BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues
APPENDIX

Samuel Albanie[1⋆], Gül Varol[1⋆], Liliane Momeni[1], Triantafyllos Afouras[1], Joon Son Chung[1,2], Neil Fox[3], and Andrew Zisserman[1]

[1]Visual Geometry Group, University of Oxford, UK
[2]Naver Corporation, Seoul, South Korea
[3]Deafness Cognition and Language Research Centre, University College London, UK
{albanie,gul,liliane,afourast,joon,az}@robots.ox.ac.uk;
neil.fox@ucl.ac.uk

This document provides additional results (Section A), details about the video pose distillation model (Section B), and about the BSL-1K dataset (Section C).

A Additional Results

In this section, we present complementary results to the main paper. Section A.1 provides a qualitative analysis. Additional experiments investigate the search window size for mouthing (Section A.2), the number of frames for sign recognition (Section A.3), the effect of masking the mouth at test time (Section A.4), ensembling part-specific models (Section A.5), the transfer to co-articulated datasets (Section A.6), and the baselines using other cues (Section A.7).

A.1 Qualitative analysis

We provide a video on our project page[1] to illustrate the automatically annotated training samples in our BSL-1K dataset, as well as the results of our sign recognition model on the manually verified test set. Figures A.1 and A.2 present some of these results. In Figure A.1, we provide *training* samples localised using mouthing cues. In Figure A.2, we provide *test* samples classified by our I3D model trained on the automatic annotations.

A.2 Size of the search window for visual keyword spotting

We investigate the influence of varying the extent of the temporal window around a given subtitle during the visual keyword spotting phase of dataset collection. For this experiment, we run the visual keyword spotting model with different search window sizes (centring these windows on the subtitle locations), and train

⋆ Equal contribution
[1] https://www.robots.ox.ac.uk/~vgg/research/bsl1k/

Fig. A.1. Mouthing results: Qualitative samples for the visual keyword spotting method for the keywords "happy" and "important". We visualise the top 24 videos with the most confident mouthing scores for each word. We note the visual similarity among manual features which suggests that mouthing cues can be a good starting point to automatically annotate training samples.

Fig. A.2. Sign recognition results: Qualitative samples for our sign language recognition I3D model on the BSL-1K test set. We visualise the top 24 videos with the highest classification scores for the signs "orange" and "business", which appear to be all correctly classified.

Table A.1. The effect of the temporal window where we apply the visual keyword spotting model. Networks are trained on BSL-1K$_{m.8}$ with Kinetics initialisation. Decreasing the window size increases the chance of missing the word, resulting in less training data and lower performance. Increasing too much makes the keyword spotting task difficult, reducing the annotation quality. We found 8 seconds to be a good compromise, which we used in all other experiments in this paper.

		per-instance		per-class	
Keyword search window	#videos	top-1	top-5	top-1	top-5
1 sec	25.0K	60.10	75.42	36.62	53.83
2 sec	33.9K	64.91	80.98	40.29	59.63
4 sec	37.6K	68.09	82.79	45.35	63.64
8 sec	38.9K	**69.00**	**83.79**	**45.86**	**64.42**
16 sec	39.0K	65.91	81.84	39.51	59.03

Table A.2. The effect of the number of frames before the mouthing peak used for training. Networks are trained on BSL-1K$_{m.8}$ with Kinetics initialisation.

	per-instance		per-class	
#frames	top-1	top-5	top-1	top-5
16	59.53	77.08	36.16	58.43
20	**71.71**	**85.73**	**49.64**	**69.23**
24	69.00	83.79	45.86	64.42

sign recognition models (following the protocol described in the main paper, using Kinetics initialisation) on the resulting annotations. We find (Table A.1) that 8-second extended search windows yield the strongest performance on the test set (which is fixed across each run)—we therefore use these for all experiments used in the main paper.

A.3 Temporal extent of the automatic annotations

Keyword spotting provides a precise localisation in time, but does not determine the *duration* of the sign. We observe that the highest mouthing confidence is obtained at the *end* of mouthing. We therefore take a certain number of frames before this peak to include in our sign classification training. In Table A.2, we experiment with this hyper-parameter and see that 20 frames is a good compromise for creating variation in training, while not including too many irrelevant frames. In all of our experiments, we used 24 frames, except in Table 6 of the main paper which combines the best parameters from each ablation, where we used 20 frames. Note that our I3D model takes in 16 consecutive frames as input, which is sliced randomly during training.

A.4 Masking the mouth region at test time

In Table A.3, we experiment with the test modes for the networks trained with (i) full-frames including the mouth, versus (ii) masking the mouth region. If

Fig. A.3. Masking the mouth: Sample visualisations for the inputs described in Table 4 of the main paper (for the signs "important" and "library"). We experiment with masking the mouth region or cropping only the face region using the detected pose keypoints.

Table A.3. We complement Table 4 of the main paper by investigating different test modes for I3D, when trained with or without the mouth pixels. The model trained with full-frames relies significantly on the mouth, whose performance drops from 65.57% to 34.74% when the mouth is masked. The models are trained on the subset of BSL-1K$_{m.8}$ where pose estimates are available.

	Test mouth-masked				Test full-frame			
	per-instance		per-class		per-instance		per-class	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
Train mouth-masked	**46.75**	**66.34**	**25.85**	**48.02**	46.21	65.34	25.83	46.23
Train full-frame	34.74	51.42	13.62	29.80	**65.57**	**81.33**	**44.90**	**64.91**

the mouth is masked only at test time, the performance drops from 65.57% to 34.74% suggesting the model's significant reliance on the mouth cues. The model can be improved to 46.75% if it is forced to focus on other regions by training with masked mouth.

Table A.4. Ensembling part-specific models from Table 4 of the main paper. We observe that combining the I3D model trained only with the face and another model without the mouth (last row) achieves superior performance than using one model that inputs the full-frame. This suggests that disentangling manual and non-manual features, which are complementary, for sign recognition is a promising direction. The models are trained on the subset of BSL-1K$_{m.8}$ where pose estimates are available.

		per-instance		per-class	
		top-1	top-5	top-1	top-5
face-crop		42.23	69.70	21.66	50.51
mouth-masked		46.75	66.34	25.85	48.02
full-frame		**65.57**	**81.33**	**44.90**	**64.91**
full-frame	& face-crop	64.50	83.01	42.30	65.58
full-frame	& mouth-masked	68.09	81.33	**46.29**	65.41
mouth-masked	& face-crop	**68.55**	**83.63**	45.29	**67.47**

A.5 Late fusion of part-specific models.

We further experiment with ensembling two I3D networks, each specialising on different parts of the human body, by averaging the classification scores, i.e., late fusion. The results are summarised in Table A.4. We observe significant improvements when combining a mouth-specific model (face-crop) with a body-specific model (mouth-masked), which suggests that forcing the network to focus on separate, but complementary signing cues (68.55%) can be more effective than presenting the full-frames (65.57%). This procedure; however, involves additional complexity of computing the human pose and training two separate models. It is therefore only used for experimental purposes.

Figure A.3 presents sample visualisations for the masking procedure. For mouth-masking, we replace the box covering the mouth region with the average pixel of the region. For face-cropping, we set pixels outside of the face region to zero (we observed divergence of training if the mean value was used).

A.6 Transferring BSL-1K pretrained model to other datasets

As explained in Section 5.3 of the main paper, we use our model pretrained on BSL-1K as initialisation for transferring to other datasets.

Additional details on fine-tuning for ASL. For MSASL [5] and WLASL [7] isolated datasets on ASL, we have used the pretraining with the mouth-masking to force the model to entirely pay attention to manual features. We also observed that some signs are identical between ASL and BSL; therefore, instead of randomly initialising the last classification layer, we have kept the weights corresponding to common words between BSL-1K and the ASL dataset. We observed slight improvements with both of these choices.

Results on co-articulated datasets. Here, we report the results of training sign language recognition on two co-articulated datasets: (i) RWTH-PHOENIX-Weather-2014-T [6,1] and (ii) BSL-Corpus [11,12], with and without pretraining on BSL-1K.

Phoenix dataset is not directly applicable to our model due to the lack of sign-gloss alignment to train I3D with short clips of individual signs. We therefore implemented a simple CTC loss [4] to adapt I3D for Phoenix and obtained 5.6 WER improvement with BSL-1K pretraining over Kinetics pretraining.

BSL-Corpus is a linguistic dataset, and has not been used for computer vision research so far. We defined a train/val/test split (8:1:1 ratio) for a subset of 6k annotations of 966 signs and obtained 24.4% vs 12.8% accuracy with/without BSL-1K pretraining. In this case, we have also kept the last-layer classification weights for which the words are in common between BSL-Corpus and BSL-1K signs. We observed this to provide small gains over completely random initialisation of classification weights.

We conclude that our large-scale BSL-1K dataset provides a strong initialisation for both co-articulated and isolated datasets; for a variety of sign languages: ASL (American), BSL (British), and DGS (German).

A.7 Dataset expansion through other cues and additional baselines

In addition to the experiments reported in the paper, we further implemented the dataset labelling technique described in [10] which searches subtitles for signs and picks candidate temporal windows that maximise the area under the ROC curve for positively and negatively labelled bags (here, a positive bag refers to temporal windows that occur within an approximately 400 frame interval centred on the target word). However, we found that without the use of the keyword spotting model for localisation, the annotations collected with this technique were extremely noisy, and the resulting model significantly under-performed all baselines reported in the main paper (that were instead trained on BSL-1K). We also experimented with dataset expansion through training ensembles of exemplar SVMs [9] for each episode on signs that were predicted as confident positives (greater than 0.8) by our strongest pretrained model (and using all temporal windows that did not include the keyword as negatives). In this case, we found it challenging to calibrate SVM confidences (we explored both the parameters of the original paper, who discuss the difficulties of this process [9] and a range of other parameters) and expanded the dataset by a factor of three, but did not achieve a boost in model performance when training on the expanded data.

B Video Pose Distillation

In this section, we give additional details of the video pose distillation model described in Section 4.3 of the main paper. The model architecture uses an I3D backbone [3] and takes as input a sequence of 16 frames at 224×224 pixels. We remove the classification head used in the original architecture and replace it with a linear layer that projects the 1024-dimensional embedding to 4160 dimensions—this corresponds to a set of 16 per-frame predictions of the xy coordinates of the 130 human pose keypoints produced by an OpenPose [2] model (trained on COCO [8]). The coordinates are normalised (with respect to the dimensions of the input image) to lie in the range $[0, 1]$ and an L2 loss is used to penalise inaccurate predictions. The training data for the pose distillation model comprises one-minute segments from each episode used in the BSL-1K training set. The model is trained for 100 epochs using the Lookahead optimizer [13] with minibatches of 32 clips using a learning rate of 0.1 (reduced by a factor of 10 after 50 epochs) and a weight decay of 0.001.

C BSL-1K Dataset Details

C.1 Sign verification tool

In Figure A.4, we show a screenshot of the verification tool used by annotators to verify or reject signs found by the proposed keyword spotting method in the test set. Annotators have the ability to view the sign at reduced speed and indicate whether the sign is correct, incorrect, or they are unsure.

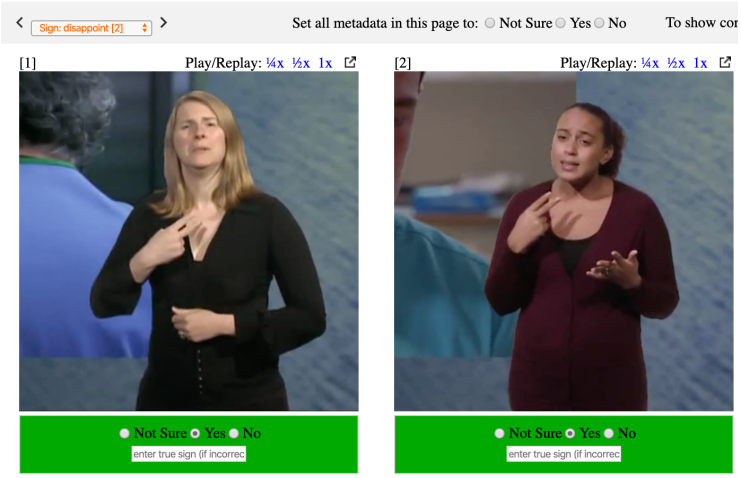

Fig. A.4. Manual annotation: A screenshot of the Whole-Sign Verification Tool.

C.2 Dataset source material

The BBC broadcast TV shows, together with their respective number of occurrences in the source material used to construct the dataset are:

Countryfile: 266, Natural World: 70, Great British Railway Journeys: 122, Holby City: 261, Junior Masterchef: 24, Junior Bake Off: 22, Hairy Bikers Bakeation: 6, Masterchef The Professionals: 37, Doctor Who Sci Fi: 23, Great British Menu: 110, A To Z Of Cooking: 24, Raymond Blanc Kitchen Secrets: 9, The Apprentice: 88, Country Show Cook Off: 18, A Taste Of Britain: 20, Lorraine Pascale How To Be A: 6, Chefs Put Your Menu Where Your: 13, Simply Nigella: 7, The Restaurant Man: 5, Hairy Bikers Best Of British: 27, Rip Off Britain Food: 20, Our Food Uk 4: 3, Disaster Chefs: 8, Terry And Mason Great Food Trip: 19, Gardeners World: 70, Paul Hollywood Pies Puds: 20, James Martin Food Map Of Britain: 10, Baking Made Easy: 6, Hairy Bikers Northern: 7, Nigel Slater Eating Together: 6, Raymond Blanc How To Cook Well: 6, Great British Food Revival: 17, Great British Bake Off: 28, Two Greedy Italians: 4, Food Fighters: 10, Hairy Bikers Mums Know Best: 9, Hairy Bikers Meals On Wheels: 6, Paul Hollywood Bread 6: 5, Home Comfort At Christmas: 1

C.3 Dataset vocabulary

The 1,064 words which form the vocabulary for BSL-1K are:

abortion, about, above, absorb, accept, access, act, activity, actually, add, address, advance, advertise, afford, afghanistan, africa, afternoon, again, against, agree, aids, alcohol, all, already, always, amazed, america, angel, angry, animal, answer, anything, anyway, apple, apprentice, approach, april, apron, arch, archery, architect, area, argue,

arm, army, around, arrive, arrogant, art, asian, ask, assess, atmosphere, attack, attention, attitude, auction, australia, austria, automatic, autumn, average, award, awful, baby, back, bacon, bad, balance, ball, ballet, balloon, banana, bank, barbecue, base, basketball, bath, battery, beach, beat, because, bedroom, beef, been, before, behind, belfast, belgium, believe, belt, better, big, billion, bingo, bird, birmingham, birthday, biscuit, bite, bitter, black, blackpool, blame, blanket, blind, blonde, blood, blue, boat, body, bomb, bone, bonnet, book, booked, border, boring, born, borrow, boss, both, bottle, boundary, bowl, box, boxing, branch, brave, bread, break, breathe, brick, bridge, brief, brighton, bring, bristol, britain, brother, brown, budget, buffet, build, bulgaria, bull, bury, bus, bush, business, but, butcher, butter, butterfly, buy, by, cabbage, calculator, calendar, call, camel, camera, can, canada, cancel, candle, cap, captain, caravan, cardiff, careful, carpenter, castle, casual, cat, catch, catholic, ceiling, cellar, certificate, chair, chalk, challenge, chance, change, character, charge, chase, cheap, cheat, check, cheeky, cheese, chef, cherry, chicken, child, china, chips, chocolate, choose, christmas, church, city, clean, cleaner, clear, clever, climb, clock, closed, clothes, cloud, club, coffee, coffin, cold, collapse, collect, college, column, combine, come, comedy, comfortable, comment, communicate, communist, community, company, compass, competition, complicated, compound, computer, concentrate, confident, confidential, confirm, continue, control, cook, copy, corner, cornwall, cottage, council, country, course, court, cousin, cover, cracker, crash, crazy, create, cricket, crisps, cross, cruel, culture, cup, cupboard, curriculum, custard, daddy, damage, dance, danger, daughter, deaf, debate, december, decide, decline, deep, degree, deliver, denmark, dentist, depend, deposit, depressed, depth, derby, desire, desk, detail, detective, devil, different, dig, disabled, disagree, disappear, disappoint, discuss, disk, distract, divide, dizzy, doctor, dog, dolphin, donkey, double, downhill, drawer, dream, drink, drip, drive, drop, drunk, dry, dublin, dvd, each, early, east, easter, easy, edinburgh, egypt, eight, elastic, electricity, elephant, eleven, embarrassed, emotion, empty, encourage, end, engaged, england, enjoy, equal, escalator, escape, ethnic, europe, evening, everything, evidence, exact, exchange, excited, excuse, exeter, exhausted, expand, expect, expensive, experience, explain, express, extract, face, factory, fairy, fall, false, family, famous, fantastic, far, farm, fast, fat, father, fault, fax, february, feed, feel, fence, fifteen, fifty, fight, film, final, finance, find, fine, finish, finland, fire, fireman, first, fish, fishing, five, flag, flat, flock, flood, flower, fog, follow, football, for, foreign, forever, forget, fork, formal, forward, four, fourteen, fox, france, free, freeze, fresh, friday, friend, frog, from, front, fruit, frustrated, fry, full, furniture, future, game, garden, general, generous, geography, germany, ginger, girl, give, glasgow, glass, gold, golf, gorilla, gossip, government, grab, grandfather, grandmother, greedy, green, group, grow, guarantee, guess, guilty, gym, hair, half, hall, hamburger, hammer, hamster, handshake, hang, happen, happy, hard, hat, have, headache, hearing, heart, heavy, helicopter, hello, help, hide, history, holiday, holland, home, hope, hopeless, horrible, horse, hospital, hot, hotel, hour, house, how, hundred, hungry, hypocrite, idea, if, ignore, imagine, impact, important, impossible, improve, income, increase, independent, india, influence, inform, information, injection, insert, instant, insurance, international, internet, interrupt, interview, introduce, involve, ireland, iron, italy, jamaica, january, japan, jealous, jelly, jersey, join, joke, jumper, just, kangaroo, karate, keep, kitchen, label, language, laptop, last, late, later, laugh, leaf, leave, leeds, left, lemon, library, lighthouse, lightning, like, line, link, list, little, live, liverpool, london, lonely, lost, love, lovely, machine, madrid, magic, make, man, manage, manchester, many, march, mark, marry, match, maybe, meaning, meat, mechanic, medal, meet, meeting, melon, member, mention, message, metal, mexico, milk, million, mind, minute, mirror, miserable, miss, mistake, mix, monday, money,

monkey, month, more, morning, most, mother, motorbike, mountain, mouse, move, mum, music, must, name, nasty, national, naughty, navy, necklace, negative, nervous, never, newcastle, newspaper, next, nice, nightclub, normal, north, norway, not, nothing, notice, november, now, number, nursery, object, october, offer, office, old, on, once, one, onion, only, open, operate, opposite, or, oral, orange, order, other, out, oven, over, overtake, owl, own, pack, page, pager, paint, pakistan, panic, paper, paris, park, partner, party, passport, past, pattern, pay, payment, pence, penguin, people, percent, perfect, perhaps, period, permanent, person, personal, persuade, petrol, photo, piano, picture, pig, pineapple, pink, pipe, place, plain, plan, plaster, plastic, plate, please, plenty, plumber, plus, point, poland, police, politics, pop, popular, porridge, portsmouth, portugal, posh, poster, potato, pound, practise, praise, prefer, pregnant, president, pretend, price, priest, print, prison, problem, professional, professor, profile, profit, project, promise, promote, protestant, proud, provide, pub, pull, pulse, punch, purple, push, pyramid, quality, quarter, question, quick, quiet, quit, rabbit, race, radio, rain, rather, read, reading, ready, really, receipt, receive, recommend, red, reduce, region, regular, relationship, relax, release, relief, remember, remind, remove, rent, repair, replace, research, resign, respect, retire, review, rhubarb, rich, ride, right, river, rocket, roll, roman, room, roots, rough, round, rub, rubbish, rugby, run, russia, sack, sad, safe, same, sand, sandwich, satellite, saturday, sauce, sausage, school, science, scissors, scotland, scratch, search, second, seed, seem, self, sell, send, sense, sensitive, sentence, separate, september, sequence, service, settle, seven, sex, shadow, shakespeare, shame, shark, sharp, sheep, sheet, sheffield, shine, shirt, shop, should, shoulder, shout, show, shower, sick, sight, sign, silver, similar, since, sister, sit, six, size, skeleton, skin, sleep, sleepy, small, smell, smile, snake, soft, some, sometimes, son, soon, sorry, south, spain, specific, speech, spend, spicy, spider, spirit, split, sport, spray, spread, squash, squirrel, staff, stand, star, start, station, still, story, straight, strange, stranger, strawberry, stress, stretch, strict, string, strong, structure, stubborn, stuck, student, study, stupid, success, sugar, summer, sun, sunday, sunset, support, suppose, surprise, swan, swap, sweden, sweep, swim, swing, switzerland, sympathy, take, talk, tap, taste, tax, taxi, teacher, team, technology, television, temperature, temporary, ten, tennis, tent, terminate, terrified, that, theory, therefore, thing, think, thirsty, thousand, three, through, thursday, ticket, tidy, tiger, time, tired, title, toast, together, toilet, tomato, tomorrow, toothbrush, toothpaste, top, torch, total, touch, tough, tournament, towel, town, train, training, tram, trap, travel, tree, trouble, trousers, true, try, tube, tuesday, turkey, twelve, twenty, twice, two, ugly, ultrasound, umbrella, uncle, under, understand, unemployed, union, unit, university, until, up, upset, valley, vegetarian, video, vinegar, visit, vodka, volcano, volunteer, vote, wait, wales, walk, wall, want, war, warm, wash, waste, watch, water, weak, weather, wednesday, week, weekend, well, west, wet, what, wheelchair, when, where, which, whistle, white, who, why, wicked, width, wild, will, win, wind, window, wine, with, without, witness, wolf, woman, wonder, wonderful, wood, wool, work, world, worry, worship, worth, wow, write, wrong, yes, yesterday.

References

1. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: CVPR (2018) 6
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018) 7
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the Kinetics dataset. In: CVPR (2017) 7
4. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: ICML (2006) 6
5. Joze, H.R.V., Koller, O.: MS-ASL: A large-scale data set and benchmark for understanding american sign language. In: BMVC (2019) 6
6. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015) 6
7. Li, D., Opazo, C.R., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: WACV (2019) 6
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 7
9. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV (2011) 7
10. Pfister, T., Charles, J., Zisserman, A.: Large-scale learning of sign language by watching tv (using co-occurrences). In: BMVC (2013) 7
11. Schembri, A., Fenlon, J., Rentelis, R., Cormier, K.: British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition) (2017), http://www.bslcorpusproject.org 6
12. Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., Cormier, K.: Building the British sign language corpus. Language Documentation & Conservation **7**, 136–154 (2013) 6
13. Zhang, M., Lucas, J., Ba, J., Hinton, G.E.: Lookahead optimizer: k steps forward, 1 step back. In: NeurIPS (2019) 7