Towards a Distributed Digital Library for Sign Language Content

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

ABSTRACT
The Internet provides access to content in almost all languages through a combination of crawling, indexing, and ranking capabilities. The ability to locate content on almost any topic has become expected for most users. But it is not the case for those whose primary language is a sign language. Members of this community communicate via the Internet, but they pass around links to videos via email and social media. In this paper, we describe the need for, the architecture of, and initial software components of a distributed digital library of sign language content, called SLaDL. Our initial efforts have been to develop a model of collection development that enables community involvement without assuming it. This goal necessitated the development of video processing techniques that automatically detect sign language content in video.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, systems issues, user issues.*

General Terms
Algorithms, Design, Experimentation, Human Factors.

Keywords
Sign language digital library; sign language detection; digital library architecture; distributed digital library collection.

1. INTRODUCTION
General purpose and specialized digital library search engines allow Internet users to search for content on most any topic. In many cases, users can select among search engines and libraries to match their primary language. But languages that do not have a written form, including sign languages, are not well supported by the existing infrastructure.

Sign languages evolved in part independently of written language and thus do not have a one-to-one mapping to any written form [14]. For example, American Sign Language (ASL) is independent from British Sign Language (BSL) and both have a different grammar and syntax than written/spoken English. Sign languages are the primary language for many in the Deaf community, particularly for those who become deaf early in life. Those who grow up with sign language as a primary language learn English (or another written language) as a second language. A study of deaf and hard-of-hearing 17-18 year olds shows that half of the population had a lower than beginning of fourth grade reading level [5]. Thus, for a large portion of the sign language community, access to content in a written language is not a substitute for access to sign language content.

The ease of recording and sharing videos has resulted in a large quantity of sign language content being available on video sharing sites, such as YouTube. But this content is most often accessed via ad-hoc mechanisms, such as people sharing URLs to videos via email and social media. Searching for content is limited to the capabilities provided by the video sharing sites, which relies on the quality of metadata and tags. As a result, locating sign language content on a particular topic is not easy.

The next section of this paper illustrates the difficulty of locating sign language videos on particular topics. We then discuss related work and present an architecture/framework for building SLaDL, a distributed digital library for sign language content. Next, we describe the initial design and evaluation of components of the architecture. We conclude with a summary of our status and a plan of work yet to be done.

2. QUANTIFYING THE PROBLEM
A more precise understanding of sign language videos (SL videos) is valuable before discussing their location. We consider a SL video one where most content in the video can be understood through sign language. These are the videos that are of value to the sign language community. Videos of one or more signers presenting to the camera or having conversations, and videos with a sign language interpreter in picture-in-picture are the most common forms. Videos that include sign language incidentally, such as a spoken-language news report that includes a brief bit of ASL, are not useful to the community and thus not our target.

Accessing content on a particular topic is a common activity on the Internet. From personal experience (the third author's primary language is ASL) we know that there is considerable sign language content passed around from person to person but there is no "go to" place for finding ASL content on particular topics. Indeed, most do not even try to locate such content. But why?

To help answer this question we previously conducted a study to quantify how hard is it to locate content in ASL for particular topics on YouTube [13]. The topics were chosen as the top 10 news queries for 2011 from Yahoo! [12]. To locate content on

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.
 ISBN 978-1-4503-3594-2/15/06…$15.00
DOI: http://dx.doi.org/10.1145/2756406.2756945

these topics in ASL, we generated queries that included the string phrase for the topic, e.g. "Arizona shooting" or "Arab spring", and added the query terms ASL, "sign language", or both.

Table 1. Number and percent of results in/not in sign language and on/off topic when ASL and "sign language" are included to locate content on top ten 2011 news queries on Yahoo!

	In SL	Not in SL	Total
On Topic	50 (46%)	27 (25%)	77 (70%)
Not on Topic	24 (22%)	9 (8%)	33 (30%)
Total	74 (67%)	36 (33%)	110 (100%)

The top 20 results were coded by hand to determine if they were on topic *and* in ASL. The best performance (precision=46%) was achieved when both terms (ASL and "sign language") were added to the query; see Table 1. The remaining videos returned were nearly evenly divided among videos not in sign language and videos not on topic; in total, 33% of the videos returned were not in sign language. When either term (but not both) was included, the proportion of videos not in sign language increased to 46%.

A larger, follow-up study [6] on 100 more varied and less time-limited topics (e.g. topics were the top 10 queries in 10 different categories, including topics like food, health, etc. instead of just news topics) found a lower precision (13% instead of 46%) but only 21% not in sign language; see Table 2. Though differences in topics and the availability of content in those topics are largely responsible for the difference in results, it is also possible that changes to the YouTube ranking algorithm played a role.

Table 2. Number and percent of results in/not in sign language and on/off topic when both terms (ASL and "sign language") are included in the query along with one of 100 topics.

	In SL	Not in SL	Total
On Topic	203 (13%)	38 (3%)	241 (16%)
Not on Topic	998 (66%)	279 (18%)	1277 (84%)
Total	1201 (79%)	317 (21%)	1518 (100%)

Why is the precision so low? Examination of the videos uncovers limitations of using text-based queries for locating SL videos. In particular, query terms are inherently ambiguous so the phrase "sign language" returns videos (1) in sign language, (2) about sign language, and (3) about the language used in signs, e.g., protest signs. An additional issue is the quality of the metadata used to match the query terms; text-based metadata can be idiosyncratic in video sharing and other social media sites [4][10].

The above results indicate that supporting access to sign language content from popular video sharing sites requires the development of specialized tools. Our work to date [11][7][13][6] has focused on developing video processing techniques for that purpose.

3. ARCHITECTURE

Creating a sign language digital library shares technical challenges with prior work on video digital libraries. Our focus is not on storage infrastructure and playback interfaces (capabilities that are already provided by video sharing sites) but on extracting metadata related to the language(s) used in the video. Thus, our research is more closely related to prior work on web services that support the location of video content. Among these it is worth noting TalkMiner [1], a webcast search engine that provides an alternative interface for accessing lecture videos hosted on other sites. TalkMiner includes a component for downloading video lectures from other sites, a component to generate new metadata and segment the videos, and an interface/portal with unique visualizations where users can go to locate video lectures and segments.

A challenge for SLaDL is that SL videos are hard to locate since most are uploaded with minimal metadata – as noted, the current practice among members of the SL community is to share the URL via email or social media. Thus, SLaDL needs a robust approach for locating videos to be included in the corpus. Figure 1 shows the current SLaDL architecture, which acknowledges this collection-building challenge.

In this model, content is added to SLaDL via two paths, one via human/community classification and the other via automatic classification. Both paths begin with the identification of a set of potential SL videos. These are identified by the crawler, either through the addition of query terms to the user's query request, resulting in queries much like described in Section 2, or by relationships to known SL videos (e.g. being in same YouTube channel, posted by same user, in same list of videos, etc.)

Community involvement in the creation of digital library collections is a common practice [8][3]. This path involves members of the community (1) providing the locations of SL videos (e.g. ones they have uploaded to video sharing sites), (2) identifying videos where there is evidence that they may include sign language content, and (3) voting them in/out of the library. Such a path involves a number of social issues that are not our current focus, and will not be discussed further.

The second path to building the collection is through automatic classification. This path involves a multistage process whereby potential SL videos are crawled from video sharing sites. The crawler currently uses hand-authored queries to the YouTube query interface, though future versions may exploit information such as YouTube channels, video lists co-occurrence, co-viewing statistics, etc. The important aspect of the crawler is that it creates a manageable set of videos for further processing.

Once there is a local cache of potential SL videos, a specialized classifier is used to detect sign language in these videos. This

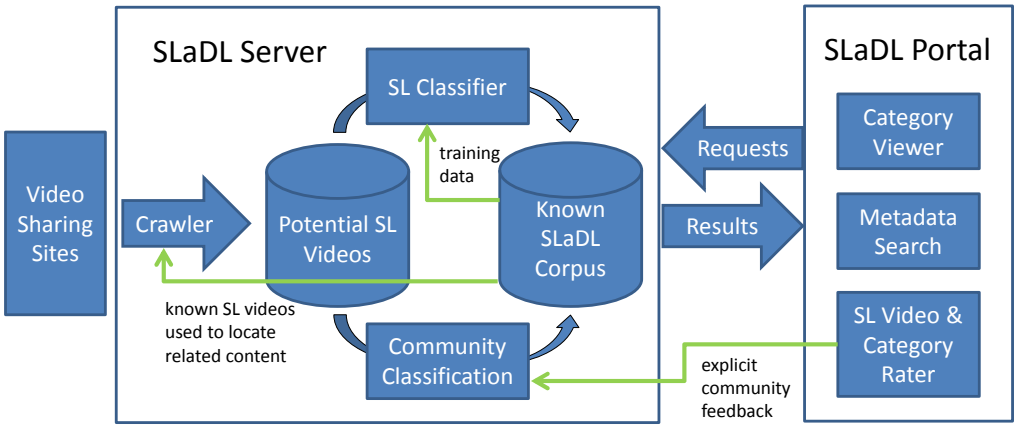

Figure 1. Architecture of the Sign Language Digital Library (SLaDL).

classification problem has been the focus of much of our activity to date and is discussed in depth in the next section.

The architecture includes feedback paths to improve the crawler and sign language classification techniques. As the known SLaDL corpus grows, the set of videos that are brought into the cache of potential SL videos will change (due to their relations to known SL videos) and the SL classifiers, which rely on machine learning techniques, will be retrained with the expanded corpus.

The SLaDL portal (see Figure 2) provides an interface for the sign language community to access and modify the collection. At present, the portal includes a category viewer to browse the collection via multiple perspectives, and a view by topic, much like the categories found in news aggregators (e.g. Google News). In coming work, metadata filters are being added to tailor views to individual users. For example, items in the collection may be filtered based on the sign language used in the video. While we cannot automatically assign the particular sign language, such metadata can be added by members of the community, or can be predicted from user comments and metadata.

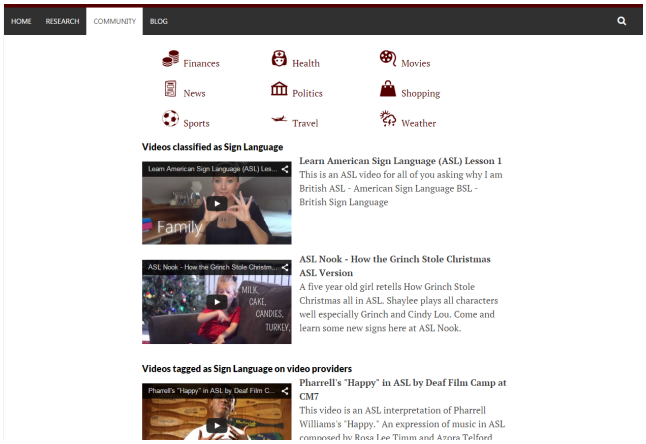

Figure 2. The SLaDL portal enables collection browsing, search, and community feedback on presented videos.

Our architecture also allows a path for mitigating the issues concerned with the initial lack of content on many topics. When users search for content on a topic, two sets of results are generated. The first set is the result of the query being compared to the videos within the known SLaDL corpus. The second set of results is generated from the list of potential SL videos (see Figure 1); these results are provided separately. In addition, the user's query may cause new requests to go to the video sharing services that add new videos to the set of potential SL videos. This happens when there are too few matching results in the potential SL video cache. In this way, expression of interest from the community helps drive the content being examined for inclusion in SLaDL by automated and/or human means.

Overall, the SLaDL architecture addresses the difficulty of locating SL videos. As with prior distributed video digital libraries, we do not store video content but only information about categories and topics that have been identified by the community and our algorithms.

4. DETECTING SIGN LANGUAGE VIDEO

Sign language is a fairly obvious activity to those around signers, since the combination of temporal structure, trajectories, velocities, hand shapes, and facial cues is hard to mistake. This observation led us to develop algorithms for the automatic detection of sign language as one path to create the SLaDL collection. While there has been prior work in identifying pauses in signing to improve bandwidth utilization [2], no prior work exists on the problem of distinguishing between videos containing sign language vs. other human gestures. The problem shares similarities with prior research on activity detection in video, in that developing a sign language detector involves selecting video or metadata features as inputs for a pattern classifier – a support vector machine (SVM) in our case.

Our video features are motivated by the fact that sign language generally happens in a region from slightly above the head (top) to the middle of the torso (bottom) and in front of or slightly to the side of the body (left/right). Our first approach [11] to characterizing activity in this region was based on this observation. The approach is illustrated in Figure 3. First, a face detector from openFrameworks [9] is used to locate people. If no faces are detected, it is unlikely that the video will have sign language content, and the process is terminated. Once at least one face is detected (Fig 3a), a dynamic background model is generated (Fig. 3c) and then used to perform foreground-background segmentation (Fig. 3d). Next, a morphological filter for the resulting content is used to remove noise so that only the larger moving objects remain (Fig. 3b). In a final step, five video features are defined from the identified foreground motion: 1) total activity in video, 2) spread of activity across video frame, 3) speed of motion, 4) symmetry of motion relative to middle of the face, and 5) amount of non-facial movement. Details on these five features and the resulting *five feature classifier (5FC)* were reported elsewhere [11].

Figure 3. Identifying activity via (a) face detection in the video frame, (b) the final foreground image, (c) the computed background model and (d) the intermediate foreground image

The 5FC classifier was evaluated on a corpus containing 78 ASL, 20 BSL, and 94 likely false positive videos (e.g. elaborately gesturing reporters) with a single signer/speaker. Overall discrimination performance between SL and non-SL videos was 82% precision and 86% recall. Given the difficulty of the corpus (in practice, most YouTube videos are not of weathermen, mimes, etc.) the performance on a more representative set of videos would almost certainly be higher still. Further evaluation of the individual features showed that symmetry of motion was the most discriminatory feature. Examination of false negatives showed that many errors originate during face detection (e.g., the detector loses track of the face) rather than at the classification stage.

These initial results led us to develop a second sign language classifier. In this new approach, the face recognition module was improved by running multiple face detectors in parallel, each of which could recognize multiple faces in the video frame. A majority voting scheme was then used to determine which ones among all candidate faces would be included in the motion analysis stage. Because symmetry of motion had proven to be so valuable in the first evaluation, a *polar-motion profile (PMP)* was developed to characterize the distance and polar orientation of the motion from the center of the face [7]. Figure 4a shows the regions of interest generated for three signers, and the PMP (the probabilities for finding a foreground pixel at particular distance or orientation) for the rightmost signer. The peaks in the graphs indicate the position of the hands relative to the face.

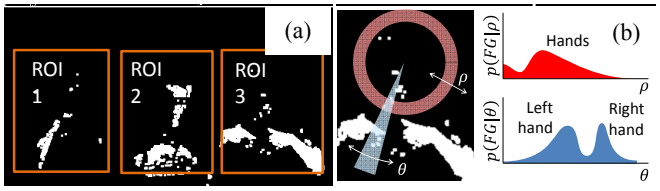

Figure 4. (a) Region of interests for the three signers in the frame. (b) PMPs for a video frame represent the probability of finding a foreground pixel.

To evaluate this second sign language classifier, a new dataset was collected that included videos with multiple signers, and also more complex backgrounds. This data set was selected based on an examination of the top 105 results from YouTube for the query "American Sign Language" and video recommendations from YouTube for these 105 results.

When applied to the original corpus, results on the PMP classifier were similar to those of the original 5FC classifier. With the second corpus, the precision of the classifiers were similar (81-82%) but the recall of the PMP classifier (94%) was significantly higher than that of the 5FC classifier (60%). This result is due to the PMP classifier's richer representation of motion, more robust face tracking, and the ability to track multiple signers.

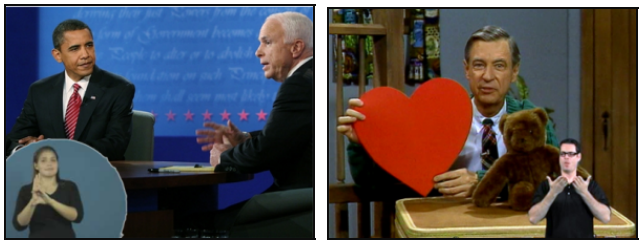

Figure 5. Frames from videos with sign language captions.

In short, we have developed robust techniques to detect sign language in videos. But there are still classes of sign language content yet to be explored. We are currently working towards using the existing techniques to detect sign language captions, examples of which are seen in Figure 5, and segmentation of video based on the existence of sign language content.

5. DISCUSSION

Our development of a sign language library started with the goal of enabling topic-based access to the sign language content found in video sharing sites. We quantified the challenges of locating sign language content via text-based queries. Towards that goal of creating a sign language digital library, we have fashioned architecture based on the unique challenges of identifying sign language content and have developed and tested components of this architecture. A distinguishing feature of SLaDL is generating a collection through a combination of automatic classification and community feedback. While deployment and engagement are required to assess the community-feedback path, initial results from the video classifiers supports our plan to initially populate SLaDL automatically.

Topic-based access to content is very different from the term-based access provided by search engines. While there is considerable research into sign language translation, the current state of those efforts cannot be applied to the vocabulary, signing speed, context (e.g. moving backgrounds), or recording quality commonly found in shared sign language video. As a result, SLaDL focuses its video analysis efforts on detecting SL content and uses metadata analysis and community feedback when assessing finer topic-oriented distinctions.

6. REFERENCES

[1] J. Adcock, M. Cooper, L. Denoue, H. Pirsiavash, and L.A. Rowe, "TalkMiner: A lecture webcast search engine", *Proc. MM*, 2010, pp. 241-250.

[2] N. Cherniavsky, R.E. Ladner, and E.A. Riskin, "Activity detection in conversational sign language video for mobile telecommunication", *Proc. FG '08,* 2008, pp.1-6.

[3] E. Fox, et al. "Ensemble PDP-8: eight principles for distributed portals". *Proc. JCDL*, 2010, pp. 341-344.

[4] M. Heckner, T. Neubauer, and C. Wolff, "Tree, funny, to_read, google: What are tags supposed to achieve?", *Proc. of Workshop on Search in Social Media*, 2008, 3-10.

[5] J. Holt, C. Traxler, and T. Allen, *Interpreting the Scores: A User's Guide to the 9th Edition Stanford Achievement Test for Educators of Deaf and Hard-of-Hearing Students*. Gallaudet Research Institute Technical Report 97-1. Washington, DC: Gallaudet University, 1997.

[6] V. Karappa, *Detection of Sign-Language Content in Video through Polar Motion Profiles*, unpublished MS Thesis, Texas A&M University, 2014.

[7] 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.

[8] M. Khoo, "Community design of DLESE's collections review policy: a technological frames analysis", *Proc. JCDL*, 2001, pp. 157-164.

[9] Z. Lieberman and T. Watson. openFrameworks. http://www.openframeworks.cc/ Accessed September 2011

[10] C. Marshall, "No Bull, No Spin: A comparison of tags with other forms of user metadata", *Proc. JCDL*, 2009, 241-250.

[11] C. Monteiro, R. Gutierrez-Osuna, F. Shipman. "Design and evaluation of classifier for identifying sign language videos in video sharing sites". *Proc. ASSETS*, 2012, 191-198.

[12] E. Osmeloski. *2011 Yahoo! In review: Top US searches in 30 categories.* http://searchengineland.com/2011-yahoo-in-review-top-us-searches-in-30-categories-103215.

[13] 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.

[14] C. Valli and C. Lucas, *Linguistics of American Sign Language: An Introduction*, Gallaudet University Press, Washington D.C., 2000.