# AUTOMATIC SIGN LANGUAGE IDENTIFICATION

*Binyam Gebrekidan Gebre*[*†] *Peter Wittenburg*[†] *Tom Heskes* [⋆]

[†]Max Planck Institute for Psycholinguistics, Nijmegen
[⋆]Radboud University, Nijmegen
{bingeb,peter.wittenburg}@mpi.nl,t.heskes@science.ru.nl

## ABSTRACT

We propose a Random-Forest based sign language identification system. The system uses low-level visual features and is based on the hypothesis that sign languages have varying distributions of phonemes (hand-shapes, locations and movements). We evaluated the system on two sign languages – British SL and Greek SL, both taken from a publicly available corpus, called Dicta Sign Corpus. Achieved average F1 scores are about 95% – indicating that sign languages can be identified with high accuracy using only low-level visual features.

***Index Terms***— Sign language, sign language identification, language identification

## 1. INTRODUCTION

Human language is expressed and interpreted in different modalities of communication. These include text, speech and sign. The task of automatic Language Identification (LID) is to quickly and accurately identify the used language as expressed in a given modality. The correct identification of a language enables efficient deployment of tools and resources in applications that include machine translation, information retrieval and routers of incoming calls to a human switchboard operator fluent in the identified language. All these applications need language identification systems that work with near perfect accuracy; but how accurate are language identification systems?

Language identification is a widely researched area in written and spoken modalities [1, 2, 3, 4, 5]. The literature shows varying degrees of success depending on the modality. Languages in their written forms can be identified to about 99% accuracy using markov models [1]. Languages in their spoken forms can be identified to an accuracy that ranges from 79-98% using different models (GMM, PRLM, parallel PRLM) [3, 6]. What are the results for automatic sign language identification?

[*]The research leading to these results has received funding from the European Commissions 7th Framework Program under grant agreement no 238405 (CLARA).

Even though extensive literature exists on sign language recognition [7, 8, 9, 10], to the best of our knowledge, little published work exists on automatic sign language identification and in this paper, we propose a sign language identification algorithm and report experimental results carried out on two sign languages (British and Greek). The best performance obtained, measured in terms of average F1-score, is 95% – much higher than 50% – a result we would expect from a random binary classifier. Interestingly, this performance is achieved using low-level visual features. The rest of the paper will give more details.

## 2. SIGN LANGUAGE PHONEMES

A signer of a given sign language produces a sequence of signs. According to Stokoe [11], each sign consists of phonemes called *hand-shapes*, *locations* and *movements*. The phonemes are made using one hand or both hands. In either case, each active hand assumes a particular *hand-shape*, a particular *orientation* in a particular *location* (on or around the body) and with a possible particular *movement*.

The aforementioned phonemes that come from hands make up the *manual signs* of a given sign language. But the whole message of a sign language utterance is contained not only in *manual signs* but also in *non-manual signs*. Non-manual signs include facial expressions, head/shoulder motion and body posture. This paper does not attempt to directly use non-manual signs for language identification.

The central idea of Stokoe's model is that signs can be broken down into phonemes and that the phonemes contrast only simultaneously. An alternative to Stokoe's model is that of Move-Hold model [12]. The Move-Hold model (M-H model) emphasizes the sequence aspect of segments of signs (i.e. signs are made up of sequences of *moves* and *holds*) and each segment is described by a set of features of *hand-shape*, *orientation*, *location* and *movement*.

This paper uses the idea than signs can be broken into phonemes, an idea that is common to both Stokoe's and M-H model. But it assumes that features extracted from frames are independent of each other. Every frame has features for hand-shapes, locations and movements. The movement features are extracted from two frames (current and previous frames).

## 3. METHOD

An ideal sign language identification system (SLID) should be independent of content, context, vocabulary and robust with regard to signer identity as well as to noise and distortions introduced by cameras. Some of the desirable features of an ideal SLID system are:

1. should be robust to intra- and inter-signer variability.
2. should be invariant to camera-induced variations (scale, translation, rotation, view, occlusion, etc).
3. increasing the number of target sign languages should not degrade performance (there are at least 200 sign languages [13]).
4. decreasing the duration of the test utterance should not degrade system performance.

Our proposed SLID system has subcomponents and each subcomponent attempts to address points 1, partly 2 (scale and translation) and 4. The system subcomponents are four: *a*) skin detection *b*) feature extraction *c*) modeling *d*) identification. We briefly describe each subcomponent in the following subsections.

### 3.1. Skin detection

We used skin color to detect hands/face [14, 15]. Skin-color has practically useful features. It is invariant to scale, orientation and it is easy to compute. But it also has two problems: *1*) perfect skin color ranges for one video do not necessarily apply to another *2*) some objects in the video have the same color as the hands/face. To solve the first problem, we did explicit manual selection of the skin-color RGB ranges in a way that is comparable to [16]; other skin detection approaches (i.e. based on parametric and non-parameteric distributions) did not perform any better on our dataset. To solve the second problem, we applied dilation operations and constraint rules to remove unexpected size of face/hands.

### 3.2. Feature extraction

Assuming the phonemes of sign language are formed from a set of *hand shapes/orientations/arrangements (N)*, in a set of *locations (L)* and with *movement types (M)*, we encode shapes using Hu-moments, locations using discrete grids (binary patterns) and movements as XORs of two consecutive location grids (binary patterns).

#### *3.2.1. Hand-shapes/Orientations*

To encode hand-shapes and orientations of the hands, we used the seven Hu set of invariant moments $(H_1 - H_7)$ [17] calculated from the gesture space of the signer, which are bounded by the external lines of the grids shown in figure 1. The seven Hu moments capture shapes and arrangements of the foreground objects (in this case, skin blobs). Formed by combining normalized central moments, these moments offer invariance to scale, translation, rotation and skew [17]. They are among the most widely used features in sign language recognition [10].

#### *3.2.2. Locations/Hand-arragements*

To encode hand locations of the signer, we used grids of *10x10* with the center of the face used as a reference. To find the center of the face, we used the Viola Jones face detector [18]. The position and scale of the detected face is used to calculate the position and scale of the grid. The center of the grid is fixed at the third and in the middle column (See figure 1). Each cell in the grid is a quarter of the height of the detected face [10]. A cell is assigned 1 if more than 50 percent of the area is covered by skin, otherwise, it will be assigned 0. These cells are changed into a single row vector of size 100 by concatenating the various rows – one after the other.

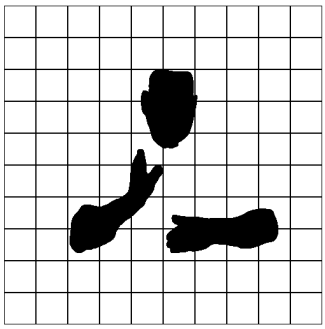

**Fig. 1**. Each cell in the grids is a square whose side is a quarter of the height of the face. The size of the face is determined by Viola Jones algorithm [18] using the data and implementation provided in the OpenCV library [19].

#### *3.2.3. Movements*

To encode the types of body movements, we compare the locations of hands/face in the current frame with respect to the previous frame. The motion is then captured by XORing (the absolute of pairwise element subtraction) two frame location vectors. The location vectors are obtained from the cell grids as described in 3.2.2.

### 3.3. Modeling – Random Forest

We use a random forest algorithm for sign language classification [20, 21]. Random forests generate many decision tree classifiers and aggregate their results [20]. Their attractive features include high performance [22], flexibility (no need

for feature normalization and feature selection) and stability (small parameter changes do not affect performance).

The random forest algorithm for classification works as follows:

1. Prepare $\{x, y\}$ pairs of training data.
2. Learn $N$ predictors as follows:

**Data**: $\{x, y\}$ pairs of data
**Result**: $N_{trees}$ predictors (Random forest)
Let $N_{trees}$ be the number of trees to build;
**for** *each of* $N_{trees}$ *iterations* **do**
  Select a new bootstrap sample from training set;
  Grow an un-pruned tree on this bootstrap:
  **for** *each node* **do**
    randomly sample $m$ of the feature variables;
    choose the best split from among those variables using gini impurity measure;
  **end**
**end**

**Algorithm 1:** Random forest training

3. Predict new data by aggregating the predictions of the $N_{trees}$ (majority votes for classification).

The random sampling of features at every node in a tree prevents random forests from overfitting and makes them perform very well compared to many other classifiers [20]. In our experiments, we fixed $N_{trees}$ to 10 and $m$ to 14 (14 is $\approx \sqrt{207}$, the size of our feature vector).

### 3.4. Identification

During identification, an unknown sign language utterance of frame length $T$ is first converted to frame vectors of length $T$, with each frame vector having features $x_t$ (207-dimension). These feature vectors are then scored against each language. With the assumption that the observations (feature vectors $x_i$)s are statistically independent of each other, the scoring function is a log-likelihood function and is defined as:

$$L(x/l) = \sum_{t=1}^{T} \log p(x_t/l), \tag{1}$$

where $T$ is the number of frames and $p(x_t/l)$ is a probability of $x_t$ for a given language $l$ (values returned by the random forest model [21]). The language $\hat{l}$ of the unknown utterance is chosen as follows:

$$\hat{l} = \arg\max_{l} (\sum_{t=1}^{T} \log p(x_t/l) + \log p(l)), \tag{2}$$

where $p(l)$ is the prior probability, which we fixed to $0.5$ (making it irrelevant in our experiments).

## 4. EXPERIMENT

We tested our sign language modeling and identification system on a part of data that is publicly accessible from the Dicta-Sign Corpus[1] [23]. The corpus has recordings for four sign languages with at least 14 signers per language and a session duration of approximately 2 hours using the same elicitation materials across languages. From this collection, we selected 19 signers for British and Greek sign languages[2]. The signers have been selected with the criterion that their skin color is distinct enough from background and their clothes. Table 1 gives details of the experiment data.

**Table 1**. Sign language identification: experiment data

| Sign Language | British | Greek | Total |
|---|---|---|---|
| Total length (in hours) | 8.9 | 7.17 | 16.07 |
| Number of signers | 9 | 10 | 19 |
| Number of clips | 186 | 209 | 395 |
| Average clip size (in minutes) | 2.86 | 2.06 | 2.46 |

## 5. RESULTS AND DISCUSSION

We evaluate performance of our identification system in terms of precision, recall and F1-score. We also evaluate the impact on performance of varying *a*) the number of training clips *b*) the length (in seconds) of test clips.

Table 2 indicates that high accuracy scores can be found by training on half the data and testing on the other. Figure 2 shows performance variations as a function of training data size and utterance length; it indicates that utterance length of 10 seconds is good enough to achieve about 90% F1 score. Ten seconds of utterance correspond to about 25 signs [24].

**Table 2**. Sign language identification: classification results

Number of training clips = 197 (random 50% of clips)
Number of test clips = 198 (rest 50%)
Clip size = 60 seconds

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| BSL | 0.94 | 0.96 | 0.95 | 94 |
| GSL | 0.96 | 0.94 | 0.95 | 104 |
| Average/total | 0.95 | 0.95 | 0.95 | 198 |

Are we identifying sign languages and not necessarily clips of the same signers? In order to answer this, we trained our system on clips of a group of randomly selected 11 signers and tested on clips of the rest 8 signers. Even though the

[1] http://www.dictasign.eu/

[2] British and Greek sign language DictaSign Corpora were immediately available online for our experiments.

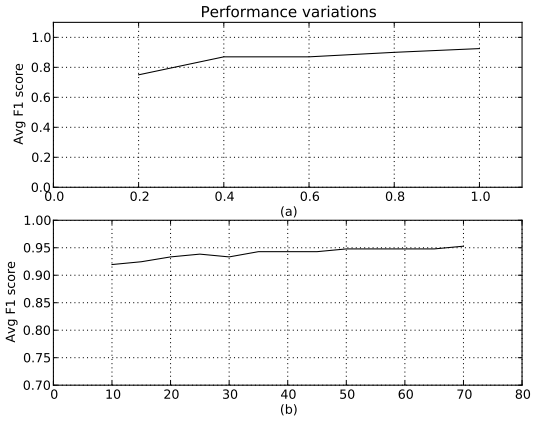


**Fig. 2**. (a) shows impact of varying the fraction of training data on average F1 score. (b) shows impact of varying test utterance length (in seconds) on average F1 score.

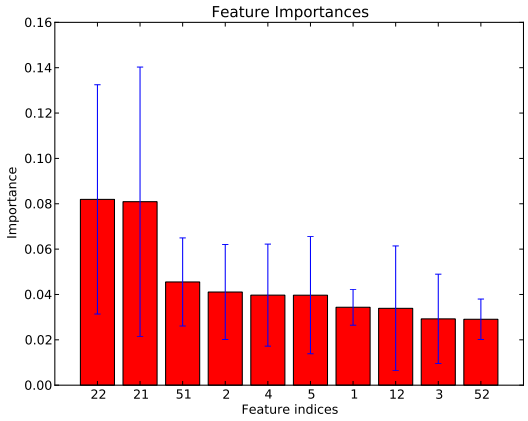


**Fig. 3**. shows the importance of the ten most informative features out of 207 features (7 for shapes, 100 for locations and another 100 for movements, indexed in that order). The error bars are variance of the feature importances for the ten trees.

scores are now less, the results show that our system is doing more than signer identity classification (See table 3).

**Table 3**. Signer independent classification results

Number of training clips = 248 (11 signers)
Number of test clips = 147 (from 8 unseen signers)
Clip size = 60 seconds

| | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| BSL | 0.77 | 0.72 | 0.74 | 64 |
| GSL | 0.79 | 0.83 | 0.81 | 83 |
| Average/total | 0.78 | 0.78 | 0.78 | 147 |

Are we really identifying sign languages and not some other random pattern? In order to answer this question, we assigned random labels to each clip and trained our system on random 50% of the clips and tested on the rest. Performance on different runs produced F1 scores that averaged to about 50% – indicating that our system is not picking any random pattern. What about systematic patterns like video or people characteristics that are unique to signers of each language?

The video characteristics of the two sign language corpora are similar as they were deliberately designed to be parallel for research purposes. But signer bodily characteristics of each sign languages could be different. How can we distinguish bodily characteristics from sign languages?

To answer this correctly, further research needs to be done with sign language clips produced by multilingual signers (the same signers producing utterances in two or more sign languages). For now, we can get insight by examining the most important features selected by the random forest classifier.

Figure 3 shows the relative importances of the ten most important features indexed by their position in the feature vector. The figure indicates that feature indices 22 and 21 are the most important. Interestingly, these refer to locations above the head slightly to the left. Indices 51 and 52 refer to locations lower (slightly right) of the chin of the signer. Most of the shape features (the Hu-moments) are also among the most important. None of the movement features ended up among the top ten.

## 6. CONCLUSIONS AND FUTURE WORK

This paper makes contribution to existing literature on automatic language identification by *a*) drawing attention to sign languages, and *b*) proposing one method for identifying them. The proposed sign language identification system (SLID) has the attractive features of simplicity (it uses low-level visual features without any reference to phonetic transcription) and high performance (it uses a random forest algorithm).

The system performs with an accuracy of about 95% (F1-score). From this performance, we can make one important conclusion: sign languages, like written and spoken languages, can be identified using low level features.

Future work should extend this work to identify several sign languages. Other possible sign language identification methods should also be explored (language identification methods that perform best in written and spoken languages are phonotactic – Ngram language models). Future work should also examine automatic phoneme extraction and clustering algorithms with the view to developing sign language typology (families of sign languages).

## 7. REFERENCES


[1] T. Dunning, *Statistical identification of language*, Computing Research Laboratory, New Mexico State University, 1994. 1

[2] Y.K. Muthusamy, E. Barnard, and R.A. Cole, "Reviewing automatic language identification," *Signal Processing Magazine, IEEE*, vol. 11, no. 4, pp. 33–41, 1994. 1

[3] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996. 1

[4] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and JR Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. ICSLP*, 2002, vol. 2, pp. 33–36. 1

[5] E. Singer, P. Torres-Carrasquillo, D. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. Sturim, "The mitll nist lre 2011 language recognition system," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012. 1

[6] E. Singer, PA Torres-Carrasquillo, TP Gleason, WM Campbell, and D.A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification," in *Proc. Eurospeech*, 2003, vol. 9. 1

[7] Thad Starner and Alex Pentland, "Real-time american sign language recognition from video using hidden markov models," in *Motion-Based Recognition*, pp. 227–243. Springer, 1997. 1

[8] Thad Starner, Joshua Weaver, and Alex Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998. 1

[9] Dariu M Gavrila, "The visual analysis of human movement: A survey," *Computer vision and image understanding*, vol. 73, no. 1, pp. 82–98, 1999. 1

[10] H. Cooper, E.J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *Journal of Machine Learning Research*, vol. 13, pp. 2205–2231, 2012. 1, 2

[11] W.C. Stokoe, "Sign language structure: An outline of the visual communication systems of the american deaf," *Journal of deaf studies and deaf education*, vol. 10, no. 1, pp. 3–37, 2005. 1

[12] S.K. Liddell and R.E. Johnson, *American sign language: The phonological base*, Gallaudet University Press, Washington. DC, 1989. 1

[13] "List of sign languages – Wikipedia," February 2013. 2

[14] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *Proc. Graphicon*. Moscow, Russia, 2003, vol. 3, pp. 85–92. 2

[15] S.L. Phung, A. Bouzerdoum Sr, and D. Chai Sr, "Skin segmentation using color pixel classification: analysis and comparison," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 1, pp. 148–154, 2005. 2

[16] J. Kovac, P. Peer, and F. Solina, *Human skin color clustering for face detection*, vol. 2, IEEE, 2003. 2

[17] M.K. Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179–187, 1962. 2

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. IEEE, 2001, vol. 1, pp. I–511. 2

[19] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*, O'Reilly Media, Incorporated, 2008. 2

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. 2, 3

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 2, 3

[22] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161–168. 2

[23] E. Efthimiou, S.E. Fotinea, C. Vogler, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, and J. Segouat, "Sign language recognition, generation, and modelling: a research effort with applications in deaf communication," *Universal Access in Human-Computer Interaction. Addressing Diversity*, pp. 21–30, 2009. 3

[24] E.S. Klima and U. Bellugi, *The signs of language*, Harvard University Press, 1979. 3