American Hand Sign Recognition in Video Streams

Domonkos Varga[1],[2]

[1]MTA SZTAKI, H-1111 Budapest, Kende u. 13-17.
[2]BME VIK Department of Control Engineering and Information Technology, H-1117 Budapest, Magyar tudósok krt. 2.

Abstract

We introduce in this paper a hand gesture recognition program based on computer vision. We detect and interpret the signs of the American Hand Sign Language. After reviewing related work a method is described for hand sign recognition in video streams. Our program follows the steps of a typical gesture-recognition algorithm - segmentation, feature extraction and classification. Then we outlined in detail the worked-out and applied algorithms. The segmentation in our case contains two steps - the spatial and the temporal segmentation. We extracted the so-called Fourier descriptors from the segmented video frames. On the grounds of Fourier descriptors we solved the classification. Finally we described our experimental results which appertain to the speed, robustness and running time of the program. CUDA is used for accelerating the running time of our application.

Categories and Subject Descriptors (according to ACM CCS): I.4 [Image Processing and Computer Vision]: Gesture recognition, segmentation, Fourier descriptor

1. Introduction

Hand gesture recognition is a new challenging means of interactions with computers. Device-based techniques such as electronic gloves can also be used for communicating with computers but our approach is a vision-based one using only a simple camera. This allows to communicate with the computer directly but recognition of shape of hand is a complex tasks which needs many algorithms.

Gestures to be recognized must be simple but different enough from each other. That is why we used the American Sign Language which satisfies these requirements.

This paper presents a hand gesture recognition system using a contour based approach. The first step is the spatial segmentation then the temporal segmentation comes. In the spatial segmentation a skin-detection algorithm will be used to determine the pixels which belongs to the hand. In the temporal segmentation we detect a moving or unmoved hand. We realize the temporal segmentation with the help of a final state machine. With morphological operations the hand contour can be extracted so the image is converted into boundary image. The first 50 coefficients of the Fourier transform are used to characterize the the hand gesture. For classification distance metric and SVM techniques are used.

We also outline the possibility of enhancing the speed of computation by using CUDA thrust. This time CUDA thrust is used for the determining the geometric properties.

The method has certain limitations. The user must wear a single colored shirt or pullover and he must hold his hand in the middle of the frame. No flashing signs are allowed in the background. In my experience these mean no severe limitations.

The paper is organized as follows. In section 2 the related work is summarized. Section 3 gives an overview of the algorithms and methods used while results and analysis comes in section 4.

2. Related Work

In hand gesture recognition studies two main approaches can be distinguished: hand model-based and appearance-based techniques[1].

Hand model-based methods detect the exact 3D pose of the hand and use a device. Device can be electronic or colored gloves which have and interface to the computer. They area excellent methods but not the topic of this paper.

For view-dependent techniques a simple built-in camera in laptops is sufficient, they are efficient in computation time but they can not distinguish fine movements as device-based methods.

Ravikiran et. al proposed a method of recognizing American Sign Language using number of fingers open in a hand gesture[2]. Mapari et. al developed a method where they crop and resize the image and count the number of peaks and valleys on the hand contour[3]. After dividing the image into sixteen parts the number of peaks and valleys are added to the counters of the image parts. Neural network is used to train the program for classifying the hand signs.

The appearance-based methods mostly differ in methods extracting content representing features of images and in classifying techniques. Region-based descriptors are for example the Hu[?] moments while Fourier descriptors[5] are widely used as contour descriptors.

For classification in simple cases multidimensional Euclidean or Manhattan distance is used while for complex cases either neural network or SVM (Support Vector Machine) techniques are applied. In the latter cases a machine learning process is involved where the number of training hand sign are important factors.

3. Overview of the methodology

The program works with a laptop built-in camera or with a simple camera but some constraints have to be made. The user should held his hand in the middle of the image, he must wear long-sleeved single color shirt or pullover. Cropping of images was not necessary of enhancing the reliability of recognition. The outline of the algorithm can be seen in Figure 1. Spatial segmentation is made by skin color detection. In spatial segmentation we detected the pixels which belongs to the hand of the user. Temporal segmentation is achieved using a finite state machine model. Hand gesture is only recognized when the hand remains still for a predefined period of time. Then we extracted from the segmented image the so-called Fourier descriptors. On the grounds of Fourier descriptors we solve the problem of classification.

3.1. Spatial segmentation

For spatial segmentation we used a skin-color-based detection of the hand. First, the image in RGB was converted to HSI color space. We applied the algorithm of histogram equalization to the "I" component of HSI color space. After this we reconverted the preprocessed image in HSI to RGB. Applying the histogram equalization on the Red, Green, and Blue components of an RGB image may yield dramatic changes in the color balance of the image. That is why we did histogram equalization in HSI.

After the preprocessing we converted to HSV. The HSV color space is more related to human color perception[6]. The skin in channel H is characterized by values between 6 and 38, in the channel S from 0.23 to 0.68 for Caucasian ethnics[7]. Pixels classified as skin were set to value 255, and non-skin pixels were set to value 0.

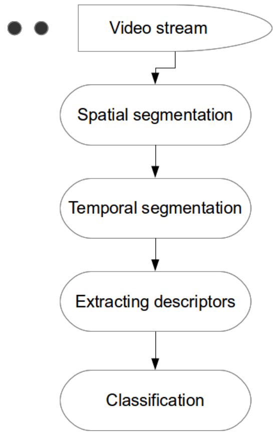

Figure 1: *Outline of the gesture recognition algorithm.*

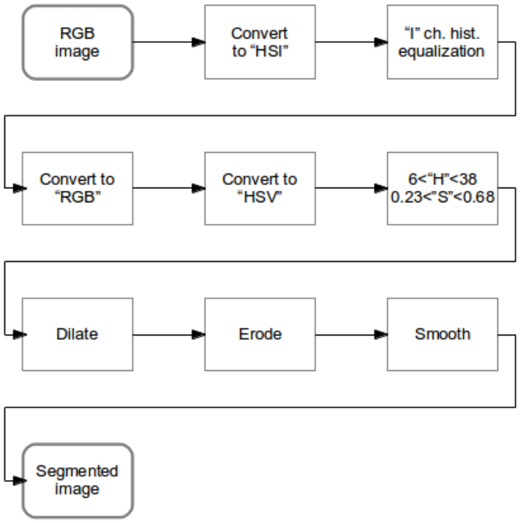

Figure 2: *The process of spatial segmentation.*

Next step minimizes noises, using a 5×5 structuring element in morphological filters. First, we used the structuring element with a dilatation filter. After that the same structuring element was used to erode the image. Then, a 3×3 median filter was used to eliminate the small noises.

The whole process of spatial segmentation can be seen in Figure 2.

3.2. Temporal segmentation

We realize the temporal segmentation with a finite-state machine. The possible states are start, hand movement, hand unmoved, hand recognized and finish. The state diagram can be seen in Figure 3. The start state is shown drawn with an arrow pointing at it from any where. The accept state is represented by thickened line.

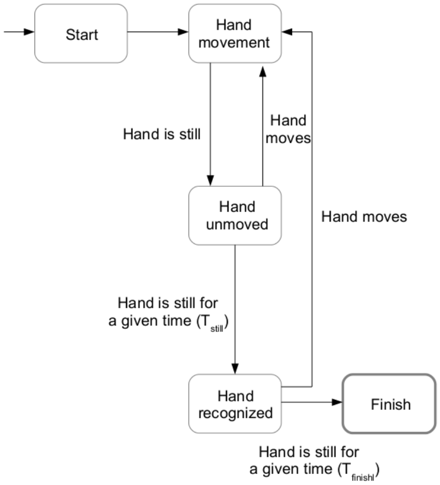
Start
Hand movement
Hand is still
Hand moves
Hand unmoved
Hand moves
Hand is still for a given time (T_{still})
Hand recognized
Finish
Hand is still for a given time ($T_{finishi}$)

Figure 3: *The state diagram of temporal segmentation.*

The system only recognizes a gesture as a hand sign if the hand is still for a predefined period (T_{still}). The performed sign is recognized in the hand-recognized state. The next hand sign is only detected when the state of the recognition turns into hand movement by the detection of the hand movement. The input mechanism can be stopped when the recognition is in the hand-recognized state but the user's hand is still for a given period (T_{finish}).

Our method continuously analysis hand movement by function Move(t)[8]:

$$\text{Move(t)} = \begin{cases} 1 & \text{if } X_t > \Omega \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$X_t = \max\{\text{median}_{\text{i=0}}{}^{N}(|v_{t-i} - v_{t-i-1}|), \text{median}_{\text{i=0}}{}^{N}(|h_{t-i} - h_{t-i-1}|)\}. \tag{2}$$

The hand moves in frame t if Move(t) $= 1$. The value N denotes the time window's length of the temporal analysis, v_t and h_t give the vertical and horizontal position of the center of the spatial segmented hand in frame t, and Ω is a threshold value to detect movement.

In our experiments, we tested different parameters: $T_{still} = 1.5$ seconds and $T_{finish} = 2$ seconds. These intervals offered enough time for users to perform gestures. We set the threshold $\Omega = 6$. We set the size of the temporal windows as $N = T_{motion} * FPS$, where T_{motion}=0.5 seconds and FPS is frames per second and gives the actual processing speed of the system.

If the motion is continuous and we cannot detect static stages, HMM based state estimation can be used to find the gesture state positions.

3.3. Fourier descriptors

Points of the hand contour can be represented with the following signatures:

- complex coordinates
- central distance
- curvature
- cumulative angular function.

Using complex coordinates each point M_i of the hand contour is represented by a complex number z_i where N is the number of the contour points:

$$\forall i \in [1,N], \qquad M_i(x_i,y_i) \Leftrightarrow z_i = x_i + jy_i \tag{3}$$

Calculating the Fourier transform with FFT leads to the N Fourier coefficients C_k:

$$C_k = \sum_{i=0}^{N-1} z_i \exp\left(\frac{-2\pi jik}{N}\right), \qquad k = 0,1,...N-1 \tag{4}$$

C_0 is discarded because it represents the translation of the shape. In order to achieve scale invariance the Fourier coefficients are divided by the magnitude of the second coefficient[9]:

$$I_k = \frac{|C_k|}{|C_1|}, \qquad k = 2,...N-1 \tag{5}$$

The computation of Fourier descriptors is not invariant to the zero point. That is why we determine the inertia tensor and the mass center of the segmented hand image. Then we rotate the image into the direction of eigenvectors. The contour of the segmented and rotated image is extracted with morphological operations. The whole process of extraction of Fourier descriptors can be seen in Figure 4.

Low frequency coefficients give information about the overall shape of the hand while high frequency coefficients represent fine details and noise. We used the first 50 descriptors in the classification process so very high frequencies are removed to decrease noise sensitivity.

3.4. CUDA thrust

CUDA thrust is a development in CUDA with the following advantages:

- incorporation STL-like algorithm
- high-level interface to GPU
- hiding the details of parallel processing
- fast switching between GPU, CPU and OpenMP.

These new characteristics make CUDA thrust especially suitable for solving image analysis problems because the

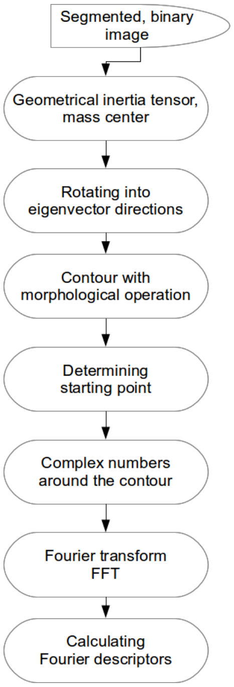
Segmented, binary image
Geometrical inertia tensor, mass center
Rotating into eigenvector directions
Contour with morphological operation
Determining starting point
Complex numbers around the contour
Fourier transform FFT
Calculating Fourier descriptors

Figure 4: *Extracting the Fourier descriptors.*

users do not need to bother with memory management functions and by writing a C-like code can concentrate on the core issues.

3.5. Classification

First we use for the classification different types of distance metric. Euclidean D_e and Manhattan D_M distances in multidimensional space:

$$D_e = \sqrt{(\hat{I}_0 - I_0)^2 + (\hat{I}_1 - I_1)^2 + \ldots + (\hat{I}_{49} - I_{49})^2} \qquad (6)$$

$$D_M = |\hat{I}_0 - I_0| + |\hat{I}_1 - I_1| + \ldots + |\hat{I}_{49} - I_{49}| \qquad (7)$$

Where $\hat{I}_i$ variables represent reference values of Fourier descriptors. We determine these constants with the analysis of the user's hand while the user performed the hand signs.

Then we use a support vector machine (SVM) for classification. SVM is a machine learning technique which uses hyperplanes to separate the characteristics of the images in multidimensional space. Properly choosing the kernel function nonlinear cases can also be solved by transforming them to linear cases.

A SVM carries out the classification by the selected feature vectors[10]. The classifier has 50 input parameters and returns the identifier of the recognized hand gesture. We used LIBSVM to build up our support vector machine[11]. The SVM kernel is a radial basis function.

For American Sign Language recognition and under stable lighting conditions there were no difference in the three classification techniques. Nevertheless with unexperienced users it is advisable to employ the SVM method.

4. Experimental results

The codes were first developed in a MATLAB environment using the official toolbox functions. Cropping of images was not necessary for enhancing the reliability of recognition. Optimizing the code and using C++/OpenCV the running times could be decreased see table 1.

Operation	Running time
Spatial segmentation	88.2 msec
Center of mass	28.2 msec
Temporal segmentation	13.4 msec
Extracting FD	352.4 msec
Classification (Euc.)	0.1 msec
Classification (Man.)	0.1 msec
Classification SVM	56.7 msec
Total with distance metrics	480.3 msec
Total with SVM	536.9 msec

Table 1: *Running time. Genius FaceCam 312 (640 × 480).*

Using CUDA thrust the running times of the spatial segmentation, computation of center of mass, temporal segmentation, and extracting of Fourier descriptors can be decreased significant see table 2. We can establish that we can reduce the running time using CUDA thrust with 378.5 msec.

Operation	Running time
Spatial segmentation	9.8 msec
Center of mass	2.8 msec
Temporal segmentation	1.7 msec
Extracting FD	89.3 msec

Table 2: *Running time using CUDA thrust. Genius FaceCam 312 (640 × 480).*

We tested the usability and the performance of our system up to the present with 3 users. These people were selected from our friends. The tests were performed with unexperienced users. The hand silhouette depends on the hand physiology and the length of the sleeve could also modify the contour. That is why preliminary training is necessary. In our first experiments we collected training samples from the users by making gesture snapshots.

We noticed that recognition rates are dependent on the users. Some unexperienced users have difficulties to realize one of the gestures. In our experiments, each user trained the system with 90 samples of hand signs. This means 9 samples per hand sign. The average recognition rates of hand signs are summarized in a so-called confusion matrix.

	A	B	C	D	G	H	I	L	V	Y
A	**88.8**	1.4	0	0	2.0	2.7	0	0	1.7	3.5
B	1.1	**91.6**	0	0	0	3.5	0	0	1.8	2.0
C	0.8	0	**91.7**	7.5	0	0	0	0	0	0
D	0.2	0	8.7	**87.9**	0	3.1	0	0	0	0
G	1.0	0	0	0	**83.3**	10.1	1.0	0	4.6	0
H	0.2	0	0	2.9	3.4	**81.2**	10.1	0	0	0
I	1.7	0.2	0	0	0	0	**89.1**	0	0	9.1
L	3.1	2.1	0	0	0	2.1	0	**83.3**	1.0	8.3
V	5.6	2.1	0	0	0	0	5.0	12.2	**75.1**	0
Y	8.5	0	0	0	0.3	0	0	0	10.4	**80.7**

Figure 5: *Confusion matrix with the classification method of SVM.*

In Figure 5 average recognition rates (%) are summarized in a confusion matrix. Rows indicate signs to be performed, while columns depict the distribution of recognition results.

Comparing our results to similar solutions, in[8] Table 1 only 7 gestures were detected with 81 - 96% efficiency. Here we demonstrated static signs only. In case of dynamic motions, like the the repeat action in[12], the error rate in Figure 5 can be significantly decreased. In[13] the performances of Fourier descriptors and Hu moments were tested for vision-based hand posture recognition. They have shown that Fourier descriptors outperform Hu moments. Hand posture gestures were detected with 60 - 96% efficiency with the help of Fourier descriptors.

5. Conclusion

The recognition of the static American Sign Language can be performed with a simple built-in camera. Running times make possible the usage in real-time environment.

CUDA thrust is a suitable way of decreasing the running time in parallel processing without the need of bothering with memory management and CPU, GPU hardware specifications.

Robustness, reliability and speed of static recognition is the base to recognize dynamic hand movements and converting them into text or spoken word.

6. Acknowledgement

I would like to express my thanks to my professor Dr. Ferenc Vajda for teaching and mentoring me. I also would like to thank Dr. Tamás Szirányi for his support and advices.

References

1. S. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 27*, (6):873-891, 2005.

2. J. Ravikiran, K. Mahesh, S. Mahishi, R. Dheeraj, S. Sudheender, and, N. V. Pujari Finger detection for sign language recognition. *In Proceedings of the International MultiConference of Engineers and Computer Scientists vol. 1*, 18:20, 2009.

3. R. Mapari and G. Kharat Hand gesture recognition using neural network. *International Journal of Computer Science and Network vol. 1, issue 6*, 2012.

4. M. K. Hu. Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory vol. 8*, pp. 179:187, 1962.

5. E. Persoon and K. Fu. Shape discrimination using Fourier descriptors. *IEEE Transactions on Man and Cybernetics vol. 8*, (3):388-397, 1986.

6. A. Albiol, L. Torres, and E. J. Delp Optimum color spaces for skin detection. *In proceedings of the 2001 International Conference on Image Processing vol. 1*, 122:124, 2001.

7. V. A. Oliveira and A. Conci Skin detection using HSV color space. *Workshops of Sibgraphi*, 1:2, 2009.

8. A. Licsár, T. Szirányi, L. Kovács, and B. Pataki A folk song retrieval system with a gesture-based interface. *IEEE Multimedia vol. 16*, (3):48:59, 2009.

9. D. Toth and T. Aach Detection and recognition of moving objects using statistical motion detection and Fourier descriptors. *In proceedings of the 12th International Conference on Image Analysis and Processing vol. 1*, 430:435, 2003.

10. K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks vol. 12*, (2):181:201, 2001.

11. C.-C. Chang and C.-J. Lin LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology vol. 2*, (3):1:27, 2011.

12. A. Licsár and T. Szirányi User-adaptive hand gesture recognition system with interactive training. *Image and Vision Computing vol. 23*, (12):1102:1114, 2005.

13. S. Conseil, S. Bourennane, and L. Martin Comparision of Fourier descriptors and Hu moments for hand posture recognition. *In proceedings of European Signal Processing Conference vol. 1*, 2007.