International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869, Volume-3, Issue-5, May 2015

Survey on Recent Techniques in Content Based Video Retrieval

Dr. H.B. Kekre, Dr. Dhirendra Mishra, Ms. P. R. Rege

Abstract— **Content based video retrieval (CBVR) utilizes the rich and varied video contents for video representation and retrieval. The contents can be broadly divided into static frame level contents, spacio-temporal contents, motion contents and high level semantic contents. Many successful techniques have been proposed in various literature which focus on these levels. Few of these recent ones are highlighted in this paper. Work done on static frame level use well researched techniques also used in CBIR. These are used when there is a wider variety of videos to retrieve. Techniques which use the dynamic contents work well on videos which have unique characteristic motions as their identifying factor. The temporal and motion information is utilized for retrieval purposes. The feature extraction methods used demonstrate varying degrees of computation complexity and performance. The extraction of high level semantic contents requires incorporation of techniques capable of utilizing the low and middle level video data to extract the topic or subject of the video. Analysis of high level semantic is performed using learning models promising higher level of satisfaction in retrieval. The recent trends in CBVR aim for this higher semantic retrieval. As low level contents are the base for extracting semantic content, improvements in handling of low level contents will also be necessary for contribution to this trend.**

Index Terms— **CBVR, frame content, motion content, spacio-temporal content, semantic content, video feature extraction**

I. Introduction

Video data is being stored in repositories in large numbers. Retrieving the relevant video from a huge repository is a difficult task. Earlier tagging of videos was performed for easy retrieval. This has limitations in terms of person interpreting the contents of video and the exhaustive textual descriptions needed to tag a particular video. Thus there was a need for identifying and describing a video based on its contents for retrieval purpose.

The main components of a CBVR as shown in figure 1 are 1) feature extraction of a video based on its content, 2) storing the feature vector obtained 3) feature extraction of a user query image/clip and 4) matching the feature vectors for retrieval purpose.

The task of feature extraction of a video, based on its content is the most important component of any CBVR. It is a major research area because of the nature of content found in videos. Video content is multimodal and multidimensional due to the visual, textual, audio and temporal data it usually contains. It also includes visual semantic data at the frame, shot, scene, clip level. This rich and varied video content is utilized in CBVR for extraction of features. Structurally, videos consists of scenes containing shots containing frames. Frames, at the lowest level structurally, are static images. Video structure analysis is a prerequisite for videos containing multiple scenes and shots of varied content and it is necessary to obtaining content at frame, shot or scene level. Structural analysis includes detection of representative key frame of a shot and detection of boundaries of a shot [1]. After structural analysis, video content can be identified at required structural levels. Content wise, video features can be categorized into low level, middle level and high level [2]. Low level content are color, shape, contour, texture, entropy, motion. Middle level content is 3D motion features like object trajectory and camera motion [2]. High level features contributing to the visual semantics are objects, actions, simple events/ activities and complex events. Further semantics involved are the concepts, stories or subject in video. Thus, video content of any level and type can become the base for extraction of feature vector. This paper majorly discusses this component of CBVR.

The further components in CBVR are comparison and retrieval. Once a representation for videos as well as query is ready they can be compared using a suitable similarity measure and relevant videos can be retrieved. An important factor affecting performance of video retrieval is the similarity measure used [3] for comparing these representative features.

Various feature extraction methods available in literature are based on the kind of video content that is focused on. As a key frame contains sample contents of a shot its contents are

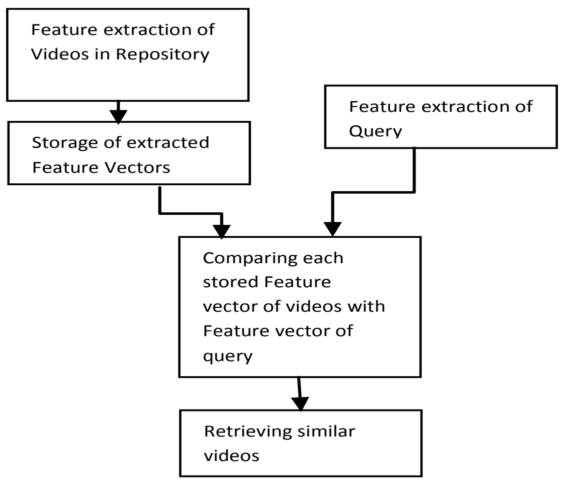

Figure 1: Main components of CBVR

used for feature extraction. Section II deals with these techniques. Section III explores video retrieval techniques which extract a video's dynamic content. They utilize spacio-temporal content and motion content extraction.

Dr. H.B. Kekre, Sr. Professor, Computer Engineering Dept., MPSTME, SVKM's NMIMS (Deemed University), Vile Parle(W), Mumbai -56

Dr. Dhirendra Mishra, Associate Professor and Head,Computer Engineering Dept, MPSTME, SVKM's NMIMS (Deemed University), Vile Parle(W), Mumbai -56

Ms. P.R.Rege,PhD Research Scholar, Computer Engineering Dept, MPSTME, SVKM's NMIMS (Deemed University), Vile Parle(W), Mumbai -56

Section IV discusses extraction/representation of high level semantic attempted successfully by researchers. The paper ends with a conclusion and comments on the recent trends noticed.

II. FRAME CONTENT BASED

At the lowest structural level of a video lies the static images or frames of a video. A video shot is comparable to a sequence of images/frames. The low level frame contents like color, texture, shape etc. have been used to represent videos. A variety of feature extraction techniques and similarity measures are employed for efficient retrieval. These techniques are comparable to video adaptations/extensions of content based image retrieval techniques. Invariance of color correlation is used as a technique for video retrieval by Yanqiang Leiet. al. in [4]. In their method each frame is divided into separate blocks. A small size frame feature is formed by sorting the red, green and blue color components of each block based on the average of their intensity values and taking percentage of color correlation. The authors show that the features of a frame thus obtained are immune to operations like noise addition, shape distortion, blurring, enhancing the contrast and strong re-encoding. Their proposed method outperforms traditional color histogram method with satisfactory time and space complexity.

The technique of evolutionary population based search algorithm can be used to reduce the number of frames required to be used for comparison with user query image. Particle swarm optimization (PSO) can be used to retrieve frames within the video library as proposed by Salahuddin, A. et. al in [5]. Their technique requires that each swarm particle be evaluated for its fitness using degree of similarity. The similarity measure used are correlation based template matching, result from scale-invariant feature transform (SIFT) and convolution. The relative best match in each generation of PSO is shown to the user. Real video library is used for experimentation purpose.

Key frames can be transformed using various transforms (DCT, DST, Haar, Hartley, Kekre, Slant, Walsh,etc) and a fraction of the coefficients obtained can be used as feature vector [6]. Kekre H.B. et. al. have used it to reduce the computational complexity. They have shown that feature vectors with a smaller fraction of coefficients give better average precision and recall for CBVR than full set of coefficients. The performance measure that they use is the crossover points of average precision and recall values for various transforms. The results on their specific choice of 500 videos shows Haar transform performing the best. Other observations made are reduction of performance of Kekre transform with reduction in size of coefficients, no change for Hartley and Slant transforms for decrease in size of coefficients and DCT, DST, Haar, Walsh transforms giving their best performance at 0.048% fraction of coefficients.

Another CBVR technique uses color feature extraction by block truncation coding (BTC). Its extension called multi-level Thepade's Sorted Ternary BTC (TSTBTC) is used in [7].S. D. Thepade et. al. apply it on even odd videos and on intermediate blocks of videos for representation videos. They show that the method performs best using KLUV color space for multi-level and even odd videos. They have found that for intermediate blocks the YIQ color space works best.

Multiple low level feature have also been employed for retrieval. Entropy of key frames along with extracted black and white edge points are used for video retrieval by B. V. Patel et. al. in [8]. S. Padmakala et. al. in their paper [9] retrieve videos for a query using feature vector generated with a combination of features obtained from two different schemes. The first scheme extracts video features by finding the color moments and texture analysis of objects obtained from video segmentation. While the other scheme uses probability of occurrence of the a particular pixel intensity at a location in every frame of video. For the query video clip, the aforesaid features are extracted and compared with the feature in the feature library. Saluja G. et. al. in [10]extract frames at fixed time intervals from the videos and hash them into feature vectors. They implement a layered filtering technique on four features, which are, corners, adjacent pixel intensity difference, color distribution and edges. For retrieval of similar videos they use histogram based comparison for each feature.

III. DYNAMIC CONTENT BASED

The dynamic content of videos, the spacio-temporal contents and motion content, can also be utilized for retrieving motion videos. There is generally a foreground and a background to every visual frame. The motion characteristics change this content of frames in time. This change occurs due to camera motion and or object motion. Camera motion, the major contributor to global motion, is generally due to zooming, panning, tilting etc. which leads to the change in background of the scene in the video. Local motion or motion of objects change the foreground. A good representation of motion in videos can be used as a query for retrieval of similar videos.

Spacio-temporal feature curves of videos are formed by taking into consideration spacial contents of each frame of a video and stringing them together to form curves as proposed by Xiuxin Chen et. al in [11]. N. Dimitrova et. al. in [12] have strung together macroblock trajectories for object motion description. Motion of objects and their trajectories, represented as freehand sketches, can be used as query mechanism. Multi-Spectro Temporal-Curvature Scale Space (MST-CSS) feature representation can be obtained for the query and matched with a set of MST-CSS features generated offline from the video clips in the database as proposed by Chiranjoy Chattopadhyay et. al. in [13]. The authors mention a disadvantage of their technique, that is, its inadequacy to capture the salient features of the MST-CSS surface leading to unsatisfactory retrieval results and enhance it in paper [14] with EMST-CSS (Enhanced MST-CSS) as a better feature representation with an improved comparison technique for CBVR. Using one synthetic and two real-world datasets they show enhanced performance with their own previous MST-CSS representation and other current methods for CBVR.

As motion content in a video can either be directional (like football, basketball, etc.) or can have magnitude (like explosion videos, volcano eruption, approaching object video etc.) authors Ying Chen et. al. in [15] use an optical flow algorithm for motion information extraction and Haar wavelet for building the representative feature vector. They use two-dimensional non standard Haar wavelet to speed up the wavelet transform process. Wavelet transform can decompose the frame data into wavelet coefficients on different scales and then wavelet coefficients can be used to represent the original data. The measure used for performance evaluation of retrieved shots are average normalized modified retrieval rank (ANMRR) and average recall (AR). ANMRR gives the rank of correct shots not retrieved and AR determines the rate of retrieving correct shots. Thus ANMRR value should be low and AR value must be high to indicate better performance.

Spacio-temporal content of videos are explored by A. Lakshmi et. al. in paper [16]. The authors present a new spatio temporal key-point detector and descriptor using 3D complex wavelet transform. Obtained key-points are then converted to spatio temporal features to represent videos.

Motion vectors in MPEG bit stream and additionally other frame contents together give more semantic information about type of motion in a video, as proposed by Chih-Wen Su et. al in [17]. The other frame contents they utilize are color distribution, consistent of motion direction and the common area between macroblocks of two consecutive frames. This additional information allows the linking of motion vectors into more meaningful "motion flow" like trajectories. They also handle the situation of large moving objects occupying multiple macro blocks by replacing similar shaped motion flows into one or more motion flows to represent the motion. They also handle motion of a large object with different parts moving independently by allowing separate representative motion flows for that object.

Video retrieval can be of help in retinal surgery as explored by Zakarya Droueche et.al in [18]. The authors use video streams of an ongoing surgery as query to a digital archive of surgery videos. The retrieved similar videos tell the surgeon what other experienced doctors have done in similar situations. For this purpose the authors used motion information contained in MPEG- 4 AVC/H.264 video standard. They base one of their techniques on motion histogram of compressed video sequence. They use this to extract motion direction and intensity statistics. For comparison with archived videos they use extended fast dynamic time warping to multidimensional time series.

IV. HIGH LEVEL CONTENT BASED

Querying for video retrieval can be enhanced by allowing semantic video retrieval. Research in the field of video analysis where task of recognizing object, action and activity/event in videos is performed can be used for this purpose. This section discusses the ongoing research for recognition of these high level video contents. Object recognition is handled in [19-22]. Action recognition (human action) is handled in [23,24,25,26]. Common human actions involving only body movements are handwaving, handclapping, running, walking, jogging etc. Human actions with object interaction are mixing, pouring, shooting, kicking-ball etc. Sports player actions are handled in [23]. Surveillance videos capture activities of people. These are mostly normal activities. However a few abnormalities or anomalies in activities can occur and need to be recognized. Papers [27,28] successfully attempt activity recognition using scene/context in which the activity is taking place. Papers [29,30,31] are aimed at event recognition.

For retrieval of video frames/videos containing object/objects in query image various techniques are proposed in literature. Video object retrieval requires object detection and recognition. More than a decade ago in [19] Di Zhong et.al. explored region-based analysis for video object segmentation and retrieval. More recently, a multi-scale segmentation strategy is proposed in [20] by Camilo C. Dorea et. al. which uses region merging technique. Region merging is progressively complex for defining increasing aptitude partition layers for object detection. Giovani Gualdi et. al. detect objects using statistical-based search method in [21]. Object representation and mining is performed in [22] by Arasanathan Anjulan et.al. The shot features are grouped into object clusters which are used to mine frequently appearing objects in video. Object mining is demonstrated on full length feature films.

Further into the semantics of video content, methods for recognizing human actions are being explored. Player action recognition in sports video is attempted in [23] by Haojie Liet. al. After player body segmentation from jump and diving videos, action recognition is performed using Hidden Markov Models as a tool for sequential pattern recognition. Xingxiao Wu et. al. perform action recognition using multilevel features and latent structural support vector machines (SVM) in [24].Chungfeng Yuan et. al. in [25] perform action recognition by using 3D covariance descriptors of local features and represent action with a spacio-temporal matrix which contains geometric-temporal information along with the appearance information. In [26] Jianzhai Wu et. al. make an observation that different actions may share few common features making it difficult to differentiate between them, however, each action has an image sequence pattern containing a crucial motion pattern for identifying that action. They also observe that for recognizing multiple human actions in real-world unconstrained videos a well trained model will be required, which in turn requires a large training dataset. In [27] Yingying Zhu et. al. aim to recognize activity and detect anomaly using trained model. They train their data with labeled normal activities. Their model captures frequent motion and context pattern for each activity class. The learned model is used to label the testing videos. MyoThida et. al. handle the problem of detecting and localizing abnormal activities in crowded scenes in [28].

Events occurring in videos are a collection of multiple human actions. Videos can be retrieved by querying for specific events. Event recognition in real-world videos is attempted by Xiang Ma et. al. in [29]. Authors use multiple interactive motion trajectories obtained from object trajectories for this task. A sample event indexed by two interacting motion trajectories is "two people meet-fight-chase". Their work is limited to a maximum of 5 trajectories. Their tensor-based reduced dimension representation of multi-object trajectories assists in fast retrieval. Michele Merler et al. in [30] recognize complex events in TRECVID MED10 dataset like "assembling a shelter", "baking a cake" and " batting a run" involving multiple human actions. Their proposed "semantic model vector" representation helps recognize semantics of complex events. Xiofeng Wang et. al. perform sports video event classification in [31]. The events handled are "bowling shot", "full swing" in golf videos etc. To avoid the limitations of hidden Markov model (HMM) they use hidden conditional random field (HCRF) model which can analyze contents of a video content better. They use independent component analysis (ICA) mixture in their proposed feature function.

V. CONCLUSION

To give a video retrieval system the desired capability, effective handling of varied and voluminous video content is required. Some techniques used are computationally intensive while some address needs of specific categories of videos only like medical videos, sport videos, news videos, surveillance videos etc. having their own peculiarities in context of video retrieval. They extract either static low level video features, dynamic spacio-temporal and motion features or high level semantics to characterize a video for retrieval purpose. They demonstrate varying computation intensity and performance.

The recent trends in research in this area aim at semantic retrieval. However faster extraction of low level features is still the basic need. Transforms on key frames and using their fractional coefficients facilitate faster feature comparison and video handling. Also, videos in compressed domain avoid decompression delays and techniques using compressed videos is also gaining attention. Future work in these areas can contribute to better performances.

REFERENCES

[1] R. Krishnamoorthy, M. Braveen, "Object Level Frame Comparison for Video Shot Detection with Orthogonal Polynomials Model", Proc. International Conference on Emerging Trends in Science, Engineering and Technology, 2012

[2] Wei Pan, Francois Deschenes, "Interpreting Camera Operations in the Context of Content-based VideoIndexing and Retrieval", Proc. 3rd Canadian Conference on Computer and Robot Vision (CRV'06), 2006

[3] Xiang Bai, Bo Wang, Cong Yao, Wenyu Liu, ZhuowenTu, "Co-Transduction for Shape Retrieval", IEEE transactions on Image Processing, vol. 21, no. 5, May 2012

[4] Yanqiang Lei, WeiqiLuo, Yuangen Wang, Jiwu Huang, "Video Sequence Matching Based On The Invariance Of Color Correlation", IEEE Transactions On Circuits And Systems For Video Technology, vol. 22, no. 9, September 2012

[5] SalahuddinA., NaqviA., MurtazaK.,AkhtarJ., "Content Based Video Retrieval Using Particle Swarm Optimization" , Proc. 10th International Conference on Frontiers of Information Technology (FIT), DOI: 10.1109/FIT.2012.23, pp. 79 – 83, 2012

[6] Dr. H.B. Kekre, Dr. Sudeep D. Thepade& Saurabh Gupta, "Content Based Video Retrieval in Transformed Domain using Fractional Coefficients", International Journal of Image Processing (IJIP), vol. 7, Issue (3), 2013

[7] S. D. Thepade, K. S. Subedarpage, A. A. Mali, "Performance Rise In Content Based Video Retrieval Using Multi-Level Thepade's Sorted Ternary Block Truncation Coding With Intermediate Block Videos And Even-Odd Videos", Proc. International Conference on Communication and Signal Processing, April 3-5, 2013, India

[8] B. V. Patel, A. V. Deorankar, B. B. Meshram, "Content Based Video Retrieval using Entropy, Edge Detection, Black and White Color Features", Proc. 2nd International Conference on Computer Engineering and Technology (ICCET), 2010, vol.6, pp.V6-272 - V6-276

[9]S.Padmakala, Dr.G.S.AnandhaMala, M.Shalini, "An Effective Content Based Video Retrieval Utilizing Texture, Color and Optimal Key Frame Features", International Conference on Image Information Processing (ICIIP 2011), 978-1-61284-861-7/11, 2011

[10] SalujaG., RokdeA.,MaruR., KondekarR., Gupta, A.,Deshpande P., "Layered filtering technique for content based video retrieval", Proc. International Conference onComputer & Information Science (ICCIS), June 2012, vol. 1, pp.420 – 424

[11] Xiuxin Chen, KebinJia, Zhipin Deng, "A Video Retrieval Algorithm Based on Spatio-temporal Feature Curves and Key Frames", Proc. Fifth International Conference on Intelligent Information Hiding and MultimediaSignal Processing, 2009

[12] N. Dimitrova, F. Golshani, "Motion Recovery for video content classification", ACM Transaction on Information Systems, vol. 13, no.4, pp. 408-439, 1995

[13] ChiranjoyChattopadhyay and Sukhendu Das, "A Motion-Sketch Based Video Retrieval Using MST-CSS Representation", Proc. IEEE International Symposium on Multimedia, 2012

[14] ChiranjoyChattopadhyay and Sukhendu Das, "Enhancing the MST-CSSrepresentationUsing Robust Geometric Features for Efficient Video Retrieval (CBVR)", Proc. International Symposium on Multimedia(ISM), 2012

[15] Ying Chen, Ou Wu, "Motion Video Retrieval Based on Optical Flow and Haar Wavelet", Proc. Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011

[16] A. Lakshmi, SubrataRakshit, "Spatio temporal key-point detection and video representation in 3D complex wavelet domain",Proc. International Conference on Signal Processing and Communications (SPCOM), pp.1-5, 2012

[17] Chih-Wen Su, Hong-Yuan Mark Liao, Hsiao-RongTyan, Chia-Wen Lin,Duan-Yu Chen and Kuo-Chin Fan ,"Motion Flow-Based Video Retrieval", IEEE Transactions On Multimedia, Vol. 9, No. 6, October 2007

[18] ZakaryaDroueche, Mathieu Lamard, Guy Cazuguel, Gw´enol´eQuellec, Christian Roux and B´eatriceCochener, " Motion-based Video Retrieval with Application to Computer-Assisted Retinal Surgery", Proc. 34th Annual International Conference of the IEEE EMBS, San Diego, California USA, 28 August - 1 September, 2012

[19] Di Zhong, Shih-Fu Chang, "An Integrated Approach for Content-Based Video Object Segmentation and Retrieval", IEEE transaction on Circuitsand Systems for Video Technology, vol. 9, no. 8, 1999

[20] Camilo C. Dorea, Montse Pardas, Ferran Marques, "Trajectory Tree as an Object-Oriented Hierarchical Representation for Videos", IEEE Transactions On Circuits and Systems forVideo Technology, vol. 19, no. 4, 2009

[21]GiovaniGualdi, Andrea Prati, Rita Cucchiara, "Multistage Particle Windows for Fast and Accurate Object Detection", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 34, no. 8, 2012

[22] ArasanathanAnjulan and NishanCanagarajah, "A Unified framework for Object Retrieval and Mining", IEEE Transaction on Circuits and Systems for Video Technology, vol 19, no.1, 2009

[23]Haojie Li, Jinhui Tang, Si Wu, Yongdong Zhang, Shouxun Lin,"Automatic Detection and Analysis of Player Action in Moving Background Sports Video Sequences", IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 3, 2010

[24]Xingxiao Wu, Dong Xu, LixinDuan, JieboLuo, "Action Recognition Using MultilevelFeatures and Latent Structural SVM", IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 8, 2013

[25] Chungfeng Yuan, Xi Li, Weiming Hu, Haibin Ling, Stephen J. Maybank, "Modeling Geometric-Temporal Context With Directional Pyramid Co-Occurence for Action Recognition", IEEE Transactions on Image Processing, vol. 23, no.1, 2014

[26] Jianzhai Wu, Dewwn Hu, "Learning Effective Event Models to Recognize a Large Number of Human Actions", IEEE Transactions on Multimedia, vol. 14, no.1, 2014

[27]Yingying Zhu, Nandita M. Nayak, Amit K. Roy-Chawdhury,"Context-Aware Activity Recognition and Anomaly Detection in Video", IEEE Journal Of Selected TopicsIn Signal Processing, vol. 7, no. 1, 2013

[28] MyoThida, How-lung Eng, Paolo Remagnino, "LaplacianEigenmapWith Temporal Constraints for Local Abnormality Detection in Crowded Scenes", IEEE Transaction on Cybernetics, vol. 43, no. 6, 2013

[29]Xiang Ma, Faisal bashir, Ashfaq A. Khokhar, Dan Schonfeld, "Event Analysis Based on Multiple Interactive Motion Trajectories",IEEE Transactions On Circits and Systems for Video Technology, vol. 19, no. 3, 2009

[30] Michele Merler, Bert Huang,LexingXie, Gang Hua, ApostolNatsev, Semantic Model Vectors for Complex Video Event Recognition", IEEE Transactions on Multimedia, vol. 14, no.1, 2012

[31]Xiofeng Wang, Xiao-Ping Zhang, "An ICA Mixture Hidden Conditional random Field Model for Video Event Classification", IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 1, 2013

Dr. H. B. Kekre received B.E. (Hons.) in Telecomm. Engg. from Jabalpur University in 1958, M.Tech (Industrial Electronics) from IIT Bombay in 1960, M.S.Engg. (Electrical Engg.) from University of Ottawa in 1965 and Ph.D. (System Identification) from IIT Bombay in 1970. He worked over 35 years as Faculty of Electrical Engineering and then HOD Computer Science and Engg. at IIT Bombay. After serving IIT for 35 years, he retired in 1995. After retirement from IIT, for 13 years he was working as a professor and head in the department of computer engineering and Vice principal at Thadomal Shahani Engg. College, Mumbai and as a senior professor at MPSTME, SVKM's NMIMS University. He guided 17 Ph.Ds., more than 100 M.E./M.Tech and several B.E. / B.Tech projects, while in IIT and TSEC. His areas of interest were Digital Signal processing, Image Processing and Computer Networking. He had more than 450 papers in

National / International Journals and Conferences to his credit. He was Senior Member of IEEE. He was Fellow of IETE, Member of ISTE and Senior Member of International Association of Computer Science and Information Technology (IACSIT).

Dr. Dhirendra S. Mishra has received his BE and ME degrees (Computer Engineering) from University of Mumbai and his PhD. degree from NMIMS University, India. He is working as an associate professor and Head of Computer Engineering Department at NMIMS University. His area of interest is Image Processing and he has more than 40 research papers published in National / International Journals. He is the reviewer for many journals.

Ms. Pallavi R. Rege has received B.E .(Computer) degree with distinction from Birla Institute of Technology, Mesra, Ranchi , in 1993 and M.E. (Computer Science and Engineering) from Bharati University (Deemed), Pune in 2009. She is currently pursuing her PhD from SVKM's NMIMS University, Mumbai, India. She has nearly 15 years of experience in teaching, research and industry. She is currently working as Assistant Professor in Computer Engineering at Vishwakarma Institute of Information technology, Pune., India.