International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Content Based Movie Scene Retrieval using Spatio-Temporal Features

Vidit Kumar, Vikas Tripathi, Bhaskar Pant

Abstract: ***Thousands of movies along with TV shows, documentaries are being produced each year around the world with different genres and languages. Making a movie scene impactful as well as original is challenging task for the director. On the other hand, users demands to retrieve similar scenes from their queries is also challenging task as there is no proper maintenance of database of movie scene videos with proper semantic tags associated with it. So to fulfill the requirement of these two (but not the least) application areas there is a need of content based retrieval system for movie scenes. Content based video retrieval is a problem of retrieving most similar videos to a given query video by analyzing the visual contents of videos. Traditional video level features based on key frame level hand engineered features which does not exploit rich dynamics present in the video. In this paper we propose a Content based Movie Scene Retrieval (CB-MSR) framework using spatio-temporal features learned by deep learning. Specifically deep CNN along with LSTM is deploy to learn spatio-temporal representations of video. On the basis of these learned features similar movie scenes can be retrieve from the collection of movies. Hollywood2 dataset is used to test the proposed system. Two types of features: spatial and spatio-temporal features are used to evaluate the proposed framework.***

Keywords: CNN, LSTM, CB-MSR, Deep learning.

I. INTRODUCTION

Thousands of movies, television shows and short documentaries in different genres and languages are produced every year all over the world. Due to the advancement in technologies in media and entertainment industry the number of movies producing each year is increasing rapidly. Making film is challenging task for a producer as well as for a director. After film script or screenplay is written and ready for movie, director plays key element of film making since the director is the one who visualize the script and controls the film from the beginning to the end while guiding technical crew and actors throughout the film making [1]. All the scenes of movies or TV shows are shots with director's vision to the movie by having dynamic script i.e. if required the director can change some points in script according to the demands of a scene. Though it is challenging for a director to ensure the scene is original, powerful or whether it will make an impact to the audience; with the help of some similar past movie scenes the director can make more powerful scene. As there is no proper maintenance of previous movie scenes as a database (since they are part of whole movie) it is a time consuming and laborious work to search the past movie scenes from whole movie one by one to ensure the originality of present movie scene. Therefore, there is need of content based movie scene retrieval system to overcome this issue by automatically search for required scenes from the past movies. On the other hand, from the user's point of view, when they want to watch a similar movie scene they can query a movie scene as query by example to retrieve similar scenes. In this paper, we proposed a Content based Movie Scene Retrieval (CB-MSR) by employing CNN and LSTM for spatio-temporal representation of movie scene. The spatial features of scene videos are extracted by CNN, and dynamic descriptors are built with LSTM network. To obtain scene-level representation time series pooling operation is done to pool the frame-level activations. So, our method exploits both the spatial and the temporal dynamics to build a movie-scene level representation. And finally on the basis of learned features similar movie scenes are retrieved from the database of movies.

This paper is presented as: literature review is presented in Section II. Section III explains the proposed framework in detail. Experimental settings are discussed in Section IV. Section V discusses the Results and discussion. Finally, Section VI summarizes the conclusion with future research directions.

II. LITERATURE REVIEW

In [2] proposes a movie genre classification method based on visual cues like color variance and motion content. In [3] video recommender system is proposed which employ fusion of high-level and low-level video features (such as textual, visual and auditory) along with feedback for further improvement. In [4] framework for automatic classification of Movie genre is proposed which uses the bag of visual word model with key frame based features. Wang Fangshi et al. [5] proposed feature extraction method based on features differences between adjacent two frames. Deldjoo et al. [6] proposed a content-based recommender system which automatically analyze the contents of video and extracts visual features as a set of visual cues (lighting, color, and motion). Simões et al. [7] use CNN for movie genre classification. Deldjoo et al [8] proposed a movie recommendation system based on MPEG-7 visual descriptors and Deep networks activations. Rimaz et al.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

Vidit Kumar, Department of Computer Science and Engineering, Graphic Era deemed to be university, Dehradun, India. Email: viditkumaruit@gmail.com

Vikas Tripathi, Department of Computer Science and Engineering, Graphic Era deemed to be university, Dehradun, India. Email: vikastripathi.be@gmail.com

Bhaskar Pant, Department of Computer Science and Engineering, Graphic Era deemed to be university, Dehradun, India. Email: pantbhaskar2@gmail.com

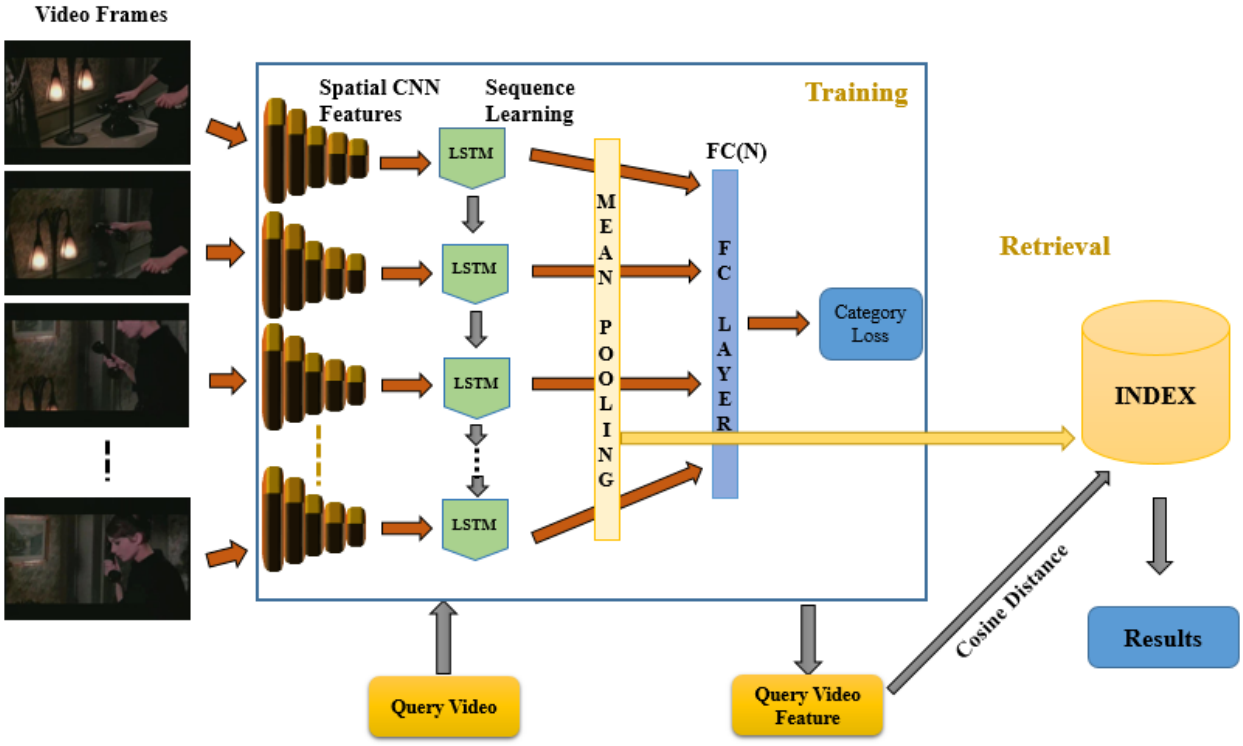

Figure 1: Overview of the proposed CB-MSR system.

In [9] visual features like color variance, Standard deviation of color variance, mean of motion across frames are employed in recommendation of movies. Srivastava et al. [10] use motion image for action representation in video.

III. PROPOSED FRAMEWORK

Figure 1 illustrates the training and retrieval process of proposed CB-MSR framework. CB-MSR system is divided into three core components: First, T CNNs takes a T individual frames of a movie scene to encode spatial information to rich features. Second, output of first component as sequence of frames level features representing the action in scene are input to the LSTM (with 256 hidden units) to learn temporal dynamics in the movie scene. Third, mean pooling is done over the output of second component the frame level LSTM responses to represent whole scene. All components of CB-MSR are described in detail in the following subsections.

A. Convolutional Neural Network (CNN)

The first component of our method is CNN. CNN is a neural network origins from the field of deep learning having large number of neurons for data processing. It is design to process 2d array data such as images or 3d volumetric data such as video. It consist of large number of hidden layers namely convolutional layers. Having properties of weight sharing and local connection, convolutional layer analyzes the spatial structure of images in layer by layer. Deep CNN requires large scale database of images to train and when it is trained, it learns to extract powerful rich features in the images or videos. In this paper we use the imagenet pretrained model of GoogLeNet [11] for spatial representation of video frames. It takes color image of 224 × 224 size as input and pass through multiple inception modules to average pool layer followed by fully connected layer and softmax. To represent frames the output of average pooling layer in the GoogLeNet as a feature descriptor of 1024 dimensional is extracted.

For each video V_i , sequence of frame features S_i are generated as follows:

$$S_i = f_{CNN}\left(V_i^{(t)}\right) \quad (1)$$

where, t = 1,... n_i (number of frames in i^{th} video)

B. Long Short Term Memory (LSTM)

Recurrent neural networks (RNN) are a class of neural networks in which current network's state is dependent to previous state i.e. output of previous time step is fed as input to the network along with input data to learn temporal dynamics present in the data as shown in figure 2.

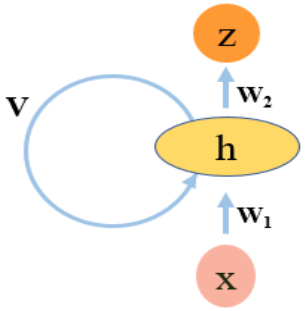

Figure 2: Basic RNN model

RNN has limitation in training to learn continuous dynamics in long-length, which is because of vanishing and exploding gradients problem. To solve this problem LSTM is developed in [12]. With the use of memory cells in LSTM (figure 3) long-range learning can be done without any issue [12]. For a given input sequence x_t at time t, cell states and hidden states of LSTM are updated for time step t as follows:

$$i_t = \sigma\left(U_{xi}x_t + U_{hi}h_{t-1} + b_i\right) \quad (2)$$

$$f_t = \sigma\left(U_{xf}x_t + U_{hf}h_{t-1} + b_f\right) \quad (3)$$

$$g_t = \sigma\left(U_{xg}x_t + U_{hg}h_{t-1} + b_g\right) \quad (4)$$

$$o_t = \sigma\left(U_{xo}x_t + U_{ho}h_{t-1} + b_o\right) \quad (5)$$

$$c_t = \left(f_t \odot c_{t-1} + i_t \odot g_t\right) \quad (6)$$

$$h_t = o_t \odot \phi\left(c_t\right) \quad (7)$$

where U is a weight matrix, b is bias, σ is sigmoid activation function, ϕ is hyperbolic tangent function, $\odot$ denotes element-wise multiplication and i, f, g, o, c are input gate, forget gate, input modulation gate, output gate, cell (memory) activation vectors respectively.

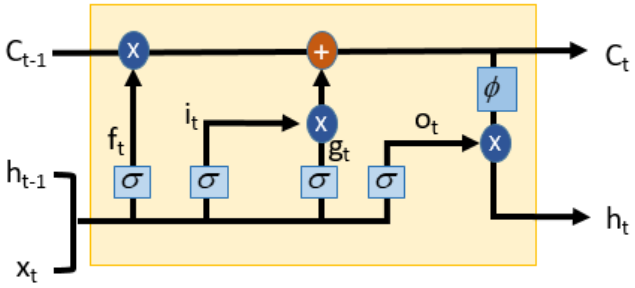

Figure 3: LSTM module

The hidden states $h_i^{(t)}$ of LSTM are generated by feeding S_i (1) into LSTM as single time step at a time using (7).

Then set of LSTM's fame level responses (8) of video is input to the third component of CB-MSR i.e. time series pooling layer discussed in next subsection.

$$H_i = \left\{h_i^{(1)}, \ldots, h_i^{(t-1)}, h_i^{(t)}\right\} \quad (8)$$

C. Time Series Pooling

Having set of frames level features H_i (8) one way to represent video with these features is to simply concatenate to form $R^{n\times d}$ vector. Computing cosine distance in $R^{n\times d}$ space is computationally expensive. So another technique is pooling to pool frame level features into same dimensional space R^d. So to represent video with a d dimensional feature vector the mean pooling is done over frame level features by mean-pooling layer in CB-MSR system as:

$$Zmean_i = f_{mean}(H_i) \quad (9)$$

where f_{mean} is mean pooling function.

IV. EXPERIMENTS

A. Dataset and Setting

To evaluate the proposed CB-MSR system we choose hollywood2 dataset [13]. It consist of 1,707 clean videos collected from 69 movies into 12 categories in which 33 and 36 movies are used to create training set and testing set respectively. Training and testing set contains clips from different movies i.e. training and testing set does not share common movies.

All the experiments are performed using Matlab 2019a's deep learning toolbox along NVidia tesla k40c gpu for training our system.

B. Dataset Preprocessing

Originally there are 1707 videos which are divided into two sets: training set consist of 823 videos and testing set consists of 884 videos. Since some videos in training set belongs to multi class-label therefore we remove all multi class-label videos from training set and retain to 750 videos which has single class-label as shown in table I. Now these set of 750 videos are used to train the proposed network. For retrieval process we use original training set as a gallery and testing set as queries as shown in table I.

C. Implementation Details

Training stage: First each video from the training set is divided into 16 continuous frames clips with 50% overlap. Then a video clip is randomly selected and resized to 256×256×3×16 followed by randomly cropped to 224×224×3×16 which is then inputs to the network. To update the network parameters Adam optimizer [14] is used along with minibatch size of 32. Learning rate is set to .0001 for entire training.

Retrieval stage: All gallery and query videos are first represented as 256 dimensional descriptor using (9). Retrieval of videos is done by measuring cosine distance between query video and gallery videos.

V. EVALUATION MEASURES AND RESULTS

Two types of features i.e. spatial features and spatio-temporal features are used for comparison purpose as well as to evaluate the efficiency of proposed CB-MSR system. For spatial representation of movie scene the average of frame level CNN features (1) is computed and spatio-temporal features of movie scene are computed using (9).

For evaluation of retrieval performance we adopted Mean Average Precision at top K retrieval i.e. MAP@K. Given a query q, average precision is computed as:

$$AP(q) = \frac{1}{G_q}\sum_{k=1}^{R} P_q(k)\,\theta_q(k) \quad (10)$$

where G_q is number of similar videos in the retrieved set for a query q. $P_q(k)$ is the precision

Table I. Training and Retrieval set

Category	Modified Training Set	Gallery set	Query Set
Answer Phone	60	66	64
Drive Car	83	85	102
Eat	35	40	33
Fight Person	42	54	70
Get Out Car	46	51	57
Handshake	24	32	45
Hug Person	34	64	66
Kiss	86	114	103
Run	120	135	141
Sitdown	92	104	108
Situp	21	24	37
Standup	107	132	146
Total	**750**	**823**	**884**

Table II. MAP@K on Hollywood2 dataset

K	Spatial (CNN) Features	Spatio-Temporal (LSTM) Features
10	0.0933	0.3772
20	0.0727	0.3655
30	0.0628	0.3582
40	0.0567	0.3519
50	0.0527	0.3459
60	0.0504	0.3430
70	0.0490	0.3428
80	0.0481	0.3439
90	0.0477	0.3456
100	0.0478	0.3466

Table III: Category wise MAP@K when using spatial features (CNN)

K	Answer Phone	Drive Car	Eat	Fight Person	GetOut Car	Hand-Shake	Hug-Person	Kiss	Run	SitDown	SitUp	Stand-Up
10	0.1121	0.1315	0.0278	0.0372	0.0898	0.0439	0.0355	0.1273	0.1089	0.0776	0.0193	0.1339
20	0.0781	0.1061	0.0192	0.0248	0.0644	0.0303	0.0271	0.0913	0.0891	0.0698	0.0139	0.1069
30	0.0667	0.0915	0.0164	0.0195	0.0530	0.0240	0.0232	0.0757	0.0787	0.0627	0.0136	0.0939
40	0.0584	0.0812	0.0138	0.0170	0.0465	0.0257	0.0206	0.0676	0.0721	0.0579	0.0149	0.0843
50	0.0531	0.0738	0.0152	0.0152	0.0423	0.0287	0.0194	0.0623	0.0671	0.0542	0.0158	0.0779
60	0.0488	0.0687	0.0165	0.0158	0.0451	0.0312	0.0184	0.0591	0.0634	0.0522	0.0179	0.0734
70	0.0483	0.0642	0.0179	0.0173	0.0480	0.0338	0.0193	0.0558	0.0605	0.0506	0.0188	0.0699
80	0.0514	0.0595	0.0189	0.0187	0.0516	0.0370	0.0212	0.0528	0.0585	0.0488	0.0201	0.0670
90	0.0538	0.0591	0.0200	0.0202	0.0543	0.0386	0.0230	0.0509	0.0563	0.0475	0.0207	0.0644
100	0.0568	0.0625	0.0207	0.0216	0.0568	0.0411	0.0245	0.0491	0.0543	0.0462	0.0218	0.0621

Table IV: Category wise MAP@K when using spatio-temporal features (LSTM)

K	Answer Phone	Drive Car	Eat	Fight Person	GetOut Car	Hand-Shake	Hug-Person	Kiss	Run	SitDown	SitUp	Stand-Up
10	0.2697	0.7624	0.1730	0.3393	0.3061	0.0908	0.3669	0.3630	0.6430	0.2636	0.1088	0.2459
20	0.2686	0.7611	0.1637	0.3052	0.2835	0.0784	0.3590	0.3469	0.6407	0.2621	0.0959	0.2218
30	0.2671	0.7587	0.1487	0.2837	0.2672	0.0630	0.3487	0.3388	0.6418	0.2614	0.0900	0.2114
40	0.2653	0.7574	0.1300	0.2558	0.2529	0.0617	0.3286	0.3348	0.6431	0.2595	0.0947	0.2057
50	0.2608	0.7563	0.1411	0.2315	0.2343	0.0633	0.2872	0.3350	0.6448	0.2584	0.1005	0.2011
60	0.2529	0.7532	0.1472	0.2303	0.2396	0.0638	0.2526	0.3365	0.6456	0.2588	0.1029	0.1970
70	0.2491	0.7499	0.1487	0.2406	0.2448	0.0647	0.2467	0.3379	0.6481	0.2564	0.1037	0.1929
80	0.2580	0.7425	0.1507	0.2501	0.2497	0.0658	0.2547	0.3372	0.6502	0.2540	0.1054	0.1905
90	0.2654	0.7430	0.1515	0.2582	0.2565	0.0672	0.2620	0.3374	0.6525	0.2514	0.1067	0.1866
100	0.2692	0.7505	0.1536	0.2645	0.2665	0.0682	0.2680	0.3319	0.6528	0.2462	0.1085	0.1834

of the top k retrieved and θ is an sign function such that, $\theta_q(k) = 1$ if the item at rank k is a relevant, else 0.

Mean average precision for all queries Q is computed as follows:

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \qquad (11)$$

MAP@K for whole dataset for spatial as well as spatio-temporal features is depicted in table II and visualized in figure 4. Also scene categorical wise MAP@K for spatial features and spatio-temoral features is depicted in table III and table IV respectively. From table II, III and IV it is clear that LSTM features performs better than CNN features. This is because LSTM models the temporal variation with use of spatial information whereas CNN concentrates on frame's spatial region and ignores the temporal dynamics.

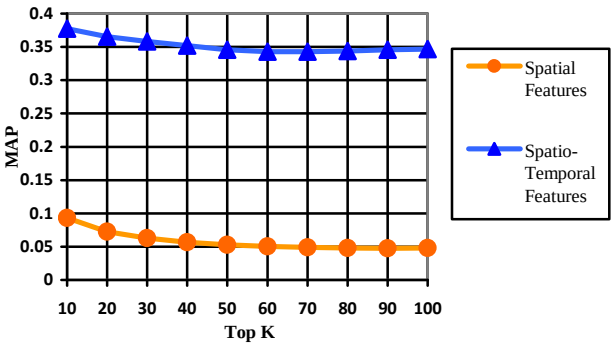

Figure 4: MAP@K curve

VI. CONCLUSION AND FUTURE WORK

This paper presents Content based movie scene retrieval (CB-MSR) framework via deep learning. Spatio-temporal dynamics is learned by LSTM to generate a descriptor to represent movie scene. Retrieval of similar movie scenes is done by using cosine distance. Use of optical flow or other motion information, hashing along with training in large dataset for further improvement of CB-MSR system are the future works.

REFERENCES

1. Mackendrick, Alexander, and Paul Cronin. "On film-making: an introduction to the craft of the director." Cinéaste 30, no. 3 (2005): 46-54.
2. Rasheed, Zeeshan, Yaser Sheikh, and Mubarak Shah. "On the use of computable features for film classification." IEEE Transactions on Circuits and Systems for Video Technology 15, no. 1 (2005): 52-64.
3. Yang, Bo, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Mingjing Li. "Online video recommendation based on multimodal fusion and relevance feedback." In Proceedings of the 6th ACM international conference on Image and video retrieval, pp. 73-80. ACM, 2007.
4. Zhou, Howard, Tucker Hermans, Asmita V. Karandikar, and James M. Rehg. "Movie genre classification via scene categorization." In Proceedings of the 18th ACM international conference on Multimedia, pp. 747-750. ACM, 2010.
5. Fangshi, Wang, Xu De, and Wu Weixin. "A Cluster Algorithm of Automatic Key Frame Extraction Based on Adaptive Threshold [J]." Journal of Computer Research and Development 10 (2005).
6. Deldjoo, Yashar, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. "Content-based video recommendation system based on stylistic visual features." Journal on Data Semantics 5, no. 2 (2016): 99-113.
7. Simões, Gabriel S., Jônatas Wehrmann, Rodrigo C. Barros, and Duncan D. Ruiz. "Movie genre classification with convolutional neural networks." In 2016 International Joint Conference on Neural Networks (IJCNN), pp. 259-266. IEEE, 2016.
8. Deldjoo, Yashar, Mehdi Elahi, Massimo Quadrana, and Paolo Cremonesi. "Using visual features based on MPEG-7 and deep learning for movie recommendation." International journal of multimedia information retrieval 7, no. 4 (2018): 207-219.
9. Rimaz, Mohammad Hossein, Mehdi Elahi, Farshad Bakhshandegan Moghadam, Christoph Trattner, and Reza Hosseini. "Exploring the Power of Visual Features for the Recommendation of Movies." In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, pp. 303-308. ACM, 2019.
10. Srivastava, Awadhesh Kr, K. K. Biswas, and Vikas Tripathi. "A Robust Framework for Effective Human Activity Analysis." In International Conference on Innovative Computing and Communications, pp. 331-337. Springer, Singapore, 2019.
11. Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.
12. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9, no. 8 (1997): 1735-1780.
13. Marszałek, Marcin, Ivan Laptev, and Cordelia Schmid. "Actions in context." In CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition, pp. 2929-2936. IEEE Computer Society, 2009.
14. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

AUTHORS PROFILE

Mr. Vidit Kumar has done B.Tech in Computer Science and Engineering from Uttaranchal University, M.Tech from Graphic Era deemed to be university, and currently pursuing PhD in Computer Science and Engineering from Graphic Era deemed to be university, Dehradun, India. His research of interest is in Machine Learning, Deep Learning, Video Analytics and Computer Vision.

Dr. Vikas Tripathi has done BE in information technology from Technocrats institute of technology, Bhopal, M. Tech in Software engineering from Indian institute of information technology Gwalior and PhD from Uttarakhand technical university, Dehradun. He is actively involved in research related to Software engineering, Computer Vision, Machine learning and Video Analytics. He has published many papers in reputed international conferences and journals. Currently he is working as an associate professor in Graphic era deemed to be university Dehradun, India.

Dr. Bhasker Pant Currently working as Dean Research & Development and Associate Professor in Department of Computer Science and Engineering. He is Ph.D. in Machine Learning and Bioinformatics from MANIT, Bhopal. Has more than 15 years of experience in Research and Academics. He has till now guided as Supervisor 3 Ph.D. candidates (Awarded).and 5 candidates are in advance state of work. He has also guided 28 MTech. Students for dissertation. He has also supervised 2 foreign students for internship. Dr.Bhasker Pant has more than 70 research publication in National and international Journals. He has also chaired a session in Robust Classification & Predictive Modelling for classification held at Huangshi, China.