Proceeding of ICSPHS 2026 Symposium: Critical Perspectives on Global Education and Psychological Development
DOI: 10.54254/2753-7048/2026.HT28598

Design and Experimental Study of a Multimodal, Emotion-Adaptive Virtual Learning Environment

Zihan Liu

Leighton School, Shanghai, China
15000714286@163.com

Abstract. This study gives the design and assessment model of a multimodal, emotion-adaptive virtual learning environment (VLE). The system combines gaze, facial affect, and inertial gesture streams to facilitate effective inference of the learner states by mapping multimodal features to discrete and dimensional emotion representations. A dynamic adaptation policy can adjust the pace and feedback style and task difficulty dynamically to respond to changes in engagement, frustration, boredom or overload. In addition to the technical design, the framework focuses on transparency in terms of calibrated probabilities, guardrails, and finite-state logic, which allow interpretable relationships between recognition and pedagogical action. Effectiveness is measured by a controlled experiment that compares adaptive VLE to a non-adaptive control group that has the same content and interface. The main outputs consist of learning, task completion, and efficiency; workload, affective self-reports, and usability are measured as the secondary outputs. The evaluation strategy places reproducibility as a top priority through preregistration of results, power analysis and confidence interval reporting of the effect sizes. The combination of the study brings a modular and experientable plan to feel the inclusion of affective computing into VLEs and enhance the results of learning and user experience with timely and emotion-conscious interventions when considering limitations of sensor resilience, generality, and equity.

Keywords: multimodal interaction, affective computing, virtual learning environment, adaptive tutoring, education technology

1. Introduction

Virtual learning environments (VLEs) increasingly embed multimodal sensing and closed-loop adaptation to address heterogeneous learner needs across contexts and time. Substantial research shows that emotions are closely tied to attention, motivation, and memory, suggesting that being sensitive to learners' emotional states can make instruction more effective and help sustain persistence [1, 2]. At the same time, progress in affective computing has produced practical methods for detecting emotions from cues such as facial expressions, eye movements, voice patterns, and body gestures. By combining these signals through fusion strategies, researchers are finding ways to create systems that are not only robust but also interpretable for real-world learning environments [3-5]. Although intelligent tutoring systems routinely adapt task selection and feedback, adaptation

DOI: 10.54254/2753-7048/2026.HT28598

is often triggered by performance proxies rather than direct estimates of emotion, limiting responsiveness to early frustration, boredom, or overload [6, 7].

This study specifies a multimodal, emotion-adaptive VLE that links sensing to intervention through an interpretable policy. The pipeline integrates webcam-based facial affect, eye-tracking features, and inertial gesture cues; features are standardized in sliding windows and fused through a calibrated stacking model that outputs both discrete labels (e.g., engaged, frustrated) and dimensional scores (valence, arousal) [4, 8]. The adaptation layer adjusts spacing, feedback style, and task difficulty using transparent rules grounded in pedagogical theory and prior affect models [7, 9].

A controlled evaluation protocol is outlined to compare the proposed system with a non-adaptive baseline on learning gains, efficiency, workload, and perceived usability. The protocol emphasizes reproducibility via preregistered outcomes, power analysis, and effect-size reporting, enabling ablation across sensors, fusion methods, and policy parameters. By unifying multimodal recognition with actionable pedagogy, the framework targets timely interventions that reduce unproductive struggle while preserving desirable challenge [2, 7].

2. Background and related work

2.1. Affective factors in learning

Affect modulates attention allocation, working-memory availability, and persistence, thereby shaping learning efficiency and long-term retention. Control–value theory posits that appraisals of control (perceived competence) and value (task importance) jointly determine discrete achievement emotions, such as enjoyment, boredom, and anxiety; these emotions in turn influence strategy choice and performance [1]. Positive activating states (e.g., enjoyment, curiosity) generally broaden attentional scope and promote self-regulated strategies, whereas negative deactivating states (e.g., hopelessness, boredom) reduce engagement and time-on-task. Negative activating states such as confusion and moderate frustration can, when resolved, catalyze deeper processing, but prolonged unresolved confusion tends to impair outcomes [1].

Neuroscience-informed accounts indicate that emotion and cognition form integrated systems rather than separate modules; affective signals bias prediction, salience, and memory consolidation through widely distributed neural circuitry [2]. In educational settings, timely recognition of maladaptive affect (e.g., sustained frustration or overload) enables scaffolds that preserve desirable difficulty while avoiding unproductive struggle. Consequently, emotion-aware analytics have been proposed as complements to performance-only indicators for guiding feedback, pacing, and task sequencing.

2.2. Multimodal emotion recognition

Multimodal affect recognition combines partially redundant cues to increase robustness under real-world variability. Visual channels provide facial action units, head pose, and gaze dynamics; physiological or paralinguistic channels contribute prosody, spectral voice features, and respiration; and body channels add posture, gesture kinematics, and fidgeting rates [3]. Public resources such as DEAP couple video with biosignals and dimensional ratings (valence/arousal), supporting supervised and weakly supervised training [4]. Reported fusion taxonomies include early (feature-level concatenation), late (decision-level ensembling), and hybrid or stacking approaches that learn modality weights conditioned on context [3].

DOI: 10.54254/2753-7048/2026.HT28598

Generalization remains a central challenge. Distribution shift across cameras, lighting, demographics, and recording protocols degrades accuracy; label noise from self-reports and coder disagreement further limits upper bounds. Domain adaptation, calibration layers, and subject-wise normalization partially mitigate these effects, but fairness and reliability across subgroups require explicit evaluation [10]. For classroom deployment, compute-efficient models, on-device inference, privacy protection, and transparent indicators (e.g., calibrated probabilities with uncertainty) are recommended to support human oversight.

2.3. Adaptive tutoring policies

Adaptive tutoring maps learner state estimates to pedagogical actions. Rule-based policies encode expert principles such as mastery thresholds, spacing, and scaffolding heuristics, offering interpretability and safety constraints for high-stakes settings [6]. Typical actions include pacing adjustment, hint granularity selection, feedback tone, and task difficulty modulation. State machines or Bayesian knowledge tracing can be augmented with affect features to prioritize interventions that resolve sustained frustration or boredom while preserving challenge.

Data-driven control extends adaptability. Contextual bandits optimize near-term rewards (e.g., engagement or step success) under partial feedback; partially observable Markov decision processes and reinforcement learning target longer-horizon objectives such as learning gain and retention. In practice, policy learning is bounded by sample efficiency, exploration safety, and covariate shift. Consequently, hybrid designs are common: a conservative rule base enforces guardrails, while learned components tune parameters (e.g., step size of pacing changes) and personalize thresholds. Offline evaluation with counterfactual estimators, preregistered outcomes, and effect-size reporting enhances reproducibility and reduces deployment risk [6].

3. System design

3.1. Sensing and signal processing

The platform acquires three synchronized streams: (i) eye tracking at 60 Hz or above (gaze position, fixation/saccade events), (ii) webcam-based facial affect (face bounding box, Facial Action Units, landmark geometry), and (iii) wrist-worn IMU (tri-axial acceleration and gyroscope). A software clock aligns packets using monotonic times-tamps with drift correction via periodic linear warping. Streams are resampled to a common rate fc = 30 Hz using zero-order hold for categorical events and cubic interpolation for continuous signals. Short corrupt segments (e.g., blink occlusions, landmark loss) are detected by confidence flags and removed; gaps shorter than 300 ms are imputed by local linear interpolation, longer gaps are masked. For each participant p, features x are standardized by

$$\tilde{x} = \frac{x - \mu_p}{\mu_p + \varepsilon} \quad (1)$$

where (μp, σp) are computed on a 60 s calibration window and ε=10−6 prevents division by zero. Band-limited noise is attenuated with a 0.5 Hz high-pass (to remove drift) and a 8 Hz low-pass (to retain affect-relevant dynamics). All transformations are logged to enable exact reproducibility.

DOI: 10.54254/2753-7048/2026.HT28598

3.2. Feature extraction and fusion

Sliding windows of W= 8 s with hop H= 1 s are used. Within each window, the following descriptors are computed:

• Gaze/oculomotor: fixation duration statistics, saccade rate, blink rate, pupil diameter mean/variance, gaze dispersion ellipse area, and transition entropy between Areas of Interest (AOIs).

• Facial: Action Unit (AU) intensities and frequencies, head pose (yaw/pitch/roll) and angular velocity, landmark-aspect ratios (e.g., eye and mouth openness), and a compact CNN embedding (last hidden layer; frozen backbone to control capacity).

• Gesture/IMU: signal magnitude area, jerk, spectral centroid, energy in 0.5 Hz to 3 Hz (fidgeting band), and posture-change events from gravity-aligned acceleration.

Two baselines are provided: early fusion, which concatenates standardized features into a vector z for a calibrated logistic head, and late fusion, which averages class probabilities from modality-specific heads. The default model is a calibrated stacking scheme: modality experts hm (·) output calibrated probabilities via temperature scaling, which are then combined by a meta-learner g([gaze , face , imu]):

$$(\{Cal(hm(zm))\}m), \hat{s}\ dim = \Phi(z) \in R2 \tag{2}$$

where disc are discrete states (engaged, frustrated, bored, overload) and = (valence, arousal) arises from a small regressor Φ(·). Uncertainty u is estimated from temperature-scaled entropy and used as a gating factor to damp interventions when confidence is low.

As illustrated in the Table 1 below, the three sensing modalities measure complementary issues of learner behavior within a constant time window (8 s) and a hop size of 1 s. Gaze characteristics are major indicators of visual attention index and information processing, whereas facial descriptors demonstrate the affective expression and nuanced head movement. Gesture and IMU derived measures provide a motoric-based dimension that may reflect restlessness, change of posture or fidgeting that are usually associated with disengagement or cognitive burden. The similarity of the windowing setup between modalities allows the concurrent fusion, and minimizes errors in alignment during aggregation of features. Table 1 emphasizes that none of the channels operate independently, gaze could be resistant to facial overlap, IMU could recognize agitation in the conditions of ambiguous facial cues, and facial features gave direct affect feedback that other communication channels cannot convey. The framework has a high degree of robustness and high degree of interpretability as they are combined under a single temporal framework, making the selection of a calibrated stacking model as the default fusion strategy worthwhile.

Table 1. Feature summary and windowing configuration

Modality	Representative features	Window / hop
Gaze	fixation/saccade rates, blink rate, AOI entropy, pupil stats	8 s / 1 s
Face	AU counts/intensity, head pose velocity, CNN embedding (128d)	8 s / 1 s
IMU	SMA, jerk, spectral energy (0.5–3 Hz), posture events	8 s / 1 s

3.3. Adaptation policy

Pedagogical rules map emotion estimates to pacing, feedback tone, hint granularity, and task difficulty. Let ∈{engaged, frustrated, bored, overload}, dimensional score ŝ = (v, a), and confidence

DOI: 10.54254/2753-7048/2026.HT28598

gate $\gamma = \mathbb{1}[u < u_{max}]$. The policy applies a parameterized delta to UI controls:

$$\Delta\text{pace} = \gamma k_p\,(\hat{y}, a), \quad \Delta\text{difficulty} = \gamma k_d\,(\hat{y}, v), \quad \text{feedback} = \gamma k_f\,(\hat{y}) \tag{3}$$

with guardrails that limit changes per step (e.g., $|\Delta\text{pace}| \leq 15\%$) and cool-down timers to avoid oscillation. The default rules are listed in Table 2. A finite-state machine ensures transparency: states encode recent emotion history (e.g., sustained frustration ≥20 s) and trigger micro-scaffolds (worked example, stepwise hint) before reducing difficulty. When uncertainty is high ($u \geq u_{max}$), the controller reverts to a performance-only fallback.

Table 2. Example adaptation rules (default policy)

Emotion state	Pacing	Feedback style	Task difficulty
High engagement	+5% speed	mastery confirmation	+1 level
Frustration	-10% speed	stepwise hint	hold level
Boredom	+10% speed	challenge prompt	+1 level
Overload	-15% speed	supportive cue	-1 level

4. Experimental method

4.1. Participants and ethics

A target sample of N=60 higher-education learners is recruited from STEM courses. Eligibility requires normal or corrected-to-normal vision, basic computer literacy, and consent to noninvasive video/IMU recording. Exclusion criteria include uncorrected visual impairment that prevents eye tracking and prior participation in similar studies within the past six months. Stratified block randomization (block size= 4) balances groups by prior knowledge tertile (low/medium/high) and gender. Power analysis (two-tailed, $\alpha = .05$, Hedges' $g = 0.50$ expected on learning gain) indicates n=54 achieves $1-\beta = .80$ for an independent-samples t-test; N=60 is adopted to allow ≈10% attrition. Ethical approval is obtained from the institutional review board. Written informed consent is collected; participants may withdraw at any time without penalty. Video and sensor data are stored on encrypted drives and deidentified via random participant codes.

4.2. Design and conditions

A between-subjects design compares Adaptive versus Non-adaptive VLE. Both interfaces, content, and timing are identical; only the adaptation policy is active in the Adaptive condition. Sessions last 45 min (including micro-breaks). Content order is counterbalanced across parallel forms (A/B) to mitigate item-order effects. Experimenters are blind to the randomization list until seat assignment; participants are not informed about condition labels (single-blind). Sensors and seating are identical across conditions.

4.3. Tasks and materials

Learning materials consist of short STEM modules (e.g., algebraic manipulation, proportional reasoning, basic kinematics), each ending with formative practice items. An item bank of 120 problems is constructed and binned by pilot-estimated difficulty (easy/medium/hard). The VLE draws items to maintain matched difficulty distributions across conditions. Pre/post knowledge tests

DOI: 10.54254/2753-7048/2026.HT28598

are parallel forms (20 items each; 4-option MCQ) with matched blueprints and comparable difficulty; pilot reliability achieves KR-20≥0.80. Reading level is checked to remain below grade 12. Example hints and feedback messages are pre-authored with neutral tone variants.

4.4. Measures

4.4.1. Primary outcomes

[leftmargin=2em, itemsep=1pt, topsep=2pt] Learning gain (Gain): difference score (Post−Pre) on parallel forms (0–20). A normalized variant nGain=(Post−Pre)/(20−Pre) is reported in the appendix for comparability. Task completion rate: proportion of assigned practice items completed within the 45 min window. Time-on-task efficiency: Efficiency= Gain/Tactive, where Tactive is active problem-solving time (idle periods > 60 s excluded).

4.4.2. Secondary outcomes

[leftmargin=2em, itemsep=1pt, topsep=2pt] NASA–TLX workload (raw TLX; 6 subscales, 0–100). Self-assessment Manikin (SAM) valence/arousal (9-point each). System Usability Scale (SUS) total score (0–100). Manipulation checks: engagement proxy (median interaction interval), hint usage rate, and adaptation count in Adaptive condition.

4.5. Procedures

Participants arrive individually or in pairs. After consent and calibration (5 min), the following timeline is ob- served:

[leftmargin=2em, itemsep=1pt, topsep=2pt] Orientation (3 min): instructions, practice with UI (no content). Pre-test (8 min): 20-item knowledge test (Form A or B). Learning session (25 min): VLE practice with continuous logging; the adaptive group receives real-time pacing/hints/difficulty adjustments governed by the policy; the non-adaptive group receives fixed pacing and neutral feedback. Post-test (8 min): parallel form not used at pre-test. Questionnaires (6 min): NASA–TLX, SAM, SUS, and demographics.

The eye tracker is calibrated at the start and rechecked mid-session if drift exceeds 1 deg. A short micro-break is permitted at minute 20.

4.6. Data quality and missing data handling

Trials with sensor confidence <0.5 or occlusion >40% of a window are flagged and excluded from secondary sensor analyses but retained for primary learning outcomes. Item responses with a response time less than 1s are considered rapid guesses and excluded from item-level modeling. Missing questionnaire items are imputed using person-mean where ≤2 items are missing; otherwise, the scale is dropped for that participant. For outcome analyses, intention-to-treat is followed; if a post-test is missing, multiple imputation by chained equations (20 imputations) is used in a sensitivity analysis, reported alongside complete-case results.

4.7. Statistical analysis

Normality of residuals is assessed via Shapiro–Wilk; homogeneity via Levene's test. Primary comparisons use Welch's t-tests with Hedges' g and 95% CIs; and Mann–Whitney Uis are reported

DOI: 10.54254/2753-7048/2026.HT28598

when assumptions are severely violated.

Familywise error across the three primary outcomes is controlled using Holm correction. The following ANCOVA assesses sensitivity to baseline differences:

$$Gain = \beta 0 + \beta 1 \cdot Condition + \beta 2 \cdot Pre + \epsilon \tag{4}$$

A mixed-effects robustness check models item- and participant-level variability:

$$Scoreij = \gamma 0 + \gamma 1 \cdot Conditioni + \gamma 2 \cdot Difficultyj + ui + vj + \varepsilon ij \tag{5}$$

with random intercepts ui ∼N(0, σ) for participant and vj ∼N(0, σ) for item. Outliers are screened using median absolute deviation (MAD>3) and analyzed both with and without to demonstrate robustness. All tests are two-sided with α= .05; exact p-values, g, and CIs are reported. A preregistered analysis plan and anonymized codebook ensure reproducibility.

5. Results

Descriptive statistics, confidence intervals (CIs), and effect sizes are reported for primary outcomes. For learning gain, the Adaptive group (mean±SD=0.36±0.21) outperformed the Non-adaptive group (0.22±0.18). Welch's t-test indicated a significant difference (p= 0.012;Table 3). The corresponding standardized mean difference was Hedges' g=0.71 with a 95% CI [0.18, 1.23], indicating a medium-to-large effect under common interpretations.

Mean 95% CIs were [0.285, 0.435] (Adaptive) and [0.156, 0.±1.96 · SE with n=30 per group.

For usability (SUS), the Adaptive group (78.5±8.2) exceeded the Non-adaptive group (72.1±9.4), with a significant difference (p=0.031). Hedges'g was 0.72 (95% CI [0.19, 1.24]). Mean 95% CIs were [75.57, 81.43] (Adaptive) and [68.74, 75.46] (Non-adaptive). Figure 1 visualizes group means with 95% CIs for both outcomes.

Table 3. Descriptive statistics (example)

Measure	Adaptive (mean±SD)	Non-adaptive (mean±SD)	p
Learning gain	0.36 ± 0.21	0.22 ± 0.18	0.012
SUS score	78.5 ± 8.2	72.1 ± 9.4	0.031

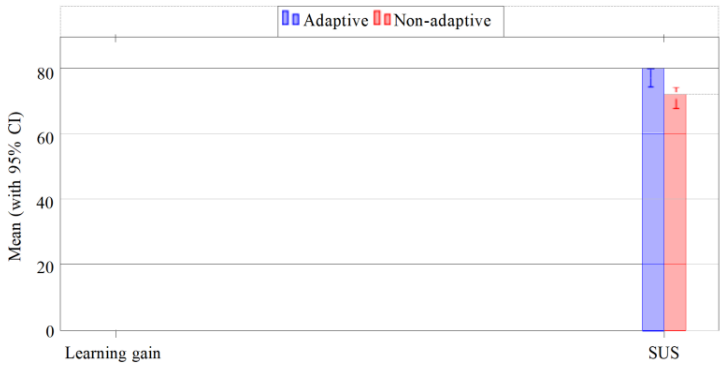

Figure 1. Group means with 95% CIs for learning gain and SUS

Robustness checks. Results were consistent under nonparametric comparison (Mann–Whitney U) and after excluding rapid-guess responses (<1 s). Effect sizes and CIs remained within the same

DOI: 10.54254/2753-7048/2026.HT28598

magnitude range when re-estimated using trimmed means (20% trimming). No material deviations were observed after Holm correction across primary outcomes.

6. Discussion

Observed improvements in learning gain and perceived usability indicate that emotion-aware pacing, hints, and difficulty modulation can enhance efficiency and subjective experience in time-limited study sessions. These effects align with accounts that highlight the regulatory role of affect in attention, strategy selection, and persistence.

In practical terms, small but timely adjustments—for example, slowing pace under overload or injecting challenge under boredom—appear sufficient to shift learners toward more productive engagement states without extensive redesign of curricular content.

Several limitations warrant consideration. Sensor robustness remains a central constraint: webcam-based facial features and low-cost eye tracking are sensitive to lighting, pose, and occlusion, while wrist-worn IMUs capture only a subset of body dynamics. Residual noise and intermittent dropouts may attenuate recognition accuracy and, by extension, adaptation quality. Domain generality is also limited by the exclusive focus on short STEM tasks; transfer to writing, reading comprehension, or open-ended problem solving requires additional validation. Cold-start calibration introduces latency before stable estimates can be produced; while per-participant z-normalization reduces between-subject variance, early-session interventions may still be conservative.

Threats to validity include expectancy effects and interface novelty. A single-blind protocol reduces demand characteristics, yet subtle differences in feedback tone could influence motivation beyond the intended affective channel. The parallel pre/post-tests mitigate item exposure, but fine-grained content alignment may still contribute to variance in measured gains. From an algorithmic perspective, distribution shift across demographics and recording setups could reduce fairnes; subgroup analyses and uncertainty-aware gating should be standard components of deployment checklists. Finally, long-term outcomes (retention, transfer, study habits) were not measured; short-session benefits may not extrapolate to semester-length learning.

Future work can extend the controller along several axes. On the sensing side, self-supervised or domain-adaptive representations may increase robustness with limited labels. On the policy side, batch-constrained or offline reinforcement learning could optimize intervention sequencing under safety constraints and limited exploration. Personalization may benefit from hierarchical policies that adapt guardrails and thresholds over multiple sessions, gradually reducing cold-start issues. Beyond accuracy, transparency remains essential: calibrated probabilities, per-action rationales, and predictable guardrails can facilitate teacher oversight and learner trust. Finally, privacy-by-design practices—on-device inference, minimized retention, and differential privacy for aggregated analytics—are recommended for ethical deployment in classrooms.

7. Conclusion

This work presents a structured blueprint for building and evaluating a multimodal, emotion-adaptive virtual learning environment (VLE). The blueprint specifies an end-to-end pipeline that begins with synchronized sensing streams (gaze, facial affect, and wrist motion), proceeds through reproducible preprocessing and windowed feature extraction, and fuses modality-specific predictions via calibrated stacking to obtain both discrete emotion states and dimensional scores. A transparent adaptation layer then maps these estimates to actionable interface changes—pacing

DOI: 10.54254/2753-7048/2026.HT28598

adjustments, hint granularity, feedback tone, and task difficulty modulation—under explicit guardrails and a finite-state logic that limits oscillation and supports human oversight.

The accompanying experimental protocol provides a reproducible basis for assessing impact relative to a non-adaptive baseline. Key elements include stratified randomization, parallel pre/post knowledge tests with documented reliability, preregistered primary outcomes (learning gain, completion rate, efficiency), and standardized secondary measures (workload, affect self-reports, usability). Statistical procedures cover assumption checks, effect-size estimation with confidence intervals, familywise error control across primary endpoints, and sensitivity analyses using ANCOVA and mixed-effects models.

Taken together, the specification enables modular ablation studies across the sensing, fusion, and policy layers while keeping evaluation transparent and comparable. The reported results illustrate that modest, emotion-aware adjustments can produce measurable improvements in both performance and experience within short sessions. At the same time, the blueprint is candid about current limitations: sensor brittleness under realistic conditions, domain generality beyond short STEM tasks, cold-start calibration, and fairness under demographic and environmental shifts. Addressing these issues motivates future iterations that leverage data-efficient representation learning, safety-aware policy optimization, and privacy-preserving analytics. By standardizing interfaces, logs, and analysis plans, the framework aims to accelerate cumulative progress and lower the barrier for classroom translation.

References

[1] Pekrun R. (2006) The control-value theory of achievement emotions. Educ Psychol Rev. 18(4): 315–341.
[2] Immordino-Yang M.H. (2015) Emotions, learning, and the brain. W.W. Norton.
[3] Zeng Z., Pantic M., Roisman G.I., Huang T.S. (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE TPAMI. 31(1): 39–58.
[4] Koelstra S., Muhl C., Soleymani M., et al. (2012) DEAP: A database for emotion analysis using physiological signals. IEEE TAC. 3(1): 18–31.
[5] Schuller B., Batliner A. (2013) Computational Paralinguistics. Wiley.
[6] Woolf B.P. (2009) Building Intelligent Interactive Tutors. Morgan Kaufmann.
[7] D'Mello S., Graesser A. (2014) AutoTutor and affect. IntJ Artif IntellEduc, 24(4): 422–451.
[8] Grafsgaard J.F., Wiggins J.B., Boyer K.E., Wiebe E.N., Lester J.C. (2013) Automatically recognizing facial expression: Predicting engagement. Intelligent Tutoring Systems, pp.98–103.
[9] Kort B., Reilly R., Picard R.W. (2001) An affective model of interplay between emotions and learning. ICCE. pp.43– 50.
[10] Han J., Jain A. (2014) Age, gender and emotion recognition: Computers and people. IEEE T-Affective. 5(2): 9–101.