Integration of Visual Features for Assessing Authenticity and Detecting Synthetic Video Content

Sergey Eremeev[1, a) [0000-0001-8482-1479]], Veronika Shikinova[1, b) [0009-0006-0030-9771]], Yulia Anokhina[1, c) [0009-0006-1696-2804]] and Ekaterina Fedoseeva[1, d) [0009-0001-5086-411X]]

[1] Murom Institute of Vladimir State University, Russia
a)sv-eremeev@yandex.ru, b)niks.shik10@gmail.com,
c)uliaanohina73@gmail.com, d) fedoseevakatya2006@mail.ru

Abstract. The paper considers the task of assessing behavioral and physiological features of a person based on video and audio streams. The main focus is placed on computer vision methods which are used to detect gaze direction, blink frequency, eye openness, facial expressions, microexpressions and changes in skin tone. Speech parameters are additionally used which makes the final assessment more complete. Each indicator is calculated using specified formulas, normalized, and compared with the individual baseline obtained at the calibration stage. The paper describes the structure of the software module, the stages of video and audio stream processing, as well as the principles of combining features into a general indicator. The developed program is intended to record features that may indicate tension, unnatural behavior, or a change in the user's emotional state. The program makes it possible to consider visual and speech features in a unified assessment system without being tied to a single parameter. As a result of the program's operation, final numerical scores are generated for individual parameters of the user's state on a scale from 0 to 1. Unlike existing approaches that analyze individual features, the developed program applies a multimodal approach, produces a more stable final assessment and can be used in the task to evaluate the authenticity and detection of synthetic video content.

Keywords: computer vision, facial expressions, gaze direction, blinking, calibration, normalization, behavioral features, integration of features

1 Introduction

The development of online communication systems has radically changed the way we study, take exams, and attend interviews. Today, automatic analysis of human behavior has become relevant far beyond classical remote monitoring systems. For example, in personnel recruitment, such technologies help assess a candidate's engagement during an interview and detect signs of severe stress or, conversely, insincerity, while in education they help ensure the integrity of knowledge assessment. In all these cases, the input data consist of standard multimedia streams

including video images from a webcam and synchronized audio recordings from a microphone.

One of the current trends in image processing is the growth of synthetic content and deepfakes. Any student or job candidate can now generate a fake video or replace their voice in real time. For this reason, relying only on basic visual markers is no longer sufficient. To detect deception in time, the system must instantly compare dozens of microsignals over time. Traditional algorithms usually analyze video and audio separately, which leads to isolated data processing and the loss of the overall picture. In addition, modern heavy neural networks [11] either process the entire recording as a whole or create a significant load on the processor when attempting to combine several models. At the same time, video frames and audio require different processing speeds, which leads to a noticeable decrease in system performance.

Therefore, fundamentally new approaches capable of operating on the fly and without delays are now urgently needed. It is proposed to divide the system into independent parallel blocks that operate simultaneously.

In the proposed approach, each type of data is processed by a separate isolated module. For example, while one block monitors the eyes and blink frequency using facial landmarks, specifically the EAR (Eye Aspect Ratio) method [4] and landmark-based facial alignment techniques [13], another module processes audio in parallel. At the initial stage, root mean square loudness calculation and spectral analysis are used to instantly distinguish silence from speech. If speech is detected, streaming speech recognition is performed using models from the Whisper family[1].

The combination of audio, text, and video is now actively used in multimodal machine learning for joint analysis of heterogeneous data [3]. Our multimodal feature-fusion approach brings together the most important information while filtering out random noise in both audio and video. Dividing the program into independent processes makes it possible to create a fast real-time system. It can effectively detect facial microexpressions and even autonomic reactions, such as short-term skin reddening, known as the flushing effect, caused by stress from difficult questions asked by examiners or HR specialists. As a result, we obtain a flexible architecture that is equally effective for protecting online exams from cheating and for assisting experts in candidate selection.

2 Review

At present, the development of systems that recognize behavior and monitor students during exams is of great interest to researchers [2, 5]. Such technologies make it possible to divide a complex video stream into separate parts and then combine them again in order to understand what a person is doing. The analysis of individual features helps detect moments when a person is distracted, remove random noise, and make monitoring more accurate.

---

[1] https://openai.com/ru-RU/index/whisper/

Classical methods usually divide data into image and sound and evaluate them separately. In this case, behavior analysis takes place in different spaces that are not connected with each other. This approach is useful because it makes it easy to detect obvious noise or track simple features such as gaze direction. At present, classical analysis of facial geometry and deep neural networks are actively combined, for example in real-time face alignment, face detection, and 3D face pose estimation methods [1, 13, 12]. However, such solutions have a drawback: they often analyze the entire process as a whole, which increases computational complexity and prevents the system from adapting to the characteristics of different data channels [3]. At the same time, different behavioral markers have their own unique properties. This creates the problem of how to teach the system to evaluate data in a narrowly focused way while taking these features into account. Such characteristics include, for example, spatio-temporal relationships between human movements.

In the proposed approach, each type of data is processed by a separate module. The eye module analyzes eye openness and blink frequency using facial landmarks and the Eye Aspect Ratio method [4], while the facial expression module uses landmark-based facial geometry and expression-related parameters [13]. This algorithm makes it possible to focus on small details, such as eyelid and pupil movements, which is very important for detecting hidden abnormalities.

Another popular approach is to combine text, speech, and facial expressions for emotion analysis. The end-to-end fusion method collects the most important information reflected in the attention weights of the neural network. The analysis of these coefficients helps effectively remove noise that appears in background audio and video. This approach also makes it possible to create specialized models for face spoofing detection and face detection [6, 7]. These new architectures form the basis for algorithms that take into account the specific features of human behavior, such as facial microexpressions or temporal inconsistency between what a person says and where they are looking at that moment [3].

The aim of the work is to develop a method for integrating visual features when used in the task of assessing authenticity and identifying synthetic video content.

To achieve this aim, the multimodal consistency between visual, physiological, and speech features is investigated. Behavioral and physiological changes are considered as changes in the state of a real user, while technical anomalies are considered as inconsistencies that may occur in generated or modified multimedia content.

3 Methodology

This work proposes the simultaneous analysis of several independent information streams: video from a webcam, audio recording from a microphone, eye coordinate data, facial landmarks, gaze direction, skin color characteristics, and speech data. The processing of all parameters at the same time makes it possible to obtain a more complete picture of the user's state.

For training, calibration, and testing of the developed software modules, a specialized dataset of video images was created. The dataset includes video

recordings of users' faces captured under different lighting conditions and with different head positions relative to the webcam.

The program is divided into modules, each of which processes features for the future multimodal system. The system starts with capturing the input video stream from the camera, after which it is processed frame by frame.

Existing APIs are used to extract primary features from video and audio streams, including facial expression parameters, gaze direction, blink frequency, skin tone changes, and speech characteristics. The main purpose of the module is to form a structured set of features for further analysis.

The proposed approach uses heuristic coefficients to calculate integral features. The optimal values of the coefficients were obtained during the experiments. These values are used for preprocessing and normalization of features within modules.

Each frame is a separate image that can be represented as a pixel matrix:

$$I = (a_{ij})_{n \times m},$$

where I is an image, a_{ij} is the pixel value in the i-th row and j-th column, n, m are the height and width of the image in pixels.

Accordingly, the video stream is an ordered set of images:

$$V = \{I_1, I_2, ..., I_T\},$$

where V is the video stream, I_t is the frame obtained at time t, T is the number of processed frames.

Then the MediaPipe Face Landmarker[2] neural network analyzer determines the key facial points and facial expression coefficients. These data are the main ones for all developed modules.

After processing a frame, the neural network analyzer forms a set of facial landmarks L and facial expression coefficients B:

$$L = \{P_1, P_2, \ldots, P_N\},$$

$$B = \{b_1, b_2, ..., b_Q\},$$

where L_t is the set of facial points in frame t, $P_i = (x_i, y_i)$ are the coordinates of the i-th facial point $(i = 1,2, \ldots, N)$, N is the number of landmarks, B_t is the blendshape vector, b_j is the value of a separate facial expression coefficient $(j = 1,2, \ldots, Q)$.

The audio signal is received in small fragments. For each fragment, the volume level is determined. These fragments are then processed for further use in the speech recognition model:

$$X = \{x_1, x_2, ..., x_M\},$$

where X is the audio signal, x_i is the i-th amplitude sample, M is the total number of samples.

---

[2] https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker

3.1 Facial Expression Analysis Module

The facial expression analysis module is designed for the quantitative assessment of facial movements. The module receives a set of blendshape coefficients as input, which are numerical features reflecting the degree of activation of individual facial areas: smiling, mouth opening, eyebrow raising, frowning, and squinting. They are used to calculate measurable facial expression features that can be associated with specific facial movements.

To calculate blendshapes, the MediaPipe Face Landmarker ML model by Google is used. It first detects the face and 3D facial key points, after which a separate blendshape prediction model calculates 52 facial expression coefficients. The model has 4 MLP-Mixer blocks, hidden layers of 384 and 256, an internal dimension of 64, and uses 1×1 convolutions for feature projection.

The initial smile score is calculated using two blendshape coefficients: mouthSmileLeft and mouthSmileRight. To take smile asymmetry into account, the average value of the left and right sides is used:

$$S_{smile} = \frac{\text{mouthSmileLeft} + \text{mouthSmileRight}}{2},$$

where S_{smile} is the initial smile intensity score, mouthSmileLeft and mouthSmileRight are the smile coefficients of the left and right sides of the mouth.

Mouth opening is determined using the jawOpen and mouthOpen coefficients. The first coefficient shows how much the jaw is lowered, while the second shows how open the mouth is in general. Since both features describe the same movement from different sides, the program selects the larger value. This is necessary in order not to lose the movement if one of the coefficients is lower due to recognition features.

$$S_{mouth} = \max(\text{jawOpen,mouthOpen}),$$

where S_{mouth} is the unprocessed mouth opening value, jawOpen and mouthOpen are the coefficients of jaw and mouth opening.

Eyebrow raising is evaluated using three coefficients at once: the movement of the inner part of the eyebrows and the movement of the outer parts of the left and right eyebrows. The inner part of the eyebrows is the main indicator of surprise, tension, or the user's reaction, so weighting coefficients are introduced in the program to give priority to this indicator with the value 0.55. The outer parts of the left and right eyebrows supplement the overall analysis, so their weight is distributed evenly.

$$S_{browRaise} = 0.55 \times \text{browInnerUp} + 0.225 \times \text{browOuterUpLeft} + 0.225 \times \text{browOuterUpRight},$$

where $S_{browRaise}$ is the eyebrow raising indicator, browInnerUp is the raising of the inner part of the eyebrows, browOuterUpLeft and browOuterUpRight are the raising of the outer parts of the left and right eyebrows.

Frowning and squinting are calculated as the average values of symmetrical features:

$$S_{browFrown} = \frac{\text{browDownLeft} + \text{browDownRight}}{2},$$

where $S_{browFrown}$ is the frowning indicator, browDownLeft and browDownRight are the lowering of the left and right eyebrows.

$$S_{squint} = \frac{\text{eyeSquintLeft} + \text{eyeSquintRight}}{2},$$

where S_{squint} is the squinting indicator, eyeSquintLeft and eyeSquintRight are the squinting of the left and right eyes.

To take into account the individual characteristics of users, the module provides calibration of the neutral state. At this stage, the program determines which values of facial expression features are normal for a particular person. This processing helps reduce the number of false detections. During calibration, 20 samples of a neutral expression are accumulated, after which the average baseline value is calculated for each feature:

$$Base = \frac{1}{n}\sum_{i=1}^{n} z_i,$$

where $Base$ is the individual baseline value of the feature, z_i is the unprocessed value of the feature in the i-th frame, n is the number of calibration samples, in the program $n = 20$.

As a result, the baseline values $Base_{smile}$, $Base_{mouth}$, $Base_{browRaise}$, $Base_{browFrown}$ and $Base_{squint}$ are formed. The program then evaluates the deviation from the individual baseline. For this purpose, normalization with a gain coefficient and a dead zone is used. The dead zone means the minimum threshold of feature change below which the program does not take fluctuations into account in the calculation.

$$\text{Score} = f\big((z - Base - DeadZone) \times Gain\big),$$

where Score is the normalized intensity of the feature, z is the current unprocessed value, $Base$ is the baseline value after calibration, $DeadZone$ is the dead zone that cuts off weak fluctuations, $Gain$ is the gain coefficient, f is a function that limits the result to the range from 0 to 1.

The dead zone excludes small natural facial movements and recognition noise, while the gain coefficient makes noticeable changes more expressive. The following formulas are used for specific features:

$$\text{Score}_{smile} = f\big((S_{smile} - B_{smile} - 0.03) \times 5.5\big), \quad (1)$$

$$\text{Score}_{mouth} = f\big((S_{mouth} - B_{mouth} - 0.03) \times 5.5\big), \quad (2)$$

$$\text{Score}_{browRaise} = f\big((S_{browRaise} - B_{browRaise} - 0.02) \times 6.0\big), \quad (3)$$

$$\text{Score}_{browFrown} = f\big((S_{browFrown} - B_{browFrown} - 0.02) \times 5.5\big), \quad (4)$$

$$\text{Score}_{squint} = f\Big(\big(S_{squint} - B_{squint} - 0.02\big) \times 6.0\Big), \quad (5)$$

where the corresponding Score$_*$ values are the normalized scores of smiling, mouth opening, eyebrow raising, frowning, and squinting; the numerical thresholds 0.02–0.03 define the dead zone, so very weak facial changes are considered noise and are not taken into account, while the coefficients 5.5–6.0 are used as selected gain values that convert small but noticeable changes in facial expression features into a normalized score from 0 to 1.

3.2 Eye and Gaze Analysis Module

The eye module uses the coordinates of facial landmarks related to the eyelids, eye contour, eye corners, and iris. Its task is to determine gaze direction and calculate blinks. These features are considered as a separate module of the program because eye movement and blink frequency may reflect the user's attention, fatigue, or tension.

First, iris points are identified for each eye. Since the iris is represented by several landmarks, its center is calculated as the average value of the coordinates:

$$C_{iris} = \frac{1}{n}\sum_{i=1}^{n} P_i,$$

where C_{iris} is the center of the iris, P_i is the coordinate of one of the iris points, n is the number of iris points.

The horizontal gaze direction is determined by the position of the iris center between the eye corners. If the center is shifted toward one of the boundaries, the gaze is considered to be directed to the corresponding side:

$$R_h = \frac{C_{iris,x} - X_{min}}{X_{max} - X_{min}},$$

where R_h is the relative horizontal position of the iris, $C_{iris,x}$ is the x-coordinate of the iris center, X_{min} and X_{max} are the coordinates of the eye corners.

The final gaze direction is formed by combining horizontal and vertical classification: "center", "left", "right", "up", or "down", as well as combined variants such as "left-up".

To count blinks, the EAR indicator is used, which represents the ratio of eye height to eye width. It is calculated using six key eye points:

$$EAR = \frac{d(P_2,P_6)+d(P_3,P_5)}{2\times d(P_1,P_4)}, \quad (6)$$

where EAR is the ratio of eye height to width, P_1 and P_4 are the horizontal eye points, P_2, P_3, P_5 and P_6 are the upper and lower eyelid points, $d(P_i, P_j)$ is the Euclidean distance between two points.

3.3 Audio Stream Processing Module

The audio module estimates volume, accumulates history, and prepares data for Whisper-based speech recognition. Operating at a sampling rate of 16000 Hz with a block size of 1024 samples, the module calculates the volume of each incoming block as the root mean square value of the amplitude:

$$Volume = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2},$$

where $Volume$ is the volume of the audio block, N is the number of audio samples, x_i is the value of the i-th audio sample.

Sound fragments are stored in a rolling audio history window to eliminate excessive RAM usage:

$$N_{window} = f_s \times T_{window},$$

where N_{window} is the number of samples in the audio window, f_s is the sampling rate, T_{window} is the duration of the window.

Recognition is started every 1.5 seconds rather than after each audio block. This makes it possible to accumulate enough speech data, reduce the load on the model, and avoid processing fragments that are too short. The number of new samples required for the next recognition launch is calculated as follows:

$$N_{hop} = f_s \times T_{hop},$$

where N_{hop} is the number of new samples required to start recognition, T_{hop} is the processing step.

When enough data has been accumulated in the window, the audio fragment is passed to the Whisper neural network model, which returns a set of segments from which the program extracts the text and combines non-empty parts into a single string:

$$\text{Text} = \text{segment}_1 + \text{segment}_2 + ... + \text{segment}_n,$$

where Text is the combined recognized phrase, segment_i is the text of the i-th speech segment, n is the number of segments.

3.4 Redness Analysis Module

The redness module analyzes color changes in facial skin. Unlike the facial expression and eye modules, which mainly work with geometric features, this module uses pixel characteristics of the image. Its task is to determine whether the skin tone in selected facial areas has changed relative to the individual baseline state.

Four zones are used for analysis: the left cheek, the right cheek, the nose, and the forehead. The center of each zone is calculated as the average coordinate of the selected facial landmarks.

A small image area is cut out around the found center. Before estimating redness, it is normalized to reduce the influence of changes in lighting. First, the average values of the R, G and B channels are calculated, and then the overall gray level is determined:

$$Gray_{mean} = \frac{R_{mean} + G_{mean} + B_{mean}}{3},$$

where $Gray_{mean}$ is the average gray level, R_{mean}, G_{mean} and B_{mean} are the average values of the red, green, and blue channels.

Since natural skin tone and lighting conditions differ, the module collects 30 samples of the normal state. For each sample, the indicator of the left cheek, right cheek, nose, and forehead is calculated, as well as the general global indicator:

$$M_{global} = 0.34 \times M_{left} + 0.34 \times M_{right} + 0.18 \times M_{nose} + 0.14 \times M_{forehead},$$

where M_{global} is the general color indicator, M_{left}, M_{right}, M_{nose} and $M_{forehead}$ are the skin redness estimates for the left cheek, right cheek, nose, and forehead.

After 30 samples are accumulated, the baseline values $Base_{left}$, $Base_{right}$, $Base_{nose}$, $Base_{forehead}$ and $Base_{global}$ are calculated. Then the program evaluates the deviation of the current color from the baseline. First, the global shift is determined, which helps separate local redness from a general lighting change:

$$Shift_{global} = \max\left(M_{global} - Base_{global}, 0\right),$$

where $Shift_{global}$ is the overall positive color shift relative to the baseline, B_{global} is the baseline global indicator.

Then, for each zone, a normalized redness score is calculated. The formulas use a dead zone $D = 0.030$ and compensation for part of the global shift:

$$R_{left} = f\left(\left(M_{left} - Base_{left} - D - 0.25 \times Shift_{global}\right) \times 8.0\right), \quad (7)$$

$$R_{right} = f\left(\left(M_{right} - Base_{right} - D - 0.25 \times Shift_{global}\right) \times 8.0\right), \quad (8)$$

$$R_{nose} = f\left(\left(M_{nose} - Base_{nose} - D - 0.20 \times Shift_{global}\right) \times 7.2\right), \quad (9)$$

$$R_{forehead} = f\left(\left(M_{forehead} - Base_{forehead} - D - 0.20 \times Shift_{global}\right) \times 6.8\right), \quad (10)$$

where R_{left}, R_{right}, R_{nose} and $R_{forehead}$ are the current redness scores for the zones, M_* is the current value, $Base_*$ is the baseline value, D is the dead zone, $Shift_{global}$ is compensation for the general lighting change.

The overall redness indicator is obtained as a weighted sum of the zone values:

$$R_{overall} = 0.34 \times R_{left} + 0.34 \times R_{right} + 0.18 \times R_{nose} + 0.14 \times R_{forehead},$$

where $R_{overall}$ is the overall facial redness indicator. The cheeks have the greatest weight because they are usually the most informative zones.

3.5 Multimodal Feature Fusion Module

The final stage of the system consists in integrating heterogeneous parameters obtained from independent analytical blocks.

Special importance in data fusion is given to the detection of increased or anomalous parameters recorded in the initial dataset both for real behavior and for synthetic content [3, 6].

The integration of redness indicators with geometric features clarifies the overall assessment of the user's state. In human physiology, strong local facial skin redness under normal conditions is almost always accompanied by corresponding changes in facial expressions, such as microexpressions of tension, squinting, or a change in blink frequency measured by the EAR index. This may indicate stress, cognitive load, or a strong emotional response.

If the system detects an increased level of skin redness, but facial activity indicators remain completely static or neutral, the combined risk index increases sharply, since such inconsistencies may be important for visual authenticity analysis [6].

Other anomalies from the dataset are taken into account in the fusion module in a similar way. When there is a sharp facial movement, such as a quick eyebrow raise, face detection, head pose estimation, and face alignment modules may become less stable [8, 9, 10].

At this moment, the system automatically lowers the sensitivity threshold of the blink and gaze detector in order to record a delay of synthetic generation.

Abnormally low eye openness, measured by the EAR index, combined with a high blink frequency is compared with the audio stream. If, at the moment of pronouncing key phrases recognized by the Whisper model, the trajectory of lip and eyelid movements does not match the patterns of real people from the training sample, the system marks the content as suspicious.

4 Results

To comprehensively verify the operability of the developed multimodal system and assess the accuracy of feature integration, a full-scale experimental study was conducted, which included 100 independent participants.

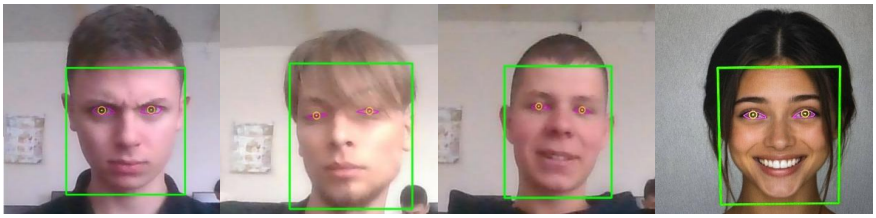
The testing involved students of higher educational institutions aged 18 to 23. This age group was chosen intentionally, as it represents the main group of users of distance learning and online interview systems.

Each experiment simulated the process of passing a critically important oral online test, such as an interview or an exam, lasting from 5 to 10 minutes.

The system continuously captured data in parallel. During multiple runs, an optimal set of parameters was selected. The effectiveness of the methodology was evaluated by verifying the operation of individual modules, comparing the obtained values with the individual calibration baseline, and analyzing the temporal consistency between visual, physiological, and speech-related indicators. Coherent changes were interpreted as behavioral or physiological reactions, while

inconsistencies between modalities were considered potential technical anomalies. Based on 100 conducted tests, the system recorded standard behavioral patterns and borderline physiological states, such as stress and excitement. Additionally, image in Figure 1d, used as part of the experimental dataset, was generated using OpenAI GPT-5.4 with the prompt “create a girl looking forward” and subsequently animated using the Yandex Alice system (based on YandexGPT) with the prompt “the girl looks forward, raises her eyebrows, and smiles.”

Examples of the experiments are presented in Figure 1. For a more detailed review of the obtained results, the data were distributed across separate modules of the program. Numerical indicators and a brief description of the observations are presented in Tables 1–4.



(a) (b) (c) (d)

Fig. 1. Program output results: (a) text was recognized, the state was calm, the confidence was likely true, the overall risk was 0.13, the facial expression was 0.76, the blink coefficient was 0.00, and the gaze was 0.00; (b) text was recognized, the state was calm, the confidence was likely true, the overall risk was 0.00, the facial expression was 0.00, the blink coefficient was 0.00, and the gaze was 0.00; (c) text was recognized, the state was moderate, the confidence was likely true, the overall risk was 0.10, the facial expression was 0.00, the blink coefficient was 0.19, and the gaze was 0.00; (d) text was recognized, the state was moderate, the confidence was likely true, the overall risk was 0.21, the facial expression was 1.00, the blink coefficient was 0.00, and the gaze was 0.00.

To test the module for eye and gaze analysis, gaze direction, the number of blinks, and the eye openness coefficient were evaluated. These parameters make it possible to determine how steadily the user holds their gaze and how eye activity changes during observation.

Table 1. Results of the eye and gaze analysis module

№	Number of blinks	Gaze direction	Eye openness coefficient
1	31	center	0.50
2	3	center	0.40
3	50	center	0.20
4	11	center	0.64

The table shows that all participants’ gaze was directed toward the center. At the same time, blink and eye openness indicators differ: the first participant showed moderate blinking, the second showed lower blinking, and the third showed the highest blink frequency with lower eye openness. These modules make it possible to compare individual features of eye activity.

Next, the operation of the facial expression module was analyzed. The program determined the intensity of individual features and identified the dominant facial expression based on them.

Table 2. Results of the facial expression analysis module obtained using equations (1–5)

№	Dominant facial expression	Mouth opening coefficient	Smile	Eyebrow raising coefficient	Frowning coefficient	Squinting
1	frowning	0.00	0.00	0.00	1.00	0.62
2	neutral	0.00	0.00	0.00	0.00	0.02
3	smile	0.09	1.00	0.00	0.11	0.99
4	smile	0.08	1.00	0.07	0.00	0.61

The results show that different facial expression states were recorded among the participants. The first participant predominantly showed frowning, the second participant's face remained neutral, and the third participant showed a smile and squinting. The fourth participant demonstrated a pronounced smile with moderate eye squinting and stable gaze direction, indicating a more active positive facial state. Therefore, the module distinguishes between calm, neutral, tense, and more active facial expression states.

The microexpression module was tested to record short-term facial changes. Unlike general facial expression analysis, it shows rapid reactions that occur at specific moments of observation.

Table 3. Results of the microexpression module calculated using using formula (6)

№	Recorded microexpression	Observation features
1	rapid eyebrow raise	A short-term facial change appeared against the background of general frowning.
2	rapid eyebrow raise	The microexpression was recorded against the background of an overall neutral facial expression.
3	rapid frowning	A short-term change in facial expression appeared during an active smile and squinting.
4	rapid smile	A short-term facial change was observed in the form of a rapid smile with noticeable squinting against a stable positive facial expression and centered gaze.

The first and second participants showed a rapid eyebrow raise, while the third participant showed rapid frowning. The fourth participant demonstrated a rapid smile accompanied by slight squinting, indicating a short-term positive facial reaction under stable gaze conditions. These data complement the results of facial expression analysis and make it possible to consider not only stable facial expressions but also brief changes that may be associated with an emotional reaction.

To assess external physiological manifestations, the facial redness analysis module was used. The program determined the overall redness level and separately evaluated several facial areas.

Table 4. Results of the facial redness analysis module obtained using equations (7–10)

№	Redness status	Overall redness indicatoropening coefficient	Left cheek	Right cheek	Nose	Forehead
1	normal	0.05	0.04	0.07	0.10	0.07
2	normal	0.06	0.04	0.27	0.00	0.08
3	strong	0.36	0.00	0.51	0.68	0.63
4	normal	0.07	0.05	0.06	0.09	0.06

In the first two participants, redness was within the normal range, although the right cheek was locally more pronounced in the second participant. The third participant showed strong redness, most noticeable in the areas of the nose, forehead, and right cheek. The fourth participant exhibited weak facial redness with all regional indicators remaining within the normal range and no pronounced asymmetry across facial zones. This shows that the module makes it possible to detect both the overall level of facial color change and individual areas where it is more pronounced.

The speech recognition module was used to convert spoken phrases into text. This is necessary to compare the speech fragment with the visual features recorded by the program at the same moment.

For quantitative comparison of existing solutions, a functional coverage score was calculated based on supported modalities and feature groups. Each function is evaluated as 1 for full support, 0.5 for partial support, and 0 for no support. The final result is calculated as the sum of all criteria, which allows systems to be compared in terms of functional coverage (Table 5).

Table 5. Quantitative comparison of functional coverage of existing solutions and our solution

Functions\ Systems	Face Tec[3]	iProo[4]	Ent rust[5]	Conve rus[6]	Our method
Bio-metrics	1	1	1	0	1
Liveness	1	1	1	0	1
Synthetic	0.5	1	0.5	0	1
Behavior	0	0	0	0.5	1
Facial movement	0	0	0	0	1
Gaze	0	0	0	1	1
Voice	0	0	0	0	1

[3] https://www.facetec.com/
[4] https://www.iproov.com/
[5] https://www.entrust.com/
[6] https://converus.com/

BioSig	0	0	0	0.5	1
Dynamics	0	0	0	0.5	1
Multimodal fusion	0	0	0	0.5	1
Total	2.5	3	2.5	3	10

5 Conclusions

Within the framework of this work, a methodology for the comprehensive analysis of human behavioral features was developed, combining geometric, physiological, and audio-linguistic modalities with mandatory individual adjustment.

Based on the proposed methodology, a program[7] was created for the comprehensive analysis of human behavioral features in real time. This work successfully solved the task of developing and studying a multimodal system for monitoring and analyzing user behavior in real time.

An original method of end-to-end was implemented, allowing geometric parameters, such as facial landmarks and the EAR eye openness index, physiological markers, such as local skin redness dynamics, and synchronized audio-linguistic data from the Whisper model to be combined in a spatio-temporal context.

The proposed methodology includes a mandatory stage of preliminary individual adjustment, namely baseline memorization based on 20–30 frames, and algorithmic compensation for external lighting using the Gray World method, which minimizes the number of false system triggers and takes into account the unique characteristics of each person.

Acknowledgements. This study was supported by the Russian Science Foundation, project no. 26-21-20099, https://rscf.ru/project/26-21-20099/.

References

1. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7617–7627 (2021). https://doi.org/10.1109/CVPR46437.2021.00753
2. Atoum, Y., Chen, L., Liu, A.X., Hsu, S.D.H., Liu, X.: Automated Online Exam Proctoring. IEEE Transactions on Multimedia 19(7), 1609–1624 (2017). https://doi.org/10.1109/TMM.2017.2656064
3. Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2), 423–443 (2019). https://doi.org/10.1109/TPAMI.2018.2798607
4. Lu, Y.: Real-time Eye Blink Detection Using General Cameras: A Facial Landmarks Approach. International Science Journal of Engineering & Agriculture 2(5), 1–8 (2023).

[7] https://github.com/Anohina-yulia/Detecting-Deepfakes-with-Visual-Features

5. Milone, A.S., Cortese, A.M., Balestrieri, R.L., Pittenger, A.L.: The impact of proctored online exams on the educational experience. Currents in Pharmacy Teaching and Learning 9(1), 108–114 (2017). https://doi.org/10.1016/j.cptl.2016.08.037
6. Song, X., Zhao, X., Fang, L., Lin, T.: Discriminative representation combinations for accurate face spoofing detection. Pattern Recognit. 85, 220–231 (2019). https://doi.org/10.1016/j.patcog.2018.08.019
7. Sun, X., Wu, P., Hoi, S.C.H.: Face detection using deep learning: an improved faster RCNN approach. Neurocomputing 299, 42–50 (2018). https://doi.org/10.1016/j.neucom.2018.03.030
8. Tsai, Y.-H., Lee, Y.-C., Ding, J.-J., Chang, R.Y., Hsu, M.-C.: Robust in-plane and out-of-plane face detection algorithm using frontal face detector and symmetry extension. Image Vis. Comput. 78, 26–41 (2018). https://doi.org/10.1016/j.imavis.2018.07.003
9. Wang, Y., Liang, W., Shen, J., Jia, Y., Yu, L.-F.: A deep Coarse-to-Fine network for head pose estimation from synthetic data. Pattern Recognit. 94, 196–206 (2019). https://doi.org/10.1016/j.patcog.2019.05.026
10. Xu, X., Kakadiaris, I.A.: Joint head pose estimation and face alignment framework using global and local CNN features. In: FG 2017 IEEE International Conference on Automatic Face & Gesture Recognition (2017), pp. 642–649. https://doi.org/10.1109/FG.2017.81
11. Zhang, Y., Tino, P., Leonardis, A., Tang, K.: A survey on neural network interpretability. IEEE Trans. Emerg. Top. Comput. Intell. 5, 726–742 (2021). https://doi.org/10.1109/TETCI.2021.3100641
12. Zhen, X., Yu, M., Xiao, Z., Zhang, L., Shao, L.: Heterogenous output regression network for direct face alignment. Pattern Recognit. 105, 107311 (2020). https://doi.org/10.1016/j.patcog.2020.107311
13. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face Alignment in Full Pose Range: A 3D Total Solution. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(1), 78–92 (2019). https://doi.org/10.1109/TPAMI.2017.2778152