Res. Lett. Inf. Math. Sci., 2011, Vol. **15**, pp. 12–20
Available online at http://iims.massey.ac.nz/research/letters/

1

A New 2D Static Hand Gesture Colour Image Dataset for ASL Gestures

A.L.C. Barczak, N.H. Reyes, M. Abastillas, A. Piccio and T. Susnjak
IIMS, Massey University, Auckland, New Zealand

It usually takes a fusion of image processing and machine learning algorithms in order to build a fully-functioning computer vision system for hand gesture recognition. Fortunately, the complexity of developing such a system could be alleviated by treating the system as a collection of multiple sub-systems working together, in such a way that they can be dealt with in isolation. Machine learning need to feed on thousands of exemplars (e.g. images, features) to automatically establish some recognisable patterns for all possible classes (e.g. hand gestures) that applies to the problem domain. A good number of exemplars helps, but it is also important to note that the efficacy of these exemplars depends on the variability of illumination conditions, hand postures, angles of rotation, scaling and on the number of volunteers from whom the hand gesture images were taken. These exemplars are usually subjected to image processing first, to reduce the presence of noise and extract the important features from the images. These features serve as inputs to the machine learning system. Different sub-systems are integrated together to form a complete computer vision system for gesture recognition. The main contribution of this work is on the production of the exemplars. We discuss how a dataset of standard American Sign Language (ASL) hand gestures containing 2425 images from 5 individuals, with variations in lighting conditions and hand postures is generated with the aid of image processing techniques. A minor contribution is given in the form of a specific feature extraction method called moment invariants, for which the computation method and the values are furnished with the dataset.

Keywords: Hand Image Dataset, Gesture Recognition, Image Dataset, Computer Vision, Feature Extraction, Colour Classification

1 Introduction

Gesture recognition is a challenging area in computer vision. It is important to have standard data that can be used to compare different algorithms and methods. The main objective of this work is to create a reasonably large set of hand images with standard gestures, and make it available to the computer vision research community.

This image dataset is based on ASL hand gestures. Although there are many hand gesture datasets available, there are some characteristics that distinguish our work from other datasets. Firstly, the images cover a large variety of hands using different illumination conditions. Secondly, the images are segmented and cropped, but not altered from the original captured images, allowing researchers to test their own combination of feature extraction methods (e.g., binary images, edge detection, etc.). Thirdly, contrary to many publicly-available datasets for gestures, there is no need to use special gloves, or any other apparatus. In the data collection, a simple wrist cover was used in order to improve the quality of the colour segmentation, but this does not mean that the wrist covers need to be considered for the development of recognition algorithms (see for

[1] Author for Correspondence: a.l.barczak@massey.ac.nz

example [1], where Haar-like features are used to detect certain gestures from hand images without wrist covers).

The images were taken at a certain angle of rotation (perpendicular to the subject), which limits the number of samples. This can be easily extended by tilting the original images, as explained in section 5. The acquisition process of the images took into consideration that the person making the gestures may tilt the hand slightly, which affects the positions of the hands and fingers. It is also possible to extend the dataset by altering the width/height ratio and producing extra images. These images would be (almost) equivalent to tilting their hands and taking another image. For certain training methods, as for example, Viola-Jones [2], it is important to have a large number of positive samples in order to train the classifier properly. This dataset can be used for this purpose if extended with the suggested methods described in section 5.

The current version of the dataset contains 2425 images of 5 individuals, with variations in lighting conditions and hand postures. The final goal is to collect 18000 images from 20 different volunteers. New versions of the dataset will be uploaded as soon as more images are collected.

In addition to the images, the new dataset includes numerical data for a specific feature extraction method. The method uses moment invariants up to the 4^{th} order. These features can also be fed directly to machine learning algorithms.

The dataset is called **MU_HandImages_ASL** The dataset is available publicly on the link: `http://www.massey.ac.nz/~albarcza/gesture_dataset2012.html`

2 Related Work

Back in 2005, a former computer vision research group at IIMS produced an image dataset for hand gesture recognition evaluation. The dataset, described in [3], is now obsolete. It did not have a complete set of standard gestures, and it was relatively small. A new image dataset was created, including all standard ASL (American Sign Language) gestures.

A google search on “hand image dataset gesture” results in few relevant results. Most of these datasets do not have the full image available or have a limited number of gestures and postures, or need special gloves to facilitate the segmentation. The most common and well-known datasets are listed in table 1. The citation of these web sites is for reference purposes only, as there is no guarantee that they will be accessible in future. Whenever possible, we linked the dataset with a proper journal or conference publication.

The dataset that is the most similar to ours is the ASL rendered dataset from Athitsos and Sclaroff [4]. However, it does not have the same variety of hands or illumination conditions.

Table 1: List of other hand datasets and gesture datasets

Description	**URL**	**References**
Various Hand and Gesture Datasets	`www.idiap.ch/resources/gestures`	[5–8]
Gesture Recognition Database Image	`www-prima.inrialpes.fr/FGnet/data/12-MoeslundGesture`	[9]
Pointing 2004 Gesture Recognition Dataset	`wwwprima.inrialpes.fr/Pointing04/datasets.html`	[10, 11]
Two-handed datasets	`http://www.idiap.ch/resources/twohanded/`	[12]
RWTHBOSTON-104 Database Video	`www-i6.informatik.rwthaachen.de/~dreuw/dataset-rwth-boston-104.php`	[13, 14]
Hand image dataset (ASL,rendered)	`http://www.cs.bu.edu/groups/ivc/data.php`	[4]
ASL video sequences	`http://www.bu.edu/asllrp/ncslgr.html`	[15]

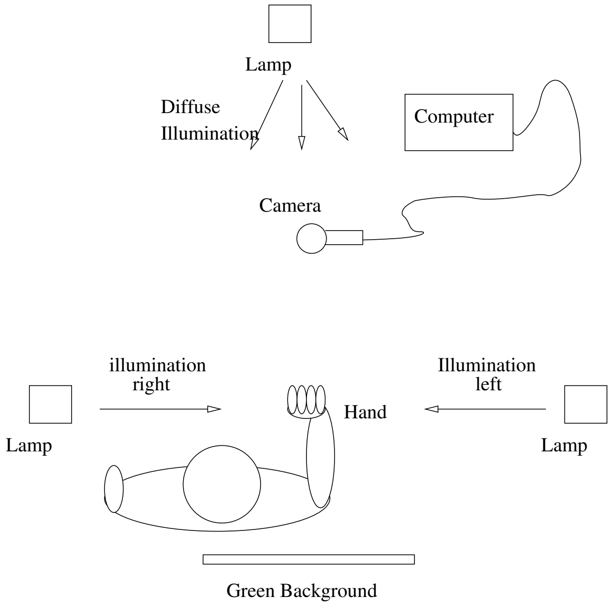

Figure 1: Lighting direction and set up for the image collection (bird's eye-view).

3 Methodology

The images were acquired as follows: volunteers gesticulated standard ASL gestures close to a camera on a tripod, with a neutral-coloured wall behind the subject, and a green background behind the hands. To add variety to the illumination conditions, a series of mounted lamps were used, from top, bottom, left and right. Diffuse lighting conditions were achieved with lamps from various directions, simulating a natural environment. Figure 1 shows the image acquisition approach.

In order to facilitate the segmentation process, the person taking the images for the training set was asked to wear a wrist cover with the same background colour (see figure 2). Once the hands are segmented, they can be used for training directly. If the experiment requires separate training and test sets, it is possible to split them and add artificial backgrounds to the hand images.

The images are furnished with the best colour segmentation on a black background (i.e., with a (R=0,G=0,B=0) background).

3.1 Defined Classes and Similarities Between ASL Gestures

In principle, there are 36 classes in the dataset. However, depending on which feature extraction method is used, there are gestures that are notoriously difficult to classify due to their similarities. For example, the difference between "M" and "N" is the thumb appearing or not between the fingers. Many simple feature extraction methods will simply not differentiate them. Other examples are "K" against "2" and "V", "S against "T", "I" against "J" etc.

Note that some gestures are identical, e.g., "O" (letter O) and "0" (digit zero). In this dataset, these gestures have had their images collected separately for completion.

The decision to coalesce classes or not has to take into consideration the image acquisition method and the feature extraction method. For example, for a rotation invariant method, "I" "J" are too similar. However, for non-invariant features they might still belong to different classes.

Two ASL gestures, namely "J" and "Z", would be dynamic in their original form. In order to standardise the dataset, we did not consider the movement. Instead, we rotated slightly the

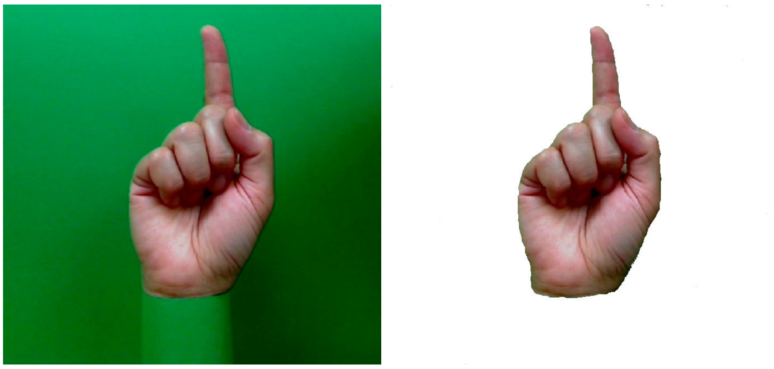

Figure 2: Acquired image with wrist cover. The images are segmented to obtain the final images stored in the dataset.

gesture to differentiate them from others that might be similar.

One can find many variations of the ASL gesture set for letters and digits. Sometimes angles are slightly different, and often, the profile of the hand can vary considerably because of the position of the camera. Researchers need to consider these variations carefully if the intention is to build an accurate gesture recognition system.

Table 2 presents a list of similarities in this dataset, as well as some of the conditions that should be considered when coalescing these classes.

Table 2: List of other hand datasets and gesture datasets

similarity	**gestures**	**Observations**
identical	“0” “O”	
	“V” “2”	
very similar	“V” “2” “K”	the difference for “K” is in the position of the thumb
	“W” “6”	
	“Z” “1”	different angle
	“M” “N”	
	“I” “J”	different angle (“J” is dynamic in the original ASL)
	“D” “1”	position of the thumb
	“S” “T”	position of the thumb

Figures 3 and 4 show image examples for all the standard ASL gestures.

As an example, in previous experiments described in [16], Haar-like features were used to train binary classifiers based on AdaBoost. In this case, one needs to have a variety of backgrounds for the same hand images, so the training can concentrate on the similarities between samples. If the features are influenced by the background, extra samples need to be created in order to train properly.

3.2 Image Nomenclature

The names of the files follow a simple convention to make it easy to automate scripts and programs. The convention is as follows:

```
handX_G_ILL_seg_crop_R.png
```

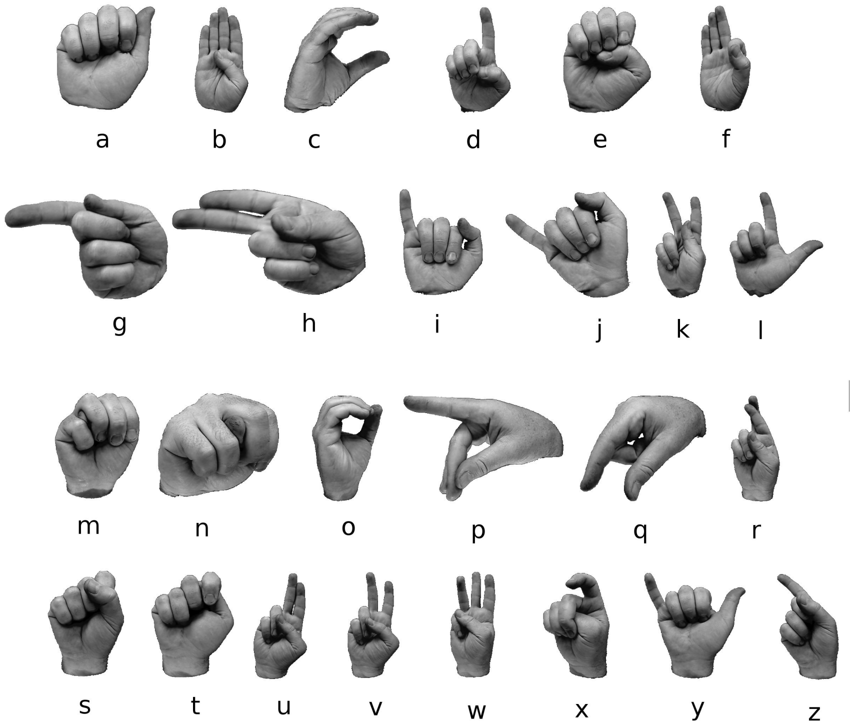

Figure 3: The complete ASL letter set with sample segmented images

where:

- X is the number of the volunteer whose images where recorded.
- G is the gesture number, from a to z, 0 to 9.
- ILL is the illumination conditions. ILL can be bot (bottom), top, left, right or diff (diffuse).
- R is the repetition, usually from 1 to 5.

The next subsection describe one example of feature extraction.

4 An example of feature extraction: moment invariants, up to the 4^{th} order

There are too many methods for feature extraction to cover them all in this article. Only moment invariants is given as an example of feature extraction that can be used with this dataset. When using this particular feature extraction method, it is assumed that that the hands are perfectly segmented, i.e., that a simple colour filter is effective. The numerical data is furnished with the dataset.

Moment invariants are invariant to rotation, scaling and translation (for translation the background needs to be dark, i.e., R=0,G=0,B=0). However, these features are not robust when contrast changes suddenly, or if there is a strong shadow over the hands. They are also prone to numerical instabilities. One notorious example happens when figures are perfectly symmetric. Therefore, great care is needed when utilising the numerical values furnished with the dataset.

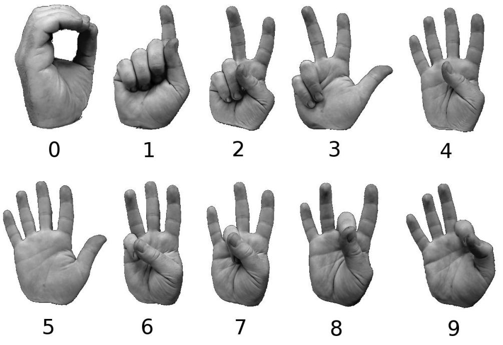

Figure 4: The complete ASL numeric set with sample segmented images

The moment invariants in this dataset are computed according to [17], up the the 4^{th} order. The formulas for moment invariants up the the 3^{rd} order are part of the well-known Hu's set (5 of the original 7). Interestingly, Flusser found an additional independent 3^{rd} order moment, and the remaining five 4^{th} order features, written as a function of η_{pq} (normalised central moments). These are:

$$\psi_7 = \eta_{40} + \eta_{04} + 2\eta_{22} \tag{1}$$

$$\psi_8 = (\eta_{40} - \eta_{04})[(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2] + 4(\eta_{31} + \eta_{13})(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21}) \tag{2}$$

$$\psi_9 = 2(\eta_{40} - \eta_{04})(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) - 2(\eta_{31} + \eta_{13})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \tag{3}$$

$$\begin{aligned}\psi_{10} = (\eta_{40} - 6\eta_{22} + \eta_{04})\{[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]^2 - 4(\eta_{30} + \eta_{12})^2(\eta_{03} + \eta_{21})^2\} \\ + 16(\eta_{31} - \eta_{13})(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21})[(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2]\end{aligned} \tag{4}$$

$$\begin{aligned}\psi_{11} = 4(\eta_{40} - 6\eta_{22} + \eta_{04})(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21})[(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2] \\ - 4(\eta_{31} - \eta_{13})\{[(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2]^2 - 4(\eta_{30} + \eta_{12})^2(\eta_{03} + \eta_{21})^2\}\end{aligned} \tag{5}$$

The moment invariants values are furnished in three sets. The first was computed directly using grey-scale images, obtained by colour segmentation from the hand images in the dataset. The second set of moment values was computed from the binary images, which were taken from the grey-scale images. The third set was computed from the contour images obtained from the binary images. A comparative study of the accuracy of these three sets of moments can be found in [18]. Invariant features are subject to errors, and it was found that the average variance in moments from contour images are quite large, followed by the moments from binary images. Moreover, it was observed that the moments from the grey-scale images are the most stable for rotation and scaling.

5 Extending the Dataset

The large variety of feature extraction methods makes it difficult to generalise a set of images that will serve for a specific experiment or application test. The training images collected for the original dataset can be extended by rotation, stretching, addition of random backgrounds and image processing operations such as contrast stretching, filtering, grey-scale conversions, etc. In line with this, the existing images can be easily processed using open source tools such as GIMP (`www.gimp.org`) or ImageMagick (`www.imagemagick.org`).

For example, using ImageMagick, the following command will rotate the image 45 degrees clockwise:

```
convert -rotate 45 -background black orig.jpg result.jpg
```

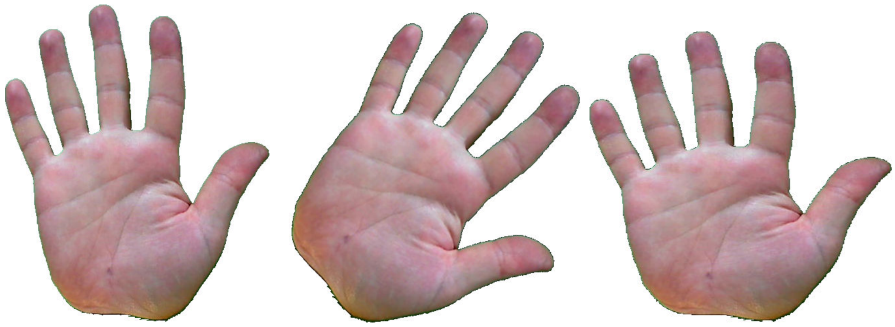

Figure 5: Extending the Dataset: a) Original image. b) Rotated 45 degrees. c) Scaling 80% on axis y.

Some feature extraction methods may use colours to segment the hands, while others (e.g., Haar-like features) do not use colours at all. As an alternative, for shorter downloading, there is a grey-scale version of the images available.

The geometric proportion and rotation of hands may or may not influence a particular feature extraction method. For example, while moment invariants are relatively robust against rotation and scaling, they are not invariant to general affine transforms. Two different approaches can tackle similar problems. The first approach is to find a different feature extraction method that is invariant to that particular transformation. The second approach is to create new images that produce the variability that is expected from the test images.

Considering the problem described above, images of the hands of different people may have longer or shorter fingers, making it difficult to generalise the data collection. Extra images can be created, in order to simulate stretched hands that will differ in geometric proportion from the original image. The resulting images enrich the training sets; therefore, helping the classifiers achieve better generalisation capability.

A possible command to create transformed images using ImageMagick is called `convert`:

```
convert -resize 100%x80% orig.jpg result.jpg
```

Figure 5 shows the two transformations, rotation (figure 5.a) and shrinking axis y to 80% of its original size (figure 5.b). Using commands similar to these examples associated with a scripting language allow for batch processing. In the utilities folder of the dataset there are examples of scripts using *bash* scripts.

Two scripts are included in the first version: *rotate.sh* and *scale.sh*.

The *rotate.sh* script will create new images rotated every 5 degrees between -45 to 45 degrees, relative to the original position of the hand. This is going to multiply the number of images of the original dataset size by 18.

The *scale.sh* script will create new images scaling them from 0.5 (50%) to 2.0 (200%) in 10% intervals. This increases the original dataset by 16 times. Combining both scaling and rotation makes the number of images grow to 288 times the original one.

6 Summary and Future Work

In this short paper we described the creation of a new image dataset that can be used by other computer vision researchers. The first version of the training dataset contains 2425 images of 5 individuals for each of the 36 ASL gestures. Future versions will eventually contain 18000 hands of 20 individuals under 5 different illuminations, with 5 repetitions, for each one of the 36 ASL gestures. The dataset is going to be updated constantly with new images until completion.

7 Conditions for using these images

The images distributed with this dataset are copyrighted by the authors of this paper. You are granted free use, distribution and publication of these images, but you are obliged to follow the following rules:

1. You may not claim that these images belong to you, or that they were taken by you. If you publish images of these dataset you should cite this paper to acknowledge the rightfull origin of the images.

2. You can alter images using the methods described in this paper. This does not mean that you own the modified images. Modified images that contain the hand images in any shape or form are still subject to the same rules hereby described. You are allowed to publish, copy, re-distribute the modified images, as long as the true origin is acknowledged.

3. You can produce research using these images, and the independent results of your original research carried out by you using our image datasets are yours. If commercial products are spawned from your research, you can freely redistribute our images with the product, but you are still not allowed to sell images that belong to this dataset. You are not allowed to sell images that contain cropped gesture images pasted onto them either.

4. If you re-distribute copies or distribute modified copies of the images contained in this dataset, you should add this notice *ipsis litteris* to the documentation, and include a link or an acknowledgement to this research paper.

References

[1] A. L. C. Barczak, F. Dadgostar, and M. J. Johnson, "Real-time hand tracking using the viola and jones method," in *SIP 2005*, Honolulu, HI, 2005, pp. 336–341.

[2] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[3] F. Dadgostar, A. L. C. Barczak, and H. Sarrafzadeh, "A colour hand gesture database for evaluating and improving algorithms on hand gesture and posture recognition," *Research Letters in the Information and Mathematical Sciences*, vol. 5, pp. 127–134, 2005.

[4] V. Athitsos and S. Sclaroff, "Estimating 3d hand pose from a cluttered image," in *Proc. Conf. Computer Vision and Pattern Recognition*, Maddison, WI, June 2003, pp. 423–439.

[5] J. Triesch and C. von der Malsburg, "Robust classification of hand postures against complex backgrounds," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, IEEE Computer Society Press. IEEE Computer Society Press, 1996, pp. 170–175.

[6] J. Triesch and C. von der Malsburg, "A system for person-independent hand posture recognition against complex backgrounds," *IEEE Trans. on PAMI*, vol. 23, pp. 1449–1453, 2001.

[7] S. Marcel, "Hand posture recognition in a body-face centered space," in *Proc. of the Conference on Human Factors in Computer Systems (CHI)*, 1999.

[8] S. Marcel, O. Bernier, J.-E. Viallet, and D. Collobert, "Hand gesture recognition using input/ouput hidden markov models," in *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition (AFGR)*, 2000.

[9] T. B. Moeslund, "Recognizing gestures from the hand alphabet using principal component analysis," Master's thesis, Aalborg University, 1996.

[10] S. Carbini, J. E. Viallet, and O. Bernier, "Pointing gesture visual recognition for large display," in *FG Net Workshop on Visual Observation of Deictic Gestures*, 2004.

[11] T. B. Moeslund, M. Stoerring, and G. E., "Pointing and command gestures for augmented reality," in *FG Net Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK, 2004.

[12] S. Marcel and A. Just, "Two-handed gesture recognition," IDIAP Research Institute, Research Report IDIAP-RR 05-24, 2005.

[13] P. Dreuw, J. Forster, T. Deselaers, and H. Ney, "Efficient approximations to model-based joint tracking and recognition of continuous sign language," in *IEEE International Conference Automatic Face and Gesture Recognition (FG)*, Amsterdam, The Netherlands, 2008.

[14] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," in *Interspeech*, Antwerp, Belgium, 2007, pp. 2513–2516.

[15] C. Neidle, "Signstream annotation: Conventions used for the american signe language linguistic research project," Boston University, Technical Report 11, August 2002.

[16] A. L. C. Barczak, F. Dadgostar, and C. H. Messom, "Real-time hand tracking based on non-invariant features," in *IMTC2005*, Ottawa, Canada, 2005, pp. 2192–2199.

[17] J. Flusser, "On the independence of rotation moment invariants," *Pattern Recognition*, vol. 33, pp. 1405–1410, 2000.

[18] A. L. C. Barczak, A. Gilman, N. H. Reyes, and T. Susnjak, "Analysis of feature invariance and discrimination for hand images: Fourier descriptors versus moment invariants," in *International Conference Image and Vision Computing New Zealand IVCNZ2011*, 2011.