An Efficient Video to Video Face Recognition using Neural Networks

Wilson S.
Research Scholar
Asst. Prof of Computer Science
CSI Jayaraj Annapackiam College
Nallur – 627853, Tamil Nadu, India.

Lenin Fred
Principal
Mar Ephraem College of Engineering and Technology, Marthandam
Tamilnadu

ABSTRACT
The interpretations at face images are difficult owing to its wide variations like appearance, individual, different facial poses and illumination. In biometrics video based face recovery is vital and this paper proposes an efficient algorithmic mode which achieves high recovery rate. The face recognition system proposed in this paper comprises of three stages video partitioning, feature extraction and neural network for recognition. The video partitioning was based on the changes in scene and feature extraction was carried out by local binary pattern and Principal Component Analysis. The algorithm is tested on four publically available datasets and the experimental results substantially prove that the proposed algorithm achieves higher face recognition rate when compared with the recent related work.

Keywords
Back propagation neural network, principal component analysis, facial features, Pearson Correlation Coefficient.

1. INTRODUCTION
Video based face recognition in image sequences has gained increased interest in the last few years. Since face recognition has some important challenges such as pose, illumination, etc. Video based face recognition is developed. A video sequence consists of continuous still images with different motion and illumination. Hence it is easy to recognize faces with all the possible challenges. The face recognition is primarily used in law (crime detection and surveillance) and commercial applications (credit cards, ATM cards and license verifications). The video based face recognition system concentrates on face with different poses and illumination, the traditional face recognition system concentrates on recognition from still images [1]. In the recognition of people from videos, efficient fusion of face, body traits and motion are done. The 3D face model or super resolved frames are determined from video sequences that improves the recognition results. The head movements and gestures play vital role in the face recognition from video sequences. The computational models for face recognition from video sequences are popular and its objective is to recognize the person from video with different pose and illumination [1]. The key challenge is exploiting the motion information available in the video and the variations in resolution, illumination, pose and facial expressions are really a threat in the development of efficient face recognition algorithm from video sequences [1-2]. Numerous techniques have been developed for the robust face recognition from video sequences [3-5].

In particular, Arandjelovic et. al. introduced three ideas: (i) photometric model of image formation is combined with a statistical model of generic face appearance variation to generalize in the presence of extreme illumination changes; (ii) the unseen head poses are identified using the smoothness of geodesically local appearance manifold structure and a robust same-identity likelihood and (iii) a reillumination algorithm is introduced to achieve robustness to face motion patterns in video. A fully automatic recognition system based on these ideas is developed [6].It also does not exploit temporal information. The recognition and tracking is concurrently performing generic video-face recognition algorithm that improves the accuracy [5] [7-11]. Zhou and Chellappa [9] presented a method for incorporating temporal information in a video sequence for human recognition. A state space model with tracking state vector and recognizing identity were used to characterize the identity. This approach aimed to integrate identity information though sequential importance sampling algorithm (SIS); This algorithm considered only identity consistency in temporal domain and thus it may not work well when the target is partially occluded. In [10], the head information is modeled as a texture mapped cylinder and tracking was formulated as an image registration problem in the cylinder's texture map image. Aggarwal et al. [11] presented a structured approach to the problem of video-based face recognition to recognize faces when both gallery and probe consists of face videos. In this framework, a moving face is represented as a linear dynamical system whose appearance changes with time. Subspace angles based distance metrics are used to get the measure of similarity between ARMA models representing moving face sequences . The choice of ARMA model is based on its ability to take care of the change in appearance while modeling the dynamics of pose, expression etc. The statistical method use subspace-based models and tools from Riemannian geometry of the Grassmann manifold improve the face recognition accuracy [12]. Intrinsic and extrinsic statistics are derived for the maximum-likelihood classification for video-based face recognition [13]. The video dictionaries comprise of temporal, pose and illumination information and it produces better results when compared to conventional video based face recognition techniques [14]. The adaptive GOP structure based on the method of frames comparison produces superior result when compared with fixed GOP structure in the improvement of coding efficiency of H.264/AVC [15-17]. The person recognition in unconstrained environment is really a challenge in video based face recognition for multimodal biometrics application and novel algorithms are required to solve the issues [18] [19] [20]. The correlations as well as coupling information between the video frames are incorporated in video based face recognition algorithm that produces good results even in the presence of noise and occlusion [21].

Chowdhury et al. [22] developed a method to estimate the pose and lighting of face images contained in video frames and compares them against synthetic 3D face models exhibiting similar pose and lighting. This method is capable to handle situations where the pose and lighting conditions in the training and testing data are completely disjoint.

Liu et al. [23] introduced an adaptive Hidden Markov Model (HMM) to perform video-based face recognition. In [24], kernel principal angles, applied on the original image space and a feature space, are used as the measure of similarity between two video sequences. Zhou et al [25] propose a tracking-and-recognition approach by resolving uncertainties in tracking and recognition simultaneously in a probabilistic framework. Lee et al [26], represent each person by a low dimensional appearance manifold, approximated by piecewise linear subspaces. They present a maximum a posteriori formulation for recognizing faces in test video sequences by integrating the likelihood that the input image comes from a particular pose manifold and the transition probability to this manifold from the previous frame. Among the methods mentioned, Lee et al [26] method seems to be the one most capable of handling large 2-D and 3-D rotations. Although many previous methods make use of temporal information present in face videos to improve recognition, there has been no attempt to model a moving face as a dynamical system. In [27], an attempt is made to explore a method for modeling a moving face as a linear dynamical system to perform recognition. Each frame of a video is, therefore, assumed to be the output of the dynamical system particular to the subject. The goal of the work is to extract features with different poses illumination. For this purpose, key frames are extracted from the video to identify the change in pose and illumination. In the previous methods, only motion are analyzed which may not identify small change in pose. In the proposed method, key frames are identified by correlation between frames. The proposed method extracts the features by grouping the frames in the video according to the pose and illumination. These features are given to neural network for classification. By grouping the frames, features for each pose is correctly extracted. The proposed method is compared with state-of-the-art methods and with the recent works.

The rest of the paper is organized as follows: Section 2 describes the proposed method with architecture. The performance of the proposed method is analysed and compared with some existing methods which are discussed in Section 3 followed by conclusion in Section 4.

2. PROPOSED ARCHITECTURE

The key point of the proposed architecture is to cluster the faces with similar poses. The input video sequence is first divided into partitions based on the changes in scene. In each partition, one representative frame is chosen which is called as key frame. From the key frames, facial features are extracted and it is given to neural network for recognition. The overall proposed system architecture is shown in Fig. 1. The video sequence is splitted into partitions based on the changes in scene as illustrated in [15]. The scene change is calculated using Pearson Correlation Coefficient (PCC). PCC is widely used to measure the similarity of two frames for cut detection [16]. The value of PCC lies in the range between 0 (no correlation) and1 (perfect correlation). Correlations value above 0.40 is considered as really high and values below will be determined as cuts [16]. The Pearson correlation coefficient for two dimensional signals like video sequences is conveyed below.

$$PCC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(g(i,j)-g^{m})(g_p(i,j)-g_p^{m})}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(g(i,j)-g^{m})^2(g_p(i,j)-g_p^{m})^2}} \quad \text{.....(1)}$$

Where and are the x- coordinates of the frames for which the correlation is calculated. and are y- coordinates of the frames for which the correlation is calculated.

Correlation is calculated between a kth frame and its nearest 10th frame. If the correlation is below threshold, 9th frame is compared; otherwise, 11th frame is compared. The process is repeated till the correct cut is detected. After the cut is detected, the video is divided into partitions. The threshold value is set to 0.4, as small changes are important to know the facial features. In each partition, the first frame is selected as the key frame. From the key frames selected, features such as LBP, Regional Directional Weighted Local Binary Pattern (RDW-LBP), 2D-PCA are extracted and given to neural network. The illustration of partitioning video sequence by identifying key frames is shown in Fig. 2.

The features obtained from key frames are classified using Back Propagation Neural Network (BPNN). Since facial images have more features to be learned, it needs some training. Neural network is the best method for training and classification.

The BPNN comprises of three layers: input layer, hidden layer and the output layer. The feature extracted are propagated to the nodes in the input layer. The input layer propagates feature vectors to each node in the middle layer. The middle layer nodes compute output values which are given to the output layer nodes. The output layer nodes compute the network output for the particular feature vector.

In BPNN method, weights can be initialized and random values of weights lead to error. Weights are initialized through series of training. By analyzing the error, weights are changed and given back to input layer. The error values for each node are computed in the output layer and middle layer nodes. This is done by assigning a part of the error due to the middle layer node which feed that output node. The amount of error due to middle layer node depends upon size of the weight assigned to the connection between the two nodes.

3. EXPERIMENTAL ANALYSIS

The proficiency of the proposed algorithm was evaluated through experiments carried out on four publicly accessible datasets: the UMD dataset [17], the Multiple Biometric Grand Challenge (MBGC) dataset [18] [19], the Honda/UCSD dataset [5] and the FOCS UT-Dallas Video.

3.1 UMD video

The UMD dataset consists of 12 videos recorded with a group of 16 subjects and it was collected in HD format (1920 × 1088 pixels). It comprises of standing sequences and walking sequences. The video of subjects standing without walking toward the camera are referred as standing sequences and sequences of each subject walking toward the camera are referred as walking sequences. The video sequences are segmented according to subjects and sequence types. After segmentation, 93 sequences are obtained comprising of 70 standing sequences and 23 walking sequences.

In the video sequences, sometimes the faces were some subjects having conversations and others were looking elsewhere, their faces were sometimes non-frontal or partially corked. The walking subject's head sometimes turned to the right or left showing a profile face. Furthermore, for both types of sequences, the camera was not always static.

Table 1 shows the recognition rate achieved by the proposed method and this method is compared with other state-of-the-art methods. It is proved that the proposed method achieves recognition rate slightly higher than the other methods.

3.2 MBGC Video version 1

The MBGC Video version 1 dataset (Notre Dame dataset) has 399 walking (frontal-face) and 371 activity (profile-face) video sequences recorded from 146 subjects. These were collected in two formats namely SD format (720 × 480 pixels) and HD format (1440 × 1080 pixels). The 399 walking sequences consist of 201 sequences in SD and 198 sequences in HD. In 371 walking video sequence, 185 sequences are in SD and 186 sequences are in HD. The challenging conditions in these videos include frontal and non-frontal faces in shadow. Some of the example faces are shown in Fig. 3.

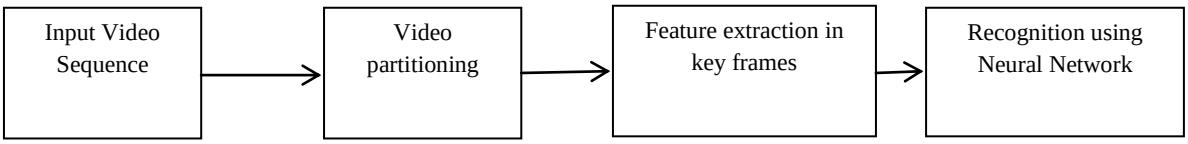

Fig. 1 Proposed System Architecture

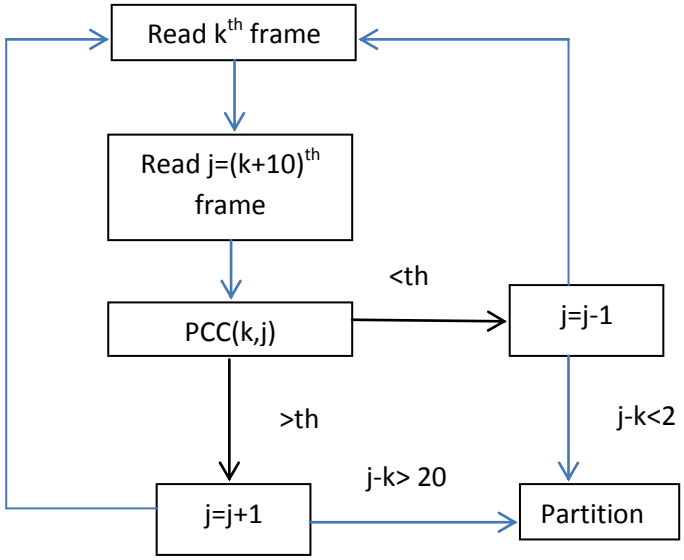

Fig. 2 Illustration of partitioning video sequence

Table 1 Recognition rate comparison of the proposed method with other methods in UMD dataset

UMD Videos	PM [12],[20]	KD [12],[20]	WGCP [12]	SANP [13]	SRV [21]	KSRV [21]	Proposed Method
S2	82.80	81.72	82.97	92.47	92.47	93.55	93.64
S3, S4, S5	84.62	83.52	83.52	93.41	94.51	94.51	94.95
S6	98.04	96.08	88.23	98.04	98.04	98.04	98.21
Average	88.49	87.11	84.91	94.64	95.01	95.37	95.6

Table 2 Recognition rate comparison of the proposed method with other methods in MBGC video dataset

MBGC walking videos	Procrustes Metric [12], [20]	Kernel Density [12], [20]	WGCP [12]	SANP [13]	DFRV [14]	SRV [21]	KSRV [21]	Proposed Method
S2	43.79	39.74	63.79	83.88	85.64	86.65	86.65	87.1
S3	53.88	50.22	74.88	84.02	88.13	87.67	88.58	88.94
S4	53.70	50.46	75	84.26	88.43	87.96	88.89	89.23
Average	50.46	46.81	71.22	84.05	87.40	87.43	88.04	88.43

Fig. 3 Examples from MBGC dataset

In this paper, leave-one-out identification experiments on 3 subsets of cropped face images from the walking videos were conducted. The 3 subsets present in the experiments are S2 (144 subjects, 397 videos), S3 (55 subjects, 219 videos) and S4 (54 subjects, 216 videos).Table 2 tells out the percentage of recognition rate for the proposed method. The result of proposed method is good when compared with other methods.

3.3 Honda/UCSD Dataset

The Honda dataset [5] contains 59 video sequences recorded from20 subjects. It follows the experimental procedure presented in [13]. The experiments are performed in three cases of the maximum set length as defined in [13]. The lengths used in analysis are 50, 100 and full length frames. Image resolution is 20 × 20 pixels. Table 3 lists the identification rates of proposed methods and other ten state-of-the-art methods [28-31], [13]. It is observed that the proposed method obtained the highest average recognition rates. The identification of keyframes in Honda dataset are shown in Fig. 4.

3.4 FOCS UT-Dallas Video

Finally, experiments are carried out in the challenging dataset: UT Dallas video sequences contained in the Face and Ocular Challenge Series (FOCS) [6]. The FOCS UT Dallas dataset has 510 walking (frontal face) video sequences and 506 activity (non-frontal face) video sequences recorded from 295 subjects having frame size of 720 × 480 pixels. These sequences were collected on different days. In walking sequences, place the subject far away from the video camera, walk towards it with a frontal pose and finally turns away from the video camera showing the profile face.

The same leave-one-out tests were tested on 3 subsets: S2 (189 subjects, 404 videos), S3 (19 subjects, 64videos) and S4 (6 subjects, 25 videos) from the UT-Dallas walking videos.

Table 4 shows recognition rate results. The proposed method achieves best recognition rates among all the compared algorithms.

Table 3 Recognition rate comparison of the proposed method with other methods in Honda dataset

Set length	50 frames	100 frames	Full Length	Average
DCC [22]	76.92	84.62	94.87	85.47
MMD [23]	69.23	87.18	94.87	83.76
MDA [24]	74.36	94.87	97.44	88.89
AHISD [25]	87.18	84.62	89.74	87.18
CHISD [25]	82.05	84.62	92.31	86.33
SANP [13]	84.62	92.31	100	92.31
DFRV	89.74	97.44	97.44	94.87
CHISD [25]	82.05	84.62	92.31	86.33
SRV[21]	94.87	97.44	97.44	96.58
KSRV [21]	94.87	97.44	97.44	96.58
Proposed Method	95.21	98.12	98.12	97.15

4. CONCLUSION

Video-to-video face recognition has gained more interest in recent years. Face with different poses and illumination are recognized using feature extraction by grouping the frames. The face is recognized using Back Propagation Neural Network. Experiments are carried out in publicly available datasets the Multiple Biometric Grand Challenge (MBGC), the Face and Ocular Challenge Series (FOCS), the Honda/UCSD and the UMD Comcast10 datasets. The performance of the proposed method is finding out by comparing the recognition rate with the state-of-the-art methods. It is proved that the proposed method achieves approximately 0.5% higher recognition rate with all other methods.

5. REFERENCES

[1] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," ACM Computing Surveys, pp. 399–458, Dec. 2003.

[2] A. J. O'Toole, P. J. Phillips, S. Weimer, D. A. Roark, J. Ayyad,R. Barwick, and J. Dunlop, "Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach", Vision Research, vol. 51, no. 1, pp. 74–83, 2011.

[3] A. Ross, K. Nandakumar, and A. K. Jain, "Handbook of Multibiometrics" Springer, 2006.

[4] M. Tistarelli, S. Z. Li, and R. Chellappa, Handbook of Remote Biometrics: For Surveillance and Security. Springer, 2009.

[5] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds", Computer Vision and Image Understanding, vol. 99, pp. 303–331, 2005.

[6] O. Arandjelovic and R. Cipolla, "Face recognition from video using the generic shape-illumination manifold," European Conference onComputer Vision, vol. 3954, pp. 27–40, 2006.

[7] G. Hager and P. Belhumeur, "Efficient region tracking with parametricmodels of geometry and illumination," IEEE Transactions on PatternAnalysis and Machine Intelligence, vol. 20, no. 10, pp. 1025–1039,Oct. 1998.

[8] A. Lanitis, C. Taylor, and T. Cootes, "Automatic interpretation andcoding of face images using flexible models," IEEE Transactions onPattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 743–756, July 1997.

[9] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking andrecognition using appearance-adaptive models in particle filters," IEEETransactions on Image Processing, vol. 13, no. 11, pp. 1491–1506,Nov. 2004. [51]

[10] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head trackingunder varying illumination: an approach based on registration oftexture-mapped 3d models," IEEE Transactions on Pattern Analysisand Machine Intelligence, vol. 22, no. 4, pp. 322–336, Apr. 2000.

[11] G. Aggarwal, A. Veeraraghavan, and R. Chellappa, "3D facial posetracking in uncalibrated videos," International Conference on PatternRecognition and Machine Intelligence, 2005.

[12] P. K. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa,"Statistical computations on grassmann and stiefel manifolds for image and video-based recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 11, pp. 2273–2286, Nov. 2011.

[13] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest pointsfor image set classification", IEEE Conference on Computer Vision and Pattern Recognition, pp. 27–40, 2011.

[14] Yi-Chen Chen, Vishal M. Patel, P. Jonathon Phillips, and Rama Chellappa, "Dictionary-based Face Recognition from Video", Springer, Computer Vision. 2012, Vol. 7577, Lecture Notes in Computer Science, pp. 766-779.

[15] S Sowmyayani and P ArockiaJansi Rani ,"Adaptive GOP structure to H.264/AVC basedon Scene change", ICTACT journal on image and video processing: special issue on videoprocessing for multimedia systems, August 2014, Vol: 5, Issue:1, pp. 868-872

[16] Lenka Krulikovsk´a and JaroslavPolec, "GOP Structure Adaptable to the Location of Shot Cuts", International Journal of Electronics and Telecommunications, 2012, vol. 58, no. 2, pp. 129–134.

[17] R. Chellappa, J. Ni, and V. M. Patel, "Remote identification of faces: problems, prospects, and progress," Pattern Recognition Letters,vol. 33, no. 15, pp. 1849–1859, Oct. 2012.

[18] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J.O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M.Lui, H. Sahibzada, J. A. Scallan III, and S. Weimer, "Overview of the multiple biometrics grand challenge," International Conference onBiometrics, 2009.

[19] National Institute of Standards and Technology, "Multiple biometric grand challenge(MBGC)." http://www.nist.gov/itl/iad/ig/mbgc.cfm

[20] P. K. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysison stiefel and grassmann manifolds with applications in computer vision," IEEE Conference on Computer Vision and Pattern Recognition,pp. 1–8, 2008.

[21] Yi-Chen Chen, Vishal M. Patel, SumitShekhar, Rama Chellappa and P. Jonathon Phillips, "Video-based Face Recognition via Joint Sparse Representation", IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 22-26 April 2013, pp. 1-8.

[22] A. Roy-Chowdhury and Y. Xu (2006), Pose and Illumination Invariant Face Recognition Using Video Sequences, Face Biometrics for Personal Identification: Multi-Sensory Multi-Modal Systems, Springer-Verlag, pp. 9-25.

[23] X. Liu and T. Chen (2003), "Video-based face recognition using adaptive hidden markov models", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 340-345.

[24] L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. In Proc. of Intl. Conf. on Computer Vision and Pattern Recognition, 2003.

[25] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. Computer Vision and Image Understanding, 91:214–245, 2003.

[26] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Videobase face recognition using probabilistic appearance manifolds. In Proc. of Intl. Conf. on Computer Vision and Pattern Recognition, 2003.

[27] Gaurav Aggarwal, Amit K. Roy Chowdhury, Rama Chellappa," A System Identification Approach for Video-based Face Recognition".

[28] M. K. Kim, O. Arandjelovic, and R. Cipolla, "Discrimitive learningand recognition of image set classes using canonical correlations", IEEE Transactions

on Pattern Analysis and Machine Intelligence,vol. 29, no. 6, pp. 1005–1018, June 2007.

[29] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distancewith application to face recognition based on image set," IEEEConference on Computer Vision and Pattern Recognition, pp. 1–8, 2008.

[30] R. Wang and X. Chen, "Manifold discriminant analysis," IEEE Conference on Computer Vision and Pattern Recognition, pp. 429–436, 2009.

[31] H. Cevikalp and B. Triggs, "Face recognition based on image sets", IEEE Conference on Computer Vision and Pattern Recognition, pp.2567–2573, 2010.

6. AUTHOR PROFILE

Prof. Dr. A. Lenin Fred received B.E in Computer Science and Engineering from Madurai Kamaraj University, India, in 1995 with First class. He has passed M.E. Degree in the same discipline in Madurai Kamaraj University, India and awarded in 2001 with First class. He received his Ph.D. (Doctor of Philosophy) in Computer Science and Engineering at Manonmaniam Sundaranar University, Tirunelveli, India in 2010. His research interest has widen over a variety of fields such as Information Technology, Digital Image Processing, Biometrics, Automatic Fingerprint Identification System, Fingerprint Feature Extraction, Biometric Fusion, Feature Level Integration in Multiple Biometrics.

S. Wilson received his MCA degree from Manonmaniam Sundaranar University in Tirunelveli, India in 1998 with First class. On 2008, he passed out M.Phil degree in the field of computer science from the same university. Again, on 2010, he has completed M.Tech (Computer Science and Information technology) degree with First class. He passed State Eligibility Test (SET) for lectureship on 2012 conducted by Bharathiar University, Coimbatore.

7. APPENDIX

Table 4 Recognition rate comparison of the proposed method with other methods in FOCS UT-Dallas video dataset

UT-Dallas walking videos	Procrustes Metric [14],[12]	Kernel Density [14],[12]	WGCP [14]	SANP [15]	DFRV	Proposed Method
S2	38.12	40.84	53.22	48.27	59.90	60.42
S3	60.94	64.06	70.31	60.94	78.13	78.88
S4	64	64	76	68.00	80.00	81.18
Average	54.35	54.97	66.51	59.07	72.68	73.49