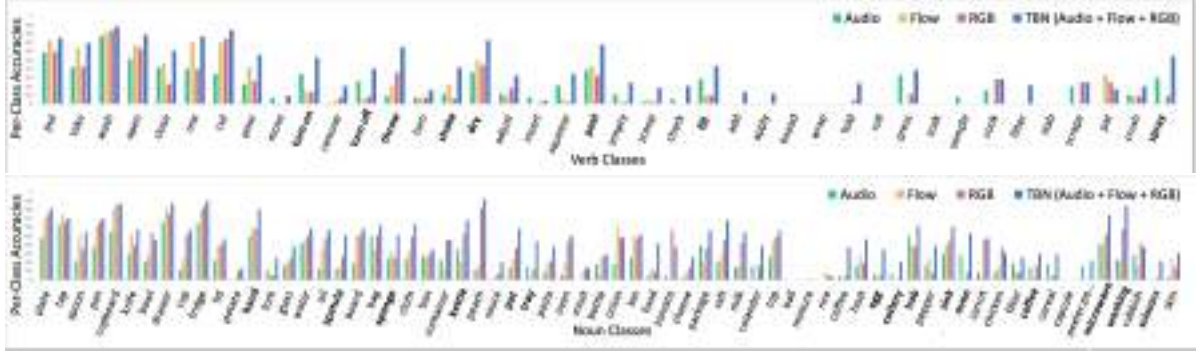


Figure 8.6: Per-class accuracies for the S1 test set for verbs (top) and nouns (bottom) for fusion and single modalities. We select verb classes with more than 10 samples, and noun classes with more than 30 samples. The classes are presented in the order of number of samples per class, from left to right. For most classes the fusion method provides significant performance gains over single modality classification (largest improvements shown in bold). Best viewed in colour.

irrelevant to the observed actions have been captured by the wearable sensor. These include music or TV playing in the background, ongoing washing machine, coffee machine or frying sounds while actions take place. To quantify the effect of these sounds, we annotated the audio in the test set, and report that 14% of all action segments in S1, and 46% of all action segments in S2 contain other audio sources. We refer to these as actions containing ‘irrelevant’ sounds, and independently report the results in Table 8.3. The table shows that the model’s accuracy increases consistently when audio is incorporated, even for the ‘irrelevant’ segments. Both models (All and RGB+Flow) show a drop in performance for ‘irrelevant’ S2 (comparing to ‘rest’), validating that irrelevant sounds are not the source of confusion, but that this set of action segments is more challenging even in the visual modalities. This demonstrates the robustness of our network to noisy and unconstrained audio sources.

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	Attention Clusters [Long, Gan, de Melo, et al., 2018]	40.39	19.37	11.09	78.13	41.73	24.36	21.17	09.65	02.50	14.89	11.50	03.41
	[Damen et al., 2018] (from leaderboard)	48.23	36.71	20.54	84.09	62.32	39.79	47.26	35.42	11.57	22.33	30.53	09.78
	Ours (TSN [L. Wang, Xiong, et al., 2016] w. Audio)	55.49	36.27	23.95	87.04	64.17	44.26	53.85	30.94	13.55	30.60	29.82	11.11
	Ours (TBN, Single Model)	64.75	46.03	34.80	90.70	71.34	56.65	55.67	43.65	22.07	45.55	42.30	21.31
	Ours (TBN, Ensemble)	66.10	47.89	36.66	91.28	72.80	58.62	60.74	44.90	24.02	46.82	43.89	22.92
S2	Attention Clusters [Long, Gan, de Melo, et al., 2018]	32.37	11.95	05.60	69.89	31.82	15.74	17.21	03.86	01.84	11.59	07.94	02.64
	[Damen et al., 2018] (from leaderboard)	39.40	22.70	10.89	74.29	45.72	25.26	22.54	15.33	06.21	13.06	17.52	06.49
	Ours (TSN [L. Wang, Xiong, et al., 2016] w. Audio)	46.61	22.50	13.05	78.19	48.59	29.13	28.92	15.48	06.47	21.58	16.61	07.55
	Ours (TBN, Single Model)	52.69	27.86	19.06	79.93	53.78	36.54	31.44	21.48	12.00	28.21	23.53	12.69
	Ours (TBN, Ensemble)	54.46	30.39	20.97	81.23	55.69	39.40	32.57	21.68	10.96	27.60	25.58	13.31

Table 8.5: Results on the EPIC-Kitchens for seen (S1) and unseen (S2) test splits. At the time of submission, our method outperformed all previous methods on all metrics, and in particular by 11%, 5% and 4% on top-1 verb, noun and action classification on S1. Our method achieved second ranking in the 2019 challenge. Screenshots of the leaderboard at submission and challenge conclusion are in the supplementary material.

Comparison of fusion strategies: As Fig 8.2 indicates, TBN performs mid-level fusion on the modalities within the binding window. Here we describe three alternative mid-level fusion strategies, and then compare their performances.

(i) Concatenation, where the feature maps of each modality are concatenated, and a fully-connected layer is used to model the cross-modal relations.

$$f_{tbw}^{concat} = \phi(W[m_{1j}, m_{2k}, m_{3l}] + b) \quad (8.15)$$

where ϕ is a non-linear activation function. When used within TBWs, shared weights f_{tbw} are to be learnt between modalities within a range of temporal shifts.

(ii) *Context gating* was used in [Miech, Laptev, & Sivic, 2017], aiming to recalibrate the strength of the activations of different units with a self-gating mechanism:

$$f_{tbw}^{context} = \sigma(Wh + b_z) \circ h \quad (8.16)$$

where \circ is element-wise multiplication. We apply context gating on top of our multi-modal fusion with concatenation, so h in Eq. (8.16) is equivalent to Eq. (8.15).

(iii) *Gating fusion* was introduced in [Arevalo et al., 2017], where a gate neuron takes as input

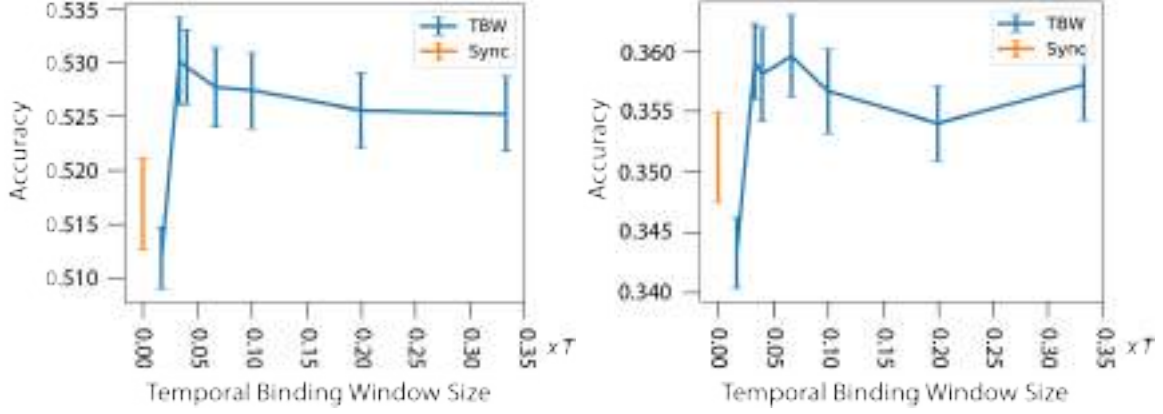


Figure 8.7: Effect of TBW width for verbs (left) and nouns (right) in the **S1** test set.

the features from all modalities to learn the importance of one modality w.r.t. all modalities.

$$h_i = \phi(W_i m_{ij} + b_i) \quad \forall i \quad (8.17)$$

$$z_i = \sigma(W_{zi}[m_{1j}, m_{2k}, m_{3l}] + b_{zi}) \quad \forall i \quad (8.18)$$

$$f_{tbw}^{gating} = z_1 \circ h_1 + z_2 \circ h_2 + z_3 \circ h_3, \quad (8.19)$$

In Table 8.4, we compare the various fusion strategies. We find that the simplest method, concatenation (Eq. 8.15) generally outperforms more complex fusion approaches. We believe this shows modality binding within a temporal binding window to be robust to the mid-level fusion method.

The effect of TBW width: Here, we investigate the effect of the TBW width in training and testing. We varied TBW width in training with $b \in \{\frac{T}{6}, \frac{T}{3}, T\}$, by training three TBN models for each respective window width. We noted little difference in performance. As changing b in training is expensive and performance is subject to the particular optimisation run, we opt for a more conclusive test by focusing on varying b in testing for a single model.

In testing, we vary $b \in \{\frac{T}{60}, \frac{T}{30}, \frac{T}{25}, \frac{T}{15}, \frac{T}{10}, \frac{T}{5}, \frac{T}{3}\}$. This corresponds, in average, to varying the width of TBW on the **S1** test set between 60ms and 1200ms. We additionally run with

synchrony $b \sim 0$.

In each case we sample a *single TBW*, to solely assess the effect of the window size. We repeat this experiment for 100 runs and report mean and standard deviation in Fig. 8.7, where we compare results for verb and noun classes separately. The figure shows that best performance is achieved for $b \in [\frac{T}{30}, \frac{T}{20}]$, that is on average $b \in [120ms \pm 190ms, 180ms \pm 285ms]$. TBWs of smaller width show a clear drop in performance, with synchrony comparable to $b = \frac{T}{60}$. Note that the ‘Sync’ baseline provides only approximate synchronisation of modalities, as modalities have different sampling rates (RGB 60fps, flow 30fps, audio 24000kHz). The model shows a degree of robustness for larger TBWs.

Note that in Fig. 8.7, we compare widths on a single temporal window in testing. When we temporally aggregate multiple TBWs, the effect of the TBW width is smoothed, and the model becomes robust to TBW widths.

Comparison with the state-of-the-art: We compare our work to the baseline results reported in [Damen et al., 2018] in Table 8.5 on all metrics. First we show that a late fusion with an additional audio stream, outperforms the baseline on top-1 verb accuracy by 7% on S1 and also 7% on S2. Second, we show that our TBN single model, improves these results even further (9%, 10% and 11% on top-1 verb, noun and action accuracy on S1, and 6%, 5% and 6% on S2 respectively). Finally we report results of an Ensemble of five TBNs, where each one is trained with a different TBW width. The ensemble shows additional improvement of up to 3% on top-1 metrics.

We compare TBN with Attention Clusters [Long, Gan, de Melo, et al., 2018], a previous effort to utilise RGB, Flow, and Audio for action recognition, using *pre-extracted features*. We use the authors available implementation, and fine-tuned features (TSN, BN-Inception), from the global avg pooling layer (1024D), to provide a fair comparison to TBN, and follow the implementation choices from [Long, Gan, de Melo, et al., 2018]. The method from [Long, Gan, de Melo, et al., 2018] performs significantly worse than the baseline, as pre-extracted

video features are used to learn attention weights.

At the time of submission, our TBN Ensemble results demonstrated an overall improvement over all state-of-the-art, published or anonymous, by 11% on top-1 verb for both S1 and S2. Our method was also ranked 2nd in the 2019 EPIC-Kitchens Action Recognition challenge.

8.5 Conclusion

We have shown that the TBN architecture is able to flexibly combine the RGB, Flow and Audio modalities to achieve an across the board performance improvement, compared to individual modalities. In particular, we have demonstrated how audio is complementary to appearance and motion for a number of classes; and the pre-eminence of appearance for noun (rather than verb) classes. The performance of TBN significantly exceeds TSN trained on the same data; and provides state-of-the-art results on the public EPIC-Kitchens leaderboard. Further avenues for exploration include a model that learns to adjust TBWs over time, as well as implementing class-specific temporal binding windows.

Acknowledgements Research supported by EPSRC LOCATE (EP/N033779/1), GLANCE (EP/N013964/1) & Seebibyte (EP/M013774/1). EK is funded by EPSRC Doctoral Training Partnership, and AN by a Google PhD Fellowship.

Additional Appendices

More details, code and appendices can be found online ².

Statement of Authorship

A statement of authorship for this work can be found in Appendix C.

²<https://ekazakos.github.io/TBN/>

9 | Use What You Have: Video Retrieval Using Representations From Collaborative Experts

Yang Liu* Samuel Albanie* Arsha Nagrani* Andrew Zisserman

Visual Geometry Group, University of Oxford

*Equal Contribution

Abstract

The rapid growth of video on the internet has made searching for video content using natural language queries a significant challenge. Human-generated queries for video datasets ‘in the wild’ vary a lot in terms of degree of specificity, with some queries describing ‘specific details’ such as the names of famous identities, content from speech, or text available on the screen. Our goal is to condense the multi-modal, extremely high dimensional information from videos into a single, compact video representation for the task of video retrieval using free-form text queries, where the degree of specificity is open-ended. For this we exploit existing knowledge in the form of pre-trained semantic embeddings which include ‘general’ features such as motion, appearance, and scene features from visual content. We also explore the use of more ‘specific’ cues from ASR and OCR which are intermittently available for videos and find that these signals remain challenging to use effectively for retrieval. We propose a *collaborative experts* model to aggregate information from these different pre-trained experts and assess our approach empirically on five retrieval benchmarks: MSR-VTT, LSMDC, MSVD, DiDeMo, and ActivityNet. Code and data can be found at www.robots.ox.ac.uk/~vgg/research/collaborative-experts/.

Published in the Proceedings of the British Machine Vision Conference, 2019.

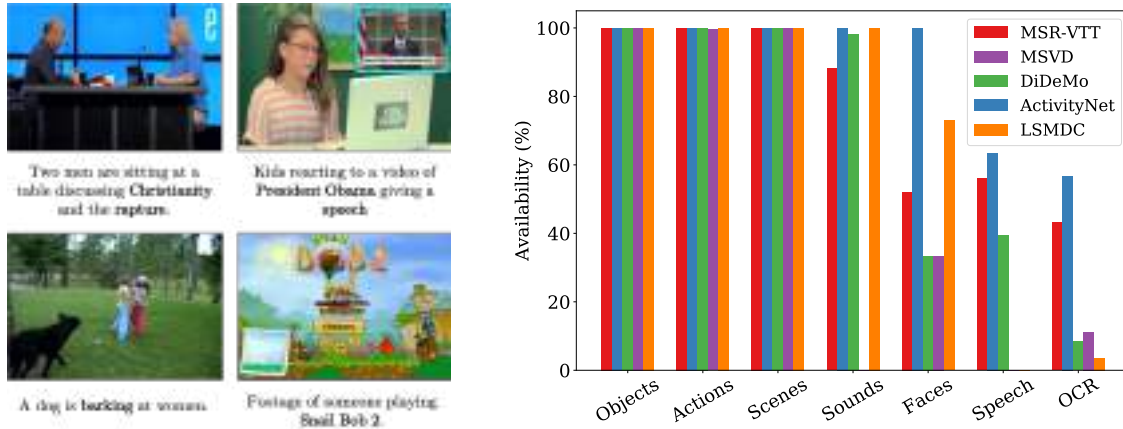


Figure 9.1: (Left): Unconstrained videos ‘in the wild’ convey information in various different ways, including (clockwise from upper-left), clues from distinctive speech, names of individuals on screen, other text clues embedded in the video and audio. (Right): For the five video datasets considered in this work, the chart portrays the video-level availability of “expert” embeddings from different domains (with potentially multiple experts per domain): certain generic embeddings can almost always be extracted via pretrained object/action/scene classification networks. Other features such as sounds, faces, speech and OCR are less consistently available and are more challenging to exploit (Sec. 9.4.3).

9.1 Introduction

Videos capture the world in two important ways beyond a simple image: first, video contains temporal information – semantic concepts, actions and interactions evolve over time; Second, video may also contain information from multiple modalities, such as an accompanying audio track. This makes videos both richer and more informative, but also more challenging to represent. Our goal in this paper is to embed the information from multiple modalities and multiple time steps of a video segment into a compact fixed-length representation. Such a compact representation can then be used for a number of video understanding tasks, such as video retrieval, clustering and summarisation. In particular, we focus on retrieval; our objective is to be able to retrieve video clips using a free form text query that may contain both general and specific information.

Learning a robust and compact representation *tabula rasa* for this task is made extremely

challenging by the high dimensionality of the sensory data contained in videos—to do so with discriminative training would require prohibitively expensive textual annotation of a vast number of videos. The primary hypothesis underpinning our approach is the following: *the discriminative content of the multi-modal video embedding can be well approximated by the set of semantic representations of the video data learnt by individual experts (in audio, scenes, actions, etc).* In essence, this approximation enables us to exploit knowledge from existing individual sources where the cost of annotation is significantly reduced (e.g. classification labels for objects and scenes in images, labels for actions in videos etc.) and where consequently, there exist very large-scale labelled datasets. These large-scale datasets can then be used to train independent experts for different perception tasks, which in turn provide a robust, low-dimensional basis for the discriminative query-content approximation described above.

The two key aspects of this idea that we explore in this paper are: (i) *General and specific features*: in addition to using generic video descriptors (e.g. objects and actions) we investigate encodings of quite specific information from the clip, for example, text from overlaid captions and text from speech to provide effective coverage of the “queryable content” of the video (Fig. 9.1, left). While such features may be highly discriminative for humans, they may not always be available (Fig. 9.1, right) and as we show through experiments (Sec. 9.4.3), making good use of these cues is challenging. We therefore also propose (ii) *Collaborative experts*: a framework that seeks to make effective use of embeddings from different ‘experts’ (e.g. objects, actions, speech) by learning their combination in order to render them more discriminative. Each expert is filtered via a simple dynamic attention mechanism that considers its relation to all other experts to enable their collaboration. This pairwise approach enables, for instance, the sound of a dog barking to inform the modulation of the RGB features, selecting the features that have encoded the concept of the dog. As we demonstrate in the sequel, this idea yields improvements in the retrieval performance.

Concretely, we make the following three contributions: (i) We propose the *Collaborative Experts* framework for learning a joint embedding of video and text by combining a collection

of pretrained embeddings into a single, compact video representation. Our joint video embeddings are independent of the retrieval text-query and can be pre-computed offline and indexed for efficient retrieval; (ii) We explore the use of both *general* video features such as motion, image classification and audio features, and *specific* video features such as text embedded on screen and speech obtained using OCR and ASR respectively. We find that strong generic features deliver good performance, but that specific, rarely available features remain challenging to use for retrieval. (iii) We assess the performance of the representation produced by combining all available cues on a number of retrieval benchmarks, in several cases achieving an advance over prior work.

9.2 Related Work

Cross-Modal Embeddings: A range of prior work has proposed to jointly embed images and text into the same space [Farhadi et al., 2010; Frome et al., 2013; Faghri et al., 2017; Kiros et al., 2014; Nam et al., 2017], enabling cross-modal retrieval. More recently, several works have also focused on audio-visual cross-modal embeddings [Arandjelović & Zisserman, 2017; Nagrani et al., 2018a], as well as audio-text embeddings [Chechik et al., 2008]. Our goal in this work, however, is to embed videos and natural language sentences (sometimes multiple sentences) into the same semantic space, which is made more challenging by the high dimensional content of videos.

Video-Text Embeddings: While a large number of works [Dong et al., 2016; Otani et al., 2016; Pan et al., 2016; Torabi et al., 2016; R. Xu et al., 2015] have focused on learning visual semantic embeddings for video and language, many of these existing approaches are based on image-text embedding methods by design and typically focus on single visual frames. Mithun et al. [Mithun et al., 2018] observe that a simple adaptation of a state-of-the-art image-text embedding method [Faghri et al., 2017] by mean-pooling features from video frames provides a better result than many prior video-text retrieval approaches [Dong et al., 2016; Otani et al.,

2016]. However, such methods do not take advantage of the rich and varied additional information present in videos, including motion dynamics, speech and other background sounds, which may influence the concepts in human captions to a considerable extent. Consequently, there has been a growing interest in fusing information from other modalities—[Mithun et al., 2018; Miech et al., 2018] utilise the audio stream (but do not exploit speech content) and use models pretrained for action recognition to extract motion features. These methods do not make use of speech-to-text or OCR for additional cues, which have nevertheless been used successfully to understand videos in other domains, particularly lecture retrieval [Radha, 2016; Yamamoto et al., 2003] (where the videos consist of slide shows) and news broadcast [Hauptmann et al., 2002] retrieval, where a large fraction of the content is displayed on screen in the form of text. Our approach draws particular inspiration from the powerful joint embedding proposed by [Miech et al., 2018] (which in turn, builds on the classical Mixtures-of-Experts model [Jordan & Jacobs, 1994]) and extends it to investigate additional cues (such as speech and text) and make more effective use of pretrained features via the robust collaborative gating mechanism described in Sec. 9.3.

Annotation scarcity: A key challenge for video-retrieval is the small size of existing training datasets, due to the high cost of annotating videos with natural language. We therefore propose to use the knowledge from existing embeddings pretrained on a wide variety of other tasks. This idea is not new: semantic projections of visual inputs in the form of ‘experts’ was used by [Douze et al., 2011] for the task of image retrieval and has also been central to modern video retrieval methods such as [Miech et al., 2018; Mithun et al., 2018]. More recently, alternative approaches to addressing the issue of annotation scarcity have been explored, which include self-supervised [C. Sun, Myers, et al., 2019] and weakly-supervised [Zhukov et al., 2019] video-text models.

9.3 Collaborative Experts

Given a set of videos with corresponding text captions, we would like to create a pair of functions ϕ_v and ϕ_t that map sensory video data and text into a joint embedding space that respects this correspondence—embeddings for paired text and video should lie close together, while embeddings for text and video that do not match should lie far apart. We would also like ϕ_v and ϕ_t to be independent of each other to enable efficient retrieval: the process of querying then reduces to a distance comparison between the embedding of the query and the embeddings of the collection to be searched (which can be pre-computed offline). The proposed Collaborative Experts framework for learning these functions is illustrated in Fig. 9.2. In this work, we pay particular attention to the design of the video encoder ϕ_v and the process of combining information from different video modalities (Sec. 9.3.1). To complete the framework, we then discuss how the query text is encoded and the ranking loss used to learn the joint embedding space (Sec. 9.3.2).

9.3.1 Video Encoder

To construct the video encoder ϕ_v , we draw on a collection of pretrained, single-modality experts. These operate on the video sensory data \mathbf{v} and project it to a collection of n variable-length task-specific embeddings $\{\Psi_{\text{var}}^{(1)}(\mathbf{v}), \dots, \Psi_{\text{var}}^{(n)}(\mathbf{v})\}$. Here $\Psi_{\text{var}}^{(i)}$ represents the i^{th} expert (we use the var subscript to denote a variable-length output when applied to a sequence of frames) whose parameters have been learned on a prior task such as object classification and then frozen. Each element of this collection is then aggregated along its temporal dimension to produce fixed-size, task-specific embeddings per video $\{\Psi^{(1)}(\mathbf{v}), \dots, \Psi^{(n)}(\mathbf{v})\}$. Any temporal aggregation function may be used here—in this work, we use simple average pooling to aggregate slow visual features such as objects and scenes, and NetVLAD [Arandjelović et al., 2016] to aggregate more dynamic audio and word features (see Sec. 9.4.1 for further details). Next, to enable their combination, we apply linear projections to transform these task-specific

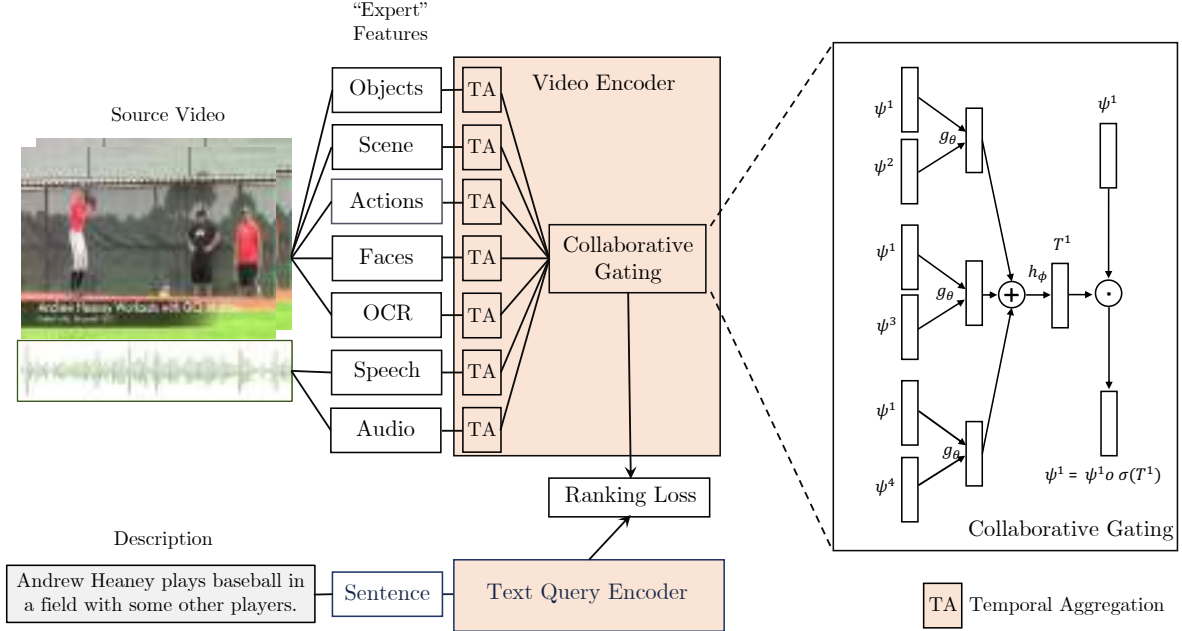


Figure 9.2: (Left): The proposed Collaborative Experts framework for learning a joint video-text embedding (coloured boxes denote learnable parameters). The information provided by each pretrained “expert” (potentially with multiple experts from a single domain) is temporally aggregated as it enters the video encoder and then refined through the use of a collaborative gating mechanism (right) to obtain the video-embedding (for visual clarity, we show the interaction of just a single expert with three others, though in practice all experts are used—see Sec. 9.3.1 for details). Note that to maintain retrieval efficiency, collaboration occurs only between video experts (the text-query and video embeddings are computed independently).

embeddings to a common dimensionality. Our goal when fusing the resulting representations together into a single condensed video representation is to capture the valuable complementary information between task-specific projections while simultaneously filtering out irrelevant noise and resolving individual expert conflicts on a *per-sample basis*. To do so, this we propose a collaborative gating module, described next.

Collaborative Gating: The collaborative gating module comprises two operations: (1) Prediction of attention vectors for every expert projection $T = \{T^{(1)}(\mathbf{v}), \dots, T^{(n)}(\mathbf{v})\}$; and (2) modulation of expert responses. Inspired by the relational reasoning module proposed

by [Santoro et al., 2017b] for visual question answering, we define the attention vector of the i^{th} expert projection T_i as follows:

$$T^{(i)}(\mathbf{v}) = h_\phi\left(\sum_{j \neq i} g_\theta(\Psi^{(i)}(\mathbf{v}), \Psi^{(j)}(\mathbf{v}))\right), \quad (9.1)$$

where functions h_ϕ and g_θ are used to model the pairwise relationship between projection $\Psi^{(i)}$ and projection $\Psi^{(j)}$. Of these, g_θ is used to infer pairwise task relationships, while h_ϕ maps the sum of all pairwise relationships into a single attention vector. In this work, we instantiate both h_ϕ and g_θ as multi-layer perceptrons (MLPs). Note that the functional form of Equation (9.1) dictates that the attention vector of any expert projection should consider the potential relationships between all pairs associated with this expert. That is to say, the quality of each expert $\Psi^{(j)}$ should contribute in determining and selecting the information content from $\Psi^{(i)}$ in the final decision. It is also worth noting that the collaborative gating module uses the same functions g_θ and h_ϕ (shared weights) to compute all pairwise relationships. This mode of operation encourages greater generalisation, since g_θ and h_ϕ are encouraged not to over-fit to features of any particular pair of tasks. After the attention vectors $T = \{T^{(1)}(\mathbf{v}), \dots, T^{(n)}(\mathbf{v})\}$ have been computed, each expert projection is modulated follows:

$$\Psi^{(i)}(\mathbf{v}) = \Psi^{(i)}(\mathbf{v}) \circ \sigma(T^{(i)}(\mathbf{v})), \quad (9.2)$$

where σ is an element-wise sigmoid activation and \circ is the element-wise multiplication (Hadamard product). This gating function re-calibrates the strength of different activations of $\Psi^{(i)}(\mathbf{v})$ and selects which information is highlighted or suppressed, providing the model with a powerful mechanism for dynamically filtering content from different experts. A diagram of the mechanism is shown in Fig. 9.2 (right). The final video embedding is then obtained by passing the modulated responses of each expert through a Gated Embedding Module (GEM) [Miech et al., 2018] (note that this operation produces l2-normalized outputs) before concatenating the

outputs together into a single fixed-length vector.

9.3.2 Text Query Encoder and Training Loss

To construct the text embeddings, query sentences are first mapped to a sequence of feature vectors with pretrained contextual word-level embeddings (see Sec. 9.4.1 for details)—as with the video experts, the parameters of this first stage are frozen. These are then aggregated, again using NetVLAD [Arandjelović et al., 2016]. Following aggregation, we follow the text encoding architecture proposed by [Miech et al., 2018], which projects the aggregated features to separate subspaces for each expert using GEMs (as with the video encoder, producing l2-normalized outputs). Each projection is then scaled by a mixture weight (one scalar weight per expert projection), which is computed by applying a single linear layer to the aggregated text-features, and passing the result through a softmax to ensure that the mixture weights sum to one (see [Miech et al., 2018] for further details). Finally, the scaled outputs are concatenated, producing a vector of dimensionality that matches that of the video embedding.

With the video encoder ϕ_v and text encoder ϕ_t as described, the similarity s_i^j of the i^{th} video, \mathbf{v}_i , and the j^{th} caption, \mathbf{t}_j , can then be directly computed as the cosine of the angle between their respective embeddings $\phi_v(\mathbf{v}_i)^T \phi_t(\mathbf{t}_j)$. During optimisation, the parameters of the video encoder (including the collaborative gating module) and text query encoder (the coloured regions of Fig. 9.2) are learned jointly. Training proceeds by sampling a sequence of minibatches of corresponding video-text pairs $\{\mathbf{v}_i, \mathbf{t}_i\}_{i=1}^{N_B}$ and minimising a *Bidirectional Max-margin Ranking Loss* [Socher et al., 2014]:

$$\mathcal{L}_r = \frac{1}{N_B} \sum_{i=1, j \neq i}^{N_B} \max(0, m + s_i^j - s_i^i) + \max(0, m + s_j^i - s_j^j) \quad (9.3)$$

where N_B is the batch size, and m is a fixed constant which is set as a hyperparameter. When

assessing retrieval performance, at test time the embedding distances are simply computed via their inner product, as described above.

9.3.2.1 Missing Experts

When a set of expert features are missing, such as when there is no speech in the audio track, we simply zero-pad the missing experts when estimating the similarity score. To compensate for the implicit scaling introduced by missing experts (the similarity is effectively computed between shorter embeddings), we follow the elegant approach proposed by [Miech et al., 2018] and simply remove the mixture weights for missing experts, then renormalise the remaining weights such that they sum to one.

9.4 Experiments

In this section, we evaluate our model on five benchmarks for video retrieval tasks. The description of datasets, implementation details and evaluation metric are provided in Sec. 9.4.1. A comprehensive comparison on general video retrieval benchmarks is reported in Sec. 9.4.2. We present an ablation study in Sec. 9.4.3 to explore how the performance of the proposed method is affected by different model configurations, including the aggregation methods, importance of different experts and number of captions in training.

9.4.1 Datasets, Implementation Details and Metrics

Datasets: We perform experiments on five video datasets: MSR-VTT [J. Xu et al., 2016], LSMDC [Rohrbach et al., 2015], MSVD [D. L. Chen & Dolan, 2011], DiDeMo [Anne Hendricks et al., 2017] and ActivityNet-captions [Krishna et al., 2017], covering a challenging set of domains which include videos from YouTube, personal collections and movies.

Expert Features: In order to capture the rich content of a video, we draw on existing powerful representations for a number of different semantic tasks. These are first extracted at a

frame-level, then aggregated to produce a single feature vector per modality per video. *RGB object* frame-level embeddings of the visual data are generated with two models: an SENet-154 model [J. Hu et al., 2019] (pretrained on ImageNet for the task of image classification), and a ResNext-101 [S. Xie et al., 2017] pretrained on Instagram hashtags [Mahajan et al., 2018]. *Motion* embeddings are generated using the I3D inception model [Carreira & Zisserman, 2017] and a 34-layer R(2+1)D model [Tran et al., 2018] trained on IG-65m [Ghadiyaram et al., 2019]. *Face* embeddings are extracted in two stages: (1) Each frame is passed through an SSD face detector [W. Liu et al., 2016] to extract bounding boxes; (2) The image region of each box is passed through a ResNet50 [K. He et al., 2016] that has been trained for the task of face classification on the VGGFace2 dataset [Cao et al., 2018]. *Audio* embeddings are obtained with a VGGish model, trained for audio classification on the YouTube-8m dataset [Hershey et al., 2017]. *Speech-to-Text* features are extracted using the Google Cloud speech API, to extract word tokens from the audio stream, which are then encoded via pretrained word2vec embeddings [Mikolov et al., 2013]. *Optical Character Recognition* is done in two stages: (1) Each frame is passed through the Pixel Link [D. Deng et al., 2018] text detection model to extract bounding boxes for text; (2) The image region of each box is passed through a model [Y. Liu et al., 2018] that has been trained for scene text recognition on the Synth90K dataset [Jaderberg et al., 2014]. The text is then encoded via a pretrained word2vec embedding model [Mikolov et al., 2013].

Temporal Aggregation: We adopt a simple approach to aggregating the features described above. For appearance, motion, scene and face embeddings, we average frame-level features along the temporal dimension to produce a single feature vector per video (we found max-pooling to perform similarly). For speech, audio and OCR features, we adopt the NetVLAD mechanism proposed by [Arandjelović et al., 2016], which has proven effective in the retrieval setting [Miech, Laptev, & Sivic, 2017]. As noted in Sec. 9.3.1, all aggregated features are projected to a common size (768 dimensions).

Text: Each word is encoded using pretrained word2vec word embeddings [Mikolov et al.,

2013] and then passed through a pretrained OpenAI-GPT model [Radford et al., 2018] to extract contextual word embeddings. Finally, the word embeddings in each sentence are aggregated using NetVLAD.

Dataset-specific details: Except where noted otherwise for ablation purposes, we use each of the embeddings described above for the MSR-VTT, ActivityNet and DiDeMo datasets. For MSVD, we extract the subset of features which do not require an audio stream (since no audio is available with the dataset). For LSMDC, we re-use the existing face, text and audio features made available by [Miech et al., 2018], and combine them with the remaining features described above.

Training Details: The CE framework is implemented with PyTorch [Paszke et al., 2017]. Optimisation is performed with the Lookahead solver [Kingma & Ba, 2014] in combination with RAdam [L. Liu et al., 2019] (implementation by [Wright, 2019]).

Evaluation Metrics: We follow prior work (e.g. [Dong et al., 2016; B. Zhang et al., 2018; Mithun et al., 2018; Y. Yu et al., 2018; Miech et al., 2018]) and report standard retrieval metrics (where existing work enables comparison) including median rank (lower is better), mean rank (lower is better) and R@K (recall at rank K—higher is better). When computing video-to-sentence metrics for datasets with multiple independent sentences per video (MSR-VTT and MSVD), we follow the evaluation protocol used in prior work [Mithun et al., 2018; Dong et al., 2018, 2019] which corresponds to reporting the minimum rank among all valid text descriptions for a given video query. For each benchmark, we report the mean and standard deviation of three randomly seeded runs.

9.4.2 Comparison to Prior State-of-the-Art

We first compare the proposed method with the existing state-of-the-art on the MSR-VTT benchmark for the tasks of sentence-to-video and video-to-sentence retrieval Tab. 9.1. Driven by strong expert features, we observe that Collaborative Experts (CE) consistently improves retrieval performance for both sentence and video queries. We next evaluate the performance

Method	Test-set	Text \Rightarrow Video					Video \Rightarrow Text				
		R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
JSFusion	1k-A	10.2	31.2	43.2	13	-	-	-	-	-	-
CE	1k-A	20.9 ± 1.2	48.8 ± 0.6	62.4 ± 0.8	6 ± 0	28.2 ± 0.8	20.6 ± 0.6	50.3 ± 0.5	64.0 ± 0.2	5.3 ± 0.6	25.1 ± 0.8
MoEE	1k-B	13.6	37.9	51.0	10	-	-	-	-	-	-
MoEE _{COCO}	1k-B	14.2	39.2	53.8	9	-	-	-	-	-	-
CE	1k-B	18.2 ± 0.7	46.0 ± 0.4	60.7 ± 0.2	7 ± 0	35.3 ± 1.1	18.0 ± 0.8	46.0 ± 0.5	60.3 ± 0.5	6.5 ± 0.5	30.6 ± 1.2
VSE	Full	5.0	16.4	24.6	47	215.1	7.7	20.3	31.2	28	185.8
VSE++	Full	5.7	17.1	24.8	65	300.8	10.2	25.4	35.1	25	228.1
Mithun et al.	Full	7.0	20.9	29.7	38	213.8	12.5	32.1	42.4	16	134.0
W2VV	Full	6.1	18.7	27.5	45	-	11.8	28.9	39.1	21	-
Dual Enc.	Full	7.7	22.0	31.8	32	-	13.0	30.8	43.3	15	-
E2E	Full	9.9	24.0	32.4	29.5	-	-	-	-	-	-
CE	Full	10.0 ± 0.1	29.0 ± 0.3	41.2 ± 0.2	16 ± 0	86.8 ± 0.3	15.6 ± 0.3	40.9 ± 1.4	55.2 ± 1.0	8.3 ± 0.6	38.1 ± 1.8

Table 9.1: Retrieval with sentences and videos on the MSR-VTT dataset. R@k denotes recall@k (higher is better), MdR and MnR denote median rank and mean rank resp. (lower is better). Standard deviations are reported from three randomly seeded runs. 1k-A and 1k-B denote test sets of 1000 randomly sampled text-video pairs used by [Y. Yu et al., 2018] and [Miech et al., 2018] resp. **Citations:** JSFusion [Y. Yu et al., 2018], MoEE [Miech et al., 2018], MoEE_{COCO} [Miech et al., 2018], VSE [Mithun et al., 2018], VSE++ [Mithun et al., 2018], Mithun et al. [Mithun et al., 2018], W2VV [Dong et al., 2018], Dual Enc. [Dong et al., 2019], E2E [Miech, Alayrac, et al., 2019]

of the CE framework on the LSMDC benchmark for sentence-to-video retrieval (Tab. 9.2, left) and observe that CE matches or outperforms all prior work, including the prior state-of-the-art method [Miech et al., 2018] which incorporates additional training images and captions from the COCO benchmark during training, but uses fewer experts. We observe similar trends in the results for the MSVD retrieval benchmark (Tab. 9.2, right). In Tab. 9.4, we compare with prior work on the ActivityNet paragraph-video retrieval benchmark (note that we compare to methods which use the same level of annotation as our approach i.e. video-level annotation), and see that CE is competitive. Finally, in Tab. 9.3 we provide a comparison with previously reported numbers on the DiDeMo benchmark and see that CE again outperforms prior work.

9.4.3 Ablation Studies

In this section, we provide ablation studies to empirically assess: (1) the effectiveness of the proposed collaborative experts framework vs other aggregation strategies; (2) the importance

Method	Text \Rightarrow Video			
	R@1	R@5	R@10	MdR
Yu et al. [Y. Yu et al., 2016] [†]	3.6	14.7	23.9	50
CCA [Klein et al., 2015] (rep. by [Miech et al., 2018])	7.5	21.7	31.0	33
JSFusion [Y. Yu et al., 2018] [‡]	9.1	21.2	34.1	36
MoEE [Miech et al., 2018]	9.3	25.1	33.4	27
MoEE _{COCO} [Miech et al., 2018]	10.1	25.6	34.6	27
CE	11.2\pm0.4	26.9\pm1.1	34.8\pm2.0	25.3\pm3.1

Method	Text \Rightarrow Video				
	R@1	R@5	R@10	MdR	MnR
CCA ([R. Xu et al., 2015])	-	-	-	-	245.3
JMDV [R. Xu et al., 2015]	-	-	-	-	236.3
VSE [Kiros et al., 2014]	12.3	30.1	42.3	14	57.7
VSE++ [Faghri et al., 2017]	15.4	39.6	53.0	9	43.8
Multi. Cues [Mithun et al., 2018]	20.3	47.8	61.1	6	28.3
CE	19.8 \pm 0.3	49.0 \pm 0.3	63.8 \pm 0.1	6 \pm 0.0	23.1 \pm 0.3

Table 9.2: Text-to-Video retrieval results on the LSMDC dataset (top) and the MSVD dataset (bottom). [†], [‡] denote the winners of the 2016 and 2017 LSMDC challenges, respectively.

Method	Text \Rightarrow Video					Video \Rightarrow Text				
	R@1	R@5	R@50	MdR	MnR	R@1	R@5	R@50	MdR	MnR
S2VT	11.9	33.6	76.5	13	-	13.2	33.6	76.5	15	-
FSE	13.9 \pm 0.7	36 \pm 0.8	78.9 \pm 1.6	11	-	13.1 \pm 0.5	33.9 \pm 0.4	78.0 \pm 0.8	12	-
CE	16.1\pm1.4	41.1\pm0.4	82.7\pm0.3	8.3\pm0.6	43.7\pm3.6	15.6\pm1.3	40.9\pm0.4	82.2\pm1.3	8.2\pm0.3	42.4\pm3.3

Table 9.3: Comparison of paragraph-video retrieval methods trained with video-level information on the DiDeMo dataset. **Citations:** S2VT [Venugopalan et al., 2014] ([B. Zhang et al., 2018]), FSE [B. Zhang et al., 2018]

of using of a diverse range of experts with differing levels of specificity; (3) the relative value of using experts in comparison to simply having additional annotated training data.

Aggregation method: We compare the use of collaborative experts with several other baselines (with access to the same experts) for embedding aggregation including: (1) simple expert concatenation; (2) CE without projecting to a common dimension, without mixture weights and without the collaborative gating module described in Sec. 9.3.1; (3) the state of the art MoEE [Miech et al., 2018] method (equivalent to CE without the common projection and collaborative gating) and (4) CE without collaborative gating. The results, presented in Tab. 9.6

Method	Text \Rightarrow Video					Video \Rightarrow Text				
	R@1	R@5	R@50	MdR	MnR	R@1	R@5	R@50	MdR	MnR
LSTM-YT)	0.0	4.0	24.0	102	-	0.0	7.0	38.0	98	-
NOCTXT)	5.0	14.0	32.0	78	-	7.0	18.0	45.0	56	-
DENSE	14.0	32.0	65.0	34	-	18.0	36.0	74.0	32	-
FSE	18.2 \pm 0.2	44.8 \pm 0.4	89.1 \pm 0.3	7	-	16.7 \pm 0.8	43.1 \pm 1.1	88.4 \pm 0.3	7	-
HSE(4SEGS) [†]	20.5	49.3	-	-	-	18.7	48.1	-	-	-
CE	18.2 \pm 0.3	47.7 \pm 0.6	91.4 \pm 0.4	6 \pm 0	23.1 \pm 0.5	17.7 \pm 0.6	46.6 \pm 0.7	90.9 \pm 0.2	6 \pm 0	24.4 \pm 0.5

Table 9.4: Comparison of paragraph-video retrieval methods trained with video-level information on the ActivityNet-captions dataset (val1 test-split). **Citations:** LSTM-YT [Venugopalan et al., 2015] ([B. Zhang et al., 2018]), NOCTXT [Venugopalan et al., 2014] ([B. Zhang et al., 2018]), DENSE [Krishna et al., 2017], FSE [B. Zhang et al., 2018], HSE(4SEGS) [B. Zhang et al., 2018]

(left), demonstrate the contribution of collaborative gating which improves performance and leads to a more efficient parameterisation than the prior state of the art.

Importance of different experts: The value of different experts is assessed in Tab. 9.5 (note that since several experts are not present in all videos, we combine them with features produced by a scene expert pretrained on Places365 [B. Zhou et al., 2017]—the expert with the lowest performance that is consistently available as a baseline to enable a more meaningful comparison). There is considerable variance in the effect produced by different choices of expert. Using stronger features within a given modality (pretraining on Instagram [Mahajan et al., 2018] rather than Kinetics [Carreira & Zisserman, 2017] (resp. ImageNet) [J. Deng et al., 2009] for actions (resp. object) experts can yield a significant boost in performance). The cues from scarce features (such as speech, face and OCR) which are often missing from videos (see Fig. 9.1, right) provide significantly weaker cues and bring a limited improvement to performance when used in combination.

Number of Captions in training: An emerging idea in our community is that many machine perception tasks might be solved through the combination of simple models and large-scale training sets, reminiscent of the “big-data” hypothesis [Halevy et al., 2009]. In this section, we perform an ablation study to assess the relative importance of access to pretrained experts

Experts	Text \Rightarrow Video				
	R@1	R@5	R@10	MdR	MnR
Scene	4.0 \pm 0.1	14.1 \pm 0.1	22.4 \pm 0.3	50.0 \pm 1.0	201.3 \pm 1.6
Scene+Speech	4.6 \pm 0.1	15.5 \pm 0.2	24.4 \pm 0.2	44.7 \pm 1.2	183.6 \pm 1.7
Scene+Audio	5.6 \pm 0.0	18.7 \pm 0.1	28.2 \pm 0.1	33.7 \pm 0.6	140.8 \pm 0.3
Scene+Action (KN)	5.3 \pm 0.3	17.6 \pm 0.8	27.1 \pm 0.9	36.0 \pm 1.7	158.7 \pm 1.6
Scene+Obj (IN)	5.0 \pm 0.2	16.6 \pm 0.7	25.5 \pm 1.0	40.7 \pm 2.1	173.1 \pm 3.3
Scene+Obj(IG)	7.2 \pm 0.1	22.3 \pm 0.3	33.0 \pm 0.2	25.3 \pm 0.6	125.1 \pm 0.1
Scene+Action (IG)	6.8 \pm 0.1	21.7 \pm 0.1	32.4 \pm 0.1	25.7 \pm 0.6	122.1 \pm 0.3
Scene+OCR	4.1 \pm 0.1	14.1 \pm 0.1	22.2 \pm 0.2	50.3 \pm 1.2	203.1 \pm 4.4
Scene+Face	4.1 \pm 0.1	14.2 \pm 0.3	22.4 \pm 0.4	49.7 \pm 0.6	194.2 \pm 5.1

Experts	Text \Rightarrow Video				
	R@1	R@5	R@10	MdR	MnR
Scene	4.0 \pm 0.1	14.1 \pm 0.1	22.4 \pm 0.3	50.0 \pm 1.0	201.3 \pm 1.6
Prev.+Speech	4.6 \pm 0.1	15.5 \pm 0.2	24.4 \pm 0.2	44.7 \pm 1.2	183.6 \pm 1.7
Prev.+Audio	5.8 \pm 0.1	19.0 \pm 0.3	28.8 \pm 0.2	32.3 \pm 0.6	136.8 \pm 1.2
Prev.+Action (KN)	6.7 \pm 0.2	21.8 \pm 0.4	32.5 \pm 0.5	25.3 \pm 0.6	115.9 \pm 1.0
Prev.+Obj (IN)	7.5 \pm 0.1	23.4 \pm 0.0	34.1 \pm 0.2	23.7 \pm 0.6	111.9 \pm 0.6
Prev.+Obj (IG)	9.5 \pm 0.2	27.7 \pm 0.1	39.4 \pm 0.1	18.0 \pm 0.0	92.6 \pm 0.4
Prev.+Action (IG)	9.9 \pm 0.1	28.6 \pm 0.3	40.7 \pm 0.1	17.0 \pm 0.0	86.4 \pm 0.4
Prev.+OCR	10.0 \pm 0.1	28.8 \pm 0.2	40.9 \pm 0.2	16.7 \pm 0.6	87.3 \pm 0.8
Prev.+Face	10.0 \pm 0.1	29.0 \pm 0.3	41.2 \pm 0.2	16.0 \pm 0.0	86.8 \pm 0.3

Table 9.5: **The importance of different experts** (Top): The value of different experts in combination with a baseline set for text-video retrieval and (bottom) their cumulative effect on MSR-VTT (here Prev. denotes the experts used in the previous row).

and additional video description annotations. To do so, we measure the performance of the CE model as we vary (1) the number of descriptions available per-video during training and (2) the number of experts it has access to. The results are shown in Tab. 9.6 (right). We observe that increasing the number of training captions per-video from 1 to 20 brings an improvement in performance, approximately comparable to adding the full collection of experts, suggesting that indeed, adding experts can help to compensate for a paucity of labelled data. When multiple captions and multiple experts are both available, they naturally lead to the most robust embedding. Some qualitative examples of videos retrieved by the multiple-expert, multiple-caption system are provided in Fig. 9.3.

Aggreg.	R@1	R@5	R@10	MdR	Params
Concat	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0	1495.5 \pm 0.0	369.72k
CE - MW,P,CG	8.5 \pm 0.1	25.9 \pm 0.3	37.6 \pm 0.2	19.0 \pm 0.0	246.22M
MoEE [Miech et al., 2018]	9.6 \pm 0.1	28.0 \pm 0.2	39.7 \pm 0.2	17.7 \pm 0.6	400.41 M
CE - CG	9.7 \pm 0.1	28.1 \pm 0.2	40.2 \pm 0.1	17.0 \pm 0.0	181.07 M
CE	10.0 \pm 0.1	29.0 \pm 0.3	41.2 \pm 0.2	16.0 \pm 0.0	183.45 M

Expert	Num. Captions	R@1	R@5	R@10	MdR
Obj(IN)	1	2.6 \pm 0.1	9.3 \pm 0.4	15.0 \pm 0.7	101.3 \pm 15.5
Obj(IN)	20	4.9 \pm 0.1	16.5 \pm 0.2	25.3 \pm 0.4	40.7 \pm 1.2
All	1	4.8 \pm 0.2	16.2 \pm 0.5	25.0 \pm 0.7	43.3 \pm 4.0
All	20	10.0 \pm 0.1	29.0 \pm 0.3	41.2 \pm 0.2	16.0 \pm 0.0

Table 9.6: (Top): Aggregation methods for text-video retrieval on MSR-VTT; (Bottom): The relative value of training with additional captions vs the value of experts.

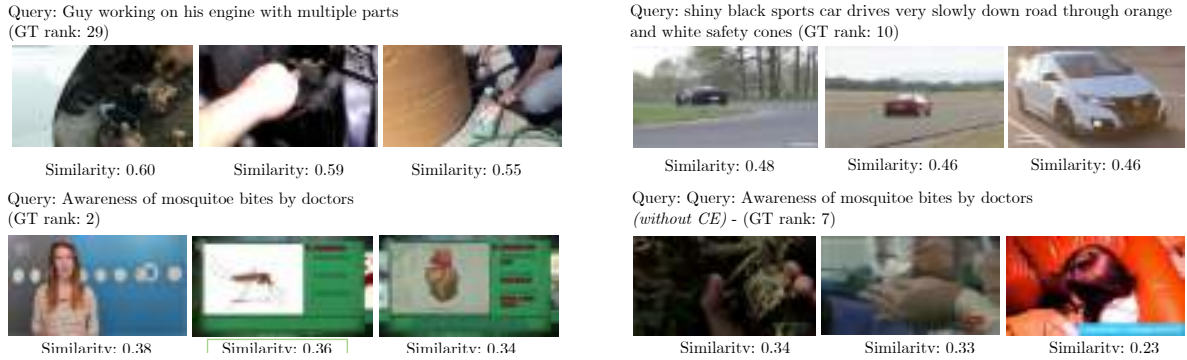


Figure 9.3: **Qualitative Results on MSR-VTT:** For each query, we show frames from the top three ranked videos (where present, the ground truth video is indicated by a green box around the similarity score). Top row: (left) Even for imperfect rankings, the model retrieves reasonable videos; Failure case (right) the embeddings can fail to differentiate between certain signals (in this case, ranking cars of the wrong colour above the ground truth video). Bottom row: (left) the videos retrieved by the proposed model (which assigns its second highest similarity to the correct video); (right) removing the proposed CE component produces a noisier ranking.

9.5 Conclusion

In this work, we introduced collaborative experts, a framework for learning a joint video-text embedding for efficient retrieval. We have shown that using a range of pretrained features and combining them through an appropriate gating mechanism can boost retrieval performance. In

future work, we plan to explore the use of collaborative experts for other video understanding tasks such as clustering and summarisation.

Acknowledgements: Funding for this research is provided by the EPSRC Programme Grant Seebibyte EP/M013774/1 and EPSRC grant EP/R03298X/1. A.N. is supported by a Google PhD Fellowship. We would like to thank Antoine Miech, YoungJae Yu and Bowen Zhang for their assistance with experiment details. We would like to particularly thank Valentin Gabeur for identifying a bug in the software implementation that was responsible for the inaccurate results reported in the initial version of the paper. We would also like to thank Zak Stone and Susie Lim for their help with cloud computing.

Additional Appendices

Code, additional results and appendices can be accessed online ¹.

Statement of Authorship

A statement of authorship for this work can be found in Appendix C.

¹<https://www.robots.ox.ac.uk/vgg/research/collaborative-experts/>

10 | Discussion

In this chapter we first summarise the main achievements and highlight the impact of the work presented in this thesis (Section 10.1) and then suggest avenues for future work (Section 10.2).

10.1 Achievements and Impact

VoxCeleb: In chapter 2, we proposed a scalable audio-visual pipeline to create a large dataset of human speech from YouTube videos, called VoxCeleb. VoxCeleb has been released in two installments (1 and 2), and consists of over 2000 hours of audio-visual human speech in the wild. We have publicly released speech segments, identity annotation and face-tracks¹.

The VoxCeleb datasets have spurred research in a number of different audio-visual domains, including cross-modal retrieval [Wen et al., 2018; Y. Chen et al., 2019; S.-W. Chung, Chung, & Kang, 2020], unsupervised landmark learning [Jakab et al., 2020; Wiles et al., 2018], source localisation and speech enhancement [Afouras et al., 2018; Owens & Efros, 2018], and visual speech synthesis [Jamaludin et al., 2019; H. Zhou et al., 2019], to name a few. Their wide use by the research community is evidenced by the fact that the datasets have been downloaded over 5,000 times and our publications have been jointly cited over 800 times.

However, by far the greatest impact has been on the field of unconstrained speaker recognition. As mentioned by [K. A. Lee et al., 2019], VoxCeleb represents a “new initiative towards speaker recognition in the wild”. By being orders of magnitude larger than traditional speaker recognition datasets, VoxCeleb has allowed the development of new deep neural networks models [Snyder et al., 2017; Okabe et al., 2018; Safari & Hernando, 2019; Z. Wang et al., 2020] for speaker recognition, enabling a paradigm shift in the community from hand crafted features to deep learning models that are trained end-to-end. International challenges in speaker recognition have also now adopted these datasets as the gold standard for train-

¹All data is available here: <http://www.robots.ox.ac.uk/~vgg/data/voxceleb/>

ing and evaluating new methods – for example, they are now an integral part of the National Institute of Standards and Technology (US Department of Commerce) Speaker Recognition challenges held annually (NIST-SRE)², as well as the main subject of the VoxCeleb Speaker Recognition Challenges (VoxSRC)³ also held annually. Our first paper on VoxCeleb1 was also awarded the Best Student Paper Award at Interspeech 2017.

Cross-Modal Learning for Action Recognition: Obtaining manual annotation for action recognition in videos is a notoriously expensive task. In chapter 3, we provide an alternative for manual supervision, by predicting action labels directly from the speech. By applying this approach to thousands of movies and TV shows, we are the first to predict actions solely from speech in the film domain, exceeding the performance of fully supervised action classification on the AVA dataset.

Cross-Modal Distillation for Emotion Recognition: In chapter 4, we trained a network to distill knowledge from a pre-trained facial emotion classifier to a speech emotion recognition model using unlabelled videos as a bridge. Such knowledge transfer from one modality to another is useful for a task like emotion recognition, where the size of face emotion datasets is much larger (two orders of magnitude) than data labelled in the speech modality. Because of the inherent ambiguity in the emotion label space, we distill information from the logits of the face model using temperature. This idea of using faces as a source of supervision has been adopted by a number of works attempting emotion recognition in speech [J. Han et al., 2019; G. He et al., 2020]; while [J. Han et al., 2019] use both redundant and complementary cross-modal information from faces to improve emotion recognition in speech, [G. He et al., 2020] propose a semi-supervised adversarial network that allows knowledge transfer from large labeled face datasets to audio datasets with only a few labels.

Face-Voice Cross-Modal Representation Learning: In chapters 5 and 6, we proposed a new face-voice matching task, where we showed that it is possible to learn a joint representation

²Further details can be found at <https://sre.nist.gov/>

³<http://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition.html>

for identity from both face and voice inputs. We learnt this embedding from videos of human speech with no identity labels, using a form of cross-modal self-supervision. This idea received great interest from the research community, prompting others to develop newer architectures and losses to solve the task [Wen et al., 2018; Y. Chen et al., 2019; S.-W. Chung, Chung, & Kang, 2020] on the benchmark we provided (data from the VoxCeleb and VGGFace datasets). Others extended our work to their own different datasets, creating the FVCeleb dataset with 1k more identities from the MS-Celeb-1M dataset [Horiguchi et al., 2018], as well as a voice-face dataset (with 1.15M face images and 0.29M audio segments) from Chinese speakers [Xiong et al., 2019], and then performed the task of face-voice matching on this new data. [S.-W. Chung, Kang, & Chung, 2020] extended our work in Chapter 7 by adding an additional training constraint that not only optimises metrics across modalities, but also enforces intra-class feature separation within either of the modalities.

We also developed a curriculum learning schedule for hard negative mining that we show is essential for learning to proceed successfully in this self-supervised setting, a concept which has also been successfully shown to be effective for other self-supervised works [Wiles et al., 2018].

After we demonstrated that it was possible to retrieve faces from just voices or vice versa, this naturally begs the question of whether it is possible to generate a face given just the voice. This was explored in detail by a group at MIT in their CVPR 2019 paper [Oh et al., 2019], and then further at ICLR 2020 [Choi et al., 2020]. Interestingly, others also attempted to generate a *3D face* from just the voice [R. Singh, 2019], while a rather unusual work also investigated the face humans visualise when they hear a *robot* (artificially generated voice) speaking [McGinn & Torre, 2019]. Finally, we also note that our work on cross-modal retrieval inspired others to apply the same architecture to different modalities, including to remote sensing image and spoken audio [Mao et al., 2018], as well as images and touch [J. Lin et al., 2019].

Modality Fusion: In chapter 8, we proposed a new audio-visual fusion neural network archi-

ture that combines audio, RGB frames and optical flow inputs within a range of temporal offsets for the task of action recognition. Unlike previous audio visual fusion methods, we fuse modalities *before* temporal aggregation, with shared modality and fusion weights over time. Since frames of different input modalities such as audio and text are never perfectly aligned, this necessitates the need for a temporal binding window, to allow for some leniency in temporal alignment of input features. Our proposed architecture is trained end-to-end, outperforming individual modalities as well as late-fusion of modalities. This idea was also explored by AudioSlowFast [Xiao et al., 2020] where they fuse an Audio pathway into the output of a SlowFast fusion [Feichtenhofer et al., 2019] architecture, which is coarser in temporal resolution. They claim that they adopt this design as it "imposes a less stringent requirement on temporal alignment between audio and visual features" which they found to be important in their experiments. [Ryu & Kim, 2020] also apply the idea of a temporal binding window to fusion, albeit for the classification of time-series multirate sensor measurements such as accelerations and audio recordings.

Our work has also impacted the growing interest in egocentric action recognition. We have released all code and models publicly⁴, and noticed that there has been a renewed interest in the area, with a number of works focusing on new architectures for egocentric action recognition [Furnari & Farinella, 2020; Y. Li et al., 2020; X. Wang et al., 2020].

In Chapter 9, we designed and implemented a fusion method for pre-extracted ‘expert’ embeddings from video. Our method consists of pairwise gating functions, allowing each modality embedding to adjust its values based on the embeddings of other modalities. This work set the new state-of-the-art results on *five* different datasets for video-text retrieval, covering a wide range of domains.

Recent work [Gabeur et al., 2020] builds upon our framework by incorporating self-attention layers to reason about cross-modal relationships, showing performance gains. A method based

⁴<https://github.com/ekazakos/temporal-binding-network>

on this work [Y. Liu et al., 2019] also won the second place at the Multi Moments in Time challenge. Additionally, our work formed the basis of the Video Pentathlon challenge held at CVPR 2020⁵.

10.2 What Comes Next?

Our ultimate goal is to be able to retrieve human-centric videos with flexible natural language queries, as well as understand long-range narratives in human stories using multimodal learning. In this section we take a step back, and outline some broader challenges in these areas, as well as describe some preliminary progress in overcoming these challenges.

10.2.1 Audio-Visual Parsing

In this thesis we show progress in transferring supervision from one modality to another (Chapters 2 to 4), learning cross-modal representations with self-supervision (Chapters 5 to 7), and modality fusion (Chapters 8 and 9). In most cases, we assume that the audio and visual data are always correlated, and sometimes even more strictly, temporally aligned within a certain temporal window. In practice, however, many videos have audible sounds, which originate outside of the frame of view (leaving no visual correspondences) but are still useful for video understanding. Examples include out-of-screen running cars, or the narration of a person.

This poses the question, is it possible to automatically determine which events in a video are audible, visible, and both audible and visible, and hence can be used as a learning source? This task is referred to as audio-visual parsing [Tian et al., 2020]. Once events are identified as being present in a given modality stream (events bound to sensory modalities), we can then analyse cross modal relationships further, i.e. by performing the spatial and temporal

⁵<https://www.robots.ox.ac.uk/~vgg/challenges/video-pentathlon/challenge.html>

localisation of such events in the video.

For example, in all our cross-modal tasks involving human faces and voices (Chapters 2 to 7), we always assume that we have access to single-speaker segments, where both the face and voice of the speaker are present in the video. In the real world, human speech often consists of challenging multispeaker segments, where the faces of speakers are sometimes not visible (off-screen narration, flashbacks, reaction shots etc). One way to overcome this problem is via perfect audio-visual diarisation, where diarisation refers to the task of breaking up multispeaker video into homogeneous single speaker segments, effectively solving ‘who is speaking when’.

Diarisation involves noting when humans are speaking, and then recognising them either by their face or their voice. Beyond being an interesting research problem in itself, it is also a valuable pre-processing step for a number of applications, including speech-to-text. To encourage research in this field, we make preliminary progress in developing an audio-visual semi-automatic diarisation pipeline. We use it to curate VoxConverse [J. S. Chung et al., 2020], an audio-visual diarisation dataset from YouTube. The speech for on-screen identities is accurately segmented automatically using active speaker detection, identified using face recognition, and further enhanced using an audio-visual speech enhancement model [Afouras et al., 2018] to better isolate and identify speaker identities. Finally, to accurately recognise *off-screen* speakers, we utilise state of the art speaker recognition embeddings that verify identities from audio alone. This will allow us to apply some of the multimodal techniques developed in this thesis to more general, noisy, multispeaker videos ‘in the wild’.

10.2.2 Richer Long-Range Temporal Context

Most of the concepts in this thesis have been applied to short video segments, of less than a minute. A long standing goal in computer vision is the ability to understand videos that cover longer timescales.

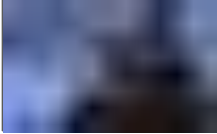
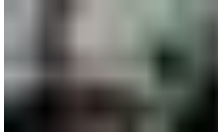



INTENT	RELATIONSHIP	EMOTION	PAST CONTEXT	FUTURE CONTEXT
				
Sean wants to sulk alone in the snow but is interrupted by Paul.	Adam meets his therapist Katherine, who is much younger than he expected.	Barbara Jean succumbs to her stress and anxiety on stage.	Ronny tries to get his camera back from Zip who is still angry about their previous altercation .	Frankie reveals his master plan to steal \$10,000 from charity, and how the group of kids will be used to help him

Figure 10.1: **Semantic descriptions:** Examples of high level semantic descriptions accompanying each video clip in the Condensed Movies Dataset. The semantic descriptions cover a number of high level concepts, including intent/motivation, relationships, emotions and attributes, and context from surrounding clips in the storyline.

An ideal source of data for long range video understanding is movie and TV show material – the high level understanding of a movie narrative requires understanding characters’ identities, motivations and behaviour over an extended period of time. Since video architectures struggle to learn over such large timescales (movies are often roughly 2 hours long), we first create a dataset from the MovieClips channel on YouTube⁶, which contains the key scenes or clips from numerous movies and also contains a semantic text description that describes the content of each clip. At 2-3 mins in length, these clips are still much longer than the input resolution of current CNN architectures. We obtain roughly 10 clips per movie, condensing a movie down to about 20 mins of video data. Our dataset is hence called the Condensed Movie Dataset (CMD) [Bain et al., 2020]. This dataset can be used for a number of tasks related to longer term video understanding, that we will explore in the future, such as video-text retrieval, video representation learning, summarisation, intelligent fast-forwards, and so on.

In preliminary work [Bain et al., 2020] on long term temporal modelling, we show that story based video-text retrieval is possible by a simple extension of the architecture introduced in Chapter 9. We introduce a Context Boosting Module (CBM), which allows the incorporation of past and future context into the video embedding. Unlike LSMDC [Rohrbach et al., 2017],

⁶<https://www.youtube.com/user/movieclips>

which is created from DVS⁷ and contains mostly low-level descriptions of what is visually occurring in the scene, e.g. ‘Abby gets in the basket’ the descriptions in the Condensed Movie dataset (Fig. 10.1), often require information from past or future scenes for correct retrieval. Developing richer models to model longer term temporal context will also allow us to follow the evolution of relationships [Kukleva et al., 2020] and higher level semantics in movies, exciting avenues for future work. Since modalities such as audio and human speech are essential to following plotlines, multimodal video understanding techniques will be crucial here.

10.2.3 Modelling Human Intent

We show how visual human action recognition can be greatly improved by using other modalities such as speech (Chapter 3) and audio (Chapter 8). An exciting avenue for future work is to model *intent* rather than just actions - both the intent of the media creator (what is the movie more broadly trying to tell us with this scene) as well as the intent of the observed agents in a scene. For edited media, this can be useful to understand desired audience reactions, while for videos in situated perception, this can enable the important goal of understanding "what to do" i.e. event/anomaly detection and the creation of natural interfaces for robotics. In many of these cases additional modalities such as audio and human speech can be a natural source of supervision to augment computer vision.

10.3 Conclusion

In this thesis we develop methods to transfer supervision from one modality to another (Chapters 2 to 4), learn cross-modal representations with self-supervision (Chapters 5 to 7), and fuse together different modalities to learn compact video representations (Chapters 8 and 9). We use text, vision, audio and speech for human-centric video understanding tasks such as human identity, emotion and action recognition, as well as free form text-video retrieval. These in

⁷Descriptive Video Services

conjunction with other modalities such as touch and remote sensing will enable us to create machines that can automatically understand and interact with our vibrant multimodal world.

Bibliography

- Afouras, T., Chung, J. S., & Zisserman, A. (2018). The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*.
- Alamri, H., Hori, C., Marks, T. K., Batra, D., & Parikh, D. (2018). Audio visual scene-aware dialog (AVSD) track for natural language generation in DSTC7. In *DSTC7 at AAAI2019 Workshop*.
- Albanie, S., Nagrani, A., Vedaldi, A., & Zisserman, A. (2018). Emotion recognition in speech using cross-modal transfer in the wild. In *ACM International Conference on Multimedia*.
- Albanie, S., & Vedaldi, A. (2016). Learning grimaces by watching TV. In *Proceedings of the British Machine Vision Conference*.
- Aldeneh, Z., & Provost, E. M. (2017). Using regional saliency for speech emotion recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2741–2745).
- Alvi, M., Zisserman, A., & Nellåker, C. (2018). Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference on Computer Vision (ECCV)*.
- Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning* (pp. 1247–1255).
- Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., & Russell, B. (2017). Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5803–5812).
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C., & Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the International Conference on Computer Vision*.
- Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Arandjelović, R., & Zisserman, A. (2017). Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 609–617).
- Arandjelović, R., & Zisserman, A. (2018). Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 435–451).

- Arevalo, J., Solorio, T., Montes-y Gomez, M., & Gonzalez, F. A. (2017). Gated multimodal units for information fusion. In *ICLRW*.
- Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6), 345–379.
- Aviezer, H., Bentin, S., Hassin, R. R., Meschino, W. S., Kennedy, J., Grewal, S., ... Moscovitch, M. (2009). Not on the face alone: perception of contextualized face expressions in huntington’s disease. *Brain*, 132(6), 1633–1644.
- Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., & Torralba, A. (2017). Cross-modal scene networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Aytar, Y., Vondrick, C., & Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems* (pp. 892–900).
- Aytar, Y., Vondrick, C., & Torralba, A. (2017). See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*.
- Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep? In *Advances in neural information processing systems* (pp. 2654–2662).
- Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011). Sequential deep learning for human action recognition. In *International workshop on human behavior understanding* (pp. 29–39).
- Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology*, 36(2), 190.
- Bain, M., Nagrani, A., Brown, A., & Zisserman, A. (2020). Condensed movies: Story based retrieval with contextual embeddings. *arXiv preprint arXiv:2005.04208*.
- Ballas, N., Yao, L., Pal, C., & Courville, A. (2015). Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*.
- Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2), 423–443.
- Bargal, S. A., Barsoum, E., Ferrer, C. C., & Zhang, C. (2016). Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 433–436).

- Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, 1(3), 295–311.
- Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. d., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of machine learning research*, 3(Feb), 1107–1135.
- Barsoum, E., Zhang, C., Canton Ferrer, C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*.
- Batliner, A., Hacker, C., Steidl, S., Nøth, E., D’Arcy, S., Russell, M. J., & Wong, M. (2004). You Stupid Tin Box - Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus. In *LREC*.
- Bell, P., Gales, M. J., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., ... others (2015). The MGB challenge: Evaluating multi-genre broadcast media recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.
- Ben-Younes, H., Cadene, R., Cord, M., & Thome, N. (2017). Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2612–2620).
- Bhattacharya, G., Alam, J., & Kenny, P. (2017). Deep speaker embeddings for short-duration speaker verification. In *INTERSPEECH*.
- Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., & Gould, S. (2016). Dynamic image networks for action recognition. In *CVPR*.
- Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., & Sivic, J. (2013). Finding actors and actions in movies. In *Proceedings of the IEEE international conference on computer vision* (pp. 2280–2287).
- Borjon, J. I., Schroer, S. E., Bambach, S., Slone, L. K., Abney, D. H., Crandall, D. J., & Smith, L. B. (2018). A view of their own: Capturing the egocentric view of infants and toddlers with head-mounted cameras. *JoVE (Journal of Visualized Experiments)*(140), e58445.
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, 77(3), 305–327.
- Brunelli, R., & Falavigna, D. (1995). Person identification using multiple cues. *IEEE transactions on pattern analysis and machine intelligence*, 17(10), 955–966.

- Bucilua, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 535–541).
- Budnik, M., Poignant, J., Besacier, L., & Quénot, G. (2014). Automatic propagation of manual annotations for multimodal person identification in tv shows. In *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on* (pp. 1–4).
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*.
- Bushnell, E. W. (1994). A dual-processing approach to cross-modal matching: Implications for development. *The development of intersensory perception: Comparative perspectives*, 19–38.
- Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., . . . Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4), 335.
- Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., . . . Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205–211).
- Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 961–970).
- Cai, W., Cai, Z., Zhang, X., Wang, X., & Li, M. (2018). A novel learnable dictionary encoding layer for end-to-end language identification. *arXiv preprint arXiv:1804.00385*.
- Cai, W., Chen, J., & Li, M. (2018a). Analysis of length normalization in end-to-end speaker verification system. *arXiv preprint arXiv:1806.03209*.
- Cai, W., Chen, J., & Li, M. (2018b). Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. *arXiv preprint arXiv:1804.05160*, 2018.
- Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: a dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.
- Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the Kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299–6308).

- Çetingül, H. E., Erzin, E., Yemez, Y., & Tekalp, A. M. (2006). Multimodal speaker/speech recognition using lip motion, lip texture and audio. *Signal processing*, 86(12), 3549–3558.
- Cetingul, H. E., Yemez, Y., Erzin, E., & Tekalp, A. M. (2005). Robust lip-motion features for speaker identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 1, pp. I–509).
- Chakravarty, P., & Tuytelaars, T. (2016). Cross-modal supervision for learning active speaker detection in video. In *European Conference on Computer Vision*.
- Chatfield, K., Lempitsky, V., Vedaldi, A., & Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference*.
- Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference*.
- Chechik, G., Ie, E., Rehn, M., Bengio, S., & Lyon, D. (2008). Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (pp. 105–112).
- Chen, D., Tsai, S., Chandrasekhar, V., Takacs, G., Chen, H., Vedantham, R., ... Girod, B. (2011). Residual enhanced visual vectors for on-device image matching. In *Asilomar*.
- Chen, D. L., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 190–200).
- Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2172–2180).
- Chen, Y., Lu, X., & Feng, Y. (2019). Deep voice-visual cross-modal retrieval with deep feature similarity learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (pp. 454–465).
- Chiu, C.-C., et al. (2018). State-of-the-art speech recognition with sequence-to-sequence models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4774–4778).
- Choi, H.-S., Park, C., & Lee, K. (2020). From inference to generation: End-to-end fully self-supervised generation of human face from speech. *arXiv preprint arXiv:2004.05830*.

- Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Chowdhury, F., Wang, Q., Moreno, I. L., & Wan, L. (2017). Attention-based models for text-dependent speaker verification. *arXiv preprint arXiv:1710.10470*.
- Chung, J., Ahn, S., & Bengio, Y. (2016). Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704*.
- Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., & Bengio, Y. (2015). A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*.
- Chung, J. S., Huh, J., Nagrani, A., Afouras, T., & Zisserman, A. (2020). Spot the conversation: speaker diarisation in the wild. *INTERSPEECH*.
- Chung, J. S., Jamaludin, A., & Zisserman, A. (2017). You said that? In *Proceedings of the British Machine Vision Conference*.
- Chung, J. S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. In *INTERSPEECH*.
- Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Chung, J. S., & Zisserman, A. (2016a). Lip reading in the wild. In *Proceedings of the Asian Conference on Computer Vision*.
- Chung, J. S., & Zisserman, A. (2016b). Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.
- Chung, J. S., & Zisserman, A. (2017). Lip reading in profile. In *Proceedings of the British Machine Vision Conference*.
- Chung, S.-W., Chung, J. S., & Kang, H.-G. (2019). Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3965–3969).
- Chung, S.-W., Chung, J. S., & Kang, H. G. (2020). Perfect match: Self-supervised embeddings for cross-modal retrieval. *IEEE Journal of Selected Topics in Signal Processing*.

- Chung, S.-W., Kang, H. G., & Chung, J. S. (2020). Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. *arXiv preprint arXiv:2004.14326*.
- Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., & Lee, L.-S. (2016). Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *INTERSPEECH*.
- Cinbis, R. G., Verbeek, J., & Schmid, C. (2011). Unsupervised metric learning for face identification in tv video. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 1559–1566).
- Cour, T., Sapp, B., Jordan, C., & Taskar, B. (2009). Learning from ambiguously labeled images. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 919–926).
- Crowley, E. J., Gray, G., & Storkey, A. (2017). Moonshine: Distilling with cheap convolutions. *arXiv preprint arXiv:1711.02613*.
- Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., & Schuller, B. W. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 2017 ACM on Multimedia Conference* (pp. 478–484).
- Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, 52(6), 555–564.
- Cvejic, E., Kim, J., & Davis, C. (2012). Recognizing prosody across modalities, face areas and speakers: Examining perceivers’ sensitivity to variable realizations of visual prosody. *Cognition*, 122(3), 442–453.
- Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1), 30–42.
- Damen, D., Doughty, H., Maria Farinella, G., Fidler, S., Furnari, A., Kazakos, E., ... others (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 720–736).
- Damen, D., Leelasawassuk, T., Haines, O., Calway, A., & Mayol-Cuevas, W. (2014). You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video. In *BMVC*.
- Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788–798.

- Deng, D., Liu, H., Li, X., & Cai, D. (2018). Pixellink: Detecting scene text via instance segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).
- Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4690–4699).
- Deng, J., Zhang, Z., Eyben, F., & Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9), 1068–1072.
- Deng, J., Zhang, Z., & Schuller, B. (2014). Linked source and target domain subspace feature transfer learning—exemplified by speech emotion recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on* (pp. 761–766).
- de Sa, V. R. (1994). Learning classification with unlabeled data. In *Advances in neural information processing systems* (pp. 112–119).
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dhall, A., Goecke, R., Joshi, J., Hoey, J., & Gedeon, T. (2016). Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 427–432).
- Dhall, A., Goecke, R., Lucey, S., Gedeon, T., et al. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3), 34–41.
- Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1422–1430).
- Dong, J., Li, X., & Snoek, C. G. (2016). Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838*.
- Dong, J., Li, X., & Snoek, C. G. (2018). Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12), 3377–3388.

- Dong, J., Li, X., Xu, C., Ji, S., & Wang, X. (2019). Dual dense encoding for zero-example video retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Douglas-Cowie, E., Cowie, R., & Schröder, M. (2000). A new emotion database: considerations, sources and scope. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.
- Douze, M., Ramisa, A., & Schmid, C. (2011). Combining attributes and fisher vectors for efficient image retrieval. In *CVPR 2011* (pp. 745–752).
- Duchenne, O., Laptev, I., Sivic, J., Bach, F., & Ponce, J. (2009). Automatic annotation of human actions in video. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 1491–1498).
- Duygulu, P., Barnard, K., de Freitas, J. F., & Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision* (pp. 97–112).
- Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection*. Basic books.
- Engilberge, M., Chevallier, L., Pérez, P., & Cord, M. (2018). Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3984–3993).
- Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., ... Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *SIGGRAPH*.
- Everingham, M., Sivic, J., & Zisserman, A. (2006). “Hello! My name is... Buffy” – automatic naming of characters in TV video. In *BMVC*.
- Everingham, M., Sivic, J., & Zisserman, A. (2009). Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 2009, 27(5).
- Fabius, O., & van Amersfoort, J. R. (2014). Variational recurrent auto-encoders. *arXiv preprint arXiv:1412.6581*.
- Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2(7), 8.
- Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *European conference on computer vision* (pp. 15–29).

- Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 6202–6211).
- Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Feng, L., & Hansen, L. K. (2005). *A new database for speaker recognition* (Tech. Rep.).
- Fernando, B., Bilen, H., Gavves, E., & Gould, S. (2016). Self-supervised video representation learning with odd-one-out networks. *arXiv preprint arXiv:1611.06646*.
- Fisher, W. M., Doddington, G. R., & Goudie-Marshall, K. M. (1986). The DARPA speech recognition research database: specifications and status. In *Proc. DARPA Workshop on speech recognition* (pp. 93–99).
- Fouhey, D. F., Kuo, W.-c., Efros, A. A., & Malik, J. (2018). From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4991–5000).
- Fraccaro, M., Sønderby, S. K., Paquet, U., & Winther, O. (2016). Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems* (pp. 2199–2207).
- Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems* (pp. 2121–2129).
- Furnari, A., & Farinella, G. (2020). Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Gabeur, V., Sun, C., Alahari, K., & Schmid, C. (2020). Multi-modal transformer for video retrieval. *arXiv preprint arXiv:2007.10639*.
- Gao, R., & Grauman, K. (2019). 2.5D visual sound. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report*.
- Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 776–780).

- Ghadiyaram, D., Tran, D., & Mahajan, D. (2019). Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 12046–12055).
- Ghalehjeh, S. H., & Rose, R. C. (2015). Deep bottleneck features for i-vector based text-independent speaker verification. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.
- Ghanem, B., Niebles, J. C., Snoek, C., Heilbron, F. C., Alwassel, H., Escorcia, V., . . . Dao, C. D. (2018). The activitynet large-scale activity recognition challenge 2018 summary. *arXiv preprint arXiv:1808.03766*.
- Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2018). A Better Baseline for AVA. In *ActivityNet Workshop at CVPR*.
- Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2019). Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 244–253).
- Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., & Russell, B. (2017). ActionVLAD: learning spatio-temporal aggregation for action classification. In *CVPR*.
- Girdhar, R., Tran, D., Torresani, L., & Ramanan, D. (2019). Distinit: Learning video representations without a single labeled video. *ICCV*.
- Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).
- Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., . . . others (2013). Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing* (pp. 117–124).
- Gordo, A., & Larlus, D. (2017). Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, 56(4), 544–549.

- Greenberg, C. S. (2012). The NIST year 2012 speaker recognition evaluation plan. *NIST, Technical Report*.
- Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative neurology and neuroscience*, 28(2), 219–236.
- Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., ... Malik, J. (2018). AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6047–6056).
- Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*.
- Gupta, S., Hoffman, J., & Malik, J. (2016). Cross modal distillation for supervision transfer. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on* (pp. 2827–2836).
- Gutmann, M., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 297–304).
- Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Hajibabaei, M., & Dai, D. (2018). Unified hypersphere embedding for speaker recognition. *arXiv preprint arXiv:1807.08312*, 2018.
- Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8–12.
- Han, J., Zhang, Z., Ren, Z., & Schuller, B. W. (2019). Emobed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. *IEEE Transactions on Affective Computing*.
- Han, J., Zhang, Z., Schmitt, M., Pantic, M., & Schuller, B. (2017). From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 2017 ACM on Multimedia Conference* (pp. 890–897).
- Han, T., Xie, W., & Zisserman, A. (2019). Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.
- Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6), 74–99.

- Hansen, J. H., Sarikaya, R., Yapanel, U. H., & Pellom, B. L. (2001). Robust speech recognition in noise: an evaluation using the spine corpus. In *INTERSPEECH*.
- Harwath, D., Torralba, A., & Glass, J. (2016). Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems* (pp. 1858–1866).
- Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, 5(1), 60–65.
- Hauptmann, A. G., Jin, R., & Ng, T. D. (2002). Multi-modal information retrieval from broadcast video using ocr and speech recognition. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries* (pp. 160–161).
- He, G., Liu, X., Fan, F., & You, J. (2020, June). Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. (2016). End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5115–5119).
- Hennebert, J., Melin, H., Petrovska, D., & Genoud, D. (2000). POLYCOST: a telephone-speech database for speaker recognition. *Speech communication*, 31(2), 265–270.
- Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.
- Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., . . . Wilson, K. (2017). Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Retrieved from <https://arxiv.org/abs/1609.09430>
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Hollien, H., & Moore, G. P. (1960). Measurements of the vocal folds during changes in pitch. *Journal of Speech, Language, and Hearing Research*, 3(2), 157–165.

- Hori, C., Hori, T., Lee, T.-Y., Zhang, Z., Harsham, B., Hershey, J. R., ... Sumi, K. (2017). Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision* (pp. 4193–4202).
- Horiguchi, S., Kanda, N., & Nagamatsu, K. (2018). Face-voice matching using cross-modal embeddings. In *Proceedings of the 26th ACM international conference on Multimedia* (pp. 1011–1019).
- Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-excitation networks. *IEEE transactions on pattern analysis and machine intelligence*.
- Hu, P., Cai, D., Wang, S., Yao, A., & Chen, Y. (2017). Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 553–560).
- Huang, C. (2017). Combining convolutional neural networks for emotion recognition. In *Undergraduate Research Technology Conference (URTC), 2017 IEEE MIT* (pp. 1–4).
- Idrees, H., Zamir, A. R., Jiang, Y.-G., Gorban, A., Laptev, I., Sukthankar, R., & Shah, M. (2017). The THUMOS challenge on action recognition for videos in the wild. *Computer Vision and Image Understanding*, 155, 1–23.
- Ignat, O., Burdick, L., Deng, J., & Mihalcea, R. (2019). Identifying visible actions in lifestyle vlogs. *arXiv preprint arXiv:1906.04236*.
- Ioffe, S. (2006). Probabilistic linear discriminant analysis. In *Proceedings of the European Conference on Computer Vision* (pp. 531–542).
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.
- Jacobs, R., Jordan, M., Nowlan, S., & Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation* 3, 79–87.
- Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.
- Jakab, T., Gupta, A., Bilen, H., & Vedaldi, A. (2020). Self-supervised learning of interpretable keypoints from unlabelled videos. *CVPR*.
- Jamaludin, A., Chung, J. S., & Zisserman, A. (2019). You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11–12), 1767–1779.

- Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., . . . others (2003). The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3304–3311).
- Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *PAMI*, 35(1), 221–231.
- Jing, L., & Tian, Y. (2018). Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*.
- Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns’ preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1-2), 1–19.
- Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2), 181–214.
- Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology*, 13(19), 1709–1714.
- Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3128–3137).
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725–1732).
- Karras, T., Aila, T., Laine, S., Herva, A., & Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4), 94.
- Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., . . . Zisserman, A. (2017). The Kinetics human action video dataset. *CoRR*, abs/1705.06950.
- Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1867–1874).

- Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The MegaFace benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal, 2005*.
- Khoury, E., El Shafey, L., McCool, C., Günther, M., & Marcel, S. (2014). Bi-modal biometric authentication on mobile phones in challenging conditions. *Image and Vision Computing*, 32(12), 1147–1160.
- Kidron, E., Schechner, Y. Y., & Elad, M. (2005). Pixels that sound. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 88–95).
- Kim, C., Shin, H. V., Oh, T.-H., Kaspar, A., Elgharib, M., & Matusik, W. (2018). On learning associations of faces and voices. *arXiv:1805.05553*.
- Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., ... Theobalt, C. (2018). Deep video portraits. *SIGGRAPH*.
- Kim, J., Englebienne, G., Truong, K. P., & Evers, V. (2017). Deep temporal models using identity skip-connections for speech emotion recognition. In *Proceedings of the 2017 ACM on Multimedia Conference* (pp. 1006–1013).
- Kim, Y., & Provost, E. M. (2016). Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 92–99).
- King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10, 1755–1758.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.
- Klein, B., Lev, G., Sadeh, G., & Wolf, L. (2015). Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4437–4446).

- Korbar, B., Tran, D., & Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*.
- Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, 38(6), 618–625.
- Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Carlos Nibbles, J. (2017). Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 706–715).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).
- Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. In *2011 International Conference on Computer Vision* (pp. 2556–2563).
- Kukleva, A., Tapaswi, M., & Laptev, I. (2020). Learning interactions and relationships between movie characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9849–9858).
- Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., . . . Berg, T. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 2891–2903.
- Kulkarni, T. D., Whitney, W. F., Kohli, P., & Tenenbaum, J. (2015). Deep convolutional inverse graphics network. In *NeurIPS* (pp. 2539–2547).
- Lachs, L., & Pisoni, D. B. (2004). Specification of cross-modal source information in isolated kinematic displays of speech. *The Journal of the Acoustical Society of America*, 116(1), 507–518.
- Lampert, C. H., & Krömer, O. (2010). Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *European Conference on Computer Vision* (pp. 566–579).
- Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, 33(4), 905.
- Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *IEEE Conference on Computer Vision & Pattern Recognition*.
- Latif, S., Rana, R., Younis, S., Qadir, J., & Epps, J. (2018). Transfer learning for improving speech emotion classification accuracy. *arXiv preprint arXiv:1801.06353*.

- Le, N., & Odobez, J.-M. (2017). Improving speaker turn embedding by crossmodal transfer learning from face embedding. *arXiv preprint arXiv:1707.02749*.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- Lee, H.-Y., Huang, J.-B., Singh, M., & Yang, M.-H. (2017). Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 667–676).
- Lee, K. A., Hautamaki, V., Kinnunen, T., Yamamoto, H., Okabe, K., Vestman, V., ... others (2019). I4u submission to nist sre 2018: Leveraging from a decade of shared experiences. *arXiv preprint arXiv:1904.07386*.
- Lee, Y. J., Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *CVPR*.
- Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., ... Zhu, Z. (2017). Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.
- Li, D., Dimitrova, N., Li, M., & Sethi, I. K. (2003). Multimedia content processing through cross-modal association. In *Proceedings of the eleventh ACM international conference on Multimedia* (pp. 604–611).
- Li, F., Neverova, N., Wolf, C., & Taylor, G. (2016). Modout: Learning to fuse modalities via stochastic regularization. *Journal of Computational Vision and Imaging Systems*, 2(1).
- Li, J., Zhao, R., Huang, J.-T., & Gong, Y. (2014). Learning small-size dnn with output-distribution-based criteria. In *Fifteenth Annual Conference of the International Speech Communication Association*.
- Li, Y., Liu, M., & Rehg, J. M. (2020). In the eye of the beholder: Gaze and actions in first person video. *arXiv preprint arXiv:2006.00626*.
- Lieberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). Ldc emotional prosody speech transcripts database. *University of Pennsylvania, Linguistic data consortium*.

- Lin, J., Calandra, R., & Levine, S. (2019). Learning to identify object instances by touch: Tactile recognition via multimodal matching. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 3644–3650).
- Lin, W., Mi, Y., Wu, J., Lu, K., & Xiong, H. (2018). Action recognition with coarse-to-fine deep feature integration and asynchronous fusion. *AAAI*.
- Lin, X., & Parikh, D. (2015). Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2984–2993).
- Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2019). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: single shot multibox detector. In *Proceedings of the European Conference on Computer Vision* (pp. 21–37).
- Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Liu, Y., Albanie, S., Chen, Q., & Zisserman, A. (2019). Team speedy multi moments in time challenge 2019 technical report.
- Liu, Y., Li, H., & Wang, X. (2017). Rethinking feature discrimination and polymerization for large-scale recognition. *arXiv preprint arXiv:1710.00870*.
- Liu, Y., Wang, Z., Jin, H., & Wassell, I. (2018). Synthetically supervised feature learning for scene text recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 435–451).
- Long, X., Gan, C., De Melo, G., Liu, X., Li, Y., Li, F., & Wen, S. (2018). Multimodal keyless attention fusion for video classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Long, X., Gan, C., De Melo, G., Wu, J., Liu, X., & Wen, S. (2018). Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7834–7843).
- Long, X., Gan, C., de Melo, G., Wu, J., Liu, X., & Wen, S. (2018, June). Attention clusters: Purely attention based local feature integration for video classification. In *CVPR*.

- Long, X., Gan, C., Melo, G., Liu, X., Li, Y., Li, F., & Wen, S. (2018). Multimodal keyless attention fusion for video classification. In *AAAI Conference on Artificial Intelligence*.
- Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.
- Ma, M., Fan, H., & Kitani, K. M. (2016). Going deeper into first-person activity recognition. In *CVPR*.
- Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), 2579–2605.
- Magavand, P., Molholm, S., Nayak, A., & Foxe, J. J. (2013). Recalibration of the multisensory temporal window of integration results from changing task demands. *PLOS ONE*, 8.
- Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., . . . van der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 181–196).
- Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems* (pp. 1682–1690).
- Mao, G., Yuan, Y., & Xiaoqiang, L. (2018). Deep cross-modal retrieval for remote sensing image and audio. In *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)* (pp. 1–7).
- Mariooryad, S., & Busso, C. (2013). Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In *Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 85–90).
- Marszałek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition* (pp. 2929–2936).
- Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The enterface’05 audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on* (pp. 8–8).
- McCool, C., & Marcel, S. (2009). *Mobio database for the ICPR 2010 face and speech competition* (Tech. Rep.). IDIAP.
- McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., . . . others (2005). The AMI meeting corpus. In *International Conference on Methods and Techniques in Behavioral Research* (Vol. 88).

- McGinn, C., & Torre, I. (2019). Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 211–221).
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.
- McLaren, M., Ferrer, L., Castan, D., & Lawson, A. (2016). The speakers in the wild (SITW) speaker recognition database. In *INTERSPEECH*.
- Miech, A., Alayrac, J., Bojanowski, P., Laptev, I., & Sivic, J. (2017). Learning from video and text via large-scale discriminative clustering. In *Proceedings of the IEEE international conference on computer vision*.
- Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., & Zisserman, A. (2019). End-to-end learning of visual representations from uncurated instructional videos. *arXiv preprint arXiv:1912.06430*.
- Miech, A., Laptev, I., & Sivic, J. (2017). Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*.
- Miech, A., Laptev, I., & Sivic, J. (2018). Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.
- Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *Proceedings of the IEEE international conference on computer vision*.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Millar, J. B., Vonwiller, J. P., Harrington, J. M., & Dermody, P. J. (1994). The Australian national database of spoken language. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Miller, D. R., Kleber, M., Kao, C.-L., Kimball, O., Colthurst, T., Lowe, S. A., ... Gish, H. (2007). Rapid and accurate spoken term detection. In *Eighth Annual Conference of the international speech communication association*.
- Misra, I., Zitnick, C. L., & Hebert, M. (2016). Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision* (pp. 527–544).

- Mithun, N. C., Li, J., Metze, F., & Roy-Chowdhury, A. K. (2018). Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (pp. 19–27).
- Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems* (pp. 2265–2273).
- Moltisanti, D., Wray, M., Mayol-Cuevas, W., & Damen, D. (2017). Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video. In *ICCV*.
- Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, Y., . . . others (2019). Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*.
- Morrison, G., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., . . . Chow, D. (2015). Forensic database of voice recordings of 500+ Australian English speakers. URL: <http://databases.forensic-voice-comparison.net>, 2015.
- Nagrani, A., Albanie, S., & Zisserman, A. (2018a). Learnable PINs: Cross-modal embeddings for person identity. In *The European Conference on Computer Vision (ECCV)*.
- Nagrani, A., Albanie, S., & Zisserman, A. (2018b). Seeing voices and hearing faces: Cross-modal biometric matching. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. In *INTERSPEECH*.
- Nagrani, A., & Zisserman, A. (2017). From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script. In *Proceedings of the British Machine Vision Conference*.
- Naim, I., Al Mamun, A., Song, Y. C., Luo, J., Kautz, H., & Gildea, D. (2016). Aligning movies with scripts by exploiting temporal ordering constraints. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 1786–1791).
- Nam, H., Ha, J.-W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 299–307).

- Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., . . . Natarajan, P. (2012). Multimodal feature fusion for robust event detection in web videos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1298–1305).
- Neverova, N., Wolf, C., Taylor, G., & Nebout, F. (2015). Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8), 1692–1706.
- Neverova, N., Wolf, C., Taylor, G. W., & Nebout, F. (2014). Multi-scale deep learning for gesture detection and localization. In *European Conference on Computer Vision* (pp. 474–490).
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 689–696).
- Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. (2017). Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*.
- Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision* (pp. 69–84).
- Oh, T.-H., Dekel, T., Kim, C., Mosseri, I., Freeman, W. T., Rubinstein, M., & Matusik, W. (2019). Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7539–7548).
- Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. *arXiv preprint arXiv:1803.10963*.
- Oller, D. K. (1973). The effect of position in utterance on speech segment duration in english. *The journal of the Acoustical Society of America*, 54(5), 1235–1247.
- Otaegui, O. (2018). Multimodal deep learning for advanced driving systems. In *Articulated Motion and Deformable Objects: 10th International Conference, AMDO 2018, Palma de Mallorca, Spain, July 12-13, 2018, Proceedings* (Vol. 10945, p. 95).
- Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., & Yokoya, N. (2016). Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision* (pp. 651–667).
- Ouyang, H., & Lee, T. (2006). A new lip feature representation method for video-based bimodal authentication. In *Proceedings of the 2005 NICTA-HCSNet Multimodal User Interaction Workshop-Volume 57* (pp. 33–37).

- Owens, A., & Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 631–648).
- Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., & Torralba, A. (2016). Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision* (pp. 801–816).
- Pan, Y., Mei, T., Yao, T., Li, H., & Rui, Y. (2016). Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4594–4602).
- Parise, C., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology*, 22(1), 46 - 49.
- Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference*.
- Parthasarathy, S., Zhang, C., Hansen, J. H., & Busso, C. (2017). A study of speaker verification performance with expressive speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (pp. 5540–5544).
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017). Automatic differentiation in pytorch.
- Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *The Journal of the Acoustical Society of America*, 109(4), 1668–1680.
- Pell, M. D. (2005). Prosody–face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior*, 29(4), 193–215.
- Pérez-Rúa, J.-M., Vielzeuf, V., Pateux, S., Baccouche, M., & Jurie, F. (2019). Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6966–6975).
- Perronnin, F., & Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1–8).
- Piaget, J. (1936). (1963). the origins of intelligence in children. *Trans. M. Cook. New York: WW Norton*.
- Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *CVPR*.

- Poria, S., Cambria, E., Hussain, A., & Huang, G.-B. (2015). Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63, 104–116.
- Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1), 145–151.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Retrieved from https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- Radha, N. (2016). Video retrieval using speech and text in video. In *2016 International Conference on Inventive Computation Technologies (ICICT)* (Vol. 2, pp. 1–6).
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3), 19–41.
- Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1), 72–83.
- Rigoulot, S., & Pell, M. D. (2014). Emotion in the voice influences the way we scan emotional faces. *Speech Communication*, 65, 36–49.
- Riley, C. (2009). *The Hollywood standard: the complete and authoritative guide to script format and style*. Michael Wiese Productions.
- Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3202–3212).
- Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., ... Schiele, B. (2017). Movie description. *International Journal of Computer Vision*, 123(1), 94–120.
- Rolnick, D., Veit, A., Belongie, S., & Shavit, N. (2017). Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.
- Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & psychophysics*, 68(1), 84–93.
- Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*.

- Roy, A., & Marcel, S. (2010). Introducing crossmodal biometrics: Person identification from distinct audio & visual streams. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on* (pp. 1–6).
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Li, F. (2015). Imagenet large scale visual recognition challenge. in: *International Journal of Computer Vision, 2015*.
- Ryu, S., & Kim, S.-C. (2020). Embedded identification of surface based on multirate sensor fusion with deep neural network. *IEEE Embedded Systems Letters*.
- Safari, P., & Hernando, J. (2019). Self multi-head attention for speaker recognition. *arXiv preprint arXiv:1906.09890*.
- Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017a). A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*.
- Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017b). A simple neural network module for relational reasoning. In *Advances in neural information processing systems* (pp. 4967–4976).
- Saon, G., Soltan, H., Nahamoo, D., & Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *ASRU* (pp. 55–59).
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).
- Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*. (Vol. 3, pp. 32–36).
- Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing, 1*(2), 119–131.
- Sekuler, R., Sekuler, A., & Lau, R. (1997, January). Sound alters visual motion perception. *Nature, 385*(6614), 308. Retrieved from <https://doi.org/10.1038/385308a0> doi: 10.1038/385308a0

- Senocak, A., Oh, T.-H., Kim, J., Yang, M.-H., & So Kweon, I. (2018). Learning to localize sound source in visual scenes. In *CVPR*.
- Shahroudy, A., Ng, T.-T., Gong, Y., & Wang, G. (2017). Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(5), 1045–1058.
- Sharma, S., Kiros, R., & Salakhutdinov, R. (2015). Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*.
- Sheffert, S. M., & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Attention, Perception, & Psychophysics*, 66(2), 352–362.
- Shinohara, Y. (2016). Adversarial multi-task learning of deep neural networks for robust speech recognition. In *INTERSPEECH* (pp. 2369–2372).
- Shon, S., Tang, H., & Glass, J. (2018). Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model. *arXiv preprint arXiv:1809.04437*.
- Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 761–769).
- Sigurdsson, G. A., Gupta, A., Schmid, C., Farhadi, A., & Alahari, K. (2018). Actor and observer: Joint modeling of first and third-person videos. In *CVPR*.
- Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision* (pp. 510–526).
- Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (pp. 568–576).
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
- Singh, R. (2019). Reconstruction of the human persona in 3D from voice, and its reverse. In *Profiling Humans from their Voice* (pp. 325–363). Springer.
- Singh, S., Arora, C., & Jawahar, C. V. (2016). First person action recognition using deep learned descriptors. In *CVPR*.

- Sivic, J., Everingham, M., & Zisserman, A. (2009). "who are you?"-learning person specific classifiers from video. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1145–1152).
- Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *null* (p. 1470).
- Smith, H. M., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016a). Concordant cues in faces and voices: Testing the backup signal hypothesis. *Evolutionary Psychology*, 14(1), 1474704916630317.
- Smith, H. M., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016b). Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, 78(3), 868–879.
- Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2), 13–29.
- Snoek, C. G., Worring, M., & Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 399–402).
- Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *Interspeech* (pp. 999–1003).
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.
- Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2, 207–218.
- Song, H. O., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Song, S., Chandrasekhar, V., Mandal, B., Li, L., Lim, J.-H., Sateesh Babu, G., ... Cheung, N.-M. (2016). Multimodal multi-stream deep learning for egocentric activity recognition. In *CVPRW*.
- Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: a dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.
- Srivastava, N., & Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems* (pp. 2222–2230).

- Stevenson, R. A., Wilson, M. M., Powers, A. R., & Wallace, M. T. (2013). The effects of visual training on multisensory temporal processing. *Experimental Brain Research*, 225(4), 479–489.
- Stoll, L. L. (2011). Finding difficult speakers in automatic speaker recognition. *Technical Report No. UCB/EECS-2011-152*.
- Sudhakaran, S., & Lanz, O. (2018). Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *BMVC*.
- Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273), 71–78.
- Sun, C., Baradel, F., Murphy, K., & Schmid, C. (2019). Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.
- Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*.
- Sun, Y., Chen, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems* (pp. 1988–1996).
- Sung, K.-K. (1996). *Learning and example selection for object and pattern detection* (Unpublished doctoral dissertation).
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).
- Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, 36(2), 219–238.
- Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014a). Deep-Face: Closing the gap to human-level performance in face verification. In *IEEE CVPR*.
- Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014b). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1701–1708).
- Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., ... Zhou, J. (2019). Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1207–1216).

- Tapaswi, M., Bäumel, M., & Stiefelhagen, R. (2012). ‘Knock! Knock! Who is it?’ probabilistic person identification in TV-series. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2658–2665).
- Tapaswi, M., Baumel, M., & Stiefelhagen, R. (2015, June). Book2movie: Aligning video scenes with book chapters. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Thornhill, R., & Møller, A. P. (1997). Developmental stability, disease and medicine. *Biological Reviews*.
- Tian, Y., Li, D., & Xu, C. (2020). Weakly-supervised audio-visual video parsing toward unified multisensory perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- Torabi, A., Tandon, N., & Sigal, L. (2016). Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*.
- Torfi, A., Dawson, J., & Nasrabadi, N. M. (2018). Text-independent speaker verification using 3d convolutional neural networks. In *ICME* (pp. 1–6).
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489–4497).
- Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 6450–6459).
- Tzeng, E., Hoffman, J., Darrell, T., & Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. In *Proceedings of the International Conference on Computer Vision* (pp. 4068–4076).
- van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In *Advances in Neural Information Processing Systems*.
- van der Vloed, D., Bouten, J., & van Leeuwen, D. A. (2014). NFI-FRITS: a forensic speaker recognition database and some first experiments. In *The Speaker and Language Recognition Workshop*.
- Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Vedaldi, A., & Lenc, K. (2014). Matconvnet – convolutional neural networks for matlab. *CoRR*, abs/1412.4564.

- Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision* (pp. 4534–4542).
- Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156–3164).
- Wallace, M. T., & Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia*, 64, 105 - 123.
- Wan, L., Wang, Q., Papir, A., & Moreno, I. (2018). Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Wang, F., Liu, W., Liu, H., & Cheng, J. (2018). Additive margin softmax for face verification. *arXiv preprint arXiv:1801.05599*.
- Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., & Liu, W. (2019). Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4006–4015).
- Wang, J., Markert, K., & Everingham, M. (2009). Learning models for object recognition from natural language descriptions. In *Proceedings of the British Machine Vision Conference*.
- Wang, L., Li, Y., & Lazebnik, S. (2016). Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5005–5013).
- Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.
- Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *CVPR*.
- Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In *Proceedings of the International Conference on Computer Vision* (pp. 2794–2802).
- Wang, X., Wu, Y., Zhu, L., & Yang, Y. (2020). Symbiotic attention with privileged information for egocentric action recognition. *arXiv preprint arXiv:2002.03137*.

- Wang, Z., Yao, K., Li, X., & Fang, S. (2020). Multi-resolution multi-head attention in deep speaker embedding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6464–6468).
- Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473–1480).
- Wells, T., Baguley, T., Sergeant, M., & Dunn, A. (2013). Perceptions of human attractiveness comprising face and voice cues. *Archives of sexual behavior*.
- Wen, Y., Ismail, M. A., Liu, W., Raj, B., & Singh, R. (2018). Disjoint mapping network for cross-modal matching of voices and faces. *arXiv preprint arXiv:1807.04836*.
- Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision* (pp. 499–515).
- Wiles, O., Koepke, S., & Zisserman, A. (2018). Self-supervised learning of a facial attribute embedding from video. In *BMVC*.
- Winer, D. R., & Young, R. M. (2017). Automated screenplay annotation for extracting storytelling knowledge. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.
- Woo, R., Park, A., & Hazen, T. J. (2006). The MIT Mobile Device Speaker Verification Corpus: Data collection and preliminary experiments. In *The Speaker and Language Recognition Workshop*.
- Wright, L. (2019). *Ranger deep learning optimizer*. <https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer>. GitHub.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Wu, Z., Jiang, Y.-G., Wang, X., Ye, H., & Xue, X. (2016). Multi-stream multi-class fusion of deep networks for video classification. In *ACM International Conference on Multimedia*.
- Xiao, F., Lee, Y. J., Grauman, K., Malik, J., & Feichtenhofer, C. (2020). Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*.

- Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).
- Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 305–321).
- Xie, W., Nagrani, A., Chung, J. S., & Zisserman, A. (2019). Utterance-level aggregation for speaker recognition in the wild. In *International Conference on Acoustics, Speech, and Signal Processing*.
- Xie, W., Shen, L., & Zisserman, A. (2018). Comparator networks. In *European Conference on Computer Vision*.
- Xiong, C., Zhang, D., Liu, T., & Du, X. (2019). Voice-face cross-modal matching and retrieval: A benchmark. *arXiv preprint arXiv:1911.09338*.
- Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., & Zhuang, Y. (2019). Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10334–10343).
- Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5288–5296).
- Xu, R., Xiong, C., Chen, W., & Corso, J. J. (2015). Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Yamamoto, N., Ogata, J., & Aiki, Y. (2003). Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition. In *Eighth European Conference on Speech Communication and Technology*.
- Yang, X., Molchanov, P., & Kautz, J. (2016). Multilayer and multimodal fusion of deep neural networks for video classification. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 978–987).
- Yapanel, U. H., Zhang, X., & Hansen, J. H. (2002). High performance digit recognition in real car environments. In *INTERSPEECH*.
- Ye, G., Liu, D., Jhuo, I.-H., & Chang, S.-F. (2012). Robust late fusion with rank minimization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3021–3028).

- Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1), 23–43.
- Yonetani, R., Kitani, K. M., & Sato, Y. (2016). Recognizing micro-actions and reactions from paired egocentric videos. In *CVPR*.
- Yu, Y., Kim, J., & Kim, G. (2018). A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 471–487).
- Yu, Y., Ko, H., Choi, J., & Kim, G. (2016). Video captioning and retrieval models with semantic attention. *arXiv preprint arXiv:1610.02947*, 6(7).
- Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (pp. 435–442).
- Yuan, L., Xu, T. L., Yu, C., & Smith, L. B. (2019). Sustained visual attention is more than seeing. *Journal of experimental child psychology*, 179, 324–336.
- Yuhas, B. P., Goldstein, M. H., & Sejnowski, T. J. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11), 65–71.
- Zhalehpour, S., Akhtar, Z., & Erdem, C. E. (2016). Multimodal emotion recognition based on peak frame selection from video. *Signal, Image and Video Processing*, 10(5), 827–834.
- Zhang, B., Hu, H., & Sha, F. (2018). Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 374–390).
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.
- Zhang, C., Koishida, K., & Hansen, J. H. (2018). Text-independent speaker verification based on triplet convolutional neural network embeddings. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(9), 1633–1644.
- Zhang, Z., Weninger, F., Wöllmer, M., & Schuller, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on* (pp. 523–528).
- Zhao, G., & Pietikäinen, M. (2013). Visual speaker identification with spatiotemporal directional features. In *International Conference Image Analysis and Recognition* (pp. 1–10).

- Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., & Torralba, A. (2018). The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 570–586).
- Zhao, H., Yan, Z., Wang, H., Torresani, L., & Torralba, A. (2017). SLAC: A sparsely labeled dataset for action classification and localization. *arXiv preprint arXiv:1712.09374*.
- Zhao, S., Ding, G., Gao, Y., & Han, J. (2017). Learning visual emotion distributions via multi-modal features fusion. In *Proceedings of the 2017 ACM on Multimedia Conference* (pp. 369–377).
- Zheng, S., Song, Y., Leung, T., & Goodfellow, I. (2016). Improving the robustness of deep neural networks via stability training. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4480–4488).
- Zhong, Y., Arandjelović, R., & Zisserman, A. (2018). Ghostvlad for set-based face recognition. In *Asian Conference on Computer Vision*.
- Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). Temporal relational reasoning in videos. In *ECCV*.
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhou, H., Liu, Y., Liu, Z., Luo, P., & Wang, X. (2019). Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 9299–9306).
- Zhou, L., Xu, C., & Corso, J. J. (2018). Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Zhou, X., & Bhanu, B. (2008). Feature fusion of side face and gait for video-based human identification. *Pattern Recognition*, 41(3), 778–795.
- Zhou, Y., & Berg, T. L. (2015). Temporal perception and prediction in ego-centric video. In *ICCV*.
- Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision* (pp. 19–27).
- Zhukov, D., Alayrac, J.-B., Cinbis, R. G., Fouhey, D., Laptev, I., & Sivic, J. (2019). Cross-task weakly supervised learning from instructional videos. *arXiv preprint arXiv:1903.08225*.

Zitnick, C., Parikh, D., & Vanderwende, L. (2013). Learning the visual interpretation of sentences. In *Proceedings of the International Conference on Computer Vision* (pp. 1681–1688).

A | Feature Aggregation

In this section we provide a more general background for research related to the spatial, temporal and multimodal aggregation of frame level features obtained from deep neural networks.

From early research in audio-visual speech recognition [Yuhas et al., 1989] to the recent blossoming of interest in language and vision models [Antol et al., 2015; C. Sun, Myers, et al., 2019], multimodal machine learning has seen decades of varying research techniques (see [Baltrušaitis et al., 2018] for a comprehensive survey). A key component of multimodal deep learning is the aggregation of frame level features (this can be temporal, spatial, or across modality) extracted from neural network or similar extractors. Note that following [Bengio et al., 2013], we use the term feature, representation and embedding interchangeably, with each referring to a vector or tensor representation of an entity, be it an image, audio sample, speech segment or a text sentence.

A.1 Spatiotemporal Aggregation

Human faces, objects, and storylines exist over time in a video, and unlike single images, tracking their motion over the length of a video is crucial to story understanding. Since most deep learning networks extract features locally, it is important to aggregate these features across space and time. The most common solutions have been Spatiotemporal Convolutional Networks [Karpathy et al., 2014; Tran et al., 2015], which mainly rely on convolution and pooling to aggregate spatio-temporal information or recurrent Spatial Networks [Baccouche et al., 2011; Ballas et al., 2015], which apply Recurrent Neural Networks such as LSTMs or GRUs to model the temporal information in videos. Early works also managed to incorporate traditional pre-deep learning encoding methods which proved effective for aggregating local image descriptions into global compact vectors (Bag of visual Words (BoW) [Sivic & Zisserman, 2003], Fisher Vectors [Perronnin & Dance, 2007] and VLAD (Vector of Locally

Aggregated Descriptors) [Jégou et al., 2010]) to large-scale video classification [Laptev et al., 2008; Schuldt et al., 2004], largely for the task of human action recognition. Inspired by this early success, [Arandjelović et al., 2016] developed NetVLAD, a differentiable version of VLAD that could be integrated into neural networks. NetVLAD was then proven to be very effective in aggregating spatial and temporal information for compact video representations [Miech, Laptev, & Sivic, 2017; Girdhar et al., 2017], and as we show in Chapter 2, for speaker recognition as well. In chapter 9, we also show that NetVlad is suitable for aggregating pre-extracted features from a number of deep learning models in different modalities, including text audio and speech. Inspired by the success of transformers in language modelling [Devlin et al., 2018], more recent approaches have also focused on attention for temporal aggregation [Sharma et al., 2015; Girdhar, Carreira, et al., 2019; C. Sun, Myers, et al., 2019].

A.2 Multimodal Aggregation

Early works distinguished early and late fusion methods [Atrey et al., 2010], respectively fusing low-level features and prediction-level features. While some works report that late fusion is largely outperformed by early fusion [Snoek et al., 2005], most multimodal fusion has been largely conducted in a late fusion manner in the video recognition literature [Simonyan & Zisserman, 2014; Long, Gan, De Melo, Wu, et al., 2018; Ghanem et al., 2018] – indeed almost all the entries that utilize audio in the 2018 ActivityNet challenge report [Ghanem et al., 2018] adopt this paradigm (they process visual and audio inputs separately, and then either concatenate the output features or average the final class scores across modalities [Xiao et al., 2020]).

Perhaps as a form of mid fusion, Zhou et al. [X. Zhou & Bhanu, 2008] use a Multiple Discriminant Analysis on concatenated features, while Neverova et al [Neverova et al., 2014] apply a heuristic consisting of fusing similar modalities earlier than the others. While Feicht-

enhofer et al. [Feichtenhofer et al., 2016] show that it is better to fuse multimodal networks spatially at the last convolutional layer than earlier, and that additionally fusing at the class prediction layer can boost accuracy, other works explore fusion at earlier levels. A widespread belief is that neural networks encode hierarchical semantic concepts, e.g. in vision, lower layers are known to serve as edge detectors, while later layers capture more complex semantic information. When fusing representations across different modalities, it is unclear at which level such fusion should take place. For example, learning to classify furry animals might require analysis of lower level visual features that can be used to build up the concept of fur, but recognition of growling from audio might require analysis of more complex attributes. Hence many multimodal fusion works are focused on finding at which depths unimodal layers should be fused. To take advantage of both low-level and high-level features, Yang et al. [Yang et al., 2016] leverage boosting for fusion across all layers, and more recently, Audio-visual Slow Fast networks [Xiao et al., 2020] fuse audio and visual features at multiple stages, from intermediate-level features to high-level semantic concepts. Another recent approach for multimodal fusion is the focus on attention mechanisms, which decide how to select part of different modalities using contextual information. The seminal work on mixture of experts by [Jacobs et al., 1991] could be viewed as a kickstarter in this direction, where the authors use a gated model that picks an expert network for a given input. As an extension, Arevalo et al. [Arevalo et al., 2017], proposed Gated Multimodal Units, applying this fusion strategy anywhere in the model and not only at the prediction-level. We build upon this seminal work in Chapter 9, where we propose a collaborative mixture of embedding experts that uses a pairwise gating fusion to fuse expert embeddings. In the same spirit, multimodal attention has also been integrated to existing architectures that use attention for temporal aggregation [Hori et al., 2017; Long, Gan, De Melo, Liu, et al., 2018].

B | Objective Functions

In this section we outline recent research related to learning constraints and objectives for multimodal and cross-modal learning.

A number of multimodal learning methods define constraints in order to control the relationship between unimodal features and/or the structure of neural network weights. Most losses for learning compact representations within a single modality have been driven by advances in face recognition. These losses can be loosely separated into two families: (i) classification-based losses such as the regular cross entropy loss, the additive angular margin loss [F. Wang et al., 2018; J. Deng et al., 2019], the center loss [Wen et al., 2016], and the congenerous cosine loss [Y. Liu et al., 2017]; and (ii) contrast based losses such as the contrastive loss [Hadsell et al., 2006], the triplet loss [Weinberger et al., 2006; Schroff et al., 2015] and the noise contrastive estimation (NCE) loss [Gutmann & Hyvärinen, 2010; Mnih & Kavukcuoglu, 2013].

In order to learn a shared representation (or a joint embedding between multiple modalities), there have been a number of different approaches. Ngiam et al. [Ngiam et al., 2011], proposed a bimodal autoencoder, forcing the hidden shared representation to be able to reconstruct both modalities, even in the absence of one of them. Andrew et al. [Andrew et al., 2013], adapted Canonical Correlation Analysis to deep neural networks, maximizing correlation between representations. Shahroudy et al. [Shahroudy et al., 2017], use cascading factorization layers to find shared representations between modalities and isolate modality-specific information. To ensure similarity between unimodal features, Engilberge et al. [Engilberge et al., 2018] minimize their cosine distance. In Chapter 6, we show that a simple contrastive loss can be used to learn a joint embedding between audio and visual modalities in a self-supervised manner. Since most of these methods are self-supervised (the only objective is to maximise similarity across corresponding samples in different modalities), many evaluate the representations learnt in a single modality for that modality alone [Arandjelović & Zisserman, 2017; Owens et al., 2016].

Structural constraints can also be applied on the very weights of the neural networks. In addition to modality dropping, [Neverova et al., 2015] zero-mask the cross-modal blocks of the weight matrix in early stages of training. Extending this idea of modality dropping, Li et al. [F. Li et al., 2016], learn a stochastic mask on the features. Another structure constraint in the form tensor factorization was proposed by [Ben-Younes et al., 2017]. Cross-modal correlations are also learnt through *cross-modal distillation*, where a teacher in one modality is used to guide weight learning of a student network, often by minimising the KL divergence between output predictions. We use an adapted version of this technique to learn good speech representations for emotion recognition in Chapter 4.

C | Statements of Authorship

For each multi-authored paper in this thesis, we provide a statement of authorship.

Each statement describes the candidate's and co-authors' independent research contributions for each thesis publications. For each publication in this thesis there is a complete statement that filled out and signed by the candidate and supervisor.

Statement of Authorship for the joint/multi-authored paper in Chapter 2: “Voxceleb: Large-scale speaker verification in the wild.”

Title of Paper	Voxceleb: Large-scale speaker verification in the wild.
Publication Status	Published
Publication Details	Published in Computer Speech and Language, 60, p.101027 2020. Arsha Nagrani* , Joon Son Chung*, Weidi Xie, and Andrew Zisserman. (* denotes Equal Contribution)

Student Confirmation

Student Name	Arsha Nagrani		
Contribution to the paper	1. Conception of the idea 2. Design and implementation of the speaker recognition models 3. Research of prior work 4. Writing and presentation of the paper		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor Name and Title	Professor Andrew Zisserman		
Supervisor Comments	This was a beneficial, joint and equal collaboration.		
Signature		Date	

Statement of Authorship for the joint/multi-authored paper in Chapter 3: “Speech2Action: Cross-modal Supervision for Action Recognition”

Title of Paper	Speech2Action: Cross-modal Supervision for Action Recognition
Publication Status	Published
Publication Details	Published in the proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2020 Arsha Nagrani , Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, Andrew Zisserman

Student Confirmation

Student Name	Arsha Nagrani		
Contribution to the paper	1. Joint conception and validation of the idea 2. Research of prior work 3. Creation of the IMSDb dataset 4. Implementation and running of all experiments 5. Writing and presentation of the paper		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor Name and Title	Professor Andrew Zisserman		
Supervisor Comments	This was an internship project.		
Signature		Date	

Statement of Authorship for the joint/multi-authored paper in Chapter 4: “Emotion Recognition in Speech using Cross-Modal Transfer in the Wild”

Title of Paper	Emotion Recognition in Speech using Cross-Modal Transfer in the Wild
Publication Status	Published
Publication Details	Published in the proceedings of ACM Multimedia, 2018 Samuel Albanie*, Arsha Nagrani* , Andrea Vedaldi, Andrew Zisserman (* denotes Equal Contribution)

Student Confirmation

Student Name	Arsha Nagrani		
Contribution to the paper	1. Joint conception of the idea 2. Design of the architecture and framework 3. Research of prior work 4. Writing and presentation of the paper		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor Name and Title	Professor Andrew Zisserman		
Supervisor Comments	This was a beneficial, joint and equal collaboration.		
Signature		Date	

Statement of Authorship for the joint/multi-authored paper in Chapter 5: “Seeing Voices and Hearing Faces: Cross-modal biometric matching”

Title of Paper	Seeing Voices and Hearing Faces: Cross-modal biometric matching
Publication Status	Published
Publication Details	Published in the proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2018 Arsha Nagrani , Samuel Albanie, Andrew Zisserman

Student Confirmation

Student Name	Arsha Nagrani		
Contribution to the paper	1. Conception of the idea 2. Design of the architecture and framework 3. Implementation and running of all experiments. 4. Research of prior work 5. Writing and presentation of the paper		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor Name and Title	Professor Andrew Zisserman		
Supervisor Comments	No comment required.		
Signature		Date	

Statement of Authorship for the joint/multi-authored paper in Chapter 6: “Learnable PINs: Cross-Modal Embeddings for Person Identity”

Title of Paper	Learnable PINs: Cross-Modal Embeddings for Person Identity
Publication Status	Published
Publication Details	Published in the proceedings of European Conference on Computer Vision (ECCV), 2018 Arsha Nagrani* , Samuel Albanie*, Andrew Zisserman (* denotes Equal Contribution)

Student Confirmation

Student Name	Arsha Nagrani		
Contribution to the paper	1. Conception of the idea 2. Design of the architecture and framework 3. Implementation and running of all experiments. 4. Research of prior work 5. Writing and presentation of the paper		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor Name and Title	Professor Andrew Zisserman		
Supervisor Comments	This was a beneficial, joint and equal collaboration.		
Signature		Date	

Statement of Authorship for the joint/multi-authored paper in Chapter 7: “Disentangled Speech Embeddings using Cross-Modal Self-Supervision”

Title of Paper	Disentangled Speech Embeddings using Cross-Modal Self-Supervision
Publication Status	Published
Publication Details	Published in the proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020 Arsha Nagrani* , Joon Son Chung*, Samuel Albanie*, Andrew Zisserman (* Equal Contribution)

Student Confirmation

Student Name	Arsha Nagrani		
Contribution to the paper	1. Joint conception of the idea 2. Research of prior work 3. Writing and presentation of the paper		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor Name and Title	Professor Andrew Zisserman		
Supervisor Comments	This was a beneficial, joint and equal collaboration.		
Signature		Date	

Statement of Authorship for the joint/multi-authored paper in Chapter 8: “EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition”

Title of Paper	EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition
Publication Status	Published
Publication Details	Published in the proceedings of International Conference on Computer Vision (ICCV), 2019 Evangelos Kazakos, Arsha Nagrani , Andrew Zisserman, Dima Damen.

Student Confirmation

Student Name	Arsha Nagrani		
Contribution to the paper	1. Joint conception of the idea 2. Joint design of the architecture and framework 3. Writing and presentation of the paper		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor Name and Title	Professor Andrew Zisserman		
Supervisor Comments	This was a beneficial and joint collaboration.		
Signature		Date	

Statement of Authorship for the joint/multi-authored paper in Chapter 9: “Use What You Have: Video Retrieval Using Representations From Collaborative Experts”

Title of Paper	Use What You Have: Video Retrieval Using Representations From Collaborative Experts
Publication Status	Published
Publication Details	Published in the proceedings of British Machine Vision Conference (BMVC), 2019 Yang Liu*, Samuel Albanie*, Arsha Nagrani* , Andrew Zisserman (* Equal Contribution)

Student Confirmation

Student Name	Arsha Nagrani		
Contribution to the paper	1. Joint Conception of the idea 2. Joint design of the architecture and framework 3. Implementation and running of some experiments. 4. Research of prior work 5. Writing and presentation of the paper		
Signature		Date	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor Name and Title	Professor Andrew Zisserman		
Supervisor Comments	This was a beneficial, joint and equal collaboration.		
Signature		Date	