Identifying Sign Language Videos in Video Sharing Sites

FRANK M. SHIPMAN, RICARDO GUTIERREZ-OSUNA, and CAIO D. D. MONTEIRO,
Texas A&M University

Video sharing sites enable members of the sign language community to record and share their knowledge, opinions, and worries on a wide range of topics. As a result, these sites have formative digital libraries of sign language content hidden within their large overall collections. This article explores the problem of locating these sign language (SL) videos and presents techniques for identifying SL videos in such collections. To determine the effectiveness of existing text-based search for locating these SL videos, a series of queries were issued to YouTube to locate SL videos on the top 10 news stories of 2011 according to Yahoo!. Overall precision for the first page of results (up to 20 results) was 42%. An approach for automatically detecting SL video is then presented. Five video features considered likely to be of value were developed using standard background modeling and face detection. The article compares the results of an SVM classifier when given all permutations of these five features. The results show that a measure of the symmetry of motion relative to the face position provided the best performance of any single feature. When tested against a challenging test collection that included many likely false positives, an SVM provided with all five features achieved 82% precision and 90% recall. In contrast, the text-based search (queries with the topic terms and "ASL" or "sign language") returned a significant portion of non-SL content—nearly half of all videos found. By our estimates, the application of video-based filtering techniques such as the one proposed here would increase precision from 42% for text-based queries up to 75%.

Categories and Subject Descriptors: I.4.9 [**Image Processing and Computer Vision**]: Applications; K.4.2 [**Computers and Society**]: Social Issues—*Assistive technologies for persons with disabilities*

General Terms: Algorithms, Design, Experimentation, Human Factors

Additional Key Words and Phrases: Sign language, ASL, video analysis, video sharing, metadata extraction

ACM Reference Format:
Shipman, F. M., Gutierrez-Osuna, R., and Monteiro, C. D. D. 2014. Identifying sign language videos in video sharing sites. *ACM Trans. Access. Comput.* 5, 4, Article 9 (March 2014), 14 pages.
DOI:http://dx.doi.org/10.1145/2579698

1. INTRODUCTION

The growing use of video sharing sites such as YouTube has resulted in a growing quantity of sign language video on the Web. Ideally, members of the sign language community would be able to search for discussions of topics in sign language, but limitations inherent in text-based search and incomplete and ambiguous tagging mean members of the sign language community often rely on ad hoc mechanisms (e.g. links in emails, blogs, and social network sites) to pass around pointers to these videos.

This material is based on work supported in part by the National Science Foundation under Grant No. DUE 09-38074 and by a gift from Microsoft Corporation. The statements made herein are solely the responsibility of the authors.

Author's address: F. M. Shipman; email: shipman@cse.tamu.edu.

DOI:http://dx.doi.org/10.1145/2579698

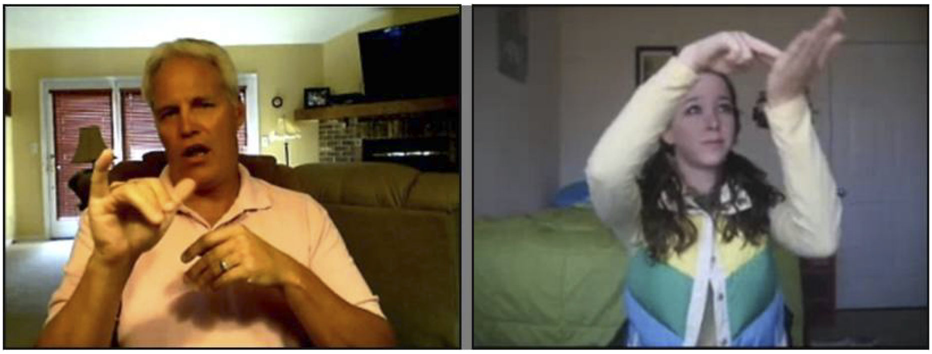

Fig. 1. Examples of SL video from video sharing sites.

In addition to the general purpose video sharing sites, sign language videos can be found at sites devoted to sign language content.[1] SignStream [Neidle et al. 2001] and the European ECHO project [Crasborn et al. 2007] also provide sign language collections, although the target users are researchers rather than the sign language community. While sign language sharing sites cover a wide range of topics, the quantity and timeliness of these sites cannot compare to the 1,320,000 videos returned from the query "sign language" on YouTube, 395,000 for "American sign language," 37,100 for "British sign language," and 24,500 for "'lenguaje de señas."

Since video sharing sites are developed for sharing all video, it is difficult to locate sign language video on a particular topic unless it is accurately tagged for both topic and language. Examinations of user-assigned tags [Heckner et al. 2008; Marshall 2009] indicate that tag quality varies for many reasons and that tags alone are unlikely to provide reliable access to the contents of a collection. In the context of locating sign language videos, tags related to sign language (e.g. "ASL") are ambiguous since they could be indicating that the video is either *in* sign language or *about* sign language. Such ambiguity is apparent in the first page of results for the query "sign language" on YouTube. This list of query results includes two music videos that include the phrase "sign language" in the title of the song, a video discussing the use of language in signs during demonstrations, and several videos about sign language recognition. The ability to identify sign language video would help resolve such ambiguities and also would be valuable when used in conjunction with tags to locate sign language videos (when tags exist). When tags are not available, which is likely when videos include sign language interpretation in a region of the video, the results could greatly improve access.

While there are plenty of videos of sign language conversations involving two or more signers, the focus here is on recordings where a single signer faces the camera and records an expression in sign language. We will use the term *SL video* to indicate this subset of videos including sign language. We consider these SL videos of an individual's message the equivalent to a document in a digital library in that they are deliberately recorded for others to access. Figure 1 provides examples of still images from such SL videos.

The work presented here aims at automatically identifying SL video found in video sharing sites. To the best of our knowledge, there is no previous work on automatically discriminating sign language from other forms of content in video. Such a capability would immediately allow members of the deaf and hard-of-hearing community to limit their searches within the large corpora of videos, to those in sign language. As a step

[1] For example, http://www.deafvideo.tv/, http://www.deafread.com/vlogs/, or http://www.aslvlog.net.

Table I. Number of Results and Number of On-Topic SL Videos on the First Page of Results

	with "ASL"		**with "sign language"**	
	matches	SL & on topic	matches	SL & on topic
Casey Anthony	6	5 of 6	2	1 of 2
Japan earthquake	8	7 of 8	56	12 of 18
Royal wedding	8	3 of 8	37	3 of 19
Osama bin Laden	19	9 of 18	18	8 of 18
Unemployment	46	9 of 18	61	11 of 19
Arizona shooting	1	1 of 1	0	0 of 0
Moammar Gaddafi	12	1 of 12	16	1 of 16
Amy Winehouse death	1	0 of 1	2	1 of 2
Arab spring	18	1 of 18	5	1 of 5
Occupy Wall Street	41	12 of 18	45	9 of 18
Total	*160*	*48 of 108* *(44.4%)*	*242*	*47 of 117* *(40.1%)*

towards this goal, in Monteiro et al. [2012] we proposed five video features and examined their use for SL video classification. In particular, we estimated the information content of each individual feature and compared it against the full feature set. This article is an expanded version of that conference paper. We extend that analysis to consider all possible feature subsets (Tables VII and VIII), and we quantify the difficulty of using current text-based approaches for locating SL videos on popular topics (Section 2). Our results show that the video-based classifier significantly outperforms text-based queries in identifying SL content. They also indicate that combining video and text features could dramatically increase access to SL content.

The next section assesses the performance of text-based approaches in retrieving SL video on video sharing sites. We then discuss related work on video analysis of sign language, and present the design of the classifier and its evaluation. Finally, we give directions for future work and conclusions from the project.

2. ASSESSING THE PROBLEM OF TEXT-BASED SEARCH FOR SL VIDEOS

Currently, unless they are sent a direct link to a video, members of the sign language community are limited to using text-based search mechanisms to locate SL videos. These mechanisms exploit relationships between query terms and the textual metadata associated with a video (e.g. title, tags, comments) to determine whether a video matches a query. We examined the results from expected text queries to assess how well current search mechanisms work for the purpose of locating SL videos on particular topics. A set of informational queries was needed for this assessment. After exploring a number of options (e.g. Google's Trends and hot searches, Bing's top searches, and Yahoo's top searches), we chose to use the top 10 news queries for 2011 from Yahoo! [Osmeloski 2011]. While many of the lists were largely made up of queries for individuals without any indication of why the person would be of interest, these queries are information-oriented and likely to be of interest to the general public. The list of queries is presented in Table I.

Queries for each of these topics were issued to YouTube as a string phrase (e.g. "Arizona shooting") with the following additional phrases ("ASL", "sign language"). Thus, two queries were issued for each of the ten topics. All queries were issued on January 8, 2013 and the results were saved for further analysis. The number of videos reported as matching the 20 queries ranged from 0 to 61 videos, with an average of 20.1 videos. Queries with ASL added to the topic averaged 16.0 matching videos while queries with "sign language" added to the topic averaged 24.2 videos.

Table II. Number and Percent of Results In/Not In Sign Language and On/Off Topic for Likely Added Query Terms

	In Sign Language	Not in Sign Language	Total
On Topic	95 (42.2%)	76 (33.8%)	171 (76.0%)
Not on Topic	26 (11.6%)	28 (12.4%)	54 (24.0%)
Total	121 (53.8%)	104 (46.2%)	225 (100%)

Each video on the first page of results for each query was examined to determine whether the video was on the desired topic and whether it was in sign language. YouTube presents a maximum of 20 videos and playlists on each results page. We did not consider YouTube playlists in our results as each of these could include dozens of videos. Because of this, no queries had more than 19 videos on its first page, as shown in Table I. For example, for the "Occupy Wall Street" and "ASL" query, YouTube reported 41 matching videos and we examined the 18 that were on the first page of query results. Of these, 12 were found to be in sign language and about the Occupy Wall Street topic. Because of the expertise required, a single person with sign language expertise did all assessments. In this examination, topic assessment was considered loosely (did the video discuss the topic at all). Table I shows how many true positives were on the first page for each of the queries.

A total of 225 videos were analyzed, i.e., those on the first page of YouTube results. Table II shows the breakdown of true positives (sign language videos on topic) and false positives (not in sign language, off-topic, or both). Ninety-five of those videos (42.2%) were on-topic and in sign language. Text-based retrieval based on metadata works reasonably well for the activity of identifying the topic of a video: 76.0% of the 225 videos were on the desired topic. Comparatively, though, only 121 (53.8%) of the videos were in sign language. Of these, 95 (78.5% of those in sign language) were on topic. Thus, text queries do not appear to be as good at determining whether a video is in sign language as they are at determining its topic.

This could be due to our choice of terms for locating sign language videos. Was it affected by how much ambiguity is associated with the term? Our two query variations ("ASL" and "sign language") might give some indication of the answer to this question.

ASL is an acronym and thus is likely to be used for other phrases (e.g. "above sea level", "age sex location" and "l'Armée syrienne libre"). This last acronym had a significant effect on the "Arab spring" query results, resulting in 13 of the 18 matches for the "ASL" version of the query being on topic but not in sign language. In contrast, none of the five matches for the "sign language" version of the query were on topic but not in sign language. The phrase "sign language" also has multiple meanings. It is used when discussing the use of language in printed signs. This meaning was most noticeable for the Occupy Wall Street queries where 9 of the top 18 "sign language" query matches were on topic but not in sign language while this was true for only four the "ASL" query's top 18 matches.

Overall, 53% of the videos returned for the "sign language" queries were in sign language; this number was increased to 55% for the "ASL" queries. Thus, it appears that the specificity of the terms we used had little effect on their ability to locate sign language content, or they were equally ambiguous; interestingly, the results for the two queries for the same topic included considerable overlap. Overall precision (the percentage of videos both in sign language and on topic) was 44.4% for "ASL" queries and 40.1% for "sign language" queries.

To explore how greater specificity in text queries might help with locating videos in sign language, we examined the results from the same ten topic queries with both "sign language" and ASL as query terms—thus, the query was of the form (topic-phrase "sign language" ASL). We expected this set of queries to result in a greater percentage

Table III. Number and Percent of Results In/Not In SL and On/Off Topic when Adding Both SL Query Terms

	In SL	**Not in SL**	**Total**
On Topic	50 (45.5%)	27 (24.5%)	77 (70.0%)
Not on Topic	24 (21.8%)	9 (8.2%)	33 (30.0%)
Total	74 (67.3%)	36 (32.7%)	110 (100%)

of videos being in sign language since the inclusion of both ASL and "sign language" query terms should disambiguate the meaning of those terms. We also expected an increase in overall precision and the number of videos to be returned to be much smaller.

The results are presented in Table III. As expected, the more specific queries did increase the percentage of top ranked videos in sign language to 67.3%. Surprisingly, the new queries resulted in 217 reported matches for the 10 queries. This is more matches than the 160 reported matches for the ASL variants of the queries, indicating that YouTube does not use a Boolean comparison when determining whether videos match query terms. Additionally, the new queries resulted in some videos not seen as matches for either of the prior queries on the same topic. These videos tended to be in sign language but not on topic. The percentage of videos on topic went down to 30.0%. The increase in sign language videos and the decrease in on-topic videos combined to result in a small effect on overall precision—it increased to 45.5%. For these topics with YouTube's approach to text search, the ambiguity of the terms did not greatly influence overall precision.

Thus, there is reason to believe that better text queries will not solve the problem of locating SL videos on particular topics. First, text-based retrieval engines often use concept-based or knowledge-based associations between terms. Our results suggest this is true for YouTube as well since all three query variations ("ASL", "sign language", and both) returned many of the same videos (including false positives). Second, for video sharing sites that do use Boolean text retrieval, increasing the specificity of the query will likely result in more false negatives. This reduces the number of SL videos on the desired topic presented to the user. While we do not know how many SL videos on the topics were not found because their metadata did not match our queries, it seems clear that techniques that increase precision by adding query terms are likely to reduce recall.

The results in Tables II and III show how well the current text-based search mechanism provides access to sign language content. This should be taken not as a critique but as an indication of how ambiguity in the query and the search mechanism itself combine to return a significant proportion of non-sign language content—nearly half of all videos found.

Our technique for identifying sign language based on video content has the potential to greatly reduce the number of videos that must be examined by those looking for SL video. No technique will work perfectly, but if a filter is applied to the results generated via text-based search (Table II) that correctly identifies 90% of SL videos (recall rate) and correctly removes 95% of non-SL videos (precision rate), the resulting overall precision for this particular set of queries would increase from 42.2% to 74.9%.

$$\textit{Predicted precision} = 0.9 \times 95\,/\,(0.9 \times (95 + 26) + 0.05 \times (76 + 28)) = 0.749.$$

An equivalent estimate of the effect on recall rates is difficult to compute since we do not know how many SL videos on-topic exist.

The existence of a technique to detect SL videos could replace the use of query terms to locate sign language content. In such a case, the set of potential videos are all those that match the topic query. For the topic queries in this analysis, YouTube reported a range from about 800 to 200,000 matching videos. This implies that the precision of sign language detection would need to be very high to avoid generating too many false positives.

3. RELATED WORK

While the problem of identifying sign language videos has not been addressed to the best of our knowledge, it is related to prior work aimed at transcribing the content of sign language videos. The latter is a much more difficult problem than identifying whether sign language is present in a video, much like speech-to-text transcription is harder than detecting speech in an audio stream.

Transcribing sign language from video is a very difficult problem, particularly with user-contributed videos. In one of the earliest studies, Starner et al. [1998] developed a hidden Markov model (HMM) classifier capable of recognizing up to 40 words in sign language. The small vocabulary and the requirement that the system be trained for each individual, limit the applicability of such an approach for our goals. Somewhat more generalizable, Somers and Whyte [2003] approached the problem as one of matching 3D models of handshapes and silhouettes but again were limited to a small set of signs. Instead of using a learned vocabulary or 3D model, other researchers have attempted to recognize handshapes through image similarity comparisons to known signs [Dimov et al. 2007] or to handshapes [Potamias and Athitsos 2008]. Combining multiple techniques is more likely to capture sufficient information to determine the five components of each sign—handshape, position, palm orientation, motion, and facial expression. Towards this goal, Caridakis et al. [2000] proposed an architecture for providing features for hand trajectory, region, and shape to a combination of self-organizing maps, Markov chains, and HMMs for recognition, although it does not appear to have been instantiated or evaluated. Approaches developed for sign language transcription are of limited value in our context in that most of them work only modestly with relatively small vocabularies, or are signer-dependent and require large amounts of training data. In addition, much of this work has focused on recognizing single signs or handshapes in isolation rather than in sentences or phrases (exceptions include Hernandez-Rebollar [2005] and Vogler and Metaxas [1999]). Finally, most of these efforts do not discuss the speed of expression—a fluent signer communicates very rapidly with other fluent signers but will drastically slow down for non-fluent signers.

Given these challenges, our approach to supporting the sign language community avoids translating sign language in the first place. Detecting sign language is a much simpler problem than translating it. As an example, Cherniavsky et al. [2008] used a simple activity detection technique for cell-phone cameras that could determine whether or not a user was signing with 91% accuracy, even in the presence of noisy (moving) backgrounds. While many measures of motion can be used to determine whether a person is watching the other signer or signing themselves, they would not be able to discriminate between signing and many other forms of motion by a person. Thus, it is unlikely this algorithm would be as successful in distinguishing between SL videos and other videos involving people gesturing.

4. DESIGN OF SL-VIDEO CLASSIFIER

Our SL-video classifier is composed by two components: a video processing module that generates video features, and a classification module that determines whether or not they have sign language content.

4.1. Video Processing

The video processing module[2] is responsible for generating video features designed to have potential value in distinguishing SL video from other video. The module performs background modeling to segment moving objects from the stationary background, and face detection to determine if a person is facing the camera (and establish their location within the video frame). For videos where many of the frames are found to have a person facing the camera, the video features are computed.

Background Modeling. A dynamic background [Cucchiara et al. 2003; Gloyer et al. 1995] is best suited for the task of identifying SL videos on video sharing sites due to the wide variety of videos that must be analyzed. These videos can have lighting changes, camera motion, and motion in the background.

Unlike surveillance or other applications requiring foreground/background separation, it is not important to identify the signer as a single foreground object. Since the signer is often seated and facing the camera, much of their body can end up included in the background model. As a result, a relatively simple dynamic background model can be used without losing information needed by the classifier.

The background model is computed as a running average of the grayscale frames of the video. In such models, a discount rate α determines how fast the background model changes over time. A high discount rate results in a highly dynamic background model where only the most abrupt movements are detected, while with a low discount rate any slight change in the image will be detected. We have found having each new frame account for 4% of the next background model works well for this task. Thus, the background model for a background pixel BP at time t is

$$BP(t) = (1 - \alpha)\, BP(t - 1) + \alpha P(t),$$

where $P(t)$ is the grayscale value of the pixel at time t, and $\alpha = 0.04$. Figure 2(c) shows the background model for the frame in Figure 2(a).

Once the background model is computed, each grayscale pixel of the next frame is compared to the pixel in the background model. If these two values differ by more than a threshold (our threshold is 45) the pixel is considered a foreground pixel. Figure 2(d) shows the results of this process for the frame in Figure 2(a). Normal body movements and changes in the background can result in noise throughout the frame. A spatial filter that removes all small regions of foreground pixels is used to locate the large moving objects in the image, the result of which is shown in Figure 2(b). As shown in Figure 2, our background model will ideally contain the whole image except for the signer's hands and arms. Once the final foreground model is computed, the feature extraction process starts.

Feature Extraction. The first step in the feature extraction process is to locate the signer's head. This is done through a face detection algorithm based on Haar-like features [Viola and Jones 2001]. Figure 2(a) shows a white box around the face location.

Once the face has been located, the next step is to compute visual features. Our approach uses five video features developed with the intuition that the location, quantity and speed of sign language motion is distinct from the motion associated with normal gesturing (as done by a politician at a podium), domain-oriented gesturing (like a weatherperson), and other forms of human motion (dance, mime, charades). The

[2] The module was developed using the openFrameworks open-source toolkit [Lieberman et al. 2011].

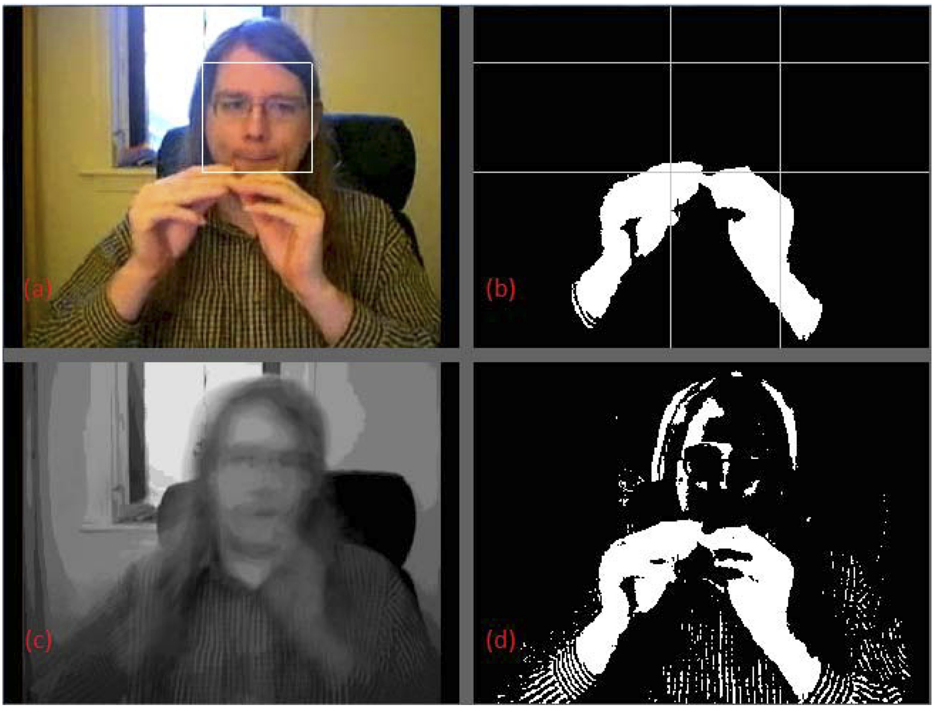

Fig. 2. (a) the incoming frame of the video, (b) the final foreground image, (c) the actual background model and (d) the intermediate foreground image.

following features contain information about the overall quantity of movement, the continuity of movement, and the location of the movement relative to the face.

(1) *Quantity of movement*. We extract two features.
   (i) VF1: total amount of activity per video, measured as the proportion of pixels on each frame that are on the foreground, averaged across frames,
   (ii) VF2: spread of activity across the scene, measured as the proportion of pixels that are included in the foreground model of at least one frame.
(2) *Continuity of motion*. We extract one feature.
   (i) VF3: proportion of foreground pixels that change between one frame and the previous one, averaged across frames.
(3) *Location of motion*. We extract two features.
   (i) VF4: symmetry of motion, measured as the proportion of foreground pixels that are in a symmetric position relative to the center of the face, averaged across frames. Some signs are symmetric but many signs are made with a single hand or using different gestures for each hand and some signs that are symmetric, SL videos are likely to fall within a symmetry band—not too much or too little symmetry when compared to other human activity (e.g. common gesturing, dance).
   (ii) VF5: amount of non-facial movement, measured as percentage of pixels outside of the facial rectangle that are part of the foreground, averaged across frames. SL videos contain significant hands/arms and torso movements relative to head movement, so the number of frames containing foreground pixels outside the face region is an important feature.

Table IV. Results from Varying the Size of the Training Set for the Classifier with All Five Visual Features as Inputs

# Videos/Class	Precision	Recall	F1 Score
15	81.73%	86.47%	0.84
30	83.62%	88.11%	0.85
45	80.67%	91.00%	0.85
60	82.21%	90.83%	0.86

4.2. Classifier

Since our goal is to discriminate SL and non-SL videos, a binary classifier is suitable. We explored several classifiers (e.g., Gaussian classifiers, nearest neighbors) at an early stage in the project but chose a Support Vector Machine (SVM) classifier [Cristianini and Shawe-Taylor 1999] due to its higher performance. The SVM is trained on a dataset containing an equal number of known SL videos and known non-SL videos, each video represented by the five features (VF1–5). The classifier is then evaluated on a different set of videos.

5. EVALUATION OF CLASSIFIER

A collection of 192 videos, including 98 SL videos (including 78 in American Sign Language and 20 in British Sign Language) and 94 non-SL videos was collected from video sharing sites like YouTube, Vimeo, and so on. The majority of the non-SL videos were selected by browsing for likely false-positives based on visual analysis (e.g. the whole video consists of a gesturing presenter, weather forecaster, or other person moving their hands and arms.) A small number of non-SL videos were included due to confounding tags or metadata indicating a relationship to sign language (e.g. videos that are located via text search when searching for videos in sign language.) However, we only included a small number of these videos since they were visually distinct from SL videos and thus not difficult for our classifier. For example, if a particular video does not include a person for most of the duration, it will be very easy to reject it based on the face-detection step alone.

5.1. Processing of Videos

Each of the selected videos (MPEG4 format) was cut to 1 minute, a duration that is long enough for feature extraction yet keeps the processing requirements bounded regardless of the length of the original video. The 1-minute interval for each video was chosen randomly, but we avoided the start or the end of the video to avoid front or back matter (e.g. credits at the end or titles or other pre-presentation content at the beginning). The video processing and feature extraction routines were then run on each resulting 1 minute video and the results were stored for use by the various classifiers considered.

5.2. Results

The classifier was tested on 1000 executions for each of the contexts assessed; in each execution the training and test data were selected randomly. The performance measures considered are the precision (number of correct SL classifications divided by all SL classifications), recall (number of correct SL classifications divided by the total number of SL videos in the test set), and the F1 score (the harmonic mean of precision and recall).

Our first assessment was to determine the impact of training set size on classifier performance. Table IV shows the results for different training set sizes when all five video features are provided to the SVM; in all cases, examples not included in the

Table V. Results when One Feature is Not Provided to the Classifier with a Training Set of 15 Videos/Class

Video Feature Removed	Precision	Recall	F1 Score
VF1	80.36%	86.25%	0.83
VF2	78.34%	85.41%	0.82
VF3	78.90%	83.62%	0.81
VF4	72.80%	74.30%	0.74
VF5	78.86%	85.60%	0.82

Table VI. Results when Only One Feature is Provided to the Classifier with a Training Set of 15 Videos/Class

Video Feature	Precision	Recall	F1 Score
VF1	70.48%	60.14%	0.65
VF2	73.57%	53.26%	0.62
VF3	65.65%	64.03%	0.65
VF4	75.95%	83.69%	0.80
VF5	56.31%	49.52%	0.53

training set were used for testing. Overall, the classifier achieved slightly above 80% precision and 90% recall, resulting in a F1 score around .85. This means that four out of five videos identified as being SL videos really were SL videos, and 9 out of 10 videos in the testing corpus were correctly identified.

As the number of videos used to train the classifier is increased, the precision stays relatively stable (irregularly varying within a 3% band). As the number of training examples per set is increased, recall improves by more than 4%. This indicates that while more training data certainly improves performance, the classifier works well with a small training set (15 videos per class). All further assessments reported are performed with 15 training videos per class and thus 162 test videos.

Given this result, we explored the relative value of the five visual features through a series of feature-subset-selection experiments. First, we estimated the loss in classification performance when each of the five features is removed. For this purpose, we used 15 videos per training class and trained a classifier using all but one of the videos features as input. The results in Table V show that VF4, a measure of the symmetry of motion relative to the face, is the most critical feature: without it, precision drops almost 9% and recall drops more than 12% compared to a classifier trained on all five features. Removing each of the remaining features led to a small reduction in performance, which suggests those features provide overlapping information to the classifier. VF1 was the least valuable feature in this context—its removal resulted in a 1.3% drop in precision and a drop of 0.2% in recall.

Next, we explored which single visual feature had the most discriminative power when used as the sole input to the classifier. Again, the classifier was trained on 15 videos from each class. The results in Table VI indicate that VF4 is again the best predictor. The difference between this feature and the other four is significant; VF4 alone outperforms the other four features combined. This result is interesting because it gives direction in the search for additional video features that might be valuable for this task. VF4 is a measure of the symmetry of motion relative to the face of the signer indicating alternative measures comparing movement on the two sides of the body should be explored.

Based on these results, VF4 is clearly the most valuable feature for SL discrimination. But which of the other features adds the most additional information to this activity? Table VII shows the results of providing all pairs of features to the SVM. Unexpectedly, it is the worst performing single feature, VF5, that increases performance

Table VII. Results when Two Features are Provided to the Classifier with a Training Set of 15 Videos/Class

Video Feature	Precision	Recall	F1 Score
VF1 & VF2	71.07%	60.69%	0.65
VF1 & VF3	69.16%	63.40%	0.66
VF1 & VF4	76.00%	82.83%	0.79
VF1 & VF5	72.18%	73.65%	0.73
VF2 & VF3	71.06%	64.43%	0.68
VF2 & VF4	75.02%	84.98%	0.80
VF2 & VF5	72.02%	63.51%	0.67
VF3 & VF4	76.67%	82.34%	0.79
VF3 & VF5	63.61%	61.01%	0.62
VF4 & VF5	79.25%	86.39%	0.83

Table VIII. Results when Three Features are Provided to the Classifier with a Training Set of 15 Videos/Class

Video Feature	Precision	Recall	F1 Score
VF1 & VF2 & VF3	70.20%	64.75%	.67
VF1 & VF2 & VF4	76.81%	86.91%	.82
VF1 & VF2 & VF5	72.88%	72.94%	.73
VF1 & VF3 & VF4	77.11%	82.08%	.80
VF1 & VF3 & VF5	72.11%	71.92%	.72
VF1 & VF4 & VF5	78.66%	84.33%	.81
VF2 & VF3 & VF4	76.11%	83.17%	.79
VF2 & VF3 & VF5	72.23%	68.68%	.70
VF2 & VF4 & VF5	79.89%	88.23%	.84
VF3 & VF4 & VF5	80.06%	83.97%	.82

the most. VF5 is a measure of overall movement outside of the facial rectangle. Thus, it turns out the two video features that make use of face detection perform best when paired together.

Finally, we compare the performance of the SVM when provided with all permutations of three of the video features. As shown in Table VIII, the best performing group is when VF2 is added to the VF4 and VF5 combination. VF2 is a measure of the spread of motion around the video frame.

Overall, these results show that, given a good feature selection, an SVM classifier can discriminate SL vs. non-SL video, even with small training sets (15 videos). Given that we selected non-SL videos to be similar to SL video, we expect that such a classifier would perform at a quite high degree of accuracy when applied to the broader collections found on video sharing sites.

Combining these results with the results of the analysis of text-based retrieval can provide a sense of the difference such a capability would provide. The best version of the SVM (provided with all five features and with the largest training set) achieved 90.83% recall. Precision on our corpus is considerably lower than precision would be for the videos returned for the informational queries because many are news stories that show scenes from events rather than a person looking at the camera. Indeed, even the videos of a person looking at the camera, such as videos of a reporter, are very different from the likely-false positive videos in our test collection where our classifier achieved 82.21% precision. A conservative estimate is that precision on such a real collection would be 95%. Using these values, our classifier would increase precision from 42% to 75% for the news queries on YouTube. As already mentioned, it is impossible to determine how many SL videos on the desired topics that were not returned as part

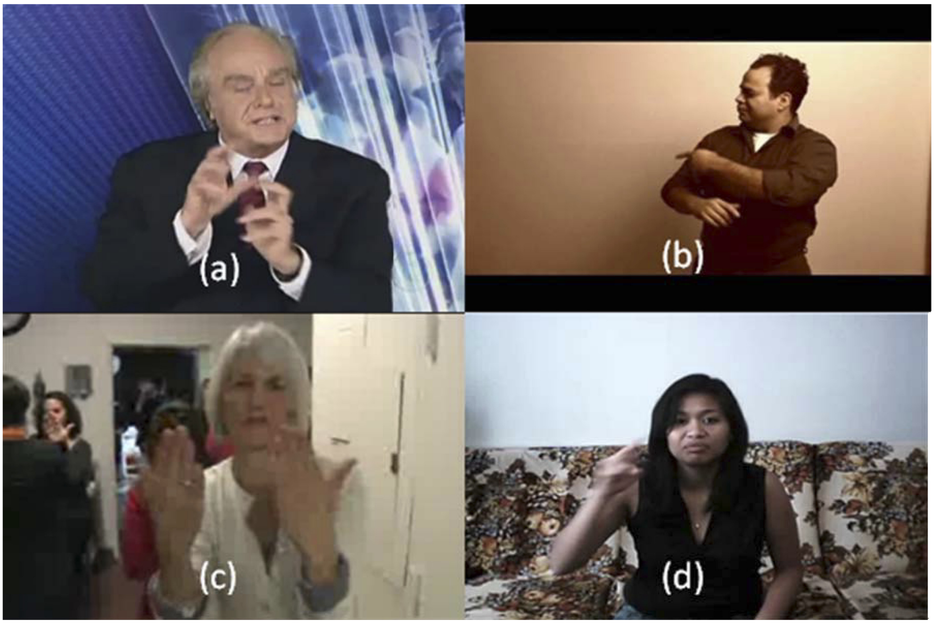

Fig. 3. Examples of videos that are difficult to successfully classify with the current approach.

of the analysis in Section 3 would be identified by this approach. Thus we cannot determine the effect on recall.

5.3. Discussion of Failures

Working with videos collected from video sharing sites results in a variety of issues that impact classification performance. Poor illumination, sudden illumination changes, and poor video resolution resulted in some videos being incorrectly classified. Examples of videos from our test collection that were difficult to classify are shown in Figure 3. We already discussed why videos may be incorrectly classified as being SL video: a presenter facing the camera and gesturing fairly constantly makes correct classification difficult, such as the newscaster in Figure 3(a). Some SL videos were not detected because the signer was sitting too far from the camera or was not facing the camera, resulting in their face not being detected, as in Figure 3(b). Additionally, signing in front of backgrounds with lots of movement (Figure 3(c)) is not detected because the hand/arm blobs get combined with the other activity. The background model also causes problems when the background includes grayscale intensities that match the intensities from the signer's skin tone or shirt color, such as the couch in Figure 3(d). These problems point to the need to improve the current process for modeling the background and to improve on our simple approach of equating face detection in a frame to face location.

6. FUTURE WORK

The success of the SL video classifier leads to a variety of directions for future work. As is clear from the discussion of the failures, it would be valuable to improve our background modeling process to increase the accuracy of the final foreground blobs: the hands and arms of the signer for SL video. Similarly, the current approach to face detection causes problems when signers are not near the camera or when they turn their heads. An approach that uses information about where a face was last detected could infer the position of the head in subsequent frames for short gaps in face tracking.

We also are exploring additional video features likely to increase the SVM performance. These include, for example, features that extract information across multiple frames, such as trajectories and velocity profiles of foreground blobs. While we compared different types of classifiers early on in the project, we plan to examine the performance of other classifiers (e.g. 1-class SVMs) with the video features we have since developed.

Finally, we are developing techniques for generating a much larger collection of SL video and non-SL video to increase the corpus size and variety available for training and evaluation.

7. CONCLUSIONS

General purpose video sharing sites such as YouTube are being used to share sign language presentations among members of the sign language community. The current text-based search provided by these sites requires the existence and accuracy of metadata indicating the video is in sign language. The resulting difficulty of locating SL videos means that pointers to videos are emailed or otherwise communicated from person to person. To address such shortcomings, we previously presented an approach to identify videos with sign language content by providing an SVM classifier with five video features extracted based on relatively simple background modeling and face detection [Monteiro et al. 2012]. This article extends that description of how the five features are extracted, quantifies the performance of current text-based retrieval, and examines the performance of all permutations of the original five video features.

To determine the performance of text-based retrieval for SL videos on particular topics, we examined the results of the top 10 news queries from Yahoo! for 2011. Two queries were issued to YouTube for each topic—one with “ASL” and one with the phrase “sign language” added to the original phrase. The results show that while 75.7% of the resulting videos were on the desired topic, only 42% of the videos were in sign language and on topic. Thus, filtering the results returned from text-based search to those videos in sign language would greatly reduce the number of extraneous videos that users must examine.

Unique to our approach is the collection of SL videos and non-SL videos that were likely false positives from video sharing sites for training and testing. The resulting classifier does not require large quantities of training data—15 examples for each category were sufficient to have greater than 81% precision and 86% recall; providing more examples improved recall but not precision. A comparison of the performance of the five video features through sequential forward/backward selection showed that a measure of the symmetry of motion relative to the center of the face was the most valuable feature for classification. Alone it was more accurate than using the other four features combined. Adding a second video feature (overall movement outside of the facial rectangle) further improves accuracy, at which point additional features only improved accuracy marginally.

Our results show ways in which image processing techniques could be used to increase access to sign language presentations for members of the deaf and hard of hearing community. The existing classifier could be applied to video sharing sites so users could filter their search results with accuracy rates much higher than reported here, since the non-SL videos in our collection were chosen to be hard to differentiate from SL videos. As an example, with the recall from the best version of the current classifier, and assuming a 95% precision on the real corpus (this is reasonable since most of the non-SL videos would be identified as such due to the lack of a person facing the camera), the precision from our initial text-based retrieval (see Table II) would increase from 42% to 75%.

We are currently improving our background modeling and use of face detection to improve on the current results. We are also identifying alternative video features that address the more common difficulties for the features described here.

REFERENCES

George Caridakis, Olga Diamanti, Kostas Karpouzis, and Petros Maragos. 2008. Automatic sign language recognition: Vision based feature extraction and probabilistic recognition scheme from multiple cues. In *Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, New York, NY, Article 89, 8 pages.

Neva Cherniavsky, Richard E. Ladner, and Eve A. Riskin. 2008. Activity detection in conversational sign language video for mobile telecommunication. In *Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE Computer Society, Los Alamitos, CA, 1–6.

Onno Crasborn, Johanna Mesch, Dafydd Waters, Annika Nonhebel, Els van der Kooij, Bencie Woll, and Brita Bergman. 2007. Sharing sign language data online. Experiences from the ECHO project. *Int. J. Corpus Ling.*, 4, 535–562.

Nello Cristianini and John Shawe-Taylor. 1999. *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK.

Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. 2003. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Trans. Pattern Anal. Mach. Intell. 25*, 10, 1337–1342.

Dimo Dimov, Alexander Marinov, and Nadezhda Zlateva. 2007. CBIR approach to the recognition of a sign language alphabet. In *Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech)*. ACM, New York, NY, Article 96, 9 pages.

Brian Gloyer, Hamid K. Aghajan, Kai-Yeung Siu, and Thomas Kailath. 1995. Video-based freeway-monitoring system using recursive vehicle tracking. In *Proceedings of SPIE*. 173–180.

Markus Heckner, Tanja Neubauer, and Christian Wolff. 2008. Tree, funny, to read, Google: What are tags supposed to achieve? A comparative analysis of user keywords for different digital resource types. In *Proceedings of the ACM Workshop on Search in Social Media (SSM)*. ACM, New York, NY, 3–10.

Jose L. Hernandez-Rebollar. 2005. Gesture-driven American Sign Language phraselator. In *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI)*. ACM, New York, NY, 288–292.

Zachary Lieberman, Theodore Watson, and Arturo Castro. 2011. http://www.openframeworks.cc/.

Catherine C. Marshall. 2009. No bull, no spin: A comparison of tags with other forms of user metadata. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*. ACM, New York, NY, 241–250.

Caio D. D. Monteiro, Ricardo Gutierrez-Osuna, and Frank M. Shipman. 2012. Design and evaluation of classifier for identifying sign language videos in video sharing sites. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. ACM, New York, NY, 191–198.

Carol Neidle, Stan Sclaroff, and Vassilis Athitsos. 2001. SignStream^{TM}: A tool for linguistic and computer vision research on visual-gestural language data. *Behav. Res. Meth. Instrum. Comput. 33*, 3, 311–320.

Elisabeth Osmeloski. 2011. 2011 Yahoo! In review: Top US searches in 30 categories. http://searchengineland.com/2011-yahoo-in-review-top-us-searches-in-30-categories-103215.

Michalis Potamias and Vassilis Athitsos. 2008. Nearest neighbor search methods for handshape recognition. In *Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*. ACM, New York, NY, Article 30, 8 pages.

G. Somers and R. N. Whyte. 2003. Hand posture Matching for Irish Sign Language interpretation. In *Proceedings of the 1st International Symposium on Information and Communication Technologies (ISICT)*. Trinity College Dublin, 439–444.

Thad Starner, Alex Pentland, and Joshua Weaver. 1998. Real-time American Sign Language recognition using desk and wearable computer based video. *IEEE Trans. Pattern Anal. Mach. Intell. 20*, 12, 1371–1375.

Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference (CVPR)*. 511–518.

Christian Vogler and Dimitris N. Metaxas. 1999. Toward scalability in ASL recognition: Breaking down signs into phonemes. In *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction (GW)*. Springer-Verlag, Berlin, 211–224.

Received January 2013; revised October 2013, November 2013; accepted November 2013