

# Modeling the Affect of “Aha!” Moments to Detect the Moment of Learning

by

Eden Adler

B.S. Neuroscience, University of Michigan, 2015

Submitted to the Integrated Design and Management Program and the  
Department of Electrical Engineering and Computer Science  
in partial fulfillment of the requirements for the degrees of

MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT

and

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER  
SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

© 2024 Eden Adler. All rights reserved.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by:	Eden Adler Integrated Design and Management and Department of Electrical Engineering and Computer Science May 17, 2024
Certified by:	Cynthia Breazeal Professor of Media Arts and Sciences, Thesis Advisor Dean for Digital Learning
Certified by:	Manish Raghavan Assistant Prof. of Electrical Engineering and Computer Science, Thesis Reader
Accepted by:	Joan Rubin Executive Director, System Design and Management Program
Accepted by:	Leslie A. Kolodziejwski Professor of Electrical Engineering and Computer Science Chair, Department Committee on Graduate Students



# Modeling the Affect of “Aha!” Moments to Detect the Moment of Learning

by

Eden Adler

Submitted to the Integrated Design and Management Program and the  
Department of Electrical Engineering and Computer Science  
on May 17, 2024 in partial fulfillment of the requirements for the degrees of

MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT

and

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER  
SCIENCE

## ABSTRACT

What if a model could pinpoint the exact moment of learning? Currently, the only way we can understand when someone has learned is by testing them afterwards, which has its limitations. In attempts to detect the moment of learning, researchers from various fields have leveraged data from methods such as Knowledge Tracing (KT) and Electroencephalograms (EEGs) to predict students’ knowledge acquisition. These methods have contributed to improving our understanding of knowledge, but not only do they fall short of detecting the exact moment of learning, they also interfere with natural learning interactions by requiring students to wear sensors or type as they learn. Often, modeling learning does not include affect and emotion data, which are key influencers of learning outcomes. One affective expression that is often observed by educators, and has evaded quantification attempts by researchers, is the moment everything suddenly clicks for the student - the “Aha!” moment. Using classroom video data of students experiencing “Aha!” moments, we created dynamic, functional handcrafted features representing the face and body position and used them to model students’ facial expressions. We then leveraged feature selection methods and statistical analysis to ultimately contribute a novel, explainable definition of the observable, affective markers of “Aha!” moments, unlocking the opportunity to use the “Aha!” moment as a signal for detecting the moment of learning. These results invite future interdisciplinary research efforts as well as applications in fields such as artificial intelligence, human-robot interaction, education, psychology, cognitive sciences, and more.

Thesis supervisor: Cynthia Breazeal

Title: Professor of Media Arts and Sciences, Thesis Advisor  
Dean for Digital Learning

Thesis supervisor: Manish Raghavan

Title: Assistant Prof. of Electrical Engineering and Computer Science, Thesis Reader



# Acknowledgments

The inspiration for this work came from an interaction I had with my friends' four year old son. While talking about a series of random, silly things I asked him, "What if the boat was made of cheese?" He froze, and his face lit up. I could tell that my question triggered some big thoughts and the wheels in his brain started spinning. Seeing his learning right in front of me made me curious if I could use the affect of his "Aha!" moment to detect the moment of learning. Thanks, Ari, for triggering my "Aha!" moment that inspired this work.

To my advisor, Cynthia Breazeal, thank you for your mentorship, leadership, and encouragement. I'm immensely grateful for the opportunity to work with and learn from you. Your passion and vision for social robots, AI, and education inspire me to dream big and strive to improve our world through education and technology. Thank you for supporting my crazy ideas and empowering me to explore this thesis work.

To Sharifa Alghowinem, thank you for sharing your knowledge, resources, and time with me, they were truly invaluable to the success of this thesis.

To everyone in the Media Lab's Personal Robots Group, I'm so grateful for your support and encouragement. Thanks for challenging and inspiring me.

To Roz Picard, your Affective Computing class solidified my desire to pursue this thesis work. Thanks for inspiring me and for all of your encouragement.

To Manish Raghavan, thanks for challenging me to carefully think through my approach, and inspiring me to consider potential business applications of this thesis work.

To all of the staff and students in the Integrated Design and Management Program, especially Tony Hu, Andy MacInnis, Sheila Pontis, and Steven Eppinger, thanks for everything you taught us, and for creating a space for people like me whose passion is at the intersection of engineering, design, and business.

To my professors and lecturers: Kimberle Koile, Randy Davis, Robert Berwick, Arvind Satyanarayan, Roz Picard, Manish Raghavan, Dylan Hadfield-Menell, Alfred Spector, David Ninio, and Samuel Dinnar, thank you for your leadership, guidance, and support in helping me achieve this milestone.

To Jay Berckley, I'm so grateful I came across your work and that you agreed to collaborate with me. It wouldn't have been possible without your support.

To my family and friends, I'm so lucky to have such an incredible support system. Forever grateful for all of your love and encouragement.

To my husband, Josh, thank you for your endless patience, love, and support, I couldn't have done this without you.

# Contents

<b>Title page</b>	<b>1</b>
<b>Abstract</b>	<b>3</b>
<b>Acknowledgments</b>	<b>5</b>
<b>List of Figures</b>	<b>9</b>
<b>List of Tables</b>	<b>11</b>
<b>1 Introduction</b>	<b>13</b>
1.1 Thesis Roadmap . . . . .	15
<b>2 Background &amp; Related Work</b>	<b>17</b>
2.1 Modeling Learning for Personalization (A) . . . . .	17
2.2 Affect & Emotion (B) . . . . .	18
2.2.1 Affective Computing . . . . .	18
2.2.2 Measuring Facial Expressions - Facial Action Coding System (FACS) and other methods . . . . .	18
2.3 "Aha!" Moments (C) . . . . .	19
2.3.1 Measurement . . . . .	19
2.4 "Aha!" Moments & Modeling Learning for Personalization (D) . . . . .	19
2.5 Affect and Emotion & Modeling Learning for Personalization (E) . . . . .	20
2.6 Affect and Emotion & "Aha!" Moments (F) . . . . .	21
2.6.1 Emotions and Learning . . . . .	21
2.6.2 Observable Correlates of "Aha!" Moments . . . . .	22
2.7 Addressing the Gap (*) . . . . .	23
<b>3 Dataset</b>	<b>25</b>
3.1 Study . . . . .	25
3.2 Benefits and Relevance . . . . .	25
3.3 Study Methods . . . . .	26
3.3.1 Insight Tasks for Inducing "Aha!" Moments . . . . .	26
3.3.2 Data Collection Context . . . . .	27

<b>4</b>	<b>Methods</b>	<b>29</b>
4.1	Data Labeling . . . . .	29
4.1.1	Manual Annotation of "Aha!" Moments . . . . .	29
4.2	Raw Feature Extraction . . . . .	31
4.2.1	Relevant Tools . . . . .	31
4.3	Data Pre-Processing . . . . .	32
4.3.1	Isolating Single Participant Face . . . . .	32
4.3.2	Data Normalization . . . . .	32
4.4	Data Composition . . . . .	33
4.5	Handcrafted Features . . . . .	33
4.5.1	Low-Level Features . . . . .	33
4.5.2	High-Level Features . . . . .	34
<b>5</b>	<b>Modeling and Feature Selection</b>	<b>37</b>
5.1	Cross-Validation Using K-Fold Grouping . . . . .	37
5.2	Architecture . . . . .	38
5.3	Comparing Model Performance . . . . .	38
5.4	Feature Selection . . . . .	38
5.4.1	T-test . . . . .	38
5.4.2	Robust Feature Selection through Aggregating Methods . . . . .	39
5.4.3	Determining the Features of "Aha!" Moments . . . . .	39
<b>6</b>	<b>Results and Discussion</b>	<b>41</b>
6.1	Model Performance . . . . .	41
6.2	Significant Selected Features . . . . .	41
6.3	Feature Explainability . . . . .	42
6.3.1	Cheek Movements . . . . .	43
6.3.2	Eyebrow Movements . . . . .	44
6.3.3	Mouth Movements . . . . .	45
6.3.4	Related Movements: Mouth, Brow, Cheek . . . . .	45
6.4	The Affect of "Aha!" Moments . . . . .	46
<b>7</b>	<b>Conclusion</b>	<b>47</b>
7.1	Limitations . . . . .	47
7.2	Future Work . . . . .	48
<b>A</b>	<b>Complete List of Handcrafted Features</b>	<b>49</b>
A.0.1	MediaPipe - Body Position Features . . . . .	49
A.0.2	Py-Feat . . . . .	50
	<b>References</b>	<b>51</b>



# List of Figures

2.1	Related work (A-F) and the gap this work addresses (*) . . . . .	17
2.2	Comparing eureka learning (left) and gradual learning (right) [1] . . . . .	20
2.3	Examples of interventions for a pedagogical agent based on student affect [2]	21
2.4	Analysis of features related to "Aha!" moments from Berckley and Hattie [3]	23
4.1	Illustrating the progression of affective components of "Aha!" moments as described by researchers [4] . . . . .	30
4.2	Overview of the process of creating the handcrafted features for modeling [5][45][46][10] . . . . .	31
5.1	Overview of the modeling and feature selection process [6], [7] . . . . .	37
6.1	Examples of before (A), during (B), and after (C) expressions which were included in the label of "Aha!" moment. . . . .	46



# List of Tables

2.1	Emotions related to learning (from Kort et. al. [38]) and their action units (AUs) . . . . .	22
4.1	Sample of action units from FACS [8] . . . . .	35
5.1	Feature selection methods used [6] . . . . .	40
6.1	Comparison of model performance at varying feature proportions . . . . .	41
6.2	Best results for each of the three models in order of performance . . . . .	42
6.3	The significant selected features from the best performing model, Combined at 10% selection, listed according to their feature region, name, statistical functional, and statistical trend. . . . .	43



# Chapter 1

## Introduction

*“When we care about something, we measure it.” [9]*

The human body produces a variety of signals to help us understand ourselves and others in real-time. These signals can be both visible and invisible to us requiring either observation or tools and tests to understand what is occurring at a given point in time. For example, there are observable signs to understand when someone is feeling tired or going into labor, and there are also physiological tests we can perform to determine if someone is in REM sleep or has high blood sugar.

Humans also produce various cognitive indicators that help us diagnose, for example, mental health conditions and learning disabilities. These indicators are evaluated using diagnostic criteria and assessments. Other cognitive tests that detect signals as they occur in real-time require physically connecting to sensors or other devices such as EEG.

There is a cognitive event, however, that we do not currently have a way to measure - learning. We don't currently have a signal or test to pinpoint the exact moment of learning in real-time.

Currently, we rely on testing to retrospectively understand if learning occurred in the past. The results of these tests, however, lack important data and context on the individual learning process. Additionally, standardized tests such as the SAT or ACT are not infallible - they cannot distinguish between true comprehension and mere guesses.

What if, instead of assessing learning using a performance-based test which has limitations, we had a visible signal indicating the moment of learning? Humans are innately programmed to learn. Could our bodies be producing a signal marking learning and we're simply not attuned to it?

One promising signal, mentioned throughout history and described by teachers and educators, is the moment of insight also known as the “Eureka moment,” “light bulb moment,” or the “Aha! Moment” as referred to here. “Aha! Moments” describe the phenomenon of experiencing sudden insight and understanding of a previously misunderstood concept. Some of the first accounts of this in history describes Archimedes' eureka moment which led to the Law of Buoyancy, Newton discovering gravity after noticing an apple fall from a tree, and Einstein's theory of relativity just hitting him one day when his mind was wandering [10].

An “Aha!” moment is a sudden moment of insight. It's also known as a Eureka moment, light bulb moments, insight moment, and is referred to using various spellings including “aha

moment,” “AHA moment,” “ah ha moment,” etc. There are various theories and models used to describe “Aha!” moments. However what is generally agreed upon across disciplines is that “Aha!” moments include the following defining features: (1) there’s a sudden comprehension of the solution, (2) the individual has no ability to access or explain the steps they took which led to the insight moment, (3) the moment creates a feeling of surprise and a burst of positive emotions [11].

In context, researchers found that there are five dimensions of an “Aha!” experience: suddenness, surprise, happiness, impasse, certainty [12]. The observable changes in facial expression and body movement that have been found to be significantly correlated with “Aha! Moments” are smiling, raising the cheeks, opening the mouth, stretching the lips, pitch or raising the head, and dropping the jaw [3].

In searching for a signal for the moment of learning, it’s important to note that insight learning is just one type of learning. Others include analytical problem-solving and memory retrieval [13]. Features of analytical problem-solving include: (1) it is effortful, deliberate, and predominantly conscious, (2) it proceeds gradually from the initial state to the solution, (3) its steps and process are available to working memory, which means that individuals are able to recall and explain how they solved the problem. Unlike analytical problem-solving where an individual gains a deep understanding of the problem, memory retrieval is simply the ability to mentally retrieve previously acquired knowledge that is relevant to the current problem [14].

The scope of this work is to identify a signal for learning specifically through moments of insight, and those that happen in the moment, not at a delay. Some “Aha!” moments can occur days or months after new information is introduced, but this work is limited to those moments of insight that are observable during learning.

Utilizing observable, affective components of “Aha!” moments are ideal because they do not require physical connections to sensors or devices, therefore allowing for measurement in natural interaction conditions. They also are more readily measurable than other types of learning, such as gradual learning, because they produce a large change in a short amount of time. They typically occur suddenly and produce observable changes in affect.

Current methods of measuring “Aha!” moments involve asking participants to record their perceived progress while problem solving, self-reporting about their experience of insight after problem solving, tracking the progress and presence of insight through a physical pressure sensor, or recording physiological responses during learning [1], [15]–[18]. Each of these methods are either cognitively disruptive, physically invasive to the natural problem solving experience, or are prone to bias. There does not yet exist an objective, non-invasive way to measure moments of insight.

Identifying this learning signal could provide benefits to a variety of fields including, but not limited to: education, cognitive sciences, psychology, artificial intelligence, neuroscience, and human-robot interaction. Detecting the moment of learning in real-time could unlock opportunities for highly valuable interventions and services that serve to benefit a variety of stakeholders. Teachers could leverage this data to review, tweak, and improve their curriculum to optimize for those moments of learning. Teachers could also keep track of frequency of learning moments and be able to more readily identify students who may require some extra support. Research has shown that affect-aware social robots deliver more engaging educational experiences [19], so this data could be used by robot peer tutors to provide more

individualized and engaging teaching experiences.

The central research question in this work is: what are the affective features that indicate an “Aha!” moment and can we model that affect computationally?

This work’s main contribution is a definition for the affective features of “Aha!” moments which serve as a signal for the moment of learning. This novel contribution has promising applications in improving education, human-robot interaction, as well as other industries such as consumer product development and more.

## 1.1 Thesis Roadmap

- **Chapter 2** - provides relevant background and reviews previous work related to “Aha!” moments, affective computing, and modeling learning for personalization.
- **Chapter 3** - introduces the study from which the dataset used for this work originated.
- **Chapter 4** - details the methods used to prepare the dataset for modeling. It describes the process of labeling the dataset, extracting raw facial and body position data from each video in the dataset, and transforming that raw data into handcrafted features used for modeling the affect of “Aha!” moments.
- **Chapter 5** - explains the details of modeling and feature selection for determining the affective features that define observable expressions of the “Aha!” moment.
- **Chapter 6** - discusses the resulting significant selected features from our model and their explainability in contributing to the “Aha!” affect.
- **Chapter 7** - summarizes the contributions and limitations of this work and proposes directions for future work.
- **Appendix A** - complete list of handcrafted features used for modeling





# Chapter 2

## Background & Related Work

Following Figure 2.1, this chapter begins by providing relevant background on (A) modeling learning for personalization, (B) “Aha!” moments, and (C) affect and emotion, then discusses previous work related to the overlap between pairs of these topics (D-F) and their gaps, and concludes by describing how this work addresses these gaps with a novel contribution (\*).

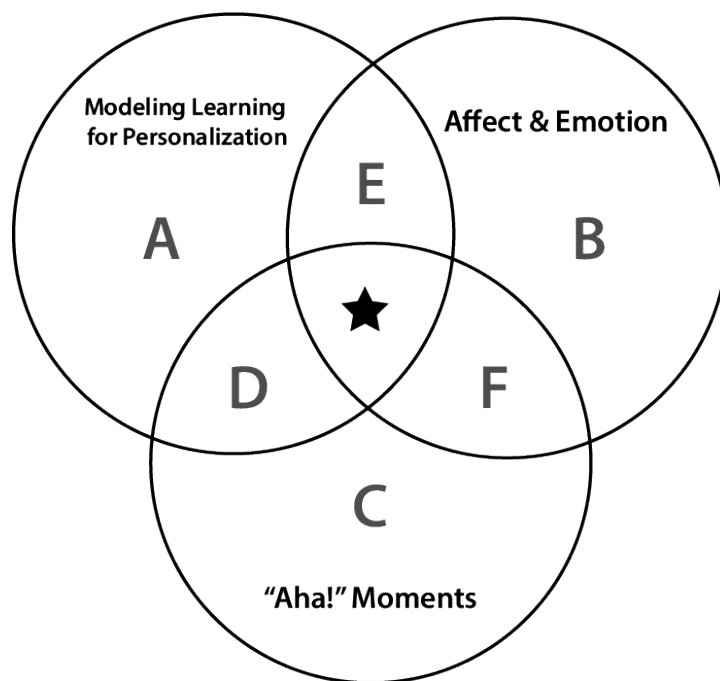


Figure 2.1: Related work (A-F) and the gap this work addresses (\*)

### 2.1 Modeling Learning for Personalization (A)

Teachers enjoy and value one-on-one time with students, but they have limited time and many students [10]. Hiring tutors is costly and not always accessible. Recognizing the

need for personalized learning companions, researchers have developed technologies such as intelligent tutoring systems (ITS), and peer robot tutors [20], [21] to provide learners with additional individualized educational support. A key component of these personalized, adaptive technologies is learner modeling, the ability to effectively estimate the knowledge state of the learner and predict future performance as they interact with the system. Most often this learner modeling is done through Bayesian Knowledge Tracing (BKT) and logistic models which relies on observational data such as response time, hint usage, history of attempts, and answer correctness [22]. The output of the models dictates the type of personalized interaction. For example, model outputs could include visualizing estimated knowledge to support self-regulated learning, choosing the order of items to present to the learner, or providing personalized explanations to the learner based on their estimated knowledge [22].

While these models aim to support student learning, they rely on result-centric data, guiding students through a series of questions, and making adjustments based on evaluating their answers. There is, however, much more to the learning process than the content. The emotional state of the student has a significant impact on their experience and learning outcomes. Without incorporating affect data into these learner models, these adaptive personalized systems will be ineffective at supporting student learning to the fullest extent.

## **2.2 Affect & Emotion (B)**

### **2.2.1 Affective Computing**

The concept of Affective Computing was proposed by Professor Rosalind Picard in 1997, and is the study and development of systems that can respond intelligently to human emotional feedback. It's an interdisciplinary field of study spanning computer science, psychology, and cognitive science [23]. Affective Computing leverages multiple modality types as input to interpret human affect and emotion including facial expressions, sentiment analysis, vocal prosody, speech, gestures and body movement. Out of all the ways humans communicate, 93% of meaning comes from nonverbal communication. Of that meaning, 7% comes from words, 38% through vocal elements, and 55% through elements such as facial expressions, posture, and gesture [24].

### **2.2.2 Measuring Facial Expressions - Facial Action Coding System (FACS) and other methods**

Ekman and Friesen developed a system for measuring unique facial expressions and behaviors called the Facial Action Coding System (FACS) which breaks down facial expressions into individual components of muscle movements called action units (AU) [8]. AUs can be understood individually or as a set of AUs that overlap in time or co-occur, and are organized by the region of the face where they occur, with few exceptions. Brow AUs, for example, include AU labels 1, 2, and 4. In FACS there is no AU 3, because it refers to a specific brow action in a specialized version of FACS for use with infants [25]. Leveraging affect and emotion data can provide powerful insights for models/digital systems. Aside from FACS, researchers employ various other methods to measure affect during learning through feelings

of warmth [26], self-report [12], [27], and visceral [16] methods. These methods, however, often contain bias, and involve disrupting the participants both cognitively and physically which in turn can impact their emotional state.

Affective computing leverages technology to gather a variety of insights that contribute to an individual's affect without requiring any action from them including facial expressions, body posture, gestures and movement, physiological signs, vocal cues, mood, and more. Using these affective computing methods, researchers have found affect to be effective in detecting individuals at risk of suicide [28], alerting a driver if they are fatigued [29], and more. These computing methods have been successful at quantifying emotion, and since emotion impacts learning, this work aims to leverage affective computing methods to fully assess student progress and performance using less disruptive and more natural interactive interventions.

## 2.3 "Aha!" Moments (C)

Since moments of insight happen suddenly and students aren't able to explain how they got to the solution, researchers say that "Aha!" Moments are one of the most difficult learning processes to understand [15]. Likewise, however, they recognize that understanding insight could reveal mysteries about human intelligence as some of the most important discoveries started as "Aha!" moments. Many of Einstein's solutions came to him as insights [30]:

I was sitting in the patent office in Bern when all of a sudden a thought occurred to me: if a person falls freely, he won't feel his own weight. I was startled. This simple thought made a deep impression on me. It impelled me toward a theory of gravitation.

### 2.3.1 Measurement

Researchers from various disciplines have sought to understand "Aha!" moments through a variety of measures including: cognitive measures with EEG and fMRI [31], [32], and pupil movements [4], physiological measures with EDA and HRV [17], and behavioral measures with text responses [1] and sudden changes in activity [33]. Very few studies have sought to understand the affective measures of "Aha!" or eureka moments, and those that have have been unsuccessful in collecting enough data for meaningful results [15], [18]. Only Berckley, has successfully measured the emotion and affect correlated with "Aha!" moments [3], [10].

## 2.4 "Aha!" Moments & Modeling Learning for Personalization (D)

Much of the research on learner modeling and ITS involves detecting the students' skill level at a given time during learning. These models do not, however, determine the exact moment the skill was acquired. The ability to determine the moment of learning could unlock more

opportunities for personalization in adaptive learning systems. Baker et. al. used text-based data from participants’ conversations with an ITS to create a model that predicts the probability that a student learned a skill at a specific step in the problem. From this they found a measure of the “spikiness” of learning, understood to represent eureka-type learning as opposed to gradual learning, and showed a correlation between spikiness and final knowledge acquired [1] Figure 2.2.

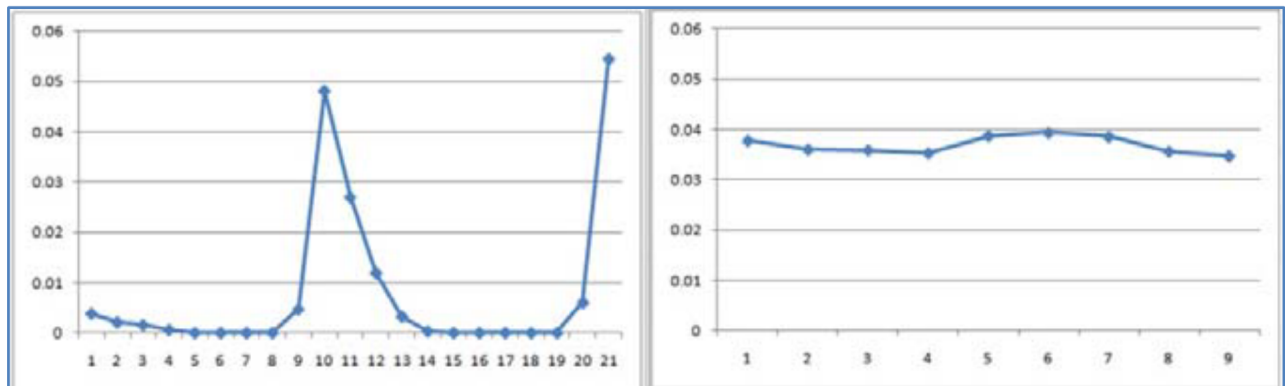


Figure 2.2: Comparing eureka learning (left) and gradual learning (right) [1]

While this work supports the ability to detect insight learning via learner modeling, it lacks important interpretability and additional context data to fully understand the students’ learning experience. Additionally, this methodology requires that students interact with learning via typed-text responses, which is limiting and detracts from typical natural cognitive learning processes.

## 2.5 Affect and Emotion & Modeling Learning for Personalization (E)

To enhance current methods and more robustly model learning, researchers should consider incorporating affective features into their existing models. When Spaulding et. al. included affective features of smile and engagement in their BKT model, this affective BKT model outperformed the traditional BKT model [19].

Emotion can serve a dual purpose with personalization, providing affective features to improve the learner models, and also providing social-emotional context to guide an agent’s response during learning. The personalized learning experience could include an embodied agent which could visually mirror the student’s affect or portray emotional reactions to the students’ progress [2], [34]. Visually representing their affect can aid the student in developing self-awareness and self-regulation skills to handle for example frustration or failure, and can also convey empathy which can improve learning [2]. Additionally, systems can use the social-emotional context to provide a more intelligent and “human-like” response to the student [35].

<i>High student frustration</i>	<i>Low student motivation</i>	<i>Low student confidence</i>
Agent looks concerned and provides an empathetic responses: “That was frustrating. Let’s move to something easier.” It gives student control: “Would you like to choose the next problem? What kind of problem would you like?”	Agent mirrors low motivation and changes its voice, motion and gestures; it may present graph, hints, adventures.	Agent provides encouragement; links performance to student effort and attributes failure to external issue (hard problem) and success to internal issues (you are doing great).
<i>Boredom because student cannot do the work</i>	<i>Boredom because work is too easy</i>	<i>Fatigue</i>
Agent moves to an easier topic and identifies material that the student can accomplish.	Agent mirrors boredom and increases the challenge level of the activity; it provides empathy messages: “Maybe this is boring? Would you like to move to something more challenging?”	Agent mirrors fatigue and presents an empathetic message: “Is this getting tiring? Shall we switch to something more fun?” It changes the activity, <i>e.g.</i> , moves to adventures, animation or game.

Figure 2.3: Examples of interventions for a pedagogical agent based on student affect [2]

Not incorporating emotion in learning personalization could even lead to detrimental effects for students [36]. For example, confusion can benefit learners in motivating deeper understanding, but prolonged, unresolved confusion could lead to negative outcomes and feelings such as frustration, boredom and eventually giving up [37]. Therefore, learner personalization must include adaptive responses to student affect in order to positively support learning outcomes. These works demonstrate the important role of emotion during learning and the positive impact of using affect and emotion data to personalize learning experiences. However, these emotions indirectly impact learning by influencing a students’ interest and motivation. They do not capture the the learning outcomes directly, and therefore could not be used to determine learning outcomes and the moment of learning.

## 2.6 Affect and Emotion & "Aha!" Moments (F)

### 2.6.1 Emotions and Learning

Teachers view student affect as essential in guiding their teaching practice. During COVID-19 when classrooms moved to virtual learning, teachers struggled to be as effective when their students had their cameras off and they couldn’t see their faces, and therefore their emotions. The limited window view also occluded their view from noticing certain body gestures and movement. Were students nervously tapping their feet, anxiously fidgeting with their hands, or pumping their fists in excitement? If students’ microphones were off,

Emotion	AUs	Ref.
Happiness	6 + 12	[38]
Engagement	1 + 10 + 45	[39]
Surprise	1 + 2 + 5 + 26	[14]
Frustration	1 + 2 + 12 + 14	[21]
Confusion	4 + 7 + 12	[21]
Boredom	43	[21]

Table 2.1: Emotions related to learning (from Kort et. al. [38]) and their action units (AUs)

often to limit distracting noises during class, the teachers also missed out on audible cues from the students. We're they talking to themselves to reason through the problem, grunting with frustration, or shrieking of excitement?

With modern education transitioning to be more virtual and multi-modal, it's important we understand what conditions are necessary for learning to occur. In our current technological state, facial expression appears to be the most reasonable and accessible emotion data available to educators. Body movement and posture are somewhat detectable but typically only includes the upper chest and head area. Audio is often disruptive to classrooms and typically students are muted. If available, physiological insights could be used, but wearing devices can alter student's natural movement, and impact their attitude.

Although affect and facial expression are helpful, not all emotions are necessarily related to learning. Kort et. al. suggests emotions that are specifically related to learning including frustration, boredom, confusion, insight, interest, etc [38](2.1). These researchers typically utilize the FACS and include recording specific AUs to classify emotions. The most commonly researched emotions related to learning are confusion, frustration, boredom, and flow [39]. Some researchers have attempted to include eureka moments [15] in studies about emotion and learning, but it occurs so rarely that they did not have enough data to find notable results.

## 2.6.2 Observable Correlates of "Aha!" Moments

Evidence from teachers supports the hypothesis that there is a unique facial affect associated with "Aha!" moments especially in educational environments. Acknowledging that "Aha!" moments can look different and that some are subtle, teachers shared how they can either see or sense the moment when it happens in their students. They describe seeing their students "light up" and noticing changes such as their eyes widening, posture straightening, smiling, clapping their hands, and sometimes their voices going higher. In more subtle cases they notice a head nod, or quick jerk in their body, and unexpected joy [10]. When they observe a student experiencing a moment of insight, teachers express feeling motivated to make adjustments to optimize for more of those moments. Sharing details about "Aha!" moments with teachers could allow them to improve their practice thereby improving student engagement and outcomes, and increase performance in schools [3].

In addition to qualitative descriptions, Berckley was able to quantify the observable affective changes that students experience before, during, and after “Aha!” moments using the FACS. Berckley used Affectiva, a proprietary software, that measures and analyzes affect and emotion in image and video data using the FACS. He reported that smiling, raising the cheeks, opening the mouth, stretching the lips, pitch or raising the head, and dropping the jaw are most significantly correlated with “Aha! Moments” [3] (Figure 2.4).

His work aimed to determine the observable correlates of “Aha!” moments, and does not involve any computation or modeling of the affect. Additionally, his work describes only a single dimension of the features, while in this work, we recognize the importance of fully understanding the dynamic behavior and context of the facial expressions [8], therefore we include first and second order derivatives of each feature as well as ten statistical methods to describe the behavior. This will be the first study to computationally model the affect and emotion of “Aha!” moments to be able to detect the moment in the wild.

TABLE 3. Means, Standard Deviations for Facial Expression Pre, Solution, and Post

Measure	Before	SD	At solution	SD	Post	SD	F	df	Sig.
Smile	1.719	7.023	55.606	42.184	10.671	17.470	49.541	2,88	<.001
Cheeks raise	0.966	6.035	32.758	30.258	1.338	6.028	44.477	2,88	<.001
Mouth open	16.297	17.053	61.050	35.034	22.662	31.929	37.450	2,88	<.001
Lip stretch	0.302	1.210	16.362	29.669	0.344	0.777	13.603	2,88	0.001
Jaw drop	7.612	11.949	25.095	22.722	9.980	24.974	10.686	2,88	0.000
Eye closure	5.273	6.799	14.144	16.895	5.226	7.437	9.835	2,88	0.002
Lid tighten	1.820	3.637	8.596	12.681	3.406	8.559	8.643	2,88	0.001
Chin raise	0.683	2.156	12.141	21.647	6.088	17.024	7.056	2,88	0.002
Dimpler	1.139	4.288	6.340	13.612	8.527	17.321	4.037	2,88	0.026
Lip press	1.442	4.670	9.304	17.092	7.674	18.342	4.008	2,88	0.028
Lip suck	2.277	9.935	13.448	27.336	7.621	22.291	3.802	2,88	0.028
Attention	96.220	5.897	89.628	19.081	92.367	11.780	3.620	2,88	0.045
Inner Brow Raise	4.342	8.407	8.535	13.876	7.269	11.815	2.775	2,88	0.072
Brow Furrow	0.341	0.737	5.554	21.275	0.350	1.189	2.769	2,88	0.103
Lip corner depressor	0.594	2.508	6.538	17.609	3.690	10.781	2.658	2,88	0.095
Brow Raise	3.046	13.040	7.115	14.044	5.343	16.003	2.099	2,88	0.132
Lip pucker	1.355	3.801	1.344	2.828	0.431	1.109	1.861	2,88	0.168
Nose wrinkle	0.100	0.304	0.605	2.416	0.398	1.512	1.445	2,88	0.242
Smirk	2.899	7.032	3.444	10.580	1.192	2.817	1.401	2,88	0.252
Upper lip raise	0.030	0.068	0.189	0.669	0.130	0.544	1.273	2,88	0.282
Eyes widen	12.578	22.895	15.103	22.950	12.401	24.172	0.536	2,88	0.533

Figure 2.4: Analysis of features related to "Aha!" moments from Berckley and Hattie [3]

## 2.7 Addressing the Gap (\*)

Previous work has explored (A) modeling learning for personalization, (B) “Aha!” moments, and (C) affect and emotion, as well as combinations of each of these topics. However, they

have yet to explore the intersection of all three of these areas. This work is the first to model learning via the affective expression of “Aha!” moments.



# Chapter 3

## Dataset

Videos from Berckley’s study on the observable correlates of “Aha!” moments [10] were shared under MIT’s data usage agreement with Eden Adler and Professor Cynthia Breazeal. His original study was approved by the Institutional Review Board of the University of Houston and all participants signed a written informed assent in order to participate, along with signed consent forms from parents or guardians (for all participants under the age of eighteen).

### 3.1 Study

The primary objective of Berckley’s study was to understand the observable correlates of “Aha!” moments and what it means for moving from surface to deep thinking. The findings from this research could be used to develop a framework for teachers to create, develop, and optimize preconditions for these moments. Berckley sought to aid educators in understanding “how to create learning environments that nurture the preconditions necessary to induce moments of insight, how to best recognize these Aha! Moments in learning, and benefits and implications of their continued development in classrooms” [10]. As an educator himself, Berckley’s desire to better understand surface to deep thinking and how it relates to “Aha!” moments is because “[o]ften, successful learning in classrooms is based solely on the need to produce high test scores, requiring teachers to frequently test and assess students’ knowledge. This process is based more on short term memory recall, and it does not favor more robust measures of long term learning that might include insight learning” [10]. The aim of his work was to show that “Aha!” moments are the gateway for moving from surface to deep thinking, therefore encouraging more engagement and richer learning experiences.

### 3.2 Benefits and Relevance

The primary reason for choosing to work with Berckley’s dataset is that he was able to successfully capture many moments of insight on video, while others recorded only a few if any [15], [18]. However, there are other noteworthy desirable characteristics unique to Berckley’s study which further supports the use of his data in this work. Firstly, when comparing Berckley’s work with others’ who study affect and emotion in learning environments [15], [18], [40], his experiments were conducted in a much shorter span of time ( $< 40$  mins)

compared to others that were between 72-90 mins long [15], [40]. The session length can significantly impact emotion and affect outcomes, for example longer sessions are positively correlated with boredom [15]. Additionally, noticeable differences in results of AUs between studies may be attributable to fatigue [40]. Secondly, other researchers recorded students in individual sessions [15], [18], [40], while Berckley’s study was done in groups. The positive result of this group effect was that students reported feeling extra motivated, even sometimes competitive with other students to be the first to answer, and that this made finding the solution more satisfying and enjoyable [10].

### 3.3 Study Methods

#### 3.3.1 Insight Tasks for Inducing "Aha!" Moments

To study insight, we first need to create conditions to induce the occurrence of an “Aha!” moment. For this, researchers have used several different types of problem solving tasks. Gestalt psychologists introduced tasks such as the “Nine-dot problem” and “Dunker candle task” which are widely used in assessing insight problem-solving, however they have significant limitations that make them not ideal for controlled experiments including: (1) being so difficult that only a small percentage of participants successfully solve them in an experimentally reasonable amount of time, (2) they cannot be retested because of their single-trial nature [31].

To overcome these challenges, researchers developed a new generation of insight problems. These problems are generally based on verbal comprehension and are easier to solve when used with participants of the appropriate linguistic background [41]. Examples of this new generation of tasks include riddle-based tasks, anagrams, matchstick arithmetic problems, Chinese logogriphs, rebus puzzles, and the Remote Associates Task (RAT) and the Compound Remote Associations (CRA or CRAT) task. Both the RAT and CRAT task consist of three words and the participant must find a fourth related word (i.e. “falling, actor, dust”, solution: “star” → falling star, movie star, stardust). In the case of CRAT, that fourth word must form a compound word with the others (i.e. “crab, pine, sauce”, solution: “apple” → crabapple, pineapple, applesauce) [31].

The advantages of the RAT and CRAT that are critical for insight research include [31]:

- Timing - solved in short time so that many attempts can be done in a single experiment session (typically 15 seconds for CRAT tasks [5])
- Solutions - single-word, unambiguous solutions make scoring responses easier
- Size - physically compact which helps in various experimental settings
- Difficulty - there are varying degrees of difficulty

### 3.3.2 Data Collection Context

Berckley shared 126 videos that each ranged from 4 to 45 seconds and featured one of the 49 different student participants in the study. These volunteer participants were students at a Texas High School, were equally representative of male and female population, had an average age of 16.5 years ( $SD = 1.5$ , range = 15 to 18 years), and are fluent English speakers with normal or corrected-to-normal vision.

Students gathered in multiple different sessions of about 40 minutes, to answer a series of 26 CRAT questions hosted on a free online quiz making platform. The CRAT [41] is an adaptation of the RAT which was created in the 1960's by Mednick [42] as a measure for creative convergent thinking and creativity in general, and which has correlated scores on classical insight solving tasks [43]. The CRAT is one of the most widely used tests to assess insight performance.

Each video clip contains one of the 26 rounds of CRAT questions. Each student answered the questions using a laptop and also activated the laptop's camera to record their faces and upper torso movements. For each round, a CRAT challenge was presented, a timer counted down on the screen, and a meme was awarded after each challenge regardless of whether they answered correctly. Students started each round at the same time and were instructed to shout out the answer when they discovered the solution. Students then received confirmation from Berckley on the accuracy of their answer. Occasionally, Berckley asked which student got the answer right, and the student then raised their hand. If none of the students got the right answer, Berckley revealed the solution at the end of the round. After the sessions, Berckley analyzed the data using Affectiva software.



# Chapter 4

## Methods

### 4.1 Data Labeling

#### 4.1.1 Manual Annotation of "Aha!" Moments

The video clips needed to be labeled prior to using them in supervised machine learning. While some annotations could have been automated, we preferred to have human annotators manually label these videos. For labeling facial affect, researchers found that human annotators perform better than automated models [44].

For each video, annotators were asked to label whether or not they observed an “Aha!” moment (‘Y’ = yes, ‘N’ = no, ‘M’ = maybe), and in cases of ‘Y’ or ‘M’ they were asked to label the approximate start and end time of the occurrence. Because “Aha!” moments happen so suddenly, and often involve subtle changes in microexpressions, I created the following guidelines to aid annotators:

- Audio context: if present, the “Aha!” moment will occur sometime between when Berckley initiated the round and just after the first student says the solution aloud. If no student gets the answer correct, then the moment may occur directly after Berckley reveals the correct solution.
- Social context: some of the compound associations were particularly funny to the high school participants, for example in round 14 where the remote associate words were: cane, daddy, plum, and the compound word was sugar. This caused many of the students to burst out laughing likely because of the silliness or tinge of awkwardness they felt upon realizing the solution. Pay careful attention in these instances to differentiate between excitement and joy due to silliness versus due to insight.
- Emotion context: part of the definition of “Aha!” moments that is widely accepted by researchers is that an individual experiences a phase of frustration, confusion, or intense thinking and concentration followed by intense joy and excitement. The “Aha!” moment is said to occur in between the period of frustration and joy [4], as shown in Figure 4.1.

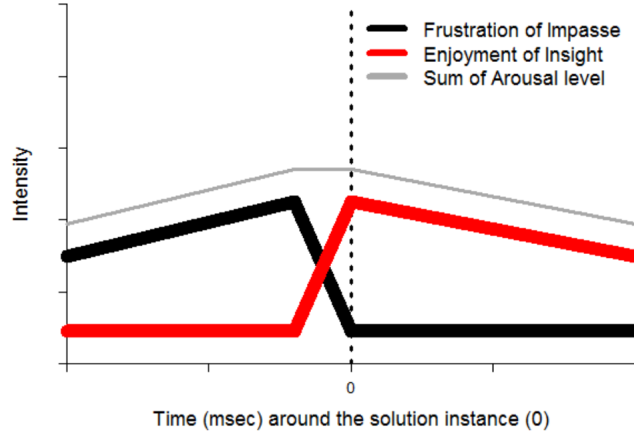


Figure 4.1: Illustrating the progression of affective components of "Aha!" moments as described by researchers [4]

In order to ensure the “Aha!” moment was captured, annotators were asked to record the start time during the initial frustration/thinking phase, about 2-3 seconds on average before the excitement phase begins, and subsequently record the end time about 2-3 seconds on average after the start of the excitement.

### Inter-Rater Reliability

Annotators manually labeled all of the videos and then tested the reliability of my labels by comparing them with 10% of the videos (randomly selected) that were labeled by another trained volunteer. The inter-rater reliability score was calculated as an average of each video’s Cohen’s Kappa scores. For each video, we compared the labels from the raters frame-by-frame. Despite being more effective at labeling affect than automated models [44], researchers report the difficulty that human annotators experience in detecting affect. Therefore, when coding for facial action units, basic gestures, and other visible behavior, researchers expect a kappa score greater than 0.6, and argue that a lower kappa score is meaningful when human judges are asked to infer complex mental states [15]. Considering the mysteriousness and lack of concrete, measurable guidelines for identifying “Aha!” moments, I’d argue that judging these moments, like we’ve done here, requires inferring complex mental states. In our case, we achieved an average Cohen’s Kappa score of 0.636 across the videos, which is expected and meaningful.

### Maybe-"Aha!" Moments

Any uncertain “Aha!” Moments (‘M’) were set aside and not used for this study. These included clips where the participant either did not answer correctly, or exhibited very low affect and it was difficult to assess their emotion and therefore label the event.

## Non-"Aha!" Moments

To balance the dataset, for each video clip containing an “Aha!” Moment (‘Y’), a subsequent non-”Aha!” Moment clip was randomly selected from the remaining video footage. Depending on the size of the remaining, non-overlapping time available in the rest of the video footage, the non-”Aha!” Moment start and end time were determined and the video length was between approximately 4-6 seconds.

## 4.2 Raw Feature Extraction

Using techniques and methods from Alghowinem [28], each video was run through MediaPipe Holistic Landmark Detection and Py-Feat pipelines to extract raw features per frame of the video. These included a total of 543 landmarks (33 pose, 468 face, and 21 hand landmarks per hand) from MediaPipe and 170 features from Py-Feat including: 4 face detection, 136 face landmarks (68 x, 68 y), 3 face and head pose estimation, 20 action units, and 7 emotion detections). Each of these features were reported per frame and approximate time in the video.

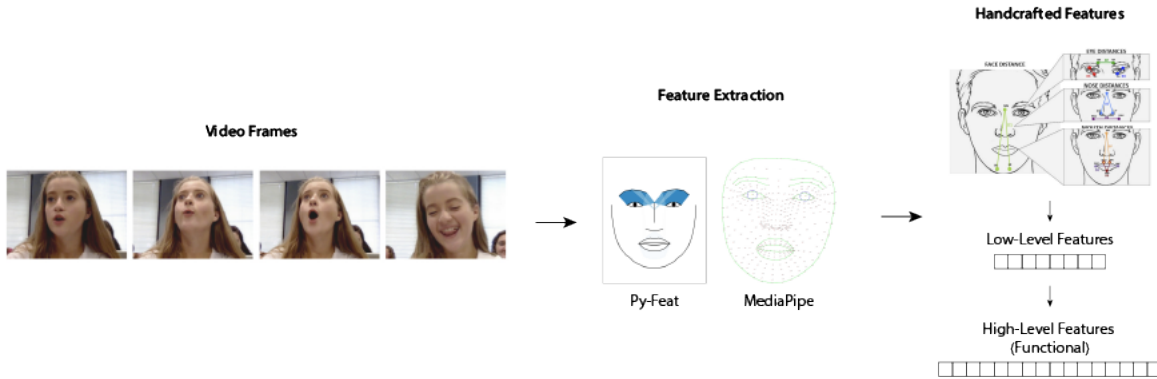


Figure 4.2: Overview of the process of creating the handcrafted features for modeling [5][45][46][10]

### 4.2.1 Relevant Tools

Several tools were utilized in this research to extract the raw data necessary for facial feature analysis. The main two discussed in this work are:

#### MediaPipe’s Holistic Landmark Detection

Google’s MediaPipe [5] provides a suite of libraries and tools to apply artificial intelligence and machine learning techniques in applications. We used one of MediaPipe’s models which are pre-trained, ready-to-run models. The model we used was a computer vision model for

holistic landmark detection, which means that it is a combination of pose, face, and hand landmarks that creates a complete landmarker for the human body. This allows us to analyze full-body gestures, poses, and actions, outputting a total of 543 landmarks (33 pose, 468 face, and 21 hand landmarks per hand) in real-time.

### **Python Facial Expression Analysis Toolbox (Py-Feat)**

To extract the action units and emotion data, we utilized the Python Facial Expression Analysis Toolbox (Py-Feat) [45]. This framework offers several different open source models as well as custom trained models to extract the desired features. For the action unit detection we used the default xbg model which is an XGBoost Classifier model trained on Histogram of Oriented Gradients extracted from a variety of datasets (BP4D, DISFA, CK+, UNBC-McMaster shoulder pain, and AFF-Wild2 datasets). For the emotion detection we used the default resmasknet model which does facial expression recognition using a residual masking network [47].

## **4.3 Data Pre-Processing**

### **4.3.1 Isolating Single Participant Face**

MediaPipe was run with a configuration to detect and process data from exactly one face. For Py-Feat, however, it outputs data for each face present per video frame. Ultimately, I needed to end up with a single frame per video and single face per video. There were several options available to isolate the desired face, including advanced computer vision techniques. Instead, I was able to notice a pattern in the data which would allow us to isolate the desired participant’s face. The target participant’s face in each video was the largest face present (other students were visible behind them and their faces were subsequently smaller). Therefore, for each set of frames, we selected the target face data as the row containing the largest face width (from the face detection features). This ensured that we ended up with only a single row of face data per video frame, and that it was of the target, desired face.

### **4.3.2 Data Normalization**

Using the raw extracted features, we use the same normalization techniques from [28]. This includes normalizing the face and body sizes using the body joints and facial landmarks, which accounts for within-participant variation (i.e. distance from the camera) and between-participant variation (i.e. differences in body size). The distance is normalized in reference to the distance between the sternum and clavicle (i.e. the distance between other points is divided by the distance between the sternum and clavicle). The two points were chosen as normalization references because they are rigid and therefore remain robust through various types of movement. Then, after the normalization, I used the Grubbs’ test [48] to detect and remove any frames containing outliers.



## 4.4 Data Composition

After including only "Aha!" and non-"Aha!" moments, and isolating the data to include only a single participant's face, we ended up with 172 video clips ('Y'=86, 'N'=86) from 36 participants.

## 4.5 Handcrafted Features

From the raw data we created both low and high-level features for each video (Figure 4.2).

### 4.5.1 Low-Level Features

Here we removed unrelated features, isolated a single face, and created meaning out of the raw data as described below:

#### MediaPipe

We needed to process the 543 landmarks (33 pose, 468 face, and 21 hand landmarks per hand) into meaningful features. The following describes the calculations by region as done by Alghowinem et. al. [28]:

- Eyes and Eyebrows:
  - Eye area: area of each eye normalized over the overall eye square area (similar to AU5 and AU7)
  - Eye openness: vertical distance between the eyelids normalized over the horizontal distance (similar to AU43)
  - Eyebrow area:
    - \* Raising and lowering eyebrows: area of each eyebrow normalized over the eye square area (similar to AU1 and AU2)
    - \* Between eyebrows: horizontal distance between inner eyebrow points normalized over the horizontal distance between the outer eyebrow points (similar to AU4)
- Nose and Cheeks:
  - Large nose area: area including tip of the nose, nose corners, and center of the eyebrow (similar to AU09)
  - Small nose area: tip of the nose, nose corners, and bottom of the nose and nostrils (similar to AU10)
  - Cheek area: area of each cheek which begins from under the eye to the edge of the face (similar to AU6 and AU11, and features related to talking and smiling)

- Mouth and Jaw:
  - Mouth openness: vertical distance between the outer lip points (features related to talking and smiling)
  - Lip corner distances: each lip corner distance from the jaw (similar to AU12 and AU15, features related to smiling)
  - Mouth width: horizontal distance between the corners of the mouth normalized over the horizontal distance between the face edges (features related to talking and smiling)
  - Inner and outer lips areas: each normalized over the whole face square area (similar to AU17, AU20, AU23, AU24, AU25, AU26, AU28)
- Face and Body:
  - Head motion and body motion
  - Head pose and body pose (pitch, roll, yaw)
- Arms and Hands:
  - Arm area: triangle area between shoulder, elbow, and wrist normalized over the body square area, not including face area (related to movements during talking and raising arm movements as common in classrooms and observed in this dataset)
  - Face touching: distance between wrist and nose (related to behaviors during thinking and concentration, frustration, anxiety which are observed in classrooms)
  - Body touching: distance between wrist and center of the body (related to behaviors during thinking and concentration, frustration, anxiety which are observed in classrooms)
  - Hands touching: distance between wrist of right and left hands (related to behaviors of fidgeting or anxiety which are observed in classrooms)

## Py-Feat

The raw data coming from Py-Feat did not require any processing as the information from the data was already clear and relevant and included: head position (pitch, roll, and yaw), and the AUs (Table 4.1). Py-Feat also includes data about seven emotions (anger, disgust, fear, happiness, sadness, surprise, neutral) and 68 face landmarks. We excluded this emotion and landmark data from our features as we hypothesized that “Aha!” moment has its own unique emotion expression, and favored MediaPipe’s more granular face mesh with 468 facial landmarks.

### 4.5.2 High-Level Features

For each video, we labeled a start and end times for both “Aha!” and non-”Aha!” moments. With high-level features, we aim to isolate the frames to include only the range between the








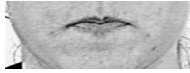




Inner brow raiser (AU1)		Chin raiser (AU17)	
Outer brow raiser (AU2)		Lip stretcher (AU20)	
Cheek raiser (AU6)		Lip Tightener (AU23)	
Nose wrinkler (AU9)		Lip pressor (AU24)	
Lip corner puller (AU12)		Lip part (AU25)	
Dimpler (AU14)		Jaw drop (AU26)	

Table 4.1: Sample of action units from FACS [8]

labeled start and end time of each video, and make sense of and summarize the low-level features through statistical analysis. This means that for each video, we ended up with two files, one for the “Aha!” moments and one for the non-”Aha!” moment.

Facial expressions are more than the presence, absence, or position of features, they also involve dynamic movement. To capture a deeper understanding of the changes in facial expressions and body movements over time, we derived first (d1) and second-order (d2) derivatives of the original features, speed and acceleration respectively. The raw features and their d1 and d2 features comprise the low-level descriptors (LLD). From the LLD, we calculated 10 different statistical functionals (minimum, maximum, range, mean, standard deviation, variance, kurtosis, skewness, peaks, valleys) to represent the dynamics of the movement and used these as inputs to the model.



# Chapter 5

## Modeling and Feature Selection

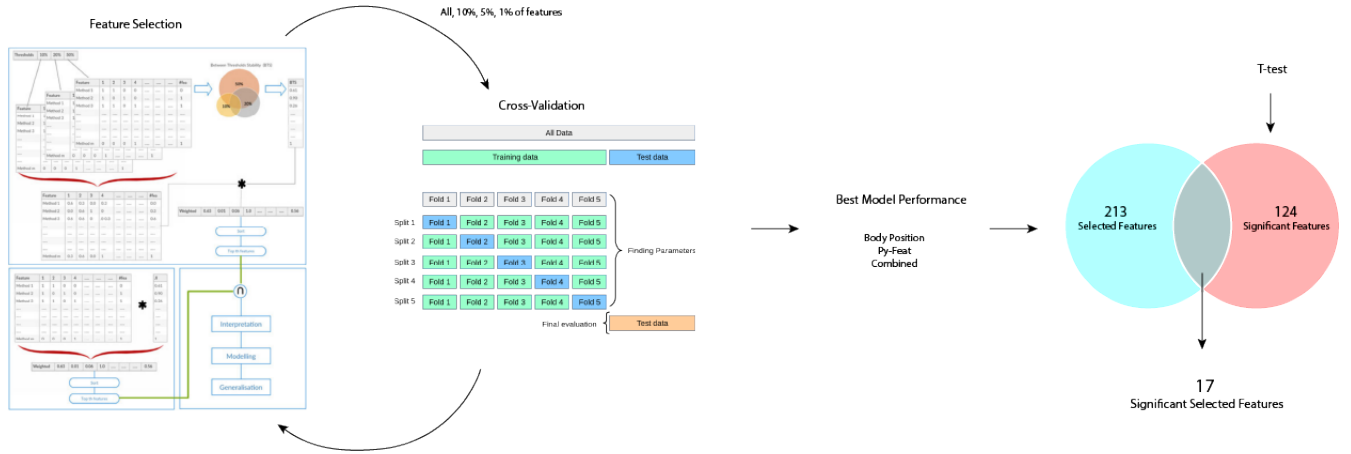


Figure 5.1: Overview of the modeling and feature selection process [6], [7]

### 5.1 Cross-Validation Using K-Fold Grouping

Cross-validation was used for modeling our dataset due to its relatively small size. K-fold grouping by participant was ideal in helping us prevent the model from overfitting to a participant. Therefore, we grouped the data by participant, and split the participants randomly to achieve approximately a 70:30 training-test split. Each participant had a different number of labeled “Aha!” and non-”Aha!” moments from the videos, therefore when the data was grouped by participant and split, the proportion of data in training versus test data may vary, which explains the approximate 70:30 split. Our validation dataset comprises 25% of the samples from our training set, which underwent a random stratification split to balance the samples for each class. Then we ran five rounds of cross-validation.

## 5.2 Architecture

Given the differences between the handcrafted features vector lengths (Table ). Like [28], we used Bayesian Optimization to search for and determine the optimal network architecture and hyperparameters. The optimization search configuration space was based on the following:

- Hidden Layers: working with already extracted features, we only needed a few dense layers. Therefore the configuration space options included 1-3 dense layers, all with ReLU activation function. An additional dense layer was included with Softmax activation function with two classes.
- Dropouts: 1-2 dropout layers were added between the hidden layers to avoid overfitting.
- Layer Size: Each hidden layer was fine-tuned using the varying input sizes.
- Optimizer and Learning Rate: We included SGD and ADAM optimizers at a range of learning rates and momentum (for SGD).

## 5.3 Comparing Model Performance

We compared the performance of three different models: (1) “Body Position” - MediaPipe features only, (2) “Py-Feat” - Py-Feat features only, and then (3) “Combined” - MediaPipe and Py-Feat features together. Comparing each of these models’ performance will help us understand and explain the resulting features and dynamics and evaluate any potential effects of feature imbalance since the body position model has more than twice the number of features as the Py-Feat model.

Comparing each model to its own performance (intra-model comparison) as well as between models (inter-model comparison) helps us identify the best performing model with which we can best interpret and explain the “Aha!” phenomenon.

## 5.4 Feature Selection

Our ultimate goal in modeling these “Aha!” moments, is to narrow down the features that are most important in detecting the affect present during “Aha!” moments. These methods will allow us to not only determine the important features but also interpret, explain, and make sense of them to aid in future research efforts. Narrowing down the features of these moments of learning includes running various feature selection methods and using results from a T-test to select the significant selected features for interpretation.

### 5.4.1 T-test

This statistical test compared features from “Aha!” moments to features from non-”Aha!” moments to determine the significant differences between them. All 2,130 features (MediaPipe + Py-Feat) were included for each group, and we applied a Bonferroni correction of

$2.35e^{-5}$  (alpha value = 0.05) to avoid the multiple comparison problem. After performing the T-test between “Aha!” and non-”Aha!” moments and applying the Bonferroni correction, 124 significant features were identified. The results of the T-test also output the T-statistic helping us understand if a significant feature describes our target affect of “Aha!” moments (positive value) or is a significant discriminant feature (negative value).

### 5.4.2 Robust Feature Selection through Aggregating Methods

We used various feature selection methods to determine the most important, discriminatory features at varying feature proportions (all features, 10%, 5%, 1%) for each of the three models. Comparing the performance of each of these models allows us to verify and validate the use of these feature selection methods in determining the best performing model. We used feature selection method results from [48], which similarly was used to assess affect features, as well as including methods that could be run in a reasonable amount of time ( $< 3$  days), to guide us in choosing the best methods to run our feature selection. Table 5.1 shows the methods used for our feature selection.

The feature selection occurred in two phases: (1) feature selection process phase, and (2) aggregation phase. The first phase ensures the stability, generalizability, and interpretability of the model by using ensembles of different splits of training data in multiple rounds of 16 feature selection methods and through the use of two different stability methods: Between Thresholds Stability (BTS) and Jaccard index (JI). Then, the aggregation phase selects the final, most robust and stable features based on the stability methods.

### 5.4.3 Determining the Features of “Aha!” Moments

Feature selection was run for each of the three model feature sets and the resulting selected features were then used to reduce the feature size for each model. The three models were then run with each of the proportional feature selections (all, 10%, 5%, 1%) and their performance was assessed for each. The results of each round of feature selection can be seen in Table 6.2. The selected features from the best performing model were then compared with the T-test results to determine the significant selected features and whether they described the “Aha!” moment or non-”Aha!” moment.

Group	Method
Tree Structured	Gradient Boosting
Statistics	T-test
	Chi Square
	CFS Continuous
Similarity-based	Fisher Score
	reliefF
	SPEC
Information Theory	JMI
	DISR
Embedded	LASSO
	L1-SVM
	Elastic nets
	HSICLasso
	Ridge
Wrappers	SVM-RFE
	SVM-Backward

Table 5.1: Feature selection methods used [6]



# Chapter 6

## Results and Discussion

### 6.1 Model Performance

Each of the three models’ performance was assessed after successive rounds of feature selection. Many performance metrics were calculated but the metric used to determine the “best” performing model was accuracy.

The results of the modeling and feature selection are in Table 6.1. Compared to using all of the features in the models, feature selection increased performance for the Combined and Py-Feat models. Despite feature selection not improving the Body Position model, from all to 5% of the features, we achieved a meaningful reduction in the number of features with just a small performance cost. These results support the benefits of using feature selection for both increasing performance and aiding in explainability as fewer features narrows to the most important features for interpretation.

	Accuracy		
Feature Size	Body Position	Py-Feat	Combined
All	82.3%	83.1%	83.6%
10%	83.4%	84.0%	89.3%
5%	78.9%	86.7%	88.0%
1%	60.3%	79.1%	78.9%

Table 6.1: Comparison of model performance at varying feature proportions

### 6.2 Significant Selected Features

Since the combined model at 10% feature selection was the best performing model, we used results from the T-test to identify the 17 significant features from the 213 selected features. They are included in Table 6.3 along with their trend as a feature to support “Aha!” moments

Model	Feature size (%)	AUC	Accuracy
Combined	213 (10%)	94.6%	89.3%
Py-Feat	35 (5%)	91.4%	86.7%
Body Position	144 (10%)	86.2%	83.4%

Table 6.2: Best results for each of the three models in order of performance

( $A > N$ ) or non-”Aha!” moments ( $N > A$ ). The resulting significant selected features came from three regions of the face: mouth, cheek, and eye.

Majority of the features came from the mouth region (13/17), and of those 13 features, only four described “Aha!” moments ( $A > N$ ), and the remaining nine described non-”Aha!” moments ( $N > A$ ). It is not surprising that majority of the significant selected features describe mouth features, since this describes behaviors such as talking and smiling which frequently occurred in this dataset as the students experienced joy when they got the correct answer, shouted out the answer, responded to questions from the teacher, or turned to their friends to talk or laugh. During the timespan labeled as “Aha!” moment, the students were mostly quiet and focused with very little mouth movement until they shouted out the answer and smiled at their success, whereas there was much more talking and laughing during the non-”Aha!” moments. Therefore, it’s also unsurprising that there are fewer significant mouth features describing “Aha!” moments.

We see the benefits of adding the first and second order derivative features to describe affective behaviors in the resulting significant features, as nine out of the 17 features described either first or second order derivative features. This gives us important dynamic descriptions of the affective behaviors of “Aha!” moments that have not previously been reported. Most of these first and second order derivative features are from the mouth region which is expected as the mouth has the largest range and dynamics of movement among all three significant regions as it describes movement related to talking, laughing, and smiling. Closely related to the mouth movements are the movements of the cheek, which understandably includes first and second order derivative feature descriptors as well.

### 6.3 Feature Explainability

We utilize the selected features from the best performing model, “both” at 10% selection, to interpret and identify the facial and body movements characteristic of the “Aha!” moment. From all of the selected features for the best performing model, we chose the features with significant differences between “Aha!” and non-”Aha!” moments to interpret the facial expression exhibited during “Aha!” moments. These 17 significant selected features are categorized by the feature region, name, statistical functional, and statistical trend (Table 6.3). The full list of features from this model can be found in Appendix ???. The significant selected features came from three regions of the face: cheek, eye, and mouth, and understanding of each selected feature within the facial region as well as identifying relationships between regions gives a richer explanation for the affect of “Aha!” moments. Examples of "Aha!"

Feature Region	Feature Name	Feature Functional	Statistic	Trend
Eye	AU01 - Inner Brow Raiser	Original	var	A >N
Cheek	Right Cheek Area	d1	range	A >N
	AU06 - Cheek Raiser	Original	std	
	AU06 - Cheek Raiser	d2	kurt	N >A
Mouth	Mouth Inner Aspect Ratio (h/w)	Original	range	A >N
	Mouth Height Right	d1	max	
	Mouth Outer Aspect Ratio (w/h)	Original	range	
	AU25 - Lips Part (mouth open)	Original	max	
	Mouth Inner Aspect Ratio (h/w)	d1	skew, kurt	N >A
		d2	skew, kurt	
	Mouth Height Left	Original	skew	
	Mouth Outer Aspect Ratio (w/h)	Original	min	
		d1	min	
	AU24 - Lip Pressor	Original	min	
		d1	min	

Table 6.3: The significant selected features from the best performing model, Combined at 10% selection, listed according to their feature region, name, statistical functional, and statistical trend.

Key: A = “Aha!” moments, N = non-“Aha!” moments, w = width, h = height, Original = raw feature, d1 = first-order derivative (speed), d2 = second-order derivative (acceleration), var = variance, std = standard deviation, kurt = kurtosis, skew = skewness, max = maximum, min = minimum.

moment facial expressions can be seen in Figure 6.1.

### 6.3.1 Cheek Movements

The resulting features described a large variability in the cheek raising intensity (AU06), which can be understood as a large variability in both raising and lowering the cheeks during “Aha!” moments. There was also a smaller range of speeds in the cheek area (particularly the right cheek) during “Aha!” moments. Together these features describe the “Aha!” moment as being characterized by large but more gradual movements of the cheeks. Smiles directly engage the cheeks, moving them up towards the eyes, and therefore creating the highest intensity of AU06. With the way we labeled the data, the “Aha!” moments include the moment right after the actual moment of insight where the person experiences immense joy and positive emotions, which accounts for the large intensity change of AU06 observed.

For non-“Aha!” moments, one of the resulting features describes cheek movements having

longer durations than those of “Aha!” moments. This is understandable, especially when considering the higher occurrence of laughter during non-“Aha!” moments which involves longer, more steady and sustained engagement of the cheeks or neutral expression which has a more steady engagement of the cheeks.

“Aha!” moments are characterized by a larger range of movement in the cheek area which is supported by feature functionals from both facial position and AU06 data. Considering how the “Aha!” moments were labeled (including a few seconds of thinking and concentration before the “Aha!” moment and followed by a few seconds of joy afterwards), the large range of movements are explainable. The moments of concentration typically do not engage the cheeks as much (AU06 minimum, and relaxed therefore larger cheek area), while the moments of joy typically involve smiling and maximally engaging the cheeks (AU06 maximum, engaged therefore smaller cheek area). These explanations align with Berckley’s findings as well. He reported AU06, cheek raiser as one of the significant features that are shown to be statistically different between the before, during, and after of “Aha!” moments, with the largest mean and standard deviation occurring during the “Aha!” moment itself (Table 6.3) [10].

### 6.3.2 Eyebrow Movements

The eyebrows, as described by the feature AU01 (inner brow raiser), have a wider variation in intensity values during the “Aha!” moment. Considering that the “Aha!” moments were labeled to include moments both before and after the moment itself, this wide variability makes sense. Beforehand, while thinking and concentrating and potentially even experiencing frustration, the eyebrows tend to be furrowed, and AU01 is at its lowest [49]. Afterwards, the eyebrows are typically raised when experiencing positive (joyful) surprise, meaning that AU01 is at its highest [49]. Moments of insight, by definition, occur suddenly, and it has been demonstrated that AU01 is a characteristic affective response to novel, sudden events [50].

Eyes are often tricky to study without use of specialized eye and gaze detection models. Students wearing glasses can also impact the detection. It’s worth noting that none of the students in this study wore glasses. The fact that these results have only one feature present representing the eye region should not deter, but rather invite further study of the eye movements using specialized models. In studies and contexts that involve speech, eyes can be especially useful features since they can be influenced by but are not directly involved in speech production [49]. Conversely, despite the majority of the resulting features representing the mouth, it is difficult to distinguish movements related to the target affect or unrelated speech. Therefore, features in the eye region can provide clearer model explainability and interpretability. Anecdotally, of the three regions represented in the significant selected features, the eyes are frequently mentioned by educators when asked about their observations of “Aha!” moments [10]. This human-level observability suggests that, when equipped with adequate feature details, models ought to identify these noticeable changes in facial expression. While the eye region can be useful, this is not to suggest it be studied in isolation. Just as we included derivative descriptors for the data, including other features and modalities is important in achieving a deeper understanding of the social, emotional behaviors.

### 6.3.3 Mouth Movements

The large majority of the significant selected features, 13 out of 17, were related to the mouth, demonstrating its importance in describing and classifying “Aha!” moments. For “Aha!” moments, the features describe a large range of motion for both the inner (without the lips) and outer (including the lips) mouth, show that AU25 (lips part, open mouth) is at its maximum intensity (mouth open with largest vertical distance), and the mouth height (right side) has the quickest change in position when the mouth is closing and the student starts to share the solution they suddenly discovered. These features describe the “Aha!” moment where the mouth opens the most (vertically) and then quickly closes as the student shares their solution (beginning speaking). These descriptions align with Berckley’s findings as well. He reported AU25 as one of the significant features shown to be significantly different between the before, during, and after of “Aha!” moments, with the largest mean and standard deviation occurring during the “Aha!” moment (Table 6.3).

AU24, lip pressor, is a significant feature that only describes movements of non-“Aha!” moments. AU24 has a larger minimum intensity and has a larger minimum speed compared to “Aha!” moments. The mouth height has a higher frequency of being open, the mouth including the lips has a larger minimum aspect ratio, and both the mouth inner and outer have more steady and even speeding up and slowing down. Together these can be understood to mean that the lip pressor does not fully disappear and it has larger changes in intensity, which are both descriptive of speaking movements. The more steady speed changes is more descriptive of laughing. AU24, lip pressor, is a lesser studied facial action unit due to its difficulty to detect (from human annotators) and the rareness of its occurrence. It is understood that an open mouth interferes with AU24 [51], therefore, AU24 should appear least frequently when the mouth is more frequently open for example during talking or laughing. More talking and open mouth laughing occurs during non-“Aha!” moments (especially after the “Aha!” moment). Speaking and laughter also explain the changes described by the body position selected mouth features for non-“Aha!” moments.

### 6.3.4 Related Movements: Mouth, Brow, Cheek

For events that have a positive valence, meaning a desirable event, AU01 and AU06 are found to cluster and exhibit similar behavior. This remains true in both high and low arousal positive valence emotions [51]. This coupling between the cheek and eyebrows, AU01 and AU06, is suggested in this data as the features describe a wide range and variation in the movements of both action units during “Aha!” moments. Further connection between AU06 and AU01 appears in the context of goal conducive appraisal, when a clear goal is present, like the participants experience during the CRAT [50].

Additionally, the features also describe the wide range and variation in eyebrow movement and mouth opening during “Aha!” moments, suggesting a connection between AU01 and AU25. Researchers have observed and reported this connection between AU01 and AU25 during sudden attention shifts [52] which is characteristic of “Aha!” moments [31].

## 6.4 The Affect of "Aha!" Moments

Based on the resulting significant selected features, it appears that the action units best characterizing the facial expression of the moment of insight are AU01, AU06, and AU25. These results, however, are preliminary pioneering work, and will require investigation with additional data (more in Chapter 7: Future Work).



Figure 6.1: Examples of before (A), during (B), and after (C) expressions which were included in the label of "Aha!" moment.

# Chapter 7

## Conclusion

Using dynamic, handcrafted features we were able to model the affect of “Aha!” moments, and leveraging feature selection and statistical methods we were able to provide explainability for our significant selected feature results. These led us to contribute a novel definition for the affective features of “Aha!” moments which include dynamic features in the eye, cheek, and mouth regions and specifically AU01, AU06, and AU25.

### 7.1 Limitations

While this work is the first to computationally model the affect of “Aha!” moments, it’s important to consider limitations impacting these results. The first limitation is that the scope of the moment of learning explored here does not fully encompass all types of learning, but rather is limited to learning by insight, and particularly insight that happens in the moment with no delay.

The second limitation comes from the dataset itself as well as the detection models used for creating the handcrafted features. The data represents a relatively small sample size with limited diversity which limits the generalizability of this study’s analysis of the affective traits of “Aha!” moments. Nonetheless, given the difficulty previous researchers experienced in capturing these moments of insight on video, we believe that this work provides meaningful preliminary insights. Additionally, our ability to extract meaningful explanations about the contribution of the eye-region to the affect of “Aha!” moments was limited due to the lack of specialized, fine-grain eye detection models used. This limited our ability to include features about eye gaze, blinks, and other micromovements in the eye region. Understanding these fine-grain eye behaviors could aid in linking the affective expression to underlying cognitive processes, thereby enabling more robust explainability.

A final limitation is the labeling of the “Aha!” moments to include moments right before and after the moment of insight. Since it is difficult to pinpoint the exact moment of insight, we used this range of time to ensure we fully captured the moment. However, this also means including behaviors and expressions which contextualizes the “Aha!” moment, but may not describe the singular moment itself.

## 7.2 Future Work

This work provides promising preliminary results for defining the affective expression of “Aha!” moments. Future work will expand upon these results and move towards the goal of detecting moments of insight in the wild. The highly sensitive nature of our facial expressions and body movements emphasizes the importance of analyzing the “Aha!” expression of people from different age groups, ethnicities, cultural upbringings, and personality types. It is also important to consider how different communication contexts may alter the “Aha!” expression. This includes comparing the changes in expressivity when students are one-on-one with a teacher, in a larger classroom environment, or one-on-one with a personalized agent such as a peer robot tutor. In addition to understanding the impact of learning contexts on expressions of “Aha!” moments, other communication contexts to consider include differentiating between the affective expression of negative and positive “Aha!” moments, and acted versus genuine expressions. Not all moments of insight are positive, individuals can also experience sudden troubling realizations [2] as well which may present a different affective expression. Students also may (intentionally or unintentionally) fake an expression of having a moment of insight to convey their understanding towards someone they are communicating with even if no learning took place. To effectively identify moments of learning, it is essential to confirm the presence of genuine learning.

In addition to exploring different contexts of “Aha!” expressions, future work should explore different technical methods to improve the detection and applicability of deploying this model in the real world. To account for varying use cases of the technology, it is important to understand the minimum distinguishable distance and granularity required for accurate detection of the “Aha!” expression. This would help determine the required proximity and specifications of camera equipment needed to accurately detect these moments. Furthermore, determining the essential features required for detection, through ablation studies or otherwise, could aid in boosting the robustness of the model and effective use in real life scenarios. These essential features may include additional modalities or detection models not used in this work, including more specialized eye detection models, and including multi-modal data such as audio.

Future work is not limited to artificial intelligence or machine learning. The nature of this work is interdisciplinary, and therefore collaborative exploration is encouraged.



# Appendix A

## Complete List of Handcrafted Features

Each feature listed below includes an original, first (d1), and second (d2) order derivative of the statistical methods. Here is one example of a complete set of features:

Head Pitch Original Min, Head Pitch Original Max, Head Pitch Original Mean, Head Pitch Original Range, Head Pitch Original Standard Deviation, Head Pitch Original Variance, Head Pitch Original Skewness, Head Pitch Original Kurtosis, Head Pitch Original Peaks, Head Pitch Original Valleys, Head Pitch d1 Min, Head Pitch d1 Max, Head Pitch d1 Mean, Head Pitch d1 Range, Head Pitch d1 Standard Deviation, Head Pitch d1 Variance, Head Pitch d1 Skewness, Head Pitch d1 Kurtosis, Head Pitch d1 Peaks, Head Pitch d1 Valleys, Head Pitch d2 Min, Head Pitch d2 Max, Head Pitch d2 Mean, Head Pitch d2 Range, Head Pitch d2 Standard Deviation, Head Pitch d2 Variance, Head Pitch d2 Skewness, Head Pitch d2 Kurtosis, Head Pitch d2 Peaks, Head Pitch d2 Valleys

### A.0.1 MediaPipe - Body Position Features

#### Body and Head Position

Head Pitch, Head Roll, Head Yaw, Body Pitch, Body Roll, Body Yaw, Head Body Pitch, Head Body Roll, Head Body Yaw

#### Arm Region

Left Hand to Face, Right Hand to Face, Left Hand to Body, Right Hand to Body, Right Hand to Left Hand, Left Arm Area, Right Arm Area

#### Eye Region

Right Eyelids Distance, Left Eyelids Distance, Eyelids Ratio, Left Eye Area, Right Eye Area, Left Eyebrows Area, Right Eyebrows Area, Eyebrows Distance, Right Eyebrow Aspect Ratio, Left Eyebrow Aspect Ratio, Average Eyebrow Aspect Ratio, Right Brow to Eyelid Distance, Left Brow to Eyelid Distance, Average Brows Raise, Brows Relative Raise

## **Nose and Cheek Region**

Large Nose Area, Small Nose Area, Left Cheek Area, Right Cheek Area

## **Mouth Region**

Mouth Inner Aspect Ratio, Mouth Frown Right, Mouth Frown Left, Mouth Frown, Nose to Mouth, Mouth Snarl Right, Mouth Snarl Left, Mouth Width, Mouth Height Right, Mouth Height Left, Mouth Height, Mouth Outer Aspect Ratio, Large Mouth Area, Small Mouth Area

## **A.0.2 Py-Feat**

### **Head Position**

Pitch, Roll, Yaw

### **Action Units**

AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU11, AU12, AU14, AU15, AU17, AU20, AU23, AU24, AU25, AU26, AU28, AU43

# References

- [1] R. S. J. D. Baker, A. B. Goldstein, and N. T. Heffernan, “Detecting the moment of learning,” in *Intelligent Tutoring Systems*, V. Aleven, J. Kay, and J. Mostow, Eds., red. by D. Hutchison, T. Kanade, J. Kittler, *et al.*, vol. 6094, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 25–34, ISBN: 978-3-642-13387-9 978-3-642-13388-6. DOI: 10.1007/978-3-642-13388-6\_7. URL: [http://link.springer.com/10.1007/978-3-642-13388-6\\_7](http://link.springer.com/10.1007/978-3-642-13388-6_7).
- [2] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard, “Affect-aware tutors: Recognising and responding to student affect,” *International Journal of Learning Technology*, vol. 4, no. 3, p. 129, 2009, ISSN: 1477-8386, 1741-8119. DOI: 10.1504/IJLT.2009.028804. URL: <http://www.inderscience.com/link.php?id=28804>.
- [3] J. Berckley and J. Hattie, “Making learning visible: Observable correlates of the aha! moment when moving from surface to deep thinking,” *The Journal of Creative Behavior*, May 6, 2023. DOI: 10.1002/jocb.589.
- [4] T. J. Wong, “CAPTURING ‘AHA!’ MOMENTS OF PUZZLE PROBLEMS USING PUPILLARY RESPONSES AND BLINKS,”
- [5] C. Lugaresi, J. Tang, H. Nash, *et al.*, *MediaPipe: A framework for building perception pipelines*, Jun. 14, 2019. arXiv: 1906.08172[cs]. URL: <http://arxiv.org/abs/1906.08172> (visited on 05/15/2024).
- [6] S. Alghowinem, T. Gedeon, R. Goecke, J. F. Cohn, and G. Parker, “Interpretation of depression detection models via feature selection methods,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 133–152, Jan. 1, 2023, ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2020.3035535. URL: <https://ieeexplore.ieee.org/document/9253541/>.
- [7] “3.1. cross-validation: Evaluating estimator performance,” scikit-learn. (), URL: [https://scikit-learn/stable/modules/cross\\_validation.html](https://scikit-learn/stable/modules/cross_validation.html).
- [8] P. Ekman and W. V. Friesen, “Facial action coding system (FACS),” *APA PsycTests*, 1978. DOI: 10.1037/t27734-000.
- [9] D. Candland, “A little learning...,” in *The Teaching of Psychology in Autobiography: Perspectives from Exemplary Psychology Teachers*, Society for the Teaching of Psychology, 2005.

- [10] J. R. Berckley, “What are the observable correlates of the aha! moment, and how does this moment relate to moving from surface to deep thinking?” Accepted: 2019-06-24T19:29:38Z, Thesis, Dec. 2018. URL: <https://uh-ir.tdl.org/handle/10657/4067> (visited on 07/05/2023).
- [11] G. Sprugnoli, S. Rossi, A. Emmendorfer, A. Rossi, S.-L. Liew, E. Tatti, G. di Lorenzo, A. Pascual-Leone, and E. Santarnecchi, “Neural correlates of eureka moment,” *Intelligence*, vol. 62, pp. 99–118, May 1, 2017, ISSN: 0160-2896. DOI: 10.1016/j.intell.2017.03.004. URL: <https://www.sciencedirect.com/science/article/pii/S0160289616302756>.
- [12] A. H. Danek, T. Fraps, A. Von MÃ¼ller, B. Grothe, and M. Ã¼llinger, “It’s a kind of magic! what self-reports can reveal about the phenomenology of insight problem solving,” *Frontiers in Psychology*, vol. 5, Dec. 8, 2014, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.01408. URL: <http://journal.frontiersin.org/article/10.3389/fpsyg.2014.01408/abstract>.
- [13] L. R. Novick and S. J. Sherman, “On the nature of insight solutions: Evidence from skill differences in anagram solution,” *The Quarterly Journal of Experimental Psychology Section A*, vol. 56, no. 2, pp. 351–382, Feb. 2003, ISSN: 0272-4987, 1464-0740. DOI: 10.1080/02724980244000288. URL: <http://journals.sagepub.com/doi/10.1080/02724980244000288>.
- [14] L. Aziz-Zadeh, J. T. Kaplan, and M. Iacoboni, ““aha!”: The neural correlates of verbal insight solutions,” *Human Brain Mapping*, vol. 30, no. 3, pp. 908–916, Mar. 2009, ISSN: 1065-9471, 1097-0193. DOI: 10.1002/hbm.20554. URL: <https://onlinelibrary.wiley.com/doi/10.1002/hbm.20554>.
- [15] S. K. D’Mello, S. D. Craig, and A. C. Graesser, “Multimethod assessment of affective experience and expression during deep learning,” *International Journal of Learning Technology*, vol. 4, no. 3, p. 165, 2009, ISSN: 1477-8386, 1741-8119. DOI: 10.1504/IJLT.2009.028805. URL: <http://www.inderscience.com/link.php?id=28805>.
- [16] K. G. Creswell, M. A. Sayette, J. W. Schooler, A. G. C. Wright, and L. E. Pacilio, “Visceral states call for visceral measures: Verbal overshadowing of hunger ratings across assessment modalities,” *Assessment*, vol. 25, no. 2, pp. 173–182, Mar. 2018, ISSN: 1073-1911, 1552-3489. DOI: 10.1177/1073191116645910. URL: <http://journals.sagepub.com/doi/10.1177/1073191116645910>.
- [17] J. Collins, H. Regenbrecht, T. Langlotz, Y. Said Can, C. Ersoy, and R. Butson, “Measuring cognitive load and insight: A methodology exemplified in a virtual reality learning context,” in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Beijing, China: IEEE, Oct. 2019, pp. 351–362, ISBN: 978-1-72810-987-9. DOI: 10.1109/ISMAR.2019.00033. URL: <https://ieeexplore.ieee.org/document/8943688/>.
- [18] S. Craig, A. Graesser, J. Sullins, and B. Gholson, “Affect and learning: An exploratory look into the role of affect in learning with AutoTutor,” *Journal of Educational Media*, vol. 29, no. 3, pp. 241–250, Oct. 2004, ISSN: 1358-1651. DOI: 10.1080/1358165042000283101. URL: <http://www.tandfonline.com/doi/abs/10.1080/1358165042000283101>.
- [19] S. Spaulding, G. Gordon, and C. Breazeal, “Affect-aware student models for robot tutors,”

- [20] S. L. Spaulding, “Lifelong personalization for social robot learning companions,”
- [21] H. Chen, H. W. Park, and C. Breazeal, “Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children’s learning and emotive engagement,” *Computers & Education*, vol. 150, p. 103 836, Jun. 2020, ISSN: 03601315. DOI: 10.1016/j.compedu.2020.103836. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0360131520300373>.
- [22] R. Pelánek, “Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques,” *User Modeling and User-Adapted Interaction*, vol. 27, no. 3, pp. 313–350, Dec. 2017, ISSN: 0924-1868, 1573-1391. DOI: 10.1007/s11257-017-9193-2. URL: <http://link.springer.com/10.1007/s11257-017-9193-2>.
- [23] R. Picard, *Affective Computing*. The MIT Press, 2000, ISBN: 978-0-262-28158-4. URL: <https://doi.org/10.7551/mitpress/1140.001.0001>.
- [24] A. Mehrabian, *Silent Messages*. 1971.
- [25] H. Oster, *Baby FACS: Facial action coding system for infants and young children*.
- [26] J. Metcalfe and D. Wiebe, “Intuition in insight and noninsight problem solving,” *Memory & Cognition*, vol. 15, no. 3, pp. 238–246, May 1987, ISSN: 0090-502X, 1532-5946. DOI: 10.3758/BF03197722. URL: <http://link.springer.com/10.3758/BF03197722>.
- [27] R. E. Laukkonen, D. J. Ingledew, H. J. Grimmer, J. W. Schooler, and J. M. Tangen, “Getting a grip on insight: Real-time and embodied aha experiences predict correct solutions,” *Cognition and Emotion*, vol. 35, no. 5, pp. 918–935, Jul. 4, 2021, ISSN: 0269-9931, 1464-0600. DOI: 10.1080/02699931.2021.1908230. URL: <https://www.tandfonline.com/doi/full/10.1080/02699931.2021.1908230>.
- [28] S. Alghowinem, X. Zhang, C. Breazeal, and H. W. Park, “Multimodal region-based behavioral modeling for suicide risk screening,” *Frontiers in Computer Science*, vol. 5, p. 990 426, Apr. 20, 2023, ISSN: 2624-9898. DOI: 10.3389/fcomp.2023.990426. URL: <https://www.frontiersin.org/articles/10.3389/fcomp.2023.990426/full>.
- [29] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, “Drowsy driver detection through facial movement analysis,” in *Human–Computer Interaction*, M. Lew, N. Sebe, T. S. Huang, and E. M. Bakker, Eds., vol. 4796, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 6–18, ISBN: 978-3-540-75772-6. DOI: 10.1007/978-3-540-75773-3\_2. URL: [http://link.springer.com/10.1007/978-3-540-75773-3\\_2](http://link.springer.com/10.1007/978-3-540-75773-3_2).
- [30] A. Einstein, A. Calaprice, and F. Dyson, *The Ultimate Quotable Einstein*. Princeton University Press, 2010, page 377.
- [31] E. Bowden, M. Jungbeeman, J. Fleck, and J. Kounios, “New approaches to demystifying insight,” *Trends in Cognitive Sciences*, vol. 9, no. 7, pp. 322–328, Jul. 2005, ISSN: 13646613. DOI: 10.1016/j.tics.2005.05.012. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1364661305001531>.

- [32] J. Kounios and M. Beeman, “The aha! moment: The cognitive neuroscience of insight,” *Current Directions in Psychological Science*, vol. 18, no. 4, pp. 210–216, 2009, Publisher: [Association for Psychological Science, Sage Publications, Inc.], ISSN: 0963-7214. URL: <https://www.jstor.org/stable/20696033>.
- [33] V. Charisi, N. Díaz-Rodríguez, B. Mawhin, and L. Merino, “On children’s exploration, aha! moments and explanations in model building for self-regulated problem-solving,”
- [34] R. Nkambou, “A framework for affective intelligent tutoring systems,” in *2006 7th International Conference on Information Technology Based Higher Education and Training*, Ultimo, Australia: IEEE, Jul. 2006, nil2–nil8, ISBN: 978-1-4244-0405-6 978-1-4244-0406-3. DOI: 10.1109/ITHET.2006.339720. URL: <http://ieeexplore.ieee.org/document/4141729/>.
- [35] L. Zhang, M. Jiang, D. Farid, and M. Hossain, “Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot,” *Expert Systems with Applications*, vol. 40, no. 13, pp. 5160–5168, Oct. 2013, ISSN: 09574174. DOI: 10.1016/j.eswa.2013.03.016. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0957417413001668>.
- [36] A. Arguel, L. Lockyer, O. V. Lipp, J. M. Lodge, and G. Kennedy, “Inside out: Detecting learners’ confusion to improve interactive digital learning environments,” *Journal of Educational Computing Research*, vol. 55, no. 4, pp. 526–551, Jul. 2017, ISSN: 0735-6331, 1541-4140. DOI: 10.1177/0735633116674732. URL: <http://journals.sagepub.com/doi/10.1177/0735633116674732>.
- [37] M. M. T. Rodrigo, R. S. Baker, M. C. Jadud, and A. C. M. Amarra, “Affective and behavioral predictors of novice programmer achievement,”
- [38] B. Kort, R. Reilly, and R. Picard, “An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion,” in *Proceedings IEEE International Conference on Advanced Learning Technologies*, Madison, WI, USA: IEEE Comput. Soc, 2001, pp. 43–46, ISBN: 978-0-7695-1013-2. DOI: 10.1109/ICALT.2001.943850. URL: <http://ieeexplore.ieee.org/document/943850/>.
- [39] M. Csikszentmihalyi, “Flow – the psychology of optimal experience,”
- [40] J. Lee, H.-J. So, S. Ha, E. Kim, and K. Park, “Unpacking academic emotions in asynchronous video-based learning: Focusing on korean learners’ affective experiences,” *The Asia-Pacific Education Researcher*, vol. 30, no. 3, pp. 247–261, Jun. 2021, ISSN: 0119-5646, 2243-7908. DOI: 10.1007/s40299-021-00565-x. URL: <https://link.springer.com/10.1007/s40299-021-00565-x>.
- [41] T. Toivainen, A.-M. Olteteanu, V. Repeykova, M. Likhanov, and Y. Kovas, “Visual and linguistic stimuli in the remote associates test: A cross-cultural investigation,” *Frontiers in Psychology*, vol. 10, p. 926, Apr. 26, 2019, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.00926. URL: <https://www.frontiersin.org/article/10.3389/fpsyg.2019.00926/full>.
- [42] S. Mednick, “The associative basis of the creative process,” *Psychological Review*, vol. 69, no. 3, pp. 220–232, 1962, ISSN: 0033-295X. DOI: 10.1037/h0048850. URL: <https://doi.apa.org/doi/10.1037/h0048850>.

- [43] J. W. Schooler and J. Melcher, “The ineffability of insight. the creative cognition approach,” in Cambridge (Mass.): MIT, 1997, pp. 97–133.
- [44] D. Dupré, E. G. Krumhuber, D. Küster, and G. J. McKeown, “A performance comparison of eight commercially available automatic classifiers for facial affect recognition,” *PLOS ONE*, vol. 15, no. 4, S. D’Mello, Ed., e0231968, Apr. 24, 2020, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0231968. URL: <https://dx.plos.org/10.1371/journal.pone.0231968>.
- [45] J. H. Cheong, E. Jolly, T. Xie, S. Byrne, M. Kenney, and L. J. Chang, “Py-feat: Python facial expression analysis toolbox,” *Affective Science*, vol. 4, no. 4, pp. 781–796, Dec. 2023, ISSN: 2662-2041, 2662-205X. DOI: 10.1007/s42761-023-00191-4. URL: <https://link.springer.com/10.1007/s42761-023-00191-4>.
- [46] E. Zane, Z. Yang, L. Pozzan, T. Guha, S. Narayanan, and R. B. Grossman, “Motion-capture patterns of voluntarily mimicked dynamic facial expressions in children and adolescents with and without ASD,” *Journal of Autism and Developmental Disorders*, vol. 49, no. 3, pp. 1062–1079, Mar. 2019, ISSN: 0162-3257, 1573-3432. DOI: 10.1007/s10803-018-3811-7. URL: <http://link.springer.com/10.1007/s10803-018-3811-7>.
- [47] L. Pham, T. H. Vu, and T. A. Tran, “Facial expression recognition using residual masking network,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy: IEEE, Jan. 10, 2021, pp. 4513–4519, ISBN: 978-1-72818-808-9. DOI: 10.1109/ICPR48806.2021.9411919. URL: <https://ieeexplore.ieee.org/document/9411919/>.
- [48] F. E. Grubbs, “Procedures for detecting outlying observations in samples,” *Technometrics*, vol. 11, no. 1, pp. 1–21, Feb. 1969, ISSN: 0040-1706, 1537-2723. DOI: 10.1080/00401706.1969.10490657. URL: <http://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657>.
- [49] P. Ekman, “About brows: Emotional and conversational signals,” in *Human ethology: Claims and limits of a new discipline*, Cambridge, UK: Cambridge University Press., 1979.
- [50] M. Mortillaro, M. Mehu, and K. R. Scherer, “Subtly different positive emotions can be distinguished by their facial expressions,” *Social Psychological and Personality Science*, vol. 2, no. 3, pp. 262–271, May 2011, ISSN: 1948-5506, 1948-5514. DOI: 10.1177/1948550610389080. URL: <http://journals.sagepub.com/doi/10.1177/1948550610389080>.
- [51] E. G. Krumhuber and K. R. Scherer, “Affect bursts: Dynamic patterns of facial expression,” *Emotion*, vol. 11, no. 4, pp. 825–841, Aug. 2011, ISSN: 1931-1516, 1528-3542. DOI: 10.1037/a0023856. URL: <https://doi.apa.org/doi/10.1037/a0023856>.
- [52] A. Gaspar and F. G. Esteves, “Preschooler’s faces in spontaneous emotional contexts—how well do they match adult facial expression prototypes?” *International Journal of Behavioral Development*, vol. 36, no. 5, pp. 348–357, Sep. 2012, ISSN: 0165-0254, 1464-0651. DOI: 10.1177/0165025412441762. URL: <http://journals.sagepub.com/doi/10.1177/0165025412441762>.