Robust Hand Gesture Recognition System Using Motion Templates

Shrikant Kulkarni, Manoj H and Sumam David
Dept of Electronics and Communication Engineering
National Institute of Technology Karnataka Surathkal
Mangalore 575025 India
Email: shrikant06,sumam@ieee.org

Venugopala Madumbu and Y Senthil Kumar
Texas Instruments
Bengaluru, India
Email: venugopala,ysenthil@ti.com

***Abstract*—This paper presents a robust hand gesture analysis system. The approach uses the video analytic technique of motion templates rather than conventional gesture recognition algorithms. Also, it utilizes background modeling and skin pixel detection which further strengthens the approach by making it tolerant to background clutter and noise. In addition it reduces the false detections to a considerable extent. The system does not necessitate the user to wear any coloured caps or gloves for the hands. Encouraging results were obtained and it was found that the methodology is flexible and can be manipulated to suit gesture based interaction as per the requirements of a system. It can also be implemented as a standalone system.**

I. Introduction

The primary goal of gesture recognition research is to create a system which can identify specific human gestures and use them to convey information or for device control. Vision based hand gesture interface has been attracting more attentions due to no extra hardware requirement except camera, which is very suitable for ubiquitous computing and emerging applications.

Most of the complete hand interactive systems can be considered to be comprised of three layers: detection, tracking and recognition. The detection layer is responsible for defining and extracting visual features that can be attributed to the presence of hands in the field of view of the camera. The tracking layer is responsible for performing temporal data association between successive image frames. Moreover, in model-based methods, tracking also provides a way to maintain estimates of model parameters, variables and features that are not directly observable at a certain moment in time. Last, the recognition layer is responsible for grouping the spatio-temporal data extracted in the previous layers and assigning the resulting groups with labels associated to particular classes of gestures.

A variety of approaches [1] are followed to implement each of the three layers above. Detection is usually done through shape or colour based techniques. Usually shape-based techniques need training data to learn the shape of hands while colour based techniques basically rely on skin pixel detection. Recognition is usually based on the definition of gestures which in turn is learnt by the system using methods such as Adaboost [2], Hidden Markov models or neural networks. Linear approximation, finite state machine/model matching and template matching have also been used.

There are a plethora of problems faced by a hand gesture recognition system in each of the above three layers such as background clutter, false detections, low frame rate and the list goes on. Many interesting solutions have been proposed to overcome them. Few algorithms assume that only the hand is present in the scene and the background is static. Some other methods require the user to wear coloured caps [3] or wrist bands or gloves [4] to aid the detection of hand in a varying background. Apart from that, many approaches require a huge number of training images from which they learn the gestures. Another approach is to use a complicated set of gestures that may be easy for the system to recognize but are equally hard for the user to enact.

Keeping an end-application in mind, it is observed that when a real-life hand gesture application is designed, it should be suitable for variety of environments without any constraints. A hand gesture recognition system for Human-Computer interaction is proposed in [5] where an application to control a powerpoint presentation is mentioned. But all of the results shown assume a plain, stagnant background without any clutter. Binh et. al. [6] proposes a robust method which is based on Pseudo Two-Dimension Hidden Markov Models (P2-DHMMs) which works well without environmental constraints. But the implementation of such complex systems as standalone can be very cumbersome.

The need to use gloves/markers or a plain background may result in a robust system, but they limit the applicability of the algorithm. Thus there arises the need to have a system that can adapt to any background and perform well without imposing too many restrictions on the user. Our system adheres to all these and does not even need to be trained. The user can wear any shirt, have any background and need not wear any gloves but still can communicate through gestures standing at a considerable distance from the camera. The combined use of Motion templates, background modeling and Haar face detection make it possible to develop such a robust system.

Such a robust system can become the backbone of many gesture-based interfaces that can find application in automatic

978-1-61284-671-2/11/$26.00 ©2011 IEEE

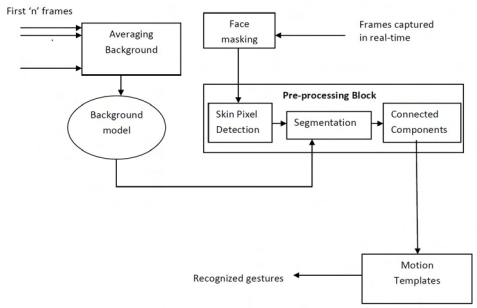

Fig. 1. Overview of proposed method

Fig. 2. An RGB Image

Fig. 3. Corresponding HSV Image

Fig. 4. Detection of Skin Pixels

ticketing, traffic monitoring, gesture-based door locking system, customer self-help kiosks at train/bus terminals, driving aids to assist pilots and other such scenarios in transportation. Wherever it is used, the user finds it more comfortable and easier to interact. Also, it helps in significant time-savings by automating some processes, which is a great advantage in public transport systems.The paper is organized as follows: Section II gives the details of the proposed algorithm, Section III discusses the implementation and results obtained . Finally, Section IV presents the conclusions that can be drawn from this investigation.

II. Proposed Method

The performance of vision based gesture interaction is prone to illumination changes, complicated backgrounds, camera movement and specific user variance. Hence, a system that uses various steps to extract a moving hand from a noisy environment and determine the direction and orientation of the motion is proposed in this paper. Fig. 1 presents an overview of the method.

The first step is the extraction of the regions of the image that resemble the skin for which a number of approaches are available. To reduce the computational burden, we have chosen a detection technique based on Double thresholding of skin pixels. A simple thresholding mechanism obviously fails due to lighting variations and other factors. To overcome this, thresholding was done in HSV colour-space [7]. A suitable upper and lower threshold was determined by trial and error method on all the three colour-spaces. This converts an image in RGB colour space (Fig.1) to HSV colour space (Fig.2). As a consequence, all the skin pixels in each of the video frames are detected. Hence, even the face is detected along with hands as seen in Fig. 4

A. Averaging Background

The target application is aimed at situations where the background is almost static. This necessitates the use of segmentation mechanism that differentiates the object from background. Simple background subtraction works fairly well for some scenes, but it does not yield satisfactory results since all the pixels are not independent.

In contrast, averaging background methods does not consider neighboring pixels while learning, but it models the variations of a pixel. To account for surrounding pixels, a multi-part model is learnt, which extends the basic independent pixel model. Thus, two models are learnt for each pixel: one when the surrounding pixels are bright the other when the surrounding pixels are dim. In this way, averaging background accounts for the surrounding context.

The background statistics are accumulated in first few frames (the number can be changed depending on the dynamics of the background)after the program starts. It is composed of four blocks:

- accumulation of frames over time;
- accumulation of frame-to-frame image differences over time;

Fig. 5. Experimental Environment

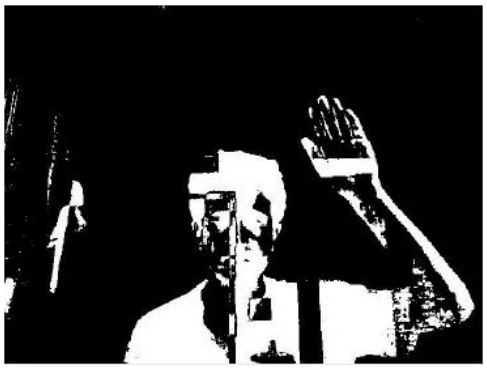

Fig. 6. Extraction of Foreground

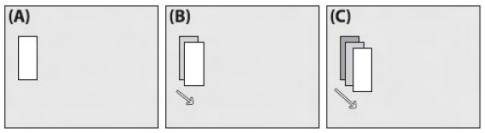

Fig. 7. Motion Templates

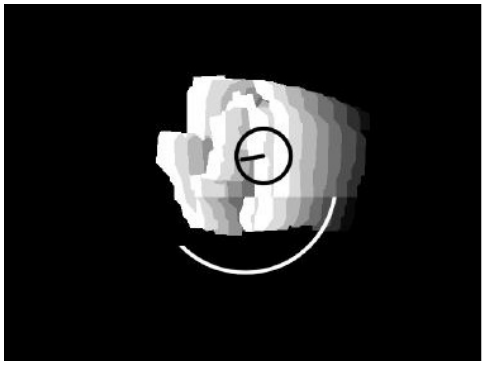

Fig. 8. Motion History Image of Hand

- segmentation of the image (once a background model has been learnt) into foreground and background regions;
- compilation of segmentations from different color channels into a single mask image

During the accumulation of the specified number of frames, the pixel values for each pixel in each of the frames is accumulated and the average pixel value p_{avg} for each pixel is determined. Similarly, the frame to frame difference for each of the successive frames is accumulated and then the average frame-to-frame absolute difference ff_{avg} is calculated. Now, any pixel *p* which is larger than the average value for that pixel p_{avg} by 7 times the average frame-to-frame absolute difference ff_{avg}, is considered to be a foreground pixel. Likewise, any pixel which is smaller than the average value for that pixel by 6 times the average frame-to-frame absolute difference, it is considered to be a background pixel. Two thresholds *t1* and *t2* are set accordingly which are used for segmentation by labeling the regions of the image as foreground and the background.

$$t1 = p_{avg} + (7 \times ff_{avg}) \quad (1)$$
$$t2 = p_{avg} - (6 \times ff_{avg})$$

The outcome of the above method is a binary image that differentiates background and foreground, the foreground in this case is the person who is using the interface through gestures. But the regions of interest are only those parts which are involved in making the gesture namely face and hands. To get this, the binary image obtained from skin pixel detection is combined with this image (using AND operation) to extract those foreground objects of the image which look like skin. But this image is prone to a lot of noise which needs to be eliminated. This is done with the help of connected component analysis.

B. Connected Component Analysis

In this operation, the morphological operations *open* (to shrink areas of small noise to 0) followed by the morphological operation *close* (to rebuild the area of surviving components that was lost in opening) are applied on a noisy input mask image. Thereafter, the large enough contours of the surviving segments are found and the statistics of all such segments are taken and the largest contour is retrieved.

At this point, a smooth and continuous estimate of moving hands and face is obtained. Fig. 6 is obtained as a result of background averaging and connected component analysis on the input frames as in Fig. 5. This marks the end of pre-processing steps and next step is the gesture recognition.

C. Motion Templates

The dual purpose of tracking the moving object to ascertain the direction of motion and fine tuning the segmentation is achieved by the application of motion templates [8]. The extraction of object silhouette acts as a pre-requisite for the application of motion templates. The binary image that underwent connected component analysis in the previous section is fed as silhouette to this. The visualization of motion templates can be seen in Fig. 7. As the white rectangle moves, new silhouettes are captured and overlaid with the (new) current time stamp. These sequentially fading silhouettes record the history of previous movement and are thus referred to as the

Fig. 9. False Detection due to face movement

Fig. 10. Masking of face

Motion History Image (MHI). Fig. 8 shows the situation when motion templates are applied to a hand in motion.

In MHI, pixels that are older than a pre-decided time duration are set to 0. Once the motion template has a collection of object silhouettes overlaid in time, gradient of the MHI image is taken. This indicates the overall motion tendency. Some gradients calculated in this manner are invalid when older or inactive parts of the MHI image are set to 0, producing artificially large gradients around the outer edges of the silhouettes. Still, a measure of global motion is obtained. The global motion can be computed from the center of mass of each of the MHI silhouettes, but summing up the pre-computed motion vectors is found to be much faster.

Now the regions of the motion template MHI image are isolated to determine the local motion within that region. When a region marked with the most current time stamp is found, the region perimeter is searched for sufficiently recent motion (recent silhouettes) just outside its perimeter. When such motion is found, a downward stepping flood-fill is performed to isolate the local region of motion that spilled off the current location of the object of interest. Once found, the local motion gradient direction in the spill-off region is calculated and that region is removed, and the process is repeated until all regions are found.

D. Haar Face Detection

This step is introduced to reduce the false detections. For example if the hand was moved in a particular way to

Fig. 11. Brightness variation in background

Fig. 12. Background motion

symbolize a gesture, the movement of the face interferes with it resulting in wrong detection as in Fig. 9. This has been overcome by masking the face using Haar classifier based face detection technique.

The core basis for Haar classifier object detection is the Haar-like features. These features, rather than using the intensity values of a pixel, use the change in contrast values between adjacent rectangular groups of pixels. The contrast variances between the pixel groups are used to determine relative light and dark areas. Two or three adjacent groups with a relative contrast variance form a Haar-like feature. Haar features can easily be scaled by increasing or decreasing the size of the pixel group being examined. This allows features to be used to detect objects of various sizes.

Once the face is detected, it is masked by using a rectangular blob as shown in Fig. 10. Thus, the face is effectively blocked and its movement does not interfere in the process.

III. Experimental Results

The environment used for testing was a room as shown in Fig. 5. The software was implemented in C and used OpenCV library for image processing. It was run on on a system with 2.1GHz Intel processor and 3 GB RAM. As stated earlier, the user does not require to use any coloured caps or wrist bands. The hand is extracted from the frames using the steps explained in the previous section and its movement is tracked using motion templates as shown in Fig. 8.

We basically distinguish four directions of motion namely UP, DOWN, LEFT and RIGHT. A threshold τ is used for this decision-making. It is nothing but the number of successive frames in which the given direction of motion has to be detected before declaring that the hand is moving in that particular direction. After initial trials, the value of τ was set to 4 for optimal functioning. A lower value results in false detections and a higher value results in non-detected motions. Table I shows the variation of successful detection for various values of τ.

As mentioned in the earlier sections, the most important factors that affect the accuracy of gesture recognition are back-

TABLE I
ACCURACY AT VARIOUS VALUES OF τ

τ	2	3	4	4
Detection percentage	62	79	92	81

TABLE II
ACCURACY AT VARIOUS DISTANCES FROM THE CAMERA

Distance (in ft.)	2	4	6	8	10
Detection percentage	91	88	84	74	63

ground noise/clutter, brightness variation, unwanted movement etc. Our methodology stays immune to these factors in most of the cases. It was tested in various noisy environments that had the above factors to assess its credibility.

Figure 11 shows one such situation in which there is a considerable amount of variation in the brightness within the frame. But as can be seen, our system is still able to extract the hand and track it with good precision. This, was possible due to the use of motion templates. It is very effective in capturing only those areas where there is motion. The skin pixel detection also aids this since it is done in HSV color space, change in brightness does not corrupt it significantly.

Similarly, Fig. 12 shows a situation in which there is a rotating fan in the background. It is an example of unwanted background motion that can deteriorate the accuracy of the system. But even in this case, good results have been obtained. In this case, the averaging background preserves the functionality. Although, the fan is rotating, the background information captured in the first few frames (which can be customized) creates a model which effectively tackles the issue.

Table II shows the accuracy of detection as the distance from the camera is increased. It can be seen that the performance is acceptable until a distance of upto 8 feet from the camera used for testing which was a 2 mega-pixel webcam. Usage of higher resolution camera may increase this distance, but the processing speed may get affected which will in turn affect real-time processing.

IV. CONCLUSION

The hand gesture analysis method presented here is based on motion templates.

The system developed is both flexible and robust one. The directions detected can be easily used in gesture-based interaction with any system that uses it as a backbone. It works well even when the camera is fixed at a distance from the user unlike other approaches which usually need the user to be as close to the camera as possible. The user need not wear any coloured caps or gloves on hands. The gestures can be simple movement of hands but don't require the fingers to be held in any particular way. All this provide enough freedom to the user and can be suitably utilized in the applications as per the requirements.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Texas Instruments(India) Ltd. and Mr.Aravindan K for their support to this technical work.

REFERENCES

[1] Prateem Chakraborty, Prashant Sarawgi, Ankit Mehrotra, Gaurav Agarwal, Ratika Pradhan, "Hand Gesture Recognition : A comparitive Study" *Proceedings IMECS*, Vol.I, pp 388-393, 2008.

[2] Yikai Fang, Kongqiao Wang, Jian Cheng and Hanqing Lu, "A real-time Hand Gesture recognition method" *Proceedings ICME*, pp 995-998, March 2007.

[3] Davis and Shah, "Visual Gesture Recognition" *Proceedings IEEE Image and Signal Processing*, Vol. 141, No. 2, pp 101-106, 1994.

[4] B. Bauer and H. Hienz, "Relevant feature for video-based continuous sign language recognition", *Proceedings Fourth International Conference on Automatic Face and Gesture Recognition*, pp 440-445, 2000

[5] S. Mohamed Mansoor Roomi, R. Jyothi Priya and H. Jayalakshmi, "Hand Gesture Recognition for Human-Computer Interaction", *Journal of Computer Science*, pp 1002-1007, 2010.

[6] Nguyen Dang Binh, Enokida Shuichi, Toshiaki Ejima, "Real-Time Hand Tracking and Gesture Recognition System", *Proceedings GVIP*, pp 362-368, December 2005.

[7] A. Albiol, L.Torres and E.J. Delp, "Optimum Color Spaces for Skin Detection", *Proceedings of International Congress on Image Processing*, 2001.

[8] A. Bobick and J. Davis, "Real-time recognition of activity using temporal templates", *IEEE Workshop on Applications of Computer Vision*, pp 39-42, December 1996.