ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition

Jun Wan and Stan Z. Li
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, China
jun.wan@ia.ac.cn, szli@nlpr.ia.ac.cn

Yibing Zhao and Shuai Zhou
Macau University of Science and Technology, Macau
xlyxl2008@163.com, shuaizhou.palm@gmail.com

Isabelle Guyon
UPSud and INRIA, Université Paris-Saclay
and ChaLearn
guyon@chalearn.org

Sergio Escalera
University of Barcelona
Computer Vision Center, ChaLearn
sergio@maia.ub.es

Abstract

In this paper, we present two large video multi-modal datasets for RGB and RGB-D gesture recognition: the ChaLearn LAP RGB-D Isolated Gesture Dataset (IsoGD) and the Continuous Gesture Dataset (ConGD). Both datasets are derived from the ChaLearn Gesture Dataset (CGD) that has a total of more than 50000 gestures for the "one-shot-learning" competition. To increase the potential of the old dataset, we designed new well curated datasets composed of 249 gesture labels, and including 47933 gestures manually labeled the begin and end frames in sequences. Using these datasets we will open two competitions on the CodaLab platform so that researchers can test and compare their methods for "user independent" gesture recognition. The first challenge is designed for gesture spotting and recognition in continuous sequences of gestures while the second one is designed for gesture classification from segmented data. The baseline method based on the bag of visual words model is also presented.

1. Introduction

The analysis of large amounts of data is part of most computer vision problems, such as image classification and location [11, 16], semantic segmentation [15], and face recognition [21, 20, 22]. As a recent example, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [18] is held every year since 2010. The ILSVRC contains many challenge tasks, including image classification (2010-2014), single-object localization (2011-2014), and object detection (2013-2014). The dataset contains 1000 object classes and approximately 1.2 million training images, 50 thousand validation images and 100 thousand test images. ILSVRC greatly promotes the the development of new techniques, particularly those based on deep learning architectures, for image classification and object localization.

The field of Looking at People recently has received special attention, and several datasets have been also presented in order to deal with different computer vision image analysis tasks, such as human pose recovery, action and gesture recognition, and face analysis, just to mention a few [7, 6, 3, 5, 1, 4, 19, 8]. However, for the RGB-D video-based gesture recognition problem, there are very few annotated datasets including a large number of samples and gesture categories. Table 1 shows the public RGB-D gesture datasets released from 2011 to 2015 in the literature. We can see that although the CGD dataset [10] has more than 50 thousands gestures, there is only one training sample per class in every batch where only 8 to 12 categories are present. The ChaLearn Multi-modal Gesture Dataset [6, 3] has more than 13 thousands gestures and 387 training samples per class, but it only has 20 classes. Besides, the other listed datasets [17, 14] only include 10 gesture classes.

In order to provide to the community with a large dataset for RGB-Depth gesture recognition, here we take benefit of the previous CGD dataset [10] by integrating all batch classes and samples to design two new large RGB-D gesture recognition datasets for gesture spotting and classification. In Table 1, the new datasets show a significant increase in size in terms of both number of categories and number of samples in comparison to state of the art alternatives. Note that although our datasets were designed for RGB-D gesture recognition, they also can be used for traditional gesture recognition by just considering its associated RGB data.

Next, we describe the design, characteristics, and associated challenges for the new datasets.

2. ChaLearn LAP IsoGD and ConGD Datasets

The CGD dataset [10] was designed for the "one-shot learning" task. By that we mean that only one training example of each gesture was available in each batch of data, the rest being used for testing. Each batch in the CGD dataset includes 100 gestures from a small vocabulary of 8 to 12 gestures. In the CGD dataset, a "lexicon" is defined as a small "vocabulary" of gestures. These are drawn from a variety of domains, including sign language for the deaf, underwater sign language, helicopter and traffic signal, pantomimes and symbolic gestures, Italian gestures, and body language. The large number of gesture labels (289 gestures from 30 lexicons), and the large number of gestures performed (54,000 gestures in about 23,000 RGB-D videos) make it good material to carve out different tasks. This is what we did by creating two large RGB-D gesture datasets: The ChaLearn LAP IsoGD dataset [1] and the ChaLearn LAP ConGD dataset [2].

Each video sequence in the original data includes the performance of one to five gestures. In order to create the datasets, first, we semi-manually segmented the whole data and labeled the temporal segmentation information (the begin and end frames of each gesture) for all the videos in the CGD dataset.

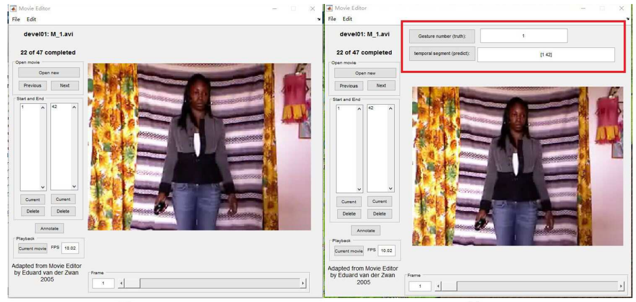
Then, we manually labeled 540 batches corresponding to 30 lexicons of gestures (480 development batches, 40 final batches, and 20 validation batches, in the original data). The total number of gesture classes is 289. However, because some gesture movements are similar in different lexicons, we finally obtained 249 gesture labels after fusing the classes having similar gestures, and deleting some batches.

Finally, we created the isolated gesture and continuous datasets. Also, we provided test protocols so participants can compare themselves on the same basis.

2.1. Semi-manual Temporal Segmentation

We used the original dataset design toolbox (see Fig. 1(a)) for labeling the begin and end frames of each gesture in a video including continuous gestures.

[1]http://www.cbsr.ia.ac.cn/users/jwan/database/isogd.html
[2]http://www.cbsr.ia.ac.cn/users/jwan/database/congd.html



(a) (b)

Figure 1. (a) Original toolbox; (b) the modified toolbox added the truth label and predicted temporal segmentation for each video (see the red box).

For the CGD dataset, there are 540 batches totalling about 25380 RGB-D videos. Each batch includes about 22 isolated videos (with only one gesture) and 25 videos must be segmented and labeled. We could easily obtain the temporal segmentation (the begin and end frames) of these isolated videos, but needed to resort to a semi-automatic tool to segment the remaining videos (about 13500 (25×540) videos). In order to accelerate the labelling process, we added some information in the toolbox as shown in Fig. 1(b). In the red box of Fig. 1(b), the first line is the truth label of the opened video, and the second line is the predicted temporal segmentation of the opened video. The predicted temporal segmentations are obtained by the dynamic time warping (DTW) algorithm [24].

For the case of isolated videos, we created the annotation file in advance and saved the temporal segmentations. For the continuous videos, we first checked the predicted temporal segmentation (see the second line of Fig. 1(b)). This allowed us to obtain the following 4 important pieces of information: the predicted temporal segmentation, the predicted gesture number, the truth labels, and the truth gesture number, which were used to guide labeling temporal information in continuous videos.

2.2. Gesture Labels

We found 30 lexicons for all 540 batches (devel01-devel480, valid01-valid20,final01-final40). However, we could not just use all gestures. Some issues had to be addressed:

(1) There were similar gesture movements in different lexicons. For example, the gesture "V" occurs in "CommonEmblems" (the "V" of Victory), "Mudra1" (Kartarimukha), "ChineseNumbers" (the number two), and so on. We manually checked all the gestures in all lexicons and found the similar gestures. The similar gesture movements were integrated as the unique label in our datasets.

(2) Some videos have only one or two frames (i.e. devel02, M_45.avi; Devel256, M_19.avi) and there are no gestures in other some videos (i.e. devel238, M_1.avi; Dev-

Dataset	Total gestures	Gesture labels	Avg. samples per class	Train samples (per class)	Data provided & Learning Task
CGD, 2011 [10]	540,000	>200	10	8∼12 (1-1-1)	RGB-D & one-shot learning
Multi-modal Gesture Dataset, 2013, 2014 [6, 3]	13,858	20	692	7,754	RGB-D, audio, skeleton & small-scale learning
ChAirGest 2013 [17]	1,200	10	120	-	RGB-D & small-scale learning
Sheffield Kinect Gesture Dataset, 2013 [14]	1,080	10	108	- -	RGB-D & small-scale learning
ChaLearn LAP IsoGD (Ours)	47,933	249	192	35,878 (144-64-851)	RGB-D & large-scale learning
ChaLearn LAP ConGD (Ours)	47,933	249	192	30,442 (122-54-722)	RGB-D & large-scale learning

Table 1. Comparison of the public RGB-D gesture datasets. The numbers in parentheses correspond to (average per class-minimum per class-maximum per class). It shows that the largest gestures and training samples and each class has at least 54 RGB-D videos in our datasets.

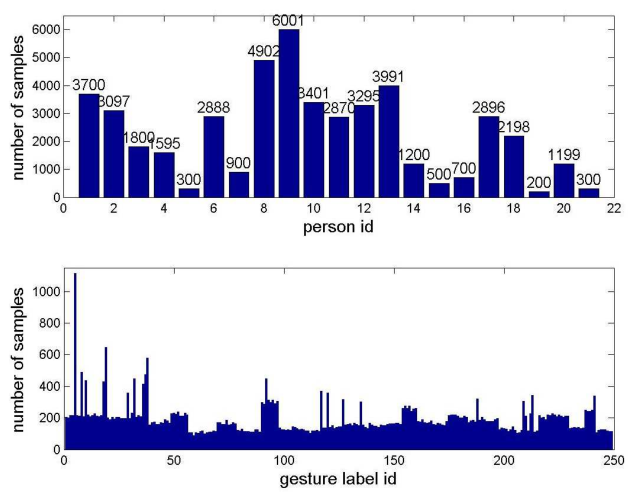


Figure 2. Top and bottom figures show the distribution of subject id and gesture label id, respectively.

el447, M_43.avi). Those videos were deleted in our datasets.

Finally, we obtained 249 unique gesture labels, 47933 gestures in 22535 RGB-D videos, which are derived from 480 batches of the original CGD dataset. The final remaining batches are shown in Table 7 of Appendix A.

2.3. Statistical Information

Here, we give some statistical information about the new datasets. Figure 2 shows the distribution of subject id and gesture label id. We can see that each subject at least has 200 gesture samples, which means some of the subjects did not perform all the types of gestures (for all gesture classes). For the distribution of gesture labels, there are at least 89 gesture samples and one gesture label has more than 1000 samples.

ChaLearn LAP IsoGD. Using the begin and end frames of each video obtained as described in Section 2.1, we split all the videos of the CGD dataset into isolated gestures. Finally, we obtained 47,933 gestures. Each RGB-D video represents one gesture instance, having 249 gestures labels performed by 21 different individuals. Details of the dataset are shown in Table 2.

ChaLearn LAP ConGD. This dataset is organized as the CGD dataset. It includes 47933 RGB-D gestures in 22535 RGB-D gesture videos. Each RGB-D video may represent one or more gestures, and there are also 249 gestures labels performed by 21 different individuals. The detailed information of this dataset are shown in Table 3.

3. Challenge Tasks

The challenge tasks proposed are both "user independent" and consist of:

- Isolated gesture recognition for the ChaLearn LAP IsoGD dataset.
- Gesture spotting and recognition from continuous videos for the ChaLearn LAP ConGD dataset.

As shown in Table 2 and 3, the datasets are split into three subsets: training, validation, and test. The training set includes all gestures from 17 subjects, the validation set includes all gestures from 2 subjects, and the rest gestures from 2 subjects are used in the test set. We guarantee that the validation and test sets include gesture samples from the 249 labels.

4. Evaluation protocol

For both datasets, we provide training, validation, and test sets. In order to make it more challenging, all three sets include data from different subjects, which means the

Sets	# of labels	# of gestures	# of RGB videos	# of depth videos	# of subjects	label provided
Training	249	35878	35878	35878	17	Yes
Validation	249	5784	5784	5784	2	No
Testing	249	6271	6271	6271	2	No
All	249	47933	47933	47933	21	-

Table 2. Information of the ChaLearn LAP IsoGD dataset. The database has been divided into three sub-datasets including different subjects (user independent task).

Sets	# of labels	# of gestures	# of RGB videos	# of depth videos	# of subjects	label provided	temporal segment provided
Training	249	30442	14134	14134	17	Yes	Yes
Validation	249	8889	4179	4179	2	No	No
Testing	249	8602	4042	4042	2	No	No
All	249	47933	22535	22535	21	-	-

Table 3. Information of the ChaLearn LAP ConGD dataset. The database has been divided into three sub-datasets including different subjects (user independent task).

gestures of one subject in validation and test sets will not appear in the training set. In the development stage, the labels of the training set are provided.

For the isolated gesture recognition challenge, we use the recognition rate r as the evaluation criteria. The recognition rate is calculated as:

$$r = \frac{1}{n}\sum_{i=1}^{n}\delta(p_l(i), t_l(i)) \tag{1}$$

where n is the number of samples; p_l is the predicted label; t_l is the ground truth; $\delta(j_1, j_2) = 1$, if $j_1 = j_2$, otherwise $\delta(j_1, j_2) = 0$.

For continuous gesture recognition, we use the Jaccard index (the higher the better), similarly to the ChaLearn Looking at People 2015 challenges [1]. The Jaccard index measures the average relative overlap between true and predicted sequences of frames for a given gesture. For a sequence s, let $G_{s,i}$ and $P_{s,i}$ be binary indicator vectors for which 1-values correspond to frames in which the i^{th} gesture label is being performed. The Jaccard Index for the i^{th} class is defined for the sequence s as:

$$J_{s,i} = \frac{G_{s,i} \cap P_{s,i}}{G_{s,i} \cup P_{s,i}} \tag{2}$$

where $G_{s,i}$ is the ground truth of the i^{th} gesture label at sequence s, and $P_{s,i}$ is the prediction for the i^{th} label at sequence s.

When $G_{s,i}$ and $P_{s,i}$ are empty, we define $J_{(s,i)} = 0$. Then, for the sequence s with l_s true labels, we can compute Jaccard Index J_s as:

$$J_s = \frac{1}{l_s}\sum_{i=1}^{L} J_{s,i} \tag{3}$$

where L is the number of gesture labels. We note that Eq. 3 is different from the definition of reference [1] $J_s = \sum_{i=1}^{L} J_{s,i} / \sum_{i=1}^{L}(1 - \delta(J_{s,i}, 0))$. We made this change because of a drawback of the original definition of reference [1] that we now explain. Suppose, for instance, that the ground truth of the sequence s with 100 frames consists of three gestures of labels [1, 2, 3] and with begin and end frames [1 40; 41 70; 71 100]. Assume that one predictor obtains as result a single gesture labels [1], with begin and end frames [1 40]. Then, $J_{s11} = 1$ by [1], but $J_{s12} = 0.33$ by Eq. 3. Assume that another predictor gets two gestures with labels [1 3] and with begin and end frames [1 40; 41 100]. Then $J_{s21} = \frac{1}{2}(\frac{40}{40} + \frac{30}{60}) = 0.75$ by [1], $J_{s22} = \frac{1}{3}(\frac{40}{40} + \frac{30}{60}) = 0.5$ by Eq. 3. The results $J_{11} > J_{21}$ by [1] indicate that the first predicted result is better than the second one. However, we obviously know that the second predicted result is more reasonable, and our result $J_{s12} < J_{s22}$ meet the requirements.

For all testing sequences $S = \{s_1, ..., s_n\}$ with n samples, the mean Jaccard Index $\overline{J_S}$ is calculated as:

$$\overline{J_S} = \frac{1}{n}\sum_{j=1}^{n} J_{s_j} \tag{4}$$

We use the recognition rate r and mean Jaccard Index $\overline{J_S}$ as the evaluation criteria for the ChaLearn LAP IsoGD and ConGD datasets, respectively.

5. Baseline Methods

We used the bag of visual words (BoVW) model in our datasets in order to compute a baseline method result. We first extracted the mixed features around sparse keypoints

Name	$translated$	$scaled$
Alfnie1	0.2255	0.2573
Alfnie2	0.2310	0.2566
BalazsGodeny	0.5636	0.5526
HITCS	0.6640	0.6066
Immortals	0.3962	0.4152
Joewan	0.2612	0.2913
Manavender	0.4252	0.4358
OneMillionMonkeys	0.4961	0.5552
Pennect	0.4888	0.4068
SkyNet	0.4693	0.4771
TurtleTamers	0.5993	0.5296
Vigilant	0.5173	0.5067
WayneZhang	0.6278	0.5834
XiaoZhuWudi	0.6986	0.6897
Zonga	0.4905	0.5776
MFSK+BoVW	**0.2120**	**0.2375**

Table 4. The results of all the top 14 results [9, 23] on the challenging subsets of CGD, such as translated and scaled data. The MFSK feature can obtain the best performances (the value shown in this table is levenshtein distance (LD) scores, the lower the better).

(MFSK[3]) [23] from RGB-D data in isolated and continuous datasets. The MFSK features were designed for local feature extraction from RGB-D videos, which have proved effective for gesture recognition. For examples, as shown in Table 4, the BoVW model with MFSK features achieved the best performances on the challenging data of CGD, such as translated and scaled subsets [23]. In addition, in order to use facial features, we first applied a Normalized Pixel Difference (NPD) detector [13] for fast face detection. We then extracted Deep hidden IDentity (Deep ID) features [21], which use a convolution neural network (C-NN). In our experiments, the Deep ID model is trained on the CASIA-WebFace dataset [25]. The Deep ID features with size 160 are extracted from RGB images only. Subsequently, we randomly selected 200,000 features to compute a BoVW codebook using the Kmeans algorithm, limiting the codebook size to 5000. Finally, we trained the classifier using Support Vector Machine (SVM) with a linear kernel [2].

5.1. Results on the ChaLearn LAP IsoGD Dataset

As shown in Table 5, the performance of MFSK features are higher than MFSK+Deep ID features. That is because the motion features (i.e. MFSK) is more effective than the static feature (i.e. deep ID) for video-based gesture recognition. The best recognition rates are 18.65% and 24.19% for validation and test sets, respectively. Table 5 shows our initial results without any optimization strategy (such as, choose different codebook size, sparse coding instead of VQ, and so on.).

[3]http://mloss.org/software/view/499/

Feature type	Set	Recognition rate r
MFSK	Validation	18.65%
MFSK+Deep ID	Validation	18.23%
MFSK	Testing	24.19%
MFSK+Deep ID	Testing	23.67%

Table 5. Experimental results on the ChaLearn LAP IsoGD dataset. All the results are obtained with linear kernel of SVM and the codebook size 5000.

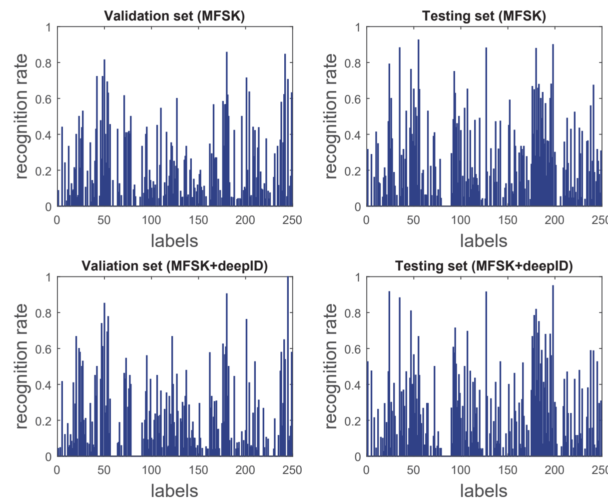


Figure 3. Recognition rate of each gesture label on the Chalearn LAP IsoGD dataset.

Furthermore, we analyze the recognition rate per gesture label (see Fig. 3) of the baseline methods on validation and test sets. As shown in Fig. 3, some gestures are failed to be recognized by our baseline methods. For example, on the validation set of the Chalearn LAP IsoGD dataset, the BoVW model with the MFSK features failed to recognize about 70 gesture labels (e.g. gesture label id: 2, 3, 4, 6, 7, 9, 14, 28). Hence, there is a margin for improvement, perhaps by incorporating more dynamic features.

Finally, the confusion matrix of the baseline method (BoVW+MFSK) for all 249 gesture labels is shown in Fig. 4. The overall recognition rate is 24.19%. We can see that some gesture labels are very difficult to recognize. For example, the gestures of label 11 (Gesture: Mudra2/Anjali) are confused with the gestures of label 26 (Gesture: ItalianGestures/Madonna). That is because some part of movements are very similar in these two kind of gestures (see Fig.5, in label 11: joint both hands-static gesture; label 26: joint both hands, figures touching, hands pointing away from you-dynamic gesture). The gesture label with its gesture name can be found in Appendix B.

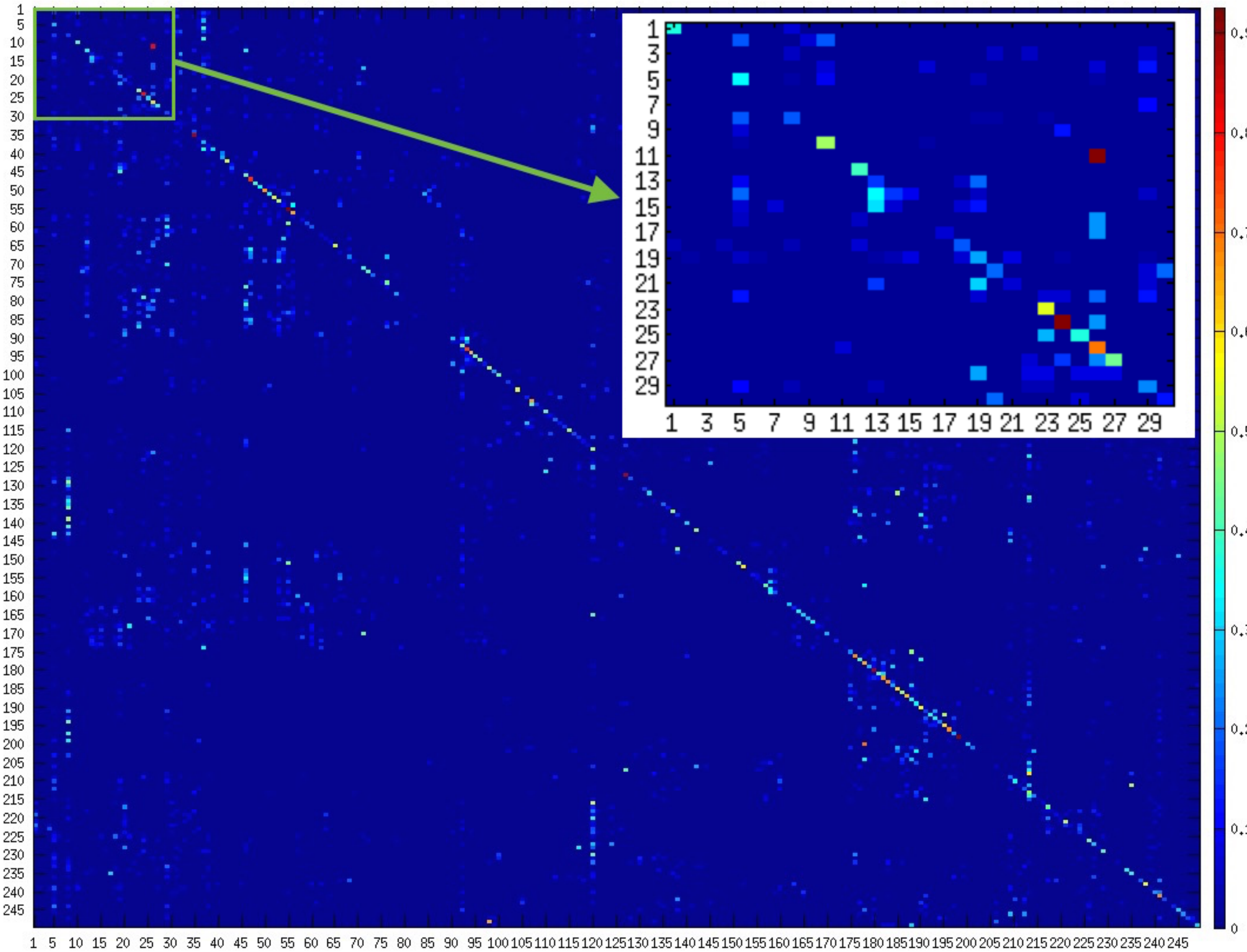


Figure 4. Confusion matrix of the baseline method (BoVW+MFSK) on the test set of the Chalearn LAP IsoGD dataset. The overall recognition rate is 24.19%.

5.2. Results on the ChaLearn LAP ConGD Dataset

For the sequence with multi-gesture, we first obtain the begin and end frames of each gesture based on motion by the work [12]. The method first measures the quantity of movement for each frame in a multi-gesture sequence and then threshold the quantity of movement to get candidate boundaries. Then, a sliding window is adopted to refine the candidate boundaries to produce the final boundaries of the segmented gesture sequences in a multi-gesture sequence. After temporal segmentation, we apply the same strategy as in the previous experiments on the ChaLearn LAP IsoGD dataset. The results are shown in Table 6, where it shows that the MFSK feature outperforms the MFSK with Deep ID features.

In the validation set, we correctly compute 2977 out of 4179 videos by the temporal segmentation method [12]. And for the test set, there are 2546 out of 4042 videos correctly computed.

6. Conclusion

In this paper, we introduced two large scale RGB-D gesture datasets. The main challenges of the released datasets are "user independent" and "large-scale" learning video-based gesture recognition. Besides, we provided the baseline methods for the two tasks. We will deploy the two challenges on the Codalab platform and set it up as an indefinitely running benchmark to allow researchers to submit their models and compare their performance with state of the art methods on the proposed RGB-D gesture recognition datasets.

A. Appendix 1

Our gestures labels are derived from the CGD dataset. After removing some batches, we finally selected 480. The considered batches are shown in Table 7.

Feature type	Set	Codebook Size	SVM kernel	Mean Jaccard Index $\overline{J_S}$
MFSK	Validation	5,000	linear	0.0918
MFSK+Deep ID	Validation	5,000	linear	0.0902
MFSK	Testing	5,000	linear	0.1464
MFSK+Deep ID	Testing	5,000	linear	0.1435

Table 6. Experimental results on the ChaLearn LAP ConGD dataset (for the value of $\overline{J_S}$, the higher the better.).

devel01	devel02	devel03	devel04	devel05	devel06	devel07	devel08	devel09	devel10	devel100	devel101	devel102	devel103	devel104	devel105	devel106	devel107	devel108	devel109	devel11	devel110	devel111	devel112
devel113	devel114	devel115	devel116	devel117	devel118	devel119	devel12	devel120	devel121	devel122	devel123	devel124	devel125	devel126	devel127	devel128	devel129	devel13	devel130	devel131	devel134	devel136	devel137
devel138	devel139	devel14	devel140	devel141	devel142	devel143	devel145	devel146	devel147	devel148	devel149	devel15	devel150	devel151	devel152	devel153	devel154	devel155	devel156	devel157	devel158	devel159	devel16
devel160	devel161	devel162	devel163	devel164	devel165	devel166	devel167	devel168	devel169	devel17	devel170	devel171	devel172	devel173	devel174	devel175	devel176	devel177	devel178	devel179	devel18	devel180	devel181
devel182	devel183	devel184	devel185	devel186	devel188	devel189	devel19	devel190	devel191	devel192	devel193	devel194	devel195	devel196	devel197	devel198	devel199	devel20	devel200	devel201	devel202	devel203	devel204
devel205	devel207	devel208	devel209	devel21	devel210	devel211	devel212	devel213	devel214	devel215	devel216	devel217	devel218	devel219	devel22	devel220	devel221	devel222	devel223	devel224	devel225	devel226	devel227
devel228	devel229	devel23	devel230	devel231	devel232	devel233	devel234	devel235	devel237	devel238	devel239	devel24	devel240	devel241	devel242	devel243	devel244	devel245	devel246	devel247	devel248	devel249	devel25
devel250	devel251	devel252	devel253	devel254	devel255	devel256	devel257	devel258	devel259	devel26	devel260	devel261	devel262	devel263	devel264	devel265	devel266	devel267	devel268	devel269	devel27	devel270	devel271
devel272	devel273	devel274	devel275	devel276	devel277	devel278	devel279	devel28	devel280	devel281	devel283	devel284	devel285	devel286	devel287	devel288	devel289	devel29	devel290	devel291	devel292	devel293	devel294
devel295	devel296	devel297	devel298	devel299	devel30	devel300	devel301	devel302	devel303	devel304	devel306	devel307	devel308	devel309	devel31	devel310	devel311	devel312	devel313	devel314	devel315	devel316	devel317
devel318	devel319	devel32	devel320	devel321	devel322	devel323	devel324	devel325	devel326	devel327	devel328	devel329	devel330	devel331	devel332	devel333	devel334	devel335	devel336	devel337	devel338	devel339	devel34
devel340	devel341	devel342	devel343	devel344	devel345	devel346	devel347	devel348	devel349	devel35	devel350	devel351	devel352	devel353	devel354	devel355	devel356	devel357	devel358	devel359	devel36	devel361	devel362
devel363	devel364	devel365	devel366	devel367	devel368	devel369	devel37	devel370	devel371	devel372	devel373	devel374	devel375	devel376	devel377	devel378	devel379	devel38	devel380	devel381	devel382	devel383	devel384
devel385	devel386	devel387	devel388	devel389	devel39	devel390	devel391	devel392	devel393	devel394	devel395	devel396	devel397	devel398	devel399	devel40	devel400	devel401	devel402	devel403	devel405	devel406	devel407
devel408	devel409	devel41	devel410	devel411	devel412	devel413	devel414	devel415	devel416	devel417	devel418	devel419	devel42	devel420	devel421	devel422	devel423	devel424	devel425	devel426	devel427	devel428	devel429
devel43	devel430	devel431	devel432	devel433	devel434	devel435	devel436	devel437	devel438	devel439	devel44	devel440	devel441	devel442	devel443	devel444	devel445	devel446	devel447	devel448	devel449	devel45	devel450
devel451	devel452	devel453	devel454	devel455	devel456	devel457	devel459	devel46	devel460	devel461	devel462	devel463	devel464	devel465	devel466	devel467	devel468	devel469	devel47	devel470	devel471	devel472	devel473
devel474	devel475	devel476	devel477	devel478	devel479	devel48	devel480	devel49	devel50	devel51	devel52	devel53	devel54	devel55	devel56	devel58	devel59	devel60	devel61	devel62	devel63	devel64	devel65
devel66	devel67	devel68	devel69	devel70	devel71	devel72	devel73	devel74	devel75	devel76	devel77	devel78	devel79	devel80	devel81	devel82	devel83	devel84	devel85	devel86	devel87	devel88	devel89
devel90	devel91	devel92	devel93	devel94	devel95	devel96	devel97	devel98	devel99	valid02	valid03	valid04	valid05	valid07	valid09	valid12	valid13	valid14	valid15	valid16	valid18	valid19	valid20

Table 7. The finally batches/folds of the CGD database are used in the Chalean LAP IsoGD and ConGD datasets.

Figure 5. Examples of failure. (a) Gesture sample (Predicted label: 26, True label: 11). It is shown in the zoom of interest regions in Figure 4; (b) Gesture sample (Ground truth label 26); (c) Gesture sample (Predicted label: 181, True label: 114); (d) Gesture sample (Ground truth label 181);

B. Appendix 2

In Fig. 6, all the gesture labels are shown with their gesture names. All the gesture names come from the CGD dataset.

Acknowledgment

We thank Microsoft Research for its sponsoring. This work was supported by the Chinese National Natural Science Foundation Projects ♯61502491, ♯61572501, ♯61203267, ♯61375037, ♯61473291, National Science and Technology Support Program Project ♯2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, AuthenMetric R&D Funds, and partially supported by Spanish project TIN2013-43478-P.

References

[1] X. Baró, J. Gonzalez, J. Fabian, M. A. Bautista, M. Oliu, H. J. Escalante, I. Guyon, and S. Escalera. Chalearn looking at people 2015 challenges: Action spotting and cultural event recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*, pages 1–9, 2015. 1, 4

[2] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152, 1992. 5

[3] S. Escalera, X. Baro, J. Gonzàlez, M. Bautista, M. Madadi, M. Reyes, V. Ponce, H. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. *ChaLearn LAP Workshop, ECCV*, 2014. 1, 3

[4] S. Escalera, J. Fabian, P. Pardo, , X. Baró, J. Gonzàlez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *International Conference in Computer Vision, Looking at People, ICCVW*, 2015. 1

[5] S. Escalera, J. Gonzàlez, X. Baro, P. Pardo, J. Fabian, M. Oliu, H. J. Escalante, I. Huerta, and I. Guyon. Chalearn looking at people 2015 new competitions: Age estimation and cultural event recognition. In *IJCNN*, 2015. 1

[6] S. Escalera, J. Gonzàlez, X. Baro, M. Reyes, I. Guyon, V. Athitsos, H. Escalante, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclarof. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. *ICMI*, pages 365–368, 2013. 1, 3

[7] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopés, I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In

Label	Gesture	Label	Gesture	Label	Gesture	Label	Gesture
1	Mudra1/Ardhachandra	2	Mudra1/Ardhapataka	3	Mudra1/Chandrakala	4	Mudra1/Chatura
5	Mudra1/Kartarimukha ChineseNumbers/er GangHandSignals1/Victory RefereeVolleyballSignals1/DoubleHit RefereeWrestlingSignals1/AwardingPoints TaxiSouthAfrica/TaxiHandSigns8	6	Mudra1/Pataka	7	Mudra1/Sarpashirsha	8	Mudra1/Shikhara CanadaAviationGroundCirculation1/ToutVaBienContinuez DivingSignals3/Ascend
		9	Mudra1/Tripataka	10	Mudra1/Trishula ChineseNumbers/san		
		11	Mudra2/Anjali				
		12	Mudra2/Chakram	13	Mudra2/Hamsapaksha	14	Mudra2/Mayura
		15	Mudra2/Mrigashirsha	16	Mudra2/Sandamsha	17	Mudra2/Swastikam
		18	Mudra2/Tamrachuda ChineseNumbers/jiu	19	Mudra2/Vitarka DivingSignals4/OK GangHandSignals2/OK	20	ItalianGestures/AndateVia
22	ItalianGestures/CheFurbo					21	ItalianGestures/Bellissima
23	ItalianGestures/ChePalle	24	ItalianGestures/CheVuoi			25	ItalianGestures/DAccordo
26	ItalianGestures/Madonna	27	ItalianGestures/NonMiFrega	28	ItalianGestures/Perfetto	29	ItalianGestures/SeiPazzo DivingSignals1/Think
30	ItalianGestures/VieniQua	31	ChineseNumbers/ba	32	ChineseNumbers/ling TaxiSouthAfrica/TaxiHandSigns5		
34	ChineseNumbers/qi	35	ChineseNumbers/shi			33	ChineseNumbers/liu
36	ChineseNumbers/si, RefereeVolleyballSignals1/FourHits	37	ChineseNumbers/wu TaxiSouthAfrica/TaxiHandSigns7	38	ChineseNumbers/yi CraneHandSignals/CableUp TaxiSouthAfrica/TaxiHandSigns4	39	GestunoColors/654_colour_couleur
						40	GestunoColors/655_black_noir
41	GestunoColors/656_white_blanc	42	GestunoColors/657_red_rouge			43	GestunoColors/658_yellow_jaune
44	GestunoColors/659_orange_orange	45	GestunoColors/660_bleu_blue	46	GestunoColors/661_green_vert	47	GestunoColors/662_purple_violet
48	GestunoColors/663_brown_brun	49	GestunoDisaster/102_thunderstorm_orage	50	GestunoDisaster/108_tide_maree	51	GestunoDisaster/109_dought_secheresse
52	GestunoDisaster/110_earthquake_tremblementdeterre	53	GestunoDisaster/111_famine_famine	54	GestunoDisaster/112_flood_inondation	55	GestunoDisaster/113_tornado_tornade
56	GestunoDisaster/114_hurricane_ouragan	57	GestunoLandscape/63_moon_lune	58	GestunoLandscape/64_sky_ciel	59	GestunoLandscape/66_star_etoile
60	GestunoLandscape/67_sun_soleil	61	GestunoLandscape/81_hill_colline	62	GestunoLandscape/82_mountain_montagne	63	GestunoLandscape/83_valley_vallee
64	GestunoLandscape/84_summit_sommet	65	GestunoLandscape/85_volcano_volcan	66	GestunoLandscape/88_desert_desert	67	GestunoLandscape/89_lake_lac
68	GestunoLandscape/90_river_fleuve	69	GestunoLandscape/91_sea_mer	70	GestunoSmallAnimals/125_bird_oiseau	71	GestunoSmallAnimals/127_butterfly_papillon
72	GestunoSmallAnimals/129_cat_chat	73	GestunoSmallAnimals/131_crab_crabe	74	GestunoSmallAnimals/132_dog_chien	75	GestunoSmallAnimals/134_fish_poisson
76	GestunoSmallAnimals/141_mouse_souris	77	GestunoSmallAnimals/143_pigeon_pigeon	78	GestunoSmallAnimals/150_worm_ver	79	GestunoTopography/65_space_espace
80	GestunoTopography/77_city_ville	81	GestunoTopography/78_suburbs_banlieu	82	GestunoTopography/79_village_village	83	GestunoTopography/80_countryside_campagne
84	GestunoTopography/86_soil_sol	85	GestunoTopography/87_ground_terre	86	GestunoTopography/92_harbour_port	87	GestunoTopography/93_peninsula_penisule
88	GestunoTopography/94_island_ile	89	GestunoTopography/95_region_region	90	MusicNotes/do	91	MusicNotes/do2
92	MusicNotes/fa DivingSignals3/Descend	93	MusicNotes/la	94	MusicNotes/mi	95	MusicNotes/re
		96	MusicNotes/sol	97	MusicNotes/ti	98	CanadaAviationGroundCirculation1/Coupez
99	CanadaAviationGroundCirculation1/DemarrezMoteurs	100	CanadaAviationGroundCirculation1/DirigezVousVers	101	CanadaAviationGroundCirculation1/FaceMe	102	CanadaAviationGroundCirculation1/Freins
103	CanadaAviationGroundCirculation1/Incendie	104	CanadaAviationGroundCirculation1/Ralentissez	105	CanadaAviationGroundCirculation1/VirezADroite	106	CanadaAviationGroundCirculation1/VirezAGauche
107	CanadaAviationGroundCirculation2/AlimentationBranchee	108	CanadaAviationGroundCirculation2/AlimentationDeBranchee	109	CanadaAviationGroundCirculation2/Avancez	110	CanadaAviationGroundCirculation2/CalesEnlevees
111	CanadaAviationGroundCirculation2/CalesMises	112	CanadaAviationGroundCirculation2/FaitesTournerLaQueueVersLaDroite,	113	CanadaAviationGroundCirculation2/FaitesTournerLaQueueVersLaGauche	114	CanadaAviationGroundCirculation2/Halte
115	CanadaAviationGroundCirculation2/RalentissezLeMoteurDuCoteIndique	116	CanadaAviationGroundCirculation2/Reculez	117	CraneHandSignals/BoomUp SwatHandSignals1/Friendly	118	CraneHandSignals/EmergencyStop
119	CraneHandSignals/EverythingSlow	120	CraneHandSignals/LowerBoom SwatHandSignals1/Hostile			122	CraneHandSignals/RaiseLoadSlowly
124	CraneHandSignals/RoutineStop			121	CraneHandSignals/LowerLoadSlowly	123	CraneHandSignals/RetractBoom
125	CraneHandSignals/SwingBoom	126	CraneHandSignals/TrolleyOut	127	CraneHandSignals/WalkCraneForward RefereeVolleyballSignals1/Substitution	128	DivingSignals1/Around
130	DivingSignals1/Danger	131	DivingSignals1/DontKnow			129	DivingSignals1/ComeHere
132	DivingSignals1/OKsurface	133	DivingSignals1/Over	134	DivingSignals1/Under	135	DivingSignals1/Watch SwatHandSignals2/LookSearch
136	DivingSignals2/CannotOpenReserve	137	DivingSignals2/Cold	138	DivingSignals2/Help		
139	DivingSignals2/Me	140	DivingSignals2/Meet	141	DivingSignals2/OutOfAir	142	DivingSignals2/PressureBalancePb
143	DivingSignals2/ReserveOpened	144	DivingSignals2/Stop	145	DivingSignals2/You	146	DivingSignals3/Boat
147	DivingSignals3/Fast	148	DivingSignals3/NotUnderstood	149	DivingSignals3/Slowly	150	DivingSignals3/SomethingWrong
151	DivingSignals3/TieUp	152	DivingSignals3/Vertigo	153	DivingSignals3/Wreck	154	DivingSignals4/HoldHands
155	DivingSignals4/HowMuchAir	156	DivingSignals4/Lead	157	DivingSignals4/LevelOff	158	DivingSignals4/MoveApart
159	DivingSignals4/StayTogether	160	DivingSignals4/WhichWay	161	GangHandSignals1/Blood	162	GangHandSignals1/ComptonCrip
163	GangHandSignals1/Crip	164	GangHandSignals1/EastSide	165	GangHandSignals1/HooberCrio	166	GangHandSignals1/MafiaCrips
167	GangHandSignals1/UndergroundCrip	168	GangHandSignals2/Killas	169	GangHandSignals2/LatinKings	170	GangHandSignals2/OKBloodKilla
171	GangHandSignals2/OKCripKilla	172	GangHandSignals2/Piru	173	GangHandSignals2/WestCoast	174	GangHandSignals2/WestSide
175	HelicopterSignals/HoldHover	176	HelicopterSignals/Land	177	HelicopterSignals/LiftOff	178	HelicopterSignals/MoveDownward
179	HelicopterSignals/MoveForward	180	HelicopterSignals/MoveLeft	181	HelicopterSignals/MoveRight	182	HelicopterSignals/MoveUpward
183	HelicopterSignals/ReleaseSlingLoad	184	RefereeVolleyballSignals1/BallOut	185	RefereeVolleyballSignals1/BallOutAfterPlayerContact	186	RefereeVolleyballSignals1/DoubleFaultOrPlayover
187	RefereeVolleyballSignals1/EndOfGame	188	RefereeVolleyballSignals1/IllegalBlockOrScreen RefereeWrestlingSignals1/NeutralPosition	189	RefereeVolleyballSignals1/Timeout	190	RefereeVolleyballSignals2/BallInBounds
191	,RefereeVolleyballSignals2/BallServedIntoNetPlayerTouchingNet			192	RefereeVolleyballSignals2/BeckoningTheServe	193	RefereeVolleyballSignals2/CenterLineViolation
194	RefereeVolleyballSignals2/HeldThrownLiftedCarried	195	RefereeVolleyballSignals2/IllegalAttackOrBlockOverNet	196	RefereeVolleyballSignals2/LossOfRallyOrPoint	197	RefereeVolleyballSignals2/OutOfRotationOrOverlap
198	RefereeWrestlingSignals1/DeferChoice	199	RefereeWrestlingSignals1/FalseStart	200	RefereeWrestlingSignals1/FlagrantMisconduct	201	RefereeWrestlingSignals1/IllegalHold
202	RefereeWrestlingSignals1/InterlockingHands	203	RefereeWrestlingSignals1/NearFall	204	RefereeWrestlingSignals1/NoControl	205	RefereeWrestlingSignals1/out ofBounds
206	RefereeWrestlingSignals2/PotentiallyDangerous	207	RefereeWrestlingSignals2/Reversal	208	RefereeWrestlingSignals2/Stalemate	209	RefereeWrestlingSignals2/Stalling SwatHandSignals1/Breacher
210	RefereeWrestlingSignals2/StartInjuryClock TractorOperationSignals/RaiseEquipment	211	RefereeWrestlingSignals2/StopInjuryClock	212	RefereeWrestlingSignals2/StopMatch TractorOperationSignals/Stop		
		214	RefereeWrestlingSignals2/TimeOut			215	RefereeWrestlingSignals2/WrestlerInControl
213	RefereeWrestlingSignals2/TechnicalViolation TaxiSouthAfrica/TaxiHandSigns1	216	SurgeonSignals/ArmyNavyRetractor	217	SurgeonSignals/CurvedForceps	218	SurgeonSignals/CurvedScissors
		219	SurgeonSignals/NeedleHolder	220	SurgeonSignals/Scalpel	221	SurgeonSignals/StraightForceps
222	SurgeonSignals/StraightScissors	223	SurgeonSignals/Syringe	224	SurgeonSignals/TissueForceps	225	SwatHandSignals1/DogNeeded
226	SwatHandSignals1/MirrorNeeded	227	SwatHandSignals1/OKClear	228	SwatHandSignals1/Quickly	229	SwatHandSignals1/Stop
230	SwatHandSignals2/CoverNeeded	231	SwatHandSignals2/DoorClosed	232	SwatHandSignals2/DoorOpen	233	SwatHandSignals2/Listen
234	SwatHandSignals2/ManDown	235	SwatHandSignals2/Obstruction	236	SwatHandSignals2/ToMe	237	TaxiSouthAfrica/TaxiHandSigns10
238	TaxiSouthAfrica/TaxiHandSigns2	239	TaxiSouthAfrica/TaxiHandSigns3	240	TaxiSouthAfrica/TaxiHandSigns6	241	TaxiSouthAfrica/TaxiHandSigns9 TractorOperationSignals/LowerEquipment
242	TractorOperationSignals/ComeToMe	243	TractorOperationSignals/MoveOut	244	TractorOperationSignals/MoveTowardMe		
245	TractorOperationSignals/SlowItDown	246	TractorOperationSignals/SpeedItUp	247	TractorOperationSignals/StartTheEngine		
248	TractorOperationSignals/StopTheEngine	249	TractorOperationSignals/ThisFarToGo				

Figure 6. Gesture labels with their gesture name in our dataset. All the gesture name can be found in the CGD dataset, which can be found in the website: http://www.causality.inf.ethz.ch/Gesture/index.html. For some gesture labels, there are more than 2 gestures from different lexicons. For example, for the gesture label 10, the videos of the similar gestures are from two lexicons ("Mudra1/Chandrakala" and "Chinese/san").

ChaLearn Multi-Modal Gesture Recognition, ICMI Workshop, 2013. 1

[8] S. Escalera, J. Gonzàlez, X. Baró, and J. Shotton. Special issue on multimodal human pose recovery and behavior analysis. *IEEE Tans. Pattern Analysis and Machine Intelligence*, 2016. 1

[9] I. Guyon, V. Athitsos, P. Jangyodsuk, and H. J. Escalante. The chalearn gesture dataset (cgd 2011). *Machine Vision and Applications*, 25:1929–1951, 2014. 5

[10] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner. Results and analysis of the chalearn gesture challenge 2012. In *Advances in Depth Image Analysis and Applications*, pages 186–204. 2013. 1, 2, 3

[11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. 1

[12] F. Jiang, S. Zhang, S. Wu, Y. Gao, and D. Zhao. Multi-layered gesture recognition with kinect. *The Journal of Machine Learning Research*, 16(1):227–254, 2015. 6

[13] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):211–223, 2016. 5

[14] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, 2013. 1, 3

[15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[16] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 1

[17] S. Ruffieux, D. Lalanne, and E. Mugellini. Chairgest: a challenge for multimodal mid-air gesture recognition for close hci. In *ACM on International conference on multimodal interaction*, pages 483–488, 2013. 1, 3

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, a. A. C. B. Michael S Bernstein, and L. Feifei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2014. 1

[19] I. G. Sergio Escalera, Vassilis Athitsos. Challenges in multimodal gesture recognition. *Journal on Machine Learning Research*, 2016. 1

[20] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014. 1

[21] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898, 2014. 1, 5

[22] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, pages 2892–2900, 2015. 1

[23] J. Wan, G. Guo, and S. Li. Explore efficient local features from rgb-d data for one-shot learning gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 5

[24] J. Wan, Q. Ruan, W. Li, and S. Deng. One-shot learning gesture recognition from rgb-d data using bag of features. *Journal of Machine Learning Research*, 14:2549–2582, 2013. 2

[25] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5