MVA2011 IAPR Conference on Machine Vision Applications, June 13-15, 2011, Nara, JAPAN

Hand Gesture Recognition using Histogram of Oriented Gradients and Partial Least Squares Regression

Arindam Misra
Indian Institute of
Technology, Roorkee

Abe Takashi
Tohoku University,
Japan

Takayuki Okatani
Tohoku University,
Japan

Koichiro Deguchi
Tohoku University,
Japan

Abstract

In this paper we propose a real-time hand gesture recognition system that employs the techniques developed for pedestrian detection to recognize a small vocabulary of human hand gestures. Our feature set comprises of grids of Histogram of Oriented Gradient (HOG) descriptors, with fine orientation binning and multi-level spatial binning for getting descriptors at the small as well as large scale. The overlapping descriptor blocks, which are contrast normalized to handle illumination changes, have a high degree of multicollinearity, resulting in a feature set of high dimensionality (more than 8000 dimensions), rendering it unsuitable for classification using the classical machine learning algorithms. Thus, we employ Partial Least Squares (PLS) regression as a 'class aware' method of dimensionality reduction, to project the feature vectors on to a lower dimensional space of 10 dimensions. We examine the results obtained by PLS as well as Principal Component Analysis (PCA) which show, that PLS outperforms PCA, and gives a better projection which preserves significant discriminative information.

1. Introduction

The recognition of human hand gestures by computers can create a new Human Machine Interface which could be used for communicating with the computer in a more natural manner than giving commands on keyboards or using mice. It finds applications in a large number of areas like sign language recognition, robotics, controlling home appliances and the gaming experience can be enhanced if the games could be controlled by hand gestures [1]. In this paper we present a vision-based method that detects hand gestures in different illuminations and cluttered backgrounds, independent of the skin colour of the user. The proposed method, which can run on a workstation in real-time, enables the user to interact with the computer using a USB camera.

Many techniques like [2][6] have been proposed in the past which employ edge and gradient based descriptors for hand gesture recognition. They are able to detect hand gestures only in a simple background and are liable to fail when the background is cluttered. Dalal and Triggs [3] have studied the question of feature sets for robust visual object recognition using HOG descriptors and have shown that HOG features significantly outperform existing feature sets, taking human detection as a test case. A detailed discussion on HOG features can be found in [3].

The HOG descriptors have been obtained using different block sizes on the same image, and concatenating these features to get the final image descriptors. The blocks are contrast normalized to make the descriptors free from illumination changes. The HOG features are computed several times for each cell in the image, resulting in multiple contributions to the final descriptor, with each cell being normalized with respect to a different block.

This redundancy adds to the dimensionality of the descriptors, which calls for a need to *reduce the dimensionality* in order to apply the classical machine learning algorithms like the k-nearest neighbor search, in real-time.

We have used Partial Least Square regression technique for dimensionality reduction as it models relations between a set of observations by means of latent variables, and is aware of the classes into which the observations are classified, the details of which can be found in [4]. The results indicate that PLS outperforms PCA in terms of classification of the training data into various hand gestures.

The plots of the classifications for both the dimensionality reduction techniques clearly point out PLS as the preferred method of reduction. Moreover, PLS has a lower execution time than PCA which saves time in the learning phase [5]. For real-time requirements and simplicity we have used the k-nearest neighbor search to classify the query points in the feature space [7].

We have developed three databases with varying degree of positional variations and backgrounds with each having 3500 images, 500 images per gesture. The results indicate the clear tradeoff between the positional variations in the training data and the classification in the feature space. We were able to achieve satisfactory performance for test images with small positional variations which were tested on classifiers which had been trained on images having small positional variations.

Our implementation performs a near perfect classification of test images with no positional variation. The performance degrades as the positional variations increases, the classifiers trained on images with high positional variations although perform reasonably well for test images with high positional variation, but sometimes fail on images with no position variation, PCA gives better results for this case but falters as compared to PLS, on the test images having no positional variation.

2. Overview of the proposed method

The proposed method uses the edge and gradient based techniques developed for human detection for the problem of hand sign recognition. Similar features have been used in works like [2][3][6], researchers in [9] have used an array of moving spots to recognize hand gestures, [8] presents a glove free solution to this problem. However, most of these works focus on recognition in controlled environments and may not perform when tested 'in the wild'.

HOG descriptors characterize the articulated gestures by the distributions of local intensity gradients. The feature extraction begins with the gradient computation for all the pixels of the image, with the largest of the gradient of three channels chosen as the gradient of the pixel. Each 'cell' in the image has a histogram which is constructed using the directions and the magnitudes of pixel gradients in the cell. The features are accumulated over a block and are normalized.

As the number of hand gestures to be classified are many, rather than the human or non human classes in pedestrian detection, the macro features have to be included by taking many block sizes for the same image and concatenating them to get the final descriptor. The descriptors are then projected on a lower dimensional space for classification using the k-nearest neighbor search. Once the classifiers are trained, the test images are used to anticipate the performance of the system.

The database was generated by taking images of the various hand gestures with a varying degree of positional variations, in front of a blue background and chroma-keying these images on various backgrounds taken from the internet. Similar strategy was used for generating the test images; however the hand images and backgrounds used were different for the test images than those which had been used for the training purpose.

3. Partial Least Square Regression

PLS constructs latent variables as a linear combination of the original variables in the matrix X containing the feature vectors of the training images and a vector Y containing the class labels of the images.

$X \subset R^N$ an N-dimensional space of variables represent the feature vector matrix and similarly Y$\subset R^M$ an M-dimensional space representing the class labels. PLS models the relationship between the two by means of score vectors. After observing n data samples from each block of variables PLS decomposes the $(n \times N)$ matrix of zero-mean variables X and the $(n \times M)$ matrix of zero-mean variables Y into the form

$$X = TP^T + E \qquad (6)$$

$$Y = UQ^T + F \qquad (7)$$

where the T, U are (n×p) matrices of the p extracted score vectors, the (N × p) matrix P and the (M × p) matrix Q represent matrices of loadings, and the (n × N) matrix E and the (n × M) matrix F are the matrices of residuals. The PLS method in its classical form is based on the nonlinear iterative partial least squares (NIPALS) algorithm.

The difference between PCA and PLS is that the latter creates orthogonal weight vectors by maximizing the covariance between the elements in X and Y. Thus, PLS not only considers the variance of the samples but also considers the class labels, making it a supervised method of dimensionality reduction.

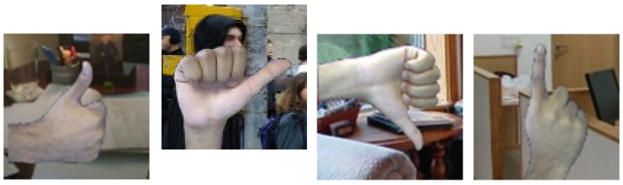

Figure 2. Various test images

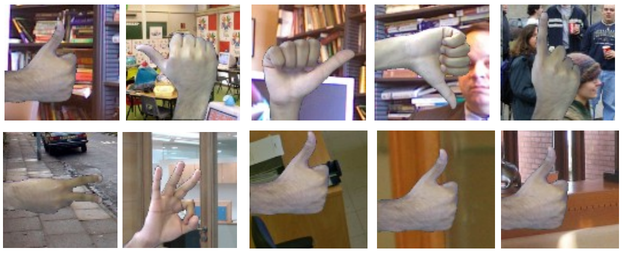

Figure 1. Various training images with positional variations.

4. Experimental Setup and Datasets

The database was developed by capturing images of the seven hand gestures to be recognized in uniformly colored backgrounds and chroma keying the images on various backgrounds obtained from Google images. As the classifier is desired to be mainly used indoors in home or office space, we collected mostly the backgrounds in the office or home environment. Image sets with varying degree of positional variations were captured and were used to generate the training images for the classifiers.

The images consisted of sets with no positional variation, slight positional variation and large positional variation. For generating the test images also, separate sets with similar attributes and different backgrounds were used. The training as well as test images were 100×100 pixels in resolution and were resized after chroma keying from the 640×480 resolution. The various test and training images have been shown in the Fig 1& 2.

The three training image sets were used to test the effect of multi-level HOG descriptors on the performance of the classifier. One set was trained using 5×5 blocks of 10×10 cells and 5×5 blocks of 20×20 cells. The other descriptors had additional descriptors of 4×4 blocks of 25×25 cells, 2×2 blocks of 50×50 cells and a single cell of 100×100. The results have been expressed in the form of confusion matrices which indicate the accuracy of detection. In addition to this to test the performance of PCA we reduced the dimensionality using PCA as well as PLS for the classifiers trained using the three set of images, which were all tested with the tes [illegible] al variations, the results of which have also been shown in the form of confusion matrices.

Table 1. Confusion matrix for classifier trained on images with slight positional variation and tested on images with large positional variation, using only low-level descriptors.

	Identified Class							
		1	2	3	4	5	6	7
Actual Class	1	77	2	0	0	5	16	0
	2	44	42	0	0	12	2	0
	3	49	20	24	0	0	7	0
	4	37	4	0	20	10	29	0
	5	30	13	0	0	55	2	0
	6	22	1	0	0	2	75	0
	7	37	4	0	0	58	0	1

The code for Hand Gesture Recognition was executed on a 64 bit Intel Core i7 machine with a maximum clock speed of 2.8 GHz, with 4GB RAM and a USB camera with 640×480 resolution running on Windows 7 Enterprise 64 bit version. The development environment used was Visual Studio 2008 and the 64 bit version of OpenCV 2.0 was used.

5. Results

The results obtained for the analysis of various block-sizes for HOG features indicate the importance of using multi-level HOG features for enhanced perfor-mance. If only 5×5 blocks of 10×10 cells and 5×5 blocks of 20×20 cells are used for HOG descriptors the classifier performs very poorly and the performance degrades with increasing positional variation, as is evident from the confusion matrices shown in Tables 1 and 2.

However for some classes, the descriptor trained with low level features give better performance when tested on images with large positional variation. The confusion matrices in Table 3 clearly indicates that the additional higher level features acquired using the 25×25, 50×50 and 100×100 cells provide essential discriminatory information for the training images. Although adding to the dimensionality of the resulting feature set, these cell sizes provide the macro-information that separates the various hand gestures. However, it was observed that removing the 100×100 cell had essentially no effect on the perfor-mance of the classifier but removing the lower size blocks degrades the performance.

Table 2. Confusion matrix for classifier trained on images with large positional variation and tested on images with large positional variation, using only low-level descriptors.

	Identified Class							
		1	2	3	4	5	6	7
Actual Class	1	66	0	0	0	18	16	0
	2	13	47	0	1	27	12	0
	3	42	0	20	0	24	12	2
	4	28	9	0	17	9	37	0
	5	3	7	0	0	85	5	0
	6	36	2	0	1	5	56	0
	7	14	0	4	0	68	3	11

Figures 3 and 4 show the plot of the first two dimensions of the reduced feature space when the reduction has been performed using both PLS and PCA. The classifiers were trained on images having slight positional variations. The plots clearly indicate the advantage of using PLS as the dimensionality reduction technique.

The confusion matrices for the classifiers trained and tested using the three image sets of test and training images clearly show the tradeoff between the accuracy of detection and the positional variation.

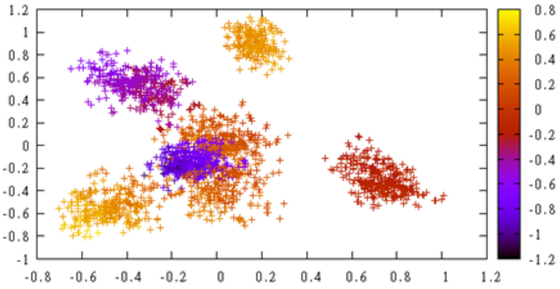

Figure 3. The first two dimensions of the reduced feature space of PCA.

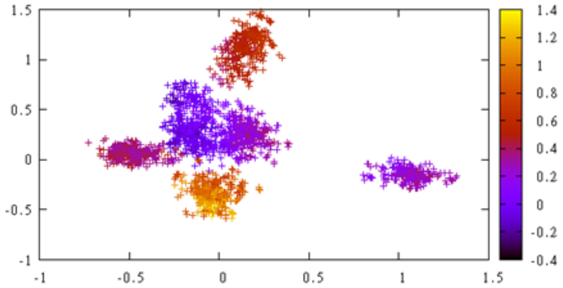

Figure 4. The first two dimensions of the reduced feature space of PLS.

Table 3. Confusion matrix for *PLS* classifier trained on images with *slight* positional variation and tested on images with *slight* positional variation, using high-level descriptors.

	Identified Class							
		1	2	3	4	5	6	7
Actual Class	1	97	0	0	0	0	3	0
	2	0	95	0	0	5	0	0
	3	0	0	100	0	0	0	0
	4	0	0	0	100	0	0	0
	5	0	4	0	0	96	0	0
	6	0	0	0	0	0	97	3
	7	0	0	0	0	0	0	100

For classifier trained using images with no positional variation, both PCA and PLS give perfect results for test images having no positional variation, which however, is of not much practical use.

For images having slight positional variations, PLS performs slightly better than PCA, and both give results degraded to almost the same extent, for test images having high positional variation. For the classifier trained using images with slight positional variation, both PLS and PCA give almost the same results with images having no and slight positional variations, with PLS giving much better results for the last three classes than PCA as shown in the confusion matrices in Tables 4 and 5, but failing for the first two classes.

The results obtained for the classifier trained using images with large positional variation shows the worst performance, with PLS performing much better for test images with no positional variation and reasonably well for test images with slight positional variation, except for the fifth class as shown in the confusion matrix in Tables 6 and 7.

However, PCA performs slightly better than PLS for test images with large positional variation; this is contrary to the other two test image sets. This can be attributed to the scattered nature of the classification and the smaller number of classes, this generates query points in the scattered region, leading to better odds of finding the point classified into the correct class as the nearest neighbor.

Thus, the classifier performs reasonably well for images having slight positional variations and the performance degrades with the increase in positional variation. We have been able to successfully recognize hand gestures in a wide variety of backgrounds and illuminations, with the degree of positional variation still being a hindrance to the accurate detection of various hand gestures, as is evident by the tradeoff which is visible in the confusion matrices.

Table 4. Confusion matrix for *PLS* classifier trained on images with *slight* positional variation and tested on images with *large* positional variation, using high-level descriptors.

		Identified Class						
		1	2	3	4	5	6	7
Actual Class	1	66	0	0	0	20	13	1
	2	2	46	0	0	47	5	0
	3	5	2	78	0	10	5	0
	4	20	0	0	80	0	0	0
	5	1	4	0	0	90	5	0
	6	4	2	0	0	3	91	0
	7	0	0	0	0	22	0	78

Table 5. Confusion matrix for *PCA* classifier trained on images with *slight* positional variation and tested on images with *large* positional variation, using high-level descriptors.

		Identified Class						
		1	2	3	4	5	6	7
Actual Class	1	83	2	0	0	8	7	0
	2	5	57	0	0	30	8	0
	3	19	2	77	0	0	2	0
	4	17	0	0	77	0	6	0
	5	18	1	0	0	76	5	0
	6	5	0	0	0	7	88	0
	7	1	0	0	0	19	1	78

Table 6. Confusion matrix for *PLS* classifier trained on images with *large* positional variation and tested on images with *slight* positional variation, using high-level descriptors.

		Identified Class						
		1	2	3	4	5	6	7
Actual Class	1	88	0	0	0	2	10	0
	2	0	92	0	0	5	3	0
	3	0	0	100	0	0	0	0
	4	0	0	0	100	0	0	0
	5	9	10	1	0	60	1	19
	6	2	1	0	0	2	95	0
	7	0	0	0	0	0	0	100

Table 7. Confusion matrix for *PCA* classifier trained on images with *large* positional variation and tested on images with *slight* positional variation, using high-level descriptors.

		Identified Class						
		1	2	3	4	5	6	7
Actual Class	1	83	0	0	0	4	13	0
	2	3	93	0	2	1	1	0
	3	0	0	100	0	0	0	0
	4	0	0	0	96	0	4	0
	5	25	7	0	0	66	1	1
	6	1	0	0	2	2	95	0
	7	0	0	0	0	0	0	100

6. Conclusion and Future Work

Through this paper we proposed a hand gesture recognition system that employs the techniques developed for pedestrian detection to recognize a small vocabulary of 7 hand gestures using Histogram of Oriented Gradients as the descriptors.~~–~~We have used PLS as a ‘class aware’ method of dimensionality reduction which performs better than PCA and preserves significant discriminative information in the lower dimensions. Three sets of databases consisting of training as well as testing image sets with varying degree of positional variation were developed. We analyzed the importance of using multi-level HOG features for robust human hand gesture recognition and showed that only low level HOG features are not sufficient for high detection rate. We were able to achieve reasonable degree of accuracy of detection of human hand gestures with performance degradation due to the evident tradeoff between the accuracy and positional variation of the hand. The simple brute-force implementation of the k-nearest neighbor search algorithm classifies the query points in real time for 3500 reference points in 10 dimensional feature space.

We have taken a small vocabulary of seven hand gestures to analyze the feasibility of HOG descriptors and PLS reduction for human hand gesture recognition, however we wish to increase this number in our future work for better interactivity.

7. References

[1] William T. Freeman, Craig D. Weissman. Television Control by Hand gestures. IEEE Intl. Workshop on Automatic Face and Gesture Recognition, Zurich, June, 1995.

[2] W. T. Freeman and M. Roth, Orientation histograms for hand gesture recognition, Intl. Workshop on Automatic Face- and Gesture- Recognition, IEEE Computer Society, Zurich, Switzerland, , pp. 296-30, June, 1995.

[3] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In CVPR 2005, San Diego, USA, pages 886 - 893, June 2005.

[4] H.Wold. Partial least squares. In S. Kotz and N. Johnson, editors, Encyclopedia of Statistical Sciences, volume 6, pages 581–591. Wiley, New York, 1985.

[5] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. ICCV, 2009

[6] Lee, H.-J. AND Chung, J.-H.. Hand gesture recognition using orientation histogram. In Proc. of IEEE Region 10 Conf. Vol. 2. 1355–1358. 1999.

[7] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. Journal of the ACM, 45(6):891–923, 1998.

[8] M. Fukumoto, K. Mase, and Y. Suenaga. Real time detection of pointing actions for a glove free interface. In Workshop on Machine Vision Applications, Tokyo, IAPR 1992.

[9] V. C. Tartter and K. C. Knowlton. Perception of sign language from an array of 27 moving spots. Nature, (239):676-678, Feb. 19, 1981.