Design and Evaluation of Classifier for Identifying Sign Language Videos in Video Sharing Sites

Caio D.D. Monteiro, Ricardo Gutierrez-Osuna, Frank M. Shipman III
Department of Computer Science and Engineering
Texas A&M University
College Station, Texas, 77843-3112
1-979-862-3216

caioduarte.diniz@gmail.com, rgutier@cse.tamu.edu, shipman@cse.tamu.edu

ABSTRACT
Video sharing sites provide an opportunity for the collection and use of sign language presentations about a wide range of topics. Currently, locating sign language videos (SL videos) in such sharing sites relies on the existence and accuracy of tags, titles or other metadata indicating the content is in sign language. In this paper, we describe the design and evaluation of a classifier for distinguishing between sign language videos and other videos. A test collection of SL videos and videos likely to be incorrectly recognized as SL videos (likely false positives) was created for evaluating alternative classifiers. Five video features thought to be potentially valuable for this task were developed based on common video analysis techniques. A comparison of the relative value of the five video features shows that a measure of the symmetry of movement relative to the face is the best feature for distinguishing sign language videos. Overall, an SVM classifier provided with all five features achieves 82% precision and 90% recall when tested on the challenging test collection. The performance would be considerably higher when applied to the more varied collections of large video sharing sites.

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.

Keywords
Sign language, ASL, video analysis, video sharing, metadata extraction.

1. INTRODUCTION
Sign languages have developed in many countries and regions. These languages have their own vocabulary, syntax and grammar that is distinct from the spoken/written language of their region [19]. Unlike written communication, however, and more like spoken communication, signing provides a wealth of non-verbal information. Thus, video sharing websites offer a great opportunity for members of the deaf and hard of hearing community to exchange signed content. Unfortunately, video sharing services do not have the ability to locate untagged or unlabeled sign language (SL) content. As a result, members of this community rely on ad-hoc mechanisms to pass around pointers to internet-based recordings, such as email, blogs, etc.

In what follows, we will use the term "SL video" to denote videos where one person faces the camera and records an expression in a SL; examples of SL video are shown in Figure 1. While other forms of videos can include sign language (e.g. video of a conversation between signers), these deliberate recordings of an individual's message are a form of sign language document meant to be accessed by others.

Figure 1. Examples of SL video from video sharing sites

While early work on recognizing SL in video is underway, most of the efforts focus on translating the sign into English (or another spoken/written language). Because of the complexity of the problem, most of this work emphasizes the recognition of hand shape and orientation but such capabilities, while necessary, are not sufficient to translate sign language. The meaning of American Sign Language (ASL) is determined by a combination of five characteristics: the shape of the hand(s), the position of the hand(s), the palm orientation of the hand(s), the direction and speed of motion of the hand(s), and the facial expression. Without taking into account all five components, true translation of ASL is not possible. For example, a hand shape and orientation recognizer could identify the sign for "help" but without position and motion information would be unable to identify who was

ASSETS'12, October 22–24, 2012, Boulder, Colorado, USA.

helping whom and without the facial expression would not know whether the help was going to happen or not, or whether it was a statement of fact or a question. Even with accurate identification of a sign, the concept of that sign must be translated based on the context and prior content, similar to translating between spoken/written languages.

The work presented here aims at a shorter-term goal: automatically identifying sign language video found in video sharing sites. 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.

The next section further describes the motivation for this project. This is followed by an overview of related work. We then describe the design of the classifier and its evaluation. Finally, we give directions for future work and conclusions from the project.

2. 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" [10]. For many that become deaf early in life, sign language is their primary means of communication. As is true with spoken languages, a unique community has formed around ASL with its own cultural norms and expectations [15]. The same is true in many other countries.

Many who grow up deaf learn English as a second language – ASL being their first learned language. Combined with late-identification of hearing loss and the lack of communication during formative periods of the brain's development, this means that the average reading/writing skills among members of the deaf population are well below average. Holt et al. [11] found that the median reading comprehension for deaf and hard-of-hearing 17- and 18-year-olds is at a 4.0 grade level, indicating half the population has a lower reading level than typical hearing students at the beginning of 4th grade. As such, for large portions of the deaf community, much of the information available on the Internet is difficult to locate and understand. For the internet to more fully support this community, information needs to be available in sign language.

Figure 2. Videos returned on the first page of results for query "sign language" that are not in sign language. Two are for songs with "sign language" in the title, one is on sign language recognition research, and one refers to the language in signs.

On-line video sharing sites, such as YouTube, have provided members of the deaf and hard of hearing community a way to publish and access content in sign language. Since these sites are developed for sharing all video, it is difficult to locate SL video on a particular topic unless it is accurately tagged for both topic and language. Studies of community-assigned tags indicate that tags alone are unlikely to provide reliable access to the contents of a collection [7][12]. Even when appropriately applied, tags related to sign language, such as "ASL", are ambiguous since they could be indicating that the video is either in sign language or about sign language. Figure 2 shows examples of such videos returned by YouTube for "sign language". 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, a likely event when videos include sign language interpretation in a region of the video, the results of our work could greatly improve access.

3. RELATED WORK

While a few web sites provide access to SL video, research related to these projects primarily concerns aspects of SL translation – either handshape or sign recognition. These efforts can be classified according to whether they rely on standard unaugmented video, require signers to wear visual markers to help tracking of hands (e.g. colored gloves, infrared tags), or use data gloves and other sensors.

3.1 Locating Sign Language Video

To locate SL video, people either go to Internet video sharing sites, such as YouTube or sites devoted to SL such as http://www.deafvideo.tv/ or http://www.deafread.com/vlogs/. To the best of our knowledge, there is no previous work on automatically discriminating SL from other forms of content in video, neither are we aware of previous work on classifying SLs based on video information. An ASL video directory, http://www.aslvlog.net/, has started to categorize ASL videos according to topic. While covering a wide range of topics, there are relatively few videos found on this site compared to the 51800 videos found on YouTube that are returned from the query "ASL", 61400 for "sign language", 2390 for "British sign language", and 3090 for "'lenguaje de señas". Several academic projects provide SL content, such as SignStream [14], or the European ECHO project [3], but these corpora are designed for researchers engaged in sign language translation efforts rather than for the deaf and hard of hearing or for detecting sign language or identifying the sign language in use.

3.2 Recognizing Sign Language in Unaugmented Video

Recognizing the content of SL from video only, as is available on video sharing sites, is a very difficult problem. In one of the earliest studies, Starner et al. [18] used hidden Markov models (HMMs) to recognize a vocabulary of 40 words for a single signer. The goal was to provide a SL-to-English translator that would allow the deaf to communicate in one-on-one situations. Thus, in addition to testing the approach with a camera facing the signer, the authors also mounted a camera on a hat worn by the signer in order to create a portable system. The resulting recognition rates were 92% for the desktop camera and 98% for the head mounted camera. Limitations for our application are the

small size of the vocabulary and the quantity of training data required for each signer.

One component of recognizing a sign is recognizing handshape. Somers and Whyte [17] used a hybrid of 3D models and silhouettes to identify the handshape (called "hand posture" in Irish Sign Language). With a set of eight images (2 for each of four handshapes) and a vocabulary of eight handshapes, their approach achieved a classification rate of 50%.

Instead of using a learned vocabulary or 3D model, other researchers have treated the problem as a lookup problem – with the goal of finding the sign in a database of known signs through image similarity. Dimov et al. [5] used a database of known signs, represented by a series of 2D projections, to do a similarity search to recognize alphabet signs. With a vocabulary of seven letter signs and an average of 49 instances of each in the database, the authors achieved a classification rate of 96%. Potamias and Athitsos [16] also used nearest neighbor search of images for handshape recognition. With a set of 20 common ASL handshapes, the accuracy was 33% on 256x256 images across a number of ASL signers.

Given the five components to each sign – handshape, position, palm orientation, motion, and facial expression – using a variety of video features and techniques is required. Caridakis et al. [2] presented an architecture for providing features for hand trajectory, region, and shape to a combination of self-organizing maps, Markov chains, and HMMs for recognition. To the best of our knowledge, the work was not implemented or evaluated so there are no accuracy results or vocabulary estimates.

3.3 Detecting Sign Language

Detecting sign language is a much simpler problem than translating it. As an example, Cherniavsky et al. [1] developed an activity detection technique for cell-phone cameras that could determine whether a user was signing or not with 91% accuracy, even in the presence of noisy (i.e., moving) backgrounds. The algorithm was used to determine when the video phone user was signing and when they were watching the video of their conversational partner in order to effectively use network bandwidth during a sign language conversation on mobile devices. Thus, it is unlikely this algorithm would be as successful in distinguishing between sign language videos and other videos involving people gesturing.

3.4 Limitations and Challenges

A challenge for the above approaches is that most approaches work only modestly with relatively small vocabularies unless they rely on data gloves or other obtrusive equipment, and are single-signer approaches that require large amounts of training data. Signer-dependent solutions are not practical for video classification. Another difficulty is that only a handful of authors (e.g. [9]; [20]) have attempted to recognize signs in sentences or phrases rather than as isolated expressions. 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 SL in the first place.

4. DESIGN OF SL-VIDEO CLASSIFIER

The SL-Video classifier is composed by two components: the first is responsible for video processing and analysis in order to generate video features and the second is responsible for using the features to classify the videos into those that are SL video as we have defined it and those that are not.

4.1 Video processing

Each video is analyzed frame by frame by a video processing subsystem, developed using openFrameworks [12], an open source toolkit that includes video processing functionality.

The first step in extracting video features is to define the background and foreground of each frame image. The background is meant to contain all non-moving elements in the video, while the foreground is the portion of the image that varies from frame-to-frame. Once the foreground and background for the video frames are determined, the results are combined with face detection to compute the video features that are used to classify the video.

4.1.1 Background Modeling

Since the identification of SL video must work with a wide variety of already existing videos, it cannot assume a pre-defined or static background model. Although plenty of SL videos contain a person signing in front of the camera with no background changes, some videos have changes in the background due to lighting changes or moving background objects. So a dynamic background [5][7] is best suited for this task.

Since the goal is to classify if the video contains sign language or not, it is not important to fully identify the signer as a single foreground object. This is good because an additional difference from some video contexts is that the person signing is often already seated in front of the camera at the beginning of the video. This allows a relatively simple dynamic background model without losing information needed by the classifier.

Our background model is built in real time as a running average of the grayscale frames of the video, with a learning rate value which can be adjusted to alter how fast the background model changes over time. Having a high learning rate results is a highly dynamic background model where just the most abrupt movements are detected, while with a low learning rate any slightly change in the image will be detected. For this task a learning rate of 0.04 has proven to be a suitable choice. Thus the background model for a background pixel BP at time t is:

$$BP_t = .96 * BP_{(t-1)} + .04\ P$$

where P is the grayscale value of the pixel at time t. Figure 3 (c) shows the background model for the video at the time shown in Figure 3 (a).

The foreground image is obtained through the subtraction of the current video frame from the current background image. Every pixel where the subtraction result is greater than a threshold (our threshold is 45) is considered as a foreground pixel. Figure 3 (d) shows the results of this process for the frame in Figure 3 (a). To avoid noise in the foreground image that can result from objects moving in the background of the signer and the results of normal body movement, a spatial filter is used to remove small regions of foreground pixels in the image, the result of which is shown in Figure 3 (b). As shown in Figure 3, our background model can contain the whole image, including the signer except for his or her hands and arms. Now with the final foreground model, we can start the feature extraction process.

Figure 3. (a) the incoming frame of the video, (b) the final foreground image, (c) the actual background model and (d) the intermediate foreground image

4.1.2 Feature Extraction

Before calculating the features used for classification, we need to identify the position of the signer's head. This is done through face detection based on Haar-like features [20]. Figure 3 (a) shows a white box around the face location.

Combining the foreground model and the results of face detection, we can calculate the amount of movement in regions relative to the face, indicated by the nine regions shown in Figure 3 (b). This provides information regarding the positions and movements of the two hands relative to the head.

Initially, we computed the quantity of movement in each of the nine regions for each frame of the video as the movement of the hands relative to the body is one of the most unique features on sign language, and probably the easiest to recognize. Because of the quantity of data (9 values for every frame of the video), we further condense the data into five single-value features per video that can be provided to a classifier.

The five features were developed with the intuition that the quantity and location 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). These include features concerning the overall quantity of movement, the continuity of movement, and the location of the movement relative to the face.

The type of SL video which we are attempting to identify has a single signer who signs fairly continuously, resulting in a large amount of movement when compared with other videos of people. With regards to the quantity of movement, we compute two features: (VF1) the number of pixels included in the final foreground model for each frame, averaged across frames, and (VF2) the percentage of pixels that are included in the final foreground model for at least one frame. VF1 is a measure of the total amount of activity in the video while VF2 is a measure of how the spatial distribution of that activity changes over multiple frames. With regards to the continuity of motion, we compute one feature: (VF3) the average difference between the final foreground pixels in one frame and in the previous frame. To differentiate SL video from other videos of fairly continuous human motion we included two features associated with the location of motion: (VF4) the symmetry of motion, measured as the average number of final foreground pixels that are in a symmetric position relative to the center of the face, and (VF5) the percentage of frames with non-facial movement, measured as the average percentage of pixels outside of the facial rectangle that are part of the final foreground. As 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. Similarly, 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. Once computed, these five features are used to classify the video as being SL video or not.

4.2 Classifier

Since the goal of the project is to classify the videos between Sign Language and non-Sign Language, a binary classifier is suitable. We explored several classifiers but chose a Support Vector Machine (SVM) classifier [3] due to its performance compared to other classifiers (e.g., Gaussian classifiers, nearest neighbors) at an early stage in the project. The SVM was trained on a dataset containing SL videos and non-SL videos, each video represented by the five feature values described in the previous section. As illustrated in Figure 4, the SVM classifier works by projecting the original feature vector (5-dimensional) into a higher dimensional space by means of a non-linear mapping. By choosing the mapping carefully, computation in the high-dimensional space can be performed implicitly (i.e., through the so-called kernel trick). Operating in this high-dimensional space also improves the probability that classes become linearly separable.

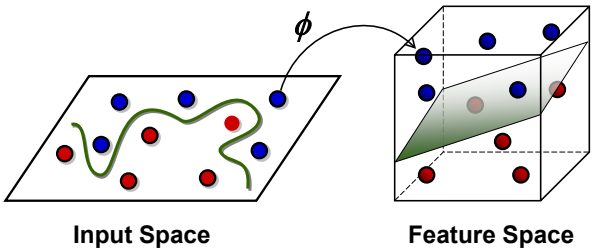

Figure 4. Example of a SVM classifier model, the black line represents the borderland between one class and another

5. EVALUATION OF CLASSIFIER

Evaluation of our approach consisted of developing a corpus of SL videos and non-SL videos, preprocessing this corpus to generate the features described for each video, and using these features to classify the video.

5.1 Developing a Training/Testing Corpus

In order to evaluate the classifier we created a collection of 192 videos, including 98 Sign Language videos (including 78 in American Sign Language and 20 in British Sign Language) and 94 non-Sign Language videos. The videos were selected from video sharing sites like YouTube, Vimeo, etc. Most of these videos were located based on tags or metadata indicating they had some relationship to sign language.

Figure 5. Examples of non-Sign Language videos that are visually similar to sign language videos and thus likely false-positives for the classifier

The majority of the non-Sign Language 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 subset of the non-Sign Language videos were chosen as they included tags or metadata indicating a relationship to sign language (e.g. videos you would likely locate when searching for videos in sign language.) While we found a number of such videos (e.g. Figure 2), we kept the number of videos collected due to tag/metadata confusion low as we found they tend to be visually distinct from the sign language videos and thus not difficult for our classifier. For example, if the video does not include a person for most of the duration, it will be very easy to classify it just based on face detection. Figure 5 shows examples of non-SL videos in the database.

5.2 Processing of Videos

The videos chosen were in the MPEG4 format and were subsampled to 1 minute length each. This time was chosen as it is long enough for feature extraction yet keeps the processing requirements bounded despite the length of the original video. We are currently not looking at cases where a single video includes segments we would consider SL video and other segments of non-SL video.

The subsampled interval for each video was randomly chosen, just assuring that this interval was not on the start of the video or at the end in order to avoid any 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 subsample and the results were stored for use by the various classifiers considered.

5.3 Results

The classifier was tested on 1000 executions for each context; in each execution the training and test data were selected randomly, accordingly to the training set size of the experimental unit.

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 testing set), and the F1 score (a combination of precision and recall). Table 1 shows the results for different training set sizes; in all cases, examples not included in the training set were used for testing.

Table 1. Results obtained when 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

As the number of videos used to train the classifier is increased, the precision stays relatively stable (irregularly varying within a 3% band) but recall increases by more than 4%. This indicates that while more training data improves the classifier, it works well with only 15 videos per training group.

Given this result, we explored the relative value of the five visual features. As a reminder, the five video features are:

VF1: the number of pixels included in the final foreground model for each frame, averaged across frames

VF2: the percentage of pixels which are included in the final foreground model for at least one frame

VF3: the average difference between the final foreground pixels in one frame and in the previous frame

VF4: the symmetry of motion, measured as the average number of final foreground pixels that are in a symmetric position relative to the center of the face

VF5: the percentage of frames with non-facial movement, measured as the average percentage of pixels outside of the facial rectangle that are part of the final foreground

We first explored the performance of the classifier when we remove each of the features. We again use 15 videos per training class and provide all but one of the videos features as input. Table 2 presents the results.

Table 2. 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

The results show that the feature that added the most discriminating power to the classifier when compared to the other features is VF4, a measure of the symmetry of motion relative to the face. Without VF4, the precision dropped almost 9% and recall dropped more than 12% from the performance of the classifier with all five features. There was not a strong effect from dropping any of the other four features, implying they may overlap in the type of information they are providing 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.

Finally, we explored which single visual feature provided 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 are shown in Table 3.

When comparing the ability of a single video feature to classify SL video the results again indicate the best predictor is VF4. The difference between this feature and the other four is significant; Feature 4 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.

Table 3. 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

As observed from the results, with a good feature selection, the SVM is successful at classifying the majority of videos as being either SL video or not, even with small training sets. Given the non-SL videos were selected to be as similar to sign language video as possible, 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.

5.4 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 that are difficult to classify are shown in Figure 6. We already have 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 6 (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 6 (b). Additionally, signing in front of backgrounds with lots of movement (Figure 6 (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 colors that match the skin tone or shirt color of the signer, such as the couch in Figure 6 (d).

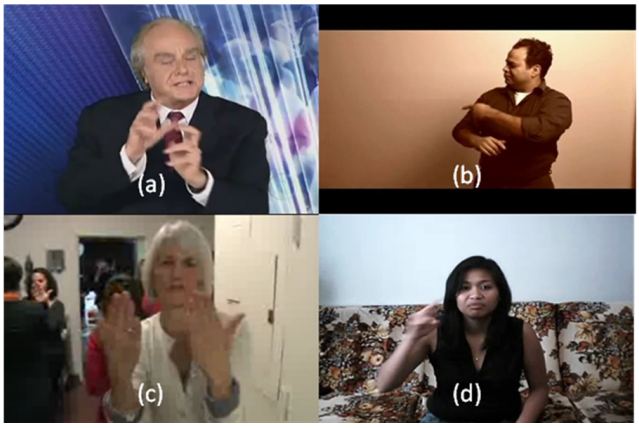

Figure 6. Examples of videos which are difficult to successfully classify with current approach.

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

Our initial success leads to a variety of directions for future work. First, we want to improve our background modeling process to increase the accuracy of the final foreground blobs being the hands and arms of the signer for SL video. Additionally, we hope to develop a more adaptive background model, capable of additional video situations, such as videos with more than one signer, videos with the signer in different positions relative to the camera, etc. Similarly, we want to improve on the use of face detection. In particular, we can use the knowledge that a face was detected in a prior frame and there is no reason to believe the person has moved or the shot has changed to infer the position of the head in subsequent frames.

We also are attempting to identify additional meaningful video features that will increase the SVM performance. While we compared different types of classifiers early on in the project, we plan to examine the performance of other classifiers (e.g. Neural Networks) with the video features we have since developed.

Additionally, we plan to develop a larger collection of SL video and non-SL video to increase the variety of videos being used for training and evaluation. Finally, we plan for the current approach to serve as a starting point to more complex tasks in this area, such as attempting to classify SL videos based on which sign language is found in the video (American Sign Language, British Sign Language, etc.) and to identify videos that have sign language translation within a region of the video.

7. CONCLUSIONS

YouTube, Vimeo and other general purpose video sharing sites are being used to share sign language presentations among the sign language community. Currently, pointers to these videos are emailed or otherwise communicated from person to person. To locate such videos using the search facilities provided by the sites requires the existence and accuracy of tags or other metadata indicating a relationship to sign language.

We have presented an approach to classifying videos as being sign language videos or not without any previous information about them. A SVM classifier was provided with five video features identified as potentially valuable for this process. The extraction of the five features relies on relatively simple background modeling and face detection.

A collection of videos for training and testing the classifier was created by selecting SL videos and non-SL videos that were likely false positives from video sharing sites. Our evaluation showed that training the classifier does not require large quantities of training data – while the classifier improved with more examples, 15 examples for each category were sufficient to have greater than 81% precision and 86% recall.

Comparison of the five video features showed that a measure of the symmetry of motion relative to the center of the face was the most accurate feature when used alone for classification. Alone it was more accurate than using the other four features combined.

The goal of this capability is to increase access to sign language presentations for members of the sign language 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 sign language videos.

Our future work looks to improve on the current classifier by improving the video processing techniques used for feature extraction and by identifying alternative features. Further, we plan to apply the classifier to new settings, such as identifying sign language translation in a region of a video and for identifying which sign language is being used in an SL video.

8. ACKNOWLEDGMENTS

This material is based upon work supported in part by the National Science Foundation under Grant No. DUE 09-38074, by a gift from Microsoft Corporation, and by NPRP grant # [08-125-2-03] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

9. REFERENCES

[1] N. Cherniavsky, R.E. Ladner, and E.A. Riskin, "Activity detection in conversational sign language video for mobile telecommunication", *Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition (FG '08),* 2008, pp.1-6.

[2] G. Caridakis, O. Diamanti, K. Karpouzis, and P. Maragos, "Automatic Sign Language Recognition: Vision Based Feature Extraction and Probabilistic Recognition Scheme from Multiple Cues", *PETRA '08: Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments*, 2008.

[3] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods, First Edition*. Cambridge University Press, 2000.

[4] O. Crasborn, J. Mesch, D. Waters, A. Nonhebel, E. van der Kooij, B. Woll and B. Bergman, "Sharing sign language data online. Experiences from the ECHO project," *International Journal of Corpus Linguistics 12(4),* 535–562, 2007.

[5] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. "Detecting moving objects, ghosts and shadows in video streams". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, pp. 1337-1342.

[6] D. Dimov, A. Marinov, and N. Zlateva, "CBIR Approach to the Recognition of a Sign Language Alphabet", *CompSysTech '07: Proceedings of the 2007 International Conference on Computer Systems and Technologies*, 2007.

[7] Brian Gloyer. "Video-based freeway-monitoring system using recursive vehicle tracking". *Proceedings of SPIE*, 1995, pp. 173-180.

[8] M. Heckner, T. Neubauer, and C. Wolff, "Tree, funny, to_read, google: What are Tags Supposed to Achieve?", *Proceedings of the 2008 Workshop on Search in Social Media*, 2008, pp. 3-10.

[9] J.L. Hernandez-Rebollar, "Gesture-Driven American Sign Language Phraselator", *ICMI '05: Proceedings of the 7th International Conference on Multimodal Interfaces*, 2005, pp. 288-292.

[10] 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.

[11] 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.

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

[13] C. Marshall, "No Bull, No Spin: A Comparison of Tags with Other Forms of User Metadata", *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 2009, pp. 241-250.

[14] C. Neidle, S. Sclaroff, and V. Athitsos. "SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data," *Behavior Research Methods, Instruments, and Computers, 33:3*, pp. 311-320, 2001.

[15] C. Padden, "The Deaf Community and the Culture of Deaf People", *Readings in Diversity and Social Justice*, edited by M. Adams, W. Blumenfeld, H. Hackman, M. Peters, and X. Zuniga, 2000, pp. 343-352.

[16] M. Potamias and V. Athitsos, "Nearest Neighbor Search Methods for Handshape Recognition", *PETRA '08: Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments*, 2008.

[17] G. Somers and R.N. Whyte, "Hand Posture Matching for Irish Sign Language Interpretation", *ISICT '03: Proceedings of the 1st International Symposium on Information and Communication Technologies*, 2003, pp. 439-444.

[18] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 1998, pp. 1371-1375.

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

[20] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features", *CVPR '01: Proceedings of the 2001 IEEE Computer Society Conference*, pp. 511-518, 2001.

[21] C. Vogler and D. Metaxas, "Towards Scalability in ASL Recognition", *Proceedings of Gesture Workshop '99*, 1999.