CONTEXT RECOGNITION IN-THE-WILD: UNIFIED MODEL FOR MULTI-MODAL SENSORS AND MULTI-LABEL CLASSIFICATION

YONATAN VAIZMAN, NADIR WEIBEL, AND GERT LANCKRIET

SUPPLEMENTARY MATERIAL

S1 MISSING LABEL INFORMATION

We composed several heuristic rules to declare labels as missing. These rules may cause losing cases of labels that were actually correct. However, these rules leave us with cleaner labels that we can be more confident in.

(1) There are examples for which the participant did not use the label reporting interface at all. For such examples, we mark as "missing" all the labels, except labels that we adjusted based on location ("At home", "At the beach", and "At main workplace").
(2) We identify subsets of labels that represent mutually-exclusive alternatives that typically cover all the possible options for a certain aspect:
    - Body posture/movement: {"Lying down", "Sitting", "Standing", "Walking", "Running", "Bicycling"}
    - Phone position: {"Phone in pocket", "Phone in hand", "Phone in bag", "Phone on table"}
    - {"Indoors", "Outside"}

    For every example, we examine each of these label subsets. If none of the labels in the set was selected, we mark all of them as missing for this example.
    For instance: if an example is not annotated with any of the body posture/movement labels, it is most likely that actually one of this subset's labels is relevant, but the participant simply did not report it. We do not want to regard all the body posture/movement labels as negative since one of them is correct, so it is better (safer) to treat them all as missing for this example.
(3) For the phone position label subset, there were cases where a participant reported two of the labels (*e.g.* "Phone in hand" and "Phone in pocket"). Most likely such cases were mistakes of label-reporting. For these cases, we mark all the phone position labels as missing, since we do not know which of the reported labels is the correct one.
(4) For every participant, we identify the subset of labels that were applied. We then mark all the other labels as missing for all the participant's examples. The reason behind this is that every participant typically used a small subset of labels during the days of participation. For these labels, we can treat the participant as an authority for when they are relevant and when they are not; but for the labels that the participant never used, it is possible the participant was not aware of them in the menu or did not bother to regard to them, so we should not rely on them to be actual negative examples.

Table S1 shows the counts of examples per label in the dataset, before and after applying the missing label information (MLI). For most labels, the number of positive examples remained the same, and the MLI simply narrowed down the collection of examples to be considered as negative.
Table S2 shows the effect of regarding to missing label information (MLI) in both training and testing of the logistic recognition system. Introducing MLI to the performance metrics (counting only non-missing entries) shows very slight increase in sensitivity (probably related to cases of wrong phone-position labels that are now marked missing) and larger increase in specificity (related to the many cases that were previously treated as negative and now as missing). The effect of MLI on training is a combination of slight decrease in specificity (a small sacrifice caused by getting rid of good negative examples) and larger increase in sensitivity, contributing to an overall increase in balanced accuracy.

			without MLI		with MLI					without MLI		with MLI	
	Label	P_l	N_l^p	N_l^n	N_l^p	N_l^n		Label	P_l	N_l^p	N_l^n	N_l^p	N_l^n
1	Lying down	47	54359	122582	54359	119880	26	Cleaning	22	1839	175102	1839	90588
2	Sitting	50	82904	94037	82904	93215	27	Laundry	12	473	176468	473	54955
3	Walking	50	11892	165049	11892	164227	28	Washing dishes	17	851	176090	851	88053
4	Running	19	675	176266	675	93692	29	Watching TV	28	9412	167529	9412	100152
5	Bicycling	22	3523	173418	3523	79920	30	Surfing the internet	28	11641	165300	11641	98028
6	Sleeping	40	42920	134021	42920	124072	31	At a party	3	404	176537	404	25876
7	Lab work	8	2898	174043	2898	24384	32	At a bar	4	520	176421	520	19986
8	In class	13	2872	174069	2872	49400	33	At the beach	5	122	176819	122	20845
9	In a meeting	34	2904	174037	2904	124578	34	Singing	6	384	176557	384	15768
10	At main workplace	26	20382	156559	20382	80114	35	Talking	44	18976	157965	18976	139394
11	Indoors	51	107944	68997	107414	7099	36	Computer work	38	23692	153249	23692	125379
12	Outside	36	7629	169312	7099	80923	37	Eating	49	10169	166772	10169	158630
13	In a car	24	3635	173306	3635	104642	38	Toilet	33	1646	175295	1646	128368
14	On a bus	24	1185	175756	1185	98751	39	Grooming	25	1847	175094	1847	109353
15	Drive (I'm the driver)	24	5034	171907	5034	93827	40	Dressing	27	1308	175633	1308	117002
16	Drive (I'm a passenger)	19	1655	175286	1655	92384	41	At the gym	6	906	176035	906	32958
17	At home	50	83977	92964	83977	91065	42	Stairs - going up	17	399	176542	399	57797
18	At a restaurant	16	1320	175621	1320	87257	43	Stairs - going down	15	390	176551	390	59749
19	Phone in pocket	31	15301	161640	14658	67960	44	Elevator	8	124	176817	124	46631
20	Exercise	36	5384	171557	5384	143467	45	Standing	51	22766	154175	22766	153353
21	Cooking	33	2257	174684	2257	127535	46	At school	39	25840	151101	25840	120042
22	Shopping	18	896	176045	896	82705	47	Phone in hand	37	8595	168346	7535	79201
23	Strolling	8	434	176507	434	25234	48	Phone in bag	22	5589	171352	5201	55473
24	Drinking (alcohol)	10	864	176077	864	41955	49	Phone on table	43	70611	106330	69929	27237
25	Bathing - shower	27	1186	175755	1186	117321	50	With co-workers	17	4139	172802	4139	62410
							51	With friends	25	12865	164076	12865	81005

Table S1. Label counts in the dataset. Counts out of the 176941 core examples (those that have all the six core sensors available). P_l is the number of participants with positive examples of the label. Without MLI presents the counts of examples (positive N_l^p and negative N_l^n) before applying MLI. With MLI presents the counts of examples (positive N_l^p and negative N_l^n) that remain after removing missing labels.

	metrics without MLI				metrics with MLI			
	accuracy	**sensitivity**	**specificity**	**BA**	**accuracy**	**sensitivity**	**specificity**	**BA**
LR (trained without MLI)	0.846	0.533	0.851	0.692	0.846	0.534	0.863	0.698
LR (trained with MLI)	0.828	0.587	0.824	0.705	0.840	0.588	0.846	0.717

Table S2. Logistic regression performance. Training without and with missing labels information. Performance scores reported with old and new metrics (without and with missing labels information, respectively).

S2 RESULTS PER-LABEL

In order to provide a complete picture, and to allow readers to examine results for different labels, we add performance scores for each of the 51 labels in tables S3–S4. These tables include results with the LR baseline, and with MLP with zero–two hidden layers. The last column refers to MLP that was trained with sensor-dropout. These tables show a general trend of improvement for many labels when progressing from the baseline to an MLP with two hidden layers. The improvement is more significant for labels that started with relatively poor performance, like "Bathing — shower", "Cleaning", "At the beach", and "Elevator".

		LR	**linear**	**(16)**	**(16-16)**	**(16-16)DO**
1	Lying down	0.870	0.871	0.874	0.874	0.876
2	Sitting	0.757	0.764	0.767	0.765	0.770
3	Walking	0.797	0.801	0.810	0.808	0.808
4	Running	0.658	0.753	0.814	0.814	0.819
5	Bicycling	0.867	0.851	0.872	0.877	0.868
6	Sleeping	0.891	0.892	0.895	0.896	0.897
7	Lab work	0.828	0.798	0.845	0.843	0.842
8	In class	0.767	0.793	0.770	0.766	0.795
9	In a meeting	0.797	0.810	0.814	0.814	0.781
10	At main workplace	0.822	0.835	0.842	0.852	0.847
11	Indoors	0.867	0.879	0.888	0.884	0.891
12	Outside	0.856	0.869	0.876	0.881	0.885
13	In a car	0.864	0.867	0.869	0.859	0.864
14	On a bus	0.809	0.835	0.866	0.865	0.858
15	Drive (I'm the driver)	0.858	0.871	0.865	0.866	0.857
16	Drive (I'm a passenger)	0.834	0.819	0.853	0.868	0.860
17	At home	0.752	0.769	0.778	0.792	0.794
18	At a restaurant	0.770	0.839	0.820	0.833	0.846
19	Phone in pocket	0.778	0.789	0.795	0.798	0.802
20	Exercise	0.821	0.813	0.812	0.829	0.821
21	Cooking	0.712	0.722	0.728	0.737	0.747
22	Shopping	0.723	0.774	0.783	0.773	0.792
23	Strolling	0.649	0.687	0.745	0.764	0.759
24	Drinking (alcohol)	0.681	0.779	0.786	0.793	0.803
25	Bathing - shower	0.632	0.706	0.731	0.734	0.746
	Average (labels 1–25)	0.786	0.807	0.820	0.823	0.825

Table S3. Balanced accuracy per label (part 1). LR is the baseline system with separate logistic regression trained per label. The other columns refer to MLP with either 0 hidden layers (linear), or with the hidden layer dimensions specified in parenthesis. The last column is for MLP trained with dropout ($p_{drop} = 0.2$).

Received February 2007; revised March 2009; accepted June 2009

		LR	linear	(16)	(16-16)	(16-16)DO
26	Cleaning	0.624	0.693	0.721	0.731	0.740
27	Laundry	0.648	0.758	0.682	0.662	0.674
28	Washing dishes	0.606	0.704	0.729	0.761	0.793
29	Watching TV	0.639	0.690	0.713	0.711	0.734
30	Surfing the internet	0.611	0.588	0.599	0.589	0.614
31	At a party	0.765	0.640	0.773	0.738	0.794
32	At a bar	0.783	0.671	0.791	0.845	0.863
33	At the beach	0.498	0.717	0.822	0.820	0.846
34	Singing	0.524	0.514	0.501	0.529	0.663
35	Talking	0.664	0.677	0.677	0.685	0.679
36	Computer work	0.705	0.724	0.732	0.730	0.727
37	Eating	0.657	0.666	0.672	0.677	0.669
38	Toilet	0.635	0.647	0.683	0.717	0.695
39	Grooming	0.632	0.667	0.698	0.702	0.735
40	Dressing	0.660	0.683	0.710	0.737	0.749
41	At the gym	0.651	0.683	0.712	0.800	0.779
42	Stairs - going up	0.595	0.708	0.757	0.755	0.731
43	Stairs - going down	0.609	0.707	0.751	0.753	0.728
44	Elevator	0.500	0.783	0.813	0.845	0.845
45	Standing	0.679	0.678	0.677	0.668	0.667
46	At school	0.739	0.748	0.751	0.754	0.751
47	Phone in hand	0.685	0.699	0.692	0.695	0.694
48	Phone in bag	0.753	0.752	0.746	0.764	0.744
49	Phone on table	0.789	0.804	0.797	0.802	0.801
50	With co-workers	0.657	0.720	0.752	0.755	0.778
51	With friends	0.608	0.613	0.617	0.636	0.635
	Average (labels 26–51)	0.651	0.690	0.714	0.725	0.736

Table S4. Balanced accuracy per label (part 2). LR is the baseline system with separate logistic regression trained per label. The other columns refer to MLP with either 0 hidden layers (linear), or with the hidden layer dimensions specified in parenthesis. The last column is for MLP trained with dropout ($p_{drop} = 0.2$).