

# Vision-Based Static Hand Gesture Recognition Using Support Vector Machines

S. Naidoo, C.W. Omlin

Department of Computer Science  
University of the Western Cape  
7535 Bellville, South Africa  
{snaidoo, comlin}@uwc.ac.za

M. Glaser

Division of Communication Sciences and Disorders  
University of Capetown  
7701 Rondebosch, South Africa  
mglaser@uctgsh1.uct.ac.za

**Abstract**— Given that advances in technology have increased rapidly over the past few years, developing a vision-based recognition system that could aid in the interaction between the Deaf and the hearing does not sound that far-fetched anymore. We can distinguish between two categories of hand gestures, namely static and dynamic. A static gesture is a particular hand configuration and pose represented by a single image. A dynamic gesture is a moving gesture, represented by a sequence of images. This makes signed language a good test case for the classification of gestures, since each gesture is assigned a particular meaning. This paper will focus on the development of a system that will recognise static hand images against complex backgrounds based on South African Sign Language (SASL). A Support Vector Recognition system will be used to classify hand postures as gestures, due to its high generalization performance without the need to add a-priori knowledge, even when the dimensions of the input space is very high.

**Index Terms**— Support Vector Machines, Local color histograms, HSV, South African Sign Language (SASL)

## I. INTRODUCTION

As a means of communicating with computers, visual communication can be significantly simpler than the more familiar devices such as keyboards, mouse devices, keypads or light pen. These devices have become quite common, but inherently limit the speed and naturalness with which we interact with the computer. It has become increasingly evident that the use of gesture recognition systems could aid in the interaction between man and machines, thus enabling machines to become more user-friendly and allowing users a richer man machine experience.

Since vision-based systems contain rich visual information which is a strong cue to infer the inner states of an object, they can provide a promising alternative to the recognition of human gestures. At the same time, vision-based systems can be very cost efficient and noninvasive, making vision systems very feasible.

Recently, there has been a surge in interest in gesture recognition, more specifically facial and hand gestures. Potential applications that use hand gesture recognition include computer games and machinery control (eg. a crane). One of the most structured sets of gestures belong to signed languages, since each sign is assigned a particular meaning. It is for this reason that signed language recognition can pose as a very good candidate for vision algorithms. We can distinguish

between two categories of hand gestures, namely static and dynamic. A static gesture is a particular hand configuration and pose, represented by a single image. A dynamic gesture is a moving gesture, represented by a sequence of images.

Natural signed languages are fully-fledged languages entirely capable of expressing all the nuances of meaning that spoken languages can express. South African Sign Language (SASL) is the primary language used by Deaf South Africans, whose number is estimated to be anywhere between 500,000 to 1,5 million people [1]. SASL has evolved naturally out of a need for communication in the Deaf community and has stabilised with continued use.

SASL has the same complex structure as spoken language including phonology, morphology, syntax and pragmatics, albeit produced in a different mode. Spoken languages are oral-aural languages using sound to convey meaning, whereas signed languages are visual-gestural using space to convey the meaning. Signed languages are neither universal, nor based on a spoken language. SASL uses the hands, body, head, eyes and face to communicate the full range of language functions. Signed languages have a set of basic phonological features, called cheremes, which make up each sign [2]. These are the handshape, the movement, the location and the palm orientation of the signer's hand. Signs are contrasted with each other by a change in any of these parameters. Signed language morphology including inflectional and aspectual information, is exhibited spatially, as is the syntax of SASL.

Given that advances in technology have increased rapidly over the past few years, developing a vision-based recognition system that could aid in the interaction between Deaf and hearing people does not sound that far-fetched anymore. Thus, it is the purpose of the South African Sign Language (SASL) project to work towards a system that will achieve this goal of signed language recognition. This paper will focus on the development of a system that will recognise static hand images based on South African Sign Language. Creating a SASL dictionary is not the desired result.

## II. RELATED WORK

The problem of hand gesture recognition has been approached with different techniques in the last few years. The vast majority of the work has used mechanical devices to sense the shape of the hand. This type of approach is known as glove based analysis: sensors are attached to a glove that transmits electromagnetic signals for determining the hand posture. Additional sensors can be used to determine the relevant position of the hand. The most common device is a thin glove mounted with binding sensors of various kinds, such as the DataGlove [3]. The shape of the hand was classified by receiving joint angle information from the DataGlove and indexing it into lookup tables. A recognition accuracy of 96% on a random sample was achieved [4], recognising ten signs using an instrumented glove and a large neural network consisting of 403 nodes. These systems provide a large amount of robustness during feature extraction, but are limited to the use of sophisticated and sometimes expensive devices. Also applications that use glove based analysis encounter a large range of problems including reliability, accuracy and electromagnetic noise.

Another method used in the recognition of hand gestures is that of vision analysis. This approach is based on the way human beings perceive information about their surroundings. Visual sensing has the potential to make gestural interaction more practical, but potentially embodies some of the most difficult problems in machine vision. The hand is a non-rigid object and, even worse, self-occlusion is very usual. The Elastic-Graph Matching technique has been employed [5] to classify hand postures against complex backgrounds. Hand postures were represented by labelled graphs with an underlying two-dimensional topology. Attached to the nodes were jets. A jet is simply a local image description based on Gabor filters. This approach can achieve scale-invariant and user-independent recognition, and it does not need hand segmentation. Since using one graph for one hand posture is insufficient, this approach is not view-independent. The recognition rate achieved against complex backgrounds was 86%.

A teleoperation system in which an articulated robot performed a block pushing task based on hand gesture commands sent through the internet was developed [6]. A Fuzzy C-Means clustering method was used to classify hand postures as gestures. The fuzzy recognition system was tested using 20 trials each of a 12 gesture vocabulary. Results revealed a recognition rate (i.e. the ratio of unclassified gestures to classified gestures) of 96% and a recognition accuracy (i.e. the percent of the classified gestures recognised correctly) of 100%. The Fuzzy C-Means (FCM) recognition system was used because of its speed in recognising gestures with sufficient accuracy in real-time operation.

## III. RECOGNITION SYSTEM

Combining relevant low level feature extraction and Support Vector Machines (SVM) can significantly increase the robustness of hand gesture recognition in real-time. Robust in the sense that the system will be user independent (i.e. once the SVM is trained, it would be able to classify various hands as a valid or invalid hand sign) and no restriction to the use of rings on fingers.

This research will evaluate the reliability and efficiency of Support Vector Machines for the real-time classification of hand signs against complex backgrounds (i.e. various objects in the background). Not only will the SVM classify, but also simultaneously learn discriminating features by way of nonlinear discriminant analysis [7]. This can be achieved by mapping the original input space to a feature space in which multiple discriminant analysis can be performed. The design of the SVM classifier architecture is very simple and mainly requires the choice of the kernel. Thus, care will be taken when selecting a kernel since an inappropriate kernel can lead to poor performance. We will also make the assumption that the background will remain constant during the feature extraction phase. The system will be broken up into two phases, a preprocessing phase and a classification phase. During the preprocessing phase features will be extracted from images in the form of a feature vector. The feature vector will then be presented to a Support Vector Machine for classification.

## IV. FEATURE EXTRACTION

Features will be extracted from color images. These images are real-world images of people signing. The images will be acquired using a digital camera. A compressed avi file will then be generated. Once the avi file is generated, a frame grabber will be used to create bitmap images of the video. We assume the sizes of the images in the database are variable to  $h \times w$  ( $h$  for the height and  $w$  for the width). The images will then be segmented into blocks of  $n \times m$ . The features to be extracted from each block in the image include color histograms of pixel values, giving an  $n \times m$  size feature vector.



Fig. 1. Static hand sign

### A. Hue, Saturation, Value (HSV)

Color is a particularly important part of most visualizations. A color space or color model has the purpose of providing a standard for people to identify, classify and visualise color. A few existing color models are RGB, HSI, HSV. The most commonly used color model to represent color images is Red, Green, Blue (RGB). However, the Hue, Saturation, Value (HSV) will be considered, since it separates the color component (HS) from the luminance component (V) and is less sensitive to illumination changes.

## B. Local Color Histograms

Local Color histograms will be used to represent the color feature. This is because the features to be extracted focus on low-level, high-dimensional translation and rotation invariance. Global feature extraction (i.e. averaging over the whole image) are mostly translation invariant. It is for this reason that local histograms will be used (i.e. the image will first be divided into blocks. Then from each block histogram representation will be taken as a feature). This way, the degree of translation can be controlled by the number of blocks into which the image is divided. The Local Color histogram technique is a very simple, low level method and has shown good results in practice [8], where feature extraction has to be as fast as possible.

## V. SUPPORT VECTOR MACHINES (SVM)

Support Vector Machines (SVM) is a linear machine with some very good properties. The main idea of a SVM in the context of pattern classification is to construct a hyperplane as the decision surface. The hyperplane is constructed in such a way that the margin of separation between positive and negative examples are maximized. The SVM uses a principled approach based on statistical learning theory know as structural risk minimization, which minimizes an upper bound on the generalization error (i.e. a parameter estimation heuristic that seeks parameter values that minimize the "risk" or "loss" that the model incurs on the training data).

### A. Optimal Separating Hyperplane

An Optimal Separating Hyperplane can be defined as the one with the maximal margin of separation between two classes (see Fig. 2.) and has the lowest capacity. In order to determine the Optimal Separating Hyperplane, we must first assume that the class is represented by the subset  $d_i = +1$  and  $d_i = -1$  are "linearly separable". We then find the weight vector  $\mathbf{w}$  and bias  $b$  so that

$$y_i(w \cdot x_i + b) > 0, i = 1, \dots, N \quad (1)$$

If there is an hyperplane that can satisfy Eq. (1), the set is said to be linearly separable. Once a linearly separable hyperplane is obtained, it is possible to rescale the weight  $\mathbf{w}$  and bias  $b$  such that the closest point to the hyperplane has a distance of  $1/\|\mathbf{w}\|$ . Eq. (1) can now be shown as

$$y_i(w \cdot x_i + b) \geq 0 \quad (2)$$

From all the separating hyperplanes, the one with the distance to the closest point is maximal is called the optimal separating hyperplane (OSH) and the quantity  $2/\|\mathbf{w}\|$  is called the margin of separation denoted by  $p$  (see Fig. 2.).

### B. Linearly non-separable case

Given a training set that is not linearly separable, it is not possible to construct a hyperplane that can separate the classes without coming across any classification errors. However, it is

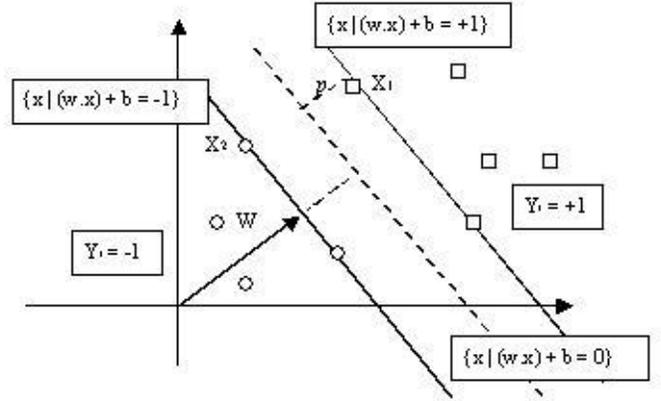


Fig. 2. Optimal Separating Hyperplane

still possible to find an optimal hyperplane that minimizes the probability of the classification error, averaged over the training set. In order to perform this task we need to introduce a slack variable  $\epsilon_i$  that measures deviation of a data point from its ideal position of linear separability as shown below :

$$y_i(w \cdot x_i + b) \geq 1 - \epsilon_i, i = 1, \dots, N \quad (3)$$

### C. Multi-Class Learning

One way to solve the K-class pattern classification problem is to extend the basic SVM formulation for binary classifiers to cover the K-class problem. There are different ways in which this can be done, these include:

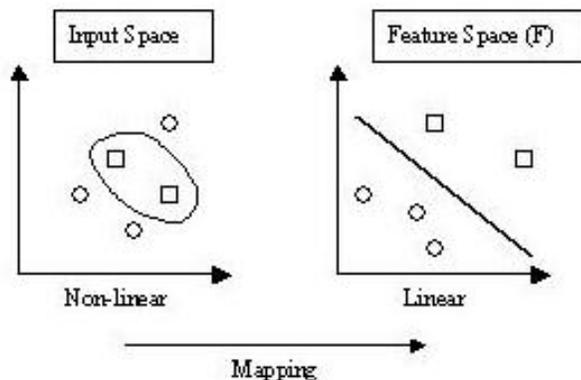
- Modifying the design of the SVM by constructing a decision surface for each class [9]. The optimization problem in the SVM formulation can be generalised to consider all the decision functions at once. In addition, the decision function does not need to give a zero value at the decision boundary. It is relaxed so that the output of the decision function for the correct class is greater than the rest of the decision function by a margin of 2.
- Combining many binary classifiers: "One against one". The classes are compared in a pairwise manner.
- "One against other" on the other hand, compares a given class with all the other classes combined.

### D. Nonlinear Support Vector Machines

Pattern classification consists of a multiclass nonlinear set of input data. As a result a nonlinear support vector needs to be implemented. This occurs by converting the nonlinear input space to a high-dimensional linear feature space via a nonlinear mapping and then apply linear algorithms within the high-dimensional feature space. For this to happen, we replace  $\mathbf{x}$  with its feature space mapping  $\Phi \mathbf{x}$  so we now get:

$$k(x, y) = (\Phi(x) \cdot \Phi(y)) \quad (4)$$

This is evaluated using dot product as shown above, but can be computationally large. Thus, it is the choice of kernel  $k$  that is important, so that large quadratic computation can be reduced. The figure below shows a mapping of the nonlinear input space to a linear feature space.



There are various kernels to choose from, for example Polynomial learning machines, Radial-basis function networks and Two-layer perceptrons. The choice of the kernel is theoretically of interest since it determines the feature space in which to work.

## VI. EXPECTED RESULTS

Current attempts to recognise hand shapes from images achieved fairly good results, but this is mainly due to either very high computation and the use of specialized devices. One such attempt is that of Pavolovic, Sharma and Huang, who have achieved good recognition rates using 2-D blob representations of the hand [10]. The aim of this system is to achieve relatively good results but at the same time a trade off must be considered between time and accuracy. However we will aim to achieve accuracies above 90%.

## VII. CONCLUSIONS

In this paper we propose a method of classifying static hand gestures using Support Vector Machines where the only features are that of low-level computation such as local color histograms. There are many Machine Learning algorithms that have been developed to choose from, for example neural networks and classification trees, in order to perform classification. However Support Vector Machine approach will be considered as it is a good candidate. This is due to its high generalization performance without the need to add a-priori knowledge, even when the dimension of the input space is very high. Chapelle, Haffner and Vapnik [11] showed that Support Vector Machines (SVM) can generalise well on difficult image classification problems where the only features are high dimensional histograms. Heavy-tailed RBF kernels were shown to far outperform traditional polynomial or Gaussian RBF kernels. The major goal of this research is to develop a system that will aid in the interaction between deaf and hearing communities. Vision-based systems provide a promising alternative for gesture recognition research.

## ACKNOWLEDGMENTS

This research has been funded by the Telkom-Cisco Centre of Excellence for IP and Internet Computing and the South African National Research Foundation.

## REFERENCES

- [1] D.Aarons, P.Akach. South African Sign Language - one language or many? A sociolinguistic question. Stellenbosch papers in Linguistics (SPIL), 31, 1-28. 1998.
- [2] C.Valli and C.Lucas. Linguistics of American Sign Language: A resource text for ASL users. Washington, D.C.: Gallaudet University Press. 1992.
- [3] Dataglove model 2 operating manual. VPL Research Inc., 1989.
- [4] K.Murakami, H.Taguchi: Gesture Recognition using Recurrent Neural Networks. In CHI '91 Conference Proceedings (pp. 237-242). ACM. 1991.
- [5] J.Triesch, C.Malsburg: Robust Classification of Hand Postures Against Complex Background, Intl Conf. On Automatic Face and Gesture Recognition, 1996.
- [6] J.Wachs, U.Kartoun, Y.Edan, H.Stern: Real-Time Gesture Telerobotic System Using the
- [7] V.Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.
- [8] M.Swain and D.Ballard, "Indexing via color histograms", Intern. Journal of Computer Vision, vol. 7, pp. 11-332, 1991.
- [9] i) V.Pavolovic, R.Sharma and T.Huang. Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19, 677-695.
- [10] j) O.Chapelle, P.Haffner, V.Vapnik. SVMs for Histogram-Based Image Classification.
- [11] k) T.M.Mitchell. Machine Learning. The McGraw-Hill Companies, Inc. 1997.