

Do Student Emotions Influence Learning During Virtual Field Trips? A Computer Vision Analysis of Emotion States in Desktop-based Virtual Reality

Ryan T. Woo¹; Foysal Ahmed²; Parth Bhadaniya³;
Ricardo Eiris, Ph.D.⁴; and Masoud Gheisari, Ph.D.⁵

¹Ph.D. Student, School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281 (corresponding author). Email: rtwoo@asu.edu

²Ph.D. Student, School of Sustainable Engineering and the Built Environment, Arizona State University. Email: fahmed55@asu.edu

³Ph.D. Student, M.E. Rinker, Sr. School of Construction Management, University of Florida, Gainesville, FL 32603. Email: parthbhadaniya@ufl.edu

⁴Assistant Professor, School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, 85281 (corresponding author). Email: reiris@asu.edu

⁵Associate Professor, M.E. Rinker, Sr. School of Construction Management, University of Florida, Gainesville, FL 32603. Email: masoud@ufl.edu

ABSTRACT

Student emotions shape learning by influencing attention, engagement, and self-regulation. Virtual field trips (VFTs) in desktop virtual reality can elicit affect, yet the emotion-learning relationship in construction VFTs remains understudied. This paper tests whether facial-expression-derived emotion metrics during a drywall-focused, instructor-guided VFT relate to learning gains. Webcam video was analyzed with a computer-vision pipeline (S3FD face detection and EmoNet classification) to label five expressions (neutral, happy, surprise, fear, sad). We computed expression-duration proportions and an emotion switch rate and related them to pre-/post-test improvement. After a face-visibility quality filter, 16 participants were analyzed. Negative expression proportion (fear + sad) correlated with learning gains ($r = 0.532$, $p = 0.034$, $R^2 = 0.283$), as did switch rate ($r = 0.506$, $p = 0.046$, $R^2 = 0.256$), while neutral proportion showed a nearly significant marginal negative association ($r = -0.484$, $p = 0.058$, $f^2 = 0.306$).

KEYWORDS

Virtual field trips; construction education; affective computing; facial expression recognition; computer vision

INTRODUCTION

Field trips to construction sites are a critical component of construction engineering and management (CEM) education, providing experiential learning opportunities that bridge theoretical classroom knowledge with real-world practice (Sippel and Blum 2022; Wang et al. 2018). These structured, instructor-led visits enable students to observe construction processes, interact with professionals, and contextualize abstract concepts within authentic settings (Wang et al. 2018). Field trips can enhance visualization of concepts and improve engagement and motivation among learners (Eiris et al. 2022). However, traditional site visits face barriers such as safety and liability concerns, large class sizes, geographic distance, scheduling conflicts, and limited resources, often restricting consistent and equitable access (Eiris and Gheisari 2017).

To address these challenges, virtual field trips (VFTs) that leverage BIM, 360-degree imagery, and point clouds have emerged as viable alternatives (Wang et al. 2018). Web-based

platforms can support avatar embodiment, spatial audio, shared whiteboards, and embedded drawings, enabling collaborative problem solving in simulated jobsite environments (Sun et al. 2022). Recent studies report promising outcomes for tasks such as hazard recognition and plan-reading learning activities (Man et al. 2024; Sun et al. 2022).

Despite the benefits of VFTs, the emotional dimension of learning within these digital environments remains underexplored in CEM VFT research. Educational psychology indicates that emotions are connected to learning processes and outcomes through mechanisms such as attention allocation, motivation, strategy use, and self-regulation (Pekrun 2006; Pekrun and Stephens 2010; Loderer et al. 2018). Immersive learning research further suggests that interactive virtual environments can intensify and shape affective responses during learning (Makransky and Petersen 2021; Harley et al. 2016; Schneider et al. 2021). Accordingly, measuring affect during VFT participation may provide additional insight into how learners engage with CEM learning content and how that engagement relates to learning outcomes.

To address this need with minimal disruption to the learning activity, this study applies automated facial expression recognition (FER) to webcam recordings captured during a drywall-focused desktop VFT. We operationalize participant-level facial expression metrics and examine their associations with pre-/post-test learning gains, providing preliminary evidence toward understanding how facially observable affective patterns relate to learning outcomes in desktop-based construction VFT settings.

BACKGROUND

Emotions, learning, and virtual reality (VR). In learning contexts, emotions (and broader affective states) are commonly discussed in terms of how learners experience and respond to instructional activities and task demands. Control-value theory and related frameworks describe how emotions can influence learning through mechanisms such as attention allocation, motivation, persistence, and self-regulation (Pekrun 2006; Pekrun and Stephens 2010).

For immersive and interactive digital learning such as VR, affect is also positioned as part of theoretical models of learning. The Cognitive Affective Model of Immersive Learning (CAMIL) and Cognitive-Affective-Social Theory of Learning in Digital Environments (CASTLE) emphasize interactions between affective factors, cognitive processing, and social mechanisms during immersive learning experiences to explain learning in digital contexts (Makransky and Petersen 2021; Schneider et al. 2021). In VFTs, learners may encounter visually complex environments and multi-step tasks, suggesting that affective responses could vary meaningfully during participation and plausibly relate to learning outcomes.

Measuring affect during learning. Affect can fluctuate within an activity, and common approaches differ in temporal resolution, intrusiveness, and scalability. Self-report is widely used but can be limited by recall bias and coarse temporal resolution when affect changes during an activity (Loderer et al. 2018). Reviews of affect detection describe multimodal approaches (e.g., self-report, observational coding, interaction logs, and physiological sensing), each with tradeoffs in intrusiveness and interpretability (Calvo and D’Mello 2010).

Computer vision for emotion and learning measurement. When face video is available, computer vision can provide a practical path to continuous measurement with relatively low additional burden. Facial expression recognition (FER) estimates expression categories and/or affective dimensions from face imagery and has been used to probe engagement and affective states during learning (Bosch et al. 2015; Tang et al. 2025). Recent progress is supported by large-scale naturalistic datasets such as AffectNet and deep neural network models that operate

on unconstrained face imagery (Mollahosseini et al. 2019). This study uses EmoNet, a convolutional neural network that outputs discrete facial expression classes and continuous valence/arousal estimates (Toisoul et al. 2021). Here, we use the discrete expression labels and derived valence categories to operationalize affect during VFT participation. Importantly, facial configurations are treated as observable facial-behavior proxies rather than direct ground-truth internal emotional experience, motivating caution when interpreting outputs as “emotion” in a psychological sense (Barrett et al. 2019).

Emotion estimation in construction engineering and management (CEM). Emotion estimation is an emerging topic within CEM educational and training research. Pilot projects have used affective computing methods to infer learners’ or workers’ affective states; for example, VR crane-training systems have integrated EEG-based affective interfaces to adapt to trainees’ cognitive load (Rezazadeh et al. 2011) and wearable EEG has been used to monitor arousal/valence during construction tasks (Hwang et al. 2018). In VR safety training, multimodal biometric responses have been used to forecast learning performance (Choi et al. 2023). However, there remains limited understanding of how FER can provide insights into learning processes in desktop-based VFT settings.

STUDY GOAL, QUESTION, AND METHODOLOGY

The goal of this study is to investigate relationships between students’ facial-expression-derived emotion metrics and learning outcomes during a desktop-based, instructor-guided VFT focused on drywall construction techniques. Research question: How do FER-derived expression proportions and emotion switch rate relate to learning gains measured via pre-/post-assessment?

Study learning context—VFT case study: drywall construction. A pedagogical jobsite model was developed to enable collaborative walkthroughs and task-based learning within an authentic jobsite context while operating entirely through a web browser to reduce technical barriers. Drywall construction was selected due to its curricular relevance and the feasibility of staging progressive construction sequences in a virtual model.

Environment creation. Platform development followed a two-step process: (1) constructing a pedagogical jobsite model using modeling software (Rhinoceros 8® and Autodesk Revit®) and (2) configuring the model within a web-based collaborative virtual environment (Frame®) (Figure 1). In Step 1, a six-story reinforced-concrete frame building with drywall systems was created with distinct drywall assemblies aligned to the learning objectives. Floors exhibited progressive construction stages, from completed paneling and openings to framing around openings and ongoing wall framing work, enabling the instructor to reference multiple levels of completion within one model.

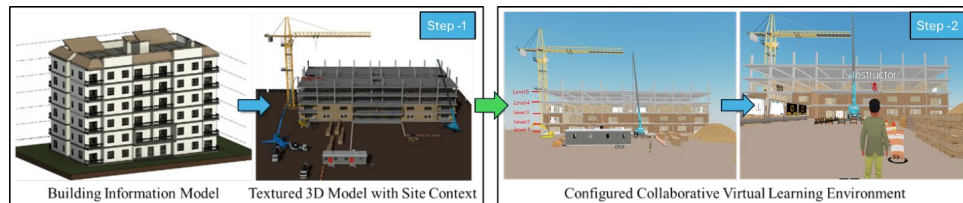


Figure 1. Environment Creation Process.

In Step 2, the consolidated model was imported into Frame® and equipped with collaboration tools. A site office served as the entry point where participants received orientation on navigation, pointer tools, whiteboard interaction, and media-sharing options. Additional tools were positioned on instructional levels to support guided explanations and collaborative tasks.

Educational intervention. The intervention targeted two learning objectives mapped to Bloom’s taxonomy: (LO1—Understand) explain drywall assembly installation sequences and (LO2—Remember) identify drywall framing component terminology.

Instruction delivery. An instructor led participants through a synchronized route from the site office through Levels 1–3. Level 2 contained spatially distributed learning zones addressing each objective (Figure 2). For LO1, assembled and exploded wall views were used to illustrate sequencing through narration and pointer guidance. For LO2, a partially constructed wall with tagged components and corresponding legends supported terminology identification. A protocol standardized waypoint sequences, time allocations, and core explanations across sessions.

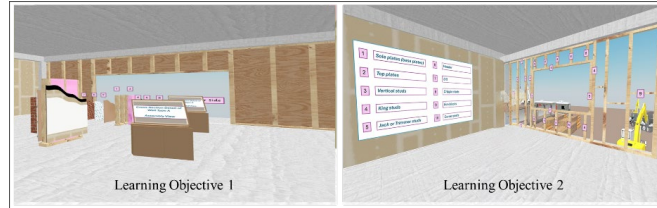


Figure 2. Instruction Delivery Setup and Learning Zones.

Participants and protocol. Thirty-four students from undergraduate and graduate construction programs participated as 17 dyads. Sessions were conducted in a controlled laboratory setting using standardized desktop computers, headsets, microphones, and webcams. Sessions followed a structured sequence including consent, demographics, pre-test, onboarding, instructor-led walkthrough, and post-test; duration ranged from 45 to 90 min.

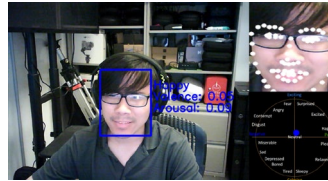


Figure 3. Webcam Facial Emotion Recognition Analysis.

Learning and facial expression data collection. Pre- and post-assessments evaluated LO1 and LO2 using identical image-based questions. LO1 required ordering multi-layer drywall assembly diagrams, and LO2 required selecting correct component names for numbered framing callouts. Webcam video recordings captured facial expressions and behaviors during VFT interaction and served as the primary data source for emotion analysis (Figure 3).

STUDY METRICS AND ANALYSIS

Emotion inference pipeline. Video was sampled at every frame (no down-sampling). For each sampled frame, a face detector (S3FD; Zhang et al. 2017) located the face region of interest, and the face crop was passed to EmoNet configured for five expression classes (neutral, happy, sad, fear, surprise) (Toisoul et al. 2021). Although EmoNet also produces continuous valence and arousal, analyses in this paper directly use the discrete expression labels only.

Frame validity and exclusions. Frames without a detected face (or with an invalid face crop) were excluded from computation. For each participant, we computed a valid-frame ratio as the proportion of processed frames producing a valid prediction; participants were excluded if this ratio was below 0.75, resulting in an analyzed sample of $n = 16$.

Feature engineering: emotional state metrics. Participant-level emotion metrics were computed from valid frames: (a) expression-category proportions (neutral, happy, sad, fear,

surprise); (b) valence-category proportions (positive = happy + surprise; negative = sad + fear; neutral = neutral); and (c) emotion switch rate (number of label transitions between consecutive valid frames divided by session duration). Before participant-level aggregation, valid frame-level predictions were temporally smoothed using a centered 1.0 s rolling window to reduce sensitivity to isolated frame-level fluctuations.

Learning outcome and statistical analysis. Pre-and post-assessments consisted of identical image-based questions aligned with LO1 and LO2. LO1 included 8 items (drywall assembly sequencing), and LO2 included 9 items (component identification). All items were equally weighted and scored as correct or incorrect for each participant to obtain a percentage of correctly answered items. Learning gain was calculated as the difference between the post-test and pre-test scores expressed in percentage points. For example, a participant scoring 50% on the pre-test and 100% on the post-test would have a learning gain of 50%. If both variables in a comparison passed a Shapiro–Wilk normality test ($\alpha = 0.05$), Pearson correlations were used. Spearman rank correlations were used otherwise.

RESULTS AND DISCUSSION

Analyses were performed for 16 participants after applying the frame-validity criterion and temporal smoothing. Across participants, neutral expressions comprised on average 47.4% of valid frames, negative expressions (fear + sad) 42.7%, and positive expressions (happy + surprise) 9.9%. Mean emotion switch rate was 0.36 switches/s. Mean LO1–LO2 pre-/post-test improvement was 42.8% (range –11.8% to 93.8%).

Table 1. Descriptive Statistics for Emotion Metrics and Learning Outcomes (n = 16).

	Student Emotional States				Learning Outcome
	Positive Emotion (%)	Negative Emotion (%)	Neutral Emotion (%)	Switch Rate (per sec.)	Pre/Post Score Increase (%)
Mean	9.9	42.7	47.4	0.36	42.8
Median	10.1	50.5	40.2	0.36	45.1
SD	3.4	25.9	27.3	0.13	29.7
Min-Max	[3.3, 16.9]	[3.6, 78.3]	[13.0, 89.9]	[0.14, 0.59]	[-11.8, 93.8]

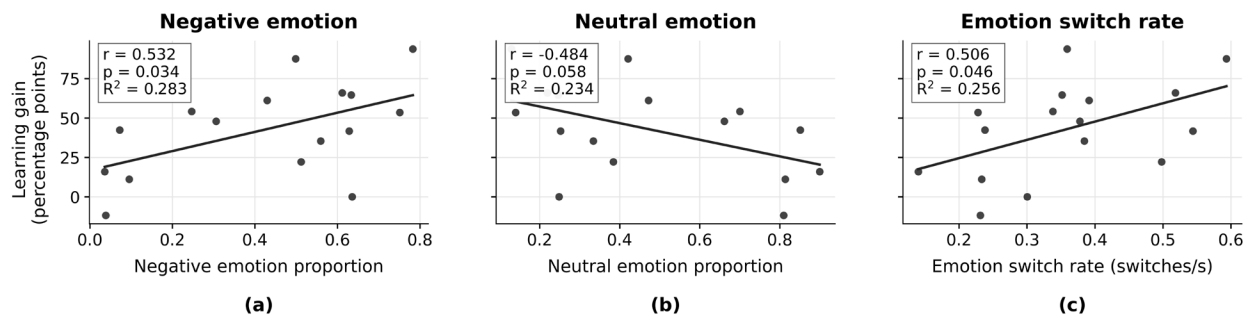


Figure 4. Correlations between Emotion Metrics and Learning Gains

Pearson correlations indicated that negative expression proportion (sad + fear) was positively correlated with learning gains ($r = 0.532$, $p = 0.034$, $R^2 = 0.283$). Emotion switch rate was also

positively correlated with learning gains ($r = 0.506$, $p = 0.046$, $R^2 = 0.256$). Neutral proportion showed a nearly significant marginally negative relationship with learning gains ($r = -0.484$, $p = 0.058$, $R^2 = 0.234$, $f^2 = 0.306$). No significant relationships were observed between learning gains and positive expression proportion (happy + surprise) or the individual expression-category proportions (all $p > 0.05$). As a data-quality check within the retained sample, missing-frame count was not significantly correlated with learning gains ($p > 0.05$), suggesting that residual between-participant variation in face-detection yield was not the primary driver of the observed correlations.

Discussion. The positive associations between learning gains and negative expression proportion in Figure 4(a) and emotion switch rate in Figure 4(c) suggest that facially observable affective dynamics during the VFT co-varied with learning improvements in this sample. One plausible interpretation is that affective states and emotion dynamics track engagement-related mechanisms (e.g., attention allocation and self-regulatory effort) during challenging learning activities (Pekrun 2006; Pekrun and Stephens 2010). A possible explanation for why negative expressions were associated with learning gains, whereas positive expressions were not, is that in an instructor-guided and cognitively demanding VFT, negative expressions may coincide with moments of uncertainty or increased effort during cognitively demanding learning. In contrast, positive expressions may be less frequent as they do not map directly onto the cognitive processing required by the activity. Similarly, higher switch rates may indicate that learners' facial expression signals changed more frequently across instructional segments and task demands, rather than maintaining a prolonged neutral display as shown in Figure 4(b).

CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

This study applied computer vision-based FER to webcam recordings collected during a desktop-based drywall VFT to derive emotion metrics and examine relationships with LO1–LO2 learning gains. Using EmoNet predictions aggregated into valence-category proportions and an emotion switch-rate measure, correlation analyses ($n = 16$) indicated that higher negative expression proportion and higher emotion switch rate were significantly associated with greater pre-/post-test improvement, while neutral proportion showed a marginal negative relationship and positive proportion showed no significant relationship.

These results demonstrate the feasibility of applying automated FER on webcam recordings to characterize facial-behavior patterns during VFT participation and to explore relationships between facially observable affective metrics and learning outcomes in construction education contexts.

Limitations. The study (1) relies on facial expressions as a single observable channel of affect that should not be interpreted as direct measurements of internal emotional experience, (2) is sensitive to face visibility and occlusion (e.g., headwear), motivating explicit frame-validity inclusion criteria, (3) applies a general-purpose FER model not calibrated to this task context or validated with manually labelled ground truths, and (4) uses correlational analyses that do not imply causality. Additionally, the analysis is based on a relatively small sample ($n = 16$) and processes only one participant from each dyad, limiting conclusions about collaborative dynamics and limiting statistical power. FER performance and potential bias can vary across contexts and participant characteristics, warranting caution in interpretation and emphasizing the value of multi-method triangulation.

Future work. Future studies should evaluate robustness with larger samples and complementary affect measures (e.g., self-report, behavioral/interaction traces), examine

temporal emotion dynamics aligned with instructional phases of the VFT, and extend analyses to collaborative settings by modeling dyadic factors (e.g., analyzing both members' affect and dyad-level learning outcomes). A multimodal approach that combines facial expression with analysis of participant dialogue (e.g., using natural language processing) may provide a more comprehensive account of affect during VFT learning.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Awards No. 2336507, 2336508, and 2336509. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. 2019. "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements." *Psychological Science in the Public Interest*, 20(1): 1–68.
- Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., et al. 2015. "Automatic detection of learning-centered affective states in the wild." *Proc., 20th Int. Conf. on Intelligent User Interfaces (IUI)*, 379–388. ACM.
- Calvo, R. A., and D'Mello, S. 2010. "Affect detection: An interdisciplinary review of models, methods, and their applications." *IEEE Transactions on Affective Computing*, 1(1): 18–37.
- Choi, D., Seo, S., Park, H., Hong, T., and Koo, C. 2023. "Forecasting personal learning performance in virtual reality-based construction safety training using biometric responses." *Automation in Construction*, 156: 105115.
- Eiris, R., Wen, J., and Gheisari, M. 2022. "iVisit-Collaborate: Collaborative problem-solving in multiuser 360-degree panoramic site visits." *Computers & Education*, 177: 104365.
- Eiris, R., and Gheisari, M. 2017. "Site visit application in construction education: A descriptive study of faculty members." *International Journal of Construction Education and Research*, 15(2): 83–99.
- Harley, J. M., Poitras, E. G., Jarrell, A., Duffy, M. C., and Lajoie, S. P. 2016. "Comparing virtual and location-based augmented reality mobile learning: Emotions and learning outcomes." *Educational Technology Research and Development*, 64(3): 359–388.
- Hwang, S., Jebelli, H., Choi, B., Choi, M., and Lee, S. 2018. "Measuring workers' emotional state during construction tasks using wearable EEG." *Journal of Construction Engineering and Management*, 144(7): 04018050.
- Loderer, K., Pekrun, R., and Lester, J. C. 2018. "Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments." *Learning and Instruction*, 70: 101162.

- Makransky, G., and Petersen, G. B. 2021. "The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality." *Educational Psychology Review*, 33(3): 937–958.
- Man, S. S., Wen, H., and So, B. C. L. 2024. "Are virtual reality applications effective for construction safety training and education? A systematic review and meta-analysis." *Journal of Safety Research*, 88: 230–243.
- Mollahosseini, A., Hasani, B., and Mahoor, M. H. 2019. "AffectNet: A database for facial expression, valence, and arousal computing in the wild." *IEEE Transactions on Affective Computing*, 10(1): 18–31.
- Pekrun, R. 2006. "The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice." *Educational Psychology Review*, 18(4).
- Pekrun, R., and Stephens, E. J. 2010. "Achievement emotions: A control-value approach." *Social and Personality Psychology Compass*, 4(4): 238–255.
- Rezazadeh, I. M., Wang, X., Firoozabadi, M., and Hashemi Golpayegani, M. R. 2011. "Using affective human–machine interface to increase the operation performance in virtual construction crane training system: A novel approach." *Automation in Construction*, 20(3): 289–298.
- Schneider, S., Beege, M., Nebel, S., Schnaubert, L., and Rey, G. D. 2021. "The Cognitive-Affective-Social Theory of Learning in Digital Environments (CASTLE)." *Educational Psychology Review*, 34(1): 1–38.
- Sippel, E., and Blum, H. 2022. "Virtual reality field trip of a steel building under construction." *Proc., 2022 ASEE Annual Conference & Exposition, Paper ID 40547. ASEE Conferences.*
- Sun, Y., Albeaino, G., Gheisari, M., and Eiris, R. 2022. "Virtual collaborative spaces for online site visits: A plan-reading pilot study." *Proc., Annual Assoc. Schools of Construction Int. Conf.*
- Tang, X., Gong, Y., Xiao, Y., et al. 2025. "Facial expression recognition for probing students' emotional engagement in science learning." *Journal of Science Education and Technology*, 34.
- Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., and Pantic, M. 2021. "Estimation of continuous valence and arousal levels from faces in naturalistic conditions." *Nature Machine Intelligence*, 3: 42–50.
- Wang, P., Wu, P., Wang, J., Chi, H.-L., and Wang, X. 2018. "A critical review of the use of virtual reality in construction engineering education and training." *International Journal of Environmental Research and Public Health*, 15(6): 1204.
- Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. 2017. "S3FD: Single Shot Scale-Invariant Face Detector." *Proc., IEEE Int. Conf. on Computer Vision (ICCV)*. arXiv:1708.05237.