SIGN VIDEO SEGMENTATION USING REGION, BOUNDARY BASED ACTIVE CONTOURS WITH SHAPE PRIORS

P.V.V.Kishore[1], E.Kiran Kumar[2] , B.Manjula[2] and P.Rajesh Kumar[1]

[1]Department of Electronics and Communication Engineering, Andhra University
pvvkishore@gmail.com
[2]Department of Electronics and Communication Engineering, DIET, Anakapalli
kiraneepuri@gmail.com,manjulabammidi@yahoo.co.in

ABSTRACT

In this paper we proposed a new and improved concept of segmentation for sign language gestures. The algorithm presented extracts signs from video sequences under various non-static backgrounds. The signs are segmented which are normally hands and head of the signing person by minimizing the energy function of the level set fused by various image characteristics such as colour, texture, boundary and shape information. Three color planes are extracted from the RGB and one color plane is used based on the environments presented by the video background. Texture edge map provides spatial information which makes the color features more distinctive for video segmentation. The boundary features are extracted by forming image edge map form the existing color and texture features. The shape of the sign is calculated dynamically and is made adaptive to each video frame for segmentation of occlude objects. The energy minimization is achieved using level sets. Experiments show that our approach provides excellent segmentation on signer videos for different signs under robust environments such as diverse backgrounds, sundry illumination and different signers.

KEYWORDS

Sign Language, Video Segmentation, Color/Texture extraction, Boundary Information, Shape Extraction, Level Sets

1. INTRODUCTION

Sign language [1,2] is the basic mode of expression and communication for deaf people. Sign language involves hand shapes, hand tracking, hand orientation with respect head and other body parts, along with head movements and facial expressions.

The primary challenge faced by any sign language recognition system is the ability to track the signer in the video of the signer with a variety of background clutter. The backgrounds used so far by many researchers are simple.

In background subtraction [3] various algorithms have been proposed such as frame differencing, adaptive mean filtering, adaptive median filtering and Gaussian Mixture Models (GMM). Of all the proposed algorithms GMM gives good segmentation levels for all kinds of complex video backgrounds. But the main factor in GMM is the speed of operation. GMM is slow compared to other methods providing a good segmentation result. GMM estimates the probabilities of foreground objects and background in to two different classes on each pixel and based on the

Natarajan Meghanathan, et al. (Eds): SIPM, FCST, ITCA, WSE, ACSIT, CS & IT 06, pp. 67–78, 2012.
 DOI : 10.5121/csit.2012.2308

maximum likelihood estimation it extracts the object class. The algorithm so far works well for surveillance videos but cannot show promising results for sign videos.

Active contours or popularly known as 'snakes' is a active research area with applications to image and video segmentation predominantly to locate object boundaries. They are also used for video object tracking applications. Chan and Vese (CV Model) [12] proposed a new level sets method based on Mumford-Shah distance for image segmentation. CV Model for level sets does not necessarily consider gradient for stopping the curve evolution.

In general the object segmentation using active contours based level sets face a few challenges considering the videos they are applied on. Firstly the contours get easily distracted if the background in the video sequence contains clutters which the case of sign language videos under real time environments. Second problem is when working with intensity images it becomes difficult to locate true boundaries of objects under varied lighting and the objects mostly blends with the background of the image.

The segmentation is devised by minimizing a force function which is a combination of color, texture, boundary and prior shape information of the objects to find their boundaries in all the video frames. The color and texture information is formulated by separating out foreground objects from background by minimizing the distance between them. The boundary information is calculated by using a gradient operator, which enables the contour to align itself to the edges of objects in the image.

2. SIGN VIDEO SEGMENTATION MODULE

The video image sequence segmentation proposed to extract hands and head segments of the signer form a variety of video backgrounds under different lighting conditions with diverse signers. A video sequence is defined as a sequence of image frames $\mathcal{I}(x, y, t): \mathfrak{D} \rightarrow \mathbb{R}$,where the images change over time. Alternatively a succession of image frames can be represented as $\mathcal{I}^{(n)}$ where $0 \leq n \leq \infty$. The basic principle behind our proposed segmentation technique is to localize the segmentation of one or more moving objects of then^{th}frame from the cues available from previous segmented frames $\mathcal{I}^{(1)}, \mathcal{I}^{(2)} \ldots\ldots\ldots\ldots \mathcal{I}^{(N)}$such that subsequent contours $\mathcal{U}^1, \mathcal{U}^2 \ldots\ldots\ldots\ldots \mathcal{U}^N$are available. Our proposed video segmentation algorithm segments hands and head of signers using color, texture, boundary and shape information about the signer given precedent understanding of hand and head shapes from $\mathcal{I}_f^{(n-1)}$and $\mathcal{I}_b^{(n-1)}$. The outline of the algorithm in the form of a block diagram is shown in figure1.

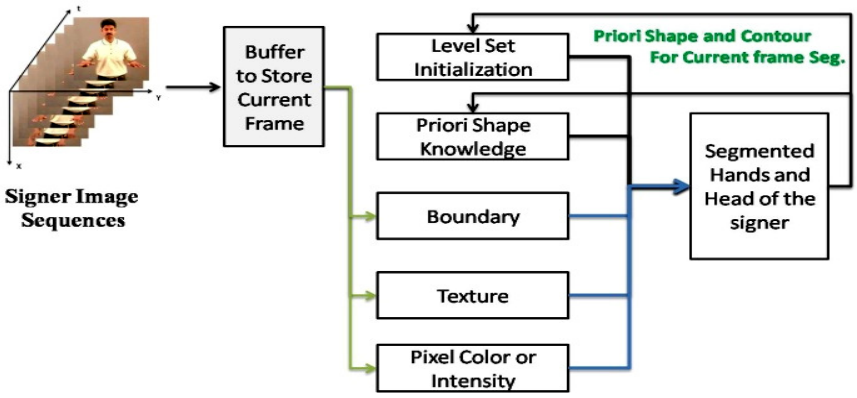

Figure 1: Process flow block representation of the segmentation algorithm

A. Color and Texture Features Module

Color plays a vital role in segmentation of complex images easily. There are various color models, but the RGB color model is most common in video acquisition. We choose manually the color plane which highlights the human object from a background of clutter. Once a color plane is identified, texture features are calculated using coorelogram of pixel neighbourhood. Texture is an irregular distribution of pixel intensities in an image. Allam.et.al established that co-occurrence matrix (CM's) produce better texture classification results than other methods. Gray Co-occurrence matrix (GLCM) presented by Haralick.et.al [11] is most effectively used algorithm for texture feature extraction for image segmentation.

Let us consider a color plane of our original RGB video. The R color plane is now considered as a $M \times N$ R coded 2D image. The element of co-occurrence matrix $C_{d,\theta}$ defines the joint probability of a pixel x_i of R color intensity r_i at a distance d and orientation θ to another pixel x_j at R color intensity r_j.

$$C^{d,\theta} = P_r\{I(Z_1) = r_i \wedge I(Z_2) = r_j : |z_1 - z_2|_\theta = d\} \tag{1}$$

where $|z_1 - z_2|_\theta$ gives the distance between pixels. For each co-occurrence matrix, we calculate four statistical properties: contrast (C), correlation (CO), energy (EN) and homogeneity (H) defined as follows

$$C = \sum_{i,j} |i - j|^2 C^{d,\theta}(r_i, r_j) \tag{2}$$

$$CO = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)}{\sigma_i \sigma_j} C^{d,\theta}(r_i, r_j) \tag{3}$$

$$EN = \sum_{i,j} \left(C_{d,\theta}(r_i, r_j)\right)^2 \tag{4}$$

$$H = \sum_{i,j} \frac{(C^{d,\theta}(r_i, r_j))}{1 + |i - j|} \tag{5}$$

Finally a feature vector $f^{vect}(x)$ is produced which is a combination of any one or all of the color planes and texture vector. Thus $f^{vect}(x) = \{f_1(x), f_2(x) \ldots \ldots \ldots f_n(x)\}$ the feature vector contains color and texture values of each pixel in the image. We classify them as background and foreground pixels using K-Means clustering algorithm. The centroids S_c of each group are used to identify each of the k –clusters where $c = 1,2 \ldots N, for\ all\ N \geq 1$. For every new classification the difference between the new vector and all the centroids is computed. The centroid corresponding to smallest distance is judged as the vector that belongs to the group.

$$d = \min_c \left| f^{vect}(x) - S_c \right| \tag{6}$$

d is the distance of every new $f^{vect}(x)$ of each frame to the previously computed centroids.

In the first frame $I_{(1)}$ all objects and background clusters are created. The object region contains three clusters of foreground $C^{(1)}{}_{f_i}$, where i =1 to 3 and background region $C^{(1)}{}_{b_j}$, where j =1 to 2 into two clusters**.** To move the contour on to the objects of interest we minimize the following energy function $\mathbf{E^{CT}}$from color and texture according to the initial object contour $\mathbf{\upsilon^{obj}}$

(7)

$$E^{CT}(\square_{obj}) = \sum_{i=1}^{3} \int_{obj} D(C^{(n-1)}{}_{f_i}, C^{(n)}{}_{f_i})dx + \sum_{j=1}^{2} \int_{bck} D(C^{(n-1)}{}_{b_j}, C^{(n)}{}_{b_j})dx$$

where$\boldsymbol{C_{f_i}^{(n-1)}}$and $\boldsymbol{C_{b_j}^{(n-1)}}$are object and background centroids of previous frame. $\boldsymbol{C_{f_i}^{(n)}}$and $\boldsymbol{C_{b_j}^{(n)}}$are object and background clusters from current frame. The $(n-1)$ frame cluster centroids will become the $\boldsymbol{n^{th}}$ frame initial centroid and the object contour is moved by minimizing the Euclidean distance between the two centroids.

B. Object Boundary Module

Poor lightning can impact image region information in a big way. Hence we use boundary edge map of the image objects which only depends on image derivatives. The way out would be to couple the region information in the form of color and texture features to boundary edge map to create a good segmentation of image objects.

We define the boundary $\boldsymbol{B^{obj}(x)}$ as pixels that are present in edge map of the object. The boundary pixels can be calculated by using gradient operator on the image. To align the initial contour $\upsilon^{\mathbf{obj}}$ from previous frame to the objects in the current frame to pixels on the boundary we propose to minimize the following energy function

$$E^{O}(I^{f}) = \int_{arc(Length\ of\ obj)} g(B^{o}(x))dx \tag{8}$$

Where $\mathrm{arc}(\mathrm{L}^{\mathrm{obj}})$ is the length of the object boundary. The function g is an edge detection function. The boundary energy reaches to a minimum when the initial contour aligns itself with the boundary of the objects in the image.

C. Shape Influence Module

Even with color, texture and boundary values of pixels in the image, the greatest challenge comes when object pixels and background pixels share the same color and texture information. The following method is used to construct the influence of shape of non-rigid objects in the image sequence. As for the first fame $\boldsymbol{I^{(u)}}$ where prior shape information is not available we just use the region and boundary information for segmentation. For $\boldsymbol{I^{(n)}\ \forall\ n \geq 1}$,the segmentation of $\boldsymbol{I^{(n)}}$is given by the level set contour υ^{n} which minimizes the energy function. The shape interaction term proposed in this paper has the from

$$E^{S}(I^{f}) = \int_{\Gamma^{int}} \phi_0(x)dx \tag{9}$$

Thus by applying shape energy to the level set we can effectively segment sign video and we could differentiate between object contour modifications due to motion and shape changes.

D. Integrated Energy Functional for Video Segmentation

By integrating the energy functions from color, texture, boundary and shape modules we formulate the following energy functional of the active contour as

$$E_{Int}(I^f) = \alpha E_{CT}(I^f) + \beta E^O(I^f) + \gamma E^S(I^f) \tag{10}$$

Where ζ, η, χ are weighting parameters that provide stability to contribution from different energy terms. The minimization of the energy function is done with the help of Euler-Lagrange equations and realized using level set functions. The resultant level set formulation is

$$\frac{d\phi^n(x,t)}{dt} = (\alpha R_{CT}(\phi^n) + \beta R^b(\phi^n) + \gamma R^S(\phi^n))\left\|\nabla\phi^n\right\| \tag{11}$$

where

$$R_{CT}(\phi^n) = -D(p_{in}, \overline{p_{in}}(x)) + D(p_{out}, \overline{p_{out}}(x)) \tag{12}$$

$$R^b(\phi^n) = g(B^O(x)) + \nabla.g(B^O(x))\left[\frac{\nabla\phi}{|\nabla\phi|}\right] \tag{13}$$

$$R^S(\phi^n) = \phi_0(x) \tag{14}$$

3. SEGMENTATION ERROR

To validate the proposed method we use a measure to compute the correctness of spatial location of segmented objects. Suppose $A^{(n)}_{(obj)}$ is the area of segmented object in the n[th] frame from the proposed method and $G^{(n)}_{(obj)}$ is the ground truth area of the object attained by hand segmenting the same object in the same frame. $A^{(n)}_{(back)}$ area of the background in the current frame and $G^{(n)}_{(back)}$ is the ground truth background area. $A^{(n)}_{(img)}$ is the total area of the image under test. The object segmentation error is calculated from the equation

$$\varepsilon = \left|\frac{(A^{(n)}_{(obj)} - G^{(n)}_{(obj)}) + (G^{(n)}_{(back)} - A^{(n)}_{(back)})}{A^{(n)}_{(img)}}\right| \tag{15}$$

The error $\varepsilon \in [0, 1]$ gives area intersections of segmented object to their background by total area of the image frame. In a sense this error tells us the percentage of misclassified pixels in each frame of the video sequence.

4. EXPERIMENTAL RESULTS

Figure 2(a) and 2(b) shows the effectiveness of the proposed algorithm against the CV model. Segmentation Error is calculated for the sequence of frames and is plotted in figure 3 for both proposed method and CV method [12].

We also experimented with noise added to the video sequence to test our proposed method. For this purpose we added to our video sequence in the previous experiment a white Gaussian noise of zero mean and standard deviation $\sigma = 2$. We also experimented with noise added to the video sequence to test our proposed method. The results are shown in figure 4. For this purpose we added to our video sequence in the previous experiment a white Gaussian noise of zero mean and standard deviation $\sigma = 2$. As we observe the level set in CV model deviated drastically from the set perimeter for segmentation. This is due to the additional prior shape from the previous frame.

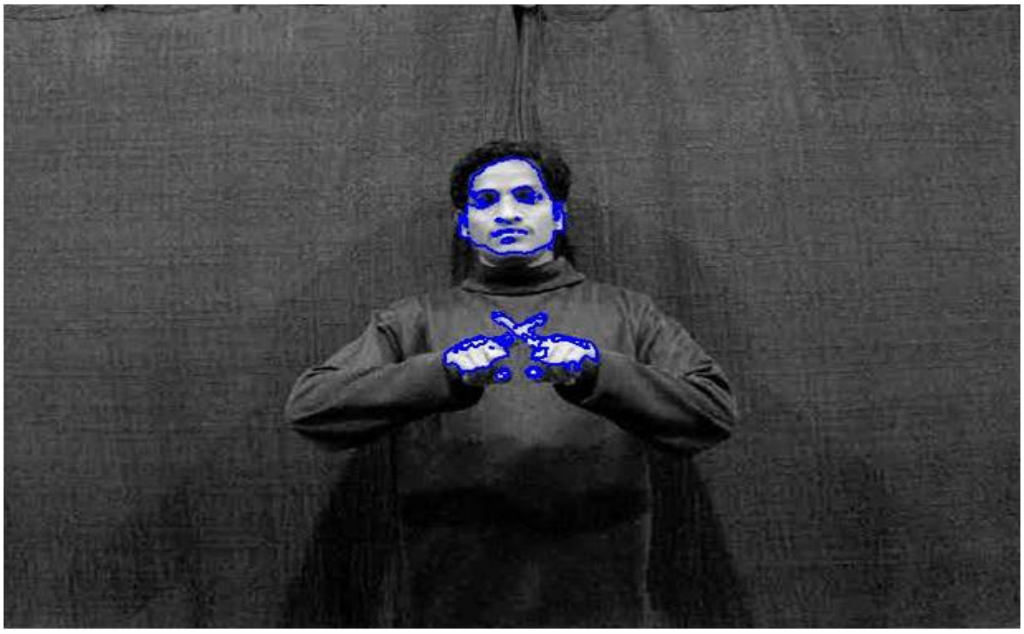

Figure 2(a). Showing the enlarged result with CV method in [12]

Figure 2(b). Showing the enlarged result with Proposed Method that is with Color, Texture, Boundary and prior shape information

Here for this video sequence we have increase the shape weighing term to $\chi = 0.43$ to influence the contour to shape information.

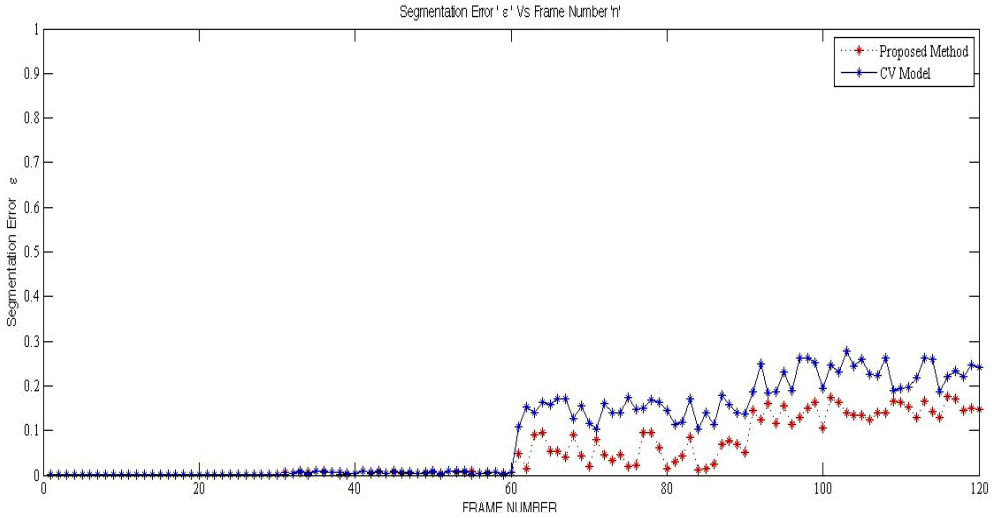

Figure 3. Segmentation Error for the sequence in figure 2(a)

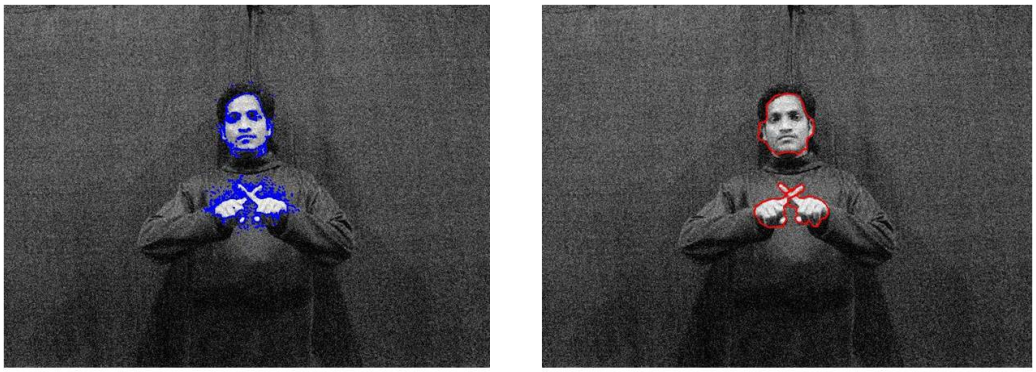

Figure 4. Example showing robustness of segmentation to noise. (a) is CV model and (b) Proposed method both with white Gaussian noise of σ =2 and μ =0

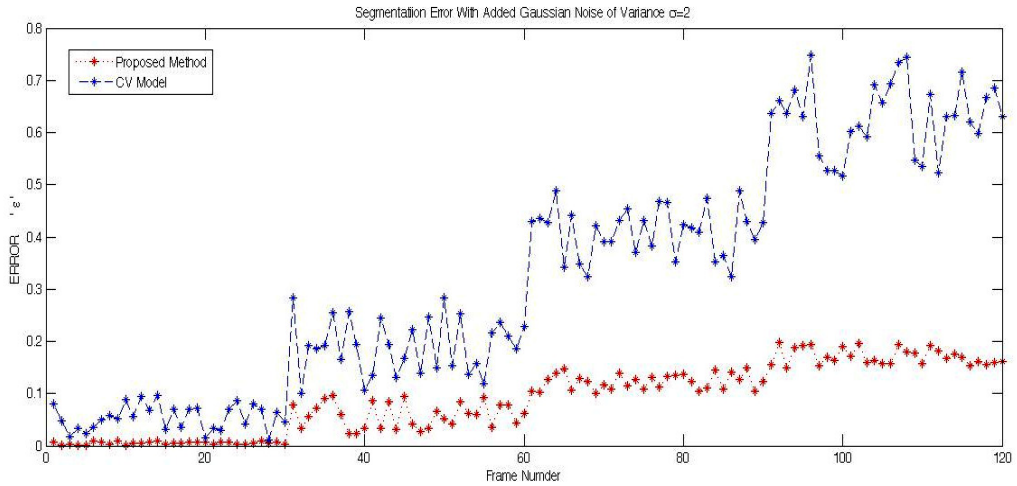

Figure 5. Segmentation Error Plot with noise added to the image sequence in figure 2(a)

The segmentation error is plotted in fig. 5 which indicated there is a larger deviation from ground truth segmentation in case of CV model when compared to level sets with prior shape knowledge.

We also experimented with more real time scenarios so that sign language recognition system can be implemented under real time. The video sequence that is considered is taken in a restaurant where it is difficult to identify the signer and the hands of the signer with multitude of background clutter. For this image sequence we have manually extracted the signer's hands and head portions from the first frame $I^{(0)}$ which is used to initialize the proposed level set. The weighing parameters in eq.26 $\zeta = 0.24, \eta = 0.21$ and $\chi = 0.63$.We observed that segmentation is good if shape term weight is increased. Because in this real time video the color and texture information does not reveal much of information. Similarly the boundary information also provides insufficient data under the influence of such a background clutter.

We observe occlusions of hands and head very frequently in sign language videos. Most sign language recognition systems insist that the signer should face the camera directly to avoid occlusions of hands largely. This problem is solved using our level set method. We initialized contour for only right hand of the signer in the first frame. With the left hand coming in the path of right hand as can be observed from the original sequence in figure 6(a), it's difficult to segment the shape of right hand. The only problem with our model, as we can observe the segmentation shape is not exact to that of right hand. This is due to the imperfect segmentation in the previous frame which is taken as a mask to current frame. This problem can be fixed by reinitializing the active contour whenever occlusion period is longer.

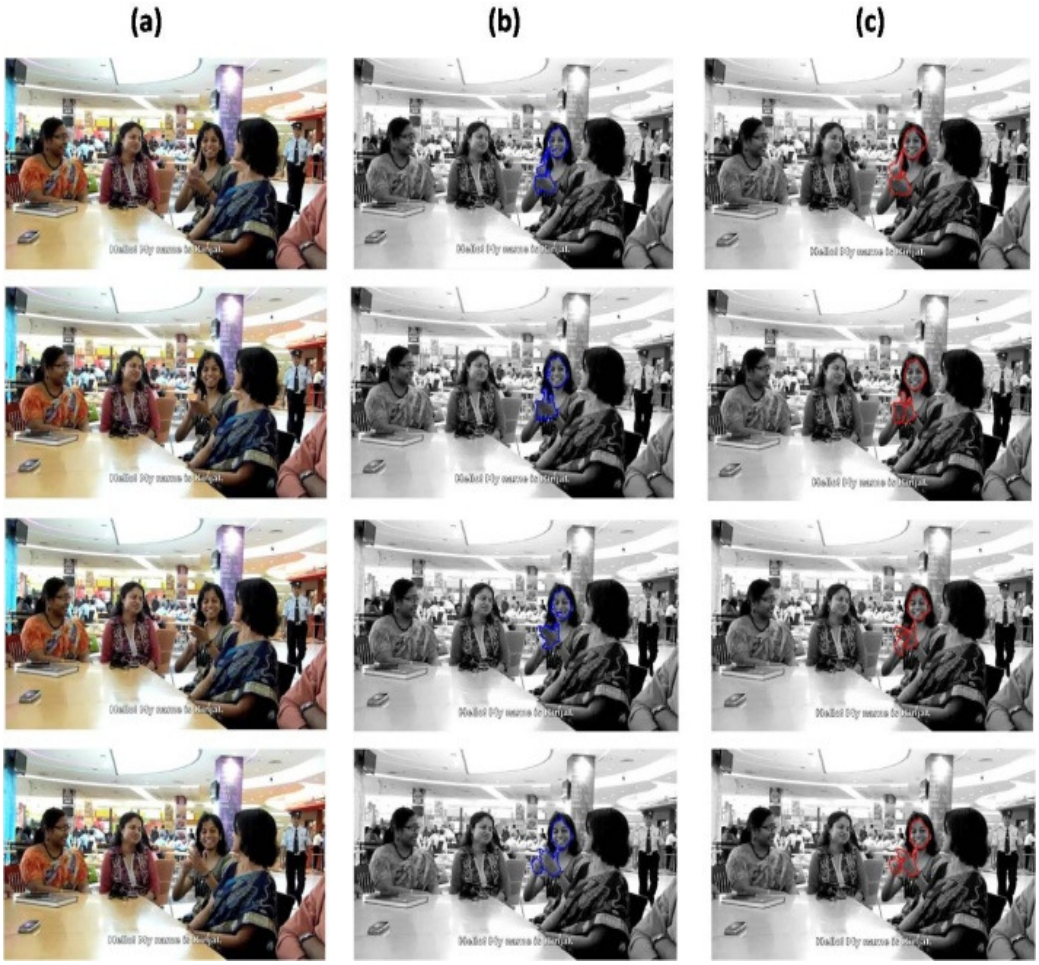

Figure 6. Comparison of Segmentation results for real time video sequence where frames 1110, 1123, 1190 and 1221 are shown. Column (a) original video sequence, column (b) results from CV model and Column (c) results from proposed method.

The final experiment shows the supremacy of our proposed technique when the video sequence contains fast moving objects in contrast to hand and head movements. The video sequence is shot on an Indian road and in the natural environment. Figure 7 shows the original sequence in column (a) along with the results of CV model in column (b) and our method in column (c).

Figure 7. Frames of a sign video sequence on a Indian road and under natural environment. Column (a) is original sequence of frames 39, 54,79 and 99. Column (b) CV method and Column (c) our proposed method.

The plot between segmentation errors calculated and frame number, i.e. error per frame shows the error increases as the bike enters the frame for CV method [12]. As pointed earlier error in our method also increasing due to re-initialization problem of the level set to the current frame. Error can be minimized by employing a re-initialization algorithm to initialize the initial contour whenever there is large change in the object of interest otherwise if there is change in the background. For the moment our method has shown that prior knowledge of shape along with other cues such as color, texture and boundary information can provide good segmentation results for segmenting hand gestures of sign language.

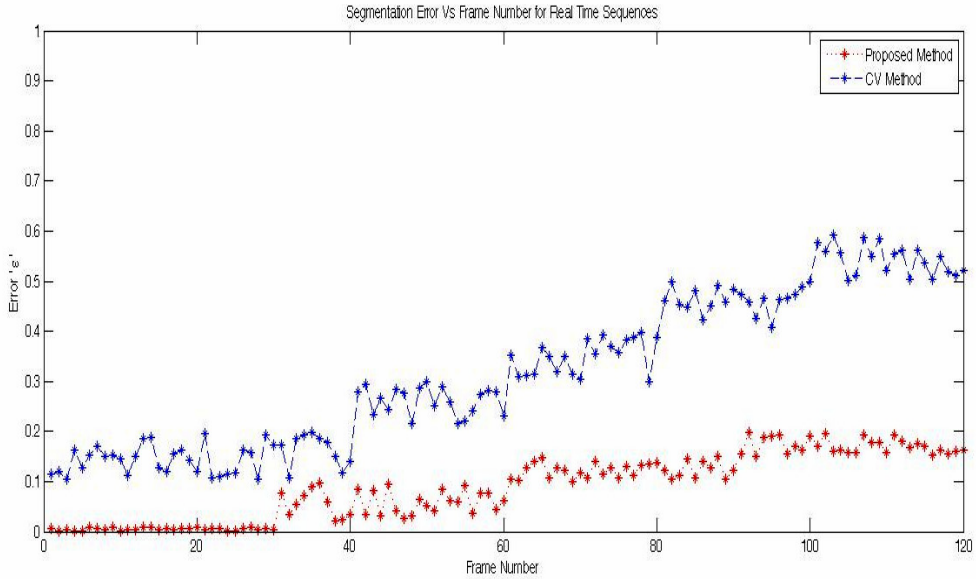

Figure 8. Segmentation Error versus frame number plot for results in 7(b) and 7(c).

5. CONCLUSION AND DISCUSSION

This paper brings us a little nearer to making sign language recognition system a reality. The method proposed combines effectively the color, texture, boundary and prior shape information to produce an effective video segmentation of sign language videos under various harsh environments such as cluttered backgrounds, poor lighting, fast moving objects and occlusions. We have effectively demonstrated by experimentation of the proposed method by applying it to video sequences under various conditions. Nevertheless there are challenges which are to addressed to apply this method to continuous sign language recognition systems to carry out the segmentation in real time.

REFERENCES

[1] W. Stokoe, D. Casterline, and C. Croneberg,(1965) "*A Dictionary of American Sign Language on Linguistic Principles*." Gallaudet College Press, Washington D.C., USA.

[2] P.V.V.Kishore, P.Rajesh Kumar, E.Kiran Kumar &S.R.C.KIshore (2011). *"Video Audio Interface for Recognizing Gestures of Indian Sign Language"*International Journal of Image Processing(IJIP), CSC Journals, Vol. 5, Issue(4), pp479-503.

[3] Qing-song Zhu, Yao-qinXie, Lei Wang (2010) *Video Object Segmentation by Fusion of Spatio-Temporal Information Based on Gaussian Mixture Model*, Bulletin of advanced technology research, vol. 5, No. 10, pp38-43.

[4] LuminitaVese and Tony Chan.(2002)"*A multiphase level set framework for image segmentation using the mumford and shah model"*. International Journal of Computer Vision, Vol.50,No. 3 pp271-293.

[5] D. Mumford and J. Shah.(1989)*"Optimal approximation by piecewise smooth functions and associated variational problems"*Comm. Pure Appl. Math, Vol.42,pp577-685.

[6] T. Chan and L. Vese(2001)"*Active contours without edges*". IEEE Transations on Image Processing,Vol. 10 No.2,pp266-277.

[7] M.Kass, A Witkin, D Terzopoulos(1987),*"Snakes: Active Contour Models"*, Int. J. of Computer Vision, pp 321-331.

[8] Tuceryan, M., Jain, A. (1998): Texture analysis. In: Chen, C.H., Pau,L.F., Wang, P.S.P. (eds) The Handbook of Pattern Recognitionand Computer Vision, 2nd edn. chap. 2.1, 207–248.World ScientificPublishing, Singapore.

[9] Allili, M.S., Ziou, D.,(2006), *"Automatic color-texture image segmentation by using active contours"*. In: Proceedings of1st IEEE International Workshop on Intelligent Computing inPattern Analysis/Synthesis, Xi'an, China, 26–27,LNCS 4153, pp. 495–504.

[10] S. Allam, M. Adel, P. Refregier, (1997)*"Fast algorithm for texture discrimination by use of a separable orthonormal decomposition of the co-occurrence matrix,"*Applied Optics, vol.36, pp.8313–8321.

[11] R. M. Haralick, K. Shangmugam, I. Dinstein, (1973)*"Textural Feature for Image Classification,"*IEEE Trans on Systems, Man, Cybernetics, 3(6), pp.610—621.

[12] Chan, T., & Zhu, W. (2005). *Level set based prior segmentation.* Proceeding CVPR, vol.(2), pp1164–1170.

[13] www.deafsigns.org

Authors

P.V.V.Kishore received his M.Tech degree in electronics from Cochin University of science and technology, and currently pursuing PhD at Andhra University College of engineering , Visakhapatnam in the department of electronics and communication engineering. His research interests are digital signal and image processing, computational intelligence, human computer interaction, human object interactions.

E. Kiran Kumar received his B.Tech degree in Electronics and Communication Engineering from Vignan's institute of information technology, Visakhapatnam, and currently pursuing M.Tech at Dadi Institute of Engineering & Technology,Anakapalli in Department of ECE. His research interests are image processing, computational intelligence, human computer interaction, human object interactions.

B. Manjula received her M.Tech degree in Digital Systems and Computer Electronics from JNTU Ananthapur. She is working as an associate professor at Dadi Institute of Engineering & Technology,Anakapalli in Department of ECE. Her research interests are image processing, computational intelligence, human computer interaction, human object interactions.

Dr. P.Rajesh Kumar received his Ph.D degree from Andhra University College of Engineering for his thesis on Radar Signal Processing. He is currentl y working as associate professor at Dept. of ECE, A.U College of engineering. His research interests are digital signal and image processing, computational intelligence, human computer interaction, human object interactions.