Supplementary Material for *Speech2Action:* Cross-modal Supervision for Action Recognition

Arsha Nagrani[1,2] Chen Sun[2] David Ross[2]
Rahul Sukthankar[2] Cordelia Schmid[2] Andrew Zisserman[1,3]
[1]VGG, Oxford [2]Google Research [3]DeepMind
https://www.robots.ox.ac.uk/~vgg/research/speech2action/

In this supplementary material we include additional details and results for training the Speech2Action model in Sec. A. In Sec. B, we show more results for the techniques used to mine training samples – i.e. the Keyword Spotting Baseline and the Speech2Action model. Finally, we show results on the UCF101 [9] dataset in Sec. C.

A. `Speech2Action` model

A.1. Screenplay Parsing

We follow the grammar created by Winer et al. [13] which is based on 'The Hollywood Standard' [8], an authoritative guide to screenplay writing, to parse the screenplays and separate out various script elements. The tool uses spacing, indentation, capitalisation and punctuation to parse screenplays into the following four different elements:

1. Shot Headings – These are present at the start of each scene or shot, and may give general information about a scene's location, type of shot, subject of shot, or time of day, e.g. `INT. CENTRAL PARK - DAY`
2. Stage Direction – This is the stage direction that is to be given to the actors. This contains the *action* information that we are interested in, and is typically a paragraph containing many sentences, e.g. `Nason and his guys fight the fire. They are CHOKING on smoke. PAN TO Ensign Menendez, leading in a fresh contingent of men to join the fight. One of them is TITO.`
3. Dialogue – speech uttered by each character, e.g. `INDY: Get down!`
4. Transitions – may appear at the end of a scene, and indicate how one scene links to the next, e.g. `HARD CUT TO:`

In this work we only extract 2. Stage Direction, and 3. Dialogue. After mining for verbs in the stage directions, we then search for the nearest section of dialogue (either before or after) and assign each sentence in the dialogue with the verb class label (see Fig. 2 for examples of verb-speech pairs obtained from screenplays).

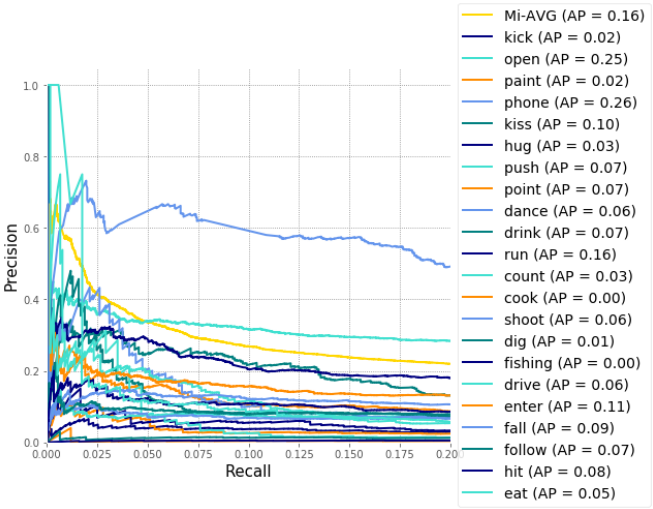

Figure 1. PR curves on the validation set of the IMSDb dataset for the `Speech2Action` model. Since the validation set is noisy, we are only interested in performance in the low recall, high precision setting. Note how some classes – 'phone', 'open' and 'run' perform much better than others.

A.2. PR Curves on the Validation Set of the IMSDb Data

We show precision-recall curves on the val set of the IMSDb dataset in Fig. 1. Note how classes such as 'run' and 'phone' have a much higher recall for the same level of precision.

We select thresholds for the `Speech2Action` model using a greedy search as follows: (1) We allocate the retrieved samples into discrete precision buckets (30%-40%, 40%-50%, etc.), using thresholds obtained from the PR curve mentioned above; (2) For different actions, we adjust the buckets to make sure the number of training ex-

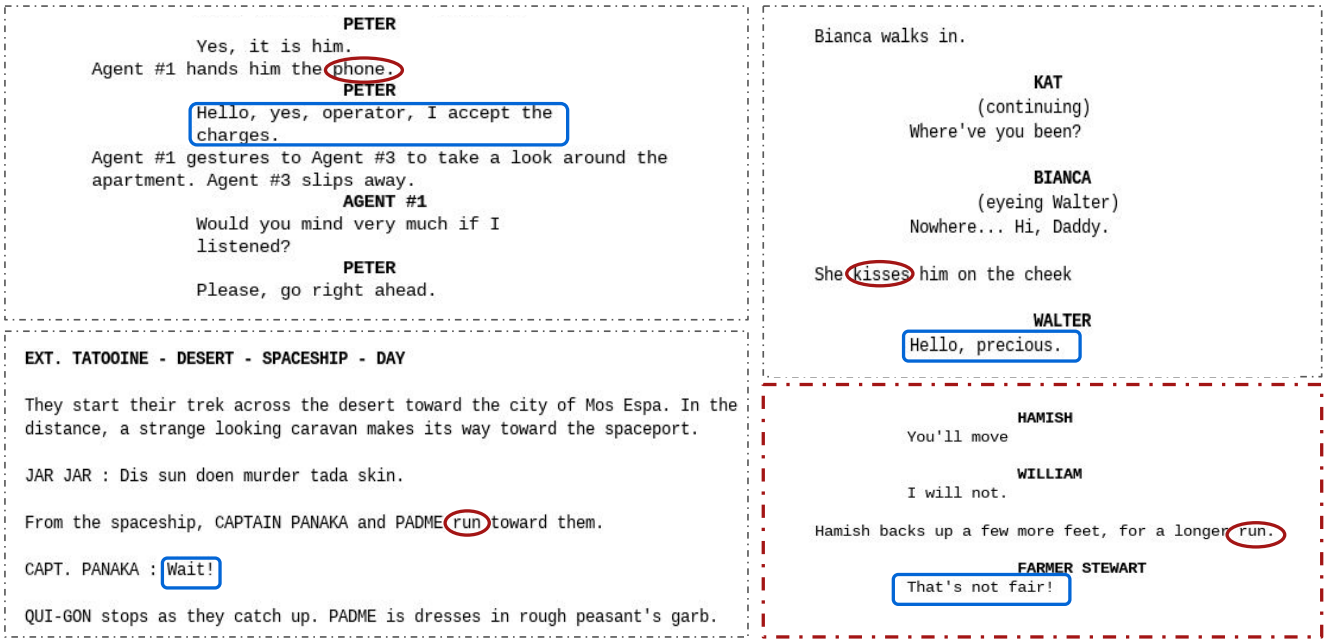

Figure 2. Examples of speech and verb action pairs obtain from screenplays. In the bottom row (right) we show a possibly negative speech and verb pair, i.e. the speech segment *That's not fair!* is assigned the action verb 'run', whereas it is not clear that these two are correlated.

phone	why didn't you return my phone calls? you each get one phone call i already got your phone line set up. but my phone died, so just leave a message, okay? i'm on the phone.. we're collecting cell phones, surveillance tapes, video we can find.
dance	she went to the dance with Harry Land do you wanna dance? and the dance of the seven veils? what if i pay for a dance? the dance starts in an hour. just dance.
drink	are you drunk? my dad would be drinking somewhere else. you didn't drink the mold. let's go out and drink. super bowl is the super bowl of drinking. i don't drink, i watch my diet, but no.
kiss	they were both undone by true love's kiss. good girls don't kiss and tell. kiss my a** it was our first kiss. i mean, when they say, "i'll call you," that's the kiss of death. i had to kiss jace.
eat	against a top notch britisher, you'll be eaten alive. eat my dust, boys! ate something earlier. i can't eat, i can't sleep. you must eat the sardines tomorrow. i ate bad sushi.
point	and you can add someone to an email chain at any point. she's got a point, buddy. the point is, they're all having a great time. didn't advance very far, i think, is mark's point. you made your point. beside the point!

Table 1. **Examples of speech samples for six verb categories labelled with the keyword spotting baseline.** Each block shows the action verb on the left, and the speech samples on the right. Since we do not need to use the movie screenplays for this baseline, unlike `Speech2Action` (results in Table. 2 of the main paper), we show examples of transcribed speech obtained directly from the unlabelled corpus. Note how the speech labelled with the verb 'point' is indicative of a different semantic meaning to the physical action of 'pointing'.

amples are roughly balanced for all classes; (3) For classes with low precision, in order to avoid picking uncertain and hence noiser predictions, we only select examples that had a precision above 30%+.

The number of retrieved samples per class can be seen in Fig. 3. The number of retrieved samples for 'phone' and 'open' at a precision value of 30% are in the millions (2,272,906 and 31,657,295 respectively), which is why we manually increase the threshold in order to prevent a large class-imbalance during training. We reiterate here once again that this evaluation is performed purely on the basis of the proximity of speech to verb class in the stage direction of the movie screenplay (Fig. 2), and hence 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. There are many cases in the movie screenplays where verb and speech pairs could

be completely uncorrelated (see Fig. 2, bottom–right for an example.)

B. Mining Techniques

B.1. Keyword Spotting Baseline

In this section we provide more details about the Keyword Spotting Baseline (described in Sec. 4.2.2 of the main paper). The total number of clips mined using the Keyword Spotting Baseline is 679,049. We mine all the instances of speech containing the verb class, and if there are more than 40K samples, we randomly sample 40K clips. The reason we cap samples at 40K is to prevent overly unbalanced classes. Examples of speech labelled with this baseline for 6 verb classes can be seen in Table 1. There are two ways in which our learned `Speech2Action` model is theoretically superior to this approach:
(1) Many times the speech correlated with a particular action does not actually contain the action verb itself e.g. *'Look over there'* for the class 'point'.
(2) There is no *word-sense disambiguation* in the way the speech segments are mined, i.e. *'Look at where I am pointing'* vs *'You've missed the point'*. Word-sense disambiguation is the task of identifying which sense of a word is used in a sentence when a word has multiple meanings. This task tends to be more difficult with verbs than nouns because verbs have more senses on average than nouns and may be part of a multiword phrase [1].

B.2. Mined Examples

The distribution of mined examples per class for all 18 classes, using the Speech2Action model and the Keyword Spotting baseline can be seen in Figures 3 and 4. We note that it is very difficult to mine examples for actions 'hug' and 'kick', as these are often accompanied with speech similar to that accompanying 'kiss' and 'hit'.

We show more examples of automatically mined video clips from unlabelled movies using the `Speech2Action` model in Fig. 5. Here we highlight in particular the diversity of video clips that are mined using simply speech *alone*, including diversity in objects, viewpoints and background scenes.

C. Results on UCF101

In this section we show the results of pretraining on our mined video examples and then finetuning on the UCF101 dataset [9], following the exact same procedure described in Sec. 5.1 of the main paper. UCF101 [9] is a dataset of 13K videos downloaded from YouTube spanning over 101 human action classes. Our results follow a similar trend to those on HMDB51, pretraining on samples mined using `Speech2Action` (81.4%) outperforms training from scratch (74.2%) and pretraining on samples obtained using the keyword spotting basline (77.4%). We note here, however, that it is much harder to tease out the difference between various styles of pretraining on this dataset, because it is more saturated than HMDB51 (training from scratch already yields a high accuracy of 74.2%, and pretraining on Kinetics largely solves the task, with an accuracy of 95.7%).

Method	Architecture	Pre-training	Acc.
Shuffle&Learn [7]⋆	S3D-G (RGB)	UCF101† [9]	50.2
OPN [6]	VGG-M-2048	UCF101† [9]	59.6
ClipOrder [14]	R(2+1)D	UCF101† [9]	72.4
Wang et al. [12]	C3D	Kinetics† [9]	61.2
3DRotNet [4]⋆	S3D-G (RGB)	Kinetics†	75.3
DPC [3]	3DResNet18	Kinetics†	75.7
CBT [10]	S3D-G (RGB)	Kinetics†	79.5
DisInit (RGB) [2]	R(2+1)D-18 [11]	Kinetics**	85.7
Korbar et al [5]	I3D (RGB)	Kinetics†	83.7
-	S3D-G (RGB)	Scratch	74.2
Ours	S3D-G (RGB)	`KSB-mined`	77.4
Ours	S3D-G (RGB)	`S2A-mined`	81.4
Supervised pretraining	S3D-G (RGB)	ImageNet	84.4
Supervised pretraining	S3D-G (RGB)	Kinetics	95.7

Table 2. **Comparison with previous pre-training strategies for action classification on UCF101.** Training on videos labelled with `Speech2Action` leads to a 7% improvement over training from scratch and outperforms previous self-supervised works. It also performs competitively with other 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 [5], we report the number achieved by their I3D (RGB only) model which is the closest to our architecture. For ⋆, we report the reimplementations by [10] using the S3D-G model (same as ours). For the rest, we report performance directly from the original papers.

References

[1] Luciano Del Corro, Rainer Gemulla, and Gerhard Weikum. Werdy: Recognition and disambiguation of verbs and verb phrases with syntactic and semantic pruning. 2014. 3

[2] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. Distinit: Learning video representations without a single labeled video. *ICCV*, 2019. 3

[3] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 3

[4] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2018. 3

[5] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 3

[6] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sort-

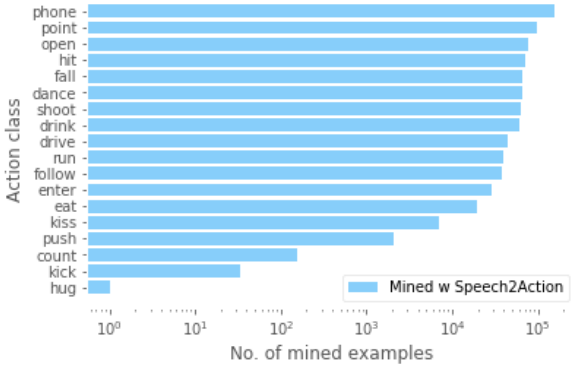

Figure 3. **Distribution of training clips mined using `Speech2Action`.** We show the distribution for *all* 18 verb classes. It is difficult to mine clips for the actions 'hug' and 'kick', as these are often confused with 'kiss' and 'hit'.

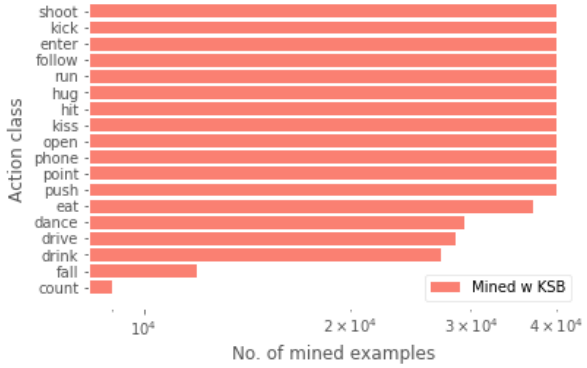

Figure 4. **Distribution of training clips mined using the Keyword Spotting baseline.** We show the distribution for *all* 18 verb classes. We cut off sampling at 40K samples for twelve classes in order to prevent too much of a class imbalance.

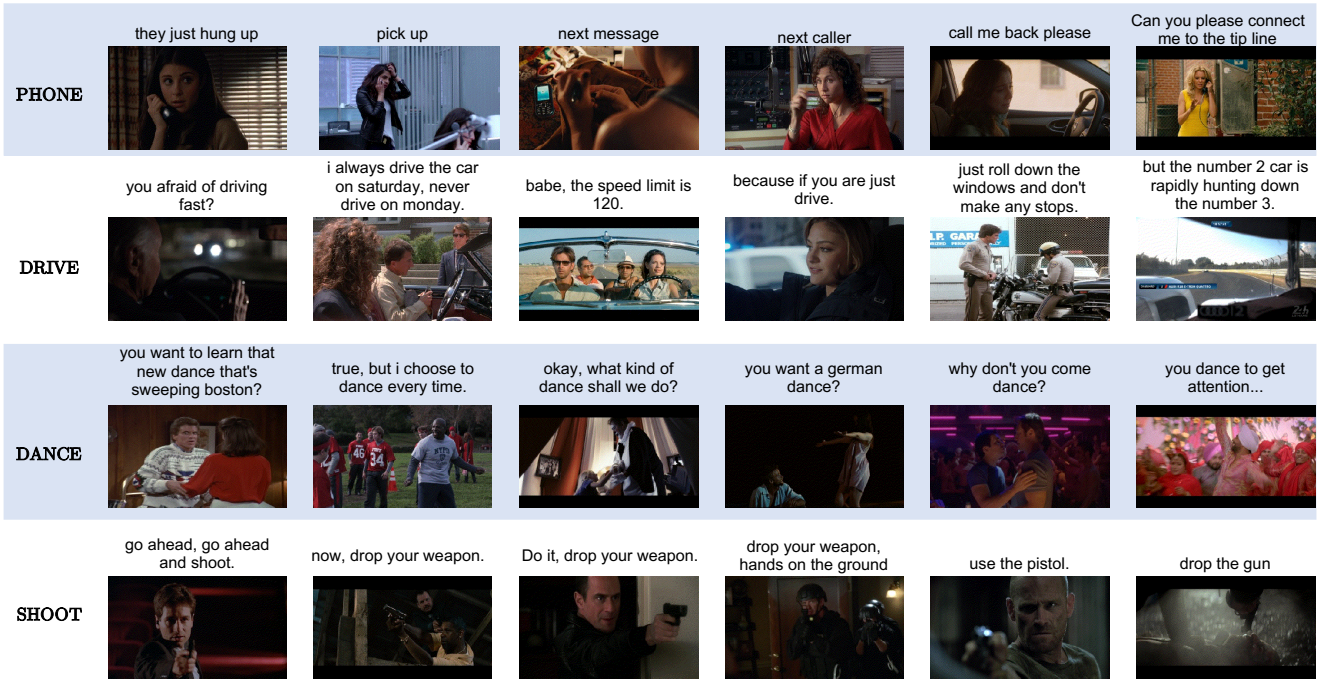

Figure 5. **Examples of clips mined automatically using the `Speech2Action` model applied to *speech alone* for 4 AVA classes.** We show only a single frame from each video. Note the diversity in *object* for the category '[answer] phone' (first row, from left to right) a landline, a cell phone, a text message on a cell phone, a radio headset, a carphone, and a payphone, in *viewpoint* for the category 'drive' (second row) including behind the wheel, from the passenger seat, and from outside the car, and in *background* for the category 'dance' (third row, from left to right) inside a home, on a football pitch, in a tent, outdoors, in a club/party and at an Indian wedding/party.

ing sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 3

[7] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3

[8] Christopher Riley. *The Hollywood standard: the complete and authoritative guide to script format and style*. Michael Wiese Productions, 2009. 1

[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 3

[10] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019. 3

[11] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

[12] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. 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*, pages 4006–4015, 2019. 3

[13] David R Winer and R Michael Young. Automated screenplay annotation for extracting storytelling knowledge. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2017. 1

[14] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 3