

HAND POSTURE DATASET CREATION FOR GESTURE RECOGNITION

Luis Anton-Canalis

*Instituto de Sistemas Inteligentes y Aplicaciones Numericas en Ingenieria
Campus Universitario de Tafira, 35017 Gran Canaria, Spain
lanton@iusiani.ulpgc.es*

Elena Sanchez-Nielsen

*Departamento de E.I.O. y Computacion
38271 Universidad de La Laguna, Spain
enielsen@ull.es*

Keywords: Image understanding, Gesture recognition, Hand dataset.

Abstract: This paper introduces a fast and feasible method for the collection of hand gesture samples. Currently, there are not solid reference databases and standards for the evaluation and comparison of developed algorithms in hand posture recognition, and more generally in gesture recognition. These are two important issues that should be solved in order to improve research results. Unlike previous hand image datasets, which creation usually involves many different people, sceneries and light conditions, we propose a simplified method that requires just a single person's hand being recorded in a controlled light environment. Our method allows the generation of thousands of heterogeneous samples within hours, thus saving time and people's efforts. The resulting dataset has been tested with a cascade classifier, although it may be used by most pattern recognition systems, and compared with a classical dataset obtaining similar results.

1 INTRODUCTION

Hand gesture recognition (Wu and Huang, 1999) implies several requirements: an appropriate hand gesture image dataset, the proper detection of hands or specific gestures on video streams and finally the recognition of the gesture. In a previous work (Anton-Canalis et al., 2005b), we presented a method for the detection of hands in video streams, formulated in terms of the integration and combination of temporal coherence information and a cascade classifier method. Temporal coherence information was supplied by a template tracker (Anton-Canalis et al., 2005a) with the aim of achieving real-time performance. Using Viola and Jones' pattern recognition method in grey level images (Viola and Jones, 2001) we were able to achieve a high detection rate (0.99) while maintaining a low error rate (0.03). The set of hand samples used in the training stage was created gathering hand images from different sources such as (Triesch, 2000). However, it was not specifically designed for our purposes, so there was a need for rebuilding the hand image dataset. Unlike face detection systems (Zhao et al., 2003), there are not many available hand datasets that meet the constraints that cascade classifiers oriented towards hand detection impose. In this paper, we propose a simplification of

the sample collecting process, reducing it to a single person's hand gesture set filmed under different light conditions, avoiding the necessity for the variation of backgrounds and subjects.

1.1 Previous Work

Cascade classifiers are currently considered the fastest and most accurate pattern detection method for faces in monocular grey-level images (Viola and Jones, 2001) Its efficiency has been proven in recent works in a wide range of conditions. However, while frontal faces share common features (eyes, eyebrows, nose, mouth, hair...), hands are not so easily described. Their flexibility makes of them highly deformable objects, so it is hard to train a cascade classifier for detecting hands. Some approaches involving cascade classifiers include training a different classifier for each recognizable gesture (Kölsch and Turk, 2004b), or a single classifier for a limited set of hands (Stenger et al., 2004) but that leads to the detection of a low number of gestures, or forces the user to perform very precise gestures (Kölsch and Turk, 2004a). We proposed a cascade classifier which was trained to detect wrists (Anton-Canalis et al., 2005b), being the main advantage of this approach the high independence from the gesture being made. Hands were de-

tected without taking into account the gesture, with no limitation in the number of gestures being detected, as long as wrists were not concealed. Additionally, there was no need for an initialization step. Figure 1 shows some detection examples.



Figure 1: Five positive results showing both wrists detections (dark rectangle) and complete hands (white rectangle).

2 SAMPLE SET CONSTRUCTION

Cascade classifiers need both positive and negative samples for their training. Negative samples should be as numerous and dissimilar as possible, and they should not contain the target object, i.e hands. Positive samples should show as many different target object views as possible, in different conditions. In relation to frontal faces, for example, it is advisable to maximize three variables: subjects, facial gestures and light source directions. There are many databases where these circumstances are met (orl,) (Georghiades et al., 2001). Hands, however, add a fourth variable: background. While in a face sample there is no need to show nothing else than a face, it is not possible to show a hand without showing pieces of background between unbended fingers. Any gesture apart from a fist will show what lies behind the hand, and thus it becomes part of the positive sample.

A set of positive hand samples created for a cascade classifier should also add so many different backgrounds as possible, allowing the classifier to infer

what is the real target object. Although there are some hand databases available, it is difficult to find the four requisites together. In (Triesch, 2000), for example, there are around 15 different backgrounds and 25 subjects, but only 9 gestures. Thus, this set is suitable for the training of a single classifier for each gesture, as in (Kölsch and Turk, 2004b), but not for a more general one. Being conscious of the high difficulty in meeting the constraints that a hand dataset imposes during its creation, we propose a method which tries to reduce them to only two requirements: different light sources and different gestures. Thus, hand gestures are performed by a single person under different light conditions, filmed against a background having a single color or a relatively narrow range of colors (chroma key or color keying). Then, each sample may suffer a slight geometrical transformation (stretching and/or rotation) and finally a random imagen, chosen from a high amount of images not showing hands, substitutes the chroma signal. Using the chroma key technique, it is possible to create a set of positive hand samples large enough to train and test a cascade classifier, avoiding the troubles of gathering many different people and backgrounds, thus saving time.

For testing purposes, we assembled a recording set with a green chroma screen, six spot lights and a single ambient light with fixed intensity and color. Lights were placed in front of the actor, on his left, right and in front of him, three at the same hand height, and three around two meters above the hand height, as seen in Figure 2. This way, it was possible to record six sequences with different light conditions. A total amount of 288 images were extracted from the recorded sequences, showing more than 20 gestures under each light setup.

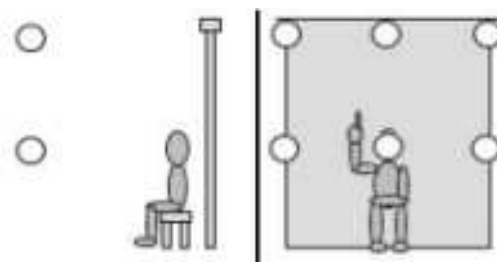


Figure 2: Side and front view of the recording set. White circles represent light sources.

Then, the chroma key was substituted four times in each image with a random background image, taken from a group of more than 5000, generating four new samples, each of them also mirrored. No geometric transformation was applied to the original image. Figure 3 shows the main process steps for a given gesture.

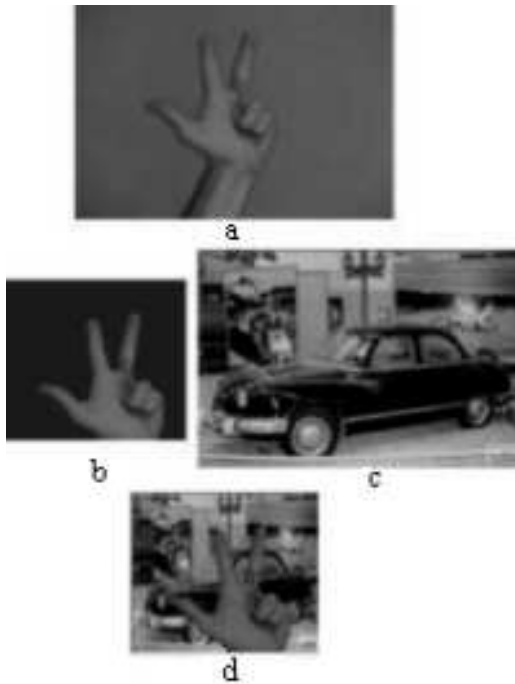


Figure 3: a) Original recorded image b) image cleaned and trimmed c) random background image d) final chroma substitution. The whole process is totally automated.

The final dataset consist in a total amount of 2304 20x20 grey level images. Figure 4 shows samples from our previous dataset and samples from the new one.



Figure 4: a) Natural in-place hand images b) samples obtained with the chroma technique.

3 SYNTHETIC DATASET PERFORMANCE

In order to analyze the goodness of a synthetic dataset (hands combined with random backgrounds) in contrast to a natural dataset (in-place taken hand images), we trained a cascade classifier with the set of new samples created with the chroma key approach. Then we applied it on a number of videos (11) that had been

previously studied in (Anton-Canalis et al., 2005b) with a cascade classifier trained with a natural dataset. Figures 5 and 6 show their respective detection rate (number of frames where there was a detection in the total amount of frames) and false positives rate. The

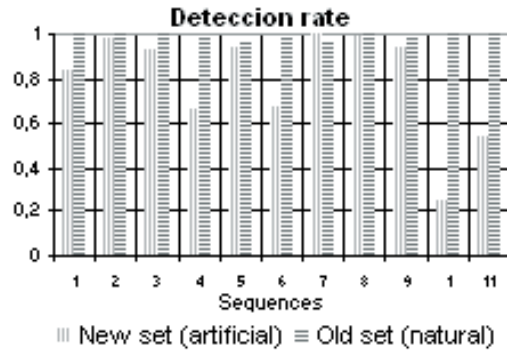


Figure 5: Detection rate for each sequence. Seq. 10 and 11 are clearly problematic, due to light conditions.

lower detection rate of the new classifier, which decreases the previous detection rate from 0.99 to 0.79, is the consequence of the poor results obtained mainly in two sequences, where we get a detection rate of 0.24 and 0.54. In both of them, ambient light conditions seem to be drastically different from the one used in the synthetic hand set. This points out the necessity for more light configurations in the recording set.

Although the global false positive rate raises slightly

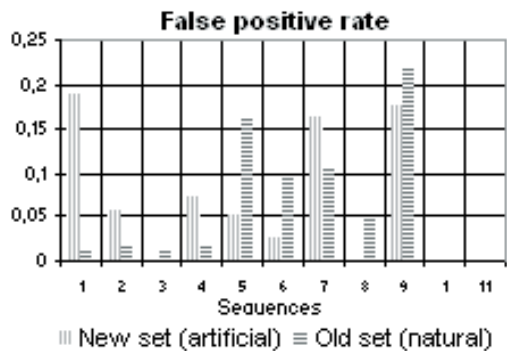


Figure 6: False positive rate. While some sequences are worse (1,2,4,6), others behave better (3,5,6,8,10,11).

from 0.062 to 0.066, there is a significant improvement in those sequences where the detection rate is maintained.

4 OVERALL RESULTS AND DISCUSSION

While results obtained with the synthetic set are in some degree not so good as those obtained with the natural set, they are quite positive. It must be taken into account that the synthetic set was created in no more than six hours, including the setup of the recording set (chroma screen, camera and lights), filming all the gestures and the final chroma substitution, and that a single person was required as a hand model, instead of a large group. The in-place picture method requires much more time. It is necessary to persuade a large group of people, vary backgrounds, get different light conditions and in some cases, images must be manually marked and segmented in order to extract hands from them.

The low detection rate, however, could be avoided maintaining the use of the tracking system when the cascade classifier failed, but it was necessary for comparison purposes to preserve the previous system structure.

5 CONCLUSIONS AND FUTURE WORK

In previous works (Anton-Canalis et al., 2005b) we proposed a cascade classifier trained to detect wrists as a simplification of the problem of finding hands in still images or videos. These kind of classifiers seem to be suitable for this task, given an appropriate training set. In our case, the training set was not created specifically for our purposes. Even though the preliminary condition of positive samples, we found results to be promising enough to develop this approach. We have presented a method for hand gesture sampling that simplifies this otherwise arduous task. Therefore, with the aid of a single person and the proper recording set, it is possible to generate a large and assorted hand image set. This method may be applied to any kind of object which morphological attributes force pieces of background to be present in positive samples (i.e. concave shapes), and used with most pattern recognition methods that require training samples.

Future research will investigate the possibility of using artificially generated hands, with a 3D rendering software, as training sets in cascade classifiers. The ultimate aim is to totally get rid of the necessity of an actor, and moreover, the meticulous setup of lights.

6 ACKNOWLEDGMENTS

This work has been supported by the Spanish Government and the Canary Islands Autonomous Government under projects TIN2004-07087 and PI2003/165.

REFERENCES

- Anton-Canalis, L., Sanchez-Nielsen, E., and Castrillon-Santana, M. (2005a). Fast and accurate hand pose detection for human-robot interaction. In *Lecture Notes in Computer Science LNCS 3522*.
- Anton-Canalis, L., Sanchez-Nielsen, E., and Castrillon-Santana, M. (2005b). Hand pose detection for vision-based gestures interface. In *IAPR MVA 2005*.
- Georghiadis, A., Belhumeur, P., and Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660. <http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html>.
- Kölsch, M. and Turk, M. (2004a). Analysis of rotational robustness of hand detection with a viola-jones detector. In *ICPR (3)*, pages 107–110.
- Kölsch, M. and Turk, M. (2004b). Robust hand detection. In *FGR*, pages 614–619.
- orl. Orlandi face database. <http://www.uk.research.att.com/facedatabase.html>.
- Stenger, B., Thayananthan, A., Torr, P. H. S., and Cipolla, R. (2004). Hand pose estimation using hierarchical detection. In *Computer Vision in Human-Computer Interaction*, pages 105–116.
- Triesch, J. (2000). Hand posture database i, ii. <http://www.idiap.ch/marcel/Databases/gestures/main.php>.
- Viola, P. A. and Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518.
- Wu, Y. and Huang, T. S. (1999). Human hand modeling, analysis, and animation in the context of HCI. In *ICIP (3)*, pages 6–10.
- Zhao, Chellappa, Phillips, and Rosenfeld (2003). Face recognition: A literature survey. *CSURV: Computing Surveys*, 35.