Speed-Accuracy Tradeoffs for Detecting Sign Language Content in Video Sharing Sites

Frank M. Shipman, Satyakiran Duggina, Caio D.D. Monteiro, Ricardo Gutierrez-Osuna
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843-3112
1-979-862-3216
shipman@cse.tamu.edu

ABSTRACT
Sign language is the primary medium of communication for many people who are deaf or hard of hearing. Members of this community access online sign language (SL) content posted on video sharing sites to stay informed. Unfortunately, locating SL videos can be difficult since the text-based search on video sharing sites is based on metadata rather than on the video content. Low cost or real-time video classification techniques would be invaluable for improving access to this content. Our prior work developed a technique to identify SL content based on video features alone but is computationally expensive. Here we describe and evaluate three optimization strategies that have the potential to reduce the computation time without overly impacting precision and recall. Two optimizations reduce the cost of face-detection, whereas the third focuses on analyzing shorter segments of the video. Our results identify a combination of these techniques that yields a 96% reduction in computation time while losing only 1% in F1 score. To further reduce computation, we additionally explore a keyframe-based approach that achieves comparable recall but lower precision than the above techniques, making it appropriate as an early filter in a staged classifier.

CCS Concepts
• **Information systems** → Multimedia and multimodal retrieval
• **Social and professional topics** → Assistive technologies

Keywords
Sign language detection; signal processing; computer vision.

1. INTRODUCTION
Many applications rely on identifying moments where communication takes place in a recording. For spoken languages, this involves voice detection in an audio signal. For sign languages (SLs), this involves detecting sign language in a video signal. This paper describes and evaluates optimizations for SL detection algorithms aimed at reducing computation without severely impacting accuracy.

Sign language is the medium of communication for many people who are deaf or hard of hearing. With the rising popularity of video sharing sites like YouTube and Vimeo, the volume of SL content available is steadily growing. The SL community often shares these videos through email or posts in social media, passing direct links to the content. However, when an information consumer, rather than an information provider, wants to locate SL content on a particular topic, they must rely on the existence of metadata that identifies both the topic and language used in the video. Unfortunately, studies have shown that metadata is frequently applied inconsistently [5]. Studies of metadata-based access to SL videos on particular topics have found precision rates of 43% [8]. As a result, many in the SL community do not search for SL content. Automatic SL detection would help this situation.

We have been exploring how to identify SL video in video sharing sites. Our earliest work explored the relative value of a set of video features [6] on detecting SL content, but was limited to videos containing a single person facing the camera. In later work [4], we relaxed the constraints to enable detection in videos including multiple visible people and improved the recall but with considerable computational cost. In this later approach, likely signing activity in each frame of a video is modeled using polar coordinates (angle and distance) in a polar motion profile (PMP) which relies on background modeling and subtraction [10] and face detection using Haar features [9]. These combine to create a computationally intensive process. Reduction in computation time is needed to scale the solution to the quantity of videos uploaded to video sharing sites. Once a fast approach to detecting sign language in videos is available, more computationally expensive techniques can be applied to detect which sign language it is.

In this work, we propose techniques to reduce the amount of computation needed to generate PMPs for a video. We evaluated the impact of using alternate face detectors, varying the length of the video segments analyzed, and detecting faces for sampled video frames on the precision and recall of the classifier. These results lead to a recommended combination of these optimizations, which is analyzed in terms of both computation and accuracy. Finally, we explore the potential for computing PMPs from sampled frames without generating a background model – an approach that further reduces computation but comes with a greater reduction in accuracy.

2. RELATED WORK
Sign language involves hand gestures, facial expressions and postures of the body to communicate. A significant amount of research has aimed at transcribing American Sign Language into written words. Such a capability would be useful for those not in the SL community to understand the videos in SL and would also

ASSETS'17, Oct. 29–Nov. 1, 2017, Baltimore, MD, USA.

DOI: http://dx.doi.org/10.1145/3132525.3132559