



國立台灣科技大學  
資訊管理系

---

碩士學位論文

Transition Motion Synthesis for Video-Based Text to ASL

研 究 生：Yulia

學 號：M10609803

指導教授：Prof. Chuan-Kai Yang

中華民國一百零八年七月二十六日





# Abstract

This research describes a novel approach to provide a text to ASL media, a Video-Based Text to ASL. The hearing impaired or we called as the Deaf are used to communicate using Sign Language. When they have to face the spoken language, they have difficulties to read the spoken words as fast as the hearing people.

The availability of a public dataset named ASL Lexicon Dataset give the challenge to make the video-based interpreter for the Deaf. The problem is on the transition from one word to another since it does not exist in the original dataset. Regarding to this case, our focus in on how to make a better transition from one word to another rather than a blink.

After the dataset has been pre-processed, they are fed to OpenPose library to extract the skeleton of the signers and save it as JSON files. The system requires the user to input some glosses by text, then it will find the JSON files and the videos for the corresponding glosses. The whole sequences of original video are also fed into the system to be used as a transition pools. Later, the corresponding frames of the glosses are input together with the transition pools to construct the sequence transition frames. After getting the sequences, a smoothing algorithm is applied to enhance the smoothness of the motion.

Since this algorithm is fully depends on the transition pulls, there are some limitation regarding to make a good transition. If the transition frames we need to make a logically and visually correct motion are not available, then the result will be not optimized. But as long as the frames we need are

available, this system can generate a logically and visually correct transitions.

***Keywords: ASL, Sign Language, Deaf Talk, OpenPose, Transition Motion Synthesis***



# Acknowledgements



# Contents

Recommendation Letter . . . . .	i
Approval Letter . . . . .	ii
Abstract . . . . .	iii
Acknowledgements . . . . .	v
Contents . . . . .	vi
List of Figures . . . . .	vii
List of Tables . . . . .	viii
List of Pseudocodes . . . . .	ix
1 Introduction . . . . .	1
1.1 Motivation . . . . .	1
1.2 Contribution . . . . .	4
1.3 Outline . . . . .	4
2 Literature Review . . . . .	5
2.1 Introduction to ASL . . . . .	5
2.2 Previous 3D Based text to ASL system . . . . .	7
2.3 ASL Lexicon Video Dataset . . . . .	7
2.4 OpenPose . . . . .	10
References . . . . .	13

# List of Figures

2.1	One of the ASL Lexicon Dataset frontal views (left), the face view (middle), and the side view (right), for a frame in a video sequence [1]. . . . .	8
2.2	Annotation of ASL Lexicon Video Dataset. Retrieved from the annotation file of [1]. . . . .	9
2.3	Left: Part Affinity Fields corresponding to the limb that connected elbow and wrist. Right: a 2D vector in each pixel of every PAF which encoded the position and orientation of the limbs [2]. . . . .	10
2.4	Overall system flow. (a) Entire image is taken as the input for a CNN to predict the join (b)confidence maps for body part is detected (c) detection of PAFs (d) bipartite matching to associate body part candidates (e) assembly of PAFs into full body poses for all detected people in the image [2]. . .	11
2.5	Output of OpenPose, body, foot, hand, and facial keypoints are detected in a real-time manner. [2]. . . . .	12



# List of Tables



# List of Pseudocodes



# Chapter 1 Introduction

## 1.1 Motivation

According to World Health Organization (WHO) [3], there are over 5% or about 466 million of world's population who have suffered of hearing disability. This number consists of 432 million adults and 34 million children. It is also predicted that by 2050, there will be over 900 million people or one in every ten people that will have the hearing disability.

A person can be said to have a hearing loss if he or she is not able to hear with a normal hearing thresholds of 25dB or better in both ears. The hearing loss may be divided into mild, moderate, severe or profound category. 'Hard of hearing' term refers to the individuals who have hearing loss ranging from mild to severe category. These people usually still can communicate using the spoken language and can get the benefit of hearing aids, cochlear implants and other devices. 'Deaf' refers to the people who mostly have profound hearing loss, which indicates the hearing ability is very little or no hearing.

Sign Language is a language developed as a primary language by the Deaf community. These people often use Sign Language to communicate with each other. Although it is used as the primary language for the Deaf, it also can be a way to communicate for the Hard of Hearing, hearing individuals which are unable to physically speak, the people who have trouble with spoken language due to a disability or condition, and the Deaf family.

Similar to spoken language, different country has different Sign Language grammar and lexicon [4]. This fact shows that sign language is

unique in each world's region, not universal and not mutually understandable.

It has been proved that the reading ability of the deaf high school students is equal to the non-deaf students which are seven years younger. The median reading skills of Deaf students between 8 and 18 years old is equal to the non-deaf students in their fourth grade. The lack of phonemic awareness does have a big impact to their reading fluency. The Deafs are at a disadvantage compared to the hearing individuals because they cannot implicitly learn the relationship between letters and sounds without direct instruction and access to sound.

American Sign Language (ASL) is a natural language which dominates sign language of Deaf communities in the United States and most of Anglophone Canada. Not only North America, the dialects of ASL and ASL-based creoles are also used in many countries, including much West Africa and parts of South East Asia. ASL have some phonemic components, including the movements of the face and torso, together with the hands.

Based on U.S. Bureau of Labor Statistics [5], the median annual wage for interpreters and translators was \$49,930 in May 2018. The growth of interpreters and translators for Deaf is projected to be 18 percent from 2016 to 2026. Meanwhile the average for all occupation is just around 7 percent. So, it means although the wage of an interpreter is quite high, the demands of them is definitely growing from time to time.

Regarding the importance of ASL for Deaf, Athitsos et al. [1] made the availability of ASL Lexicon Video Dataset, a public dataset containing video sequences of thousands of distinct ASL signs, along with the anno-

tations of those sequences. This dataset was made of purpose to provide a sign language dictionary for the hearing people and it hopes to be able to provide a baseline for sign language recognition. In this research, we want to try to make use of this dataset to provide a video-based interpreter for the Deaf.

The mentioned ASL dataset is a big collection of Sign Language videos presented by several native signers. A video contains gestures of several words based on Gallaudet Dictionary of American Sign Language. So, for each gesture to be presented in a video, the signer has to look into the guide and remember the gesture, and demonstrate it later. While the signer was looking to the guide, the camera was also still recording. Because of this reason, before and after a word was demonstrated, the signer will always stay in his/her resting position.

Detecting human body skeleton is a challenging research that has been done by various researchers. Many methods and implementations were proposed, including a method named Part Affinity Fields. This approach uses a nonparametric representation to learn to associate body parts with individuals in the image. An open library called OpenPose has made an implementation of it. The library can help us to detect human body skeleton from a 2D image or a 2D video in a real time manner.

In this research, we try to develop a transition motion selection algorithm to join some words from the ASL dataset into a video. We try to connect the gestures from the words to remove the signer's resting pose by finding the best candidates of frames collection and put them in-between to make the movement looks like they are continuous. We utilize the OpenPose algorithm to help us retrieve the similar frames.

## 1.2 Contribution

In this research, we develop a system that can automatically generated a video-based human interpreter for the Deaf. In our system, the user has to input some gloss and define the threshold to the system. The contributions of this research are as follow:

1. This research introduces a new idea of using The American Sign Language Lexicon Video Dataset [1] to provide a video-based human interpreter for the Deaf.
2. An approach to calculate similarity using the help of 2D skeleton detection library is implemented to support the proposed algorithm for the frame selection.
3. The quality of the proposed method is evaluated by a user study, participated by native ASL signers and a smoothness evaluation method is also presented.

## 1.3 Outline

The remainder of this thesis is organized as follow: introduction to ASL, previous text-to-ASL works, ASL Lexicon dataset, OpenPose Library and smoothness evaluation method are described in Chapter 2. The proposed system including the system architecture design and the methodology that are used in this research are explained in Chapter ???. The results of the experiments conducted are discussed in Chapter ??. Finally, this research is concluded in Chapter ??.

## **Chapter 2      Literature Review**

### **2.1    Introduction to ASL**

In the early of 1800s, the population of deaf Americans was only about a few thousand. The Deaf communities made various signing systems and at that time, there was no standard sign language existed. These situation are now called as an Old American Sign Language. The current used American Sign Language is actually related to this language [6].

The beginning of the current used American Sign Language was started in 1814 with Dr. Thomas Hopkins Gallaudet, a minister from Hartford, Connecticut. The neighbor of him, named Mason Fitch Cogswell, had a nine years old deaf daughter, Alice Cogswell. Although Alice could not speak or hear, Dr. Gallaudet found that she was very smart, so that he wanted to teach her how to communicate. He then had a little success in making Alice enabled to read and spell, but he felt the method he used to teach was not effective. Therefore, Dr. Gallaudet gained a community support and money to go to Europe, in order to learn the best educational methods for the deaf, since there was a history of deaf education in Europe.

Arrived at Europe, Gallaudet met three other figures, Abbe Sicard, Jean Massieu, and Laurent Clerc. Sicard was a successor of Abbe de l'Epee's at the National Insitute for Deaf-Mutes. The other two learned the deaf education from Sicard and became expert deaf educators. Gallaudet then learned how to teach the deaf from these instructors and yet he took private lessons with Clerc, the one of the best teachers at the institute.

After accomplished the study, Gallauded travelled back to America,

accompanied by Clerc. He proposed to Clerc to join him because he knew Clerc would be a huge help in starting a school for the deaf, which is now known as the American School for the Deaf. This school was established in Gallaudet hometown, Hartford, in 1817 as the first public free deaf school in the U.S [7]. At this point, Gallaudet and Clerc had successfully make a huge milestone in American Deaf history.

The name of the school was quickly spread around the U.S., making deaf students all over the country came together to study in this school. Just like the school in Europe, the students also brought their own local sign language to the school. American Sign Language derived from these signs and from French Sign Language that was learned from Clerc. Gallaudet retired in 1830, but the taught for the deaf was still continued by Clerc until the 1850s. By 1863, there were twenty two deaf schools in the U.S and most of them were founded by Clerc's students. The school founded by Clerc's students continued to use Clerc's deaf education methods.

After Thomas Hopkins Gallaudet passed away in 1851, his youngest son, Edward Miner Gallaudet continued the work of him in deaf education. Edward then became a teacher at his father's school and in 1857, he was asked to be the headmaster of the Columbia Insitution for the Deaf and Dumb and the Blind in Washington, D.C. Edward also had a big role in deaf education since he presented an idea to provide a deaf college to Congress and in 1864, the permit to issue a college degrees for Columbia Insitute was passed. In 1864, the Columbia Insitute's first college division for the deaf named the National Deaf-Mute College was opened. The college was renamed to Gallaudet College tributed to Thomas Hopkins Gallaudet. Later in 1986, it was renamed to Gallaudet University, and it is known today as



the first and only deaf university in the world.

The other schools spreaded around the U.S also took roles to pass down their education from the American School for the Deaf to the next generation of deaf students the contact language that has known as today's American Sign Language. In 1900s, the nationwide network of deaf schools was completed. The community were given the opportunity to be with other deaf to share their sign language and cultural experiences without any communication barriers. The American Sign Language used today is a result of almost two hundred years of deaf people passing down their language from one to another generation that now become one of the most used languages in the U.S [8].



## **2.2 Previous 3D Based text to ASL system**

## **2.3 ASL Lexicon Video Dataset**

The American Sign Language Lexicon Video Dataset [1] was made primarily to provide a dictionary of Sign Language. Usually when we face an English word that we do not know, we will look it up in a dictionary. But, when an American Sign Language user encounters an unknown sign, the media for s/he to look it up is very limited.

The ASL Lexicon Video Dataset is a public dataset which contains thousands of distinct ASL signs in the form of high-quality video sequences. This dataset was made as a part of a project to provide a vision-based system used to look up the meaning of an ASL sign. The main part of this collection is its comprehensiveness. Some approaches to recognize signs in

a vision-based manner only could cover small vocabularies (20-300 signs) and often rely on color markers [9], [10]. Therefore, this dataset was hoped to be able to make a trend for developing vision-based methods that operate on markerless images and can cover a broader vocabulary.

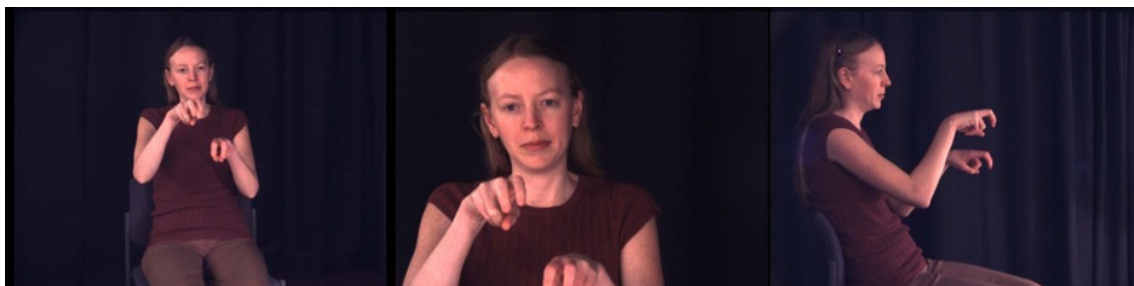


Figure 2.1: One of the ASL Lexicon Dataset frontal views (left), the face view (middle), and the side view (right), for a frame in a video sequence [1].

The total number of signs contained in the dataset has a similar scale and scope with the existing English-to-ASL dictionaries [11], [12]. This dataset has at least one sign example figured by a native signer, for almost about 3,000 sign contained in the Gallaudet Dictionary of American Sign Language [12].

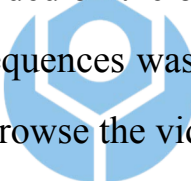
Usually for their communication, ASL users sometimes spell out English words using the alphabet. Along with the existing dataset [11], [12], in The ASL Lexicon Video dataset, the fingerspelled signs were not included, as they were counted as "loan signs".

In the described dataset, each sign was performed by native ASL signers. The number of involved native ASL users for this dataset was four people. The video sequences were collected using a four-camera system that concurrently captured two-frontal views, one side view and one view zoomed in on the face of the signer. For the side view and the two frontal views, the upper body took over a relatively large part of the scene frames.

For the face view, the frontal view of the face which took over the large part of the scene frames. All video sequences were recorded and provided in color.

The video of the side view, the first frontal view, and the face view was captured at 60 frames per second, and each frame has a resolution of 640x480 pixels. Then, for the video of the second frontal view, it was captured at 30 frames per second with a resolution of 1600x1200 pixels per frame. The second frontal view was purposely made to facilitate a more detail display of the hands, compared to the 640x480 views.

The dataset were stored in a lossless compression format and it can be read using a C++ code provided on the official website. A compressed QuickTime version of video sequences was also created, with purpose to enable the viewers to quickly browse the video.



Main New Gloss	Consultant	Combined	Session	Scene	Start	End
=====	=====	=====	=====	=====	=====	=====
TWENTY						
	Liz		ASL_2008_01_11	2	2635	2661
	Tyler	MOV	ASL_2008_05_12a	1	2400	2480
	Naomi		ASL_2008_08_04	1	2279	2353
	Brady		ASL_2011_06_08_Brady	5	5170	5198
ALONE	=====	=====	=====	=====	=====	=====
	Liz		ASL_2008_02_01	35	3707	3782
	Liz		ASL_2008_02_29	32	2890	2936
	Tyler		ASL_2008_05_29b	9	3801	3830
	Naomi	MOV	ASL_2008_08_13	29	4065	4141
	Brady		ASL_2011_06_14_Brady	42	1430	1498
	Brady		ASL_2011_07_19_Brady	83	3186	3211
	Lana		ASL_2006_10_10	2	2076	2120
	Dana		ASL_2007_05_24	5	2431	2476
LONELY	=====	=====	=====	=====	=====	=====
	Liz		ASL_2008_02_01	52	751	805
	Tyler	MOV	ASL_2008_06_10	4	845	902
	Naomi		ASL_2008_08_13_session2	7	540	575
	Brady		ASL_2011_06_14_Brady	51	80	119
BACHELOR	=====	=====	=====	=====	=====	=====
	Liz	MOV	ASL_2008_02_29	32	2629	2675
	Brady		ASL_2011_07_19_Brady	83	2957	2992

Figure 2.2: Annotation of ASL Lexicon Video Dataset. Retrieved from the annotation file of [1].

An excel file showing the annotation was also provided. Since ASL

signs have no written form, so the class label for each sign were written in an approximated English, called a "gloss". A gloss could possibly assign to two different signs if they corresponded to the same ASL Lexical item.

A video sequence was stored multiple signs. For each sign in a video sequence, the annotation was written its start and end frames, the gloss, whether the sign is one-handed or two-handed, and a signer ID. The signer IDs were written to allow researchers to do experiments for user-dependent and user-independent sign recognition.

## 2.4 OpenPose

OpenPose [2] is an open library to detect multi-person pose estimation. It was claimed as the first bottom-up representation of association scores using a set of 2D vector fields that encode the location and orientation of limbs shown on the image domain, named Part Affinity Fields.



Figure 2.3: Left: Part Affinity Fields corresponding to the limb that connected elbow and wrist. Right: a 2D vector in each pixel of every PAF which encoded the position and orientation of the limbs [2].

This research had some improvements compared to its previous version [13]. From this newer version of work, Cao et. al proved that to achieve a maximum accuracy, the most important part in the work is to

refine the Part Affinity Fields, while ignoring the body part prediction refinement. The first body and foot keypoint detector combination was also proposed in this research, created based on foot dataset annotation that will be publicly released. The OpenPose was claimed as the first open source library which able to detect 2D multi-person pose, including the body, foot, hand, and facial keypoints, in a realtime manner.

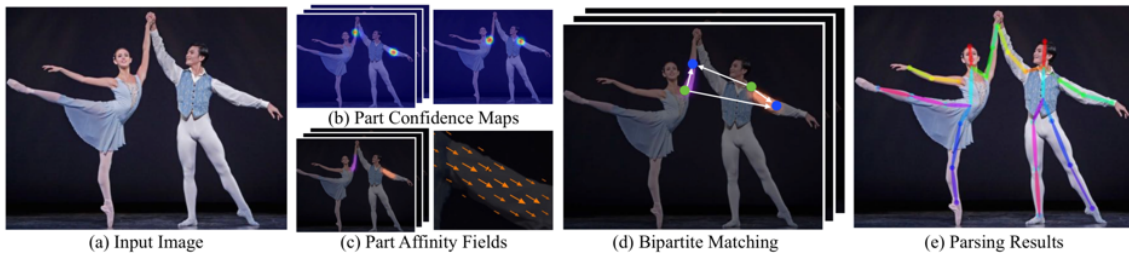


Figure 2.4: Overall system flow. (a) Entire image is taken as the input for a CNN to predict the joint (b) confidence maps for body part is detected (c) detection of PAFs (d) bipartite matching to associate body part candidates (e) assembly of PAFs into full body poses for all detected people in the image [2].

The OpenPose overall system can be seen at Fig. 2.4. The input is a color image (Fig. 2.4a) and the output is the 2D locations of the anatomical keypoints for each person in the image (Fig. 2.4e). The input image, firstly is forwarded to a feedforward network to predict a set of 2D confidence maps of body part locations (Fig. 2.4b), and a set of 2D vector PAFs (Fig. 2.4c), which encode the association degree between body parts. Lastly, greedy inference (Fig. 2.4d) will parse the confidence maps and the affinity fields to output the 2D keypoints for all people in the input image.

Compared to the existing 2D body pose estimation libraries, such as Mask R-CNN or Alpha-Pose, OpenPose has many superiorities. The existing libraries require the users to implement most of their pipeline, pro-

vide their own frame reader, a media to display the results, and provide the file generation for storing the results in JSON or XML, etc. The existing facial and body keypoint detectors are also not combined, so different libraries are required for each purpose. Different from the mentioned works, OpenPose can overcome all of the problems. It is able to be run on many different platforms, such as, Ubuntu, Windows, Mac OSX, and embedded systems. Different hardware, such as CUDA GPUs, OpenCL GPUs, and CPU only devices are also supported by this framework. The input types, images, video, webcam, and IP camera streaming are also served to be chosen by the user. The users can also select to display the results or save them on the disk. They can also choose which part of the human body they want to detect. Pixel coordinate normalization, number of GPUs to be used and frames skipping options are also the superiorities of this library.

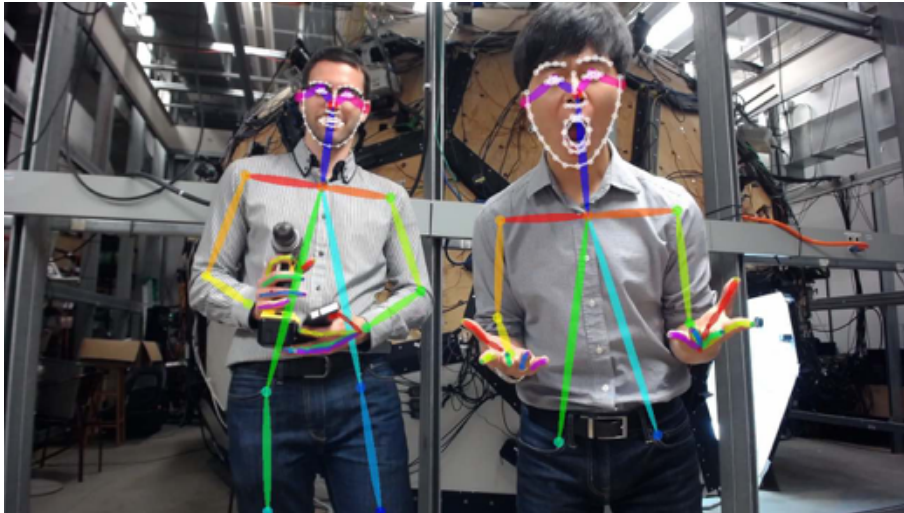


Figure 2.5: Output of OpenPose, body, foot, hand, and facial keypoints are detected in a real-time manner. [2].



# References

- [1] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, , and A. Thangali, “The american sign language lexicon video dataset,” in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, June 2008.
- [2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields,” in *arXiv preprint arXiv:1812.08008*, 2018.
- [3] W. H. Organization, “Deafness and hearing loss,” 03 2019.
- [4] L.-M. D. Sandler Wendy, *Sign Language and Linguistic Universals*. 2006.
- [5] U. D. o. L. Bureau of Labor Statistics, “Occupational outlook handbook: Interpreters and translators,” 04 2019. [Online; accessed 01-July-2019].
- [6] M. Jay, “History of american sign language,” 10 2010. [Online; accessed 09-July-2019].
- [7] Wikipedia contributors, “American sign language.” [https://en.wikipedia.org/w/index.php?title=American\\_Sign\\_Language&oldid=904706391](https://en.wikipedia.org/w/index.php?title=American_Sign_Language&oldid=904706391), 2019. [Online; accessed 9-July-2019].
- [8] DawnSignPress, “History of american sign language,” 08 2016. [Online; accessed 09-July-2019].
- [9] B. Bauer and K.-F. Kraiss, “Towards an automatic sign language recognition system using subunits,” in *Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction, GW ’01*, (London, UK, UK), pp. 64–75, Springer-Verlag, 2002.
- [10] Jiangwen Deng and H. T. Tsui, “A pca/mda scheme for hand posture recognition,” in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pp. 294–299, May 2002.
- [11] M. G. B. R. A. Tennant, *The American Sign Language Handshape Dictionary*. Gallaudet University Press, 2010.
- [12] C. Valli, *The Gallaudet Dictionary of American Sign Language*. Gallaudet University Press, 2006.
- [13] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017.
- [14] A. Cui, M. Costello, and S. J. Stolfo, “When firmware modifications attack: A case study of embedded exploitation.,” in *NDSS* [14].