6

A Study of Multimodal Colearning, Application in Biometrics and Authentication

Sandhya Avasthi[1]*, Tanushree Sanwal[2], Ayushi Prakash[1] and Suman Lata Tripathi[3]

[1]*Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad, India*
[2]*Department of Computer Science and Engineering, Krishna Group of Institution Delhi-NCR, New Delhi, India* Q1
[3]*Department of Computer Science and Engineering, Lovely Professional University, Jalandhar, India*

Abstract

"Multimodality" refers to utilizing multiple communication methods to comprehend our environment better and enhance the user's experience. Using multimodal data, we may provide a complete picture of an event or object by including new information and perspectives. Improvements in single-mode apps' performance have been possible thanks to developments in deep learning algorithms, computational infrastructure, and massive data sets. Using many modalities is superior to using a single modality, according to research dating back to 2009. The study explains the limitations of single biometric-based methods in providing security and efficiency. The multimodal architecture is based on different forms of data, such as video, audio, images, and text. Combining these kinds of data is utilized to help people learn and imitate. We provide discussions on various methods to fuse different modalities of data. Recent studies have shown that cutting-edge deep-learning techniques can give even better results in multimodal biometrics and authentication systems on mobile devices. The chapter explains different problems in multimodal colearning, various multimodal fusion methods, existing challenges, and future directions.

***Keywords*:** Multimodality, machine learning, multimodal colearning, speech recognition, multimodal biometrics, deep learning, fusion levels

**Corresponding author*: sandhya_avasthi@yahoo.com

Sandeep Kumar, Deepika Ghai, Arpit Jain, Suman Lata Tripathi and Shilpa Rani (eds.)
Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, (103–128) © 2023 Scrivener Publishing LLC

6.1 Introduction

From the beginning, human cognitive development depended on multisensory, multimodal perceptions. For instance, a person can learn the meaning of words through visual and acoustic reinforcements along with semantic or syntactic structure. The learning through elements from multisensory experiences can be further applied to a situation where modalities are missing, for example, reading a newspaper [1–3]. A general practice in machine learning is to use unimodal information based on chosen mode after due diligence in researching the domain. However, a better approach would be to apply to education using multimodal information more aligned with human cognitive development. The application of multimodal data for learning can be referred to as multimodal colearning (MCI) in this chapter [4–6]. Naturally, unstructured data from the real world can exist in various modalities, often known as formats, and frequently includes textual and visual material. Researchers in deep learning continue to be motivated by the need to extract valuable patterns from this type of data. The study presented in this chapter investigates multimodal machine learning and colearning and also explores how to develop deep learning models that integrate and mix various forms of visual inputs across different sensory modalities. In addition, it describes multiple approaches and fundamental concepts of deep multimodal learning. According to related surveys [7, 8], general image matching aims to recognize and match the same or similar structure/content from two or more images [9, 10].

The modern world faces difficulties due to a pandemic and numerous other healthcare needs due to its rising life expectancy [11]. As the field of information technology expands exponentially, users' top concerns are security, privacy, and healthcare applications [12, 13]. More inventive patient care is made possible by improved diagnostic technologies, and innovative medical equipment's real-time monitoring of vital signs raises the standard of care. Competent health care aims to inform individuals about their health conditions and treatment options [14–16]. Individuals are better prepared for potential medical emergencies thanks to intelligent healthcare. A remote check-up service is given, which reduces treatment costs and provides medical practitioners with additional options to serve patients in different regions [17]. A robust intelligent healthcare infrastructure is required to ensure patients' access to necessary medical care as smart cities proliferate. Every year, many computer vision researchers work on making systems that let machines act like humans. Using computer vision technology to map their behavior, intelligent devices like mobile phones

can find obstacles and track locations [18, 19]. Complex operations can be automated in multimodal applications, including computer vision applications. The main challenge of this research is to extract visual attributes from one or more data streams (also called "modalities") with different shapes and sizes. This is done by learning to combine extracted heterogeneous features and project them into a common representation space. This is called "deep multimodal learning." In many situations, a mix of different cues from different modalities and sensors can give context-relevant information about a single activity [20, 21]. In multimodality, a modality's place in conceptual architecture is determined by the media and the qualities that make it up. Some of these modalities are textual, visual, and auditory. They use specific methods or procedures to encode different kinds of information in a way that makes sense [22].

6.1.1 Need for Multimodal Colearning

Multimodal applications incorporate information from several sources at the signal or semantic levels, making them more accurate and dependable than single-modality applications. Applying knowledge gained through one (or more) modalities to tasks involving a different one is the goal of colearning. This typically involves learning a joint representation space, learning external modalities during training, and evaluating the cooperative model's suitability for unimodal tasks. The fusion of multiple data sources is referred to as multimodal fusion [23, 24]. Multimodal systems are those that facilitate communication between users via a variety of channels. An additional definition of multimodality is the capacity of a plan to do automated information processing and to communicate in more than one mode. Six different relationships exist between the modalities: equivalence, transfer, specialization, redundancy, complementarity, and concurrency. "put-that-there" was the first system developed in the 1980s to investigate multimodal systems. This system made inferences about the user's context based on their voice and the cursor's position [25, 26].

As explained, colearning is crucial to maximizing the effectiveness of applications in real-world multimodal. Since currently mobile devices, physiological devices, cameras, medical imaging, and all kinds of sensors are available quickly, multimodal data collection is relatively easy now. Nowadays, multimodal applications range from practical computing, decision-making, control systems, multimedia, autonomous systems, medical devices, military equipment, and satellite systems [27, 28]. The multimodal systems used in these contexts must be dependable and capable

of producing accurate predictions in imperfect signals or environmental variation, as shown in Figure 6.1. Doing so will prevent potentially fatal or otherwise disastrous outcomes. Multimodal machine learning aims to develop models capable of processing and connecting input from multiple sources. Multimodal machine learning is not limited to only audiovisual speech recognition applications; it is used in language and computer vision applications, indicating its enormous potential [29, 30].

Using supplementary or auxiliary information, multimodal data enables us to explain things or phenomena from various perspectives or angles. Applications using a single modality have achieved substantially higher performance thanks to developments in deep learning techniques, computer architecture, and massive data sets [31]. Research from 2009 [1] showed that using multiple senses rather than one can improve performance. Recent research has shown that the most recent deep learning methods lead to additional improvements. Consequently, multimodal machine learning and deep learning are becoming increasingly important.

6.1.2 Why Multimodal Biometric Systems?

Individual identification is vital to biometric authentication, security management, and video surveillance systems. Primary physical biometrics that identifies people include the face, iris, and fingerprints [16]. However, using any of them effectively in an open environment with a typical surveillance system is complex. Facial biometrics captured at a distance, for instance, are

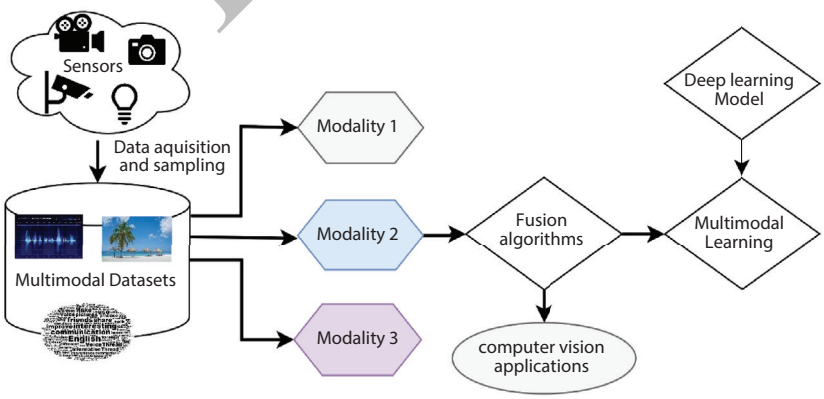

Figure 6.1 A simple pipeline of a multimodal system.

unsuccessful due to the low quality of the face photos. The only biometrics insensitive to space and the quality of the capturing equipment is a person's gait characteristics, which cannot be imitated easily. However, it imposes certain restrictions on dress, carrying cases, and environmental variables [17]. The recognition efficiency of a unimodal biometric is diminished due to numerous difficulties in acquiring feature patterns. A multimodal biometrics surveillance system allows for more precise information extraction than unimodal systems. Diverse fusion-level techniques are employed for the merging of information from various modalities [32, 33].

Although machine learning (ML) techniques frequently extract biometric features and classify objects from raw data, they could perform better in feature discrimination and selection tasks across multiple application domains. Artificial neural networks (ANN) with several hidden layers are used in deep learning (DL), a recent branch of machine learning, to extract data from the lowest level to the most abstract. DL techniques include flexible feature learning, dependable fault tolerance, and robustness features [18]. In recent years, deep convolutional neural networks (deep CNN) have been used in biometric recognition systems [19].

6.1.3 Multimodal Deep Learning

Despite significant advancements, not all facets of human learning have been incorporated into unimodal learning. Multimodal learning enhances comprehension and analysis by actively involving multiple senses in processing information [34–36]. A wide range of media is examined in this paper, including body language, facial expressions, physiological signals, images, videos, text, and audio. Along with a thorough analysis of the foundational approaches, a detailed analysis of recent developments in multimodal deep learning applications over the previous 5 years (2017–2021) has been given. A fine-grained taxonomy of multiple multimodal deep learning approaches is published, focusing on the applications. Finally, the primary concerns for each domain are described separately, along with possible future research directions [20, 21].

This chapter examines multimodal colearning from every conceivable viewpoint, current state, obstacles, data sets, and potential uses. This first effort extends the colearning taxonomy beyond the parallelism of data depicted in Figure 6.2 and into the realm of multimodal colearning [37–39]. We analyzed the preexisting categories, created new ones based on current research and introduced the most up-to-date frameworks that accommodate multimodal colearning and modality circumstances throughout the learning and assessment processes.

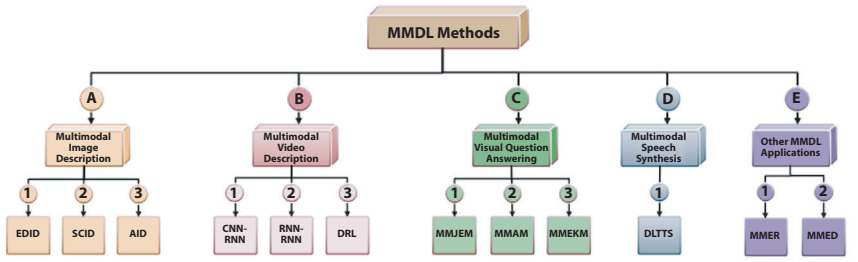

Figure 6.2 Different applications of multimodal deep learning.

6.1.4 Motivation

Recent widespread use of low-cost sensors has led to an explosion of visual data, which has improved the performance of a wide range of computer vision applications (see Figure 6.1). These visual data can be still photos, video sequences, etc., and they can be used to build multimodal models. In contrast to a static image, a video stream contains much meaningful information that considers how successive frames look in space and time. This makes it easy to use and analyze for real-world applications, such as facial expression recognition [22] and video synthesis and description [23]. The term "spatiotemporal concept" refers to analyzing video clips with different lengths in space and time. Multimodal learning analytics combine a video clip's audiovisual and textual features into a single part [40, 41].

The chapter will introduce multimodal colearning, classifying it according to the issues it addresses and the applications it enables. Section 6.1 provides an overview of Multimodal learning, Multimodal Deep Learning, application areas, and motivation for writing this chapter. Section 6.2 gives an outline of Multimodal Deep Learning and its different applications. The Deep Learning Architecture and various techniques are described in Section 6.3. Section 6.4 provides an overview of fusion levels in Multimodal Systems. Section 6.5 gives an outline of a multimodal authentication system in mobile devices. The sixth section discusses challenges, issues, and open problems related to multimodal learning. Concluding remarks and future scope are presented in section seven.

6.2 Multimodal Deep Learning Methods and Applications

In the case of single modalities, applications based on text, images, or audio deep learning models have been successfully applied. Many applications use data in

Table 6.1 Description of different multimodal learning applications.

SN	Application	Full name	Description
1	MMDL	Multimodal Deep Learning	It focuses on developing models that combine multiple data modes with varying structures.
2	EDIT	Encoder-Decoder-Based Image Description	After reading the input photo, a network model that decodes the photo's content into a fixed-length vector.
3	SCID	Semantic Concept-based Image Description	The concept layer is primarily responsible for resolving the meaning expressed by images via scene, knowledge, and emotion.
4	AID	Attention-based Image Description	The program was able to generate each word of the caption by paying attention to the area of the image that was the most important.
5	DRL	Deep Reinforcement Learning	Combines reinforcement learning and deep learning.
6	MMJEM	Multimedia Joint-embedding Models	Joint embedding aims to develop a model representing different media types in a single format.
7	MMAM	Multimodal Attention-based Models	fusion of multiple modalities, where each modality has its sequence of feature vectors.
8	MMEKM	Multimodal External Knowledge-Based Models	Knowledge evaluation and verification can be made more accessible with the help of multi-source knowledge reasoning.
9	DLTTS	Deep Learning Text to speech	to figure out how to use the audio input to guess what the words and sentences said.

(*Continued*)

Table 6.1 Description of different multimodal learning applications. (*Continued*)

SN	Application	Full name	Description
10	MMER	Multimodal Event Recognition	Multimodal social event detection finds events in vast amounts of data, like words, photos, and video clips.
11	MMED	Multimodal Emotion Detection	Combining different modalities offered an excellent viewpoint and successfully revealed hidden emotions from perceptible sources.

various forms to improve features, and those applications are based on multimodal deep-learning techniques. Table 6.1 summarizes multiple Multimodal Deep Learning applications, and the detail is provided in subsections.

6.2.1 Multimodal Image Description (MMID)

The primary purpose of image description is to produce a textual description of the visual information contained in an input image. Deep learning-era picture descriptions are conducted by combining CV and NLP. This process makes excellent use of both text and image [42, 43]. Figure 6.3 shows the visual description's general structure diagram. There are three types of image description frameworks. They are based on retrieval, templates, and description logic (DL). Two of the first ways to describe an image's visual information are "retrieval" and "template-based." This article has three DL-based approaches to describing pictures: encoder-decoder-based, semantic concept-based, and attention-based. Frameworks based on retrieval, templates, or deep learning can all be used to describe images. One of the oldest ways [25, 26] was to use a template to get visual data from a picture and describe it. This article offers a thorough analysis of DL-based methods for image description. These techniques are further divided into encoder-decoder, semantic concept, and attention-based.

6.2.2 Multimodal Video Description (MMVD)

Like image description, video description creates a textual description of what is visible in an input video. This section discusses in depth how DL can be utilized to describe the visual content of videos. When conditions

improve in this field, they can be used in various ways. Video stream and text are the two primary modalities utilized in this procedure. This study categorizes video description methods using the following architectural combinations to extract visual features and generate text [27].

The majority of early works on visual description focused on describing still images. Early attempts at providing automated video descriptions relied on a two-stage pipeline that first recognizes semantic visual concepts before stitching them together in a "subject, verb, object" template. Although a template-based solution separates the tasks of idea identification and description development, such templates need to recreate the language richness found in human-generated descriptions of films or situations [44, 45].

6.2.3 Multimodal Visual Question Answering (MMVQA)

VQA is a multimodal task that aims to correctly produce a natural language response as output after being presented with an image and a related natural language question. VQA is a new method that interests both the CV and NLP communities. It focuses on creating an artificial intelligence (AI) system that can answer questions in natural language [28]. It involves understanding and connecting the image's content to the question's context. VQA involves a diverse set of CV and NLP sub-problems due to the need to compare the semantics of information present in both modalities (the image and the natural language question related to it) (such as object detection and recognition, scene classification, counting, and so on). This means that it is a problem that can be solved entirely by artificial intelligence. Figure 6.3 displays three examples of images and accompanying questions [46, 47].

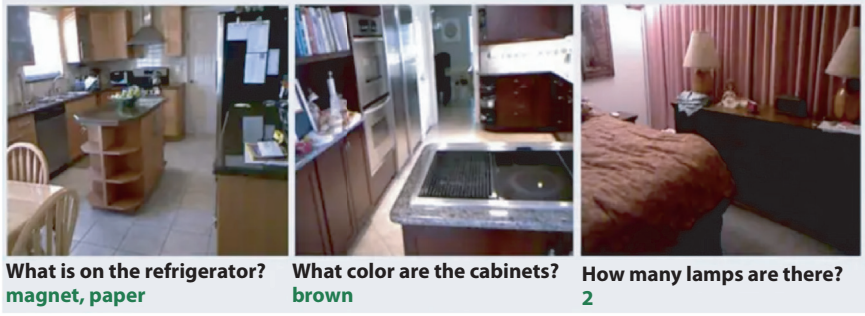

Figure 6.3 Sample examples of images and questions-answer (accessed from https://medium.com/data-science-at-microsoft/visual-question-answering).

6.2.4 Multimodal Speech Synthesis (MMSS)

Human behavior is comprised of two forms of communication: writing and speaking. Speech synthesis refers to the complex process of creating natural language spoken by a machine. Speech synthesis, also called TTS, converts text data into standardized, natural speech in real-time. It encompasses numerous academic disciplines, such as computer science, linguistics, digital signal processing, and acoustics. It is a cutting-edge information processing technology [29], especially for modern intelligent speech interaction systems. Early efforts to create speech synthesis technology heavily relied on parametric synthesis methods. Wolfgang von Kempelen, a Hungarian scientist, invented a device that could synthesize simple words in 1971. It uses a series of delicate bellows, springs, bagpipes, and resonance boxes.

Examples of speech synthesis in use today include screen readers, talking toys, talking video games, and human-computer interactive systems. The imitation of human speech is currently TTS systems' main research goal [30]. The effectiveness of the TTS system is assessed in several ways by using the quality of generated speech timing structure, rendering emotions, and pronunciation, quality of each word produced, synthetic speech preferences (listener preference for a better TTS system in terms of voice and signal quality), and human perception factors like comprehensibility [48].

6.2.5 Multimodal Event Detection (MMED)

Social event detection is the analysis of actual events in massive amounts of social media data never before seen. Even if the single-media-focused efforts produced satisfactory results, the current environment makes them difficult to manage because social media sites frequently host large amounts of multimodal data. Thanks to the widespread use of media sharing on the Internet, individuals can share their events, activities, and thoughts at any time. Multimodal event detection (MMED) systems attempt to recognize actions and occurrences in various media, including images, videos, audio files, text documents, etc. According to statistics, millions of tweets are sent daily, while more than 30,000 hours of video are uploaded to YouTube every hour. Finding events and activities within this volume of data is a complicated task. It has numerous applications in fields such as disease monitoring, governance, and business, and it enables internet users to comprehend and track global events [31, 32].

Whether or not a message input is part of a social event is determined by event inference, a stage of event discovery. Several works have been

inspired by single-modal social event detection works to directly convert non-textual media into textual tags and then use conventional methods for multimodal social event detection. The "media gap"—a situation where descriptions of various media types are inconsistent and cannot be directly measured—between different modalities makes multimodal social event detection difficult. In any event, detection system effectiveness measures how well inference is made. The inference mechanism in such a system is grouped according to social event attributes [49].

6.2.6 Multimodal Emotion Recognition

Emotions are one way that people show how they feel. Multimodal Emotion Recognition (MMER) is very important for improving the way people and computers work together. Machine learning aims to let computers learn and recognize new inputs from training data sets. Because of this, it can be used to effectively train computers to detect, analyze, respond to, interpret, and recognize human emotions. So, the main goal of affective computing is to give machines and systems emotional intelligence. It wants to learn about learning, health, education, communication, gaming, a custom user interface, virtual reality, and data retrieval. The AI/ML model prototype extracts emotional information considering different modalities, for instance, image, text, video, body gesture, body position, facial expression and other forms of data. Using facial expressions and EEG (electroencephalogram) signals, the paper [33, 34] developed a fusion method for figuring out how someone is feeling. A neural network classifier can distinguish between happy, neutral, sad, and afraid feelings.

6.3 MMDL Application in Biometric Monitoring

Multimodal biometric systems identify and verify individuals based on many physiological features. The system stores a person's fingerprint patterns, face geometry, and iris patterns for user identification. Keeping a person's numerous physiological traits is suitable when it is crucial to preserve sensitive data [50].

6.3.1 Biometric Authentication System and Issues

Knowledge-based (based on something the user knows), possession-based (based on something the user possesses), and biometric-based are the three primary methods by which a user can be authenticated

and verified (something a user is). IT systems have widely adopted the first two methods, even though they have several well-known drawbacks. Using a person's unique biological and behavioral characteristics for authentication has become increasingly common [35, 36]. Physical characteristics (such as fingerprints and facial features) serve as the basis for physiological factors, while behavioral factors (such as gait analysis and keystroke dynamics) reflect an individual's behavior and personality pattern [37].

The authentication procedure begins with collecting unique biometric features, continues with preprocessing, finds the area of focus, uses feature extraction techniques to pull out the predefined characteristics, and finally uses classification algorithms to reach a verdict [38]. In addition, numerous feature extraction and classifier construction strategies are available. You can classify a biometric system as either unimodal or multimodal based on the number of biometric modalities it supports [51]. Making a unimodal biometric system is less complicated because it only requires one identity and verification method. Problems, such as noisy data, poor recognition performance, less accurate results, and spoofing attacks [35–38], are more likely to occur in a unimodal system where the authentication metric acts as a single point of failure. These unimodal biometric systems rely on data from a single source to authenticate a person. Even though unimodal biometric systems have many benefits, they must overcome many challenges:

a) **Intra-class variation:** The biometric information collected during verification will not be the same as the information used to make a template for a person during enrolment. This is called variation within the same class. A biometric system's false Rejection Rate (FRR) increases when a category has many differences.
b) **Noisy data:** Biometric sensors that are sensitive to noise make it hard to match people because noisy data can lead to a false rejection.
c) **Interclass similarities:** Inter-class similarity is when the space of features for more than one person overlaps. A biometric system's false Acceptance Rate (FAR) increases with many class similarities.
d) **Non-universality:** Some individuals cannot provide the required biometric alone due to illness or disability.
e) **Spoofing** threatens unimodal biometrics because it allows the data to be imitated or forged.

Using a multimodal biometric system based on multiple sources of information for personal authentication is the best way to solve these issues with the unimodal biometric system [52].

6.3.2 Multimodal Biometric Authentication System and Benefits

Rather than relying on a single characteristic, a multimodal biometric system uses many or complementary characteristics (such as voice and face features). This makes it considerably more robust and difficult to fool. It has a high identification rate, is less subject to external influences, is more reliable and potent, and is more resistant to spoofing attacks [38]. Since it uses more than two biometric indications for authentication, multimodal biometrics must answer the following questions when merging data from numerous modalities: It is possible to develop a multimodal biometric authentication system by mixing specific parts at specific moments [39]. Choosing which biometric characteristics to combine, such as face and voice, fingerprints, and keystroke dynamics, requires selecting two or more biometric characteristics. How successfully the various biometric components may be integrated depends on when they are fused [53, 34]. This is performed during the pipeline phases of the biometric authentication system. How to unite describes the information's organization. The general multimodal biometrics framework is depicted in Figure 6.4.

A multimodal biometric system will make either a "genuine individual" or "imposter" determination. Fundamentally, the system's accuracy is determined by the genuine acceptance rate (GAR), false rejection rate (FRR), false acceptance rate (FAR), and equal error rate (EER) (ERR). The enrolment phase and the authentication phase are the two primary phases of operation for multimodal biometrics, and each is described as follows:

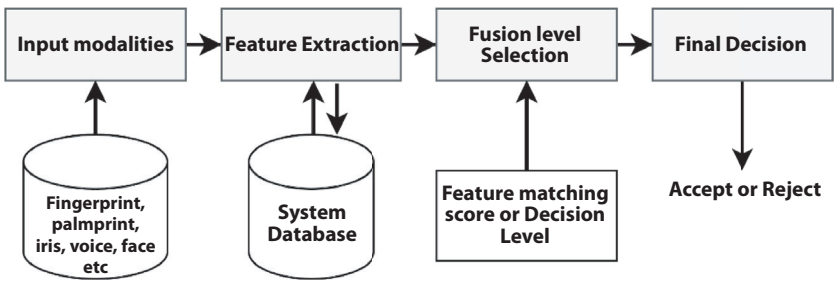

Figure 6.4 A generic process of multimodal biometric system.

- ***Enrolment phase:***
  A user's biometric characteristics are recorded during the enrolment phase and used as a template for that user during the authentication phase by being stored in the system database.
- ***Authentication phase:***
  To verify a user's identity, the system takes another look at their unique set of traits. In identification, data is matched to templates for all users in a database called a "one-to-many" match. In verification, data is only matched to the claimed identity template, called a "one-to-one" match [40].

6.4 Fusion Levels in Multimodal Biometrics

Technically, multimodal fusion refers to merging data from multiple modalities to predict an end measure, either as a constant value (e.g., sentiment positivity) via regression or as a class (happy vs sad) via classification. Interest in multimodal fusion is sparked by its ability to provide three significant advantages [55]. First, having access to many observational modalities that capture the same phenomenon could aid in developing more accurate forecasts. Combining two or more modalities to complete a task is the first step in creating multimodal systems. Fusion techniques are divided into three categories: early (feature), late (decision), and intermediate (hybrid) fusion, depending on the level of the network at which the representations are fused [56, 57]. There are no hard and fast rules for fusion; instead, it is always different depending on the data, the domain, and the objective. Since early fusion does not consider intra-modality features and late fusion does not consider inter-modality particulars, hybrid fusion is the more popular option.

a) ***Early Fusion:*** This merging occurs when the AI model's input data from various sources are combined. Further investigation reveals that the data set is first subjected to the fusion technique before being used as input to the DL algorithm. The fusion process is likely performed on the raw data itself. When raw data undergoes a feature-extraction phase before merging, we say the merging is performed at the feature level.

b) ***Late Fusion:*** The AI algorithm is used before fusing. In this case, data are dealt with uniquely and multimodally. This

method looks at the different ways of doing things as separate streams. The possible conditional links between the other modalities are not considered during the learning process. There are a lot of different ways to merge.

c) ***Intermediate Fusion:*** When the various input data types are combined before and after the relevant AI algorithm is run, it is known as hybrid fusion. This approach may be efficient when combining modalities with similar dimensions or modalities that must be preprocessed before being merged during the training phase.

There are three fusion levels in multimodal biometrics, as described by Jain and Ross [6]: the feature level, the matching score level, and the decision level. It is commonly held that applying the combination scheme as early as possible in the recognition system yields the best results [8, 9]. The following is a breakdown of the three fusion stages:

6.4.1 Fusion at Feature Level

Signals from various biometric characteristics are individually processed, and then their feature vectors are fused into a single vector via the feature-level fusion procedure. The feature vectors are combined to create a composite feature vector for classification in the subsequent step [58]. For feature-level fusion to function, redundant features must be eliminated via reduction techniques. Researchers have utilized fusion at the level of features. Figure 6.5 is a demonstration of feature fusion. The primary advantage of feature-level fusion is the discovery of correlated feature values generated by distinct biometric algorithms. This helps identify a small set of significant features that can improve the recognition's accuracy. Typically, reducing the number of dimensions is required to obtain this

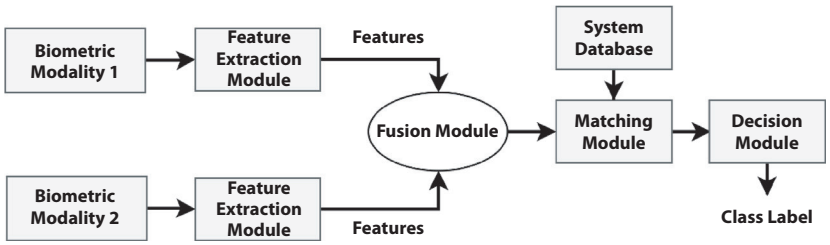

Figure 6.5 Fusion at the feature level.

set of characteristics. Consequently, feature-level fusion typically requires a large amount of training data.

6.4.2 Fusion at Matching Score Level

The feature vectors still need to be put together. Instead, each one is looked at individually to figure out its score [58]. There are many ways to combine match scores, such as logistic regression, highest rank, Borda count and weighted sum, weighted product, Bayes rule, mean fusion, linear discriminant analysis (LDA) fusion, k-nearest neighbour (KNN) fusion, and hidden Markov model (HMM). Normalizing scores from different sources [6] is a critical issue that must be dealt with at the level of Matching scores. The match scores can be normalized with min-max, z-score, median-MAD, double-sigmoidal, tan-h, and piecewise linear. The matching score is the most played fusion level because it is easy. Several researchers [10–12] have used fusion at the Matching score level. Figure 6.6 shows the merging of scores that are the same.

6.4.3 Decision-Level Fusion

In this type of fusion, each modality is independently classified, meaning each biometric attribute is captured that follows the extraction of features from that specific trait. Further, these traits are classified as accept or reject based on the extracted features. The final classification relies on the integration of outputs from numerous modalities. The fusion of decision levels is illustrated in Figure 6.7. Fusion was utilized at the level of decision-making [20]. The advantage of this type is that prediction can be possible even if one of the modality data is unavailable.

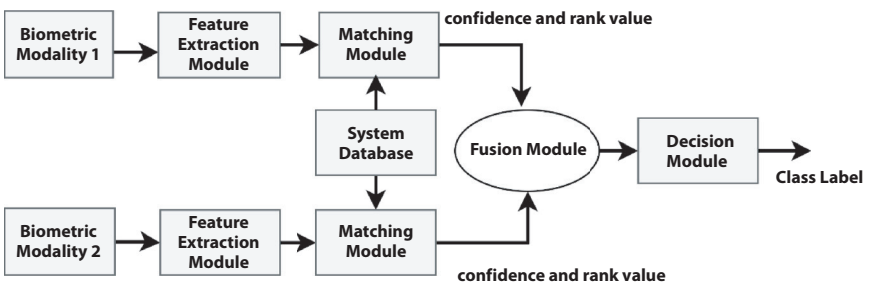

Figure 6.6 Fusion at matching score level.

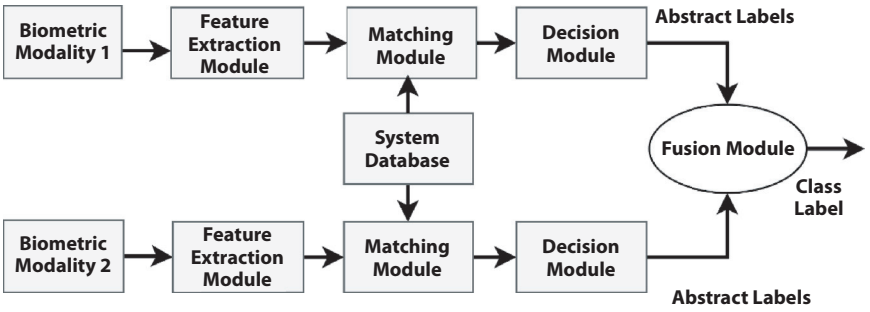

Figure 6.7 Fusion at the decision level.

6.5 Authentication in Mobile Devices Using Multimodal Biometrics

Implementing secure user authentication for mobile devices to protect users' personal information and data is becoming increasingly vital. Due to their enormous benefits over conventional authentication methods, biometric approaches have gained popularity in academics and business. This section discusses the development of existing biometric identification systems on mobile phones, namely touch-enabled devices, concerning eleven biometric methods. The types of user authentication are physiological and behavioral. In general, physiological biometrics refers to a person's physical characteristics, like their fingerprint, face, iris or retina, or hand or palm, whereas behavioral biometrics refers to their behavioral qualities, like their voice, signature, stride, keystroke dynamics, or touch dynamics.

These techniques use the entry-point authentication model, which can be biometric or based on PINs and passwords. The user only needs to be verified at the beginning of the session. Since attacks can happen after the first authentication, the session authentication paradigm has gotten much bad press. Because of these things, a new method of user authentication based on the "something that the user is" paradigm has been suggested. Continuous authentication (CA) and behavioral biometrics (BB) are used in this method [41, 42]. Mobile device sensors can capture most users' behavior quickly and accurately, allowing behavioral biometric user authentication [43]. Mobile device sensors enrol BB templates, including walking style, gestures, dynamics of keystrokes, hand motions, battery usage, and user profiles. With ongoing authentication, BB can give each user something unique. CA technology adds an extra layer of security on top of the login process by keeping an eye on what users do and frequently

re-verifying their identities during a session. CA was first thought about in the early 2000s. Since then, the business and academic worlds have become more interested in this technology. People are becoming more interested in BB and CA technology because sensor costs are expected to decrease, systems are improving, and there is political pressure for stricter security controls. People are eager to use biometric authentication solutions to protect their privacy.

6.5.1 Categories of Multimodal Biometrics

Some popular categories of BB and CA are described in this section. Some common biometric patterns are touch gestures, keystrokes dynamics, behavioral profiling, the gait of a person, and hand waving. In addition, we examine how behavioral biometrics are collected and how features are extracted.

- **Walking gait**—Smartphones' accelerometer, gyroscope, and magnetometer sensors allow them to recognize walking patterns. The main advantage of this method is that users' CAs can be deployed without their involvement. The device's orientation moving when walking, uneven ground, potential injuries, footwear, weariness, human features, etc., can all reduce accuracy. The accelerometer records information about people walking normally, slowly, and quickly. The participant's smartphone's orientation in their pocket is estimated using gyroscope data. One can calculate the movement of humans by integrating sensory data from the accelerometer, magnetometer, and gyroscope.
- **Touch gestures—**Recent mobile phones and other intelligent devices are touch-enabled, which means one can draw shapes on the touch screen using one or more strokes. Each stroke is composed of a series of numerical coordinates. The direction and duration of touches, movement velocity, and acceleration are analyzed and measured individually or in combination. They are utilizing a smartphone's touch screen sensor to collect touch data. A gesture output template is generated from input actions using speed, velocity, size, length, and direction variables. These factors vary among users and represent their unique behaviors, making them the foundation of touch gesture authentication systems.

- **Keystroke dynamics—**Keystroke dynamics is recording a user's keyboard inputs on a mobile device and attempting to recognize him by analyzing his tapping patterns. Some studies on keystroke dynamics collect information from specific texts, like writing text messages or entering passwords during a login session. Others obtain data for research purposes without using passwords or particular phrases. The outcomes are precise in both cases.
- **Behavioral profile—**Based on the idea that people use their phones in a certain way when they use apps and digital services, manipulating data from a mobile device can verify an individual's behavior. A profile of a user's behavior could be made based on how he interacts with hosts or a network. In the first scenario, users' connecting patterns to Wi-Fi networks, service providers, etc., are watched. In the second scenario, users' use of apps at different times and places is observed. Data about a device's use can be combined to make user profiles. The paper [45] used the self-created behavioral mobile application Track Maison to find out how people used five social networking sites, such as their location, the length of their sessions, and how often they used them.
- **Hand waving—**People are paying more attention to how a person's wrist moves when using or just holding a mobile phone. This method does not need the user to do anything other than hold the device. There are several ways to use it, such as twisting your wrist, waving fast, waving far, or waving often. Different people can be told apart by how they wave their hands [45].

6.5.2 Benefits of Multimodal Biometrics in Mobile Devices

Implementing multimodal biometrics on mobile devices is feasible, as many already support face, voice, and fingerprint recognition. A robust, user-friendly strategy is required for these technologies to be consolidated. In the mobile consumer market segment, multimodal biometrics is a popular authentication method with multiple benefits.

- **Mobile security.** Attackers can take down unimodal biometric systems by spoofing the system's single biometric modality. Attackers must simultaneously impersonate numerous

distinct human characteristics to establish identity-based on multiple modalities, which is more complicated.

- **Mobile authentication.** One specific modality can be used to improve the quality issues in other modalities' results. For instance, Proteus evaluates the face-image and voice recording quality and gives more weight to the sample with the highest quality.
- **Accuracy**- When multimodal biometrics are used, they make it much easier to identify a person.
- **Universality**- A multimodal biometric system works for everyone, even if a person is sick or disabled and cannot give one type of biometric. Instead, the system can use a different biometric to verify the person's identity.

6.6 Challenges and Open Research Problems

The data is highly diverse, making Multimodal Machine Learning a challenging area of computational study. Understanding natural processes on a deeper level and capturing correspondences between modalities are made possible by learning from multimodal sources. This paper identifies and explores five primary technological obstacles (and sub-challenges) associated with multimodal machine learning. The following five difficulties make up the basis of our taxonomy, which extends beyond the conventional division between early and late fusion.

a) **Representation**-The first challenge in taking advantage of multimodal data is describing and summarizing it to improve the learning process. The variety of multimodal data makes it challenging to create such representations. Language, for instance, frequently reflects symbolic aural and visual modalities, whereas signals do not.
b) **Translation**-The second obstacle is mapping (translating) data from one modality to another. In addition to the diverse data, the relationship between the modalities is frequently vague or subjective. For example, several accurate ways exist to describe an image, yet there may not be a perfect translation.
c) **Alignment**-The second impediment is determining how to map (translate) data from one modality to another. Aside from the heterogeneous data, the relationship between the

modalities could be more transparent and subjective. For example, several accurate ways exist to describe an image, but no perfect translation exists.

d) Fusion-The fourth issue is combining information from two or more modalities for forecasting. In audiovisual speech recognition, for instance, the speech signal and the visual description of lip motion are merged to predict uttered words. The predictive capacity and noise structure of the information received from many modalities may differ, and at least one may have missing data.

e) Colearning-The transfer of information between different modalities is complex, and so is representing multimodal data. Cotraining, conceptual grounding, and zero-shot learning are examples of such algorithms. Colearning studies how information gleaned from one modality may help a computer model created with a different modality. This issue is especially critical when one modality has limited resources (like annotated data).

6.7 Conclusion

The way something occurs or is experienced is referred to as its modality. Artificial intelligence must be able to process all of these various types of information concurrently to learn more about the environment around us. The main objective of multimodal machine learning is to utilize data in multiple forms to provide improved results. The newest changes to MMDL and brand-new ideas were covered in this chapter. In addition, the chapter reviews numerous applications using various modalities, including body gestures, facial expressions, physiological signals, images, audio, and video. This review contrasts it with earlier surveys of a similar nature. An overview of Multimodal biometric systems, ideas, and unresolved biometric security issues is given in this chapter. Multimodal biometrics should be the next step for mobile consumer device biometric authentication. Implementing multimodal biometrics on standard mobile devices has been little advancement.

References

1. Meltzoff, A.N., Origins of the theory of mind, cognition and communication. *J. Commun. Disord.*, 32, 4, 251–269, 1999.

2. Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J., Image matching from handcrafted to in-depth features: A survey. *Int. J. Comput. Vis.*, 129, 23–79, 2021.
3. Zhou, H., Sattler, T., Jacobs, D.W., Evaluating local features for day-night matching, in: *Proceedings of the European Conference on Computer Vision*, Springer, pp. 724–736, 2016.
4. Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L., ContextDesc: Local descriptor augmentation with cross-modality context, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2527–2536, 2019.
5. Zhou, H., Ma, J., Tan, C.C., Zhang, Y., Ling, H., Cross-weather image alignment via latent generative model with intensity consistency. *IEEE Trans. Image Process.*, 29, 5216–5228, 2020.
6. Naseer, T., Spinello, L., Burgard, W., Stachniss, C., Robust visual robot localization across seasons using network flows, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2564–2570, 2014.
7. Aubry, M., Russell, B.C., Sivic, J., Painting-to-3D model alignment via discriminative visual elements. *ACM Trans. Graph.*, 33, 2, 1–14, 2014.
8. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F., Multimodality cross attention network for image and sentence matching, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10941–10950, 2020.
9. Avasthi, S. and Sanwal, T., Biometric authentication techniques: A study on keystroke dynamics. *International Journal of Scientific Engineering Applied Science (IJSEAS)*, *2*, 1, 215–221, 2016.
10. Gupta, A. and Avasthi, S., An image-based low-cost method to the OMR process for surveys and research. *International Journal of Scientific Engineering Applied Science (IJSEAS)*, *2*, 7, 91–95, 2016.
11. Avasthi, S., Chauhan, R., Acharjya, D.P., Information extraction and sentiment analysis to gain insight into the COVID-19 crisis, in: *International Conference on Innovative Computing and Communications*, pp. 343–353, Springer, Singapore, 2022.
12. Avasthi, S., Chauhan, R., Acharjya, D.P., Topic modeling techniques for text mining over a large-scale scientific and biomedical text corpus. *International Journal of Ambient Computing and Intelligence (IJACI)*, *13*, 1, 1–18, 2022.
13. Avasthi, S., Chauhan, R., Acharjya, D.P., Extracting information and inferences from a large text corpus. *Int. J. Inf. Technol.*, 15, 1, 435–445, 2023.
14. Tiulpin, A., Klein, S., Bierma-Zeinstra, S., Thevenot, J., Rahtu, E., Meurs, J.V., Saarakkala, S., Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Sci. Rep.*, *9*, 1, 1–11, 2019.
15. Mullick, T., Radovic, A., Shaaban, S., Doryab, A., Predicting depression in adolescents using mobile and wearable sensors: Multimodal machine learning–Based exploratory study. *JMIR Form. Res.*, *6*, 6, e35807, 2022.

16. Buddharpawar, A.S. and Subbaraman, S., Iris recognition based on PCA for person identification. *Int. J. Comput. Appl.*, 975, 8887, 2015.
17. Han, J. and Bhanu, B., Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28, 316–322, 2005.
18. Wang, P., Fan, E., Wang, P., Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognit. Lett.*, 141, 61–67, 2021.
19. Boucherit, I., Zmirli, M.O., Hentabli, H., Rosdi, B.A., Finger vein identification using deeply-fused convolutional neural network. *J. King Saud Univ. Comput. Inf. Sci.*, 34, 346–656, 2020.
20. Belo, D., Bento, N., Silva, H., Fred, A., Gamboa, H., ECG biometrics using deep learning and relative score threshold classification. *Sensors*, 20, 15, 4078, 2020.
21. Mekruksavanich, S. and Jitpattanakul, A., Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models. *Electronics*, 10, 3, 308, 2021.
22. Zarbakhsh, P. and Demirel, H., 4D facial expression recognition using multimodal time series analysis of geometric landmark-based deformations. *Vis. Comput.*, 36, 951–965, 2020.
23. Dilawari, A. and Khan, M.U.G., ASoVS: Abstractive summarization of video sequences. *IEEE Access*, 7, 29253–29263, 2019.
24. Summaira, J., Li, X., Shoib, A.M., Li, S., Abdul, J., *Recent advances and trends in multimodal deep learning: A review*, 2021, https://arxiv.org/abs/2105.11087.
25. Avasthi, S., Sanwal, T., Sharma, S., Roy, S., VANETs and the use of IoT: Approaches, applications, and challenges, in: *Revolutionizing Industrial Automation Through the Convergence of Artificial Intelligence and the Internet of Things*, pp. 1–23, 2023.
26. Praharaj, S., Scheffel, M., Drachsler, H., Specht, M., Literature review on Co-located collaboration modelling using multimodal learning analytics—Can we go the whole nine yards? *IEEE Trans. Learn. Technol.*, 14, 3, 367–385, 2021.
27. Rahman, M. M., Abedin, T., Prottoy, K. S., Moshruba, A., Siddiqui, F. H., Video captioning with stacked attention and semantic hard pull. *PeerJ Comput. Sci.*, 7, e664, 2021.
28. Lobry, S., Marcos, D., Murray, J., Tuia, D., RSVQA: Visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote Sens.*, 58, 12, 2020, 2020.
29. Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S. *et al.*, *Tacotron: Towards end-to-end speech synthesis*, 2017, https://arxiv.org/abs/1703.10135.
30. Taigman, Y., Wolf, L., Polyak, A., Nachmani, E., *Voiceloop: Voice sitting and synthesis via a phonological loop*, 2018, https://arxiv.org/abs/1707.06588.

31. Huang, S., Huang, D., Zhou, X., Learning multimodal deep representations for crowd anomaly event detection. *Math. Prob. Eng.*, 2018, 1–13, 2018.
32. Koutras, P., Zlatinsi, A., Maragos, P., Exploring cnn-based architectures for multimodal salient event detection in videos, in: *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, IEEE, 2018.
33. Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., Zhou, Y., Deep voice 2: Multi-speaker neural text-to-speech. *Adv. Neural Inf. Process. Syst.*, 30, 2017, 2017.
34. Chauhan, R., Avasthi, S., Alankar, B., Kaur, H., Smart IoT systems: Data analytics, secure smart home, and challenges, in: *Transforming the Internet of Things for Next-Generation Smart Systems*, pp. 100–119, IGI Global, USA, 2021.
35. Al Abdulwahid, A., Clarke, N., Stengel, I., Furnell, S., Reich, C., Continuous and transparent multimodal authentication: Reviewing state of the art. *Cluster Comput.*, 19, 1, 455–474, Mar. 2016.
36. Ayeswarya, S. and Norman, J., A survey on different continuous authentication systems. *Int. J. Biom.*, 11, 1, 67, 2019.
37. Gad, R., El-Fishawy, N., El-Sayed, A., Zorkany, M., Multibiometric systems: A state of the art survey and research directions. *Int. J. Adv. Comput. Sci. Appl.*, 6, 6, 128–138, 2015.
38. Dargan, S. and Kumar, M., A comprehensive survey on the biometric recognition systems based on physiological and behavioural modalities. *Expert Syst. Appl.*, 143, Art. no. 113114, Apr. 2020.
39. Singh, M., Singh, R., Ross, A., A comprehensive overview of biometric fusion. *Inf. Fusion*, 52, 187–205, Dec. 2019.
40. Ross, A. and Jain, A., Information fusion in biometrics. *J. Pattern Recognit. Lett.*, 24, 2115–2125, 2003.
41. Stylios, I.C., Thanou, O., Androulidakis, I., Zaitseva, E., A review of continuous authentication using behavioural biometrics. *Conference: ACM SEEDA-CECNSM*, Kastoria, Greece, 2016.
42. *Biometric authentication: The how and why*, Available: https://about-fraud.com/biometric-authentication, accessed on 21/2/2019.
43. Morency, L.P., Liang, P.P., Zadeh, A., Tutorial on multimodal machine learning, in: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pp. 33–38, 2022, July.
44. Liang, P.P., Zadeh, A., Morency, L.P., *Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions*, 2022, https://arxiv.org/abs/2209.03430.

45. Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J., Multimodal deep learning for biomedical data fusion: A review. *Brief. Bioinformatics*, *23*, 2, bbab569, 2022.
46. Anjomshoa, F., Catalfamo, M., Hecker, D., Helgeland, N., Rasch, A., Kantarci, B., Schuckers, S., Mobile behaviometric framework for sociability assessment and identification of smartphone users, in: *2016 IEEE Symposium on Computers and Communication (ISCC)*, pp. 1084–1089, 2016, June.
47. Kumar, S., Rani, S., Jain, A., Verma, C., Raboaca, M.S., Illés, Z., Neagu, B.C., Face spoofing, age, gender and facial expression recognition using advance neural network architecture-based biometric system. *Sens. J.*, 22, 14, 5160–5184, 2022.
48. Sandeep, K., Jain, A., Agarwal, A.K., Rani, S., Ghimire, A., Object-based image retrieval using the u-net-based neural network. *Comput. Intell. Neurosci.*, 2021, https://www.hindawi.com/journals/cin/2021/4395646/.
49. Kumar, S., Haq, M., Jain, A., Jason, C.A., Moparthi, N.R., Mittal, N., Alzamil, Z.S., Multilayer neural network based speech emotion recognition for smart assistance. *CMC-Comput. Mater. Contin.*, 74, 1, 1–18, 2022. Tech Science Press.
50. Bhola, A. and Singh, S., Visualization and modeling of high dimensional cancerous gene expression dataset. *J. Inf. Knowl. Manag.*, 18, 01, 1950001–22, 2019.
51. Bhola, A. and Singh, S., Gene selection using high dimensional gene expression data: An appraisal. *Curr. Bioinform.*, 13, 3, 225–233, 2018.
52. Rani, S., Gowroju, Kumar, S., IRIS based recognition and spoofing attacks: A review, in: *10th IEEE International Conference on System Modeling &Advancement in Research Trends (SMART)*, December 10-11, 2021.
53. Swathi, A., Kumar, S., Venkata Subbamma., T., Rani, S., Jain, A., Ramakrishna, K. M.V.N.M, Emotion classification using feature extraction of facial expressiona, in: *The International Conference on Technological Advancements in Computational Sciences (ICTACS – 2022)*, Tashkent City Uzbekistan, pp. 1–6, 2022.
54. Rani, S., Lakhwani, K., Kumar, S., Construction and reconstruction of 3D facial and wireframe model using syntactic pattern recognition, in: *Cognitive Behavior & Human Computer Interaction*, pp. 137–156, Scrivener & Willey Publishing House, 2021.
55. Rani, S., Ghai, D., Kumar, S., Kantipudi, M.V.V., Alharbi, A.H., Ullah, M.A., Efficient 3D AlexNet architecture for object recognition using syntactic patterns from medical images. *Comput. Intell. Neurosci.*, 1–19, 2022.
56. Rani, S., Ghai, D., Kumar, S., Reconstruction of simple and complex three dimensional images using pattern recognition algorithm. *J. Inf. Technol. Manag.*, 14, (Special issue: Security and Resource Management challenges for Internet of Things), 235–247, 2022.

57. Bhaiyan, A.J.G., Shukla, R.K., Sengar, A.S., Gupta, A., Jain, A., Kumar, A., Vishnoi, N.K., Face recognition using convolutional neural network in machine learning, in: *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 456–461, IEEE, 2021.
58. Bhaiyan, A.J.G., Jain, A., Gupta, A., Sengar, A.S., Shukla, R.K., Jain, A., Application of deep learning for image sequence classification, in: *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 280–284, IEEE, 2021.

Answer all Queries (Q) in the margin of the text. When requested, provide missing intext reference citation as well as intext reference for figure/table. Please annotate the PDF and read the whole chapter again carefully. Careful proofing is key to optimal output.

Author Query

Q1 Inserted "New Delhi" as city for affiliation 2. Edit OK?