Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Newsgist: video generation from news stories

M. S. Karthika Devi & R. Baskaran

To cite this article: M. S. Karthika Devi & R. Baskaran (2023) Newsgist: video generation from news stories, Automatika, 64:4, 1026-1037, DOI: 10.1080/00051144.2023.2241774

To link to this article: https://doi.org/10.1080/00051144.2023.2241774

Published online: 02 Aug 2023.

AUTOMATIKA
2023, VOL. 64, NO. 4, 1026–1037
https://doi.org/10.1080/00051144.2023.2241774

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Newsgist: video generation from news stories

M. S. Karthika Devi and R. Baskaran

Department of Computer Science and Engineering, College of Engineering, Guindy, Anna University, Chennai, India

ABSTRACT
Digital transition has started to change the way people read news articles more through a digital device and less on paper. Youngsters today do not spend enough time reading news articles. In this work, a knowledge-driven news story generation using collaborative learning to represent the gist of news is proposed. The entire work focuses on two major concerns. Initially, the dialogues associated with the corresponding speaker are extracted from the news. Secondly, the audio of the mapped dialogues is incorporated into the final video. Logistic Regression is deployed to identify the theme the news. Deep learning techniques are employed to identify the main characters in a supervised manner using Named Entity Recognition (NER) tagging algorithm, suitable cartoon dispositions and their semantic relations. This approach improves not the reader's comprehension and creativity but also improves mutual goals, opportunities for peer discussion and engaging the underachievers to think reflexively. In addition, it also improves the learner's motivation and participation. The proposed framework outperforms an accuracy of 83.98% when compared with conventional methods also suggests that the readers found the packages interesting and informative on digital devices. Moreover, this method can be used efficiently in real-time for various applications.

ARTICLE HISTORY
Received 19 May 2023
Accepted 22 July 2023

KEYWORDS
News extraction; theme identification; face swapping; video reels; CaffeNet

1. Introduction

"A Picture Speaks a Thousand Words". But if a picture is worth a thousand words, then imagine the massive worth of videos. Watching a video to understand anything is far easier than reading a whole lot of paragraphs as it is time efficient. Reading news article is a very good/important activity at all stages of life to keep ourselves up to date about day to day affairs. Short video stories can improve the news consumption experience so that people can catch up on what's happening rapidly. Thus, using machine learning and Natural Language Processing techniques it is possible to extract the core atomic events, key time, place, character information and identify the relationship between entities of events from news, blog and other text that can effectively help people understand the events.

As a result of technological advancements, it is now possible to create personalized visual multimedia. Particularly, image generation based on deep learning has been the subject of extensive research in a variety of contexts. In contrast, video generation, especially when based on conditional inputs, is still challenging, and understudied. In contrast to creating an image, which only requires consideration for how everything fits together in a single frame, creating a movie requires consideration of how everything fits together in each frame. It does not matter how high the quality of each individual image is if the continuity between frames is not maintained. To fill in the voids in our knowledge of this topic, we will train our model to construct a video that is analogous to a given study.

Exploration in language comprehension instigated the hierarchical organizational framework of contents which are determined by the grammar (a set of conventions), representing the abstract structural components of the plot. As every human being tends to use memory to apprehend any news or sequence of events, the approach of generative grammar and plot structures has resulted to a greater extent. Thus the proposed work attempts to conceive a set of conventions given in Table 1 which are found to be the basis for constructing the plot structure shown in Figure 1 for news video reel generation. By adhering to the set of conventions and plot structures, the proposed work attempts to generate a News video reel using state-of-the-art techniques like Natural Language Processing and deep learning framework.

This work concentrates on automatic video generation from a news app. The final video generated will include characters (corresponding characters in the news) explaining a series of news of different categories. In short, the final output is a video semantically aligned with descriptive scripts.

CONTACT M. S. Karthika Devi ✉ mskarthikadevi2014@annauniv.edu Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

Table 1. Rules for video reel generation.

Rule #		Rule
1	Video Reels	Setting ∪ Theme ∪ Resolution
2	Setting	Character ∪ Background ∪ Audio ∪ Time
3	Character	Protagonist/Antagonist
4	Audio	Dialogue ∪ Gender Identification
5	Time	Dialogue ∪ Duration Calculation
6	Theme	Event ∪ Goal
7	Event	Video Reels
8	Goal	Event
9	Resolution	State/Event

The entire work involves main phases namely:

- News extraction and Dialogue mapping.
- Image extraction and character creation.
- Incorporating audio to the final video (semantically aligned).

Initially, it is intended to choose to extract news of diverse categories from a news app named in shorts, which aggregates news and summarizes them in 60 words or less. It is found to be one of the highest-rated news applications. Extracting relevant information from the website is found to be the first and foremost step of our proposed work. After news extraction, the dialogues are extracted using regular expressions and the respective speakers using Named Entity Recognition tagging. A basic news feed will fundamentally include dialogues voiced out by the speaker either in a direct or indirect manner. For direct/indirect dialogues, voices are chosen based on gender, whereas for the news headline another voice will be reading it out. Both the dialogues are taken into account and the dialogues are parallel mapped with the speaker.

After speaker identification, their images are scraped from the web, their face is located and their gender is identified. This is done using deep Convolutional Neural Networks (CNN) using CaffeNet models. Then the extracted face is swapped with a cartoon body of the appropriate gender using Delaunay Triangulation and smoothening techniques. The final video including the characters' images along with the voice and respective background images aligned properly will then be generated.

Usually, people prefer visual content to big paragraphs. Providing the characters' images along with the audio will create an impact in their minds. To achieve this, presenting them with a video of the events will be more useful than reading all the news content. As the manual creation of videos for each news article can be tedious, this work focuses on creating an automated system for the same.

1.1. Rule generation and hierarchy

A set of rules which provides the basis for constructing the video reels is given in Table 1. The ensued rules would suffice in delivering the crux of the news and ultimately achieving the purpose of the work. Figure 1 shows the hierarchy of video reel generation which was the corollary of the proposed rules.

Rule 1 provides the indispensable features for contriving video reels which include Setting, Theme and Resolution. The symbol "∪" symbolizes the combination of all the features. For video reel construction of news Gist, rule 2 lists imperative setting features which include background, characters involved, audio and time. According to Rule 3, the characters involved may be a protagonist or an antagonist. To achieve appropriateness in audio-visual generation, rule 4 and 5 specify the process of dialogue extraction and gender identification. Also, dialogue duration is computed for the integration of audiovisual and the dialogues uttered by a character. In Rules 6 and 7, the theme is to achieve the goal of contriving video reels by generating audio-visuals with dialogues, appropriate backgrounds and the actual characters in a cartoonish manner. Rule 8 Resolution specifies the conclusive result of the video reel generation to make an individual understand the content in a progressive manner. It conforms to the

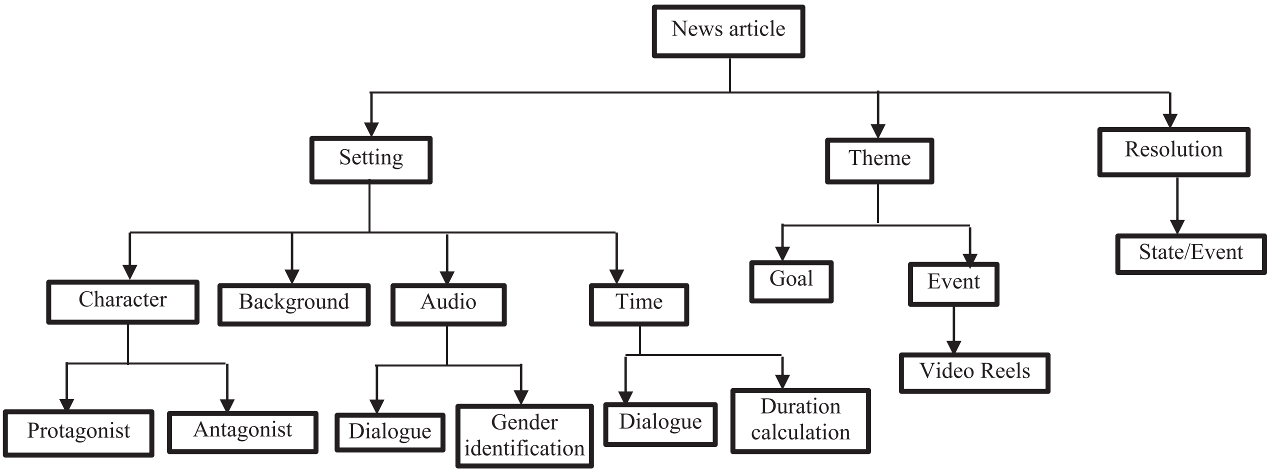


Figure 1. Hierarchy for video reel generation.

retrieval of actual characters (face swapped with cartoon images), extraction of dialogues, gender identification and the process of incorporating appropriate background for video reel generation to provide an individual with clarity in the news.

This research is formulated as follows. Section 2 presents related works. An automated system for generating news video generation is discussed in Section 3. Section 4 examines the obtained experimental results. Section 5 finishes the research with a discussion on possible improvements.

2. Related works

2.1. Text extraction

Ziyi et al. [1] proposed a method where blocking tags of the web site are automatically detected and parsed for retrieving news content. Complexity is reduced, and there is no need for web page templates or extensive training sets. Three steps namely preprocessing, page dividing and text-recognition are used once the correct blocking tag is received. The final step is finding out the news with the largest number of words.

The procedure described in [2] illustrates how to autonomously extract data from the web rather than manually duplicating it. It accomplishes this by utilizing the website's Uniform Resource Locator (URL) and associated operations such as Hypertext Transfer Protocol (HTTP) scripting, Document Object Model (DOM) parsing and HTML parsers. Using a content analysis of monthly news headlines, [3] investigated the development of relationships over time. Headline id, Entities, Keywords and Publication Date are the data fields in which entities and key phrases extracted from the news were entered. According to their frequency, entity pairings are assigned weights to determine the most prevalent set of relations. The Google Knowledge Graph, the classification of news stories in light of the current relationship and numerous other applications could benefit from this information.

Liu et al. [4] proposed two Yuan classification models are constructed to put the N-gram language model which considers the statistics of the frequency of occurrence of word strings in the text instead of word frequency to increase the accuracy of the classification model. Azzopardi et al. [5] introduced a highly efficient tree structure analysis chosen to produce effective results and to further solve problems like structure-based page classification, extractor generation and data labelling. Karthika Devi et al. [6] use state-of-the-art natural language tools and libraries to extract text from the stories and also identify the source of each text which aids in generating a text conversed by a cartoon speaker. Karthika Devi et al., [7] deploy a multimodal network to generate text from a stream of images by interrupting the meaning of the images. For a graphic novel book, a synopsis using an unsupervised abstractive dialogue is generated [8]. A custom ensemble method that consists of an inception model and 2 layer Long Short Term Memory (LSTM) model was created by Alzubi et al. [9] for generating captions from the image automatically.

In addition, a method was developed to simulate the manner in which individuals absorb and process news accounts. The concept of using content function, spatial consistency and format continuity to identify news zones was introduced in [10]. It locates, formats and identifies the semantic value of news areas in order to determine the news content. Given the large number of studies conducted on the topic of paraphrasing, Alzubi et al. [11] proposes a Collaborative Adversarial Network (CAN) model for extracting the most essential information from two phrases, which performed better than the current MaLSTM model. The title-based web content extraction model (TWCEM) proposed in [12] makes use of title-based characteristics that can filter the noises in websites and collect accurate content. Using Rich Site Summary (RSS) feeds, Newspaper python library and the Uniform Resource Locators (URLs), [13] demonstrates an advanced automated process for extracting and categorizing online news articles. In this investigation, three supervised learning algorithms were combined to create an ensemble classifier for news article classification.

2.2. Gender prediction

A gender prediction method using Bag-of-Words(BoW) is proposed in [14]. BoW is a technique that extracts features from the text for modelling or it simply represents the text from a document that describes the occurrence of words. They have developed a vocabulary for their model, which is a dictionary mapping of words and count, making it easy to update and query.

Ito et al. [15] propose a gender prediction method from face images using CNN namely WideResNet. This method outperforms the other four CNNs. It also makes use of Deep Multi-Task Learning (DMTL) to enhance the accuracy and decrease the computation time for gender and age prediction. Using the YOLO v3(You Look Only Once v3) object algorithm, darknet for training, Keras and OpenCV for testing over a combination of Google, IMDb and mobile camera images, gender classification is done in [16].

Component Analysis method is used in [17] as an image representation technique for gender classification making use of images from the FERET dataset as a feature vector in a low dimensional subspace. Face and gender recognition systems in [18] make good use of the global average pool in ResNet 50nn for efficient results on various databases. In [19], full-duplex capability to improve the spectral efficiency of unidirectional traffic has been identified. Skin colour distribution function is used in [20] to identify similar colour areas from face images considered as 2D images which were primarily extracted using a fuzzy pattern finally to get reliable results.

SVMs performance in [21] is found to be superior to traditional pattern classifiers while performing classification on images from the FERET dataset. This paper generally provides an evaluation of different classifiers. In a trainable COSFIRE filter for which selectivity is determined in an automatic configuration process combined with Support Vector Machine (SVM) is used for gender classification. Ng et al. [22] provide an extensive survey on facial feature extraction which aids in gender identification.

2.3. Text-to-video generation

The adversarial examples are generated using Bo-GAN [23]. A video generator, a semantic alignment module at the region level, a coherence-aware discriminator at the frame level, and a semantic-aware discriminator at the video level are the four components of this method. This technique produces high-resolution films with semantic and temporal coherence, but it struggles with phrases that contain multiple nouns or are overly complex.

Liu et al. [24] presented a Cross-Modal Dual Learning (CMDL) approach that takes spatial–temporal coherence between individual video frames and semantic consistency between description phrases and produced films into consideration. It consists of two components: the first generates videos from sentences, and the second re-embeds the videos into sentences. The sentence-to-video generator contains both the text-to-visual feature encoder and the conditional video generator. Despite being time-consuming and complex, this method produces high-quality films with great precision.

A technique for video creation is proposed by Li et al. that divides the generation effort in half [25]. Using a conditional VAE (Variational Auto Encoder) model, the "gist" of the video is generated from the supplied text. Next, the content and motion of the video are constructed based on the relationship between the summary and text input. The purpose of this method is to simulate the human creative process. Utilised is the Generative Adversarial Network, which is based on a minimax game between a "generator" and a "discriminator."

TiVGAN (Text-to-Image-to-Video Generative Adversarial Network) is a model for text-to-video generation presented in [26] that is based on the concept of connecting frames of movies with significant continuity. The training procedure consists of two distinct phases: text-to-image generation and evolutionary generation. Independent Samples Pairing is utilised to prevent the modes of the GAN from collapsing, resulting in enhanced precision.

2.4. Theme identification

The category browsing functionality, which is based on the text classification technique, was used to identify news categories [27]. In this machine learning classification method, features are randomly selected words from a given text corpus. To efficiently train the classification model, the search space is reduced to a manageable few thousand words using feature selection techniques. SVM is defined over a vector space with the objective of discovering the optimal decision surface with high performance. It is founded on the Structural Risk Minimization principle and error-bound analysis. A count vectorizer can transform a corpus of text into a vector of word / token counts. To diminish the significance of individual elements within a corpus, a Term Frequency-Inverse Document Frequency (TF-IDF) transformer may be applied. In [28], the categorization model Latent Dirichlet Allocation (LDA) is proposed for news text. This method employs topic models for dimension reduction and feature extraction in text.

2.5. Face extraction and face swapping

A system for autonomously deriving human facial features from colour images is presented in [29]. The system detects the position of a person's face and its characteristics (such as the nose, eyes, etc.) and then hierarchically extracts the face's contours and feature points. The location of the face is determined using integral projection. Despite the complexity of their contexts, it has been demonstrated that the proposed system is both effective and dependable in face image processing. In the paper "Constrained Generative Model" [30], two techniques for identifying facial geometry are presented.

Korshunova et al. [31] performed face swapping using a trained convolutional neural network to capture the appearance of the target identity from an unstructured collection of his/her photographs. By combining neural networks with simple pre- and post-processing steps, a face swap method in real-time with no input from the user is proposed. Xingjie et al. [32] proposed a blending method involving adaptive weight values to avoid boundary lines during face swap and handle differences in brightness which involves giving adaptive weights to extracted synthesis regions. A face-swapping method proposed by Sadu and Das [33] uses 81 facial landmark points of the human face by combining two methods – detecting the landmarks using an ensemble of regression trees and the Surrey Face Model, which involves the use of a 3D morphable face model. In addition, colour correction using RGB scaling is used in [34]. Chen and Ni [35] proposes a method using a pipeline with three parts-expression transfer, details refining and merging back with efficient results by combining the 3D model and Generative Adversarial Model even with limited input.

Nirkin et al. [36] show that a face segmentation method using Fully Convolutional Network(FCN) can achieve remarkable and fast results. Matsuhashi et al. [37] proposed a method based on the Hue-Saturation-Value colour system for extraction of human faces from a general indoor background. Sushama and Rajinikanth [38] proposed a method to extract face features using Scale Invariant Feature Transform(SIFT) and detecting faces using Back Propagation Network (BPN). Wu et al. [39] used 3D stereographic images to detect the head and face. Bhuvaneshwari et al. [40] used the Synthesized Face Recognition (SFR) model to degrade the effect of low resolution, blurriness and illumination in recognizing facial images. Smitha et al. [41] worked on proper extraction and analysis of face images irrespective of illumination, pose and face shape. In [42], a method using genetic algorithms after computing the non-zero feature space of the training set scatter matrix to search the face for discriminating features are proposed.

2.6. Text to audio conversion

Acero [43] explained several well-known speech coding techniques (including LPC vocoders, waveform interpolation, harmonic coding and layered coding) that have been used in speech synthesis.

3. Proposed method

The framework shown in Figure 2 depicts the seven modules of this proposed work.

3.1. Data extraction and pre processing

This module fetches news articles from the *Inshorts* app. *Inshorts* is a news app that selects the latest and best news from multiple national and international sources. The response website is parsed using an HTML parser and the news headline and description is extracted using appropriate values of “itemprop” attributes.

3.2. Prominent character and theme extraction

The characters mentioned in the news are extracted using NER -tagging corresponding to persons’ names explained in Algorithm 1. The theme of the news is recognized using TF-IDF vectorization and a logistic regression model trained on the “News” dataset, created by extracting older news from the Inshorts app.

3.3. Dialogue extraction

All the extracted prominent characters using NER tagging algorithm and dialogues are then inserted into the final output video. The news feed matches the following pattern: [Dialogue – direct or indirect quote] [Speaker]. The dialogues may be expressed directly (speaker’s exact words) or indirectly (rephrased by journalist) in news topics. The dialogues and their respective speakers are extracted from the news description with the use of regular expressions as in Algorithm 2.

Algorithm 1 Extraction of characters from title and description

```
Input: Dictionary Newslist[i] = (Headline H, Description D)
Output: List of characters C
nlp ← Load Spacy NER Model for 'English'
C ← Empty list
doc ← Tokenize nlp(H, D)
for each e ∈ doc do
    l ← Find label(e)
    if l = 'PERSON' then
        C ← C + e
    end if
end for
return C
```

Algorithm 2 Dialogue Extraction

```
Input: Dictionary Newslist = (Headline H, Description D)
Output: Dialogue List Dia
P1 ← Regex with pattern r '"(.*?)"'
D ← Empty List
Di ← Dialogues with matching pattern P1
for each e ∈ Di do
    l ← Find label(e)
    if l = 'PERSON' then
    S = e
    D ← D + (Di, S)
    end if
end for
if ':' ∈ H then
    if l after ':' = 'PERSON' then
        S = l after':'
        d = l before ':'
    end if
end if
D ← D + (S, d)
P2 ← Pattern with(said that\announced that\declared
   that\stated that)
Di ← Dialogues with matching pattern P2
for each e∈ Di do
    S ← regex group(1)
    d ← regex group(4)
    D ← D + (S, d)
end for
return Dia
```

3.4. Image retrieval and face recognition

This module retrieves the images of the characters from the Web. The images are then extracted by HTML parsing using appropriate CSS selectors. The face of the character is detected and extracted from the retrieved image. Then the face of the character is extracted using the cvlib module with predefined weights using a deep

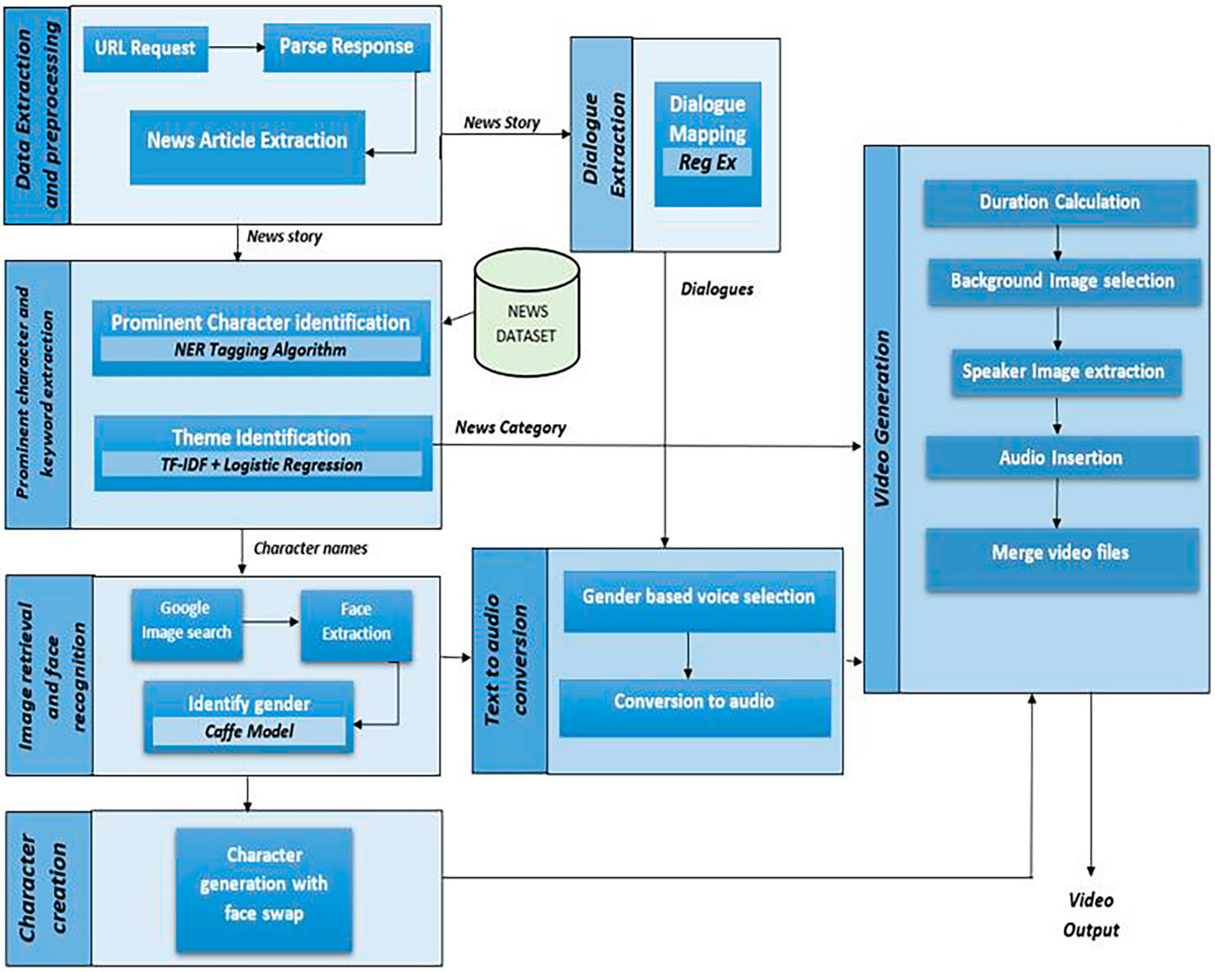


Figure 2. NewsGist framework.

Convolutional Neural Network called Caffe (Convolutional Architecture for Fast Feature Embedding) model as given in Algorithm 3.

In recent years, deep convolutional neural networks have consistently outperformed human specialists in a variety of challenging visual analysis tasks. The success of training and deploying deep models with parameter-heavy architectures on a wide range of applications can be attributed, in part, to the growing power of graphics processing units. Due to their large dimensions and excessive power consumption, such devices cannot be utilized in robotics. Recent efforts in the field have concentrated on devising architectures for deep learning that can operate efficiently on low-powered devices. Training efficient and effective models requires modules with fewer parameters and fewer floating point operations, in addition to cautious adjustment.

Algorithm 3 Face Extraction

```
Input: Image i
Output: Face Coordinates list Flist
$L \leftarrow$ Load model layers
$W \leftarrow$ Load model weights
$Caffe \leftarrow$ Initialize DNN model$(L, W)$
$i \leftarrow resize(i)$
height, width, channel count $\leftarrow i.shape$
$R, G, B \leftarrow$ Average pixel densities of RGB$-$values
for each pixel in i do
    $pixel.red \leftarrow pixel.red - R$
    $pixel.green \leftarrow pixel.green - G$
    $pixel.blue \leftarrow pixel.blue - B$
end for
$F\ list \leftarrow [\,]$
$detections \leftarrow Caffe(i)$
for each detection in detections do
    if $detection.confidence > threshold$ then
        $F\ list \leftarrow F\ list + detection$
    end if
end for
return Flist
```

3.5. Character creation

The face extracted using Algorithm 3 is then utilized to determine the gender of the character using a pre-trained Caffe model as in Algorithm 4. The retrieved face is attached to a suitable body based on gender. For this, facial key points are extracted and Delaunay triangulation of the source character's face and the destination cartoon character's face is done. The source and destination triangles are swapped and blended to match the destination face colour. Thus the whole character is created and placed at suitable places on the background image for the final video.

3.6. Text-to-audio conversion

This step involves the artificial synthesis of human speech. It converts human language text into human-like speech. Voices are allocated based on the characters' genders and in addition to the default voice that is allocated for the news reader.

A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. Audio files for the news and the dialogues are generated separately and finally concatenated which is then finally added to the output video.

Algorithm 4 Gender Prediction

```
Input: Character Image C
Output: Gender G
  L←Load model layers
  W←Load model weights
  Caffe←Initialize DN N model(L, W)
  i←resize(i)
  F←Detect face using Caffe(i)
  R, G, B←Average pixel densities of RGB−values
  for each pixel in i do
  pixel.red ←pixel.red−R
  pixel.green ←pixel.green−G
      pixel.blue←pixel.blue−B
  end for
  x, y, w, h←Extract coordinates width, height off ace(F)
  G←Caffe.predict([x, y, w, h])
  return G
```

3.7. Video generation

The background image is chosen based on the category of the news in the first place and the duration of the audio files are calculated. Face-swapped character images are added to the video and display duration is set based on the length of the audio files. Background images and the characters are sequentially added in the final output video along with the corresponding gender-based audio file at appropriate places. Finally, all the top news videos are combined to give the final video output as given in Algorithm 5.

4. Experimental results and analysis

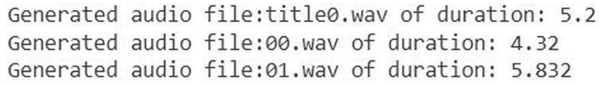
The first step is the scraping of news from the web from *Inshorts* App. One of the news extracted from the website is as follows:

Headline: "England women's Hartley praises Ashwin over tweet on women's cricket match"

Description: England woman Alexandra Hartley praised Ravichandran Ashwin for his tweet during Women-South Africa Women T20. Ashwin questioned "What is the procedure to send a soft signal to a player?". "This is what we like to see and this matters. A genuine interest in the women's game", Hartley replied.

Algorithm 5 Video Generation

```
Input: Category C + Speaker Images list I + Audio file durations list
    A + Final Audio File F
Output: Video V
  L←List of filename−category maps' category' :' background image'
  B←Background Image L(C)
  imageClip1←Video(image = B, duration = duration(A))
  titleDuration←A[0]
  imageClip2←Video(image = Reporter, duration = titleDuration)
  clipsList←[ ]
  clipsList←clipsList + imageClip1
  clipsList←clipsList + imageClip2
  audioFileIndex = 1
  startIndex←A[0]
  for each character in I do
      clipsList←clipsList + Video(image = character, duration = audioFileIndex, start = startIndex)
      startIndex←startIndex + A[audioFileIndex]
      audioFileIndex←audioFileIndex + 1
  end for
  V←Combine Videos(final Video)
  V←Add audio(F)
  return V
```



```
Generated audio file:title0.wav of duration: 5.2
Generated audio file:00.wav of duration: 4.32
Generated audio file:01.wav of duration: 5.832
```

Figure 5. Generation of audio files for the news.

Next, the category of this news article is identified as "sports" using TF-IDF vectorization and logistic regression model. Thirdly, the dialogues and names of the respective speakers are identified using regular expressions and NER tagging. Here, two dialogues are identified, one by Ashwin and Harley each. The extracted news title, description and identification of the category, dialogues and speakers are shown in Figure 3.

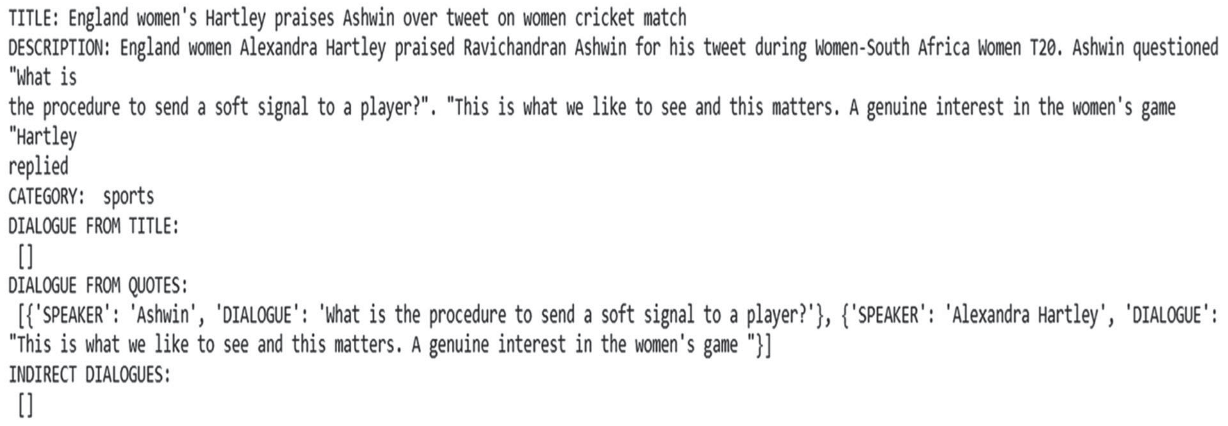
After the extraction of categories and dialogues, the cartoon images for the speakers are created. For this, firstly the image of the speaker is extracted from the web and then the gender is identified. The face is then swapped with the appropriate cartoon character body. The actual extracted image and face-swapped cartoon character for the first speaker "Ashwin" is inserted. Similarly, the actual image fetched from the web and face-swapped cartoon character for the second speaker Alexander Hartley is shown in Figure 4.

After the characters are created, the audio files for the dialogues and the headline are generated as shown in Figure 5.

After creating the character images, the video is generated with the background image. Here, since the category is "sports", a background image of a stadium is chosen. Finally, the video is generated by merging the audio files, the created characters and the background image. Finally, a video of 17 s is generated. A snapshot from the final video for different categories is shown in Figure 6. For news articles with different categories, the background images will be changed representing the corresponding category thus giving the user a nice view of the news video. Furthermore, it keeps them engaged.

4.1. Dataset – NEWS dataset

For predicting the category of news, "News dataset" is used. This news dataset consists of about 4797 tuples



```
TITLE: England women's Hartley praises Ashwin over tweet on women cricket match
DESCRIPTION: England women Alexandra Hartley praised Ravichandran Ashwin for his tweet during Women-South Africa Women T20. Ashwin questioned
"What is
the procedure to send a soft signal to a player?". "This is what we like to see and this matters. A genuine interest in the women's game
"Hartley
replied
CATEGORY:  sports
DIALOGUE FROM TITLE:
 []
DIALOGUE FROM QUOTES:
 [{'SPEAKER': 'Ashwin', 'DIALOGUE': 'What is the procedure to send a soft signal to a player?'}, {'SPEAKER': 'Alexandra Hartley', 'DIALOGUE':
"This is what we like to see and this matters. A genuine interest in the women's game "}]
INDIRECT DIALOGUES:
 []
```

Figure 3. Snapshot of news dialogue – one male and one female speaker.

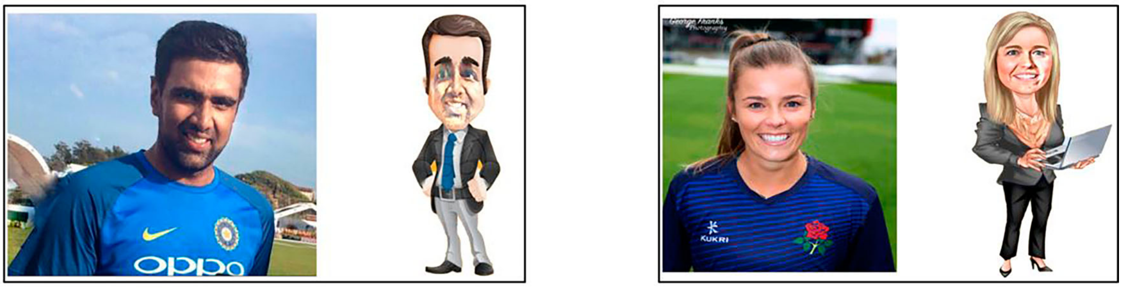


Figure 4. Actual and face swapped image of Ashwin and Hartley.

with three columns namely news headline, news article and the news category. The news articles are from various categories and are ordered sequentially. News articles listed in this dataset are from 9 different categories. The news categories are business, sports, automobile, entertainment, politics, technology, world and science. These tuples are extracted and used for the classification of news categories.

4.2. Theme identification

The performance of the model trained for predicting the category news is evaluated using two metrics namely Precision and Recall. Categories may be predicted correctly or incorrectly during the testing phase. The complete performance of the model on a set of test data is known from the confusion matrix.

Confusion Matrix gives us a matrix as output and gives the visualization of the complete performance of the model. This table layout includes four classes:

- **True Positive:** Outcome where the model predicts positive class correctly. In this case, news is classified under the “sport” category which is also actually under the “sport” category.
- **True Negative:** Outcome where the model predicts negative class correctly. In this case, news that is not under the “sport” category is classified correctly under another category.
- **False Positive:** Outcome where the model predicts positive class incorrectly. In this case, news which is not under the “sport” category is classified under the “sport” category.
- **False Negative:** Outcome where the model predicts negative class incorrectly. In this case, news which

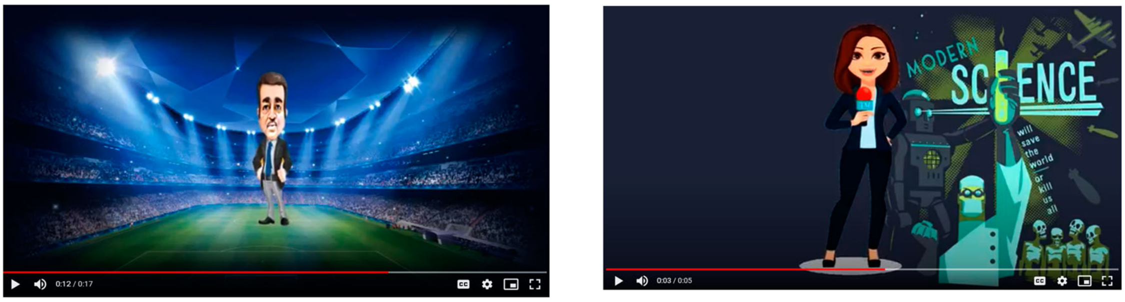
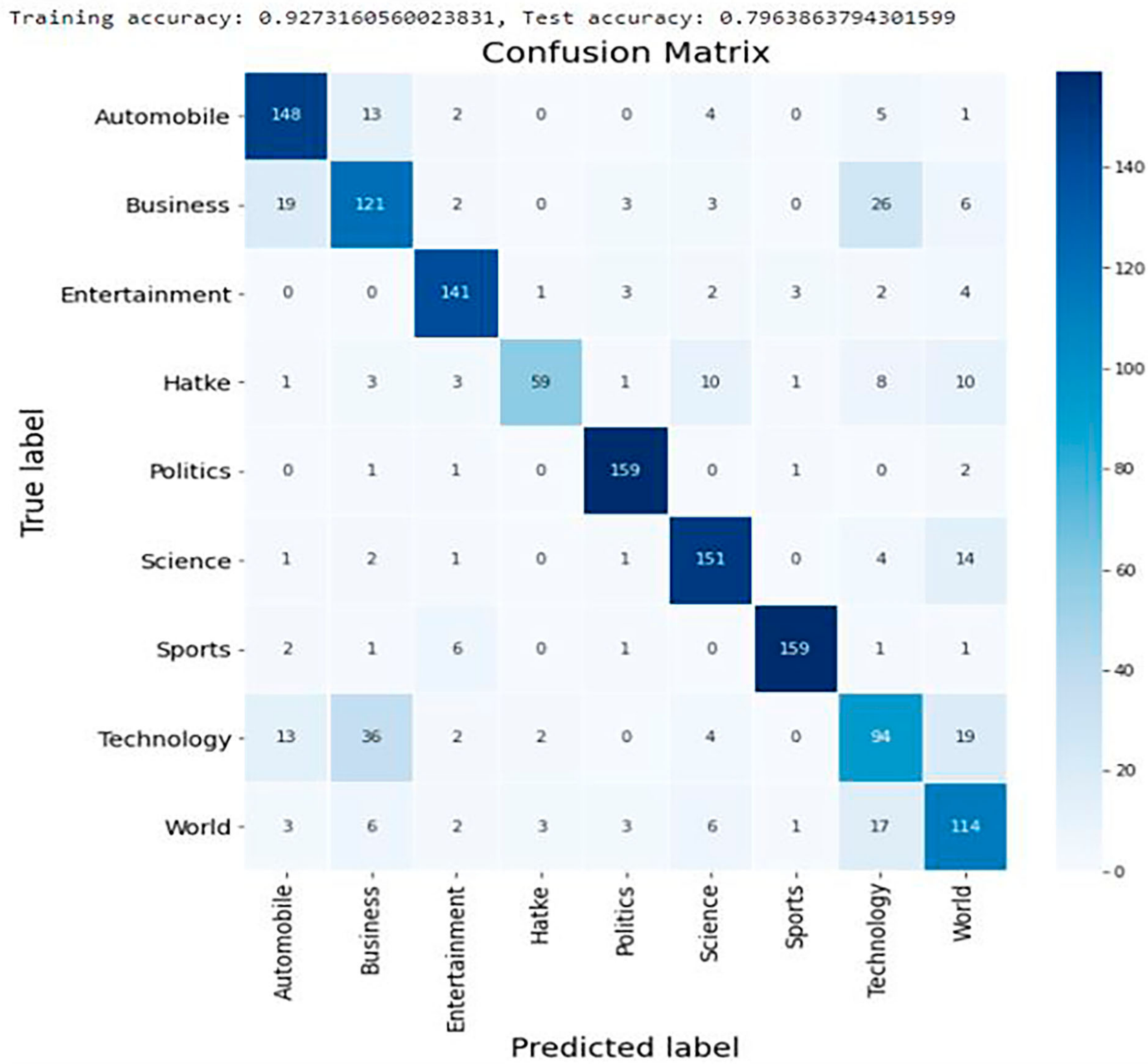


Figure 6. Snapshot from final video of category – “Sport” and science.



```
Training accuracy: 0.9273160560023831, Test accuracy: 0.7963863794301599
```

	Automobile	Business	Entertainment	Hatke	Politics	Science	Sports	Technology	World
Automobile	148	13	2	0	0	4	0	5	1
Business	19	121	2	0	3	3	0	26	6
Entertainment	0	0	141	1	3	2	3	2	4
Hatke	1	3	3	59	1	10	1	8	10
Politics	0	1	1	0	159	0	1	0	2
Science	1	2	1	0	1	151	0	4	14
Sports	2	1	6	0	1	0	159	1	1
Technology	13	36	2	2	0	4	0	94	19
World	3	6	2	3	3	6	1	17	114

Figure 7. Confusion matrix for theme identification – logistic regression.

is under the “sport” category is classified under another category.

Precision: Precision gives the fraction of correctly identified positive results out of all predicted positives.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (1)$$

Recall: Recall is the number of correct positive results divided by the number of all relevant samples.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (2)$$

Logistic Regression is chosen since it has higher accuracy compared to the other models. The values obtained from our model for Precision and Recall are 84 and 79, respectively.

A confusion matrix is generated for the identification of the theme using logistic regression as shown in Figure 7. Looking at the high concentration of correct predictions along the diagonal, it is inferred that the model performs well in identifying all the categories except for technology and business for which there were a lot of interchanged predictions. This is due to the similar keywords found in news belonging to both categories.

Table 2. Comparison of results for theme identification.

Models	Accuracy(%)	Precision	F1 Score
Support Vector Classifier (SVC)	83.32	83	83
Multinomial Naive Bayes	83.91	84	84
Logistic Regression	83.98	84	84

In order to obtain a suitable model for theme identification, three models were chosen and metrics were compared as shown in Table 2. From the table, it can be observed that Logistic regression performed with the highest accuracy, precision, and F1-score when compared with Support vector classifier (SVC) and multimodal naive Bayes.

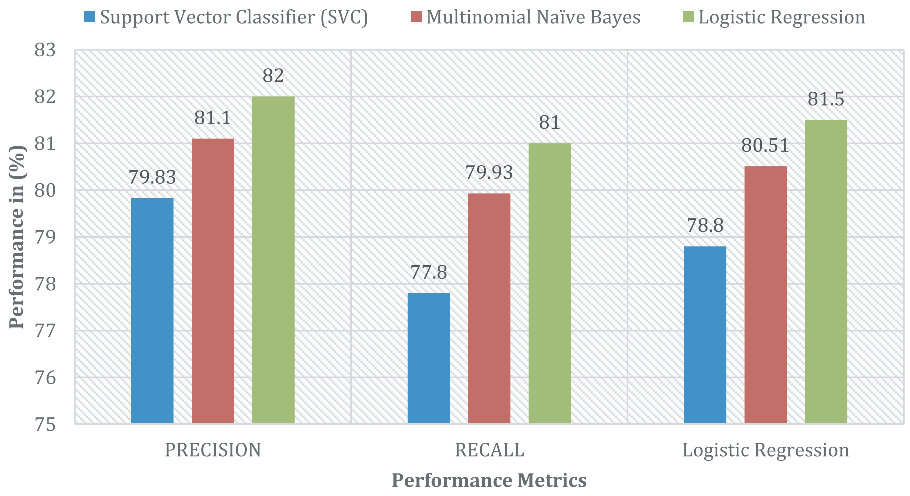


Figure 8. Comparison graph for gender prediction.

4.3. *Gender prediction*

The performance for gender prediction is evaluated using two metrics namely Precision and Recall. Gender may be predicted correctly or incorrectly during the testing on a set of test data. The values obtained from our model using logistic regression for Precision and Recall are 82 and 81 and the same has been shown in Figure 8.

5. Conclusion

In the proposed work, the automatic generation of video from the extracted news feed from the Inshorts app helps in grasping a gist of the daily news in a short span of time. It also helps to understand the news in a simpler manner. Various challenges arise in video generation. Yet, it is made possible with the help of machine learning techniques. This will be of greater use for people of all ages, children especially find it easy to grasp information because of the incorporated cartoon characters in the final output video. This improves the news consumption experience so people can catch up on what is happening quickly. The final generated output will include approximately 5 news snippets ordered sequentially with semantically aligned gender-based audio. The usage of gender-based audio makes users engaged. The proposed work creates a way for future works by generating videos with better animation effects. Also, 3D characters can be incorporated which further makes the video much more lively. It can be enhanced further by providing news articles regarding people's interests. The efficiency of the present work can be increased further with much more efficient machine learning techniques. Research is being carried out for finding the best way in generating video from the text. This work can be extended for making videos with many more animation effects.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Ziyi L, Beijun S, Xinhuai T, et al. Automatic web news extraction using blocking tag. 2009 Second International Conference on Machine Vision; 2009; Dubai, United Arab Emirates. p. 74–78.

[2] Singrodia V, Mitra A, Paul S. A Review on web scraping and its applications. 2019 International Conference on Computer Communication and Informatics (ICCCI); 2019; Coimbatore, India. p. 1–6.

[3] Sinha N, Barua J, Niyogi R, et al. Time-aware relation extraction for entities using news headlines. International Conference on Advances in Computing, Communications and Informatics(ICACCI); 2017; Udupi, India. p. 86–90.

[4] Liu X, Rujia G, Liufu S. Internet news headlines classification method based on the N-Gram language model. 2012 International Conference on Computer Science and Information Processing (CSIP); 2012; Xi'an, Shaanxi. p. 826–828.

[5] Azzopardi G, Greco A, Vento M. Gender recognition from face images with trainable COSFIRE filters. 13th IEEE International Conference on Advanced Video and Signal Based surveillance(AVSS); 2016; Colorado Springs, CO, USA. p. 235–241.

[6] Karthika Devi MS, Umaa Mahesswari G, Baskaran R. Dialogue extraction and translation from stories on Thirukural using verb cue quote content source identifier. ICT with Intelligent Applications; 2022; Ahmedabad, India. p. 525–537.

[7] Karthika Devi MS, Shahin Fathima S, Baskaran R. SYNC—Short, Yet Novel Concise Natural language description: Generating a short story sequence of album images using multimodal network. ICT Analysis and Applications. Lecture Notes in Networks and Systems. Vol. 93. 2020; Goa, India. p. 235–245.

[8] Karthika Devi MS, Shahin Fathima S, Baskaran R. CBCS-Comic Book Cover Synopsis: Generating synopsis of a comic book with unsupervised abstractive dialogue. 9th World Engineering Education Forum, Procedia Computer Science. Vol. 172. 2020; Chennai, India. p. 701–708.

[9] Alzubi JA, Jain R, Nagrath P, et al. Deep image captioning using an ensemble of CNN and LSTM based deep neural networks. J Intell Fuzzy Syst. 2021;40(4):5761–5769. doi:10.3233/JIFS-189415

[10] Chen J, Xiao K. Perception-oriented online news extraction. Proceedings of the 8th ACM/IEEE-CS joint

conference on Digital libraries; 2008; PA Pittsburgh, USA. p. 363–366.

[11] Alzubi JA, Jain R, Kathuria A, et al. Paraphrase identification using collaborative adversarial networks. J Intell Fuzzy Syst. 2020;39(1):1021–1032. doi:10.3233/JIFS-191933

[12] Tan Z, He C, Fang Y, et al. Title-based extraction of news contents for text mining. IEEE Access. 2018;6:64085–64095. doi:10.1109/ACCESS.2018.2877592

[13] Dandeniya D. An automatic e-news article content extraction and classification. 18th International Conference on Advances in ICT for Emerging Regions (ICTer); 2018; Colombo, Sri Lanka. p. 196–202.

[14] Mamgain S, Balabantaray RC, Das AK. Author profiling: prediction of gender and language variety from document. International Conference on Information Technology (ICIT); 2019; Bhubaneshwar, India. p. 473–477.

[15] Ito K, Kawai H, Okano T, et al. Age and gender prediction from face images using convolutional neural network. Asia- Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); 2018; Honolulu, HI, USA. p. 7–11.

[16] Varnima EK, Ramachandran C. Real-time gender identification from face images using you only look once (yolo). 4th International Conference on Trends in Electronics and Informatics (ICOEI); 2020; Thirunelveli, India. p. 1074–1077.

[17] Jain A, Huang J, Fang S. Gender identification using frontal facial images. 2005 IEEE International Conference on Multimedia and Expo; 2005; Amsterdam, Netherlands. p. 4.

[18] Zhou Y, Ni H, Ren F, et al. Face and gender recognition system based on convolutional neural networks. 2019 IEEE International Conference on Mechatronics and Automation (ICMA); 2019; Tianjin, China. p. 1091–1095.

[19] Mittal S, Mittal S. Gender recognition from facial images using convolutional neural network. Fifth International Conference on Image Information Processing (ICIIP); 2019; Shimla, India. p. 347–352.

[20] Chen Q, Wu H, Yachida M. Face detection by fuzzy pattern matching. Proceedings of IEEE International Conference on Computer Vision; 1995; Cambridge, MA, USA. p. 591–596.

[21] Moghaddam B, Yang MH. Gender classification with support vector machines. Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580); 2000; Grenoble, France. p. 306–311.

[22] Ng CB, Tay YH, Goi BM. A review of facial gender recognition. Pattern Anal Appl. 2015;18(4):739–755. doi:10.1007/s10044-015-0499-6

[23] Chen Q, Wu Q, Chen J, et al. Scripted video generation with a bottom-up generative adversarial network. IEEE Trans Image Process. 2020;29:7454–7467. doi:10.1109/TIP.2020.3003227

[24] Liu Y, Wang X, Yuan Y, et al. Cross- Modal Dual Learning for sentence-to -video generation. Proceedings of the 27th ACM International Conference on Multimedia (MM '19). France: Association for Computing Machinery; 2019. p. 1239–124711.

[25] Li Y, Min M, Shen D, et al. Video generation from text. AAAI. 2018;32(1).doi:10.1609/aaai.v32i1.12233

[26] Kim D, Joo D, Kim J. TiVGAN: text to image to video generation with step-by-step evolutionary generator. IEEE Access. 2020;8:153113–153122. doi:10.1109/ACCESS.2020.3017881

[27] Haruechaiyasak C, Jitkrittum W, Sangkeettrakarn C, et al. Implementing news article category browsing based on text categorization technique. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology; 2008; Sydney, NSW. p. 143–146.

[28] Li Z, Shang W, Yan M. News text classification model based on topic model. IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS); 2016; Okayama, Japan. p. 1–5.

[29] Wu H, Yokoyama T, Pramadihanto D, et al. Face and facial feature extraction from color image. Proceedings of the Second International Conference on Automatic Face and Gesture Recognition; 1996; Killington, VT, USA. p. 345–350.

[30] Brunelli R, Poggio T. Face recognition: features versus templates. IEEE Trans Pattern Anal Mach Intell. 1993;15(10):1042–1052. doi:10.1109/34.254061

[31] Korshunova I, Shi W, Dambre J, et al. Fast Face-Swap using convolutional neural networks. 2017 IEEE International Conference on Computer Vision (ICCV); 2017; Venice. p. 3697–3705.

[32] Xingjie Z, Song J, Park J. The image blending method for face swapping. 4th IEEE International Conference on Network Infrastructure and Digital Content; 2014; Beijing, China. p. 95–98.

[33] Sadu C, Das PK. Swapping face images based on augmented facial landmarks and its detection. IEEE Region 10 Conference (TENCON); 2020; Osaka, Japan. p. 456–461.

[34] Mahajan S, Chen L, Tsai T. SwapItUp: A face swap application for privacy protection. IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)2017; Taipei, Taiwan. p. 46–50.

[35] Chen R, Ni B. Face swapping with limited input. International Conference on Intelligent Computing, Automation and Systems (ICICAS); 2019; Chongqing, China. p. 40–43.

[36] Nirkin X, Masi I, Tran Tuan A, et al. On face segmentation, face swapping, and face perception. 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018); 2018; Xi'an, China. p. 98–105.

[37] Matsuhashi S, Nakamura O, Minami T. Human-face extraction using modified HSV color system and personal identification through facial image based on isodensity maps. Proceedings 1995 Canadian Conference on Electrical and Computer Engineering. Vol. 2. 1995; Montreal, QC, Canada. p. 909–912.

[38] Sushama M, Rajinikanth E. Face recognition using DRLBP and SIFT feature extraction. 2018 International Conference on Communication and Signal Processing (ICCSP); 2018; India. p. 994–999.

[39] Wu J, Tse R, Shapiro LG. Automated face extraction and normalization of 3d mesh data. 36th Annual International 35 Conference of the IEEE Engineering in Medicine and Biology Society; 2014; Chicago, IL, USA. p. 750–753.

[40] Bhuvaneshwari R, Geetha P, Karthika Devi MS, et al. A novel deep learning SFR model for FR-SSPP at varied capturing conditions and illumination invariant. Congress on Intelligent Systems: Proceedings of CIS 2021. Vol. 114. 2022; India. p. 407.

[41] Smitha E, Sendilkumar S, Hepsibah Sharon C, et al. Effective emotion recognition from partially occluded

facial images using deep learning. International Conference on Computational Intelligence in Data Science. Vol. 578. Chennai, India. Cham: Springer; 2020. p. 213–221.

[42] Yin J, Li Y, Li J. Face feature extraction based on principle discriminant information analysis. 2007 IEEE International Conference on Automation and Logistics; 2007; Jinan, China. p. 1580–1584.

[43] Acero A. An overview of text-to-speech synthesis. IEEE Work-shop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No.00EX421); 2000; Delavan, WI, USA. p. 1.