Visual gesture recognition for ground air traffic control using the Radon transform

Meghna Singh, Mrinal Mandal, Anup Basu

Visual Gesture Recognition for Ground Air Traffic Control using the Radon Transform

Meghna Singh[1], Mrinal Mandal[1] and Anup Basu[2]
[1]*Department of Electrical and Computer Engineering*
[2]*Department of Computing Science*
University of Alberta
Edmonton, T6G 2V4, Canada
{meghna,mandal}@ece.ualberta.ca, anup@cs.ualberta.ca

***Abstract* - Human gesture recognition is an active topic of vision research which has applications in diverse fields such as collaborative virtual environments and robot tele-operation. We propose a novel method for the recognition of hand gestures, used by air marshals for steering aircraft on the runway, using the Radon transform. Various aspects of the algorithm, including acquisition, segmentation, labeling and recognition using the parametric Radon transform are addressed in this paper. A binary skeleton representation of the human subject is computed. The Radon transform is used to generate maxima corresponding to specific orientations of the skeletal representation. Feature vectors are extracted from the transform space by computing the normalized cumulative projections of the Radon transform on the angle axis. K-means clustering is then applied to recognize static gestures from the extracted features. This technique has the potential to provide information about the exact orientation of gesture segments and can find use in ground control of unmanned air vehicles. Experiments with image data corresponding to the various ground air traffic control gestures used in directing aircrafts, highlight the potential application of this approach.**

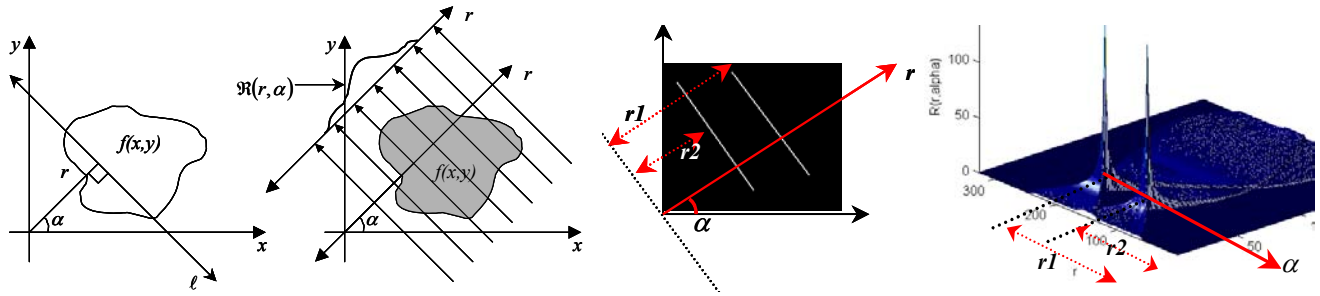
Index Terms – Gesture recognition, pose recognition, Radon transform, robot tele-operation.

I. INTRODUCTION

The increasing interaction between man and machine has lead to research in the area of man–machine interfacing. The various approaches to this interfacing can be divided into intrusive (such as haptic devices and data gloves) and non-intrusive approaches (based on computer vision techniques). These two approaches have their own pros and cons. The intrusive approach provides a higher spatial and temporal resolution to the man–machine interface, while the non-intrusive approach is more practical for cases where it is difficult to interact with devices such as markers and haptic gloves. Human gesture recognition is one of the many domains of man–machine interfacing and involves understanding and recognition of the gestures involved in either interpersonal communication (sign language) or man–machine communication (specific dictionary of gestures/poses for robot tele-operation).

Robot tele-operation has been conducted in the past via sensor gloves, keyboard interface and joysticks. Other approaches have been to learn a particular dictionary of visual commands, which are acquired and acted upon by the robot. In the latter approaches, gesture recognition is an integral part of non-verbal communication with the robot. Iba *et al*. [1] use a data glove and HMMs to recognize six hand gestures that are used to control a robot. Fong *et al*.[2] use gesture driven control and haptic devices for robot tele-operation. Their gesture commands however appear to be view dependent such that forward and reverse commands will be distinguishable only when viewed at an offset angle. In non intrusive works, Ugalde *et al*. [3] classify segmented hand gestures based on the correlation coefficient of wavelet coefficients of trained and test sequences. Hu *et al*. [4] use a set of seven gestures to issue commands such as turn left, start and stop. They evaluate various existing techniques to determine the most suitable algorithm for recognition. Bradiski *et al*. [5] use motion history images to compute the current gesture of the object for recognition. Chang *et al*. [6] use Mealy machines and Markov processes to interpret commands issued by human leg movement for virtual reality applications. The above-mentioned techniques have either high complexity thus making them unsuitable for real time operations or have low correct recognition rates. Some of the approaches such as haptic devices and data gloves are intrusive, and are not suitable for a more dynamic environment. In this paper, we propose a visual gesture recognition algorithm based on the Radon transform of skeletal lines of the arm gestures/poses. The proposed technique provides a high correct recognition rate and has low complexity. We also present a particular application of our proposed method in ground air traffic control. Another anticipated application is gesture based control of robots and unmanned aircrafts/vehicles.

This paper is organized as follows. In Section II we review the Radon transform. We present our proposed technique for pose recognition in Section III. Performance evaluation is presented in Section IV. Finally, conclusions are presented in Section V.



(a) (b) (c) (d)

Fig.1: (a) Pictorial representation of line used in (1), (b) Illustration of RT as an integration of the image; RT of two parallel lines - (c) Skeleton image, (d) Mesh plot of RT of parallel lines in (c).

II. Radon Transform

J. Radon first introduced the Radon transform (RT) in a classic paper in 1917. The RT of a function $f(x,y)$, can be defined as

$$\Re_f(r,\alpha) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(r - x\cos\alpha - y\sin\alpha)dxdy \qquad (1)$$

where r is the perpendicular distance of a line ($r=x\cos\alpha + y\sin\alpha$) to the origin and α is the angle formed by the distance vector to the line (see Fig.1(a)). The delta function in (1) evaluates to a non-zero value only for points that lie on this line. The RT is essentially the description of a function in terms of its integral projections (see Fig.1 (b)). For example, two parallel lines in the image space will translate to two maxima at the same angle α but different distances from the origin r (Fig.1 (c) and (d)) in the Radon transform space.

In the past, Radon transform (RT) and Hough transform (HT) have been used in various pattern recognition applications such as line detection, facial expression recognition and symbol recognition [7–9]. It has also been proven that while the HT can be considered a special case of the RT in continuous space, they are non-identical in discrete form [10]. The deficiencies with using the HT are the distortion in shape and location of peaks in transform space, and problems in transforming straight lines that pass through the origin. Also, the standard implementation of the discrete HT via accumulator bins creates issues. If the bins are too small then intersection of the transform curves of collinear image points may fall in separate bins, and if the bins are too large, then parallel lines close together will lie in the same bin. It is because of these associated issues with the HT that we use the RT instead, in our technique for gesture recognition. To the best of our knowledge the RT has never been used before for human gesture recognition.

III. Proposed Technique

In this section we describe the proposed technique for gesture recognition using the RT. The flowchart in Fig.2 illustrates the primary steps of this technique (these are described in detail in the following subsections). Our methodology depends on the following two assumptions:

1. For the purpose of gesture recognition, any gesture of the arms can be replaced by the medial skeleton representation of that gesture.
2. The representation of the medial skeleton of a gesture in RT space provides information concerning the classification of that gesture.

We acquire images in which subjects perform the arm gestures corresponding to ground air traffic control gestures. These images are processed to compute a binary representation of the arm segments in each image. Medial skeletonization is then applied to these binary images to obtain a skeleton representation of each gesture. The parametric representation of these skeleton images is obtained by using the RT. Feature vectors are subsequently extracted from the Radon transform and clustered for the recognition of unknown gestures. The proposed technique can be applied to a large number of applications such as gait analysis and robot-teleoperation. However, in this paper we demonstrate the application of the proposed technique for ground air traffic control.

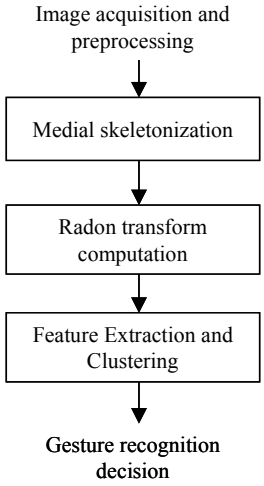
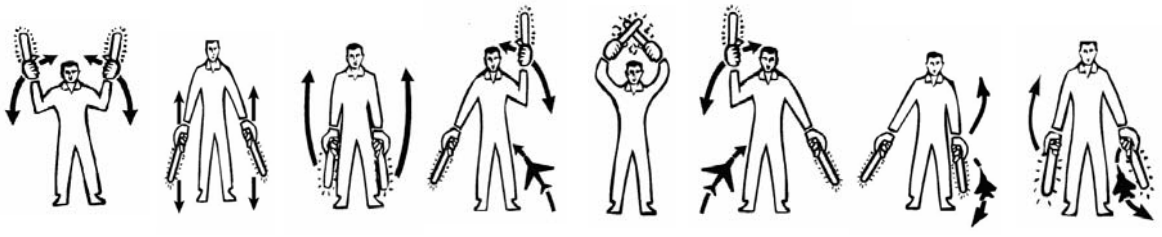


Fig.2: Flowchart of proposed technique.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig.3: Ground air traffic control gestures as designated by the United States Air Force [11]. (a) Move ahead, (b) Slow down, (c) Reverse, (d) Turn Left, (e) Stop, (f) Turn Right, (g) Reverse Left, and (h) Reverse Right.

A. Image Acquisition and Preprocessing

The proposed algorithm is designed to recognize eight ground air traffic control gestures. Images are acquired such that they correspond to the gestures indicated in Fig.3. We impose restrictions such that the motion of the upper torso and legs is minimal and most of the motion occurs only in the arms. More robust techniques such as human body models can also be implemented to acquire motion of just the arm segments; however, complexity will increase with the addition of a body model. Once the images have been acquired, the foremost step of any vision based recognition algorithm is background-foreground separation.

We perform background separation by using a statistical background modeling approach. We assume a relatively stationary background (assuming noise at each pixel to be time varying) and use an adaptive threshold $\tau(x,y)$ for each pixel $p(x,y)$. Each pixel of the current frame is thesholded against the corresponding pixel of the background model to extract foreground information. The mean intensity $\mu(x,y)$ at location (x,y), corresponding to the '*N*' initial frames is computed as: $\mu(x,y)=\frac{1}{N}\sum_{i=1}^{N} p(x,y;i)$

where, $p(x,y;i)$ is the pixel value at location (x,y) in the 'i^{th}' frame. The threshold for each pixel in the background model is calculated using the following equation:

$$\tau(x,y)=\max\left\{\left|\mu(x,y)-p(x,y;i)\right|\right\} \quad \text{for} \quad (1\le i\le N)$$

To obtain the foreground we classify each pixel in frame '*k*' ($k>N$) according to the following inequality:

if $\left|p(x,y;k)-\mu(x,y)\right|<\tau(x,y)$

pixel is background $p(x,y;k)=0$

else

pixel is foreground $p(x,y;k)=255$

This preprocessing generates binary images for arm poses, which are morphologically processed to fill any holes that may be present in the foreground (Fig.4).

B. Medial Skeletonization

Morphological operation called thinning is then applied to these binary images to obtain the medial skeleton of each segment (Fig.4). We can now consider these medial skeleton images of size $X\times Y$ to be a special discrete function $f_S(x,y)$ of the general continuous function $f(x,y)$ (see (1)), where

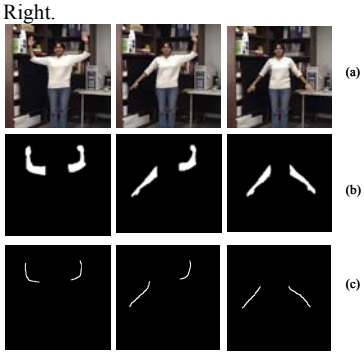


Fig.4: Sample images showing (a) original data, (b) binary image data and (c) results of the medial skeletonization operation.

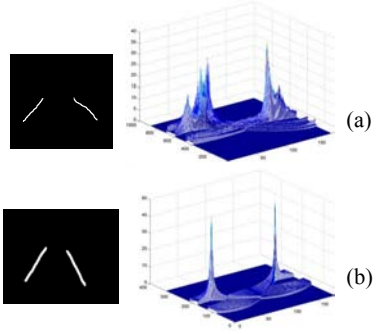
$$f_S(x,y)=\begin{cases}1, & \textit{if part of skeleton}\\ 0, & \textit{otherwise}\end{cases} \qquad (2)$$

C. Radon Transform Computation

Subsequent to the extraction of the skeletal lines, the RT is used to detect the orientation of these lines. The discrete version of the RT of skeletal images is then computed as the following summations:

$$\Re_f[r,\alpha]=\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1} f_s(x,y)\delta(r-x\cos\alpha-y\sin\alpha) \qquad (3)$$

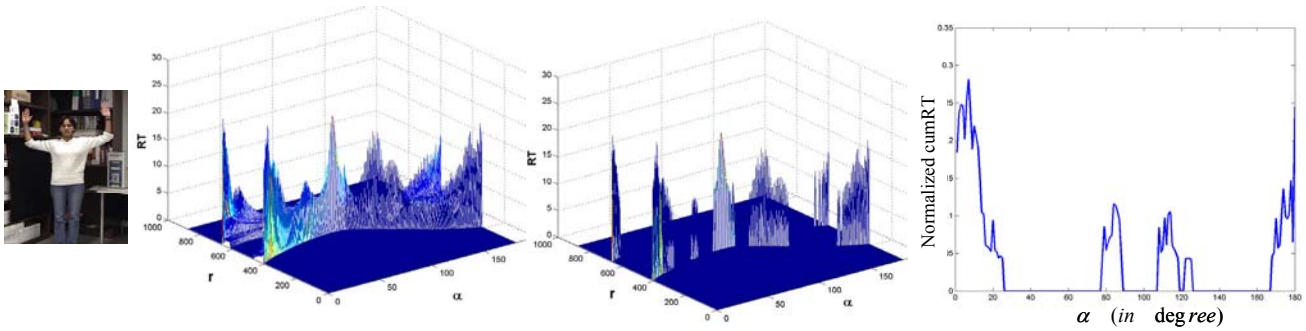
where $\delta(.)$ denotes the Kronecker delta function. It is observed in RT space (Fig.5) that there are maxima corresponding to the location and angle of the medial skeleton in image space. In real cases the maxima are not very localized and tend to be slightly diffused. This is because the skeletons are rarely straight lines but rather a collection of connected line segments varying slightly in their orientation (Fig.5). Feature vectors must now be extracted from the RT space for further processing.



(a)

(b)

Fig.5: RT of skeletal lines. (a) Actual data and its diffused RT, (b) Ideal data and its sharp RT.



(a) (b) (c) (d)

Fig.6: (a) Acquired image, (b) RT of (a), (c) Thesholded RT and (d) Normalized cumRT

D. *Feature Extraction and Clustering*

Once the Radon transform has been computed, we threshold the Radon transform coefficients to extract the most significant local regions. The threshold τ is computed as the mid value of the range of $\Re$.

$$\tau = \frac{1}{2}\left(\max\{\Re[r,\alpha]\} - \min\{\Re[r,\alpha]\}\right) \tag{4}$$

The Radon transform space is then thresholded as follows: *If* $\Re[r,\alpha] \geq \tau$ *then* $\Re[r,\alpha] = \Re[r,\alpha]$ *else* $\Re[r,\alpha] = 0$. Since it is desired that this technique be translation independent we compute the cumulative projections of the Radon transform on the α axis denoted hence by cumRT as follows:

$$cumRT[\alpha] = \sum_{i=0}^{r} \Re[i,\alpha] \tag{5}$$

We also normalize these projections ($normRT$) such that the distributions lie between 0 and 1 (see Fig.6).

$$normRT[\alpha] = \frac{cumRT[\alpha]}{\sqrt{\sum_{\alpha=0}^{180} cumRT^2[\alpha]}} \tag{6}$$

We implement K-means clustering with K=8 clusters, one cluster for each gesture that is to be recognized. From the set of acquired image data, 8 images (one for each gesture) are selected as the known gesture set. The remaining image data is selected as the unknown or test data set. All steps of the algorithm (Fig. 2) till K-means clustering are computed for both the known and unknown gesture sets. The clustering however is done only on the normalized cumRT of the unknown gesture set. In order to recognize the gestures that are clustered together, we compute the nearest neighbor of the cluster centroid in the known gesture set.

IV. Performance Evaluation

In order to evaluate the performance of our proposed method, we captured test images of size 640x480 using a SONY DCR-PC100 CCD camera. The images were then converted into 8-bit gray scale format using MATLAB. Five subjects were asked to perform randomly a set of arm gestures from the ground air traffic control gesture set. Subsequent to image acquisition the proposed algorithm was implemented in MATLAB6.5 R13 on a Pentium IV 3.2GHz processor, 1GB RAM.

Overall 86 test images were used for performance evaluation and 76 of those test images were correctly recognized giving a correct recognition rate (CRR) of 88% (see Table-I). For relative performance comparison we also implemented two techniques proposed by Hu *et al.* [4] for gesture recognition, which include template matching (via block matching, geological transform) and skeletonization (via vectorization and angle computation). The recognition rates achieved with [4] are also indicated in Table I.

TABLE 1: RECOGNITION RATE OF PROPOSED ALGORITHM. **CRR**: CORRECT RECOGNITION RATE.

True gesture	# of test images	Correct recognition rate in %		
		Template matching [4]	Vectors [4]	Proposed RT
Move ahead	12	100	75	80
Right turn	10	100	100	90
Reverse right	10	40	100	100
Reverse left	10	40	80	100
Reverse	12	0	100	92
Left turn	10	100	100	80
Stop	10	100	0	60
Slow	12	50	100	100
Overall	86	**65%**	**82%**	**88%**

Fig. 7 shows a scatter plot representation of the performance of the proposed algorithm for 43 test images. The horizontal axis represents the test image number of the unknown gesture image. The vertical axis represents the eight gestures being identified. The legend for the gestures is as follows:
RT: *Right Turn,* ***ST***: *Stop,* ***RR***: *Reverse Right,* ***LT***: *Left Turn*
MA: *Move Ahead,* ***RL***: *Reverse Left,* ***RV***: *Reverse,* ***SW***: *Slow*
It can be seen from Fig.7 that most of the gestures in the unknown images were clustered to their correct known gesture. The images whose gestures were erroneously clustered are indicated with a cross symbol ⊗. A confusion matrix of the results is presented in Table II.

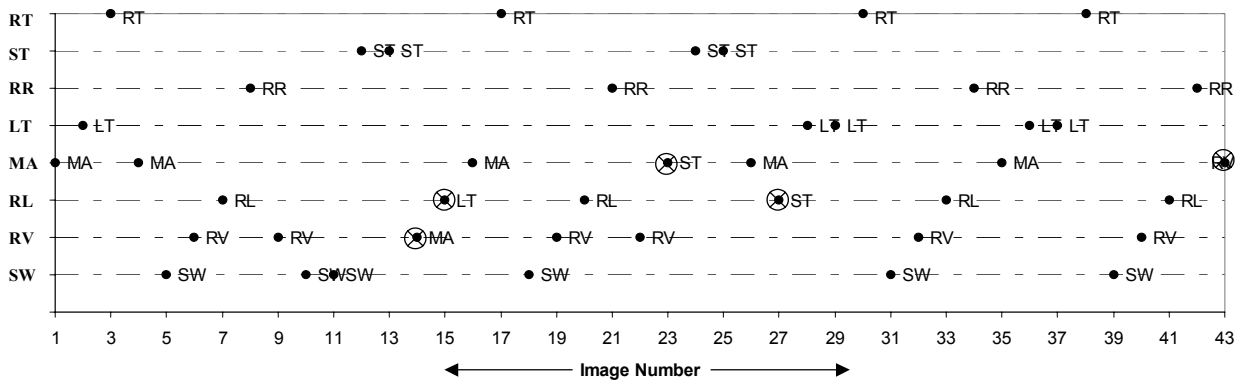
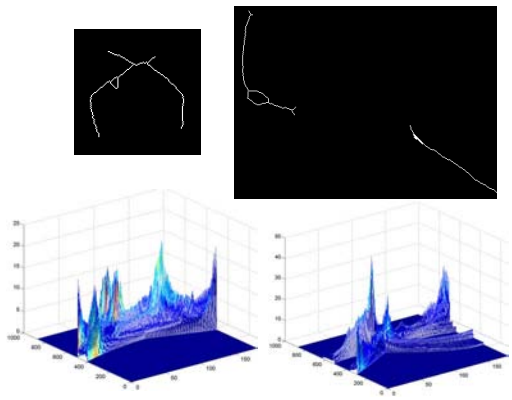


Fig.7: Scatter plot representation of the performance of the proposed algorithm for the first 43 test images. The horizontal axis represents the unknown gesture image number. The vertical axis represents the eight gestures that were identified. The script next to each test image point represents the actual gesture performed in the image. The legend for the gestures is as follows: ***RT***: *Right Turn,* ***ST****: Stop,* ***RR****: Reverse Right,* ***LT****: Left Turn,* ***MA****: Move Ahead,* ***RL****: Reverse Left,* ***RV****: Reverse,* ***SW****: Slow.* The images whose gestures were erroneously clustered are indicated with a cross symbol ⊗, all other recognition decisions are correct.

TABLE II: CONFUSION MATRIX OF THE ACTUAL AND PREDICTED GESTURES.

PREDICT→ / ACTUAL↓	RT	ST	RR	LT	MA	RL	RV	SW
RT	9	1	0	0	0	0	0	0
ST	0	6	0	0	2	2	0	0
RR	0	0	10	0	0	0	0	0
LT	0	0	0	8	0	2	0	0
MA	0	0	0	0	10	0	2	0
RL	0	0	0	0	0	10	0	0
RV	0	0	0	0	1	0	11	0
SW	0	0	0	0	0	0	0	12



(a) (b)

Fig. 8: Two incorrectly clustered images and their Radon transforms. (a) Stop image and (b) Turn right image.

Fig. 8 shows two of the test images that were incorrectly classified. It is observed that due to poor segmentation and thinning the medial skeleton is no longer a unitary segment but has loop like regions. These act as noise sources and cause a larger than usual Radon transformation. Thus when the Radon transform is thresholded (see section III D) these noise sources are not eliminated and get incorporated into the cumulative Radon transform and subsequently the normalized cumRT. Thus, noisy feature vectors lead to misclassification. However, these loop formations were rare and did not bias the clustering to one particular class. As part of future work, we plan to address this problem. It was also observed during evaluation that the actual RT representation of gesture for 'stop' was much different from ideal (see Fig.9), but to our advantage it was also much different from all the other gestures used. Therefore, gestures corresponding to stop were clustered satisfactorily.

V. CONCLUSIONS

In this paper, we have proposed an algorithm for the recognition of human arm gestures using the Radon transform. A particular application of recognizing ground air traffic control gestures is identified and tested. While there are other template-based approaches in contemporary literature that are being used for gesture recognition, these approaches are computationally more expensive and also require normalization of the images. Our approach on the other hand does not require any normalization of the image itself. The promising recognition rates achieved with the parameterized Radon transform emphasize the feasibility of the proposed method. In the future, we plan to use the orientation information derived from the Radon transform space to reconstruct motion for virtual models.

ACKNOWLEDGEMENT

The participation of volunteers in the experiments is gratefully acknowledged.

REFERENCES

[1] S. Iba, M.V. Weghe, C. Paredis and P.Khosla, "An Architecture for Gesture Based Control of Mobile Robots", *Proc. of the IEEE/RSJ IROS'99*, Vol.2, pp 851-857, Oct 1999.

[2] T.W. Fong, F. Conti, S. Grange and C. Baur, "Novel Interfaces for Remote Driving: Gesture, Haptic and PDA", *SPIE Telemanipulator and Telepresence Technologies VII*, Boston, MA, Nov 2000.

[3] F. Garcia-Ugalde, D. Gatica-Perez and V. Garcia-Garduno, "Segmentation and Classification of Hand Gestures for Man-machine Communication", *Proc of 2nd European Conference on Disability, Virtual Reality and Associated Technologies,* Sweden, Sept 1998.

[4] C. Hu, M.Q. Meng, P.X. Liu and X. Wang, "Visual Gesture Recognition for Human Machine Interface for Robot Teleoperation", *Proc. of 2003 IEEE/RSJ IROS'2003,* Las Vegas, Nevada, 2003.

[5] G. R. Bradski and J.W. Davis, "Motion Segmentation and Pose Recognition with Motion History Gradients", *Machine Vision and Applications*, Vol. 13, Issue3, 2002.

[6] C. Chang and H.W. Tsai, "Vision Based Tracking and Interpretation of Human Leg Movement for Virtual Reality Applications", *IEEE Trans on CSVT*, Vol.11, No.1, Jan 2001.

[7] V. F. Leavers, "Use of Two-Dimensional Radon Transform to Generate a Taxonomy of Shape for the Characterization of Abrasive Powder Particles", *IEEE Trans. on PAMI*, Vol. 22, No. 12, Dec 2000.

[8] N. Tsapatsoulis, Y. Avirithis and S. Kollias, "On the use of Radon Transform for Facial Expression Recognition", *Proc. of ISAS '99*, Orlando, FL, USA.

[9] S. Tabbone and L. Wendling, "Technical Symbol Recognition Using the Two-dimensional Radon Transform", *Proc. of 16th ICPR*, pp.200-203, Vol. 3, 2002.

[10] K.V. Hansen and P.A. Toft, "Fast Curve Estimation Using Preconditioned Generalized Radon Transform", *IEEE Trans. on IP*, Vol5. No12. Dec 1996.

[11] *http://www.e-publishing.af.mil/pubfiles/af/11/afva11-224/afva11-224.pdf*

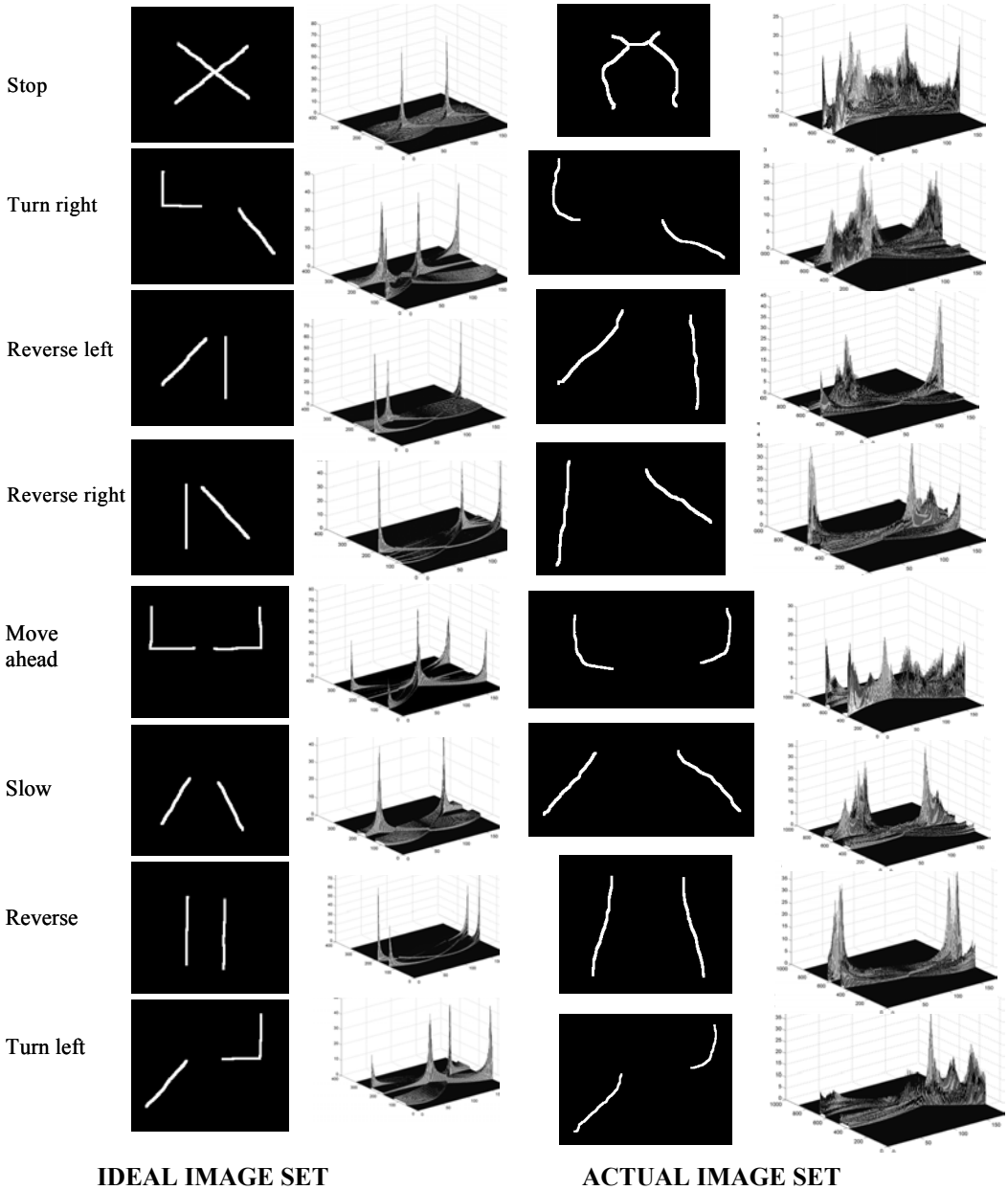


Fig.9: Ideal image set and actual image set for eight gestures. Ideal image set generates sharper peaks in the Radon transform space. Actual image set illustrates the non-linearity of the computed medial skeletons and the similar representations in Radon transform space.