DETECTION OF SIGN LANGUAGE IN PICTURE-IN-PICTURE VIDEO

An Undergraduate Research Scholars Thesis

by

MAHAK MITHANI

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Frank Shipman

May 2017

Major: Computer Science

TABLE OF CONTENTS

ABSTRACT

Detection of Sign Language in Picture-in-Picture Video

Mahak Mithani
Department of Computer Science
Texas A&M University

Research Advisor: Dr. Frank Shipman
Department of Computer Science
Texas A&M University

The internet enables almost anyone to locate content on almost any topic. This ability, however, is not easily available for those who sign. In order to provide resources to those whose primary language is sign language, a digital library, called SLaDL, has been created. In order to ensure maximum efficiency of the video-processor that detects sign language, it is important to check that the program works on all video resolutions. Picture-in-picture videos pose a challenge, as they contain fewer pixels and possess different characteristics than standard webcam sign language videos. However, these videos are very important to test as they are less likely to be retrieved otherwise through tags or other metadata. This project aims to detect and identify sign language in picture-in-picture videos through polar motion profiles, working to expand the corpus of videos on which the processor is successful.

KEYWORDS

Picture-in-picture video	One video is displayed on main screen, another is displayed in inset window
SLaDL	Sign Language Digital Library
Video analysis	Efforts to detect sign language content
Video sharing sites	E.g., YouTube

CHAPTER I

INTRODUCTION

Those who are deaf or hard of hearing rely primarily on sign language to communicate [4]. A visual form of communication, sign language consists of hand gestures, facial expressions, and bodily postures. Video sharing websites are very useful for the deaf community to exchange information with each other. Though the number of sign language videos uploaded to the internet is increasing rapidly, it is difficult for the community to find the videos relevant to them. This is because query results are heavily dependent on the presence and accuracy of metadata for both type of sign language and topic discussed in the video. Thus, it becomes imperative that we create algorithms to detect and identify sign languages in order to automatically tag these videos based on the form of communication they utilize [3].

I will be utilizing a technique developed by Dr. Frank Shipman's lab that relies on face detection, background modeling, and polar representation of hand movements. Polar motion profiles capture the signing activity in the frames of each video [2]. In order to properly evaluate the polar motion profiles, we must test the technique on a corpus of prerecorded videos. Many videos on the internet include a picture-in-picture sign language translation. These videos have fewer pixels than normal and are less likely to be retrieved by tags or other metadata. Such videos must also be automatically tagged as sign-language video by the algorithm. This project will test the developed algorithms and techniques on lower resolution videos, allowing the programmers to develop the most effective and efficient sign language recognition techniques [2].

Objectives

In my research, I will evaluate whether or not the current classifier can be applied to picture-in-picture videos and other cases of lower resolution. If the technique works on these videos, then the algorithm is effective; if not, then the algorithm must be reformed in order to properly handle picture-in-picture videos. Even if effective, it may become clear that the current approach can be improved in either accuracy or resource usage for this special class of videos. I will learn more about sign language recognition techniques through my research while broadening the corpus of videos used on the classifier.

Methodology

In order to conduct my research, I will gather from the internet a set of videos that employ sign language in a picture-in-picture format (Figure 1) and another set where picture-in-picture is used for other purposes (Figure 2). The challenge will be that such videos are not tagged as "picture-in-picture," therefore a method for locating such composite video arrangements must be developed. Then, I will test the existing Sign Language recognition algorithm on these videos to evaluate whether or not the technique is effective on lower resolution videos. If the polar motion profiles successfully recognize sign language in the videos, then we can deem the algorithm successful. If not, we will have to reevaluate the technique to optimize it for the smaller videos. I will be working with Caio Duarte Diniz Monteiro to expand the current corpus of videos used on the classifier.

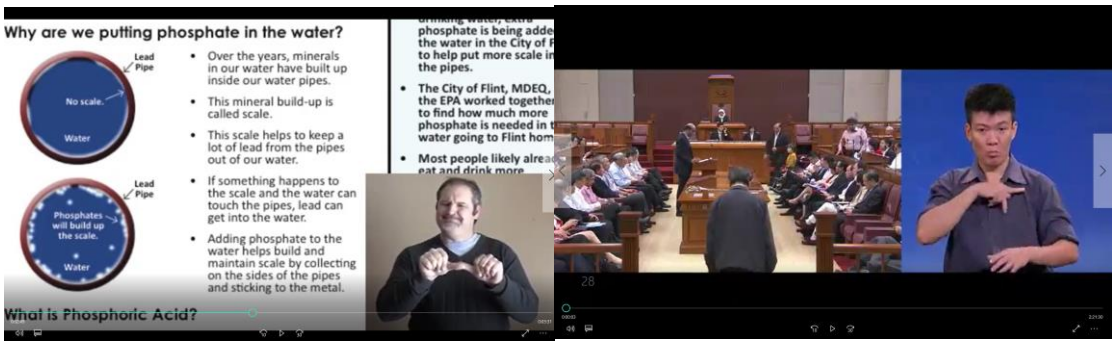

Figure 1: Picture-in-Picture Employing Sign Language

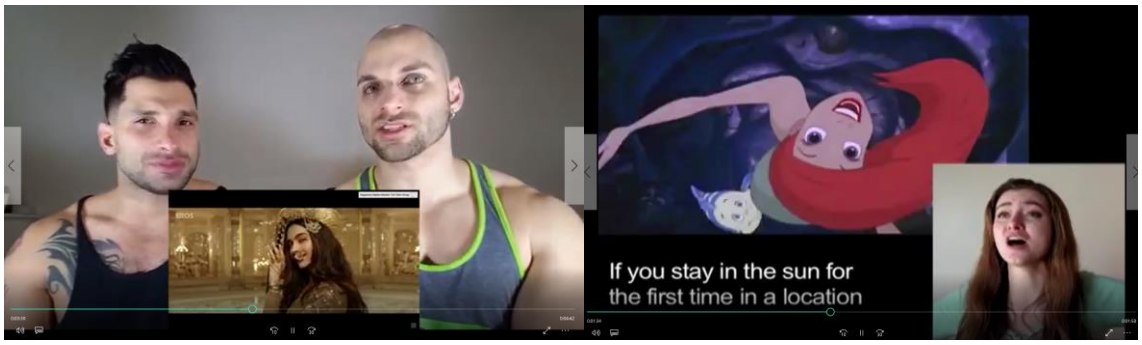

Figure 2: Picture-in-Picture not Employing Sign Language

Research Compliance

I collected data by analyzing digital video recordings from internet databases such as YouTube. I have completed the IRB Social and Behavioral Research Investigators and Key Personnel course. I do not need approval from the IRB to collect data, as I am making use of publicly available videos to test algorithms for locating and classifying sign language in video.

CHAPTER II

BACKGROUND

The target videos for the Sign Language Digital Library include those in which most of the content can be understood through sign language [2]. It is important to note that videos that incidentally contain a brief section of American Sign Language are not to be included in the library. Many internet users utilize the web to access information about specific topics. However, there did not seem to be a database online for purely ASL content. After noting the void, it was discovered by a group of researchers in the Computer Science department at Texas A&M University that specialized tools would have to be developed in order to allow users to easily and efficiently access sign language content.

The most arduous task at hand when recognizing sign language based videos is transcription, which involves recognizing the specific signs in the video [1]. Since the SLaDL does not attempt to translate the signs, but merely detect them, populating the corpus became much simpler. The video is first processed to locate regions of interest using face detection. A background model is simultaneously created. At this stage, the foreground objects and regions of interest have been identified for each frame. A polar motion profile is then extracted for each region of interest which "represents the probability of foreground objects at each polar coordinate [1]." Each video is assigned an average polar motion profile that is computed using the frames and regions of interest. A support vector machine classifier uses this average to determine whether sign language content is present in the video. The task at hand was to determine whether this classification method runs successfully on picture-in-picture videos that contain sign language, testing the algorithm on videos with fewer pixels.

CHAPTER III

METHODS

The algorithm uses a Haar-cascade recognizer to effectively detect faces [5]. The cascade will return a list of rectangles which represent bounds for potential face locations. To determine whether a face actually exists in the location, the algorithm checks whether three or more rectangles from overlap at each potential location. This removes any false positives that may be returned.

If and when a face has been detected, the next step is to subtract the background in order to extract foreground objects within a region of interest. The color distribution is then modeled for each pixel in the video. Since each pixel may have different statistics across the video, a separate probability density function is used per pixel. This helps to account for the dynamism of non-stationary backgrounds. An adaptive Gaussian Mixture Model is used to build a background model for each pixel [6]. Finally, morphological erosion and dilation are used to remove small objects in the foreground.

Lastly, the results of the face detection and background modeling algorithms are combined to extract the range of hand motions around the face. For each frame, a region of interest is defined, which encompasses the moving objects around each face. Once defined, a polar motion profile is generated per frame, to measure the signing activity. For each video, the average polar motion profile across regions of interest is calculated, thus determining whether sign language is being used in the video [1]. This entire process is more challenging with picture-in-picture videos as the number of pixels is drastically lower than that of a full-sized video.

CHAPTER IV

RESULTS

The sign language detection method was performed on two separate classes within a dataset. The first class consisted of videos that utilized sign language in an inset video (picture-in-picture). The second class also utilized picture-in-picture, but no sign language. This was done in order to maintain consistency among the dataset.

Table 1. Picture-in-Picture Results

Training Set Size	Precision	Recall	F1 Score
5	0.6989	0.7311	0.7034
10	0.7426	0.7404	0.7340
15	0.7572	0.7446	0.7463
20	0.7665	0.7341	0.7450
25	0.7768	0.7380	0.7509

There are three parts to our results: precision, recall, and F1 score. Precision is the ratio of true positives over selected elements ($\frac{true\ positives}{true\ positives+false\ positives}$). Recall is the ratio of true positives over relevant elements ($\frac{true\ positives}{true\ positives+false\ negatives}$). F1 score is the harmonic mean of precision and recall ($2 * \frac{precision*recall}{precision+recall}$). The results can be seen in Table 1.

Our conclusive F1 score for picture-in-picture videos containing sign language was approximately 75%. In non-picture-in-picture videos containing sign language, the algorithm returns an F1 score of approximately 78%, therefore we can conclude that the quality of the detection algorithm does not significantly decrease when used on picture-in-picture videos [1].

CHAPTER V

DISCUSSION

The approach presented to detect sign language in videos which generates polar motion profiles conclusively works on picture-in-picture videos. This method of utilizing face detectors to identify regions of interest produces similar results when performed on standard videos containing sign language as well as when sign language is present in an inset video.

After performing the algorithm on picture-in-picture videos, it has been confirmed that the sign language detection method is scalar invariant. This means that the algorithm will work on videos of different scales or resolutions, so long as there are enough pixels to identify faces. This threshold is not unsurpassed in picture-in-picture videos.

One of the concerns raised when analyzing the results of this detection method is the averaging of faces. If the algorithm detects five faces, four of which are speaking orally and one of which is using sign language (be it in an inset video or otherwise), the algorithm will classify the video as non-signing, since the average activity points towards oral communication. This is detrimental to the classification process, as it will mislabel videos that contain sign language as non-signing videos. Further work would have to revise the method in order to disregard face averaging.

REFERENCES

[1] V. Karappa, C. Monteiro, F. Shipman, and R. Gutierrez-Osuna, “Detection of sign-language content in video through polar motion profiles”, *Proc. ICASSP*, 2014, pp. 1299-1303.

[2] F. Shipman, R. Gutierrez-Osuna, T. Shipman, C. D. D. Monteiro, and V. Karappa, "Towards a Distributed Digital Library for Sign Language Content,".

[3] F. Shipman, R. Gutierrez-Osuna, and C. Monteiro, “Identifying sign language in video sharing sites”, *ACM Trans. On Accessible Computing*, 2014, 9:1-9:14.

[4] NIH, "American Sign Language," *NIH Publication No. 11- 4756*, June 2011.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," presented at the Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2001.

[6] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," presented at the Proc. 17th Intl. Conf. on Pattern Recognition (ICPR), 2004.

[7] Mengren Qian, Luntian Mou, Jia Li, and Yonghong Tian, “Video Picture-in-Picture Detection using Spatio-Temporal Slicing,” Proc. ICME’2014 Workshop on Emerg. Multimedia Sys. and Appl., Chengdu, China, 2014.

[8] S. Purushotham, Q. Tian, and C.-C. J. Kuo. Picture-in-picture copy detection using spatial coding techniques. In Proc. of the ACM AIEMPro Workshop, 2011.