2016 IEEE International Symposium on Multimedia

Detecting and Identifying Sign Languages through Visual Features

Caio D. D. Monteiro, Christy Maria Mathew, Ricardo Gutierrez-Osuna, Frank Shipman
Department of Computer Science and Engineering
Texas A&M University
College Station, Texas, 77843-3112
{caioduarte.diniz, cmaria_91, rgutier, shipman }@tamu.edu

Abstract— **The popularity of video sharing sites has encouraged the creation and distribution of sign language (SL) content. Unfortunately, locating SL videos on a desired topic is not a straightforward task. Retrieval depends on the existence and correctness of metadata to indicate that the video contains SL. This problem gets worse when considering a particular type of sign language (e.g. American Sign Language - ASL, British Sign Language - BSL, French Sign Language – LSF, etc.), where metadata needs to be even more specific. To address this problem, we have expanded a previous SL classifier to distinguish videos in different SLs. The new classifier achieves an F1 score of 98% when discriminating between BSL and LSF videos with static backgrounds, and a 70% F1 score when distinguishing between ASL and BSL videos found on popular video sharing sites. Such accuracy with visual features alone is possible when comparing languages with one-handed and two-handed manual alphabets.**

Keywords-sign language detection; language identification.

I. INTRODUCTION

Sign language (SL) communicators rely on a combination of hand gestures, body posture, and facial expressions to convey meaning. SL is the primary medium of communication for many deaf or hard-of-hearing people [1]. Due to its visual form, video sharing websites are beneficial for storing and conveying SL commentaries and explanations. However, finding relevant SL content in such sites depends on accurate metadata, not just about the content but also about the language of expression [2].

Here we examine the problem of locating SL videos in video sharing sites. The numerous videos posted online every minute make manual tagging of the videos infeasible. Automatic detection techniques enable the development of SL digital libraries [3]. A pilot study by Monteiro et al. [4] showed that SL videos within video sharing sites can be heuristically identified using a SVM classifier provided with only video features (i.e., without metadata). Later work by Karappa et al. [5] relaxed some of the constraints on the videos in the prior work, improving its recall and applicability.

An accurate classification technique for detecting SL is a good first step. The shortcoming is that, like oral and written languages, there are many SLs in the world. For example, an expression in American Sign Language (ASL) is not understandable by British Sign Language (BSL) signers and vice-versa. Sign languages have developed in many countries and regions, each of these languages has their own vocabulary, syntax, and grammar, apart from the surrounding spoken and written language of the region [6]. Thus, the SL video detectors of Monteiro and Karappa can be viewed as the equivalent of voice detection for audio recordings. The next step is to identify the language in use.

Identification of the language in use in a SL video can make use of multiple types of features. For example, the language of the metadata (e.g. title, description, comments, etc.) is likely to give considerable insight. But, given that parts of the world that use the same written language use different SLs, additional methods are necessary. In this work, we present a method to classify different SLs based purely on visual features of the recorded videos. In particular, we explore the classification of BSL and LSF videos taken from the Dicta-Sign corpus [7] and ASL and BSL videos taken from video sharing sites.

II. MOTIVATION

Approximately 0.5% of the US population is functionally deaf using the definition "at best, can hear and understand words shouted in the better ear" [8]. On-line video sharing sites, such as YouTube and Vimeo, have provided members of the deaf and hard of hearing community a way to publish and access content in SL. Currently, there are about 218,000 results when searching for "American Sign Language" on YouTube and about 44,700 when searching for "British Sign Language". Search engines on these sites heavily rely on tags and metadata, making it is difficult to locate SL video on a particular language and topic, unless it is accurately tagged for both topic and language. In earlier work, we found that only 40-50% of videos resulting from tag-based queries to video sharing sites are both in SL and on topic [2]. This confirms that community-assigned tags alone do not provide reliable access to the contents of a digital collection [9].

When identifying the particular SL used in a video, the problem gets worse. Many videos are tagged with generic terms like "sign language", making it impossible for metadata to identify if it is an ASL or BSL video. Even when appropriately applied, tags related to SL, such as "ASL", are ambiguous since they could indicate that the video is in SL, about SL, or about something unrelated (e.g. ASL also stands for American Soccer League). The ability to identify SL videos and distinguish between particular languages based on video content would help resolve such ambiguities.

III. RELATED WORK

While some websites provide access to SL video, most of the current research concerning SL software is focused on aspects of SL learning [10], SL communications [11] or SL

978-1-5090-4571-6/16 $31.00 © 2016 IEEE
DOI 10.1109/ISM.2016.71

transcription [12]. These techniques assume that the video being analyzed includes SL and the SL being used is known.

In practice, much SL video is passed around from person to person. When looking for SL videos that have not already been viewed, people either go to generic Internet video sharing sites, such as YouTube or they go to sites devoted to SL (e.g. http://www.deafvideo.tv/). While general purpose video sharing sites have much more content due to their popularity, they also rely on tagging for retrieval.

So, instead of trying to transcribe SL videos, our work focuses on detecting and classifying them in video sharing sites. In previous work [4], we developed a technique for detecting SL based on video analysis. The SL video corpus was composed of videos in both ASL and BSL, but that included a single signer and a relatively fixed background. In follow-up work [5], we relaxed the constraints by including videos with multiple signers and dynamic backgrounds. We build on these efforts to detect SL video and present an investigation into whether similar or related techniques can be used to discriminate ASL from BSL.

Considering the task of identifying a particular type of SL, Gebre et al. [13] developed an automatic classifier using skin detection and Hu moments to extract video features, obtaining a F1 score of 95% when comparing BSL and Greek Sign Language – GSL videos from the Dicta-Sign corpus. Here we compare our approach to that of Gebre and examine its performance on the types of SL video found on video sharing sites.

IV. Sign Language Classifier

The remainder of this paper investigates using video analysis to distinguish different sign languages, performing comparisons on the pairs BSL or LSF, and ASL or BSL.

A. *Visual Differences*

The English dialects used in the US and the UK are fairly similar, bringing no problems for communication between people from these places. The same is not true when considering the SL used in each country. British Sign Language evolved from within deaf communities present on England, while American Sign Language is strongly influenced by French Sign Language (LSF). Studies demonstrate that ASL and BSL share just about 30% of their signs [14]. One of the most prominent differences between ASL/LSF and BSL is their approach to fingerspelling. While ASL and LSF manual alphabets are one-handed and letters are signed with the signer's dominant hand, the BSL alphabet is two-handed and most signs are performed around the center of signer's body. Fig. 1 shows these differences on the sign for letter E. Considering this difference, we investigated whether features based on aggregate visual behavior were likely to be able to distinguish between the languages. As first step, a corpus of 95 BSL videos was collected and compared against a similar ASL corpus.

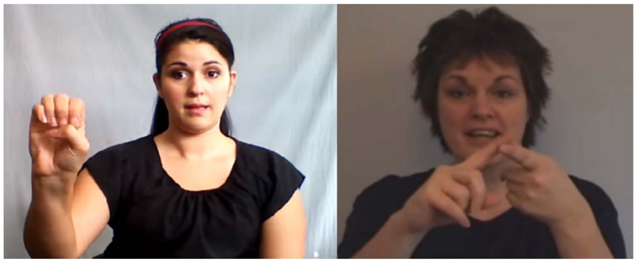

Figure 1. Letter E signed in ASL (left) and BSL (right).

The process for comparing the ASL and BSL collections relies on the face and foreground detectors used by our previous system [5]. For each video frame, we use a face detector to identify a region of interest (ROI) around each person. These ROIs are scaled based on the size of the face. Foreground activity in each ROI is likely to represent signing activity in SL videos. The result is a binary image where ON pixels represent foreground motion for a frame and OFF pixels represent the absence of motion. We scaled the ROIs from all frames in each collection to a standard width and height, and then computed the average activity map for each SL as:

$$I(x,y) = \frac{1}{\sum_{n=1}^{N}\sum_{t=1}^{T} R(t)} \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{R(t)} i(x,y) \quad (1)$$

where *N* is the number of videos, *T* is the number of frames for the n-th video, *R(t)* is the number of ROIs on frame *t*, *i(x, y)* is the pixel value on the given region *r* at coordinate *(x, y)*, and *I(x, y)* is the average pixel intensity.

Fig. 2 shows the resulting activity maps for the ASL and BSL collections. Not surprisingly, the activity maps are similar, but there is a distinct pattern for signing activity location for each language. ASL has a broader signing area and the heaviest region of activity is to the right-hand side of the signer. BSL has a more compact region of activity with heavy signing activity on the top central part of signers' torso. The difference in the location of the highest activity is largely due to the difference in fingerspelling.

B. *Polar Motion Profiles*

Considering the difference between ASL and BSL, Karappa et al's approach to representing videos as Polar Motion Profiles (PMPs) [5] seems promising. PMP is a translation and scale invariant measure of the amount of signing activity computed on a polar coordinate system (ρ, θ) centered on the signer's face.

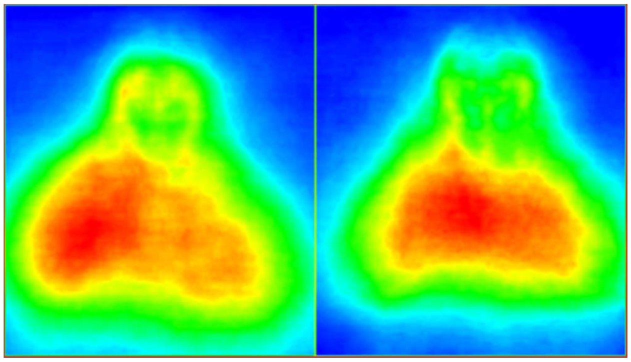

Figure 2. Activity maps for ASL (left) and BSL (right) videos.

Computed PMPs are centered on signers' faces. Both Karappa et al. [5] and our own work use a Haar-cascade face-detector to locate faces on each frame. Karappa et al. used five cascade classifiers and a majority voting scheme to more accurately determine face locations, reducing the occurrence of false positives at the expense of added computation. Instead, we use a single face detector (i.e., the Alt2 face detector in openCV), which we found yielded the best results among the five detectors used by Karappa et al.

In addition to locating the faces of potential signers in the video, computing the PMPs requires identifying those motions most likely to be signing activity. Since we deal with videos posted online, we use foreground-background separation, as it can reasonably work for a wide variety of video types and avoids the manual tuning of parameters. The foreground-background separation computes an adaptive background model based on Gaussian mixture model (GMM). Each video frame is then compared with the current state of the background model, and pixels further from any of the GMM components than a preset threshold are marked as foreground. A filter that removes small foreground objects (i.e. determined by contiguous foreground pixels) is then used to reduce noise.

Based on the face locations identified, ROIs are defined in every frame of the video. Each ROI is designed to be large enough to cover the likely span of arm and hand movements. Generation of the PMP for the video aggregates activity within each ROI identified using the foreground-background separation as described in [5].

C. Classifier

The classifier uses a Support Vector Machine (SVM) with a radial basis function kernel. The SVM is trained on labeled data containing the SLs that are going to be classified, with each video represented by its PMP.

V. Evaluation

To validate our solution, we perform two experiments. First, we use our classifier on the Dicta-Sign corpus of BSL and LSF videos. Videos on this corpus were recorded in a controlled environment with a static background and the same set of tasks per signer. As such, this corpus is not similar to videos found on video sharing sites, but it provides a good test case to validate our hypothesis on the difference in signing activity patterns between ASL/LSF and BSL. It also enables a comparison with results from Gebre et al., whose results were based on this corpus [13]. In a second experiment, we explore the performance of our classifier on real-world ASL and BSL videos, a task for which we find no prior approach to compare.

A. BSL vs. LSF Classification

For this experiment we used 128 videos from the Dicta-Sign corpus: 64 in BSL and 64 in LSF. BSL videos were recorded with a resolution of 480x384 and LSF videos 720x576. Both BSL and LSF videos were composed of one frontal capture for each of two signers, and a side capture containing both signers.

TABLE 1. BSL vs. LSF accuracy for varied video lengths

Video length	Precision	Recall	F1
10 s	94.7%	91.7%	93.0%
30 s	96.7%	93.5%	95.0%
60 s	99.6%	95.6%	97.5%

Our primary question was whether different types of SL could be distinguished based on the visual features encoded in the PMPs. The differences between the languages discussed in Section 4.1 and shown in Fig. 2 imply that it should be possible.

To evaluate the PMP-based classifier, we ran our classifier on the corpus, considering three different video segment lengths of 60, 30, and 10 seconds, respectively. All the executions used half of each language corpus for training and the other half for testing. Table 1 presents the average results of 1000 executions for each classifier configuration. In each execution, the training set was randomly selected and the remaining data was used for testing. While longer video segments improve the classifier performance, segments as short as 10 seconds are enough to correctly classify most videos, with about 95% precision, 92% recall, and 93% F1 score. In comparison, Gebre et al. reported 95% precision, recall, and F1 score when distinguishing between BSL vs. GSL (Greek Sign Language) [13] using 60 second clips. Like ASL and LSF, GSL has a one-handed alphabet.

B. ASL vs. BSL Classification

The second experiment focused on assessing classifier performance on real-world videos. The dataset was composed of ASL videos from Karappa et al. [5] and new BSL videos collected for this study. The entire dataset was collected from online video sharing sites like YouTube; videos were manually labeled as ASL or BSL. The corpus contains 100 ASL videos and 95 BSL videos. ASL videos have a resolution of 320x240 and BSL videos 424x240. These videos represent a wide variety of scenarios, including dynamic background and multiple signers. As such, this dataset closely resembles the set of videos that would be encountered in real-world situations.

The primary question on this experiment was how much harder it is to classify typical shared videos. Results on the Dicta-Sign corpus indicate that our approach can discriminate between SLs, but videos in this second corpus are much more challenging due to the lack of a controlled environment. Additionally, whereas the Dicta-Sign corpus has scripted contents aimed at teaching SL, this second corpus is mostly extemporaneous communication.

Once again, we analyze the effects of different video segment lengths on classifier performance, presenting the average results of 1000 executions for each configuration. The training set was randomly selected on each execution. For the experiments, the training set size was fixed at 60 videos per class. Tests were performed on segment lengths of 60, 45, 30, 15, 10, and 5 seconds.

Shortening the length of video segments necessary for classification provides two advantages: first, computation

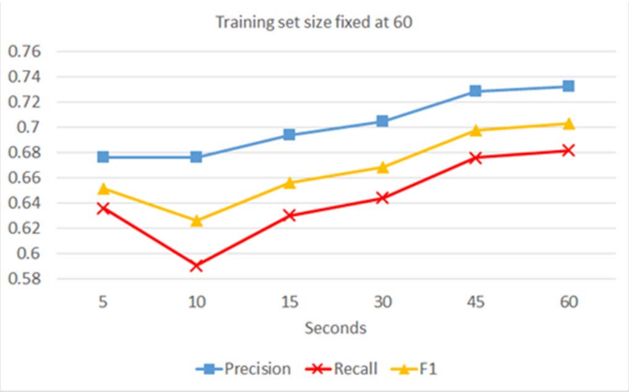

Figure 3. Precision, recall, and F1 score for video segment lengths.

time in the feature extraction stage can be substantially reduced; second, it enables the classification of shorter videos and video segments, which in turn enables the detection and diarization of videos where more than one language is present. To select segments with different sizes, we first take a one-minute base segment at the center of the full video, computing background-foreground separation and face locations; shorter segments with the corresponding PMPs are obtained from the base segment. Fig. 3 shows the results for the multiple segment lengths tested. As expected, using a one-minute segment yields the best results. All three measures experienced an overall improvement of around 5% when video segments are increased from 5 to 60 seconds.

Analyzing shorter segments has the advantage of reducing the computation time for video feature extraction. Of the three tasks performed during feature extraction, background subtraction, face detection, and PMP generation, face detection and PMP generation dominate the time required for feature extraction and thus classification. Using a 6th generation Intel i7 processor (2.50GHz, 1866MHz, 4MB) and 8GB of DDR3 RAM, the face detection and PMP generation each take about 3x real time for the 424x240 videos, with the total time being about 6x real time. Future work will explore replacing PMP generation with video features specifically tailored to distinguish between SLs.

VI. Conclusions

We have presented an approach to discriminating between ASL or BSL videos without any metadata information beyond features extracted from video itself. Differences in the activity heat maps for ASL and BSL –see Fig. 2—indicated they could potentially be discriminated by the spatial distribution of the signing activity. Considering this, Polar Motion Profiles were used as a model of aggregate signing activity.

A pilot experiment using the Dicta-Sign corpus and the PMP classifier was performed to validate our hypothesis. The results obtained (up to 99.95% precision, 97.02% recall, and 98.42% F1 score) indicate that visual features alone are able to identify a particular type of SL.

A more challenging classification task was then explored using a collection of ASL and BSL videos gathered from video sharing sites. The PMP classifier obtained 73% precision and 68% recall, with a F1 score of 70% when trained with 60 videos per class and analyzing 60 second video segments. Comparisons of different segment lengths indicate that even when using small video segments it is possible to discriminate ASL from BSL videos. Being able to use shorter video segment lengths brings many benefits, including reduced computation time to extract video features, being able to classify short videos, and diarization.

The goal of this research is to help the SL community to locate useful and intelligible content. While the previous classifier was able to successfully detect SL videos, identifying the SL used is similarly important, increasing the ability of signers to locate content in their own SL.

Distinguishing between two sign languages is a first step towards being able to label SL videos with the sign language being used in the video. A next step towards that end is to make use of the metadata attached to the videos.

References

[1] NIH, "American sign langauge," *NIH Publication No. 11-4756*, June 2011.

[2] 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.

[3] F. Shipman, R. Gutierrez-Osuna, T. Shipman, C. Monteiro, and V. Karappa, "Towards a distributed digital library for sign language content", *Proc. of JCDL*, 2015, 187–190.

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

[5] 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.

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

[7] Efthimiou, Eleni, et al. "Sign language recognition, generation, and modelling: a research effort with applications in deaf communication.", *Int. Conf. on Universal Access in HCI*. Springer Berlin Heidelberg, 2009, 21-30.

[8] J. Holt, S. Hotto, and K. Cole, *Demographic Aspects of Hearing Impairment: Questions and Answers, Third Edition*, Center for Assessment and Demographic Studies, Gallaudet University, 1994.

[9] C. Marshall, "No bull, no spin: A comparison of tags with other forms of user metadata", *Proc. JCDL*, 2009, 241-250.

[10] M. Huenerfauth, E. Gale, B. Penly, M. Willard, and D. Hariharan, "Comparing methods of displaying language feedback for student videos of American Sign Language." *Proc. ASSETS*, 2015, 139-146.

[11] J. Tran, B. Flowers, E. Risken, R. Ladner, and O. Wobbrock, "Analyzing the intelligibility of real-time mobile sign language video transmitted below recommended standards." *Proc. ASSETS*, 2014, 177-184.

[12] D. Dimov, A. Marinov, and N. Zlateva, "CBIR approach to the recognition of a sign language alphabet", *Proc. CompSysTech*. 2007, Article 96.

[13] B. G. Gebre, P. Wittenburg, and T. Heskes. "Automatic sign language identification.", *Proc. ICIP*, 2013, 2626-2630.

[14] D. McKee and G. Kennedy. "Lexical comparison of signs from American, Australian, British and New Zealand sign languages." *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima* (2000): 49-76.