Sign Language Recognition

Inflated 3D ConvNet

Nguyen To

Master's Thesis

UNIVERSITY OF
EASTERN FINLAND

Faculty of Science and Forestry

School of Computer Science

Abstract:

Effective communication is considered as the foremost fundamental of human skills. However, more than 5% of the world's population is suffering from disabling hearing loss, as indicated by the World Health Organization (WHO). There is hence a communication gap between the hearing-impaired community, whose primary means of communication is sign language, and others who are not privy with this language. To this end, Sign Language Recognition could be an essential instrument which utilizing vision-based technology and helps the hearing-impaired communicate with the society with ease, thereby diminishing the verbal exchange barrier. The first step in interpreting and analyzing communication via gestures is word-level sign language recognition (WSLR). Recognizing signs from recordings maybe a tough challenge due to the fact that the meaning of a word is determined by a combination of subtle body movements, hand gestures, and other actions. Be that as it may, with the significant advancement of technology, notably Convolutional Neural Network (CNN), in this paper, an Inflated 3D Networks (I3D), a 3D video categorization solution are used to method as an answer of WSLR.

Preface

The basis for the project initially originated from my ardor for developing better methods of sign language recognition. The hearing-impaired community's life has changed considerably over the past half-century as a result of policy adjustments and latest technological tendencies, yet the road to find a viable answer for word-level sign language recognition (WSLR) keeps on being a drawn out circumstance. It is my passion to not solely find out, however to develop tools to bridge the communication gap between the deaf community and the society. This project follows the reference and citation guidelines of the "Quo Valdis, Action Recognition? A New Model and the Kinetics Datasets" by a group of João Carreira and Andrew Zisserman.

List of Abbreviations

ACM Association for Computing Machinery

ISY Itä-Suomen yliopisto

UEF University of Eastern Finland

WSLR Word-level Sign Language Recognition

I3D Two -Stream Inflated 3D Convolutional Network

CNN Convolutional Neural Network

LSTM Long-short Term Memory

TGCN Temporal Graph Convolutional Network

Contents

List of Tables

List of Figures

1   Introduction

Sign Language, any methods of communicating by bodily motions, particularly with hands and arms, that is utilized when verbal communication is either difficult or undesirable. Sign language can consist of a series of overly-exaggerated facial expressions, shrugs, or hand gestures; or it can be a fine and delicate mix of hand signals that are complemented by facial expressions and words spelt out using a manual alphabet. When a deaf person or someone speaking a different language is communicating with someone who is hearing, using sign language can help connect the parties. (Britannica, 2020, November 12). The public has neither the time nor the patience to learn sign language, which is complicated and time-consuming to learn and practice. Additionally, there are also many language and culture-specific (Holtz, 2014) (e.g Germany, Japanese) constraints which will hinder the widespread adoption of sign language. Significant advances in deep learning (DL) and improvements in device capabilities, such as computation power, memory capacity, power usage, sensor resolution, and optics, have improved the performance and cost-effectiveness of vision-based applications, allowing them to spread more quickly in the market place. For this reason, it is interesting to examine sign language recognition (SLR), which automatically translates sign language and aids deaf-mute individuals in communicating with others in their life

Back to the history of 90s, Yann LeCun et al. published "Gradient-Based Learning Applied to Document Recognition", which is widely considered to be the most popular AI article from the era. This paper was the first modern application of convolutional neural networks to be developed. Since then, more and more sophisticated models trained on ever-larger datasets have been built using the convenient approach of convolutional neural networks. Especially in the field of Computer Vision - Human-based activity recognition, there are many methods can be applied to solve the problem, from traditional convolutional neural networks such as CNN-RNN, CNN-LSTM to RestNetCRNN, Conv3D and state-of-the-art networks e.g Pose-TGCN, I3D. Inheriting the idea of using two-stream I3D network, which is based 2D ConvNet (Carreira & Zisserman, 2017), presented by Carreira and Zisserman, this project is re-implement the model with a slightly modification inside. It might not be better when comparing with other models, however during the project, I have got many experience and broaden my

knowledge on the field of Deep Learning.

1.1 Problem

As same as other human-based activity recognition, SRL also shares some common problems such as background clutter, lightning or lightning changing in a video, motion blur, angle of camera, changing scale. SRL, on the other hand, is a more difficult task than ordinary action recognition. Firstly, sign language relies on a combination of global body movement and subtle hand/arm gesture. Additionally, depending on how many times they are repeated, same gestures might have different meanings. SRL might be more difficult to examine because of different states of motions and signers such as localism, gesture speed, preferred hand or physical form. Finally, it is also expensive to collect additional data from many signers even though it is desirable (Jiang et al., 2021).

As described above, the datasets that uses for training SLR are limited, even the number of samples inside each dataset. The table below describes some datasets that normally use for researching.

Table 1: Sign Language Datasets.

Dataset	Language	Classes	Samples	Data Type	Language Level
CSL Dataset I	Chinese	500	125,000	Video & Depth from Kinect	Isolated
CSL Dataset II	Chinese	100	25,000	Videos & Depth from Kinect	Continuous
RWTH-PHOENIX-Weather 2014	German	1,081	6,841	Videos	Continuous
RWTH-PHOENIX-Weather 2014 T	German	1,066	8,257	Videos	Continuous
ASLLVD	American	3,300	9,800	Videos(multiple angles)	Isolated
ASLLVD-Skeleton	American	3,300	9,800	Skeleton	Isolated
SIGNUM	German	450	33,210	Videos	Continuous
DGS Kinect 40	German	40	3,000	Videos(multiple angles)	Isolated
DEVISIGN-G	Chinese	36	432	Videos	Isolated
DEVISIGN-D	Chinese	500	6,000	Videos	Isolated
DEVISIGN-L	Chinese	2000	24,000	Videos	Isolated
LSA64	Argentinian	64	3,200	Videos	Isolated
GSL isol.	Greek	310	40,785	Videos & Depth from RealSense	Isolated
GSL SD	Greek	310	10,290	Videos & Depth from RealSense	Continuous
GSL SI	Greek	310	10,290	Videos & Depth from RealSense	Continuous
IIITA -ROBITA	Indian	23	605	Videos	Isolated
PSL Kinect	Polish	30	300	Videos & Depth from Kinect	Isolated
PSL ToF	Polish	84	1,680	Videos & Depth from ToF camera	Isolated
BUHMAP-DB	Turkish	8	440	Videos	Isolated
LSE-Sign	Spanish	2,400	2,400	Videos	Isolated
Purdue RVL-SLLL	American	39	546	Videos	Isolated
RWTH-BOSTON-50	American	50	483	Videos(multiple angles)	Isolated
RWTH-BOSTON-104	American	104	201	Videos(multiple angles)	Continuous
RWTH-BOSTON-400	American	400	843	Videos	Continuous
WLASL	American	2,000	21,083	Videos	Isolated

Time segmentation is another issue for SLR as it is difficult to distinguish different kinds of sign language while signers make gestures continuously to describe a

phrase or a sentence (Xiao et al., 2020). Word-level sign recognition, an integral part of comprehending sign language phrases and sentences, is also extremely difficult task itself:

- The meaning of signals is primarily determined by the mix of bhand actions, body motions and head positions, and small variations in these elements can result in a variety of interpretations.

- With the same gesture, depend on the natural languages and context, they might have different meaning. It is also possible for nouns and verbs from the same lemma to share the same sign. These nuances are not effectively reflected by the small-scale datasets that are currently available (Figure 1) (D. Li et al., 2020).

- The number of signs that are used on daily basis is enormous, it could be thousands. In comparison, tasks such as gesture and action recognition have just had a few hundred categories. The scalability of recognition algorithms is significantly hampered as a result.
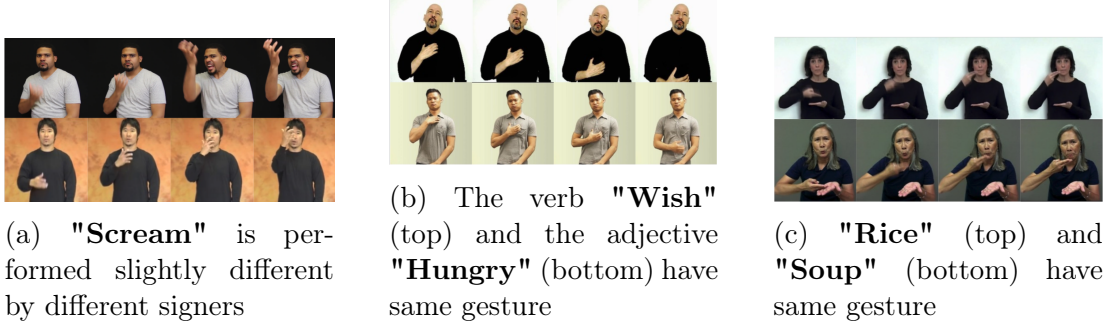

(a) **"Scream"** is performed slightly different by different signers

(b) The verb **"Wish"** (top) and the adjective **"Hungry"** (bottom) have same gesture

(c) **"Rice"** (top) and **"Soup"** (bottom) have same gesture

Figure 1: Common occurrence of ambiguity and variation signing

1.2 Thesis scope and target

In fact, the researching of SLR, including isolated (word-level) and continuous (phrase or sentences), has yielded significant results and developed rapidly in recent years (Lim, Tan, Lee, & Tan, 2019)(Y. Li, Wang, Liu, & Feng, 2018)(Mercanoglu Sincan, Junior, CS, Escalera, & Yalim Keles, 2021). So this project aims to understand the architecture of I3D model, focuses on the methodology of training, evaluating the network with given dataset. A further objective of the project

is getting familiar with libraries and frameworks of Deep Learning which is built for Python environment, and also other related topics such as image processing, data labelling, development tools e.g Google Colab, Anaconda, Jupyter Notebook. Regarding data and dataset, the project gave a comprehensive overview of organizing data, loading, splitting data for training and testing purpose.

Finally, in terms of the topic's aim, this thesis primarily focuses on enhancing the accuracy of the dataset with the two-stream I3D network. Because of the limitation of facilities, such as hardware, and time-consuming of training process, the project only used the WLASL dataset. In order to reduce the consumption of training time, as well as to test the effectiveness of I3D network, the dataset is divided into many sub-datasets which have 100 classes, 300 classes, 1000 classes and 2000 classes.

2 Model

2.1 General algorithm for video classification

A video is a collection of multi images which are represented by sequential. Therefore, the main idea of clarifying an action inside a video is analyzing frame by frame. In details, the general procedure of video recognition (Sivic & Zisserman, 2003)(Niebles et al., 2010) contains three key stages. The video is first divided into regions (Liu et al., 2009) based on the places that are easy to characterize in visual terms. This is done by either sampling regions densely (Wang & Schmid, 2013) or, if there are few areas of interest (Laptev, 2005), by using a sparse sampling technique. In the next step, the characteristics are merged into a video level description with defined size. One common method is to train a K-means dictionary and then evaluate all the features. This allows to gather visual words during the duration of the video and then arrange them into histograms of various spatio-temporal positions and extents (Karpathy et al., 2014)(Laptev et al., 2008).

References

Britannica, T. E. o. E. (2020, November 12). Sign language. encyclopedia britannica [Computer software manual]. Retrieved from `https://www.britannica.com/topic/sign-language`

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6299–6308).

Holtz, R. (2014). Reading between the signs: Intercultural communication for sign language interpreters. *Journal of International Students*, *4*(1), 106–108.

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton aware multi-modal sign language recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3413–3423).

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1725–1732).

Laptev, I. (2005). On space-time interest points. *International journal of computer vision*, *64*(2), 107–123.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *2008 ieee conference on computer vision and pattern recognition* (pp. 1–8).

Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1459–1469).

Li, Y., Wang, X., Liu, W., & Feng, B. (2018). Deep attention network for joint hand gesture localization and recognition using static rgb-d images. *Information Sciences*, *441*, 66–78.

Lim, K. M., Tan, A. W. C., Lee, C. P., & Tan, S. C. (2019). Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools and Applications*, *78*(14), 19917–19944.

Liu, J., Luo, J., & Shah, M. (2009). Recognizing realistic actions from videos "in the wild". In *2009 ieee conference on computer vision and pattern recognition* (pp. 1996–2003).

Mercanoglu Sincan, O., Junior, J., CS, J., Escalera, S., & Yalim Keles, H. (2021). Chalearn lap large scale signer independent isolated sign language recognition challenge: Design, results and future research. *arXiv e-prints*, arXiv–2105.

Niebles, J. C., Chen, C.-W., & Fei-Fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on computer vision* (pp. 392–405).

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Computer vision, ieee international conference on* (Vol. 3, pp. 1470–1470).

Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the ieee international conference on computer vision* (pp. 3551–3558).

Xiao, Q., Chang, X., Zhang, X., & Liu, X. (2020). Multi-information spatial–temporal lstm fusion continuous sign language neural machine translation. *IEEE Access*, *8*, 216718–216728.