Journal of Information Systems and Telecommunication
Vol.10, No.3, July-September 2022, 211-221

http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Rough Sets Theory with Deep Learning for Tracking in Natural Interaction with Deaf

Mohammad Ebrahimi[1], Hossein Ebrahimpour-Komeleh[2*]

[1]. Electrical and Computer Engineering Kashan University
[2]. Faculty of Electrical and Computer Engineering Kashan University

Received: 03 May 2021/ Revised: 01 Feb 2022/ Accepted: 07 Feb 2022

Abstract

Sign languages commonly serve as an alternative or complementary mode of human communication Tracking is one of the most fundamental problems in computer vision, and use in a long list of applications such as sign languages recognition. Despite great advances in recent years, tracking remains challenging due to many factors including occlusion, scale variation, etc. The mistake detecting of head or left hand instead of right hand in overlapping are, modes like this, and due to the uncertainty of the hand area over the deaf news video frames; we proposed two methods: first, tracking using particle filter and second tracking using the idea of the rough set theory in granular information with deep neural network. We proposed the method for Combination the Rough Set with Deep Neural Network and used for in Hand/Head Tracking in Video Signal DeafNews. We develop a tracking system for Deaf News. We used rough set theory to increase the accuracy of skin segmentation in video signal. Using deep neural network, we extracted inherent relationships available in the frame pixels and generalized the achieved features to tracking. The system proposed is tested on the 33 of Deaf News with 100 different words and 1927 video files for words then recall, MOTA and MOTP values are obtained.

Keywords: Natural Interaction with Deaf; Machine Vision; Persian Deaf News Hand Tracking; Sign Language; Rough Sets Theory; Deep Learning.

1- Introduction

Recognition of states and hand gestures are very important in a natural interaction with a computer. Its importance is due to its widespread applications in virtual reality, sign language recognition and computer games. Fast and robust hand gesture recognition remains an open problem [1].

By tracking the hand in the video, it is simpler to partition it from the image frames. The purpose of tracking methods is to discover and track one or more objects in the sequence of images. Tracking can be thought of as a kind of object discovery in a set of similar images. Many tracking methods are used to discover and track objects in video films, in which a large number of images have to be processed. Various kinds of probabilistic inference models haves been applied to multi-object tracking, such as Kalman filter, Extended Kalman filter and Particle filter. In the case of linear system and Gaussian-distribution object states, Kalman filter is proved to be the optimal estimator. It has been applied. Extended Kalman filter, for the nonlinear case, extended Kalman filter is a solution. It approximates the nonlinear system by Taylor Expansion. Particle filter, Monte Carlo sampling-based models becomes popular in tracking, especially after the introduce of Particle filter Typically, the strategy of Maximum A Posteriori (MAP) is adopted to derive a state with the maximum probability [2,3].

About 466 million deaf people live in the world, this is approximately 5.3% of the world population1, their natural language is the sign language. They are restricted in reading and writing the official language. Education, work, use of computers and the Internet are affected for them. Diagnosing the sign language, if used in interaction with the computer and in the translation of texts to hand gestures, can support them well [2].

Deep learning is a kind of hierarchical learning. In layered hierarchical learning, nonlinear features are extracted, then

🖂 **Hossein Ebrahimpour-Komeleh**
ebrahimpour@kashanu.ac.ir

the output layer is usually formulated depending on how many groups that are needed [4]. The output layer is a classifier. It combines all features to make predictions. The layers' hierarchy is deeper, the more nonlinear features are extracted. That is why the number of layers in deep learning is used. Sometimes these complex features cannot be obtained directly from the input image.

A Convolutional neural network, CNN is a popular deep learning architecture that automatically learns useful feature representations directly from image data. CNNs, or ConvNets, are essential tools for deep learning, and are especially useful for image classification, object detection, and recognition tasks. CNNs are implemented as a series of interconnected layers.

A semantic segmentation network classifies every pixel in an image, resulting in an image that is segmented by class. Semantic segmentation networks like DeepLab make extensive use of dilated convolutions, also known as Atreus convolutions, because they can increase the receptive field of the layer without increasing the number of parameters or computations.

Although years have passed since the design of target tracking, this topic is still an active research field with many applications in the world's universities and scientific circles. This issue is of particular importance in tracking the targets that move with quick maneuvers, because the dynamic of target motions is complex and its nature is nonlinear. Given that the targets we are interested in track down have high-level maneuvers, various intelligent methods have all been in line with tracking the best.

The "rough sets" approach to estimate sets has led to beneficial aspects of the grain calculations, and is part of computational intelligence. The basic idea of the rough sets for aggregated information implies that how much the subsets can be used to find the objects of interest for estimating [5]. Also rough sets theory is convenient for picking up irrelevant and redundant features from a dataset [6]. Here the computational intelligence of rough sets is used. The causes of the lack of information in a particular application are identified in order to overcome the problem of the lack of information in a particular application. Then, necessary relationships are used to compensate for the lack of information. In fact, subsets of classes are characterized by rough sets, then the boundary and negative members obtained from the definition of the following sets are guided to their proper position with the definition of functions.

Tracking is very important. Machine learning is used for tracking. Dongxu Li et al. used deep learning for sign language recognition [7]. Literature findings of Wadhawan et al. indicated that the major research on sign language recognition has been performed on static, isolated and single-handed signs using camera [3].

In the case of deaf communication, it is necessary to recognize the signs expressed by the deaf. Facial gesture, trajectory and hand gesture are the three basic features for recognizing the language sign expressed by a deaf person. Hand and head tracking is used to find the trajectory and segment them from the background of the video in the frames. So, the problem is accurately tracking the hands and head in videos of signs expressed by the deaf.

The sign language of countries is different. In this work, a Persian dataset of sign language videos has been collected, which is available at Kashan University. The system proposed are tested on 33 videos of Deaf News with 100 different words and 1927 video files for words, and recall, MOTA and MOTP values are obtained. We used rough set theory with deep Neural network for sign language tracking. The novelty of this paper is the use of rough set theory with deep neural network for tracking. This is the first work on this topic. In this paper, at first, tracking using particle filter is explained. At second, tracking using rough sets and deep learning is explained. In the first proposed method, we used a particle filter, which has high accuracy but is very time consuming. The second proposed method responds much faster but is less accurate. To increase the accuracy of the second proposed method, we used the rough set theory.

2- Proposed Algorithm

Sign language recognition is one of the issues that have been used in many applications. Some of them are the transcription, video rebuilding, and deaf of sign language.

In this regard, we have tried to create a system for sign language recognition for Persian, so that ordinary people and the deaf can easily interact with each other. The sign language recognition uses a variety of sub-systems, each of which has its own characteristics and procedures, and the relationship between the various components of the system is an important issue that cannot easily be ignored. The purpose of this research is to design and train the "deep learning network" to sign language recognition for Persian.

The first part that the system focuses on is the multi-tracking. The development of a new multi-tracking method used the theory of rough sets in such a way that it automatically tracks objects in a video signal. The objects in this system are two hands and face. The geometric feature of the object's presence at different times, in other words, the trajectory, can be effective in selecting the area appropriately, improving fragmentation, and identifying the results in this application [1,2,8].

The rough sets approach in estimating collections has led to beneficial benefits from granular calculations and is part of computational intelligence. The basic idea of rough sets for granular information implies that how much of the subcategories can be used to estimate in the discovery and fragmentation of favorite objects. In this system, the computational intelligence of the rough sets will be used. To overcome the problem of the lack of information in any particular application, it explains the causes of the lack of information. Then, relationships are used to compensate for the lack of information. In fact, with rough sets, the subsets of the categories are determined, and then the boundary and negative members obtained from the definition of rough sets, with the definition of the function, are directed to their proper place [9].

$$g(i,j) = \begin{cases} \frac{I(i,j)}{128} \; I(i,j) < 128 \\ \frac{255-I(i,j)}{128} \; I(i,j) \geq 128 \end{cases} \qquad (1)$$

$$g_p(i,j) = \frac{255}{e^{-1}-1} \times e^{g(i,j)^{0.75}-1}$$

There, the folding function, equation (1) is used. In the equation (1), i and j are location coordinates of pixels in image I. when two hands overlapped or the hands and face overlapped, Weak boundaries are created. At this time the tracker fails, means tracker going from right hand to left hand or to face. The g function shows up a very weak boundary overlapping regions. The g function converts the intermediate values of the gray area of the boundary to completely white values. In this case, the tracker does not cross the boundary and continues to track in its area truly.

2-1- Proposed Method 1: Multi-Tracking using Particle Filter

In simple terms, the filtering method refers to the process of obtaining and accessed targets during the movie screenings. This issue, filter for target, is very important in tracking because the targets move with quick maneuvers by means that dynamic of target motions is complex and its nature is nonlinear. lately, particle filtering has appeared as a tracking approach as compared with meanshif. It is a stochastic approach that models nonlinear motion with non-Gaussian noise.

General approaches in the tracking with filters have two stages: prediction and update. In prediction stage the model must predict the location of the hand in the next frame using motion model, after arriving to next time, the exact location is achieved and update the motion model using observation model. In the particle filter method, this is done pixel-to-pixel, and it raises the computational complexity.

For each position in frame at each time, local score is calculated. The global score is the total score for the best path until now, which ends to each position. For each position in image, the best predecessor is searched for among a set of possible previous scores. This best predecessor is then stored in table of back pointers which is used for the trace back.

Principal Component Analysis (PCA) performed by the Karhunen-Lokve transform produces features that are mutually uncorrelated. The obtained by the KL transform solution is optimal when dimensionality reduction is the goal and one wishes to minimize the approximation mean square error.

Mean face difference images (MFDI) are difference images between the mean face and the tracked face patch computed over a sentence or word segment.

The motion energy feature is used for silence detection in the presented system. Additionally, the use of motion energy as feature for sign language recognition is investigated.

hand position normalized with respect to shoulder and vertical body axis. Gabor wavelet transform is one of the most effective texture feature extraction techniques and has resulted in many successful practical applications.

PCA, MFDI, motion energy, hand position, hand texture speed and RGB to YcBcR and GRAY are features for sign language recognition.

Vision based communicating with compare of speech-based communicating is more complex and meaningful. Direct communication between deaf and other people is very difficult, so there are attempts for making a sign language interpreting system You can see a diagram of it in Fig. 1. In the first proposed method, we used a particle filter, which has high accuracy but is very time consuming. The second proposed method responds much faster but is less accurate. To increase the accuracy of the second method, we used the rough set theory.

2-2- Proposed Method 2: Multi-Tracking using Rough Sets

2-2-1- Rough Set

By using fuzzy and in particular the theory of rough sets with uncertainties in the trace problem, the best trace is attempted. All video frame points are included in the database table as examples in the first column. The properties of each point are stored as a separate column in the table. The value of the attributes for each point is recorded. Due to the fact that a camera mounted in a single place arranges the data, each frame is calculated for each frame as the changes in the positive, negative and boundary sets are added. Each time, the matrix of the hand region in the matrix is multiplied by the general relationship and the matrices of the intermediate and the primary are obtained. By using the definition and use of proper conversion functions, the tracing method improves. If it works online, it is necessary to process the same as the film. This is called active learning [10,11].

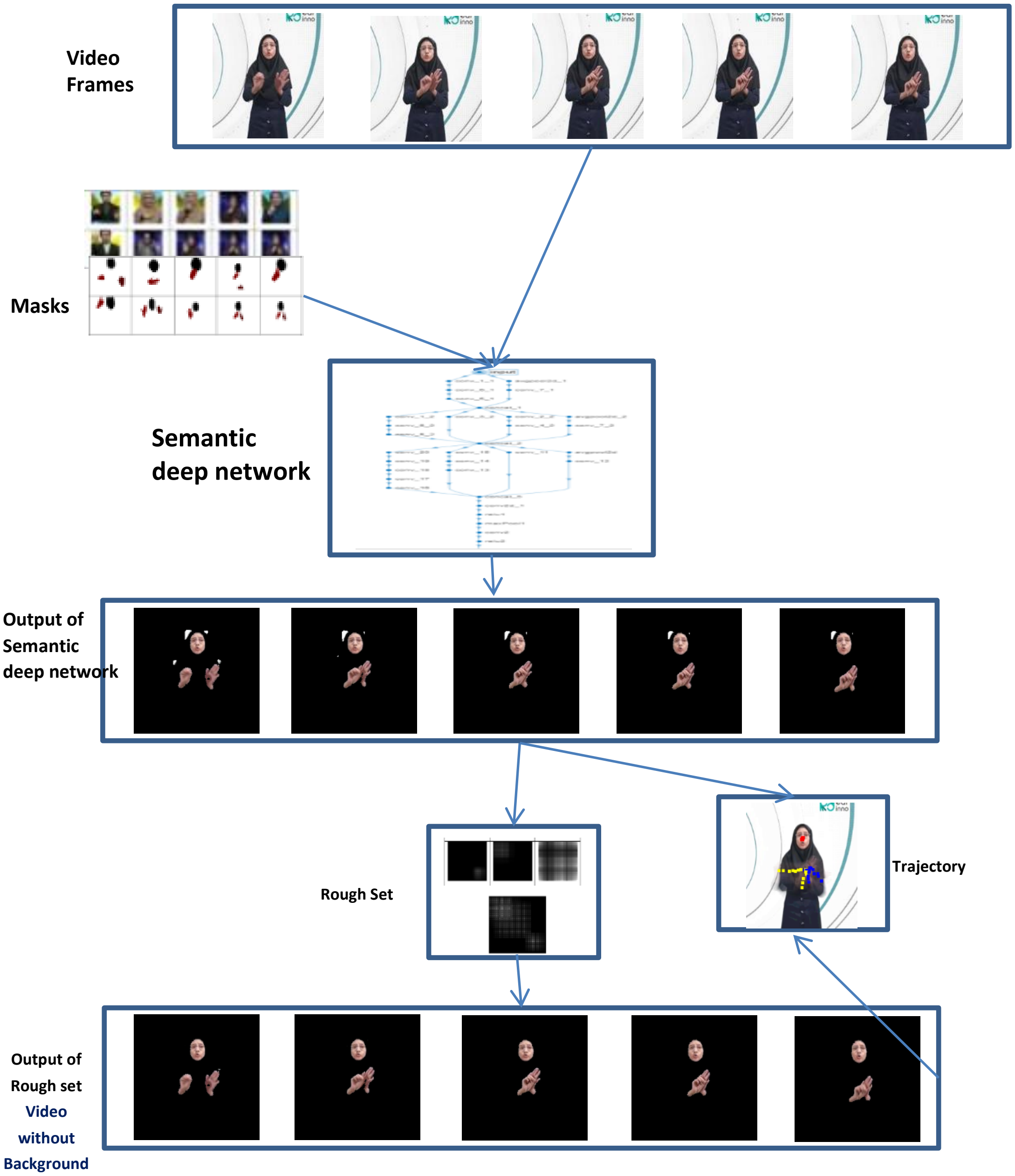

Video Frames
Masks
Semantic deep network
Output of Semantic deep network
Rough Set
Trajectory
Output of Rough set Video without Background

Fig. 1: The outline of Proposed method 2: multi-tracking using Rough sets

Fig. 1 shows the outline of proposed method 2, multi-tracking using Rough sets.

The most important features of rough set theory are:

- Finding relationships that are not discovered by statistical methods.
- Ability to use quantitative and qualitative information.
- Finding a minimum set of data that is useful for categorization (such as minimizing dimensions and number of data).
- Assessing the importance of data.
- Generate decision rules on data [3].

Using the definition and use of convenient conversion functions, the method of tracking is improved. The composite decision is listed in table 1.

Table 1: The composite decision table

U	a1	a2	a3	a4	a5	a6	a7	D
x_1	0	0	0	1	{0,1,..,255}	{0,1,..,255}	{0,1,..,255}	No
x_2	0	0	0	1	{0,1,..,255}	{0,1,..,255}	{0,1,..,255}	No
…	…	…	…	…	…	…	…	…
x_k	0	0	0	1	{0,1,..,255}	{0,1,..,255}	{0,1,..,255}	No

The k is all pixel of every frame of video. The size of frame is 208×186 and k=208×186 = 38688, a_i is 0 or 1, if $x_j \in B_i$ then a_i is 1, otherwise ai is 0. $0 < j \leq 38688$ and $i = 1..4$.

$B_1 = \{a_1, a_5, a_6, a_7\}$, $B_2 = \{a_2, a_5, a_6, a_7\}$, $B_3 = \{a_3, a_5, a_6, a_7\}$, $B_4 = \{a_4, a_5, a_6, a_7\}$ and $B = \cup_{k=1,2,3} B_k$.
If $a_1 = 1$ *or* $a_2 = 1$ *or* $a_3 = 1$, D is *Yes* and othewise No.
R_B is an equivalence relation:

- $R_{B_1}(x_i) =$ *set of pixels* x_i *region of right hand*
- $R_{B_2}(x_i) =$ *set of pixels* x_i *region of left hand*
- $R_{B_3}(x_i) =$ *set of pixels* x_i *region of face hand*
- $R_{B_4}(x_i) =$ *set of pixels* x_i *region of background*
- $CR_B(x_i) =$ *set of pixels* x_i *region of* $B_1 \cup B_2 \cup B_3$

The size of matrixes M_* are $k \times k$ and k=38688. For relations RB define M_*.

$$M^{R_{B_i}}_{k\times k}(i,j) = M^{R_{B_i}}_{k\times k}(j,i) = a_i(x_j) \tag{2}$$

$M^{R_{B_1}}_{k\times k}$, $M^{R_{B_2}}_{k\times k}$, $M^{R_{B_3}}_{k\times k}$, $M^{R_{B_4}}_{k\times k}$ and $M^{CR_B}_{k\times k}$ obtain equation (2). Fig. 2 and Fig. 3 show matrixes $M^{R_{B_1}}_{k\times k}$, $M^{R_{B_2}}_{k\times k}$, $M^{R_{B_3}}_{k\times k}$, $M^{R_{B_4}}_{k\times k}$ and $M^{CR_B}_{k\times k}$.

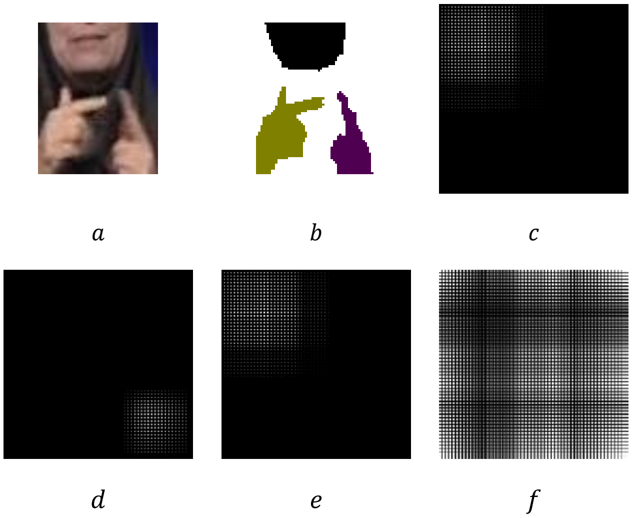


Fig. 2: a. image, b. label of image (a), c. $M^{R_{B_3}}_{k\times k}$, d. $M^{R_{B_2}}_{k\times k}$, e. $M^{R_{B_1}}_{k\times k}$ and f. $M^{R_{B_4}}_{k\times k}$

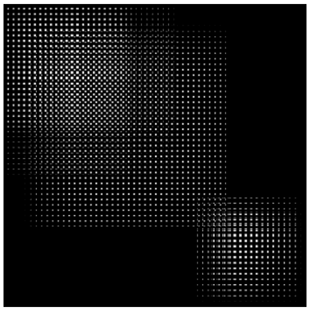


Fig. 3: $M^{CR_B}_{k\times k}$ for Fig. (2. b)

$\Lambda^{CR_B}_{k\times k}$ be an induced diagonal matrix of $M^{CR_B}_{k\times k}$, then:

$$\Lambda^{CR_B}_{k\times k} = diag\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \frac{1}{\lambda_3}, \dots \frac{1}{\lambda_k}\right) \tag{3}$$

That $\lambda_i = \sum_{j=1}^{k} m_{ij}$, $1 \leq i \leq k$.

The calculation of the intermediate matrix:

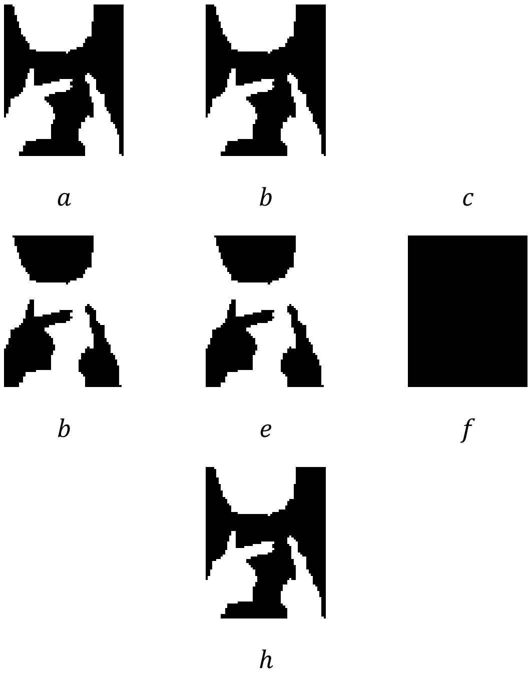
$$\Omega_{k\times 2}^{CR_B} = M_{k\times k}^{CR_B} \bullet [(D_{1\times k})^T \quad (\overline{D_{1\times k}})^T] \qquad (4)$$

The calculation of the basic matrix:

$$HD_{k\times 2} = \Lambda_{k\times k}^{CR_B} \bullet \Omega_{k\times 2}^{CR_B} \qquad (5)$$

The n-column Boolean vector $G\left(\underline{CR_B}(D_j)\right)$ of the lower approximation $\underline{CR_B}(D_j)$:

$$G\left(\underline{CR_B}(D_j)\right) = H^{[1,1]}(D_j) \qquad (6)$$



a *b* *c*

b *e* *f*

h

Fig. 4: a. $G\left(\underline{CR_B}(D_{skin})\right)$, b. $G\left(\overline{CR_B}(D_{skin})\right)$, c. $G\left(POS_{CR_B}(D)\right)$, d. $G\left(\underline{CR_B}(D_{background})\right)$,e. $G\left(\overline{CR_B}(D_{background})\right)$,f. $G\left(BND_{CR_B}(D)\right)$ and h. $G\left(NEG_{CR_B}(D)\right)$

The n-column Boolean vector $G\left(\overline{CR_B}(D_j)\right)$ of the upper approximation $\overline{CR_B}(D_j)$:

$$G\left(\overline{CR_B}(D_j)\right) = H^{(0,1]}(D_j) \qquad (7)$$

The n-column Boolean vector $G\left(POS_{CR_B}(D)\right)$ of the positive region:

$$G\left(POS_{CR_B}(D)\right) = \sum_{j=1}^{r} H^{[1,1]}(D_j) \qquad (8)$$

The n-column Boolean vector $G\left(NEG_{CR_B}(D)\right)$ of the negative region:

$J = (1,1,\ldots,1)^T$,

$$G\left(NEG_{CR_B}(D)\right) = J - \sum_{j=1}^{k} H^{(0,1]}(D_j) = \left(\sum_{j=1}^{r} H^{(0,1]}(D_j)\right)^{[0,0]} \qquad (9)$$

The n-column Boolean vector $G\left(BND_{CR_B}(D)\right)$ of the boundary region:

$$G\left(BND_{CR_B}(D)\right) = \sum_{j=1}^{r} H^{(0,1]}(D_j) - \sum_{j=1}^{k} H^{[1,1]}(D_j) = \quad \sum_{j=1}^{k} H^{(0,1]}(D_j) \qquad (10)$$

The r is number of x_j in the set of D that is Yes for D_1, or No for D_2 . Fig. 4 show matrixes $G\left(\underline{CR_B}(D_{skin})\right)$, $G\left(\overline{CR_B}(D_{skin})\right)$, $G\left(POS_{CR_B}(D)\right)$, $G\left(\underline{CR_B}(D_{background})\right)$, $G\left(\overline{CR_B}(D_{background})\right)$, $G\left(BND_{CR_B}(D)\right)$ and $G\left(NEG_{CR_B}(D)\right)$.

2-2-2- Semantic Segmentation Network using Deep Learning

A semantic segmentation network classifies every pixel in an image, resulting in an image that is segmented by class. To illustrate the training procedure, in this paper we train deep learning net, one type of convolutional neural network (CNN) designed for semantic image segmentation. Other types of networks for semantic segmentation include fully convolutional networks (FCN), SegNet, and U-Net.

We used a semantic segmentation network with 56 layers (Fig.5). We use the dataset [2] training. This dataset is a collection of images containing Right and left Hand Face. But this approach, used with dynamic programming, is very time-consuming.

A 2-D convolutional layer applies sliding convolutional filters to the input. The layer convolves the input by moving the filters along the input vertically and horizontally and computing the dot product of the weights and the input, and then adding a bias term. An average pooling layer performs down-sampling by dividing the input into rectangular pooling regions and computing the average values of each region. A concatenation layer takes inputs and concatenates them along a specified dimension. The inputs must have the same size in all dimensions except the concatenation dimension. An image input layer inputs 2-D images to a network and applies data normalization. An average pooling layer performs down-sampling by dividing the input into rectangular pooling regions and computing the average values of each region. A relu layer performs a threshold operation to each element of the input, where any value less than zero is set to zero.

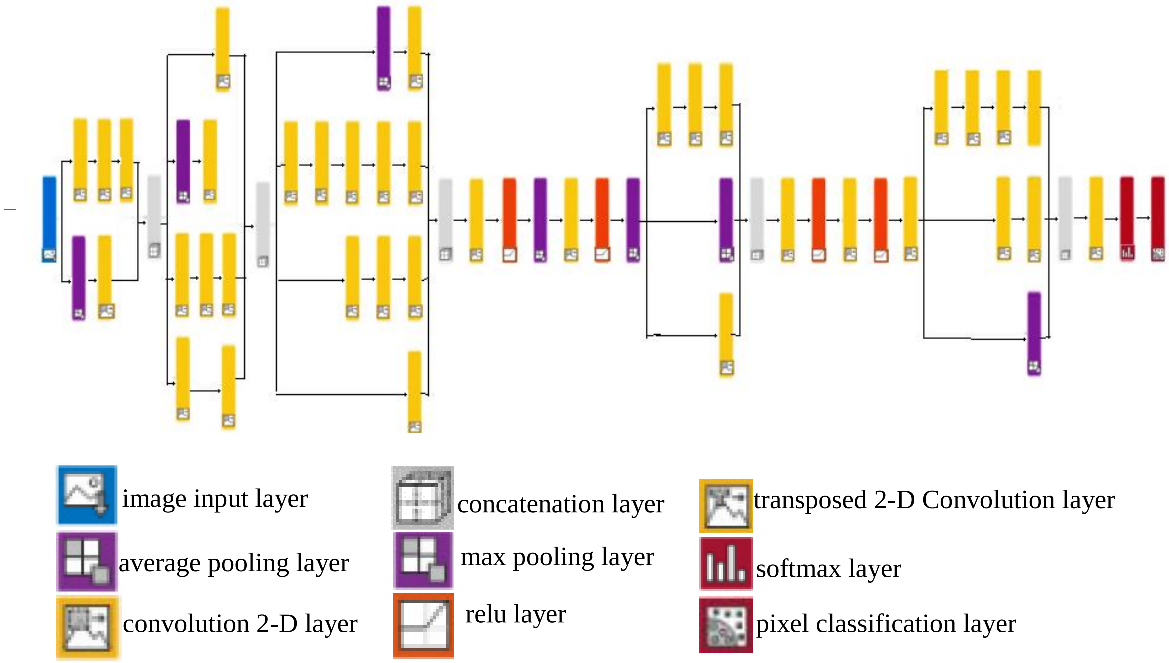


Fig 5. The 56-layer semantic segmentation network.

A max pooling layer performs down-sampling by dividing the input into rectangular pooling regions, and computing the maximum of each region. A transposed 2-D convolution layer upsamples feature maps.This layer is sometimes incorrectly known as a "deconvolution" or "deconv" layer. This layer is the transpose of convolution and does not perform deconvolution. A softmax layer applies a softmax function to the input.

A pixel classification layer provides a categorical label for each image pixel or voxel. The pixel classification layer creates a pixel classification output layer for semantic image segmentation networks. The layer outputs the categorical label for each image pixel or voxel processed by a CNN. The layer automatically ignores undefined pixel labels during training.

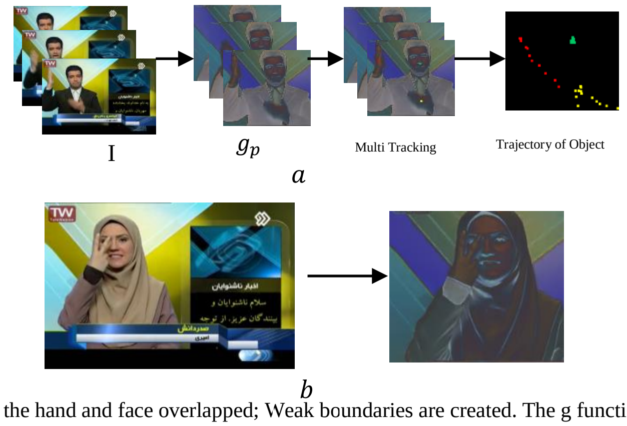


Fig. 6: a. tracking system. b. the hand and face overlapped; Weak boundaries are created. The g function converts the intermediate values of the gray area of the boundary to completely white values

3- Result

The system is tested on the deaf news. The data that will work on the system will be related to the deaf news. 33 different video files, with five different News presenters: three women and two men News presenters. At first, parts of the video associated with each word are separated from the video files

As you can see in the Fig. 10 experiments are performed on the Persian Sign Language Video Dataset. Then it will be tagged for each. Using a series of tracks and rough sets, the two-handed and headed areas will be detected and deployed to the grid. The seven general methods are based on the components and methods used for multi-tasking. In fact, each multi-tasking system consists of two main components of the observational model and the dynamic model. Dynamic model refers to the motion in sequential frames, and the observed model refers to the detection of an object in each frame of the video. Some results of deep learning with rough are shown in Fig. 6.

Recall is ratio of correctly matched detections to ground-truth detections. Multiple objects tracking accuracy (MOTA), Accuracy means the closeness of measurement values to each other, whether these values of reality are no. MOTA is obtained (TP+TN)/(P+N). MOTA combines false negative, false positives and mismatch rate. TP is True Position rate that means the right hand is there and is tracked right. TN is True Negative rate that means the right hand does not exist and has not been tracked. P is Positive that means there is a right hand. N is Negative that means there is no right hand.

Multiple Object Tracking Precision (MOTP), Precision means that in measured amounts of the same value, how close are the measured values. MOTP is obtained TP/(TP+FP). MOTP overlap between the estimated positions and the ground truth averaged over the matched. FP is False Positive, means that the right hand is there but it’s not traced. FN, False Negative, means that there is no right hand in the image but it’s tracked.

The system is tested on the Deaf News, with 100 different words, with approximately 10 to 20 samples for each word, each word is between 7 to 30 frames. For example, there are 20 samples for the "Deaf" sign, with a minimum number of frames for this word of 9 frames and a maximum number of frames of 30 frames. In total for 100 words, there are 1927 video files. And 23124 is the number of available frames.

Recall, Accuracy and Precision values are obtained in tables 2 to 3.

Table 2: right hand: Rough and Net1: Recall=0.968, Accuracy=0.979, Precision=0.977

Net1 and Rough	TP:18823	FP: 447
	FN:35	TN: 3819

In both methods “Particle Filter” and “Net1 and Rough” for ‘Face’ region: Recall=1, Accuracy=1 and Precision=1. The both methods obtain good results but the method “Net1 and Rough” answer is in a shorter time.

Table 3: left hand: Rough and Net1: Recall=0.952, Accuracy=0.977, Precision=0.938

Net1 and Rough	TP:946	FP: 62
	FN:473	TN: 21653

In Table 4 Deep learning with rough sets tracking system is compared with other methods.

3-1- Result of Proposed Method 1: Multi-Tracking using Particle Filter

The results of the particle filter show in the Fig. 7.

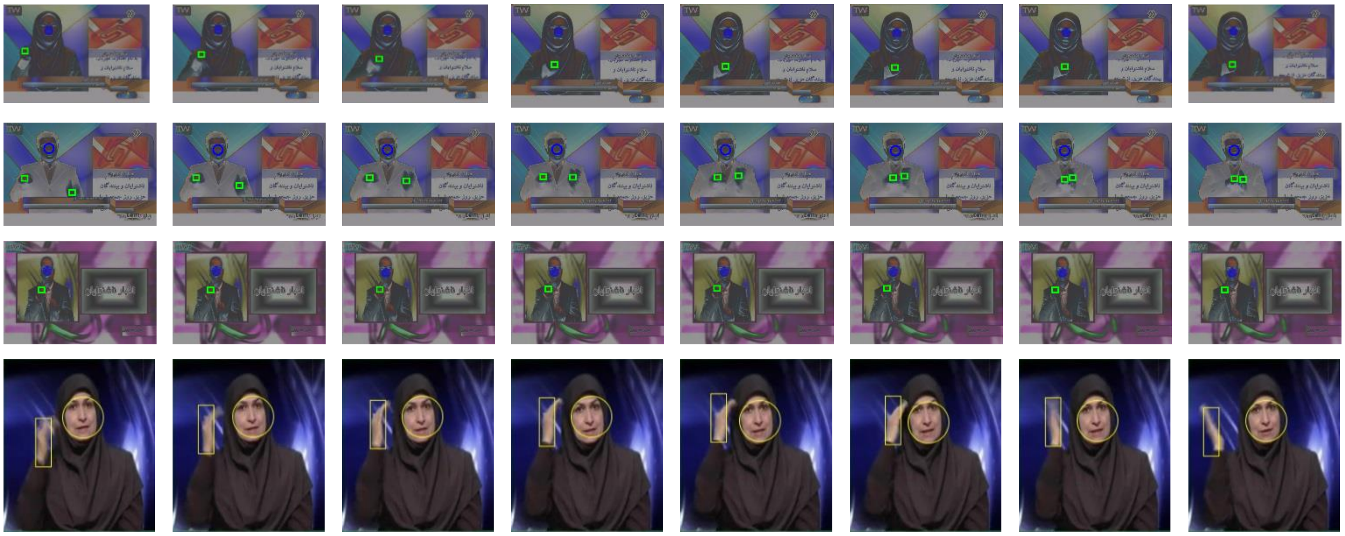


Fig. 7: Some results of Multi-tracking using particle filter on video signal DeafNews dataset

3-2- Result of Proposed Method 2: Semantic Segmentation Network using Deep Learning

Number of layers is 56, number of connections is 66, input is image and output are semantic segmentation. This 56-layer deep learning network, segment the Hands and Face areas from the background. The second-deep learning network is for tracking the right hand, face, and left hand, which in the preceding stage have their areas. Percent accuracy 97.35 on dataset. The dataset that works on the system will be related to the deaf news. The results of deep-learning tracking show in Fig. 9 and Fig. 10. and table 4. In this paper used two separate deep learning for tracking.

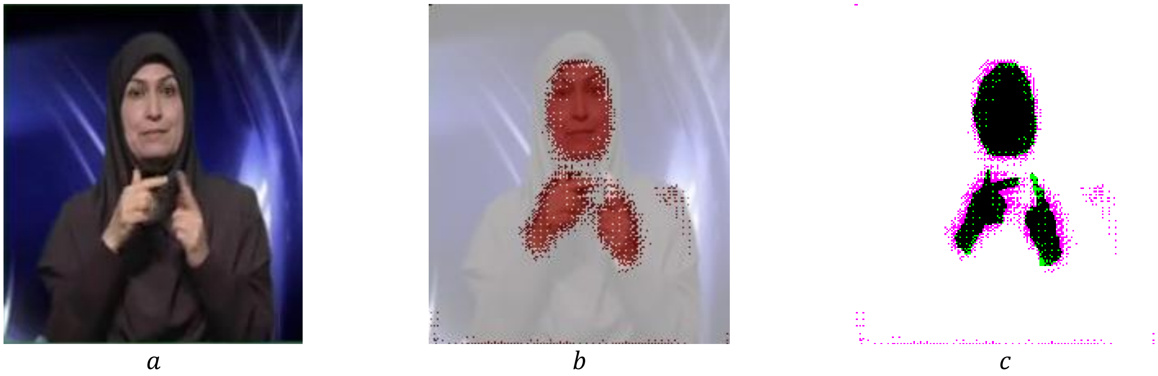


Fig. 8: result of semantic segmentation network using deep learning, a. one frame of video, b. result of semantic segmentation network using deep learning, c. black is true, green is skin false and pink is background false

In Fig. 8. shows result of semantic segmentation network using deep learning. Fig. 8.a. is one frame of video, Fig. 8.b. result of semantic segmentation network using deep learning, in Fig. 8.c. black is true, green is skin false and pink is background false.

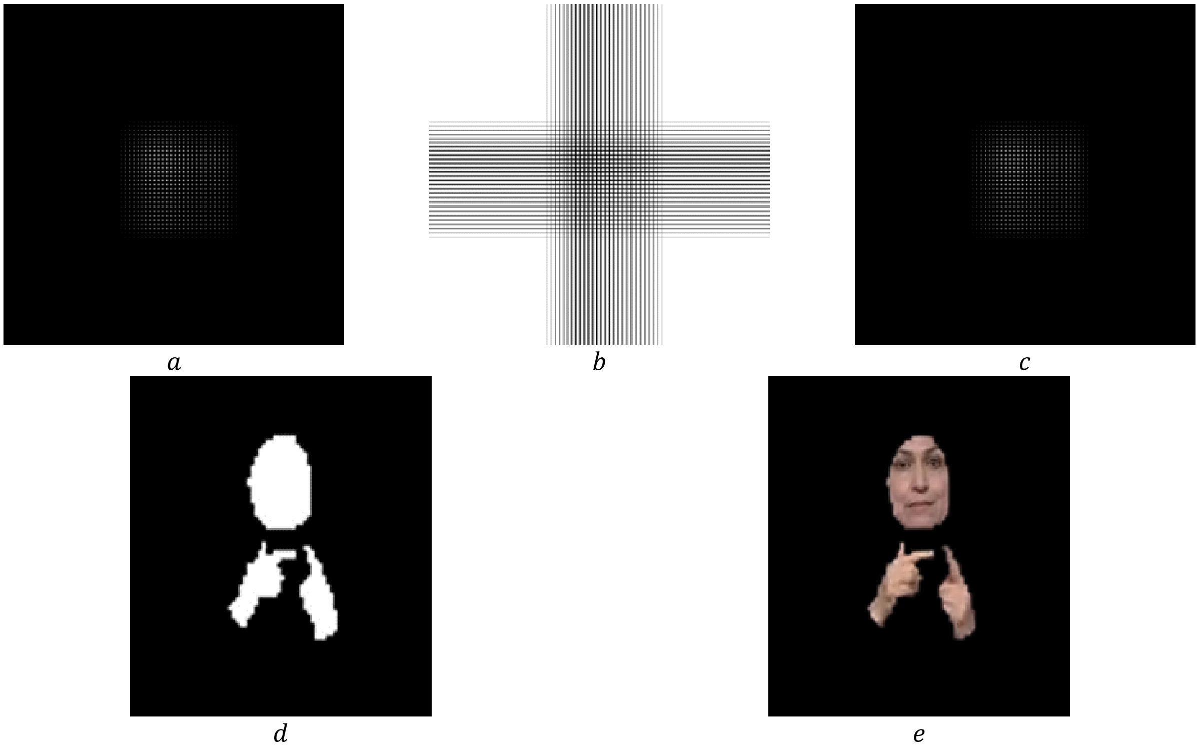


Fig. 9: a. $\boldsymbol{M}^{R_{B_3}}_{k\times k}$,b. $\boldsymbol{M}^{R_{B_2}}_{k\times k}$,c. $\boldsymbol{M}^{CR_B}_{k\times k}$,d and e. result method rough on Fig. 6.b.

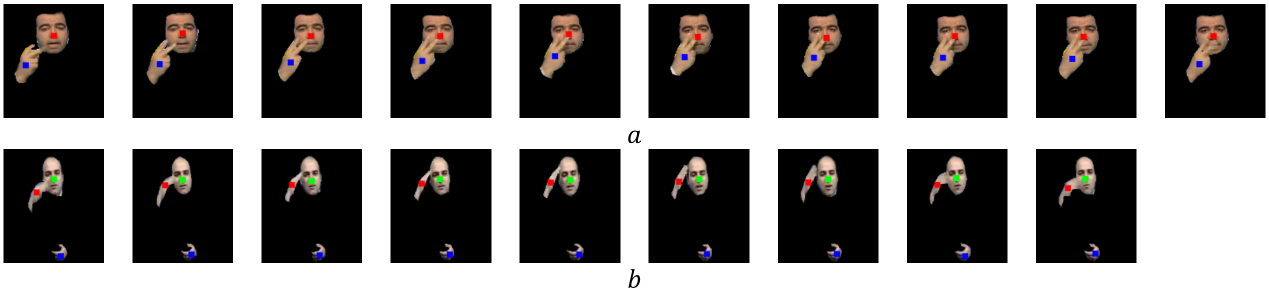


Fig. 10: Tracking with use semantic segmentation network using deep learning, then rough. a. “viewers” sign, b. “hello” sign.

Table 4: Method, MOTA, MOTP and Recall for Dataset

	Method	**MOTA**	**MOTP**	**Recall**	**Dataset**
proposed method1	Particle filter	0.971	0.935	0.948	Persian Deaf News
proposed method2	deep learning with rough	0.980	0.971	0.974	Persian Deaf News

Table 5: Method, MOTA, MOTP MOT16 Dataset

	Method	**MOTA**	**MOTP**
DeepMOT-Tracker [12]	DeepMOT-Tracker	0.548	0.772
Proposed method	Deep learning with rough	0.476	0.765

To test the proposed algorithm, another network is trained using MOT16 dataset. The results are shown in Table 5. It is compared with best result of [12].

3- Conclusion

Tracking is one of the most fundamental problems in computer vision, and use in a long list of applications such as sign languages recognition. We used rough set theory with deep Neural network for sign language tracking. The novelty of this paper is the use of rough set theory with deep neural network for tracking. This is the first work on this topic. This is the first on this topic. In this paper, at first, tracking using particle filter is explained. At second, tracking using rough sets and deep learning is explained. In the first proposed method, we used a particle filter, which has high accuracy but is very time consuming. The second proposed method responds much faster but is less accurate. To increase the accuracy of the second proposed method, we used the rough set theory. The system proposed are tested on 33 of Deaf News with 100 different words and 1927 video files for words, and recall, MOTA and MOTP values are obtained. Also, it with new mask is used for MOT16 dataset for comparing.

We focused our efforts on optimizing tracking with semantic deep network and rough set theory, but we want to use our proposed methods for sign language recognition.

References

[1] H. Tang, H. Liu, W. Xiao and N. Sebe, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion", Neurocomputing, Vol. 331, 2019, pp. 424-433.

[2] L. Kraljevi´c, M. Russo, M. Paukovi´c and M. Šari´c, “A Dynamic Gesture Recognition Interface for Smart Home Control based on Croatian Sign Language”, Appl. Sci. 2020, 10, 2300.

[3] A. Wadhawan and P. Kumar, "Sign Language Recognition Systems: A Decade Systematic Literature Review", Arch Computat Methods Eng 2019, https://doi.org/10.1007/s11831-019-09384-2.

[4] P. Kim, "MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence", Apress, 2017.

[5] A.E. Hassanien, A. Abraham, J.F. Peters, G. Schaefer, Henry C., "Rough sets and near sets in medical imaging: a review", IEEE Transactions on Information Technology in Biomedicine, Vol. 13(6), 2009, pp. 955-968.

[6] V. Sattari-Naeini, and A. Moaref, "Fuzzy-rough Information Gain Ratio Approach to Filter-wrapper Feature Selection", International Journal of Engineering, Vol. 30(9), 2017, pp. 1326-1333.

[7] D. Li, C. R. Opazo, X. Yu and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1459-1469.

[8] A. H. Mazinan, J. Hassanian, “A Hybrid Object Tracking for Hand Gesture Approach based on MS-MD and its Application,” Journal of Information Systems and Telecommunication (JIST), 2015, Vol. 3, No. 4.

[9] S. Ildarabadi, M. Ebrahimi, H. R. Pourreza, "Improvement Tracking Dynamic Programming using Replication Function for Continuous Sign Language Recognition", International Journal of Engineering Trends and Technology (IJETT), Vol 7(3), 2014, pp. 97-101.

[10] P. Chiranjeevi, S. Sengupta, “Rough-Set-Theoretic Fuzzy Cues-Based Object Tracking Under Improved Particle Filter Framework”, IEEE transactions on fuzzy systems, Vol. 24, 2016, No. 3.

[11] J. B. Zhang, T. R. Li, & H. M. Chen. "Composite rough sets for dynamic data mining", Information Science, Vol. 257, 2014, pp. 81–100.

[12] Y. Xu, A. Sep, Y. Ban, R. Horaud, L. Leal-Taixe and X. Alameda-Pineda, “How to train your deep multi-object tracker", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6787-6796, Jun. 2020.