Towards A Framework for Recognising Sign Language Alphabets Captured Under Arbitrary Illumination

Zhao Wei Yun

***Abstract*—**Our work addresses the problem of automatically recognising a Sign Language alphabet from a given still image obtained under arbitrary illumination. To solve this problem, we designed a computational framework that is founded on the notion that shape features are robust to illumination changes. The statistical classifier part of the framework uses a set of weighted, self-learned features, i.e., binary relationship between pairs of pixels. There are two possible pairings: an edge pixel with another edge pixel, and an edge pixel with a non-edge pixel. This two-pairing arrangement allows a consistent 2D image representation for all letters of the Sign Language alphabets, even if they were to be captured under varying illumination settings. Our framework, which is modular and extensible, paves the way for a system to perform robust (to illumination changes) recognition of the Sign Language alphabets. We also provide arguments to justify our framework design in term of its fitness for real world application.

***Keywords*—**Sign Language alphabet recognition, raster image classification, unsupervised learning, illuminant-invariant features.

I. Introduction

An ideal Sign Language alphabet recognition system would accept a live video feed as its direct input, fully automated, and provide response in real-time. These requirements are a must due to the end goal of such systems, which is to fully replace the human translator.

In addition, the embedded algorithm should be robust to all kind of noises found inside the live video feed. It should be able to occlude the object of interest, i.e., the gesticulating signer's palm, from the background and from the signer's wrist. To further complicate the matter, variations in illumination and distance/angle of capture increase the arbitrary nature of the input.

In our proposed framework, we selectively ignored several decidedly peripheral problems from the list mentioned in the first paragraph, concentrating only on those that we believe as integral to our work. See II for our arguments on why we arrived at such decision.

II. Allowed Limitations

We assume that the palm gestures are captured as still images, extracted from a video feed. If the requirement is to detect the complete Sign Language vocabulary, then the usage of still images becomes inadequate. The Sign Language comprises of dynamic hand and palm configurations, body movements and positions, as well as facial expressions [1]. These elaborate, multi-step actions require more than a single frame to be properly captured. As for finger spelling the Sign Language alphabets, it only requires one unique palm configuration for each letter, and as such, can be succinctly captured using a single frame.

The problem of isolating the object of interest, i.e., the gesticulating signer's palm, from a static 2D input image is a well-explored subject, and successful methods are aplenty. Examples of such methods can be found here [2, 3, 4, 5, 6]. These methods can easily be implemented as an extra step in our extensible and modular framework. Hence, we decided to exclude this particular problem and concentrate fully on the image analysis part.

The two remaining problems are the varying illumination and camera set-up. The latter can be resolved via a proper positioning and orientation of both camera and signer. Since the translation itself is a deliberate act, i.e., both signer and translator are aware, and willing participants, regulating the above two variables is possible. As for illumination, it is a harder variable to regulate. This is particularly true during an outdoor image capture, with the presence of shadows, bright spots, et cetera.

III. Previous Work

There are two main approaches for Sign Language alphabet recognition: i) vision-based, and ii) glove-based.

Vision-based approaches exploit machine vision and image processing techniques to reveal object-specific features that can be learned using a statistical learning method. In [10], a video-based continuous Sign Language recognition system was developed for the German Sign Language. Using Hidden Markov Models (HMMs), a set of predetermined parameters regarding size, shape and position of the fingers and hands are learned as discriminative features. Recognition rate was reported at 95%, working on a 52 signs vocabulary. A visual-based approach for the recognition of American Sign Language sentences was presented in [9]. A small feature vector, consisting of parameters regarding position and orientation of the signer's palm is constructed from the input image. The feature vectors are then fed to HMMs for recognition purpose. Recognition rate was reported at 97% per word on a 40-word lexicon. In [11], SIFT features [17]

are used with Linear Discriminant Analysis (LDA) to recognise Arabic Sign Language. Dardas et al. [12] used k-means clustering and Support Vector Machine (SVM) to classify individual sign language characters, also using SIFT [17] as the discriminative features.

Glove-based approaches use sensors attached to the signer's palm. In [13], electromyography (EMG) electrodes are attached to the signer's palm and the produced signals are used to recognise the signed letter. In [16], the signer wears coloured gloves and the x and y position is tracked in real-time.

The problem of recognising the Sign Language alphabets is now reduced to a task of interpreting the data generated by the sensors. Although the level of precision is higher (than vision-based methods), the usage of gloves is obtrusive and may restrict the signer's palm and fingers' motion.

IV. Our Proposed Method

Our framework consists of three sequential and strictly independent modules: i) the *input normalisation* module, ii) the *binary features profiling* module, and iii) the *statistical classifier* module. Each module has a process output, which is then feed into the next module.

The *input normalisation* module removes any irregularities introduced by the illumination factor. An input image is subjected to an edge detector algorithm; with the remaining pixel values binarised to either value 0 or 1, based on a fixed threshold value. This two-step procedure ensures that each letter of the Sign Language is quasi-consistently represented inside the binary features space. The only variable here is the differing spatial information of the edge pixels due to the unique physical characteristics of the signer's palm, which we expect to learn from the positive training examples.

In the second module: the *binary features profiling* module, we build a profile containing binary values of all possible features. We now know the number of edge pixels, n, and non-edge pixels, m, with

$$m = (20 \times 20) - n. \quad (1)$$

Similar to [10], we use the simplest possible features, i.e., relationships between pixels. Our approach differs from theirs in term of the number of binary features used, and the slightly different pixel pairing arrangement. We pair each edge pixel with all the other edge pixels, and with all non-edge pixels. For each pixel pair (i, j), we effectively compute two binary features: $(pixel_i == pixel_j)$ and $(pixel_i \neq pixel_j)$. Thus, this process generates $p \times 2$ features, with

$$p = [n \times (n-1)] + [n \times [(20 \times 20) - n]]. \quad (2)$$

The output of this module is a list containing $p \times 2$ lines of 2-bit binary code.

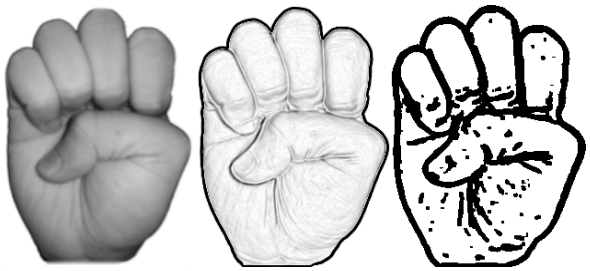
In the third module: the *statistical classifier* module, we feed the trained classifier with the binary feature profile obtained from the second module. The classifier can be trained using any machine-learning algorithms.

We used 8-bit greyscale images with 20×20 dimension as our input. The framework imposes 3 conditions on the state of the input:

1) The first requirement is the background noise must be zero. We have different strategies to achieve this with training and test examples. For training examples, we used a green screen with controlled ambient lighting during exemplar image capture. For test examples, we captured images with background noise (manually removed afterwards) and under random lighting conditions to introduce illumination artefacts, such as shadows, bright spots, et cetera.
2) For the second requirement, the signer's palm must be tightly cropped inside the captured image, with no in-plane and out-of-plane rotations. In both training and test examples, the signer's wrist is manually removed from the final image.
3) As for the third requirement, the image pixel intensities are scaled linearly to span between 0 and 255.

V. Results

To validate our framework, we trained a classifier for the letter 'e' from 50 positive examples (5 different signers with 10 examples each) using Adaboost [7]. The examples were edge-filtered using an image editing software, and the remaining pixel values binarised via thresholding, before being fed to an ad hoc binary features profiling program. Using a modified MATLAB implementation of the Adaboost algorithm [7], we built a classifier from the obtained profiles.



(a) (b) (c)

Fig. 1 A greyscale image of a signer's palm gesturing for the letter 'e', before (a), and after (b) being subjected to an edge detector algorithm. The binarised image (c), with the global threshold value set to 128.

We obtained 10 additional positive examples, and 10 negative examples from the same set of five signers (each contributing two extra positive examples, and two negative examples). For the positive set, the classifier correctly classified 7 out of 10 examples (70% accuracy). For the negative set, the classifier correctly classified 9 out of 10 examples (90% accuracy).

VI. Discussions and Conclusion

Our framework paves the way for a complete system to perform a robust (to illumination changes) recognition of the

Sign Language alphabets. The framework design is modular and extensible, allowing easy integration with additional steps when required.

We validated our framework by demonstrating a single case of the letter 'e'. It is plausible for the classifier to score a better accuracy rate if it were to train using a larger set of positive examples. We observed that methods such as [8] using Adaboost [7] to train their classifiers, used an average of 3000 positive examples per class.

As for our future work, adopting a cumulative voting scheme into our framework, as implemented in [14] and [17], might improve the classification accuracy further. We observed that in most false positive classification results, a majority of the top 5 returned matches belongs to the letter 'e' (the top match being another letter besides 'e'). By using cumulative voting, we can reduce such anomalous results.

REFERENCES

[1] American Sign Language, National Institute on Deafness and Other Communication Disorders, http://www.nicdcd.nih.gov/health/hearig/asl.asp

[2] Binh N.D., E. Shuichi and T. Ejima, "Real-time hand tracking and gesture recognition system", In Proceedings of International Conference on Graphics, Vision and Image Processing (GVIP-05) Conference, Dec 2005, pp. 362-368. Available: http://nguyendangbinh.org/publications.html

[3] Bradski G.R. and S. Clara, "Computer vision face tracking for use in a perceptual user interface", Intel Technology Journal, Vol. 2 Issue 2, 1998. Available: http://developer.intel.com/technology/itj/q21998/articles/../pdf/camshift.pdf

[4] Jennings C., "Robust finger tracking with multiple cameras", In Proceedings of International Workshop on Recognition, Analysis, and tracking of Faces and Gestures in Real- Time Systems, 1999, pp. 152-160.

[5] Segen J. and S. Kumar, "Shadow gestures: 3D hand pose estimation using a single camera", In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR99), Vol.1, 1999, pp. 479–485. Available: http://www.cis.udel.edu/~chandra/sigai/kumar.cvpr.pdf

[6] Zhu X., J. Yang, and A. Waibel, "Segmenting hands of arbitrary color", In Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition, March 2000, pp. 446-453. Available: http://isl.ira.uka.de/publications/FG2000-jerry.pdf

[7] R. Schapire, Y. Freund, P. Bartlett, W. Lee, "Boosting the Margin: A New Explanation for the effectiveness of voting Methods", Inn Proceedings of the 14th International Conference on Machine Learning (ICML97), July 1997, pp. 322-330. Available: http://www.iro.umontreal.ca/~kegl/ift3390/2006_1/Lectures/l07_ExplanationBoostingBartlett.pdf

[8] Baluja, S. and Rowley, H.A., "Boosting sex identification performance", In Proceedings of the International Journal of Computer Vision (ICIV07). Vol.71, Issue 1, January 2007, pp111-119. Available: http://www.cs.cmu.edu/~har/iaai2005.pdf

[9] Starner, T., Weaver, J., and Pentland, A., "Real-time American sign language recognition using desk and wearable computer-based video", IEEE Transactions on Pattern Analysis and Machine Intelligence, December 1998, Vol. 20, No. 12, pp. 1371-1375

[10] Hienz, H., Bauer, B., and Kraiss, K.-F, "HMM-Based Continuous Sign Language Recognition Using Stochastic Grammars", GestYvette, France (1999).

[11] Tharwat A., Gaber T., Hassanien A. E., Shahin M., Refaat B. (2015) Sift-based arabic sign language recognition system. In: Afro- European conference for industrial advancement, Springer, pp 359–370

[12] Dardas N, Chen Q, Georganas ND, Petriu EM (2010) Hand gesture recognition using bag-of-features and multi-class support vector machine. In: Haptic audio-visual environments and games (HAVE), 2010 IEEE international symposium, IEEE, pp 1–5

[13] Kainz O, Jakab F (2014) Approach to hand tracking and gesture recognition based on depth-sensing cameras and EMG monitor- ing. Acta Inf Prag 3:104–112

[14] I. Hipiny and W. Mayol-Cuevas, "Recognising Egocentric Activities from Gaze Regions with Multiple-Voting Bag of Words," Technical Report CSTR12-003, University of Bristol, 2012.

[15] Lowe, David G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints". International Journal of Computer Vision. 60 (2): 91–110. doi:10.1023/B:VISI.0000029664.99615.94

[16] Grobel K, Assan M (1997) Isolated sign language recognition using hidden Markov models. In: Systems, Man, and Cybernetics, 1997. Computational cybernetics and simulation. 1997 IEEE international conference, IEEE, pp 162–167

[17] H. Ujir, L. C. Sing and I. Hipiny, "A modular approach and voting scheme on 3D face recognition," in International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2014.