Video Summarization Techniques: Evolution and Observation

Mrinal Jyoti Sarma

Research Scholar

Department of Computer Science and Engineering

Rajiv Gandhi University, Arunachal Pradesh

Email: mrinaljyotisarma@gmail.com

Marpe Sora

Associate Professor

Department of Computer Science and Engineering

Rajiv Gandhi University, Arunachal Pradesh

Email:marpe.sora@rgu.ac.in

Bhaskar Jyoti Chutia

Assistant Professor

Department of Computer Science and Engineering

Rajiv Gandhi University, Arunachal Pradesh

Email: bhaskar.chutia@rgu.ac.in

Abstract:

In the present scenario of digital advancement, due to extensive use of multimedia applications, the digital video repositories are growing in a big scale. Video, being one of the robust sources of information; utilization and consumption of video (both online and offline) is practicing massively in the fields of education, surveillance, business, entertainment, news etc. In this busy world people want only the sufficient information of their interest, for example, a highlight of 30 minutes of an ODI cricket match is sufficient for a general viewer where all the wicket falls, boundaries along with some other high intensive clips of the match are covered. But for a coach, the information about field placements or some other strategies of a game may also be required to be contained by the highlight. The process of finding only the informative frames of interest (key-frames) from a video is called key-frame extraction and the process of keeping the selected key-frames together is known as video summarization. A good video summarization output must be able to represent the input video in terms of having all the crucial and sufficient information about the video. The key-frames are selected based on the interesting features extracted from the frames of a video. When only the set of

key-frames are summarized it is known as static video summarization (or a storyboard) but when a small video clip is summarized by taking small clip collections of some more consecutive frames before and after the key-frames, refers to dynamic video summarization or video skimming (like a highlight of a match). In this paper, we thoroughly survey different image and video features and their extraction techniques along with how different video summarization methods use them to get better solution in different situations. Also we have summarized a situation specific comparison of different existing video summarization methods.

Keywords- Video summarization, Key-frame extraction, Feature-extraction, Feature-detection.

1. INTRODUCTION

The term feature is not a well defined entity. It depends on the area of interest and the situation under which it is considered. Digitally an image is displayed in terms of the pixels, which contains the various information like color, contrast, energy, hue etc. or may be information less at that point. These set of pixel level information are termed as local features.

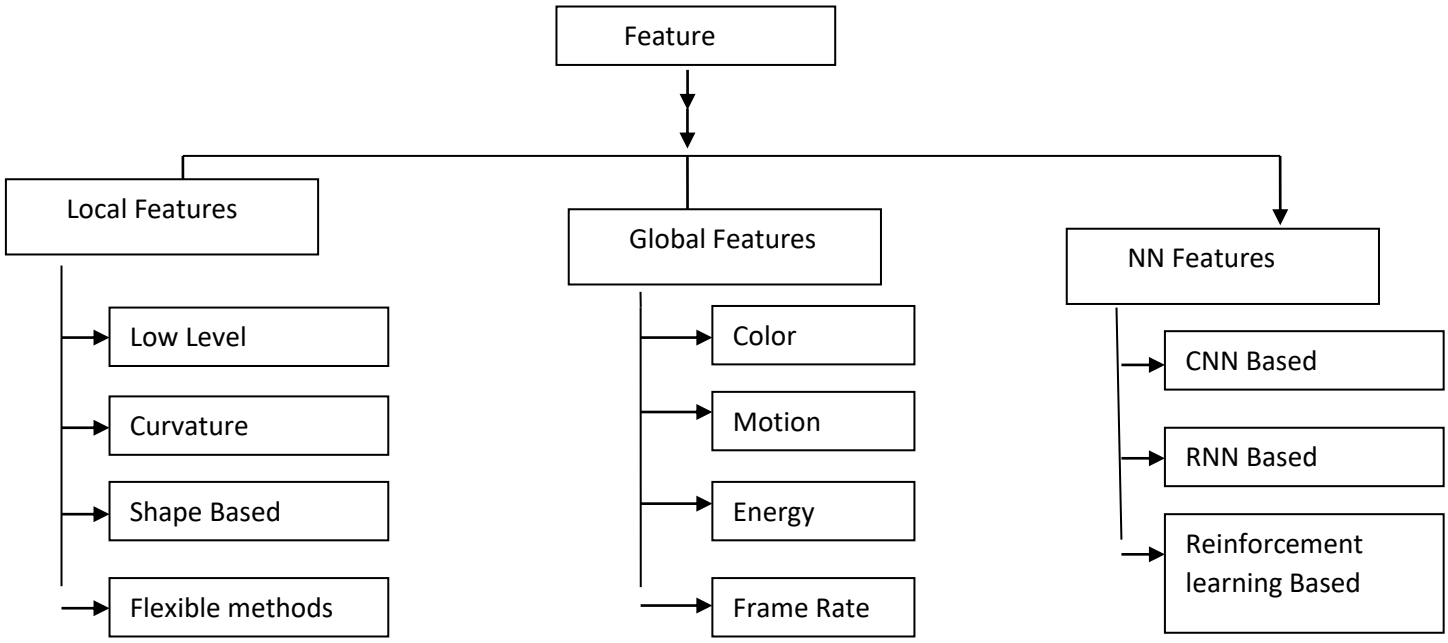

Fig 1: Classification of Features

As described in the Fig1, features can be divided into three categories: i) Local features, ii) Global features and iii) Neural Network (NN) based features. Local features are the features computed as a selected part of a frame whereas global features are computed for the whole frame [17][2]. NN based features are defined and computed by the machine itself based on the type of network the system is using. Local features can be classified into: i) Low-level features, ii) Curvature, iii) Shape based and iv) Flexible methods. Low level features include Edge detection, Corner detection, Blob detection, Ridge detection and Scale-invariant feature transform (SIFT)[8]. Curvature features contains Edge direction, changing

intensity and auto Co-relation. Different Shape based features are- Thresholding, Blob extraction, Template matching, Hough transform and Generalized Hough transform. Another type of local feature called Flexible methods contains Deformable, Parameterized shapes and Active contours. Global features can also be classified into: i) Color based (brightness, hue), ii) Motion based (motion detection, optical flow), iii) Energy based and iv) Intensity Based (frame rate). Similarly, NN based features can be classified into: i) CNN features (ResNet), ii) RNN features (LSTM) and iii) Reinforcement Learning features (DSNet).

A video consists of sequences of frames, so frame rate, sampling rate etc. are some basic features of a video. Though we normally use the term video processing separately, but inside the system it is carried out as a loop of image processing. Image processing uses feature descriptor as its first operation which is guided with higher algorithms for efficient evaluation of dominant features of that area of interest. A good number of different feature descriptors are available, based on the requirements of different interests. Feature descriptors detect the low level features like: edges, corners, blobs, ridges etc. to represent the image. Well-known feature descriptors are listed here along with the features detected by them : i) Canny(Edge) ii) Sobel (Edge) iii) Harris & Stephens/ Plessey (Edge, Corner), iv) SUSAN(Edge, Corner), v) Shi & Tomasi (Corner), vi) Level curve curvature (Corner) , vii) FAST (Corner, Blob), viii) Laplacian of Gaussian (Corner, Blob), ix) Difference of Gaussians (Corner, Blob), x) Determinant of Hessian (Corner, Blob), xi) Hessian Strength feature measures (Corner, Blob), xii) MSER (Blob), xiii) Principal curvature ridges(Ridge), xiv) Grey-level blobs (Blob).

In the next phase of image processing the output of a feature descriptor is taken as input in terms of a histogram or texture. A histogram is a pixel wise representation of a frame. A color (RGB) histogram [1] contains the pixel wise color (brightness, hue etc.) information for a frame with 16x16x16 matrix representation. An LBP (Local Binary Patterns) histogram is hierarchy mask to RGB histogram so that each pixel is compared with its eight neighbors in a 3x3 neighborhood matrix space by subtracting the central pixel value. Then the negative values and positive values coming for each neighborhood pixel after subtraction are encoded with 0 and 1 respectively and as a result a binary number is obtained by merging all these binary values in clock-wise direction which represents the central pixel.

Then the dimensionality is reduced with some higher level algorithms like PCA (Principal Component Analysis) [18][19] for faster execution. PCA is a process of computing the principal features under a real world co-ordinate by calculating the eigen-decomposition of feature set’s covariance matrix or singular value decomposition of feature matrix. Normally first few principal components are considered and rests are discarded which actually reduces the dimensionality of a feature set. Robust PCA (RPCA) L1-norm-based PCA are two improved variations of PCA [4] to achieve more robustness and less computing complexity of the process. In paper[5], different RPCA implementations have been introduced: i)RPCA via PCP(Principal Component Pursuit), ii)RPCA via Outlier Pursuit, iii)RPCA via Iteratively Reweighted Least Squares, iv)Bayesian RPCA(BRPCA), v) Variational BRPCA and vi) Approximated RPCA.

If we combine the processes of abstracting features set using feature detector and then reducing the dimensionality of that feature set, together we can term as feature extraction. The feature extraction also requires a pre-processing phase for an input image where image is smoothed by a Gaussian kernel in a scale-space representation (eg.-Gabor filter).

Then clusters are made by clustering method using the feature sets of each image using. Normally generalized clustering methods, like Gaussian Mixture Model (GMM), K-means clustering, Support vector Machine (SVM) etc. are used for easy implementation and lesser computing complexity. Based on the working procedure of a clustering method the summarization technique can be classified into two categories: i) Supervised Video Summarization and ii) Unsupervised Video Summarization. In case of supervised video summarization, a supervised clustering model is used (e.g.- SVM,CNN based deep learning models etc.) where as unsupervised video summarization uses unsupervised clustering model(e.g.- K-means, GMM etc.).

Clusters are normally evaluated on basis of the dissimilarity measures of the different feature sets. Finally key frames are selected by taking most dominant images from each cluster. This phase is known as key-frame extraction. The set of these key-frames are known as story board or static summary. We can convert this storyboard to a dynamic summary [7] or video skim [16] by taking some consecutive frames before the key-frames and some after. We can simply visualize the process by taking a window clip of k-time size keeping the key-frame at midpoint of the clip where k is taken as few second (normally 1.8 sec). Some other methods of taking important clips of a video are also introduced based on the video features like motion, intensity, density etc. for video summarization. A hierarchical structure of different video summarization techniques is described in [15].

2. LITERATURE REVIEW

The main challenge of video summarization is to separate or extract the important contents from a video. For different situation or observation, the definition of “important content” may differ based on the interest of the viewers. The extracted frames or clips must preserve all the key features of a video and yet to confirm a good computable complexity during execution of the baseline system. The computational complexity depends on the features and its size, presentation and manipulation. Features even do not have a well-defined definition as it is based on the interest of the criteria through which features are selected and considered to form a feature-set. Features can be classified into two categories: i) Local Features and ii) Global Features. The features taken over a detector (edge, corner, blob or ridge) are normally termed as local features and the features which are computed over the entire frame are known as global features (color, texture, motion, energy etc.). These pixel level features (both local and global) are also known as low level features. Some higher level information like optical flow, frame rate, density, and trajectory are also used for feature extraction. Moreover, features obtained from any machine learning application like Convolutional Neural Network (CNN) may also be considered and used for feature extraction. Some key-frame extraction techniques can be divided into following categories:

1) Color Based Technique
2) Motion Based Technique
3) Local Features Based Technique
4) Event Based Technique
5) Time Based Technique
6) Density Based Technique
7) Hierarchical clustering-based techniques
8) Neural Network Based Technique

2.1.1 Color **Based Technique**: Color is one of the most expressive, simple, stable and effective feature of a frame. The color features are computed in terms of a histogram, which contains pixel level color information of an image. Color histograms are widely used due to its simplicity and robustness against small camera motion. Redundancy elimination can be done at low computational cost but it is sensitive to noise. Normally first frame is taken as a key-frame and the next key-frames are selected on basis of dissimilarity measures of color and texture features from the color histograms of consecutive frames. Video SUMMarization (VSUMM) is one example of color-based technique that uses Hue components of color features in HVS color space to form a color histogram for performing a static video summarization.

2.1.2 Motion Based Technique: Motion based key-frame extraction techniques are gaining importance due to its expressiveness and informativeness. Motion consideration for a video can be two types: i) Object motion and ii) Camera motion. In case of video surveillance the camera is normally fixed mounted and hence camera motion is not considered. In this situation, only object motion is computed and hence computational complexity is in reduced form. But, in case of moving camera the computational complexity increases highly, which is a challenging task for using this technique. Motion estimation can be computed either by calculating pixel to pixel frame difference or by calculating the optical flow. Optical flow of each frame is calculated and results are stored in a simple motion metric, which is used for selecting the key-frames by finding the local minima of motion for a frame. Lucas- Kanade and Horn-Schunck are two popular optical flow algorithms which uses two different criteria of selecting key-frames.

This approach is also known as domain specific approach due to its capability of catching high activity contents of sports domain videos. It is independent of skimming threshold but it fails to extract key-frames accurately when the video contains high level of motion or the video is motionless. This approach is suitable for surveillance videos with medium level of motion.

2.1.3 Local Features Based Techniques: As the name suggests itself this approach uses the local features of the key-points of a frame. Scale Invariant Feature Transform (SIFT) and Speeded-Up-Robust-Features are two prominent algorithms that uses local features. In implementation of SIFT, key-points(important locations) of an image frame are defined first by finding the maximum and minimum responses of features in scale space representation of Gaussian functions calculating the differences. Only the distinct and interesting key-points are considered thereafter and rests are discarded. All the local low level features from the

selected key-points are then extracted to form a SIFT feature-set. In case of SURF a reduced feature-set is considered based on the dominance and robustness of features to secure less computational cost.

2.1.4 Event Based Technique: This approach deals with the highest semantic level of features, called events, detecting the interesting events and organizing them in essence of the original video. Normally events are detected by optical flow analysis and/or computing the energy difference of successive frames. Methods used for summarizing rare (important) events are [6]: i) RPCA-KFE, ii) Unified Framework, iii) Key-point-based Key-frame selection, iv) AJ Theft Prevention, v) Graph Modeling, vi) Two-Level Redundancy detection for personal video recorders and vii) CAFKF [14].

2.1.5 Time Based Technique: This approach uses a simple method of implementation by taking a constant time interval between two key-frames. Uniform Sampling [8] is one good example of time-based key-frame extraction. Here, every k^{th} frame of a video is selected as key-frame where k is evaluated from the length of video by which percentage the summary is required. For example, if we need 10% of summary of a video then every 10^{th} frame is taken as the key frame. Similarly, for 5%, 15% or 20% of summary video the selected frames will be of $20^{th,}$ 7^{th} and 5^{th} positions respectively. This approach doesn't require any feature to be extracted or analyzed and hence computing complexity is very less.

2.1.6 Density Based Technique: This approach uses density (number of frames) as criteria for clustering the frames. Normally clustering is done by feature-set similarity distances and a cluster grows when the neighborhood frame achieves the threshold of similarity distance. But here, a cluster grows when the density of its neighbors is greater than threshold and this approach is capable of discovering any arbitrary-shaped cluster and noise. Density-Based Clustering of applications with Noise (DBSCAN), DENsity-based CLUstEring(DENCLUE) and Ordering points to identify the clustering structure (OPTICS) are three well known algorithms that falls under this category. Like density, trajectory based techniques are also used in clustering of a feature extraction method [2].

2.1.7 Hierarchical-clustering based techniques: In this approach, hierarchy of clusters of frames is constructed based on distance, density or continuity with the independency of pre-defined clusters number [2]. Two different types of hierarchy clustering approaches are: i) Agglomerative and ii) Divisive. This approach poses a high complexity cost and the execution process gets slower when the video size increases.

2.1.8 Neural Network Based Technique: Neural Network uses regression method to train data and correlations between different features within the dataset are identified. Neural networks (NN) represent deep learning and use a number of hidden layers in between the input layer and output layer. Neural networks can be considered as the process of deep learning (DL) using artificial intelligence (AI). Different types of NN are available; common example includes: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long short-term memory (LSTM) [5], Residual Network (ResNet) [8][2] etc. These deep learning networks use the training set to learn the procedure for getting predefined desired output and apply the learned procedure on new data. Reinforcement Learning (RL) [9] is a

dynamically learning network that uses continuous feedbacks for adjusting actions to maximize a reward.

After getting the extracted features using any of the above listed feature extraction techniques, we have to make clusters of similar frames based on these features. Normally, K-means clustering and GMM are used in clustering the feature-sets (represent frames) for video summarization.

2.2.1 K-means clustering: Practically K-means is the most simple and easy-to-implement clustering method that has gained popularity due to its simplicity. It makes k partitions (cluster) by grouping each feature-set (representing a frame) from the pool of feature-sets (represents the whole video) with the condition that each cluster must have at least one object [8]. An object is kept in a cluster if it overcomes the similarity threshold of the function, used as the criteria of clustering. The clustering criteria must achieve the goal that an object of a cluster is similar to the other objects in that same cluster and dissimilar to any object from other cluster. Each cluster is represented by centroid point, measures as the average of all points in a cluster that changes after each step . K-medoids is an another clustering method similar to K-means which overcomes the problem of interpreting the final centroid of a cluster by taking the final centroid as actual data point (median).

2.2.2 Gaussian Mixture Model (GMM): This is one of the most popular data clustering method that involves the mixture of multiple Gaussian distributions. Each cluster is represented as a Gaussian distribution [8]. In K-means clustering only mean value is computed; but in GMM, along with mean, variance is also computed to make the clustering more goal oriented. GMM is considered as a probabilistic soft clustering as it can assign a data point to more than one cluster, when needed.

2.3 Dataset: Some popular and well organized image databases are: JAFEE, Cohn-Kanade, CK+ etc. JAFEE dataset have 213 images of different well annotated facial expressions of 10 different Japanese female. Cohn-Kanade image dataset is extended to CK+ dataset containing 593 video sequences of 123 different subjects between 18 to 50 years of age with a variety of gender and heritage. Popular Video dataset includes- SumMe [10], TVSum, OVP[17], VSUMM [17], YouTube etc. SumMe and TVSum are the benchmarked datasets used for video summarization. SumMe contains 25 well annotated (15 per video) videos where TVSum contains 50 well annotated (20 per video) videos.

2.4 Evaluation Method: After getting the summary as an output of a video summarization method we have to compare it with the human generated summary to evaluate the efficiency of method. For comparing we must need a measure and can be calculated as F1 score from the components: precision and recall. Precision is the ability of a classification model to return only relevant instances where recall is the ability to identify all relevant instances. F1 score combines both precision and recall using the harmonic mean:

$$F1=2(precision.recall)/(precision+recall)$$

3. OBSERVATION

Based on the experimental outcome of different researchers a comparative study of different approaches has been done and a summary has been drawn in a generalized way. Different video can have different situations like i) way of capturing, which include number of camera, movement of camera, stability of camera viewpoint etc., ii) quality of video that includes sampling rate, frame rate, motion inside video, noise etc., Moreover videos from different domain have different distinguishable characteristics. So a single approach cannot perform with same efficiency for all. That is, one single algorithm cannot be the solution of video summarization. In [16], different video domains and their summarizing criteria are well described.

For a situation where camera is fixed or with stable viewpoint and motion inside the video is very less, even the simplest methods like uniform sampling with K-means clustering can provide better results. In [6], a simple method of co-relation is used for video summarization and subtraction of matrices for object detection. In [3], a moving object detection-based video summarization method is implemented. No doubt, the other algorithm combinations will also provide better results but the complexity (and hence execution time) will increase according to the weight of the algorithm and size of the video.

Take another situation where camera is with almost stable viewpoint and the video contains higher motion inside it then SIFT/SURF algorithms seems to perform well. For clustering we can choose a motion-based clustering (DENCLUE) along with GMM. When there is a moving point video (camera is moving), convolutional neural network (ResNet) or reinforcement learning network (DSNet) [13] along with GMM performs well than any other methods.

In today's world maximum videos in Internet, specially in social media are coming from hand-held devices or wearable cameras and are being recorded for logging activities of interest. These videos are called egocentric videos and are considered as most challenging video genre for making summary from it [12]. Multi-view video summarization is yet another challenging situation where more than one camera is being used from different angles [11].

It is also observed that, object detection is used now a day as a method for video summarization. It provides very good accuracy but for a busy environment it becomes insufficient as the summary size almost same as the raw video.

4. CONCLUSION

Video summarization is one of the key procedures of video processing that reduces the cost of further video processing by eliminating unnecessary information of a video. Video summarization is a final product for many operations itself. Video summarization depends on the perception of its viewer and also the need of use (that is on what purpose it is being used?). It has a wide range of implementations from a very simple one to a computationally complex another. It is also challenging to choose a right method that meets the requirements of that specific situation. In future we will try different possible combinations of sub-methods

(feature-detector, feature-extraction and clustering) on the benchmarked datasets SumMe and TVSum to check if any combination can perform averagely well in all the situations.

References:

1. Cayllahua-Cahuina, E. J. Y., G. Cámara-Chávez, and D. Menotti. "A static video summarization approach with automatic shot detection using color histograms." *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.
2. Chamasemani, Fereshteh Falah, et al. "A study on surveillance video abstraction techniques." *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE, 2015.
3. Chung, Yi-Nung, et al. "Applying the video summarization algorithm to surveillance systems." *Journal of Image and Graphics* 3.1 (2015).
4. Bouwmans, Thierry, and El Hadi Zahzah. "Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance." *Computer Vision and Image Understanding* 122 (2014): 22-34.
5. Fajtl, Jiri, et al. "Summarizing videos with attention." *Asian Conference on Computer Vision*. Springer, Cham, 2018.
6. Balakrishnan, Aashika, Lijitha Govindankutty, and Namrata Patel. "Video Summarization: Correlation for Summarization and Subtraction for Rare Event." *Journal for Research| Volume* 2.02 (2016).
7. Elkhattabi, Zaynab, Youness Tabii, and Abdelhamid Benkaddour. "Video summarization: techniques and applications." *International Journal of Computer and Information Engineering* 9.4 (2015): 928-933.
8. Jadon, Shruti, and Mahmood Jasim. "Video summarization using keyframe extraction and video skimming." *arXiv preprint arXiv:1910.04792* (2019): 117.
9. Workie, Ashenafi, Rajesh Sharma Rajendran, and Yun Koo Chung. "Digital video summarization techniques: A survey." *Int. J. Eng. Technol* 9 (2020): 81-85.
10. Gygli, Michael, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. "Creating summaries from user videos." In *European conference on computer vision*, pp. 505-520. Springer, Cham, 2014.
11. Hussain, Tanveer, Khan Muhammad, Weiping Ding, Jaime Lloret, Sung Wook Baik, and Victor Hugo C. de Albuquerque. "A comprehensive survey of multi-view video summarization." *Pattern Recognition* 109 (2021): 107567.
12. Sreeja, M. U., and Binsu C. Kovoor. "A unified model for egocentric video summarization: an instance-based approach." *Computers & Electrical Engineering* 92 (2021): 107161.
13. Zhu, Wencheng, Jiwen Lu, Jiahao Li, and Jie Zhou. "Dsnet: A flexible detect-to-summarize network for video summarization." *IEEE Transactions on Image Processing* 30 (2020): 948-962.
14. Yasmin, Ghazaala, Sujit Chowdhury, Janmenjoy Nayak, Priyanka Das, and Asit Kumar Das. "Key moment extraction for designing an agglomerative clustering

algorithm-based video summarization framework." *Neural Computing and Applications* (2021): 1-22.

15. Chavan, Tejal, Vruchika Patil, Priyanka Rokade, and Surekha Dholay. "Superintendence Video Summarization." In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1-7. IEEE, 2020.
16. Sen, Debashis, and Balasubramanian Raman. "Video skimming: Taxonomy and comprehensive survey." *arXiv preprint arXiv:1909.12948* (2019).
17. Mei, Shaohui, Genliang Guan, Zhiyong Wang, Shuai Wan, Mingyi He, and David Dagan Feng. "Video summarization via minimum sparse reconstruction." *Pattern Recognition* 48, no. 2 (2015): 522-533.
18. Shakya, Subarna, Suman Sharma, and Abinash Basnet. "Human behavior prediction using facial expression analysis." In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 399-404. IEEE, 2016.
19. Kauser, Nazima, and Jitendra Sharma. "Automatic facial expression recognition: a survey based on feature extraction and classification techniques." In *2016 international conference on ICT in business industry & government (ICTBIG)*, pp. 1-4. IEEE, 2016.