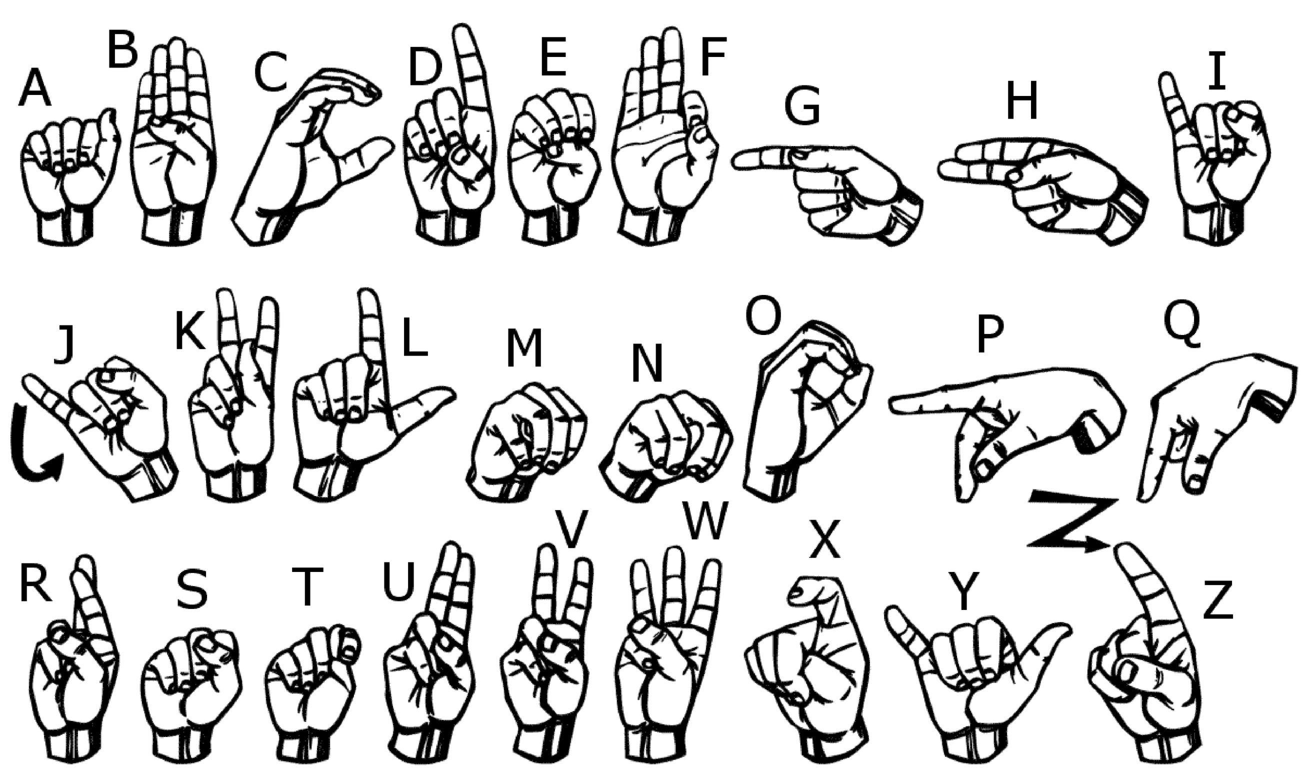
American Sign Language Recognition System

Jason Atwood, Justin Farrell
Carnegie Mellon University, Robotics Institute

Abstract

Vision-based gesture recognition shows promise for many human-computer interaction applications [1]. One such application is sign language translation, which can be used as an interactive educational tool or simply to help a hearing-impaired person communicate more effectively with someone who does not know sign language. For this project, the team created an American Sign Language (ASL) recognition system which classifies sign gestures captured by a webcam in real time. For the scope of this project the team focused on the static signs for letters of the alphabet.

Introduction



A graphical user interface was created to capture training data and perform classification on live video. In the current approach, the algorithm tracks the user's hand using mean-shift, places a bounding box around the hand, extracts SIFT features from the hand in the bounding box, and performs SIFT matching to classify the test image.

Prior work performed by the group includes two alternate approaches to this recognition problem, PCA classification with grayscale pixel intensities, and neural network classification with binary pixel intensity [2]. Results for static-image tests are presented for comparison to the current SIFT descriptor approach in the results section.

Both previous methods were reasonably successful recognizers on static-images, but video recognition was difficult because neither method is invariant to scale, rotation, or orientation. The previous work required the user to align their hand at exactly the same position and orientation as the training images for accurate detection in video. The motivation for this work is to overcome these issues.

PROPOSED METHODS/DESCRIPTION

Tracking

The camera capture window, as with many vision challenges, contains non-important features. The scene background and user's head and body interfere with proper sign recognition. To constrain the feature descriptors to just the user's hand, a dynamic bounding box tracks the hand. Tracking is accomplished by a mean-shift tracker, computed in the five dimensional (x,y,R,G,B) space over a 120 x 80 pixel window. Mean shift tracking was chosen over other techniques for its ability to track non-rigid objects, in this case a continually deforming hand.

Feature Description

Sign recognition in live video presents several additional challenges due to changes in scale, rotation, orientation, and illumination. Scale Invariant Feature Transform (SIFT) descriptors were selected to provide a robust feature-set describing the hand sign. SIFT works by localizing scale, orientation, and rotation invariant keypoints and describing the region around them with histograms of gradients. This results in features which are invariant to uniform scale, rotation, orientation, and partially invariant to illumination [3].

Feature Classification

Given a set of labeled training images, new test images are classified using SIFT matching. A set of 128 dimensional SIFT descriptors are extracted from each training and test image. A descriptor in one image is considered a 'match' with a descriptor in another image, if the Euclidean distance between the two 128 dimensional vectors divided by the distance between the test vector and the next closest training vector is above a threshold, which is set to 0.8 here [3]. The test image is given the same classification as the training image that shares the greatest number of matches.

1. Wu Y and Huang TS. Vision-Based Gesture Recognition: A Review. *Lect Notes Comput Sci* 1739, 1999.
2. J. Atwood, M. Eicholtz, and J. Farrell, "American Sign Language Recognition," Dept. Mech. Eng., Carnegie Mellon Univ., Pittsburgh, PA, Project Report., May. 2012.
3. David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* 60, 2 (2004), pp. 91-110

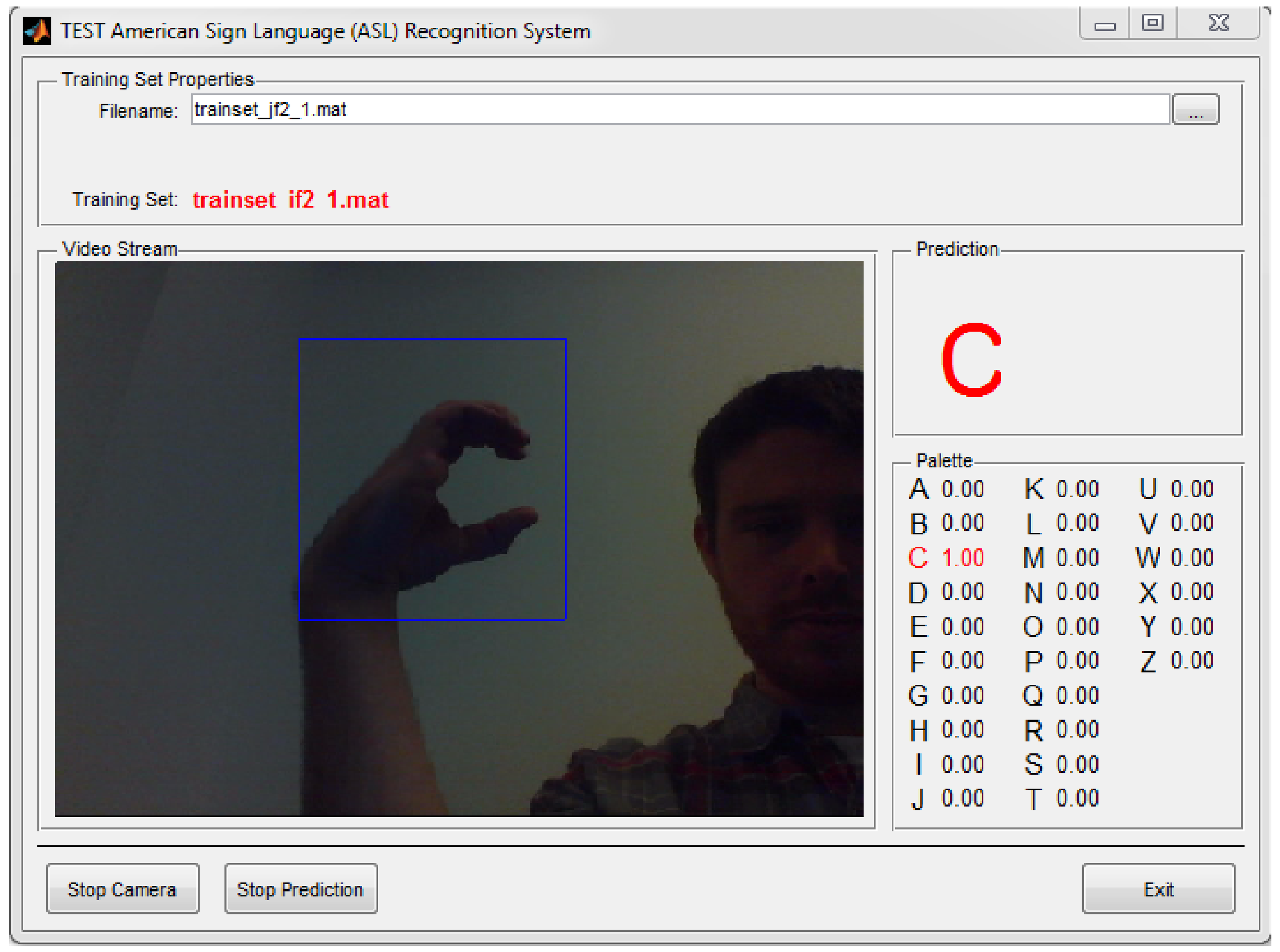
RESULTS

Static Testing

To evaluate the recognition system, classification is performed on a set of static images. Since the test is performed on static images and not live video, the tracking effects are not taken into account. The test contains 360 unlabeled test images and 1,080 labeled training images, from three different users. Each user in the test set has data in the training set as well. The results for the SIFT implementation is presented with comparison to previous methods.

	SIFT	Neural Network	PCA
Percent Accuracy	**99.2%**	96.1%	95.8%

When a test user is not represented in the training data, the recognition accuracy drops dramatically to 35-45%. Because new users have different size/shaped hands compared to training data, the relation to existing training data is no longer uniform scaling, but non-uniform scaling. Unfortunately SIFT is not invariant to this non-uniform scaling.



CONCLUSIONS

SIFT descriptors provide a robust feature set for recognition problems in video where invariance to uniform scale, rotation, and orientation are important. The features are suitable in the application of sign language recognition if a user is part of the training set. Under this constraint, the recognition system outperforms implementations of PCA and neural networks in previous work [2].

Mean shift tracking is well suited for tracking continually-deforming human hands. The method is robust to scale and orientation variations.