Multi-level Video Captioning based on Label Classification using Machine Learning Techniques

J. Vaishnavi, Dr. V. Narmatha
Department of Computer and Information Science
Faculty of Science, Annamalai University, Annamalai Nagar, India

Abstract—**Video captioning is the heuristic and most essential task in the current world to save time by converting long and highly content-rich videos into simple and readable reports in text form. It is narrating the events happening in videos in natural language sentences. It makes the way to many more interesting tasks by the use of labels, tags, and terms such as video content retrieval, video search, video tagging, etc. Video captioning is currently being attempted by many researchers using some exciting Deep learning techniques. But this approach is to find the best of machine learning for the process of captioning videos in a different way. The novel part of the proposed approach is classifying videos by using the labels existing in video frames that belong to the various categories and producing consecutive Multi-Level captions that describe the entire video in a round-robin way. Informative features are extracted from the video frames such as Gray Level Co-occurrence Matrix (GLCM) features, Hu moments, and Statistical features to provide optimal results. This model is designed with two superior and optimal classifiers such as Support Vector Machine (SVM) and Naive Bayes separately. The models are demonstrated with the prevailing standard dataset Microsoft Research Video Description corpus (MSVD) and evaluated by the benchmark classification metrics such as Accuracy, Precision, Recall, and F1-Score.**

Keywords—Video captioning; label classification; Hu moments; GLCM; statistical features; SVM; Naive Bayes

I. INTRODUCTION

Video is the most used and needed multimedia which is preferred over images and text. Social media influences and increased the usage of videos. It plays an essential role in the everyday lives of people. Video is the combination of audio and scenes which narrates with a lively touch. Hearing-impaired people suffer to extract the complete information from the video. Video captioning is the current attention-seeking computer vision [1] task which is focused on by many budding researchers. Video captioning solves the issue by narrating the video in human-understandable language. It conquers almost every field, especially from education to entertainment. The vast application of video captioning is increasing every day with prime applications such as content summarization, human-robot interaction, reports, tagging, classification, video indexing, and video surveillance [2].

Video captioning is the process of describing the frames of the input video with natural sentences. It is also a task with some complexity such as, it must capture every frame of the input video and extracting the essential features by which the frames are classified with labels to provide accurate captions. The video captioning task has more different views and possibilities to generate captions. Machine learning techniques contribute to various methods in this field. Here is the attempt to attain the best of machine learning classifiers to classify the frames along with captions based on the labels which are categorized into different categories. Firstly, videos are converted into frames and the input frames are resized by 200×200 as constant. Some essential edge, shape and texture features are extracted from processed frames. Two standard multi-class machine learning classifiers are utilized to classify the frames with appropriate captions.

The proposed model is designed to generate the captions for the videos with less time complexity and high accuracy and also to create captions for each input video frame with particular timestamps. It will be utilized in the crime branch and hearing-impaired people will learn about the happenings of the video fruitfully. Video captioning is approached by various researchers in various methods which include different techniques and cost-effective machines. But this approach is to provide a different dimension of view towards video captioning. The proposed approach is introduced to overcome the issue of utilizing cost-effective resources like GPU (Graphical Processing Unit). The proposed method is constructed with superior machine learning techniques for both feature extraction and classification and the comparative results are analyzed to declare the best model.

The article is constructed based on the following sections. Section II discusses the related works of video captioning. The proposed method is clearly explained with the needed equation and structure in Section III. The preprocessing step is discussed in sub-section IIIC. The elaborate information about the utilized benchmark dataset is given in Section IV. Section V shows the comparative study of two standard classifiers with their computed results. The article is concluded in Section VI.

II. RELATED STUDY

The field of video captioning is evolved with different techniques which various researchers develop. Still, it is considered a challenging task due to its complexities. It includes some sub-tasks such as event detection, localization, object classification, etc. The model discussed the event detection from the high-content sports video. The events are detected by using audio-visual features and classified from the multiple genres of sports video with the employment of the standard Machine learning classifier such as the Support vector machine. Video captioning is performed by utilizing the template-based techniques [3, 4, 5] for the generation of captions for the videos in the earlier stages.

Dynamic captioning model is introduced to show the variation of the speech signal based on the volume of the audio. It positions the caption which indicates the speaker for a better understanding of hard-of-hearing people. It explores various techniques for the betterment of the model to provide satisfactory captioning, especially for hearing impaired people such as face detection and recognition, speech–text correspondences and visual alignment, etc. The model is tested in real-time with 20 video clips of 60 hearing-impaired people. Video captioning is the extension of an image captioning task. Image captioning is the task of illustrating the content of the image. Graph-based automatic captioning of the image is proposed in which is superior to other methods such as LDA, HAM, and EM with the advantages of not defining the parameters but initializing values for only two constant parameters. This model is demonstrated with the dataset Corel image database. Video is the most preferred media by people when compared to other media. It has various choices. More types of videos are available nowadays effortlessly. The need for an automatic choice selection of videos is essential to avoid wasting energy and conserve time. To overcome the issue, the preferred choice of video for the viewers is automated by the model proposed by the researchers. It is modeled by extracting the fused features of visuals and closed captions with the Hidden Markov model.

In the era of video captioning, traditional machine learning techniques are utilized in the foremost phase of video captioning evolution such as the Hidden Markov model and post-action grammar techniques [6] with various features such as hand-crafted features and object-centric features [7], etc. Videos are captioned easily by initially classifying them where it belongs to and categorized based on the content of the video [8, 9] and activities [10] that exist in the video. Visual features [11, 12, 13] are considered one of the essential features extracted to generate captions for the video. The embedding spaces are constructed between the input videos and the natural language sentences [14, 15].

Visual grounding [16, 22] is also one of the major video-related tasks which utilize visual reasoning. Hidden features [17] are extracted by using pos sequence features. A gated fusion network [18] is utilized for captioning videos. Image captioning techniques are highly functional in advanced techniques in the current time which motivate many tasks such as video captioning [19]. The superior models are designed with an attention mechanism [20] to caption the videos for boosting the performance and also by utilizing convolutions in both encoding and decoding phases. One of the major differences between image captioning and video captioning is extracting the temporal information [21] which is essential to generating captions for videos. Various features are extracted for captioning videos such as visual, object, spatial, etc. Audio features [23] are also extracted for captioning as multi-model features [24] along with speech features. Additional memory modules [25] are exploited to evade coherent captions for the video. Videos are pre-processed to generate appropriate captions in which temporal segmentation [26] in a video between various events is essential in video captioning tasks. The memory attended recurrent network (MARN) model [27] aims to match the visuals and term that describes the visual. Visual reasoning is adopted in [28] Reasoning module network (RMN) for location and time. Transformers are utilized for the caption generation process. An accelerated masked transformer [29] is utilized in the decoding phase which generates captions especially with localizing tasks. The videos are categorized based on domains designed as domain-specific decoders [30]. Image captioning is also the one of the cause of video captioning. It makes the video captioning processes easier. Image captioning process is taken as the subsidiary content for video captioning process to enlarge the diversity [31]. Activity net captions and Microsoft coco image datasets are utilized for the captioning process which actually enlarges the diversity.

III. Proposed Approach

In the Proposed Model, the videos are categorized into various categories which are based on labels and generate Multi-Level natural language sentences based on the extracted features. The Multi-Level captions are generated as it describes the content of the video with 7 to 10 sentences successively. The videos are illustrated in a nutshell that truly meets the need of the model as saves time than spending more watching and understanding videos. The process of providing captions in this model consists of two vital parts Feature extraction and classification. The essential features are extracted such as Gray Level Co-occurrence Matrix (GLCM) features, Hu moments, and statistical features to detect and classify the frames. Two foremost and appropriate classifiers which work better in classifying images and videos are employed in this model such as Support Vector Machine (SVM) and Naive Bayes. Preprocessing of video frames is performed for the enrichment of informative frames to generate appropriate captions for the videos. Fig. 1 shows some of the video frames from the MSVD dataset.

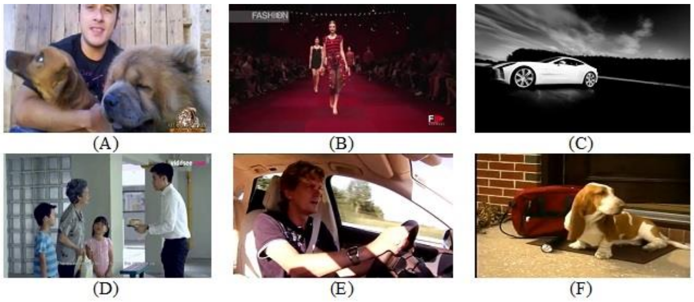


Fig. 1. Frames of different Categories of MSVD Dataset. (A). A Man Sits with Two Dogs, (B). A Girl Walks on the Parade, (C). A Car on the Road, (D). A Man is Talking with Children and an Old Lady, (E). A Man is Driving a Vehicle, (F) A Dog is Sitting Near a Bag.

A. Feature Extraction

1) GLCM Features: Gray Level Co-occurrence Matrix (GLCM) shows the various combinations of pixels based on the brightness values as gray level values in an image. GLCM also has another name "Gray Tone Spatial Dependency Matrix”. GLCM matrix is performed based on the orders of the texture calculations. Normalize GLCM is needed to get the value one as the sum of its elements. An element in GLCM after normalization defines the probability of pair of pixels which explores the gray values of spatial relationship in the image. The following steps are involved to create a Normalized GLCM Matrix. Fig. 2 shows the extracted features for the further classification.

a) Arrange and quantize the specific parameters (make the intensities of the pixels arranged in a needed number of gray levels) in an image data.

b) Create the square matrix of GLCM with specific order N×N, where N denotes the Number of Levels.

c) A symmetric matrix is introduced here with GLCM Matrix.

d) Dividing each element to normalize GLCM Matrix by the sum of all elements.

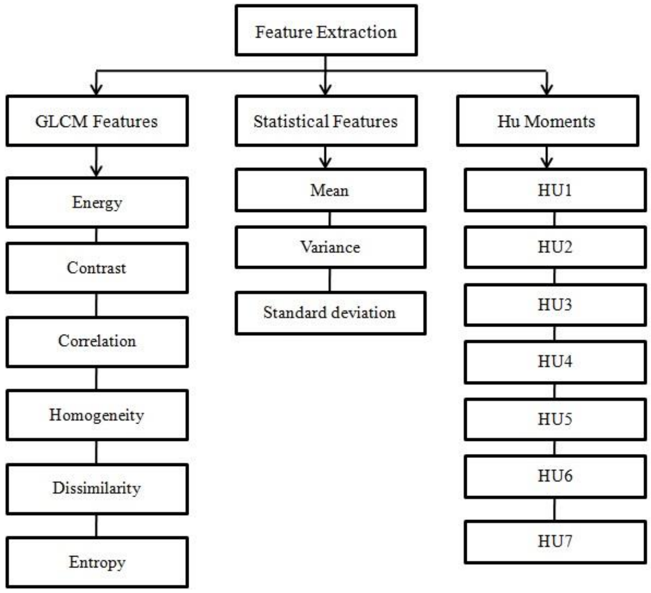


Fig. 2. Features Extracted for Classification.

Here, six different GLCM features are extracted for the betterment of predicting exact captions for the input videos such as Energy (Angular second moment), Contrast, Correlation, Dissimilarity, Entropy, and Homogeneity.

2) Energy: It is the sum value of the squared elements in the matrix of the occurrence of gray. It is also named Angular Second Moment (ASM) and Uniformity. It ranges the values from zero to one. The constant image has one as the energy value. Eq. (1) explains the calculation of the energy.

$$Energy = \sum_{x,y=0}^{N-1} R(x,y)^2 \quad (1)$$

3) Contrast: It mainly considers the intensity between the pixels that existed in the entire image. Here frames are considered images. So, the intensity between one pixel and the nearby pixel is calculated to get the contrast value. It is also named variance and inertia. The intensity contrast is measured for the entire pixels with the neighboring pixel of an image. The contrast value ranges from zero as per the constant image has the value zero as its contrast, the range of contrast value is from zero to (size (GLCM, 1) -1) two. Contrast is calculated using Eq. (2).

$$Contrast = \sum_x \sum_y |x-y|^2 R(x,y) \quad (2)$$

4) Correlation: Correlation is the estimation of how each pixel is related to the other pixel. It calculates the relation between the neighborhood pixels in the entire image. It values one for the positivity-related pixel and -1 for the negatively related pixel of an image and the value zero is assigned for a constant image. Correlation ranges the values between -1 and +1. By using Eq. (3), correlation is calculated.

$$Correlation = \sum_x \sum_y \frac{\left((x-\mu_x)(y-\mu_y)R(x,y)\right)}{\sigma_x \sigma_y} \quad (3)$$

5) Homogeneity: Homogeneity is the evaluation of similarity and nearness of the distribution of various elements in the Gray Level Co-occurrence matrix to its diagonals. It calculates the similarity between pixels of an image. It compares a pixel with the neighbor one. It ranges the values between zero and one. The diagonal of GLCM has one as the homogeneity value. Homogeneity is calculated between two pixels are calculated by using the Eq. (4).

$$Homogeneity = \sum_x \sum_y \frac{1}{1+|x-y|^2)} R(x,y) \quad (4)$$

6) Entropy: It is the average image information that calculates the amount of randomness in an image. It is measured based on the position of the pixel in a region (x, y). The Entropy of an image is computed by measuring the Entropy of the pixel values inside the two-dimensional region centered at each pixel position (x, y). By using equation (6), entropy is calculated. Table I shows the values of various GLCM features for sample frames. By using Eq. (5), entropy is calculated.

$$Entropy = \sum_x \sum_y R(x,y) \log R(x,y) \quad (5)$$

7) Dissimilarity: It concentrated on the interesting region of the frame. It requires calculating the mean absolute difference and distance between two pixels of an image. It is the calculation of local intensity variation between two pixels which is nearby one another. The larger value means the higher differences; meanwhile smaller value resembles less difference between two pixels of the region. Dissimilarity is calculated by using the Eq. (6).

$$Dissimilarity = \sum_x \sum_y |x-y| R(x,y) \quad (6)$$

TABLE I. THE GLCM VALUES FOR SAMPLE FRAMES

Frames	**Contrast**	**Correlation**	**Energy**	**Homogeneity**	**Entropy**	**Dissimilarity**
1	3.59	9.74	4.21	7.89	1.56	1.10
2	5.79	8.05	4.18	8.24	2.87	2.81
3	2.35	9.21	3.49	9.17	1.17	1.35
4	8.14	8.03	7.59	9.13	6.62	8.14
5	4.73	8.57	3.16	8.28	1.59	3.22

8) Hu Moments: Hu moments are specially for shape matching. Hu moments are the essential features to detect the object from an image or frame which is existed in any direction or orientation. It makes the model identify the object or anything that existed despite any direction. Image moments are the weighted sum of the pixel of intensities in an image. These moments take into account the fundamental measures of objects such as area, centroid, orientation, and other needed properties. The central moment of an image gives the information that is constant to translation in an image. It needs to be invariant to translation, rotation, and scale. Hu moments are able to get seven values despite any transformation of an image based on the central moments. Central moments are calculated by the equation (7).

$$\mu_{\boldsymbol{xy}} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{y}} (\boldsymbol{i} - \bar{\boldsymbol{\imath}})^{x} (\boldsymbol{j} - \bar{\boldsymbol{\jmath}})^{y} \boldsymbol{I}(\boldsymbol{i}, \boldsymbol{j}) \tag{7}$$

Here comes, Hu moments that characterize the existence in a frame. It detects the object through its shape. The strong point of Hu moments is stated that it traces the outline or edges of an object by using the shape feature vector to detect its shape. The difference between two different shapes of the object is measured by using the similarity metric. In seven moments, six moments are computed depending on the constant measures of scale, translation, reflection, and rotation. The seventh moment measures the changes in the image reflection. The seven Hu moments are calculated by the equations (8 - 14), Table II shows the values of Hu moments features for sample frames.

$$Hu_1 = H_{02} + H_{20} \tag{8}$$

$$Hu_2 = (H_{20} - H_{02})^2 + 4(H_{11})^2 \tag{9}$$

$$Hu_3 = (H_{30} - 3H_{12})^2 + 3(H_{03} - 3H_{21})^2 \tag{10}$$

$$Hu_4 = (H_{03} + H_{21})^2 + (H_{30} + H_{12})^2 \tag{11}$$

$$Hu_5 = (H_{30} - 3H_{12})(H_{30} + H_{12})[(H_{30} + H_{12})^2 - 3(H_{03} + H_{21})^2] + (3H_{21} - H_{03})(H_{03} + H_{21})[3(H_{30} + H_{12})^2 - (H_{03} + H_{21})^2] \tag{12}$$

$$Hu_6 = (H_{20} - H_{02})[(H_{30} + H_{12})^2 - 7(H_{03} + H_{21})^2] + 4H_{11}(H_{30} + H_{12})(H_{03} + H_{21}) \tag{13}$$

$$Hu_7 = (3H_{21} - H_{03})(H_{30} + H_{12})[(H_{30} + H_{12})^2 - 3(H_{03} - H_{21})^2] + (H_{30} - 3H_{12})(H_{03} + H_{21})[3(H_{30} + H_{12})^2 - (H_{03} + H_{21})^2] \tag{14}$$

9) Statistical features: Statistical features are one of the vital features which are needed to extract the content of an image for classification. It classifies based on the orders. Some of the most utilized statistical features by existing researchers are mean, median, standard deviation and variance, etc. A statistical feature depends on the texture of an image or frame. Intensity distribution information of an image is provided as the textual feature. The probability of intensity level distribution in histograms is utilized for the statistical feature calculation. Mean, Variance and standard deviation are the statistical features that are extracted from the input video frames for further purposes. Mean is the measure of the average intensity value in an image that represents the brightness of an image based on the results of the mean calculation. The image is bright if it has a high mean value otherwise the image will have a low mean value. Variance is the measure of the difference between every pixel point and the mean value, and it is estimated by determining the difference between each pixel point and the mean value and squaring the differences and then dividing the square sums by the data points. All disparities from the mean in all directions are calculated by variance. Standard deviation is the measure of the contrast of intensity values in gray. The high contrast has resulted in a high value and the low contrast has resulted in a low standard deviation value. Table III shows the values of statistical features for sample frames.

TABLE II. THE HU MOMENT VALUES FOR SAMPLE FRAMES

Frames	HU1	HU2	HU3	HU4	HU5	HU6	HU7
1	1.25	3.54	2.02	5.00	-6.13	-6.78	2.13
2	2.75	2.03	1.85	6.97	1.69	3.52	-2.63
3	2.14	1.42	1.72	2.03	-2.14	6.55	3.04
4	6.97	1.96	5.98	2.14	-6.12	1.23	1.17
5	3.72	5.89	1.53	3.56	1.96	4.56	-1.06

TABLE III. THE STATISTICAL VALUES FOR SAMPLE FRAMES

Frames	Mean	Variance	Standard deviation
1	1.00	2.15	4.85
2	6.15	1.56	4.02
3	1.11	3.58	5.99
4	2.44	0.83	0.83
5	9.47	1.74	4.01

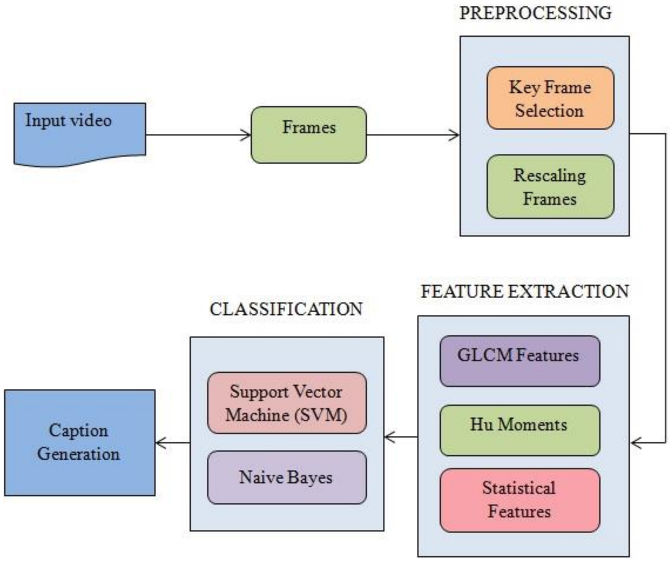


Fig. 3. Overall Architecture of the Model.

B. Classification

1) Support vector machine: Support Vector Machine (SVM) is the standard Machine learning classifier that is abundantly used for many image-related tasks which also gives high accuracy for the most classification tasks. It is a type of supervised model that splits data as training and testing. The training part is taken to learn the logic of the frames and the testing part is utilized to check whether the learned phase matches the actual frame. It works better in high-dimensional spaces. Fig. 3 explains the overall architecture of the model proposed.

2) Naïve Bayes: Naive Bayes is the most utilized multi-class classifier that can decide in less time compared to other classifiers. Probability of the existing object influences more in classifying video frames. It is the probabilistic classifier that works better in the high dimensional space. It classifies the input video frames into various categories which are categorized based on the actions. Finally, the input video is classified with the particular appropriate captions.

C. Preprocessing

Preprocessing is the foremost step that simplifies the process by reducing the computational cost of processing the entire frames of a single video. It makes the model simpler and more comfortable for the further process of generating captions. Videos are converted into various numbers of frames. Fig. 4 shows the conversion of video into frames.

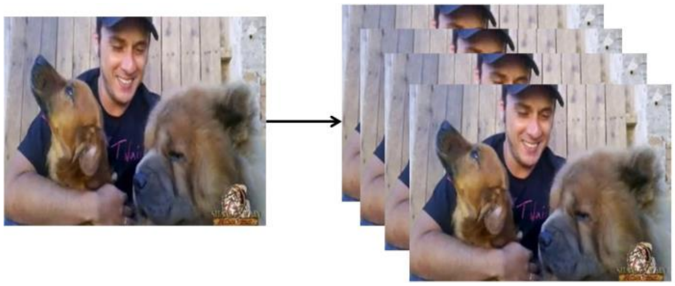


Fig. 4. Conversion of Video Into Frames.

Videos are collected from the MSVD dataset with an average duration of 10 to 25 seconds. The numbers of frames are varied for every video based on the events happening and the duration of the video. The sizes of the frames are different for extracting features. It can reduce the accuracy of predicting sentences for a particular video. To overcome the issue, the Rescaling process is acquired by changing the size of every frame of input videos to a constant size of 200×200.

Fig. 5. Rescaling of Video Frames Into Standard Size.

The size and also the number of frames are made constant. The 50 constant key frames are extracted from every video frame. This step boosted the model by speeding up the learning process. Rescaling of video frames into standard size of 200x200 is shown in Fig. 5.

IV. Dataset

The Microsoft Research Video Corpus (MSVD) dataset is the benchmark video dataset with multi-lingual captions which attracts the researchers to demonstrate their attempts at video-related tasks. MSVD Dataset is collected by AMT (Amazon Mechanical Turk) workers during the summer of 2010. It is a collection of 1970 short random YouTube video clips of a duration of 10 to 25 seconds. Each video consists of seven to ten ground truth sentences that describe the contents of the videos simultaneously. In the sum of 20000 parallel descriptions are collected for video clips with 16000 exclusive vocabularies. The data has the split as per the ratio of 80:20 for training and testing.

V. Performance Comparison

The performance of two standard techniques such as Naive Bayes and Support vector machine are evaluated using the metrics Accuracy, Precision, Recall, and F1 – score. Table IV shows the evaluation of Naive Bayes with performance metrics. Accuracy for Naive Bayes is 76.99 %. Prediction, Recall, and F1–score are scored as 72.52%, 69.89%, and 68.13% respectively. Fig. 6 shows the performances of the Naive Bayes classifier.

Table V shows the performance evaluation of the Support vector machine classifier with 80.12% accuracy, 73.85% for precision, 71.56% for recall, and 70.95 F1–score. Fig. 7 is the chart that represents the performance evaluation of the Support vector machine.

TABLE IV. Performance Evaluation of Naive Bayes

Performance Metrics	Values (%)
Accuracy	76.99
Precision	72.52
Recall	69.89
F1-Score	68.13

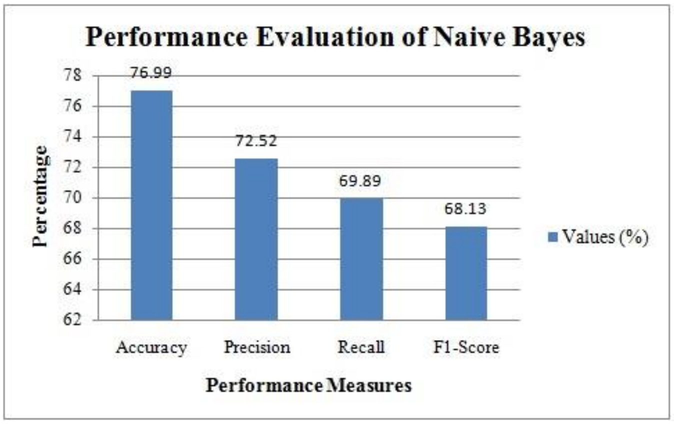


Fig. 6. Performance Evaluation of Naive Bayes.

TABLE V. PERFORMANCE EVALUATION OF SUPPORT VECTOR MACHINE

Performance Metrics	**Values (%)**
Accuracy	80.12
Precision	73.85
Recall	71.56
F1-Score	70.95

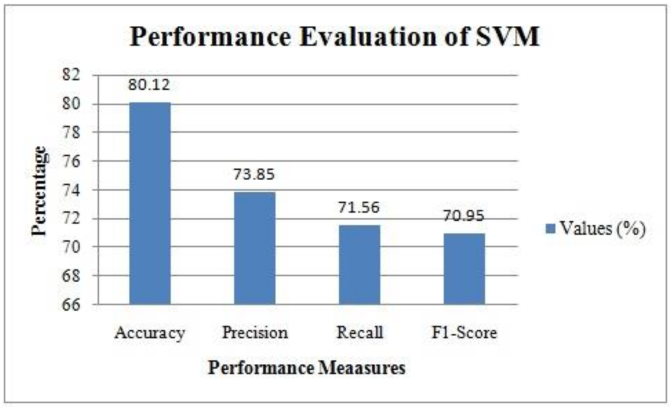


Fig. 7. Performance Evaluation of Support Vector Machine.

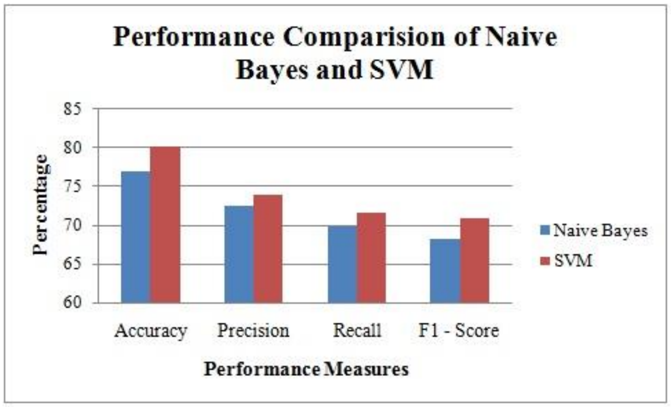
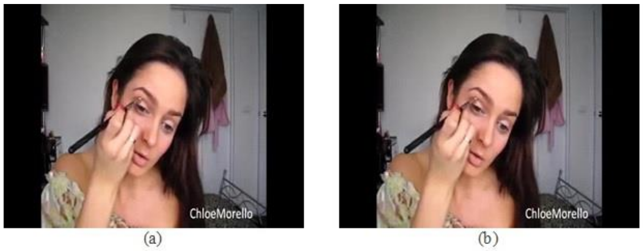


Fig. 8. Comparative Study of the Performances of Naive Bayes and Support Vector Machine.



a girl applied eye makeup to a guy

a woman is applying makeup

Fig. 9. The Result of Two Models. Fig (a). The Result of Naive Bayes, Fig (b). The Result of the Support Vector Machine.

The comparison of the performances of two classifiers such as the Support vector machine and Naive Bayes is clearly shown in Fig. 8. The chart depicts the performances of two classifiers which shows that the support vector machine outperforms Naive Bayes classifier. The results of the two models are shown in Fig. 9. The result of SVM (i.e. (A)) describes the video more precisely than Naive Bayes (i.e. (B)).

VI. CONCLUSION

Video Captioning is a fascinating task that made many researchers make an impact in the field in a unique way. This attempt is to show the different approaches to caption videos with machine learning techniques. The proposed architecture is designed based on classifying video frames with labels that are categorized into different categories and generating Multi-Level captions that explain the happenings of the videos consecutively. The model is structured with two major parts, Feature Extraction and Classification. The feature extraction part involves extracting the essential features to provide exact results such as some Gray Level Co-occurrence Matrix features, Hu moments, and statistical features. The classification process has experimented with two standard classifiers separately which show the comparative performance for classifying videos. The comparative results show that the Support Vector Machine (SVM) overperforms the Naive Bayes classifier. It also shows that the SVM is the appropriate classifier for classifying frames by using labels for other video-related tasks. It also provides the route to some other interesting video-related tasks by labels and tags. The proposed model is examined with the MSVD benchmark dataset and the classification performance is evaluated using Accuracy, Precision, Recall, and F1-Score.

REFERENCES

[1] He, X. Deng, “Deep Learning for Image-to-Text Generation: A Technical Overview,” IEEE Signal Processing Magazine, 34(6), 109–116, 2017.

[2] N. Aafaq, S. Main, W. Liu, S. Z. Gilani, M.Shah, “Video description: A survey of Methods, Datasets, and Evaluation Metrics,” ACM Computing Surveys, Vol. 52, no. 6, 2019.

[3] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, K. Saenko, “YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition,” IEEE International Conference on Computer Vision (ICCV), pp. 2712-2719, 2013.

[4] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, S. Guadarrama, “Generating Natural-Language Video Descriptions Using Text-Mined Knowledge,” AAAI Conference on Artificial Intelligence, 2013.

[5] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, B. Schiele, “Translating Video Content to Natural Language Descriptions,” IEEE International Conference on Computer Vision (ICCV), pp. 433-440, 2013.

[6] H. Pirsiavash, D. Ramanan, “Parsing videos of actions with segmental grammars,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 612– 619, 2014.

[7] L. Zhou, C. Xu, J. J. Corso, “Towards automatic learning of procedures from web instructional videos,” AAAI, 2018.

[8] A. Gaidon, Z. Harchaoui, C. Schmid, “Temporal localization of actions with actoms,” IEEE transactions on pattern analysis and machine intelligence, 35(11), 2782–2795, 2013.

[9] G. Gkioxari, J. Malik, “Finding action tubes,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 759–768, 2015.

[10] L. Wang, Y. Qiao, X. Tang, “Video action detection with relational dynamic-poselets,” European Conference on Computer Vision, pp. 565–580, Springer, 2014.

[11] M. Gygli, H. Grabner, L. Van Gool, “Video summarization by learning sub modular mixtures of objectives,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 3090–3098, 2015.

[12] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, B. Guo, “Unsupervised extraction of video highlights via robust recurrent auto-encoders,” IEEE International Conference on Computer Vision, pp. 4633–4641, 2015.

[13] T. Yao, T. Mei, Y. Rui, “Highlight detection with pairwise deep ranking for first- person video summarization,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 982–990, 2016.

[14] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkila, N. Yokoya, “Learning joint representations of videos and sentences with web image search,” European Conference on Computer Vision, pp. 651–667, Springer, 2016.

[15] R. Xu, C. Xiong, W. Chen, JJ. Corso, “Jointly modeling deep video and compositional text to bridge vision and language in a unified framework,” AAAI, vol. 5, page 6, 2015.

[16] R. Hong, D. Liu, X. Mo, X. He, H. Zhang, “Learning to compose and reason with language tree structures for visual grounding,” T-PAMI, 2019.

[17] J. Hou, X. Wu, W. Zhao, J. Luo, Y. Jia, “Joint syntax representation learning and visual cue translation for video captioning,” IEEE International Conference on Computer Vision (ICCV), 2019.

[18] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, W. Liu, “Controllable video captioning with pos sequence guidance based on gated fusion network,” IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[19] X. Yang, H. Zhang, J. Cai, “Learning to collocate neural modules for image captioning,” IEEE International Conference on Computer Vision (ICCV), 2019.

[20] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, T. Mei, “Temporal deformable convolutional encoder- decoder networks for video captioning,” AAAI Conference on Artificial Intelligence, pp. 8167–8174, 2019.

[21] J. Yu, J. Li, Z. Yu, Q. Huang, “Multimodal transformer with multi-view visual representation for image captioning,” IEEE Transaction Circuits Systems Video Technology, 4467– 4680, 30, 2020.

[22] Z. Yu, Y. Song, J. Yu, M. Wang, Q. Huang, “Intra and inter modal multilinear pooling with multitask learning for video grounding,” Neural Processing Letter, 1–17, 2020.

[23] V. Iashin, E. Rahtu, “A better use of audio-visual cues: dense video captioning with bi- modal transformer,” British Machine Vision Conference,2020.

[24] V. Iashin, E. Rahtu, “Multi-modal dense video captioning,” Computer Vision and Pattern Recognition Workshops, pp. 958–959, 2020.

[25] J. Lei, L. Wang, Y. Shen, D. Yu, TL. Berg, M. Bansal, “Mart: memory-augmented recurrent transformer for coherent video paragraph captioning,” ACL, 2020.

[26] A. Sasithradevi, S. Mohamed, R. Mansoor, “A new pyramidal opponent color-shape model based video shot boundary detection,” Journal of Computer Visual Communication & Image Represention, vol. 67, 102754, 2020.

[27] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, YW. Tai, “Memory-Attended Recurrent Neural Network for Video Captioning,” 2020.

[28] G. Tan, D. Liu, M. Wang, ZJ. Zha, “Learning to Discretely Compose Reasoning Module Networks for Video Captioning,” 2020.

[29] Y. Zhou, H. nanjia, “Accelerated masked transformer for dense video captioning,” Elsevier, Neuro computing, 2021.

[30] M. Hemalatha, C. Chandra Sekhar, “Domain-specific semantics guided approach to video captioning,” IEEE, 978-1-7281-6553-0, 2020.

[31] J. Vaishnavi, Dr. V. Narmatha, “Video Captioning based on Image Captioning as Subsidiary Content”, International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), IEEE, 2022.