Automatic Recognition of Teacher Engagement from Facial Expressions and Voice Patterns

A

Major Project

Submitted

In partial fulfillment

For the award of the Degree of

Bachelor of Technology

In Department of Computer Science & Engineering

Submitted to:

Dr. Surendra Kumar Yadav

Guided by:

Jaskirat Singh

Submitted By:

Jitender Singh Virk

1302041054

Department of *Computer Science* & Engineering

JECRC UNIVERSITY

Ramchandrapura, Jaipur

CERTIFICATE

This is to certify that Project Report entitled “Automatic Recognition of Teacher Engagement from Facial Expressions and Voice Patterns” which is submitted by Jitender Singh Virk in partial fulfillment of the requirement for the award of B. Tech. degree in department of Computer Science and Engineering is a record of the candidates own work carried out by him under my supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Jaskirat Singh

Assistant Professor-II

10 Jan, 2017

DECLARATION

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Jitender Singh Virk

1302041054

10 Jan, 2017

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Jaskirat Singh for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Prof. Naveen Hemrajani Head, Department of Computer Science and Engineering, for his full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

ABSTRACT

Teacher engagement is a key concept in contemporary education, where it is valued as a goal in its own right. In this project we explore approaches for automatic recognition of engagement from teachers' facial expressions. We studied whether human observers can reliably judge engagement from the face; analysed the signals observers use to make these judgments; and automated the process using machine learning. We found that human observers reliably agree when discriminating low versus high degrees of engagement (Cohen's _ = 0:96). When fine discrimination is required (2 distinct levels) the reliability decreases, but is still quite high (_ = 0:56). Furthermore, we found that engagement labels of 10-second video clips can be reliably predicted from the average labels of their constituent frames (Pearson r = 0:85), suggesting that static expressions contain the bulk of the information used by observers.

We used machine learning to develop automatic engagement detectors and found that for binary classification (e.g., high engagement versus low engagement), automated engagement detectors perform with comparable accuracy to humans. Finally, we show that both human and automatic engagement judgments correlate with task performance. In our experiment, teacher post-test performance was predicted with comparable accuracy from engagement labels (r = 0:47) as from pre-test scores (r = 0:44).

CONTENTS

Chapter 1

INTRODUCTION

"The test of successful education is not the amount of knowledge that pupils take away from school, but their appetite to know and their capacity to learn."
Sir Richard Livingstone, 1941.

Student engagement has been a key topic in the education literature since the 1980s. Early interest in engagement was driven in part by concerns about large drop-out rates and by statistics indicating that many students, estimated between 25% and 60%, reported being chronically bored and disengaged in the classroom. Statistics such as these led educational institutions to treat student engagement not just as a tool for improving grades but as an independent goal unto itself.

Nowadays, fostering student engagement is relevant not just in traditional classrooms but also in other learning settings such as educational games, intelligent tutoring systems (ITS) and massively open online courses (MOOCs).

The education research community has developed various taxonomies for describing student engagement.

Fredrick, et al analysed 44 studies and proposed that there are 3 different forms of engagement: behavioural, emotional, and cognitive. Anderson, et al organized engagement into behavioural, academic, cognitive, and psychological dimensions. The term behavioural engagement is typically used to describe the student's willingness to participate in the learning process, e.g., attend class, stay on task, submit required work, and follow the teacher's direction. Emotional engagement describes a student's emotional attitude towards learning – it is possible, for example, for students to perform their assigned work well, but still dislike or be bored by it. Such students would have high behavioural engagement but low emotional engagement. Cognitive engagement refers to learning in a way that maximizes a person's cognitive abilities, including focused attention, memory, and creative thinking.

The goal of increasing student engagement has motivated the interest in methods to measure it. Currently the more popular tools for measuring engagement include: (1) Self-reports, (2) Observational checklists and ratings scales, and (3) Automated measurements.

1.1 Self-reports: Self-reports are questionnaires in which students report their own level of attention, distraction, excitement, or boredom These surveys need not directly ask the students explicitly how "engaged" they feel but instead can infer engagement as an explanatory latent variable from the survey responses, e.g., using factor analysis. Self-reports are undoubtedly useful. For example, it is of interest to know that between 25% and 60% of middle school students report to be bored and disengaged. Yet self-reports also have well-known limitations. For example, some students may think it is "cool" to say they are non-engaged; other students may think it is embarrassing to say so. Self-reports may be biased by primacy and regency memory effects. Students may also differ dramatically in their own sense of what it means to be engaged.

1.2 Observational checklists and rating scales: Another popular way to measure engagement relies on questionnaires completed by external observers such as teachers.

These questionnaires may ask the teacher's subjective opinion of how engaged their students are. They may also contain checklists for objective measures that are supposed to indicate engagement. For example, do the students sit quietly? Do they do their homework? Are they

on time? Do they ask questions? in some cases, external observers may rate engagement based on live or pre-recorded videos of educational activities.
Observers may also consider samples of the student's work such as essays, projects, and class notes.
While both self-reports and observational checklists and ratings are useful, they are still very primitive: they lack temporal resolution, they require a great deal of time and effort from students and observers, and they are not always clearly related to engagement. For example, engagement metrics such as "sitting quietly", "good behaviour", and "no tardy cards" appear to measure compliance and willingness to adhere to rules and regulations rather than engagement per se.

1.3 Automated measurements: The intelligent tutoring systems (ITS) community has pioneered the use of automated, real-time measures of engagement. A popular technique for estimating engagement in ITS is based on the timing and accuracy of students' responses to practice problems and test questions. This technique has been dubbed "engagement tracing" in analogy to the standard "knowledge tracing" technique used in many ITS. For instance, chance performance on easy questions or very short response times might be used as an indication that the student is not engaged and is simply giving random answers to questions without any effort.
Probabilistic inference can be used to assess whether the observed patterns of time/accuracy are more consistent with an engaged or a disengaged student.
Another class of automated engagement measurement is based on physiological and neurological sensor readings. In the neuroscience literature, engagement is typically equated with level of arousal or alertness.
Physiological measures such as EEG, blood pressure, heart rate, or galvanic skin response have been used to measure engagement and alertness. However, these measures require specialized sensors and are difficult to use in large-scale studies.
A third kind of automatic engagement recognition – which is the subject of this paper – is based on computer vision. Computer vision offers the prospect of unobtrusively estimating a student's engagement by analysing cues from the face, body posture and hand gestures. While vision-based methods for engagement measurement have been pursued previously by the ITS community, much work remains to be done before automatic systems are practical in a wide variety of settings.
If successful, a real-time student engagement recognition system could have a wide range of applications:
(1) Automatic tutoring systems could use real-time engagement signals to adjust their teaching strategy the way good teachers do. So-called affect-sensitive ITS are a hot topic in the ITS research community and some of the first fully-automated closed-loop that ITS use affective sensors for feedback are starting to emerge. (2) Human teachers in distance-learning environments could get real-time feedback about the level of engagement of their audience.
(3) Audience responses to educational videos could be used automatically to identify the parts of the video when the audience becomes disengaged and to change them appropriately.
(4) Educational researchers could acquire large amounts of data to data-mine the causes and variables that affect student engagement.
These data would have very high temporal resolution when compared to self-report and questionnaires. (5) Educational institutions could monitor student engagement and intervene before it is too late.

1.4 Contributions: In this project we document one of the most thorough studies to-date of computer vision techniques for automatic teacher engagement recognition. In particular, we

study techniques for data annotation, including the timescale of labelling; we compare state-of-the-art computer vision algorithms for automatic engagement detection; and we investigate correlations of engagement with task performance.

1.5 Conceptualization of engagement: Our goal is to estimate perceived engagement, i.e., teacher engagement as judged by an external observer. The underlying logic is that since teachers rely on perceived engagement to adapt their teaching behaviour, then automating perceived engagement is likely to be useful for a wide range of educational applications. We hypothesize that a good deal of the information used by humans to make engagement judgements is based on the teacher's face.

Our project is organized as follows: First we study whether human observers reliably agree with each other when estimating teacher engagement from facial expressions.

Next we use machine learning methods to develop automatic engagement detectors. We investigate which signals are used by the automatic detectors and by humans when making engagement judgments. Finally, we investigate whether human and automated engagement judgments correlate with task performance.

Chapter 2

ENGAGEMENT

2.1 What Is Engagement?

In this section, we describe how the three types of engagement have been defined, how the definitions vary, and where they overlap. Although we present behavioral, emotional, and cognitive engagement separately, we note where studies combine components of engagement. Finally, we discuss how these definitions resemble other motivational and cognitive constructs and how the literature on those constructs can inform the research on engagement.

2.1.1 Behavioral Engagement

Behavioral engagement is most commonly defined in three ways. The first definition entails positive conduct, such as following the rules and adhering to classroom norms, as well as the absence of disruptive behaviors such as skipping school and getting in trouble (Finn, 1993; Finn, Pannozzo, & Voelkl, 1995; Finn & Rock, 1997). the second definition concerns involvement in learning and academic tasks and includes behaviors such as effort, persistence, concentration, attention, asking questions, and contributing to class discussion (Birch & Ladd, 1997; Finn et al. 1995; Skinner & Belmont, 1993). A third definition involves participation in school-related activities such as athletics or school governance (Finn, 1993; Finn et al., 1995).In general, these definitions do not make distinctions among various types of behavior, such as participation in academic and nonacademic school activities.

One exception is Finn's (1989) definition of behavioral engagement. He divides participation into four levels, which range from responding to the teacher's directions to activities that require student initiative, such as involvement in extracurricular activities and student government. The assumption is that participation at the upper levels indicates a qualitative difference in engagement in terms of greater commitment to the institution. From research on classroom participation, there also is evidence of differences in typologies of behavior. Some studies separate cooperative participation, or adhering to classroom rules, from autonomy participation, or self-directed academic behaviors (Birch & Ladd, 1997; Buhs &
Ladd, 2001).

2.1.2 Emotional Engagement

Emotional engagement refers to students' affective reactions in the classroom, including interest, boredom, happiness, sadness, and anxiety (Connell & Wellborn,
1991; Skinner & Belmont, 1993). Some researchers assess emotional engagement by measuring emotional reactions to the school and the teacher (Lee & Smith, 1995;
Stipek, 2002). Some conceptualize it as identification with school (Finn, 1989;
Voelkl, 1997). Finn defines identification as belonging (a feeling of being important to the school) and value (an appreciation of success in school-related outcomes).

The emotions included in these definitions duplicate an earlier body of work on attitudes, which examined feelings toward school and included survey questions about liking or disliking school, the teacher, or the work; feeling happy or sad in school; or being bored or interested in the work (Epstein & McPartland, 1976;
Yamamoto et al., 1969). Emotions that were included in this construct, such as interest and value, also overlap considerably with constructs used in motivational research. In fact, the authors of a recent report entitled *Engaging Schools* (National Research Council & Institute

of Medicine, 2004) consider motivation and engagement as synonyms and use the words interchangeably. However, the definitions used in engagement studies are much less elaborated and differentiated than those used in the motivational literature. For example, motivational studies of interest distinguish between situational and personal interest. The former is transitory, aroused by specific features of an activity, such as novelty. The latter is a relatively stable orientation that is more likely to involve consistent choices to pursue an activity or studying a topic and willingness to undertake challenging tasks (Krapp,
Hidi, & Renninger, 1992). The conceptualization of personal interest assumes that interest is directed toward a particular activity or situation. In contrast, the definitions in the engagement literature tend to be general and not differentiated by domain or activity. As a consequence, the source of the emotional reactions is not clear. For instance, it may not be clear whether students' positive emotions are directed toward academic content, their friends, or the teacher.

The theoretical work on values also outlines finer distinctions than are currently present in the engagement literature. Eccles et al. (1983) describe four components of value: *interest* (enjoyment of the activity), *attainment value* (importance of doing well on the task for confirming aspects of one's self-schema), *and utility value/importance* (importance of the task for future goals), and *cost* (negative aspects of engaging in the task). Furthermore, definitions of emotional engagement do not make qualitative distinctions between positive emotions and high involvement or investment. The concept of flow makes this distinction: Flow is a subjective state of complete involvement, whereby individuals are so involved in an activity that they lose awareness of time and space (Csikzentmihalyi, 1988). The definition of flow provides a conceptualization that represents high emotional involvement or investment.

2.1.3 Cognitive Engagement

Research on cognitive engagement comes from the literature on school engagement, which stresses investment in learning, and from the literature on learning and instruction, which involves self-regulation, or being strategic. One set of definitions focuses on psychological investment in learning, a desire to go beyond the requirements, and a preference for challenge 1992; Wehlage et al., 1989). For example, Connell and Wellborn's conceptualization of cognitive engagement includes flexibility in problem solving, preference for hard work, and positive coping in the face of failure. Other researchers have outlined general definitions of engagement that emphasize an inner psychological quality and investment in learning, implying more than just behavioral engagement. For example, Newmann et al. define engagement in academic work as the "student's psychological investment in and effort directed toward learning, understanding, mastering the knowledge, skills or crafts that the academic work is intended to promote" (p. 12). Similarly, Wehlage et al. define engagement as "the psychological investment required to comprehend and master knowledge and skills explicitly taught in schools" (p. 17).

These definitions are quite similar to constructs in the motivation literature, such as motivation to learn (Brophy, 1987), learning goals (Ames, 1992; Dweck & Leggett, 1988) and intrinsic motivation (Harter, 1981). Brophy describes a student who is motivated to learn as valuing learning and striving for knowledge and mastery in learning situations. Similarly, students who adopt learning rather than performance goals are focused on learning, mastering the task, understanding, and trying to accomplish something that is challenging. Intrinsically motivated students prefer challenge and are persistent when faced with difficulty. Each of these concepts emphasizes the degree to which students are invested in and value learning and assumes that the investment is related to, but separate from, strategic learning.

The learning literature defines cognitive engagement in terms of being strategic or self-regulating. Whether described as cognitively engaged or self-regulated, strategic students use

metacognitive strategies to plan, monitor, and evaluate their cognition when accomplishing tasks (Pintrich & De Groot, 1990; Zimmerman, 1990).
They use learning strategies such as rehearsal, summarizing, and elaboration to remember, organize, and understand the material (Corno & Madinach, 1983; Weinstein & Mayer, 1986). They manage and control their effort on tasks, for example, by persisting or by suppressing distractions, to sustain their cognitive engagement (Corno, 1993; Pintrich & De Groot, 1990). A qualitative distinction is made between deep and surface-level strategy use. Students who use deep strategies are more cognitively engaged; they exert more mental effort, create more connection among ideas, and achieve greater understanding of ideas (Weinstein & Mayer). The school engagement literature could benefit from incorporating ideas from the strategy literature to specify what more general terms such as "hard work," "mental effort," and "flexibility" actually entail.
In addition, the use of the term *effort* is problematic in that it is included in definitions of both cognitive and behavioral engagement. A distinction needs to be made between effort that is primarily behavioral, a matter of simply doing the work, and effort that is focused on learning and mastering the material. Research in the motivational literature that addresses the concept of volition can inform these distinctions. It emphasizes cognitive, or psychological, effort, characterizing volition as "psychological control processes that protect concentration and directed effort in the face of personal and/or environmental distractions, and so aid learning and performance" (Corno, 1993, p. 16). Similarly, it is important to distinguish among various types of "going beyond requirements" to further differentiate behavioral and mental effort. In summary, definitions of cognitive engagement draw from two different literatures.
One group specifically highlights a psychological investment in learning; another targets cognition and emphasizes strategic learning. Neither definition alone adequately deals with the qualitative aspects of engagement. Students may be both highly strategic and highly invested in learning; they may be strategic only when it is necessary to get good grades, not because they are motivated to learn; or they may be motivated to learn but lack skills or knowledge about how or when to use strategies.
Overall, the idea of cognitive engagement would be more valuable for understanding school success if scholars integrated the specificity of cognitive processes provided by the self-regulated learning literature with definitions of psychological investment found in the motivational literature.

2.1.4 Summary

We have noted several strengths and limitations of current conceptualizations of behavioral, emotional, and cognitive engagement. First, definitions of engagement incorporate a wide variety of constructs. For example, behavioral engagement encompasses doing the work and following the rules; emotional engagement includes interest, values, and emotions; and cognitive engagement incorporates motivation, effort, and strategy use. This inclusiveness comes at a price. Some of the definitions overlap almost completely with prior literatures, such as those on attitudes toward school or those that use teachers' ratings of behavior to predict achievement. In addition, many of the definitions in the engagement literature are more general than those in other bodies of research from which it draws. The engagement literature is also marked by duplication of concepts and lack of differentiation in definitions across various types of engagement. For example, effort is included as part of definitions of behavioral and cognitive engagement, and no distinction is made between effort aimed merely at fulfilling behavioral expectations and that aimed at understanding the material and mastering the content. Finally, many conceptualizations of engagement include only one or two of the three types.

2.2 Measurement of Engagement

In this section, we present measures of behavioral, emotional, and cognitive engagement; discuss varying approaches to measuring the same types of engagement; and look at the duplication of questionnaire items across the three types. Finally, we discuss limitations of current measurement techniques.

2.2.1 Measuring Behavioral Engagement

There have been several teacher ratings and self-report surveys of behavioral engagement. These include a variety of indicators of conduct, work involvement, and participation, although few studies measure all types of behavior. Aspects of behavior are sometimes separated into different scales (Finn, Folger, & Cox, 1991; Ladd, Birch, & Buhs, 1999). However, the majority of studies combine conduct, persistence, and participation in a single scale. This combination may be problematic because students who are poorly behaved but persist and complete the work are different from those who conform to classroom rules but do not meet academic requirements. Conduct measures include positive behaviors such as completing homework and complying with school rules (Birch & Ladd, 1997; Finn et al., 1995). Other measures incorporate negative behaviors, at both the classroom and school levels, which are indicative of disengagement, such as the frequency of absences and tardiness,

2.2.2 School Engagement

Fighting or getting into trouble, and interfering with others' work (Finn, 1993; Finn et al., 1995; Finn & Rock, 1997). To assess work-related behaviors, some scales include effort, attention, and persistence. For example, teachers are asked to rate the extent to which a particular student "is persistent when confronted with difficult problems" and "approaches new assignments with sincere effort" (Finn et al., 1995). The Rochester School Assessment Package (Wellborn & Connell, 1987) has been used by many researchers to measure behavioral engagement. It contains questionnaire items about effort and attention, such as "I work very hard on my schoolwork" and "When I'm in class I usually think of other things." Finally, some studies have used teachers' reports of helpless behavior as indicators of engagement (Rudolph, Lambert, Clark, & Kurlakowsky, 2001). Other scales focus on students' participatory behaviors. For example, teachers are asked to rate students' level of participation with items such as "Student participates actively in class discussions" and "Student is withdrawn and uncommunicative" (Finn et al., 1995; Wellborn & Connell, 1987). In addition, students are asked to report on their level of initiative with survey items such as "I ask questions to get more information" (Birch & Ladd, 1997; Finn et al., 1995; Wellborn & Connell, 1987). Participation at the school level is assessed with survey questions about involvement in extracurricular activities and governance decisions (Finn, 1993;
Finn & Rock, 1997).

Observation techniques also are used to assess behavioral engagement (Lee & Anderson, 1993; Newmann, 1992; Stipek, 2002). For example, Stipek had observers rate students' engagement by using scales ranging from *off-task* to *deeply involved*, where behaviors included student attentiveness, doing the assigned work, and showing enthusiasm. One potential problem with observational measures is that they provide limited information on the quality of effort, participation, or thinking. Peterson et al. (1984) found that some students judged to be on-task by observers reported in subsequent interviews that they were not thinking about the material. In contrast, many of the students who appeared to be off-task actually were highly cognitively engaged, that is, they were trying to relate new ideas to what they had already learned.

2.2.3 Measuring Emotional Engagement

Most of the studies of emotional engagement use self-report measures, which include survey items about a variety of emotions related to the school, schoolwork, and the people at school. The Rochester School Assessment Package also contains items about positive and negative emotions such as being happy, interested, sad, bored, frustrated, and angry (Connell & Wellborn, 1991; Skinner & Belmont, 1993). Others assess emotional engagement by asking young children to report on their general feelings about their teacher and their school (Stipek, 2002; Valeski & Stipek, 2001). Finn and Voelkl take a different approach, operationalizing emotional engagement as identification with school (Finn, 1989; Voelkl, 1997). In Finn's research, indicators of emotional engagement include student–teacher relations (e.g., "Students get along well with teachers at this school") and values (e.g., "Math will be useful to my future"). Finally, Steinberg, Brown, and Dornbush (1996) measure emotional engagement by assessing students' work orientation (e.g., "I find it hard to stick to anything that takes a long time to do") and their orientation toward school (e.g., "I feel satisfied with school because I am learning a lot").

We noted several issues with how emotional engagement has been measured. First, items that tap behavioral engagement and emotional engagement are often combined in a single scale. This practice makes it more difficult to identify the precursors and consequences of each type of engagement.

Second, the survey items do not specify the source of the emotions. For example, one student may be happy because of the school community, whereas another may be happy because of classroom processes. Third, the measures of emotional engagement tend to be more general than related constructs such as interest and value

(Eccles et al., 1983; Krapp et al., 1992). Finally, the quality and intensity of emotion may vary depending on the type of class activity and setting (Larson & Richards, 1991). Experience-sampling techniques (see Csikzentmihalyi, 1988) are one way to determine the extent to which emotional engagement is a function of stable and enduring qualities or a function of contextual factors.

2.2.4 Measuring Cognitive Engagement

The measures of cognitive engagement, conceptualized as a psychological investment in learning, are limited. In a theoretical piece, Connell and Wellborn (1991) describe measures of cognitive engagement such as survey items about flexible problem solving, preference for hard work, independent work styles, and ways of coping with perceived failure. However, we were unable to find any published studies using these measures. Many of the items parallel those used in the intrinsic motivation literature to tap preference for challenge and independent mastery attempts (e.g., Harter, 1981). This is another example of the overlap of engagement literature with previous research. One area of literature that can inform the measurement of a psychological investment in learning is goal theory. Although a variety of terms have been used, such as learning, mastery, and task-focus, the measurement of goals tends to be very consistent. The measurement scales include items such as being committed to understanding the work, in contrast to wanting to get a good grade or wanting to look smart. The different types of investment lead to different levels of strategy use. For example, students who endorse mastery goals are more likely to use deep-level strategies such as elaboration or organization than are students who endorse performance goals (Ames & Archer, 1988; Pintrich & De Groot, 1990; Wolters, Yu, & Pintrich, 1996). Other studies have assessed a psychological investment in learning by rating the quality of instructional discourse in classrooms. Nystrand and Gamoran (1991) distinguish between *substantive engagement*, a sustained commitment to the content of schooling, which is similar to cognitive engagement, and *procedural engagement*, or trying to complete task requirements,

which lasts only as long as the task itself. In this research, substantive engagement is inferred from the frequency of highlevel evaluation and authentic questions (Gamoran & Nystrand, 1992; Nystrand & Gamoran, 1991). Although the quality of discourse is a measure of engagement at the classroom level, these indicators also could be used to assess an individual's level of engagement. Researchers who write about "cognitive engagement" or "self-regulation," or both, using the terms interchangeably, have developed several measures of student strategy use. One common method for assessing strategy use is self-report questionnaires. These instruments typically measure metacognition, volitional and effort control, and cognitive strategy use. Students are asked about their metacognitive

2.2.5 Summary

In addition to the specific problems that we have noted concerning the measurement of each type of engagement, there are measurement problems that span all three. Some scholars include conceptually distinct and discrete scales for each type of engagement (e.g., Miller et al., 1996; Nystrand & Gamoran, 1991; Patrick, Skinner, & Connell, 1993; Skinner & Belmont, 1993); others combine these into a single, general engagement scale (e.g., Connell, Halpern-Felsher, Clifford, Crichlow, & Usinger, 1995; Marks, 2000; Lee & Smith, 1995). The practice of combining items into general scales precludes examining distinctions among the types of engagement. In addition, conceptual distinctions are blurred because similar items are used to assess different types of engagement. For example, questions about persistence and preference for hard work are included as indicators of both behavioral engagement (Finn et al., 1995) and cognitive engagement (Connell & Wellborn, 1991).

An additional problem is that most measures do not distinguish a target or source of engagement. In some measures the target is quite general, such as "I like school"; in others, the social and academic aspects of school are combined. This melding makes it impossible to determine the actual source of engagement. In addition, most of the self-report measures of behavioral, emotional, and cognitive engagement do not specify subject areas. Incorporating domain-specific measures can help to determine to what extent engagement represents a general tendency and to what extent it is content specific. Recent research has begun to address this problem; observational methods and discourse analysis are being used to examine emotional and cognitive engagement in math (Helme & Clarke, 2001), science (Blumenfeld & Meece, 1988; Lee & Anderson, 1993) and reading (Alvermann, 1999; Guthrie & Wigfield, 2000). Furthermore, measures are rarely attached to specific tasks and situations, instead yielding information about engagement as a general tendency. Thus it is difficult to ascertain to what extent engagement is a function of individual differences or contextual factors. Finally, current measures do not tap qualitative differences in the level of engagement, making it difficult to distinguish the degree of behavioral, emotional, or cognitive investment or commitment. Each type of engagement combines several constructs that are usually measured individually. As a consequence, the measures of the constructs in engagement scales are less well developed than when each construct is examined separately. For example, emotional engagement scales typically include one or two items about interest and values along with items about feelings. Other measures that focus only on interest and value include many items that make distinctions within interest, such as intrinsic versus situational interest, and within value, such as intrinsic, utility, and attainment value (Eccles et al., 1983; Krapp et al., 1992). Obviously, to measure every construct in detail is not practical, because of time and resource constraints. If the goal is to study and understand a particular construct in depth, then the typical measures of engagement that are more inclusive are insufficient. However, if the goal is to predict staying in school or academic success, then any disadvantages of using only a few items to tap each construct may be offset by the increased

predictive strength of a streamlined single measure. The benefits of the tradeoff remain to be determined by researchers who study engagement.

2.3 Outcomes of Engagement

2.3.1 Achievement

Several studies have demonstrated a positive correlation between behavioral engagement and achievement-related outcomes (e.g., standardized tests, grades) for elementary, middle, and high school students (Connell, Spencer, & Aber, 1994; Marks, 2000; Skinner, Wellborn, & Connell, 1990; Connell & Wellborn, 1991). Discipline problems also have been associated with lower school performance across grade levels (Finn et al., 1995; Finn & Rock, 1997). For example, Finn et al. categorized fourth-grade elementary school students as disruptive, inattentive, or withdrawn and contrasted them with students who displayed none of these types of behavior. The authors found that disruptive and inattentive students had lower scores on achievement tests. In addition, Finn and Rock documented large, significant differences on

behavioral engagement measures among high school students classified as resilient (still in school and academically successful), nonresilient completers (still in school and not academically successful), and noncompleters (dropouts). Although much of the research in this field has been cross-sectional, longitudinal studies show that early problems with behavioral engagement have long-lasting effects on achievement.

For example, the Beginning School Study (Alexander, Entwisle, & Dauber, 1993;

Alexander, Entwisle, & Horsey, 1997) showed that teachers' ratings of behavioral engagement in the first grade were related to achievement test score gains, grades over the first 4 years, and decisions to drop out of high school.

In general, there is a consistent association between teacher and student reports of behavioral engagement and achievement across a variety of samples. The strength of this correlation varies across studies. One possible reason is the variety of students studied, ranging from at-risk to gifted students. Another is the use of various achievement measures, including self-reports of grades, teachers' grades, nationally standardized achievement tests, and tests administered by schools, districts, or states. The correlation may be overestimated in the case of grades because teachers take behaviors that indicate effort, such as completing work and paying attention, into account when assigning grades. In addition, the association may be overestimated in the case of tests, which often assess memory and low-level skills, where simply doing the work and paying attention (indicators of behavioral engagement) may be sufficient for success. In contrast, behavioral engagement may not be a very good predictor of performance on assessments that require deep understanding of the material. Much less research exists on emotional engagement and achievement. Some studies show a correlation between achievement and a combined measure of emotional and behavioral engagement (Connell et al., 1994; Skinner et al., 1990). However, these studies do not allow for an examination of the unique contribution of emotional engagement on academic outcomes because they combine different types of engagement.

Voelkl (1997) documented that school identification, measured by value and school belonging, was significantly correlated with achievement test scores in fourth and seventh grades for White students but not for African American students. Studies of the relationship of specific constructs combined under the term emotional engagement, such as interest and value, also show varying associations with achievement (Pintrich & De Groot, 1990; Schiefele, Krapp, & Winteler, 1992). Achievement benefits are found when students are rated as going beyond, doing more work than is required, or initiating discussions with the teacher about school subjects (Fincham, Hokoda, & Sanders, 1989). Research on instructional discourse also demonstrates the achievement benefits of cognitive engagement. Nystrand and

Gamoran (1991) documented that substantive engagement (similar to cognitive engagement) in the classroom was positively related to scores on an achievement test developed to measure students' in-depth understanding and synthesis. Numerous studies from the field of learning also have shown the achievement benefits of strategy use. Children who use metacognitive strategies, such as regulating their attention and effort, relating new information to existing knowledge, and actively monitoring their comprehension, do better on various indicators of academic achievement (Boekarts et al., 2000; Zimmerman, 1990). In conclusion, the research reviewed shows that behavioral engagement (e.g., participation, work behavior, and conduct) is correlated with higher achievement across various samples and ages. Similarly, the link between one aspect of cognitive engagement—strategy use—and achievement in the middle and high school years has been well documented. There also is some evidence of a correlation between emotional engagement and achievement. However, support for this correlation comes mainly from the literature on specific constructs incorporated into definitions of emotional engagement, such as interest and value. Because much of this research is crosssectional, one concern is that the causal direction has not been identified and that any causality may be bidirectional over time. Moreover, measurement problems make it impossible to disentangle the unique contribution of each type of engagement to achievement. Finally, the correlation between engagement and achievement varies depending on how achievement is assessed. Behavioral engagement is likely to be associated with teacher grades and scores on tests that tap basic skills, whereas links with cognitive engagement are more likely to emerge when tests measure synthesis, analysis, and deep-level understanding of content. Although these problems make it difficult to draw firm conclusions, there is evidence from a variety of studies to suggest that engagement positively influences achievement.

2.3.2 Dropping Out

Engagement may help to protect individuals from dropping out of school. Most of the research on this correlation explores the impact of behavioral engagement on the decision to drop out of school. Ekstrom, Goertz, Pollack, and Rock (1986) showed that students who eventually drop out do less homework, exert less effort in school, participate less in school activities, and have more discipline problems at school. Other studies of urban minority samples demonstrate a correlation between low behavioral engagement and cutting class, skipping school, suspension, and retention (Connell et al., 1994; Connell et al., 1995). Involvement in these risky behaviors is a precursor to dropping out. Further evidence comes from the research on extracurricular participation, an aspect of behavioral engagement in school. Involvement in extracurricular activities has been associated with a decreased likelihood of dropping out of school and may be particularly important for certain populations, such as students who are academically at risk and low-income girls (Ekstrom et al., 1986; Mahoney & Cairns, 1997; McNeal, 1995). Other research has shown that behavioral engagement can reduce the likelihood of dropping out and the likelihood of school-age pregnancy among teenage girls (Manlove, 1998; Pillow, 1997). Behavioral engagement in the early years of schooling is a critical mediator in the dropout process (Rumberger, 1987). The Beginning School Study provides the most extensive research documenting the longitudinal effects of early school behaviors on decisions to drop out (Alexander et al., 1997; Ensminger & Slusarcick, 1992;
Entwisle & Alexander, 1993). Teachers' ratings of children's behavioral engagement and academic adjustment in the first grade were related to the decision to drop out of high school (Alexander et al., 1997). Dropouts are more likely than other students to have poor attendance, display disruptive behaviors, and exhibit early school failure (Barrington & Hendricks, 1989; Cairns, Cairns, & Neckerman, 1989).

Students' emotional engagement also has impact on the decision to drop out.

Some scholars have claimed that alienation, or feelings of estrangement and social isolation, contribute to the dropout problem (Finn, 1989; Newmann, 1981). Ethnographic studies support this claim; perceiving an emotional connection to the school or teachers can be a protective factor that keeps at-risk children in school (Fine, 1991; Mehan, Villanueva, Hubbard, Lintz, Okamato, & Adams, 1996; Wehlage et al., 1989). Studies that have examined specific concepts related to engagement point to similar findings. Students who have social difficulties and negative attitudes toward school are more likely to drop out of school (Cairns & Cairns, 1994; Ekstrom et al., 1986; Wehlage & Rutter, 1986).

Several conceptual models have been developed to explain how and why engagement is related to the decision to drop out, but to date there are few empirical studies testing the validity of these models. Finn's (1989) participation–identification model assumes that patterns of engagement and disengagement in the early grades have long-term effects on students' behavior and academic achievement in the later years. According to this model, lack of participation (i.e., lack of behavioral engagement) leads to unsuccessful school outcomes, which in turn lead to emotional withdrawal and lack of identification with the school. Lack of identification is related to nonparticipation in school-related activities, resulting in even less academic success. The process is cyclical: Participation and identification reciprocally influence each other.

Other researchers argue that the dropout process is influenced jointly by engagement and school membership (Newmann et al., 1992; Wehlage et al., 1989). These models assume that the decision to drop out is shaped by individuals' social relationships, commitment to the institution, and belief in the value and legitimacy of school.

In summary, several studies show that behavioral disengagement is a precursor of dropping out. These findings have been based on various measures of behavior (participation, work involvement, and conduct) across ethnically diverse samples in the elementary and high school years. There is less empirical evidence of a correlation between emotional engagement and dropping out. However, the ethnographic research indicates that an emotional connection to teachers and peers can help to reduce dropout rates. We found no studies of cognitive engagement and dropping out. In addition, we know very little about the process by which disengagement influences the decision to drop out. Longitudinal research that explores the mediating processes between behavioral and emotional disengagement and dropping out is critical for intervention efforts. Furthermore, dropout rates vary dramatically by school, even after controlling for demographic characteristics (Rumberger, 1995). An important issue for future study is which aspects of the school and classroom context can promote engagement. Some possible answers to this question can be found in the next section, where we review factors in the school and classroom that are related to engagement.

2.4 Antecedents of Engagement

Family, community, culture, and educational context influence engagement (Connell & Wellborn, 1991; Mehan et al., 1996; Ogbu, 2003). However, a discussion of the first three factors is beyond the scope of this article. Here, we focus on the impact of the educational context on engagement. First, we describe the school-level factors that are associated with engagement. Next, we review the research on classroom context and engagement. Finally, we discuss how individual needs may mediate the relation between the classroom context and engagement. We include findings from studies in major journals cited by engagement researchers as supporting a link between engagement and specific aspects of context when the amount of research on that aspect is relatively small. Our goal is not to provide a comprehensive review of the related literatures but to determine whether these aspects of context merit attention in future research on engagement.

2.4.1 School-Level Factors

In a review article, Newmann (1981) outlined characteristics of high schools that can reduce student alienation and "increase students' involvement, engagement, and integration in school" (p. 546). These include voluntary choice, clear and consistent goals, small size, student participation in school policy and management, opportunities for staff and students to be involved in cooperative endeavors, and academic work that allows for the development of products. There is evidence to support many of these principles. For instance, school size influences behavioral and emotional engagement. In a classic study, Barker and Gump (1964) found that students' opportunities to participate and develop social relations were greater in small schools than in large ones. Researchers who specifically study engagement report similar findings.

Students in small schools participate more in extracurricular and social activities (Finn & Voelkl, 1993). Wehlage and Smith (1992) concluded that small alternative high schools were more likely to have the conditions that promote engagement for at-risk students, including an emphasis on building school membership and a curriculum characterized by authentic work. The school restructuring movement, which supports changing from a bureaucratic to a communal structure, embodies many of the principles outlined by Newmann (1981). Communal structures encourage shared responsibility and commitment to common goals, lateral decision making, and greater individual discretion. Using the National Educational Longitudinal Study, Lee and Smith (1993, 1995) found that students in schools with more elements of communal organization showed higher engagement and greater gains in engagement over time. Other research has examined disciplinary practices, school engagement, and the decision to drop out. Fairness and flexibility in school rules are assumed to reduce the risk of disengagement (Finn & Voekl, 1993; Miller, Leinhart, & Zigmond, 1988; Natriello, 1984). However, the results concerning this assumption are mixed. Natriello (1984) interviewed students about disciplinary and evaluation practices in their schools and found that students who perceived lack of fairness in implementing rules were more likely to be behaviorally disengaged. In contrast, Finn and Voelkl did not find that rigid rules and an emphasis on discipline had a negative impact on behavioral engagement. Other work shows that schools that hold students accountable for behavioral standards have a lower incidence of dropping out (Bryk & Thum, 1989; McDill, Natriello, & Pallas, 1986).

The goal of some current school reforms is to increase engagement. One example is the First Things First model (Institute for Research and Reform in Education, 2003), developed to increase engagement and achievement in under-performing urban and rural areas. This reform model focuses on teachers to decrease the student/ adult ratio and to increase continuity of care; on academics to instantiate high standards and enriching and diverse learning tasks; and on staff to enhance collective responsibility and opportunities for instruction. Initial evaluations demonstrate positive effects on behavioral engagement (e.g., attendance, persistence, and misconduct) and emotional engagement (e.g., school connectedness and support from teachers). Another intervention model is the School Development Program, intended to mobilize the entire school community to support students' holistic development (Comer, 1980). Evaluations of this model in urban schools show increases in positive affect and attitudes toward school, which are aspects of emotional engagement, and decreases in truancy and disciplinary problems, which are aspects of behavioral engagement (Cook, Habib, Phillips, Settersten, Shagle, & Degirmencioglu, 1999).

In summary, this research suggests that school-level factors are associated with behavioral engagement. There is less evidence about the link between school-level factors and emotional and cognitive engagement. Future investigations need to systematically examine the impact of school-level factors, such as those noted by Newmann (1981), on the three types of engagement across diverse populations and ages.

Longitudinal tracking of changes in engagement as a result of attempts to alter the school context also are needed. There are several widely implemented school reforms that focus on increasing achievement and not explicitly on engagement (see Borman, Hewes, Overmann, & Brown, 2003, for a review of school reforms). Although evaluations of these reforms do not specifically measure it, engagement may be the mediator that links reforms to outcomes. Including engagement measures in these intervention studies can provide insight into the degree to which engagement is responsive to variations in the environment and can point to the specific school and classroom changes that have the largest effects on behavioral, emotional, and cognitive engagement.

2.4.2 Classroom Context

In this section, we discuss classroom context and engagement. We focus on factors that have been studied in the engagement literature, including teacher support, peers, classroom structure, autonomy support, and task characteristics.

Teacher support has been shown to influence behavioral, emotional, and cognitive engagement. Teacher support can be either academic or interpersonal, although the majority of studies do not make this distinction and many studies combine items about the two into one scale (Wenztel, 1997). Teachers' reports of the quality of the teacher–child relationship in the early school years have been associated with teachers' ratings of behavioral engagement, such as cooperative participation and self- directedness (Birch & Ladd, 1997; Valeski & Stipek, 2001). Children's initial behavioral engagement also influences their relationship with the teacher (Ladd et al., 1999). In fact, an extensive literature suggests that teachers prefer students who are academically competent, responsible, and conform to school rules over students who are disruptive and aggressive (see Kedar-Voivodas, 1983). This preference is likely to lead teachers to provide different opportunities to behaviorally engaged and disengaged students. However, the majority of the research on teacher support and engagement has been cross-sectional, making it difficult to test these reciprocal links. One exception is the research by Skinner and Belmont (1993). They documented that teacher involvement was positively associated with engagement, and that, in turn, higher student engagement elicited greater teacher involvement.

Other work has examined the effect of perceived teacher support in the elementary, middle, and high school years. Teacher support and caring has been correlated with various aspects of behavioral engagement, including higher participation in learning and on-task behavior (Battistich, Solomon, Watson, & Schaps, 1997), lower disruptive behavior (Ryan & Patrick, 2001), and a lower probability of dropping out of school (Croninger & Lee, 2001) among samples of ethnically diverse elementary, middle, and high school students. Furthermore, Marks (2000) demonstrated that a classroom environment in which students received support from both teachers and peers was associated with higher engagement among elementary, middle, and high school students in schools undergoing reforms. Additional evidence of the importance of teacher support comes from the ethnographic research; students are more likely to drop out of school when they feel they do not have a positive or supportive relationship with their teachers (Farrell, 1990; Fine, 1991; Wehlage et al., 1989).

Teacher support has been correlated with emotional engagement in a primarily White middle-class sample (Connell & Wellborn, 1991; Skinner & Belmont, 1993). This research replicates an earlier literature on classroom climate that related perceived teacher support and student attitudes (Fraser & Fisher, 1982; Moos, 1979). It is also similar to research on the middle school transition, which shows a decline in the quality of teacher–student relations and may explain the decrease in adolescents' interest during this period of their lives (Feldlaufer, Midgley, & Eccles, 1988; Midgley, Feldlaufer, & Eccles, 1989).

Another body of literature has investigated teacher support and cognitive engagement.

A sample of middle school students reported higher cognitive engagement and greater use of learning and metacognitive strategies in classrooms where teachers presented challenging work and pressed for understanding (Blumenfeld & Meece, 1988; Blumenfeld, Puro, & Mergendoller, 1992). Observational studies illustrate the benefits of a socially supportive and intellectually challenging environment. In classrooms where teachers created respectful and socially supportive environments, pressed students for understanding, and supported autonomy, students were more strategic about learning and had higher behavioral engagement and affect (Stipek, 2002; Turner, Meyer, Cox, Logan, DiCintio, & Thomas, 1998). If teachers focus only on academics but create a negative social environment, students are likely to experience emotional disengagement and be more apprehensive about making mistakes. In contrast, if teachers focus only on the social dimension but fail to attend to the intellectual dimensions, students are less likely to be cognitively engaged in learning.

In summary, numerous studies have illustrated a link between teacher support and behavioral engagement. These studies are based on a variety of measures of behavior (e.g., participation, work involvement, and conduct) across diverse samples in the elementary, middle, and high school years. Most of the evidence concerning the association between teacher support and emotional engagement comes from related literatures.
Findings concerning the impact of teacher support on cognitive engagement are beginning to accumulate and point to the importance of a combination of academic and social support. Determining whether the effects of social or academic support on engagement vary with student age and background requires further study.
Finally, because the majority of research has been cross-sectional rather than longitudinal, we know very little about the long-term consequences of teacher support on behavioral, emotional, and cognitive engagement.

2.5 Peers

Researchers have focused less on the peer group than on teachers as a factor in the socialization of engagement (Ryan, 2000). Children in elementary and middle school cluster together in peer groups with similar levels of engagement, and this clustering strengthens existing differences (Kindermann, 1993; Kindermann, McCollam, & Gibson, 1996). For example, Kindermann (1993) used social composite mapping to document that elementary school children who were affiliated with high engagement peer groups increased their level of behavioral engagement across the school year. The bodies of literature on peer acceptance and rejection have been used as theoretical justification for studying peers and engagement. Peer acceptance in both childhood and adolescence is associated with satisfaction in school, which is an aspect of emotional engagement, and socially appropriate behavior and academic effort, which are aspects of behavioral engagement (Berndt & Keefe, 1995; Ladd, 1990; Wentzel, 1994). In contrast, children who are rejected during the elementary school years are at greater risk for poor conduct and lower classroom participation, both elements of behavioral engagement, and lower interest in school, an aspect of emotional engagement (Buhs & Ladd, 2001; DeRosier, Kupersmidt, & Patterson, 1994). Peer support and engagement are likely to be reciprocal. Children who do not conform to school rules or who dislike school are less likely to perceive peers as supportive (Ladd et al., 1999; Ladd & Coleman, 1997). Peer rejection in both childhood and adolescence increases the probability of dropping out of school (French & Conrad, 2001; Parker & Asher, 1987). Other work has focused on the negative effect of the peer group on adolescents' commitment to doing well in school, especially among minority youth. Ogbu's cultural ecological model attempts to explain the academic failure of involuntary minority groups (Ogbu, 1987, 2003). Ogbu claims that students in these groups disengage from school because they perceive limited

opportunities to attain school success and they fear peer rejection for "acting White" in trying to get good grades.

Several scholars have criticized Ogbu's theory for its failure to explain why some minority students do try to succeed whereas others disengage from school (Conchas, 2001; Mehan et al., 1996; O'Connor, 1997).1 Recent qualitative descriptions of resistance and resilience examine minority youth's perceptions of discrimination, social support, and school engagement. Students who perceive that race and class constrain their educational opportunities, but who also have social supports that promote the development of agency and strategies for confronting discrimination, are more likely to remain engaged in school (Conchas, 2001; Deyhle, 1995; Mehan et al., 1996; O'Connor, 1997; Stanton-Salazar, 2001).

Newer work on cognitive engagement and learning communities illustrates how peers can be more than friends or associates. Cognitive engagement is enhanced when class members actively discuss ideas, debate points of view, and critique each other's work (Guthrie & Wigfield, 2000; Meloth & Deering, 1994; Newmann, 1992). For example, Guthrie and colleagues created a year-long intervention program that emphasized peer interactions and the use of interesting materials as crucial aspects of enhancing engagement in reading (Guthrie, McGough, Bennett, & Rice, 1996). In conclusion, the primary evidence for the effect of peers on engagement comes from studies of naturally occurring peer groups (Kindermann, 1993; Kindermann et al., 1996). Other work has shown that the peer group can contribute to school disengagement among minority youth. Related studies that use constructs and measures similar to those used in the engagement literature also illustrate the link between peers and engagement. For example, peer acceptance and peer rejection are predictors of outcomes that are aspects of behavioral engagement (e.g., participation, conduct, work involvement) and emotional engagement (e.g., interest, satisfaction in school). Future investigations should examine the impact of peers on cognitive engagement. They should also consider whether there are developmental and group differences in how peers affect engagement. For example, whether the relationship is stronger for older children, as they develop gender, racial, and cultural identities, remains to be explored.

Chapter 3

DATASET COLLECTION AND ANNOTATION FOR AN AUTOMATIC ENGAGEMENT CLASSIFIER

The data for this study were collected from 19 online c++ teaching video lectures. We collected these videos randomly from different sources like YouTube, edx.org and Coursera. The purpose of this experiment was to measure the importance of teaching by seeing the teacher's face. We use the computer vision for detecting faces from the video and collecting the features of the face by applying the Gabor filter using the skimage module provided by scipy in python.
The features extracted from the face images are mean and variance of the image data. By using gabor filter we can apply different kernels which are based on the different frequencies and orientation angles and sigma. The extracted data was stored in the txt file using the numpy module functions which can be accessed further for using it in machine learning algorithm.

3.1 Data annotation

Given the recorded videos, the next step was to label them for engagement. We organized a team of labelers consisting of undergraduate and graduate students from computer science, cognitive science, and psychology from the two universities where data were collected. These labelers viewed and rated the videos for the appearance of engagement. Note that not all labelers labeled the exact same sets of images/videos. Instead, we chose to balance the goals of obtaining many labels per image/video, and annotating a large amount of data for developing an automated detector. When labeling videos, the audio was turned off, and labelers were instructed to label engagement based only on appearance. In contrast to the more thoroughly studied domains of automatic basic emotion recognition (happy, sad, angry, disgusted, fearful, surprised, or neutral) or facial action unit classification (from the Facial Action Coding System), affective states that are relevant to learning such as frustration or engagement may be difficult to define clearly. Hence, arriving at a sufficiently clear definition and devising an appropriate labeling procedure, including the timescale at which labeling takes place, is important for ensuring both the reliability and validity of the training labels. In pilot experimentation we tried three different approaches to labeling:
1) Watching video clips (at normal viewing speed) and giving continuous engagement labels by pressing the the Up/Down arrow keys.
2) Watching video clips and giving a single number to rate the entire video.
3) Viewing static images and giving a single number to rate each image.
We found approach (1) very difficult to execute in practice.
One problem was the tendency to habituate to each subject's recent level of engagement, and to adjust the current rating relative to that subject's average engagement level of the recent past. This could yield labels that are not directly comparable between subjects or even within subjects. Another problem was how to rate short events, e.g., brief eye closure or looks to the side: should these brief moments be labeled as "non-engagement",or should they be overlooked as normal behavior if thesubject otherwise appears highly engaged? Finally, it was difficult to provide continuous labels that were synchronized in time with the video; proper synchronization would require first scanning the video for interesting events, and then re-watching it and carefully adjusting the engagement up or down at each moment in time. We found the labeling task was easier using approaches (2) and (3), provided that clear instructions were given as to what constitutes "engagement".

3.2 Engagement categories and instructions

Given the approach of giving a single engagement number to an entire video clip or image, we decided on the following approximate scale to rate engagement:

1) Interesting: Teacher seems interesting by its facial expressions and way of representing lectures.
2) Boring: Teacher seems to be boring by their facial expression or not engaged in the task.

3.3 Timescale

An important variable in annotating video is thetimescale at which labeling takes place. For approach (2) , we experimented with two different time scales: clips of 60 sec and clips of 10 sec. Approach (3) (single images) can be seen as the lower limit of the length of a video clip. In a pilot experiment we compared these three timescales for intercoder reliability.

For the 60 sec labeling task, all the video sessions (_ 45 minutes/subject) from the HBCU subjects were watched from start to end in 60 sec clips, and 2 labelers entered a single engagement score after viewing each clip. For the 10 sec labeling task, 505 video clips of 10 sec each were extracted at random timepoints from the session videos and shown to 7 labelers in random order (in termsof both time and subject). Between the 60 sec clips and the 10 sec labeling tasks, we found the 10 sec labelling task more intuitive. When viewing the longer clips, it was difficult to know what label to give if the subject appeared non-engaged early on but appeared highly engaged at the end. The inter-coder reliability of the 60 sec clip labeling task was _ = 0:39 (across 2 labelers); for the 10 sec clip labeling task _ = 0:68 (across 7 labelers).

For approach (3), we created custom labeling software in which 7 labelers annotated batches of 100 images each. The images for each batch were video frames extracted at random timepoints from the session videos. Each batch contained a random set of images spanning multiple timepoints from multiple subjects. Labelers rated each image individually but could view many images and their assigned labels simultaneously on the screen. The labeling software also provided a Sort button to sort the images in ascending order by their engagement label. In practice, we found this to be an intuitive and efficient method of labeling images for the appearance of engagement. The inter-coder reliability for image-based labelling was _ = 0:56. This reliability can also be increased by averaging frame-based labels across multiple frames that are consecutive in time (see Section 3.4).

3.4 Static versus motion information

One interesting question is how much information about students' engagement is captured in the static pixels of the individual video frames compared to the dynamics of the motion. We conducted a pilot study to examine this question. In particular, we randomly selected 120 video clips (10 sec each) from the set of all HBCU videos. The random sample contained clips from 24 subjects. Each clip was then split into 40 frames spaced 0:25 sec apart. These frames were then shuffled both in time and across subjects. A human labeler labeled these image frames for the appearance of engagement, as described in "approach (3)". Finally, the engagement values assigned to all the frames for a particular clip were reassembled and averaged; this average served as an estimate of the "true" engagement score given by that same labeler when viewing that video clip as described in "approach (2)" above. We found that, with respect to the true engagement scores, the estimated scores gave a _ = 0:78 and a Pearson correlation r = 0:85. This accuracy is quite high and suggests that most of the information about the appearance of engagement is contained in the static pixels, not the motion per se.

We also examined the video clips in which the reconstructed engagement scores differed the most from the true scores. In particular, we ranked the 120 labeled video clips in decreasing order of absolute deviation of the estimated label (by averaging the frame-based labels) from the "true" label given to the video clip viewed as a whole. We then examined these clips and attempted to explain the discrepancy: In the first clip (greatest absolute deviation), the subject was swaying her headfrom side to side as if listening to music (although she was not). It is likely that the coder treated this as non-engaged behavior. This behavior may be difficult to capture from static frame judgments. However, it was also an anomalous case.

In the second clip, the subject turned his head to the side to look at the experimenter, who was talking to him for several seconds. In the frame-level judgments, this was perceived as off-task, and hence non-engaged behavior; this corresponds to the instructions given to the coders that they rate engagement under the assumption that the subject should always be looking towards the iPad. For the video clip label, however, the coder judged the student to be highly engaged because he was intently listening to the experimenter. This is an example of inconsistency on the part of the coder as to what constitutes engagement and does not necessarily indicate a problem with splitting the clips into frames. Finally, in several clips the subjects sometimes shifted their eye gaze downward to look at the bottom of the iPad screen. At a frame level, it was difficult to distinguish the subject looking at the bottom of the iPad from the subject looking to his/her own lap or even closing his/her eyes, both of which would be considered non-engagement. From video, it was easier to distinguish these behaviors from the context. However, these downward gaze events were rare and can be effectively filtered out by simple averaging. In spite of these problems, the relatively high accuracy of estimating video-based labels from frame-based labels suggests an approach for how to construct an automatic classifier of engagement: Instead of analyzing video clips as video, break them up into their video frames, and then combine engagement estimates for each frame. We used this approach to label both the HBCU and the UC data for engagement. In the next section, we describe our proposed architecture for automatic engagement recognition based on this frame-by-frame design.

Chapter 4

AUTOMATIC RECOGNITION ARCHITECTURES

Based on the finding from Section 3.4 that video clipbased labels can be estimated with high fidelity simply by averaging frame-based labels, we focus our study on **frame-by-frame** recognition of student engagement. This means that many techniques developed for emotion and facial action unit classification can be applied to the engagement recognition problem. In this project we proposed a 3-stage pipeline.
1) Face registration: the face and facial landmark (eyes, nose, and mouth) positions are localized automatically in the image; the face box coordinates are computed; and the face patch is cropped from the image [35]. We experimented with 36 _ 36 and 48 _ 48 pixel face resolution.
2) The cropped face patch is classified by four binary classifiers, one for each engagement category l 2. The outputs of the binary classifiers are fed to a regressor to estimate the image's engagement level.
Stage (1) is standard for automatic face analysis, and our particular approach is described in [35]. Stage (2) is discussed in the next subsection, and stage (3) is discussed in Section 3.11. This architecture is reminiscent of an automated head pose estimation system we developed previously [57], which combines the outputs of multiple binary classifiers to form a real valued judgment.

4.1 Binary classification

We trained 4 binary classifiers of engagement – one for each of the 4 levels described in Section 3.1. The task of each of these classifiers is to discriminate an image (or video frame) that belongs to engagement level l from an image that belongs to some other engagement level l0 6= l. We call these detectors 1-v-other, 2-v-other, etc. We compared three commonly used and demonstrably effective feature type + classifier combinations from the automatic facial expression recognition literature:
GentleBoost with Box Filter features (**Boost(BF)**): this is the approach popularized Viola and Jones in for face detection.
_ Support vector machines with Gabor features (**SVM(Gabor)**): this approach has achieved some of the highest accuracies in the literature for facial action and basic emotion classification._ Multinomial logistic regression with expression outputs from the Computer Expression Recognition Toolbox (**MLR(CERT)**): here, we attempt to harness an existing automated system for facial expression analysis to train engagement classifiers. Our goal is not to judge the effectiveness of each feature type (or each learning method) in isolation, but rather to assess the effectiveness of these state-of-the-art computer vision architectures for a novel vision task. As relatively little research has yet examined how to recognize the emotional states specific to students in real learning environments, it is an open question how well these methods would perform for engagement recognition.We describe each approach in more detail below.

4.1.1 Boost(BF)

Box Filter (BF) features measure differences in average pixel intensity between neighboring rectangular regions of an image. They have been shown to be highly effective for automatic face detection [53] as well as smile detection. For example, for detecting faces, a 2-rectangle Box Filter can capture the fact that the eye region of the face is typically darker than the upper cheeks. At runtime, BF features are fast to extract using the "integra image" technique.

At training time, however, the number of BF features relative to the image resolution is very high compared to other image representations (e.g., a Gabor decomposition), which can lead to overfitting. BF features are typically combined with a boosted classifier such as Adaboost or GentleBoost (Boost), which performs both feature selection during training and actual classification at run-time. In our GentleBoost implementation, each weak learner consists of a nonparametric regressor smoothed with a Gaussian kernel of bandwidth _, to estimate the log-likelihood ratio of the class label given the feature value. Each Gentle- Boost classifier was trained for 100 boosting rounds. For the features, we included 6 types of Box Filters in total, comprising two-, three-, and four-rectangle features similar to those used in [53], and an additional tworectangle "center-surround" feature (see Figure 3). At a face image resolution of 48_48 pixels, there were 5397601 BF features; at a face resolution of 36 _ 36 pixels, there were 1683109 features.

4.1.2 SVM(Gabor)

Gabor Energy Filters [44] are bandpass filters with a tunable spatial orientation and frequency. They model the complex cells of the primate's visual cortex. When applied to images, they respond to edges at particular orientations, e.g., horizontal edges due to wrinkling of the forehead, or diagonal edges due to "crow's feet" around the eyes. Gabor Energy Filters have a proven record in a wide variety of face processing applications, including face recognition and facial expression recognition. In machine learning applications Gabor features are often classified by a soft-margin linear support vector machine (SVM) with parameter C specifying how much misclassified training examples should penalize the objective function. In our implementation, we applied a "bank" of 40 Gabor Energy Filters consisting of 8 orientations (spaced at 22.5deg intervals) and 5 spatial frequencies ranging from 2 to 32 cycles per face. The total number of Gabor features is N _ N _ 8 _ 5, where N is the face image width in pixels.

4.1.3 MLR(CERT)

The Facial Action Coding System is a comprehensive framework for objectively describing facial expression in terms of Action Units, which measure the intensity of over 40 distinct facial muscles. Manual FACS coding has previously been used to study student engagement and other emotions relevant to automated teaching. In our study, since we are interested in automatic engagement recognition, we employ the Computer Expression Recognition Toolbox (CERT), which is a software tool developed by our laboratory to estimate facial action intensities automatically. Although the accuracies of the individual facial action classifiers vary, we have found CERT to be useful for a variety of facial analysis tasks, including the discrimination of real from faked pain, driver fatigue detection, and estimation of students' perception of curriculum difficulty. CERT outputs intensity estimates of 20 facial actions as well as the 3-D pose of the head (yaw, pitch, and roll). For engagement recognition we classify the CERT outputs using multinomial logistic regression (MLR), trained with an L2 regularizer on the weight vector of strength. We use the absolute value of the yaw, pitch, and roll to provide invariance to the direction of the pose change. Since we are interested in real-time systems that can operate without baselining the detector to a particular subject, we use the raw CERT outputs (i.e., we do not z-score the outputs) in our experiments. Internally, CERT uses the SVM(Gabor) approach described above. Since CERT was trained on hundreds to thousands of subjects (depending on the particular output channel), which is substantially higher than the number of subjects collected for this study, it is possible that CERT's outputs will provide an identity independent representation of the students' faces, which may boost generalization performance.

4.2 Data selection

1) From the online video lectures we have selected randomly 10 seconds clips.
2) From one video 3 clips of 10 sec each are selected. So, there are 57 total 10sec clips used in this project.
3) From each clip we extracted 10 frames i.e. still images at the rate of one image per second
4) So, at the end we have a total of 570 images to work on.
5) Audio is extracted from 10 sec clip and at each 100ms we took an audio frame.
6) Then we combined the audio and video features per second and per video lecture.

Chapter 5

IMAGE FEATURES

In the realms of image processing and computer vision, Gabor filters are generally used in texture analysis, edge detection, feature extraction, disparity estimation (in stereo vision), etc. Gabor filters are special classes of bandpass filters, i.e., they allow a certain ‘band’ of frequencies and reject the others.

In the course of this tutorial, we shall first discuss the essential results that we obtain when Gabor filters are applied on images. Then we move on to discuss the different parameters that control the output of the filter. This tutorial is aimed at delivering a practical overview of Gabor filters; hence, theoretical treatment is omitted (a tutorial that provides the essential theoretical rigor is currently in the pipeline).

At each stage of the discussion, results of relevant filters have been displayed. The implementation, though contained in the tutorial itself, draws heavily from the Python script that comes along with OpenCV. It has been simplified further so that it is simple for the beginners to work with.

To start with, Gabor filters are applied to images pretty much the same way as are conventional filters. We have a mask (a more precise (cooler) term for it would be ‘convolution kernel’) that represents the filter. By a mask, we mean to say that we have an array (usually a 2D array since 2D images are involved) of pixels in which each pixel is assigned a value (call it a ‘weight’). This array is slid over every pixel of the image and a convolution operation is performed (you can refer to the following link for more information on how a mask is applied to an image.
When a Gabor filter is applied to an image, it gives the highest response at edges and at points where texture changes. The following images show a test image and its transformation after the filter is applied.

A Gabor filter responds to edges and texture changes. When we say that a filter responds to a particular feature, we mean that the filter has a distinguishing value at the spatial location of

that feature (when we're dealing with applying convolution kernels in spatial domain, that is. The same holds for other domains, such as frequency domains, as well).

There are certain parameters that affect the output of a Gabor filter. In OpenCV Python, following is the structure of the function that is used to create a Gabor kernel.

Each parameter is described very briefly in the OpenCV docs. Here's a brief introduction to each of these parameters.

Ksize is the size of the Gabor kernel. If ksize = (a, b), we then have a Gabor kernel of size a x b pixels. As with many other convolution kernels, ksize is preferably odd and the kernel is a square (just for the sake of uniformity).

Sigma is the standard deviation of the Gaussian function used in the Gabor filter.

Theta is the orientation of the normal to the parallel stripes of the Gabor function.

Lambda is the wavelength of the sinusoidal factor in the above equation.

Gamma is the spatial aspect ratio.

Psi is the phase offset.

ktype indicates the type and range of values that each pixel in the Gabor kernel can hold.
Now that we've got a quaint feel of what each parameter means, let us delve deeper and understand the practical implication of the variation of each of these parameters.

5.1 ksize
On varying ksize, the size of the convolution kernel varies. In the code above we modify the parameter ksize, while keeping the kernel square and of an odd size. We observe that there is no effect of the size of the convolution kernel on the output image. This also implies that the convolution kernel is scale invariant, since scaling the kernel's size is analogous to scaling the size of the image. Here are a few results with varying ksize. For all the following images, sigma = 4.0, theta = 0, lambd = 10.0, gamma = 0.5, psi = 0, and ktype = cv2.CV_32F (i.e., each pixel of the convolution kernel holds a weight which is a 32-bit floating point number).

Input Image ksize = 31 x 31 ksize = 51 x 5 ksize = 151 x 151

5.2 Sigma

This parameter controls the width of the Gaussian envelope used in the Gabor kernel. Here are a few results obtained by varying this parameter.

sigma = 2 sigma = 3 sigma = 4 sigma = 5

5.3 Theta

This is perhaps one of the most important parameters of the Gabor filter. This parameter decides what kind of features the filter responds to. For example, giving theta a value of zero means that the filter is responsive only to horizontal features only. So, in order to obtain features at various angles in an image, we divide the interval between 0 and 180 into 16 equal parts, and compute a Gabor kernel for each value of theta thus obtained. Note that we've chosen 16 just because it was the default value in the OpenCV implementation. These parameter values could be modified to suit specific purposes. Following are the results of varying theta on the above input image.

theta = 11.25 theta = 22.5 theta = 33.75 theta = 45

theta = 56.25 theta = 67.5 theta = 78.75 theta = 90

theta = 101.25 theta = 112.5 theta = 123.75 theta = 135

theta = 146.25 theta = 157.5 theta = 168.75 theta = 180

5.4 Lambda

Here's the variation with lambda (theta is set to zero).

lambda = 8 lambda = 9 lambda = 10 lambda = 11

5.5 Gamma

Gamma controls the ellipticity of the gaussian. When gamma = 1, the gaussian envelope is circular.

gamma = 0.3 gamma = 0.4 gamma = 0.5 gamma = 0.6

5.6 Psi

This parameter controls the phase offset.

psi = 0 psi = 10 psi = 50 psi = 90

So, we've examined the observable effects of various parameters on the output of the Gabor filter.

Chapter 6

AUDIO FEATURES

6.1 Mel Frequency Cepstral Coefficient (MFCC) Features

The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc.

The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. This page will provide a short tutorial on MFCCs.

Mel Frequency Cepstral Coefficents (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) and were the main feature type for automatic speech recognition (ASR). This page will go over the main aspects of MFCCs, why they make a good feature for ASR, and how to implement them.

6.1.1 Steps at a Glance

We will give a high level intro to the implementation steps, then go in depth why we do the things we do. Towards the end we will go into a more detailed description of how to calculate MFCCs.

1. Frame the signal into short frames.
2. For each frame calculate the periodogram estimate of the power spectrum.
3. Apply the mel filterbank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filterbank energies.
5. Take the DCT of the log filterbank energies.
6. Keep DCT coefficients 2-13, discard the rest.

There are a few more things commonly done, sometimes the frame energy is appended to each feature vector. Delta and Delta-Delta features are usually also appended. Liftering is also commonly applied to the final features.

6.1.2 Why do we do these things?

We will now go a little more slowly through the steps and explain why each of the steps is necessary.

An audio signal is constantly changing, so to simplify things we assume that on short time scales the audio signal doesn't change much (when we say it doesn't change, we mean statistically i.e. statistically stationary, obviously the samples are constantly changing on even short time scales). This is why we frame the signal into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame.

The next step is to calculate the power spectrum of each frame. This is motivated by the human cochlea (an organ in the ear) which vibrates at different spots depending on the frequency of the incoming sounds. Depending on the location in the cochlea that vibrates (which wobbles small hairs), different nerves fire informing the brain that certain frequencies are present. Our periodogram estimate performs a similar job for us, identifying which frequencies are present in the frame.

The periodogram spectral estimate still contains a lot of information not required for Automatic Speech Recognition (ASR). In particular the cochlea cannot discern the difference between two closely spaced frequencies. This effect becomes more pronounced as the frequencies increase. For this reason we take clumps of periodogram bins and sum them up to get an idea of how much energy exists in various frequency regions. This is performed by our Mel filterbank: the first filter is very narrow and gives an indication of how much energy exists near 0 Hertz. As the frequencies get higher our filters get wider as we become less concerned about variations. We are only interested in roughly how much energy occurs at each spot. The Mel scale tells us exactly how to space our filterbanks and how wide to make them. See below for how to calculate the spacing.
Once we have the filterbank energies, we take the logarithm of them. This is also motivated by human hearing: we don't hear loudness on a linear scale. Generally to double the percieved volume of a sound we need to put 8 times as much energy into it. This means that large variations in energy may not sound all that different if the sound is loud to begin with. This compression operation makes our features match more closely what humans actually hear. Why the logarithm and not a cube root? The logarithm allows us to use cepstral mean subtraction, which is a channel normalisation technique.

The final step is to compute the DCT of the log filterbank energies. There are 2 main reasons this is performed. Because our filterbanks are all overlapping, the filterbank energies are quite correlated with each other. The DCT decorrelates the energies which means diagonal covariance matrices can be used to model the features in e.g. a HMM classifier. But notice that only 12 of the 26 DCT coefficients are kept. This is because the higher DCT coefficients represent fast changes in the filterbank energies and it turns out that these fast changes actually degrade ASR performance, so we get a small improvement by dropping them.

6.2 What is the Mel scale?

The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear.

The formula for converting from frequency to Mel scale is:

$$M(f) = 1125 \ln(1 + f/700) \qquad (1)$$

To go from Mels back to frequency:

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \qquad (2)$$

6.2.2 Implementation steps

We start with a speech signal, we'll assume sampled at 16 kHz.

1. Frame the signal into 20-40 ms frames. 25ms is standard. This means the frame length for a 16 kHz signal is 0.025*16000 = 400 samples. Frame step is usually something like 10ms (160 samples), which allows some overlap to the frames. The first 400 sample frame starts at sample 0, the next 400 sample frame starts at sample 160 etc. until the end of the speech file is reached. If the speech file does not divide into an even number of frames, pad it with zeros so that it does.

The next steps are applied to every single frame, one set of 12 MFCC coefficients is extracted for each frame. A short aside on notation: we call our time domain signal $s(n)$. Once it is framed we have $s_i(n)$where n ranges over 1-400 (if our frames are 400 samples) and i ranges over the number of frames. When we calculate the complex DFT, we get $S_i(k)$- where the i denotes the frame number corresponding to the time-domain frame. $P_i(k)$ is then the power spectrum of frame i.

2. To take the Discrete Fourier Transform of the frame, perform the following:

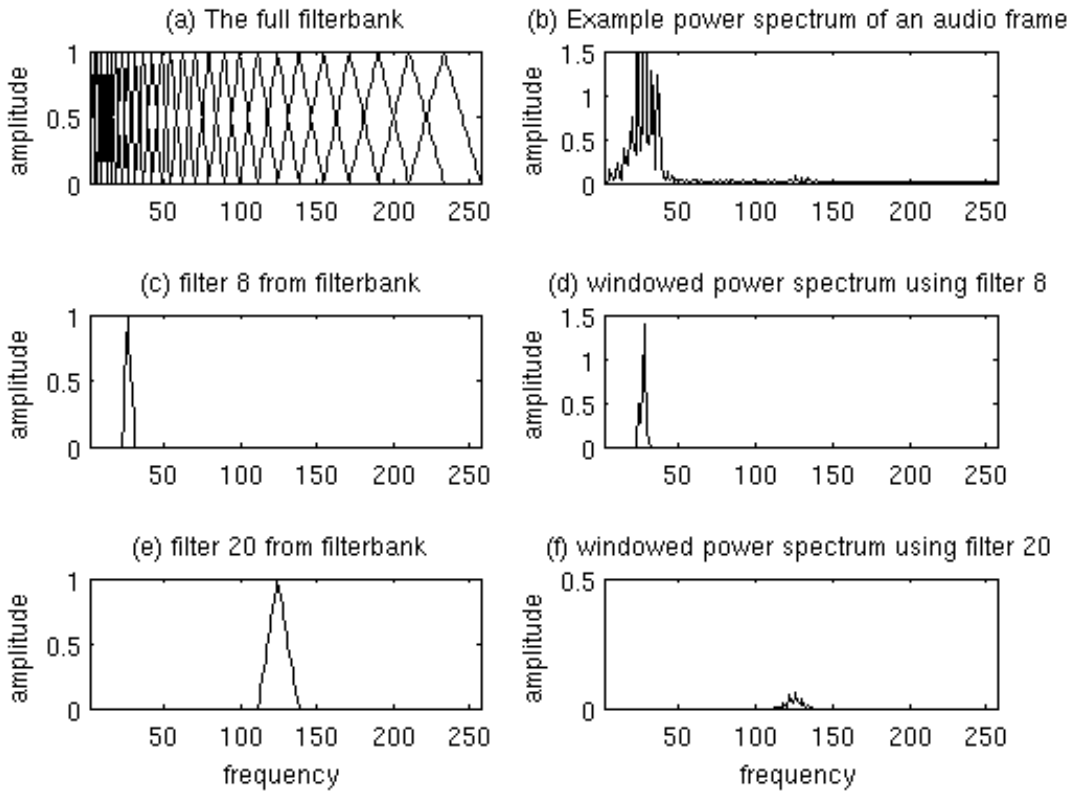
$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \qquad 1 \leq k \leq K$$

Where $h(n)$ is an N sample long analysis window (e.g. hamming window), and K is the length of the DFT. The periodogram-based power spectral estimate for the speech frame $s_i(n)$ is given by:

$$P_i(k) = \frac{1}{N}|S_i(k)|^2$$

This is called the Periodogram estimate of the power spectrum. We take the absolute value of the complex fourier transform, and square the result. We would generally perform a 512 point FFT and keep only the first 257 coefficents.

3. Compute the Mel-spaced filterbank. This is a set of 20-40 (26 is standard) triangular filters that we apply to the periodogram power spectral estimate from step 2. Our filterbank comes in the form of 26 vectors of length 257 (assuming the FFT settings fom step 2). Each vector is mostly zeros, but is non-zero for a certain section of the spectrum. To calculate filterbank energies we multiply each filterbank with the power spectrum, then add up the coefficents. Once this is performed we are left with 26 numbers that give us an indication of how much energy was in each filterbank. For a detailed explanation of how to calculate the filterbanks. Here is a plot to hopefully clear things up:



Plot of Mel Filterbank and windowed power spectrum

4. Take the log of each of the 26 energies from step 3. This leaves us with 26 log filterbank energies.

5. Take the Discrete Cosine Transform (DCT) of the 26 log filterbank energies to give 26 cepstral coefficents. For ASR, only the lower 12-13 of the 26 coefficients are kept.

The resulting features (12 numbers for each frame) are called Mel Frequency Cepstral Coefficients.

6.2.3 Computing the Mel filterbank

In this section the example will use 10 filterbanks because it is easier to display, in reality you would use 26-40 filterbanks.

To get the filterbanks shown in figure 1(a) we first have to choose a lower and upper frequency. Good values are 300Hz for the lower and 8000Hz for the upper frequency. Of course if the speech is sampled at 8000Hz our upper frequency is limited to 4000Hz. Then follow these steps:

1. Using equation 1, convert the upper and lower frequencies to Mels. In our case 300Hz is 401.25 Mels and 8000Hz is 2834.99 Mels.
2. For this example we will do 10 filterbanks, for which we need 12 points. This means we need 10 additional points spaced linearly between 401.25 and 2834.99. This comes out to:

3. m(i) = 401.25, 622.50, 843.75, 1065.00, 1286.25, 1507.50, 1728.74,

   1949.99, 2171.24, 2392.49, 2613.74, 2834.99

4. Now use equation 2 to convert these back to Hertz:
5. h(i) = 300, 517.33, 781.90, 1103.97, 1496.04, 1973.32, 2554.33,

   3261.62, 4122.63, 5170.76, 6446.70, 8000

   Notice that our start- and end-points are at the frequencies we wanted.
6. We don't have the frequency resolution required to put filters at the exact points calculated above, so we need to round those frequencies to the nearest FFT bin. This process does not affect the accuracy of the features. To convert the frequncies to fft bin numbers we need to know the FFT size and the sample rate,

   f(i) = floor((nfft+1)*h(i)/samplerate)

   This results in the following sequence:

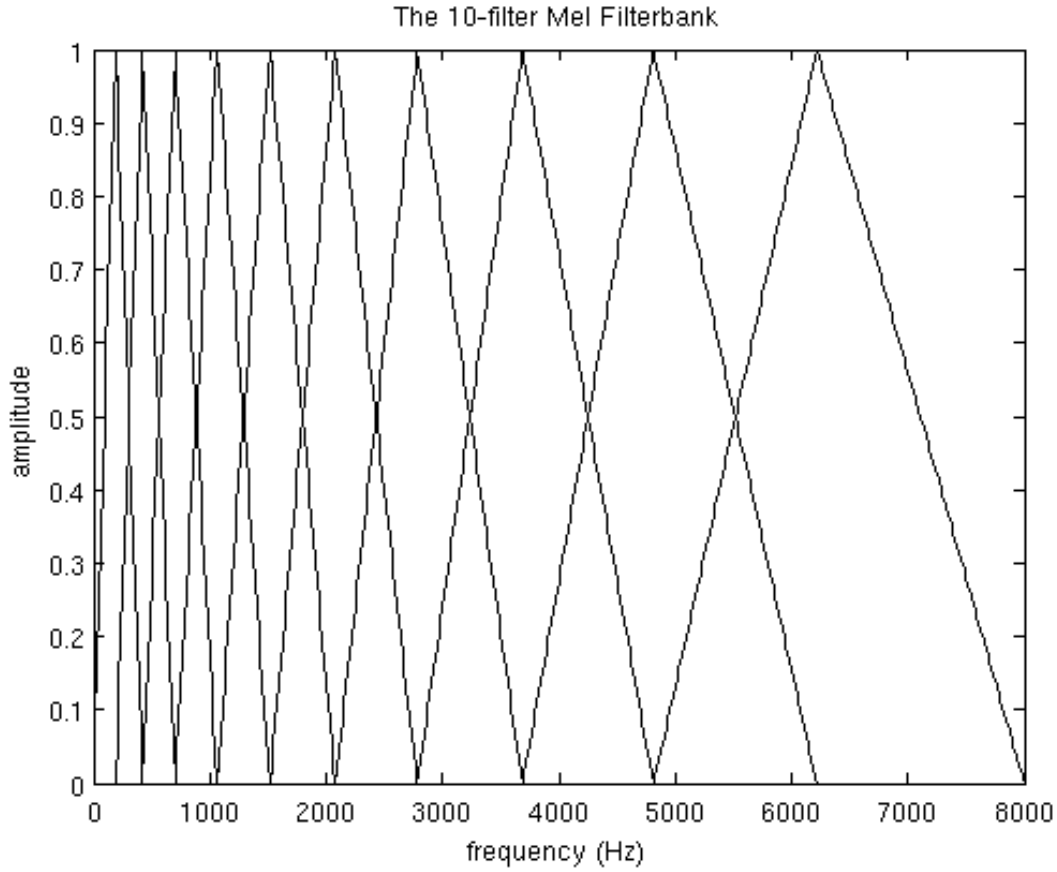
   f(i) = 9, 16, 25, 35, 47, 63, 81, 104, 132, 165, 206, 256

   We can see that the final filterbank finishes at bin 256, which corresponds to 8kHz with a 512 point FFT size.
7. Now we create our filterbanks. The first filterbank will start at the first point, reach its peak at the second point, then return to zero at the 3rd point. The second filterbank will start at the 2nd point, reach its max at the 3rd, then be zero at the 4th etc. A formula for calculating these is as follows:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1) \\ 0 & k > f(m+1) \end{cases}$$

   where M is the number of filters we want, and $f()$ is the list of M+2 Mel-spaced frequencies.

The final plot of all 10 filters overlayed on each other is:



A Mel-filterbank containing 10 filters. This filterbank starts at 0Hz and ends at 8000Hz. This is a guide only, the worked example above starts at 300Hz.

6.3 Deltas and Delta-Deltas

Also known as differential and acceleration coefficients. The MFCC feature vector describes only the power spectral envelope of a single frame, but it seems like speech would also have information in the dynamics i.e. what are the trajectories of the MFCC coefficients over time. It turns out that calculating the MFCC trajectories and appending them to the original feature vector increases ASR performance by quite a bit (if we have 12 MFCC coefficients, we would also get 12 delta coefficients, which would combine to give a feature vector of length 24).

To calculate the delta coefficients, the following formula is used:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

Where d_t is a delta coefficient, from frame t computed in terms of the static coefficients c_{t+N} to c_{t-N}. A typical value for N is 2. Delta-Delta (Acceleration) coefficients are calculated in the same way, but they are calculated from the deltas, not the static coefficients.

6.4 MFCC Features

The default parameters should work fairly well for most cases, if you want to change the MFCC parameters, the following parameters are supported:

```
Python

def mfcc(signal,samplerate=16000,winlen=0.025,winstep=0.01,numcep=13,

                nfilt=26,nfft=512,lowfreq=0,highfreq=None,preemph=0.97,

    ceplifter=22,appendEnergy=True)
```

Parameter	Description
signal	the audio signal from which to compute features. Should be an N*1 array
samplerate	the samplerate of the signal we are working with.
winlen	the length of the analysis window in seconds. Default is 0.025s (25 milliseconds)
winstep	the step between successive windows in seconds. Default is 0.01s (10 milliseconds)
numcep	the number of cepstrum to return, default 13
nfilt	the number of filters in the filterbank, default 26.
nfft	the FFT size. Default is 512
lowfreq	lowest band edge of mel filters. In Hz, default is 0
highfreq	highest band edge of mel filters. In Hz, default is samplerate/2
preemph	apply preemphasis filter with preemph as coefficient. 0 is no filter. Default is 0.97
ceplifter	apply a lifter to final cepstral coefficients. 0 is no lifter. Default is 22
appendEnergy	if this is true, the zeroth cepstral coefficient is replaced with the log of the total frame energy.

Parameter	Description
returns	A numpy array of size (NUMFRAMES by numcep) containing features. Each row holds 1 feature vector.

6.5 Filterbank Features

These filters are raw filterbank energies. For most applications you will want the logarithm of these features. The default parameters should work fairly well for most cases. If you want to change the fbank parameters, the following parameters are supported:

```
Python

def fbank(signal,samplerate=16000,winlen=0.025,winstep=0.01,

    nfilt=26,nfft=512,lowfreq=0,highfreq=None,preemph=0.97)
```

Parameter	Description
signal	the audio signal from which to compute features. Should be an N*1 array
samplerate	the samplerate of the signal we are working with
winlen	the length of the analysis window in seconds. Default is 0.025s (25 milliseconds)
winstep	the step between seccessive windows in seconds. Default is 0.01s (10 milliseconds)
nfilt	the number of filters in the filterbank, default 26.
nfft	the FFT size. Default is 512.
lowfreq	lowest band edge of mel filters. In Hz, default is 0
highfreq	highest band edge of mel filters. In Hz, default is samplerate/2
preemph	apply preemphasis filter with preemph as coefficient. 0 is no filter. Default is 0.97

Parameter	Description
returns	A numpy array of size (NUMFRAMES by nfilt) containing features. Each row holds 1 feature vector. The second return value is the energy in each frame (total energy, unwindowed)

Chapter 7

REGULARIZATION AND MODEL SELECTION

Suppose we are trying to select among several different models for a learning problem. For instance, we might be using a polynomial regression model h_(x) = g(θ0 + θ1x + θ2x2 + · · · + θkxk), and wish to decide if k should be 0, 1, . . . , or 10. How can we automatically select a model that represents a good tradeoff between the twin evils of bias and variance1? Alternatively, suppose we want to automatically choose the bandwidth parameter τ for locally weighted regression, or the parameter C for our ℓ1-regularized SVM.

How can we do that?
For the sake of concreteness, in these notes we assume we have some finite set of models M = {M1, . . . ,Md} that we're trying to select among.
For instance, in our first example above, the model Mi would be an i-th order polynomial regression model. (The generalization to infinite M is not hard. Alternatively, if we are trying to decide between using an SVM, a neural network or logistic regression, then M may contain these models.

7.1 Cross validation

Let's suppose we are, as usual, given a training set S. Given what we know about empirical risk minimization, here's what might initially seem like a algorithm, resulting from using empirical risk minimization for model selection:
1. Train each model Mi on S, to get some hypothesis hi.
2. Pick the hypotheses with the smallest training error.

This algorithm does not work. Consider choosing the order of a polynomial. The higher the order of the polynomial, the better it will fit the training set S, and thus the lower the training error. Hence, this method will always select a high-variance, high-degree polynomial model, which we saw previously is often poor choice. Here's an algorithm that works better. In hold-out cross validation (also called simple cross validation), we do the following:

1. Randomly split S into Strain (say, 70% of the data) and Scv (the remaining 30%). Here, Scv is called the hold-out cross validation set.

2. Train each model Mi on Strain only, to get some hypothesis hi.

3. Select and output the hypothesis hi that had the smallest error ˆεScv(hi) on the hold out cross validation set. (Recall, ˆεScv(h) denotes the empirical error of h on the set of examples in Scv.)

By testing on a set of examples Scv that the models were not trained on, we obtain a better estimate of each hypothesis hi's true generalization error, and can then pick the one with the smallest estimated generalization error. Usually, somewhere between 1/4 − 1/3 of the data is used in the hold out cross validation set, and 30% is a typical choice. Optionally, step 3 in the algorithm may also be replaced with selecting the model Mi according to argmini ˆεScv(hi), and then retraining Mi on the entire training set S. (This is often a good idea, with one exception being learning algorithms that are be very sensitive to perturbations of the initial

conditions and/or data. For these methods, Mi doing well on Strain does not necessarily mean it will also do well on Scv, and it might be better to forgot.
The disadvantage of using hold out cross validation is that it "wastes" about 30% of the data. Even if we were to take the optional step of retraining the model on the entire training set, it's still as if we're trying to find a good model for a learning problem in which we had 0.7m training examples, rather than m training examples, since we're testing models that were trained on only 0.7m examples each time. While this is fine if data is abundant and/or cheap, in earning problems in which data is scarce (consider a problem with m = 20, say), we'd like to do something better. Here is a method, called k-fold cross validation, that holds out less data each time:
1. Randomly split S into k disjoint subsets of m/k training examples each.
Let's call these subsets S1, . . . , Sk.

2. For each model Mi, we evaluate it as follows:
For j = 1, . . . , k
Train the model Mi on S1 ∪ · · · ∪ Sj−1 ∪ Sj+1 ∪ · · · Sk (i.e., train on all the data except Sj) to get some hypothesis hij .
Test the hypothesis hij on Sj , to get ˆεSj (hij).
The estimated generalization error of model Mi is then calculated as the average of the ˆεSj (hij)'s (averaged over j).

3. Pick the model Mi with the lowest estimated generalization error, and retrain that model on the entire training set S. The resulting hypothesis is then output as our final answer.

A typical choice for the number of folds to use here would be k = 10.
While the fraction of data held out each time is now 1/k—much smaller than before—this procedure may also be more computationally expensive than hold-out cross validation, since we now need train to each model k times.
While k = 10 is a commonly used choice, in problems in which data is really scarce, sometimes we will use the extreme choice of k = m in order to leave out as little data as possible each time. In this setting, we would repeatedly train on all but one of the training examples in S, and test on that held-out example. The resulting m = k errors are then averaged together to obtain our estimate of the generalization error of a model. This method has its own name; since we're holding out one training example at a time, this method is called leave-one-out cross validation. Finally, even though we have described the different versions of cross validation as methods for selecting a model, they can also be used more simply to evaluate a single model or algorithm. For example, if you have implemented some learning algorithm and want to estimate how well it performs for your application (or if you have invented a novel learning algorithm and want to report in a technical paper how well it performs on various test sets), cross validation would give a reasonable way of doing so.

7.2 Feature Selection
One special and important case of model selection is called feature selection. To motivate this, imagine that you have a supervised learning problem where the number of features n is very large (perhaps n ≫ m), but you suspect that there is only a small number of features that are "relevant" to the learning task. Even if you use a simple linear classifier (such as the perceptron) over the n input features, the VC dimension of your hypothesis class would still be O(n), and thus overfitting would be a potential problem unless the training set is fairly large. In such a setting, you can apply a feature selection algorithm to reduce the number of features. Given n features, there are 2n possible feature subsets (since each of the n features

can either be included or excluded from the subset), and thus feature selection can be posed as a model selection problem over 2n possible models. For large values of n, it's usually too expensive to explicitly enumerate over and compare all 2n models, and so typically some heuristic search procedure is used to find a good feature subset. The following search procedure is called forward search:
1. Initialize F = ∅.

2. Repeat {
(a) For i = 1, . . . , n if i 6∈ F, let Fi = F ∪ {i}, and use some version of cross validation to evaluate features Fi. (I.e., train your learning algorithm using only the features in Fi, and estimate its generalization error.)
(b) Set F to be the best feature subset found on step (a).
}

3. Select and output the best feature subset that was evaluated during the entire search procedure. The outer loop of the algorithm can be terminated either when F = {1, . . . , n} is the set of all features, or when |F| exceeds some pre-set thresh- old (corresponding to the maximum number of features that you want the algorithm to consider using). This algorithm described above one instantiation of wrapper model feature selection, since it is a procedure that "wraps" around your learning algorithm, and repeatedly makes calls to the learning algorithm to evaluate how well it does using different feature subsets. Aside from forward search, other search procedures can also be used. For example, backward search starts off with F = {1, . . . , n} as the set of all features, and repeatedly deletes features one at a time (evaluating single-feature deletions in a similar manner to how forward search evaluates single-feature additions) until F = ∅. Wrapper feature selection algorithms often work quite well, but can be computationally expensive given how that they need to make many calls to the learning algorithm. Indeed, complete forward search (terminating when F = {1, . . . , n}) would take about O(n2) calls to the learning algorithm. Filter feature selection methods give heuristic, but computationally much cheaper, ways of choosing a feature subset. The idea here is to compute some simple score S(i) that measures how informative each feature xi is about the class labels y. Then, we simply pick the k features with the largest scores S(i). One possible choice of the score would be define S(i) to be (the absolute value of) the correlation between xi and y, as measured on the training data. This would result in our choosing the features that are the most strongly correlated with the class labels. In practice, it is more common (particularly
for discrete-valued features xi) to choose S(i) to be the mutual information MI(xi, y) between xi and y:
MI(xi, y) = X xi2{0,1} X y2{0,1} p(xi, y) log p(xi, y) p(xi)p(y). (The equation above assumes that xi and y are binary-valued; more generally the summations would be over the domains of the variables.) The probabilities above p(xi, y), p(xi) and p(y) can all be estimated according to their empirical distributions on the training set. To gain intuition about what this score does, note that the mutual information can also be expressed as a Kullback-Leibler (KL) divergence: MI(xi, y) = KL (p(xi, y)||p(xi)p(y))
You'll get to play more with KL-divergence in Problem set #3, but infor- mally, this gives a measure of how different the probability distributions p(xi, y) and p(xi)p(y) are. If xi and y are independent random variables, then we would have p(xi, y) = p(xi)p(y), and the KL-divergence between the two distributions will be zero. This is consistent with the idea if xi and y are independent, then xi is clearly very "non-informative" about y, and thus the score S(i) should be small. Conversely, if xi is very "informative" about y, then their mutual information MI(xi, y) would be large. One final detail: Now that you've ranked the features

according to their scores S(i), how do you decide how many features k to choose? Well, one standard way to do so is to use cross validation to select among the possible values of k. For example, when applying naive Bayes to text classification— a problem where n, the vocabulary size, is usually very large—using this method to select a feature subset often results in increased classifier accuracy.

7.3 Bayesian statistics and regularization

In this section, we will talk about one more tool in our arsenal for our battle against overfitting.

At the beginning of the quarter, we talked about parameter fitting using maximum likelihood (ML), and chose our parameters according to θML = argmax_m Yi =1 p(y(i)|x(i); θ).

Throughout our subsequent discussions, we viewed θ as an unknown parameter of the world. This view of the θ as being constant-valued but unknown is taken in frequentist statistics. In the frequentist this view of the world, θ is not random—it just happens to be unknown—and it's our job to come up with statistical procedures (such as maximum likelihood) to try to estimate this parameter.

An alternative way to approach our parameter estimation problems is to take the Bayesian view of the world, and think of θ as being a random variable whose value is unknown. In this approach, we would specify a prior distribution p(θ) on θ that expresses our "prior beliefs" about the parameters. Given a training set S = {(x(i), y(i))}m i=1, when we are asked to make a prediction on a new value of x, we can then compute the posterior distribution on the parameters.

The procedure that we've outlined here can be thought of as doing "fully
Bayesian" prediction, where our prediction is computed by taking an average with respect to the posterior p(θ|S) over θ. Unfortunately, in general it is computationally very difficult to compute this posterior distribution. This is because it requires taking integrals over the (usually high-dimensional) θ as in Equation (1), and this typically cannot be done in closed-form.

Thus, in practice we will instead approximate the posterior distribution for θ. One common approximation is to replace our posterior distribution for θ (as in Equation 2) with a single point estimate. The MAP (maximum a posteriori) estimate for θ is given by θMAP = argmax_ m Yi =1 p(y(i)|x(i), θ)p(θ).

Note that this is the same formulas as for the ML (maximum likelihood) estimate for θ, except for the prior p(θ) term at the end.

In practical applications, a common choice for the prior p(θ) is to assumethat θ ∼ N(0, τ 2I). Using this choice of prior, the fitted parameters θMAP will have smaller norm than that selected by maximum likelihood.In practice, this causes the Bayesian MAP estimate to be less susceptible to overfitting than the ML estimate of the parameters. For example, Bayesian logistic regression turns out to be an effective algorithm for text classification, even though in text classification we usually have n ≫ m.

Chapter 8

SUPPORT VECTOR MACHINES

8.1 Introduction

Machine Learning is considered as a subfield of Artificial Intelligence and it is concerned with the development of techniques and methods which enable the computer to learn. In simple terms development of algorithms which enable the machine to learn and perform tasks and activities. Machine learning overlaps with statistics in many ways. Over the period of time many techniques and methodologies were developed for machine learning tasks.

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. Support vector machine was initially popular with the NIPS community and now is an active part of the machine learning research around the world. SVM becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task. It is also being used for many applications, such as hand writing analysis, face analysis and so forth, especially for pattern classification and regression based applications. The foundations of Support Vector Machines (SVM) have been developed by Vapnik and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior, to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SRM minimizes an upper bound on the expected risk, whereas ERM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems.

8.2 Statistical Learning Theory

The statistical learning theory provides a framework for studying the problem of gaining knowledge, making predictions, making decisions from a set of data. In simple terms, it enables the choosing of the hyper plane space such a way that it closely represents the underlying function in the target space.

In statistical learning theory the problem of supervised learning is formulated as follows. We are given a set of training data $\{(\mathbf{x}_1,y_1)... (\mathbf{x}_l,y_l)\}$ in $R^n \times R$ sampled according to unknown probability distribution $P(\mathbf{x},y)$, and a loss function $V(y,f(\mathbf{x}))$ that measures the error, for a given $\mathbf{x}$, $f(\mathbf{x})$ is "predicted" instead of the actual value y. The problem consists in finding a function f that minimizes the expectation of the error on new data that is, finding a function f that minimizes the expected error: $\int V(y, f(\mathbf{x}))\, P(\mathbf{x}, y)\, d\mathbf{x}\, dy$

In statistical modeling we would choose a model from the hypothesis space, which is closest (with respect to some error measure) to the underlying function in the target space. More on statistical learning theory can be found on introduction to statistical learning theory.

8.3 Learning and Generalization

Early machine learning algorithms aimed to learn representations of simple functions. Hence, the goal of learning was to output a hypothesis that performed the correct classification of the training data and early learning algorithms were designed to find such an accurate fit to the data [8]. The ability of a hypothesis to correctly classify data not in the training set is known as its generalization. SVM performs better in term of not over generalization when the neural networks might end up over generalizing easily. Another thing to observe is to find where to make the best trade-off in trading complexity with the number of epochs; the illustration brings to light more information about this. The below illustration is made from the class notes.

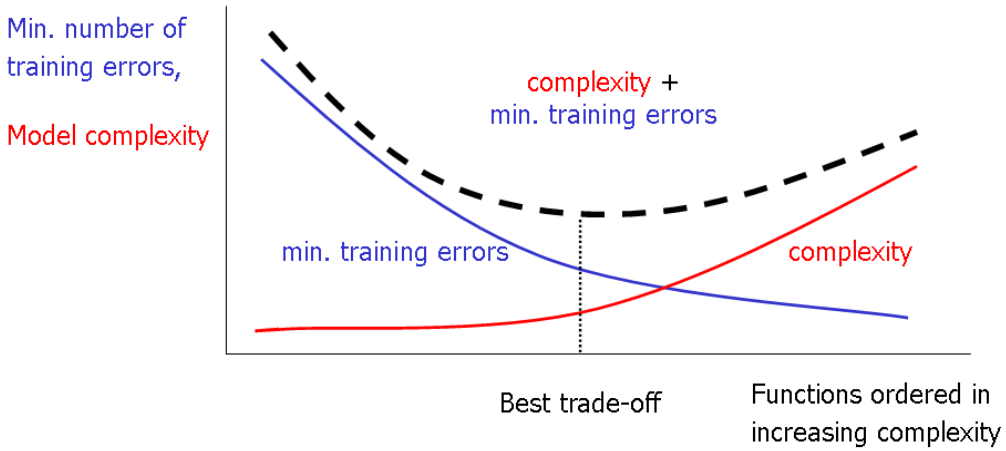


Figure 1: Number of Epochs Vs Complexity.

8.4 Introduction to SVM: Why SVM?

Firstly working with neural networks for supervised and unsupervised learning showed good results while used for such learning applications. MLP's uses feed forward and recurrent networks. Multilayer perceptron (MLP) properties include universal approximation of continuous nonlinear functions and include learning with input-output patterns and also involve advanced network architectures with multiple inputs and outputs.

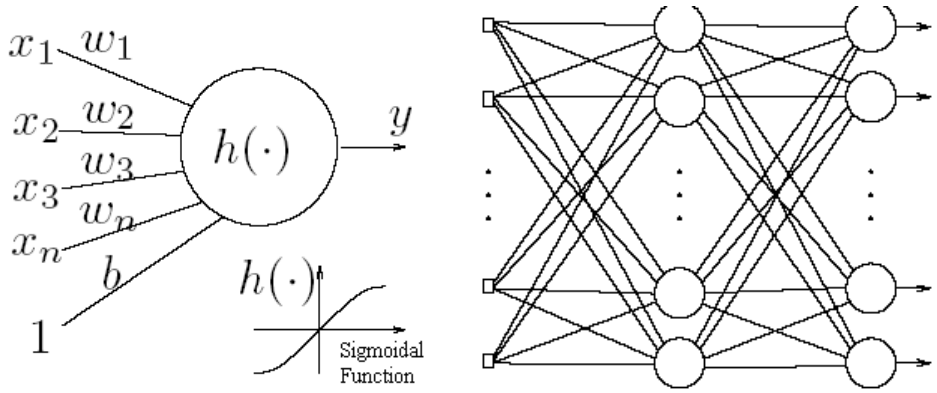


Figure 2: a] Simple Neural Network b]Multilayer Perceptron. These are simple visualizations just to have a overview as how neural network looks like.

There can be some issues noticed. Some of them are having many local minima and also finding how many neurons might be needed for a task is another issue which determines whether optimality of that NN is reached. Another thing to note is that even if the neural network solutions used tends to converge, this may not result in a unique solution. Now let us look at another example where we plot the data and try to classify it and we see that there are many hyper planes which can classify it. But which one is better?

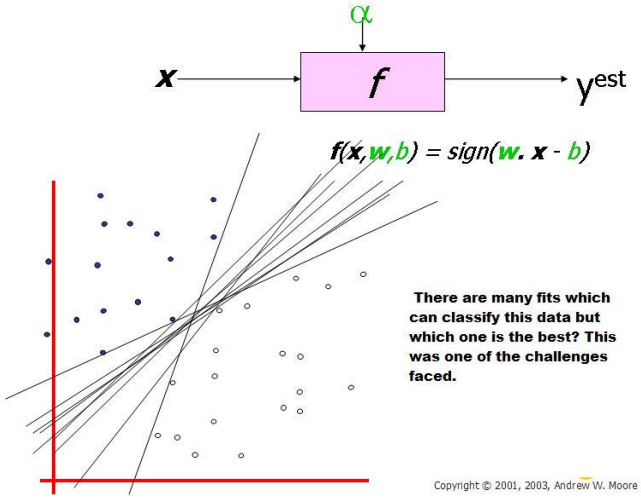


Figure 3: Here we see that there are many hyper planes which can be fit in to classify the data but which one is the best is the right or correct solution. The need for SVM arises. (Taken Andrew W. Moore 2003). Note the legend is not described as they are sample plotting to make understand the concepts involved.

From above illustration, there are many linear classifiers (hyper planes) that separate the data. However only one of these achieves maximum separation. The reason we need it is because if we use a hyper plane to classify, it might end up closer to one set of datasets compared to others and we do not want this to happen and thus we see that the concept of maximum margin classifier or hyper plane as an apparent solution. The next illustration gives the maximum margin classifier example which provides a solution to the above mentioned problem.

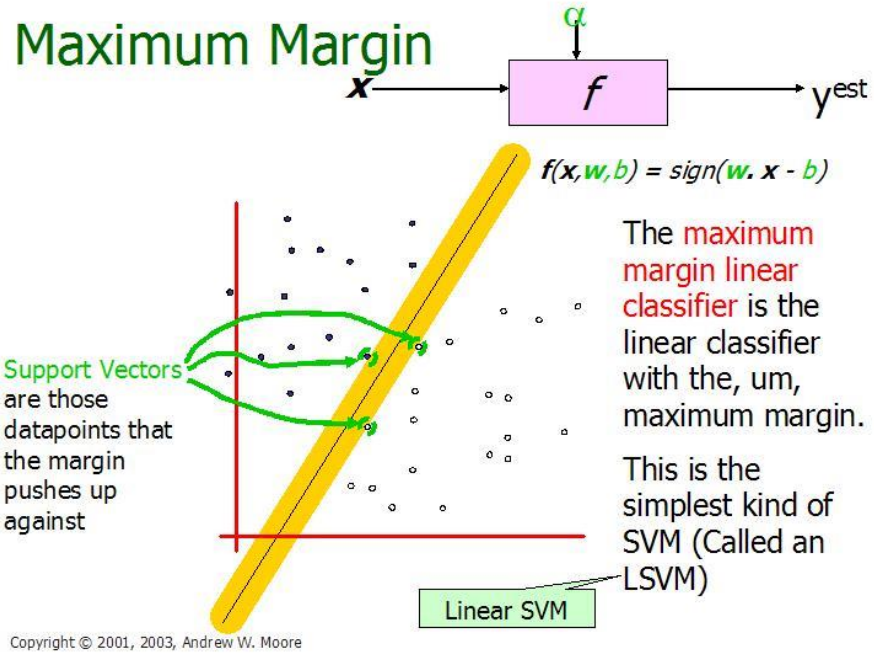


Figure 4: Illustration of Linear SVM. (Taken from Andrew W. Moore slides 2003). Note the legend is not described as they are sample plotting to make understand the concepts involved.

Expression for Maximum margin is given as (for more information visit):

$$\text{margin} \equiv \arg\min_{\mathbf{x}\in D} d(\mathbf{x}) = \arg\min_{\mathbf{x}\in D} \frac{|\mathbf{x}\cdot\mathbf{w}+b|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$

The above illustration is the maximum linear classifier with the maximum range. In this context it is an example of a simple linear SVM classifier. Another interesting question is why maximum margin? There are some good explanations which include better empirical performance. Another reason is that even if we've made a small error in the location of the boundary this gives us least chance of causing a misclassification. The other advantage would be avoiding local minima and better classification. Now we try to express the SVM mathematically and for this tutorial we try to present a linear SVM. The goals of SVM are separating the data with hyper plane and extend this to non-linear boundaries using kernel trick. For calculating the SVM we see that the goal is to correctly classify all the data. For mathematical calculations we have,

[a] If $Y_i = +1$; $wx_i + b \geq 1$

[b] If $Y_i = -1$; $wx_i + b \leq 1$

[c] For all i; $y_i (w_i + b) \geq 1$

In this equation x is a vector point and w is weight and is also a vector. So to separate the data [a] should always be greater than zero. Among all possible hyper planes, SVM selects the one where the distance of hyper plane is as large as possible. If the training data is good and every test vector is located in radius r from training vector. Now if the chosen hyper plane is located at the farthest possible from the data. This desired hyper plane which maximizes the margin also bisects the lines between closest points on convex hull of the two datasets. Thus we have [a], [b] & [c].

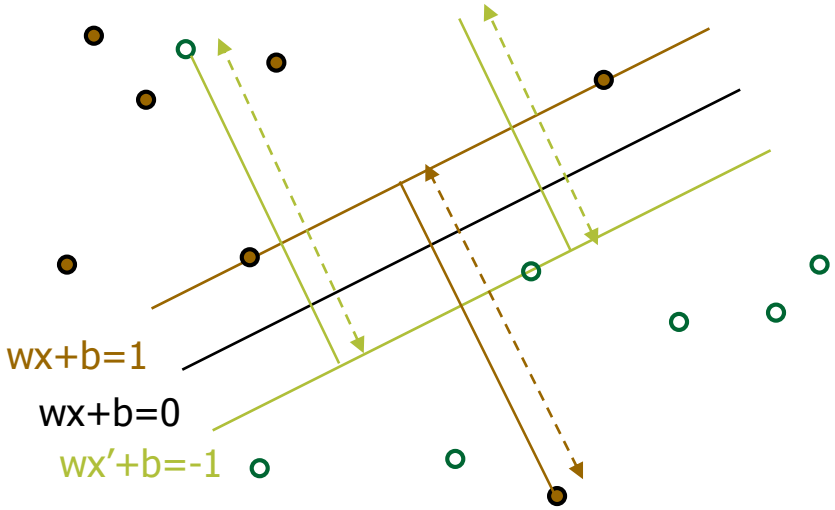


Figure 5: Representation of Hyper planes.

Distance of closest point on hyperplane to origin can be found by maximizing the x as x is on the hyper plane. Similarly for the other side points we have a similar scenario. Thus solving and subtracting the two distances we get the summed distance from the separating hyperplane to nearest points. Maximum Margin = M = 2 / ||w||

Now maximizing the margin is same as minimum. Now we have a quadratic optimization problem and we need to solve for w and b. To solve this we need to optimize the quadratic function with linear constraints. The solution involves constructing a dual problem and where a Langlier's multiplier α_i is associated. We need to find w and b such that Φ (w) =½ |w'||w| is minimized; And for all $\{(x_i, y_i)\}$: $y_i (w * x_i + b) \geq 1$.

Now solving: we get that $w = \Sigma \alpha_i * x_i$; $b = y_k - w * x_k$ for any x_k such that $\alpha k \neq 0$

Now the classifying function will have the following form: $f(x) = \Sigma\alpha_i\ y_i\ x_i\ *\ x + b$

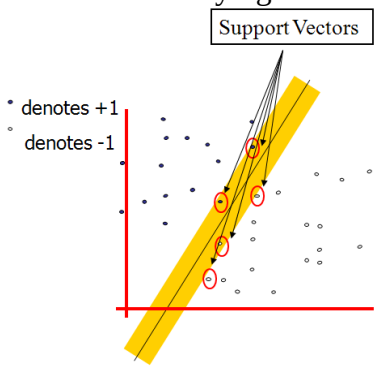


Figure 6: Representation of Support Vectors (Copyright © 2003, Andrew W. Moore)

8.5 SVM Representation

In this we present the QP formulation for SVM classification. This is a simple representation only.

SV classification:

$$\min_{f,\xi_i} \|f\|_K^2 + C\sum_{i=1}^{l}\xi_i \qquad y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \text{ for all i} \quad \xi_i \geq 0$$

SVM classification, Dual formulation:

$$\min_{\alpha_i} \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad 0 \leq \alpha_i \leq C, \text{ for all i;} \qquad \sum_{i=1}^{l}\alpha_i y_i = 0$$

Variables ξ_i are called slack variables and they measure the error made at point $(\mathbf{x}_i, y_i)$. Training SVM becomes quite challenging when the number of training points is large. A number of methods for fast SVM training have been proposed.

8.6 Soft Margin Classifier

In real world problem it is not likely to get an exactly separate line dividing the data within the space. And we might have a curved decision boundary. We might have a hyperplane which might exactly separate the data but this may not be desirable if the data has noise in it. It is better for the smooth boundary to ignore few data points than be curved or go in loops, around the outliers. This is handled in a different way; here we hear the term slack variables being introduced. Now we have, $y_i(w'x + b) \geq 1 - S_k$. This allows a point to be a small distance S_k on the wrong side of the hyper plane without violating the constraint. Now we might end up having huge slack variables which allow any line to separate the data, thus in such scenarios we have the Lagrangian variable introduced which penalizes the large slacks.

$$\min L = \tfrac{1}{2}\, w'w - \sum \lambda_k (y_k (w'x_k + b) + s_k - 1) + \alpha \sum s_k$$

Where reducing α allows more data to lie on the wrong side of hyper plane and would be treated as outliers which give smoother decision boundary.

8.7 Kernal Trick

Let's first look at few definitions as what is a kernel and what does feature space mean?

Kernel: If data is linear, a separating hyper plane may be used to divide the data. However it is often the case that the data is far from linear and the datasets are inseparable. To allow for this kernels are used to non-linearly map the input data to a high-dimensional space. The new mapping is then linearly separable. A very simple illustration of this is shown below in figure 7.

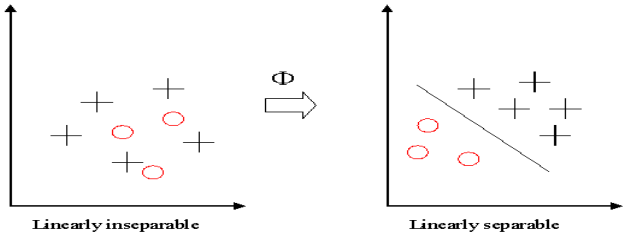


Figure 7: Why use Kernels?

This mapping is defined by the Kernel:

$$K(x,y) = \Phi(x) \cdot \Phi(y)$$

Feature Space: Transforming the data into feature space makes it possible to define a similarity measure on the basis of the dot product. If the feature space is chosen suitably, pattern recognition can be easy [1].

$$\langle x_1 \cdot x_2 \rangle \leftarrow K(x_1, x_2) = \langle \Phi(x_1) \cdot \Phi(x_2) \rangle$$

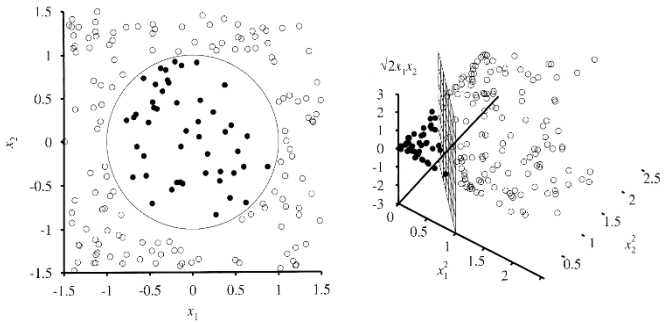


Figure 8: Feature Space Representation

Note the legend is not described as they are sample plotting to make understand the concepts involved.

Now getting back to the kernel trick, we see that when w,b is obtained the problem is solved for a simple linear scenario in which data is separated by a hyper plane. The Kenral trick allows SVM's to form nonlinear boundaries. Steps involved in kernel trick are given below.

[a] The algorithm is expressed using only the inner products of data sets. This is also called as dual problem.
[b] Original data are passed through non linear maps to form new data with respect to new dimensions by adding a pair wise product of some of the original data dimension to each data vector.
[c] Rather than an inner product on these new, larger vectors, and store in tables and later do a table lookup, we can represent a dot product of the data after doing non linear mapping on them. This function is the kernel function. More on kernel functions is given below.

Kernal Trick: Dual Problem
First we convert the problem with optimization to the dual form in which we try to eliminate w, and a Lagrangian now is only a function of λ_i. There is a mathematical solution for it but this can be avoided here as this tutorial has instructions to minimize the mathematical equations, I would describe it instead. To solve the problem we should maximize the L_D with respect to λ_i. The dual form simplifies the optimization and we see that the major achievement is the dot product obtained from this.

Kernal Trick: Inner Product summarization
Here we see that we need to represent the dot product of the data vectors used. The dot product of nonlinearly mapped data can be expensive. The kernel trick just picks a suitable function that corresponds to dot product of some nonlinear mapping instead. Some of the most commonly chosen kernel functions are given below in later part of this tutorial. A particular kernel is only chosen by trial and error on the test set, choosing the right kernel based on the problem or application would enhance SVM's performance.

Kernel Functions

The idea of the kernel function is to enable operations to be performed in the input space rather than the potentially high dimensional feature space. Hence the inner product does not need to be evaluated in the feature space. We want the function to perform mapping of the attributes of the input space to the feature space. The kernel function plays a critical role in SVM and its performance. It is based upon reproducing Kernel Hilbert Spaces.

$$K(x, x') = \langle \phi(x), \phi(x') \rangle,$$

If K is a symmetric positive definite function, which satisfies Mercer's Conditions,

$$K(x, x') = \sum_{m}^{\infty} a_m \phi_m(x) \phi_m(x'), \quad a_m \geq 0,$$

$$\iint K(x, x') g(x) g(x') dx dx' > 0, \quad g \in L_2$$

Then the kernel represents a legitimate inner product in feature space. The training set is not linearly separable in an input space. The training set is linearly separable in the feature space. This is called the "Kernel trick" [8] [12].

The different kernel functions are listed below: More explanation on kernel functions can be found in the book [8]. The below mentioned ones are extracted from there and just for mentioning purposes are listed below.

1] *Polynomial:* A polynomial mapping is a popular method for non-linear modeling. The second kernel is usually preferable as it avoids problems with the hessian becoming Zero.

$$K(x, x') = \langle x, x' \rangle^d.$$

$$K(x, x') = \left(\langle x, x' \rangle + 1\right)^d.$$

2] *Gaussian Radial Basis Function*: Radial basis functions most commonly with a Gaussian form

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

3] *Exponential Radial Basis Function*: A radial basis function produces a piecewise linear solution which can be attractive when discontinuities are acceptable.

$$K(x, x') = \exp\left(-\frac{\|x - x'\|}{2\sigma^2}\right)$$

4] *Multi-Layer Perceptron*: The long established MLP, with a single hidden layer, also has a valid kernel representation.

$$K(x, x') = \tanh\left(\rho\langle x, x' \rangle + \varrho\right)$$

There are many more including Fourier, splines, B-splines, additive kernels and tensor products. If you want to read more on kernel functions you could read the book.

8.8 Controlling Complexity in SVM: Trade-offs

SVM is powerful to approximate any training data and generalizes better on given datasets. The complexity in terms of kernel affects the performance on new datasets. SVM supports parameters for controlling the complexity and above all SVM does not tell us how to set these parameters and we should be able to determine these Parameters by Cross-Validation on the given datasets. The diagram given below gives a better illustration.

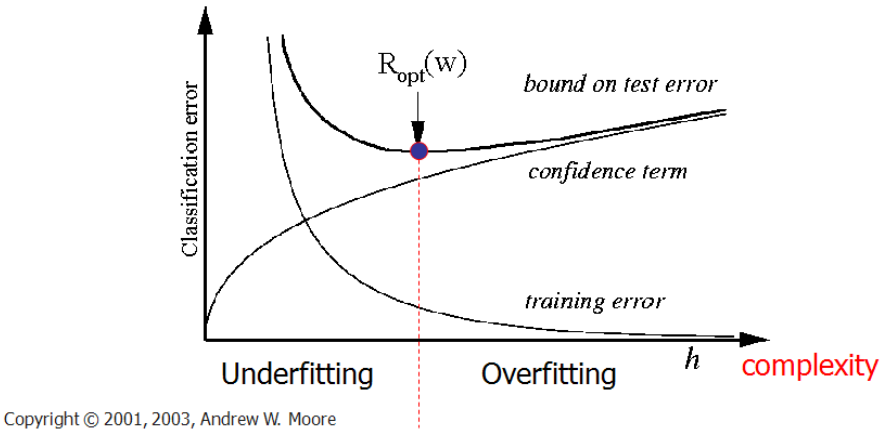


Figure 9: How to control complexity. Note the legend is not described as they are sample plotting to make understand the concepts involved.

8.9 SVM for Classification

SVM is a useful technique for data classification. Even though it's considered that Neural Networks are easier to use than this, however, sometimes unsatisfactory results are obtained. A classification task usually involves with training and testing data which consist of some data instance. Each instance in the training set contains one target values and several attributes. The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes.

Classification in SVM is an example of Supervised Learning. Known labels help indicate whether the system is performing in a right way or not. This information points to a desired response, validating the accuracy of the system, or be used to help the system learn to act correctly. A step in SVM classification involves identification as which are intimately connected to the known classes. This is called feature selection or feature extraction. Feature selection and SVM classification together have a use even when prediction of unknown samples is not necessary. They can be used to identify key sets which are involved in whatever processes distinguish the classes.

8.10 SVM for Regression

SVMs can also be applied to regression problems by the introduction of an alternative loss function. The loss function must be modified to include a distance measure. The regression can be linear and non linear. Linear models mainly consist of the following loss functions, e-intensive loss functions, quadratic and Huber loss function. Similarly to classification problems, a non-linear model is usually required to adequately model data. In the same manner as the non-linear SVC approach, a non-linear mapping can be used to map the data into a high dimensional feature space where linear regression is performed. The kernel approach is again employed to address the curse of dimensionality. In the regression method there are considerations based on prior knowledge of the problem and the distribution of the noise. In the absence of such information Huber's robust loss function, has been shown to be a good alternative.

8.11 Applications of SVM

SVM has been found to be successful when used for pattern classification problems. Applying the Support Vector approach to a particular practical problem involves resolving a number of questions based on the problem definition and the design involved with it. One of

the major challenges is that of choosing an appropriate kernel for the given application. There are standard choices such as a Gaussian or polynomial kernel that are the default options, but if these prove ineffective or if the inputs are discrete structures more elaborate kernels will be needed. By implicitly defining a feature space, the kernel provides the description language used by the machine for viewing the data. Once the choice of kernel and optimization criterion has been made the key components of the system are in place. Let's look at some examples.

The task of text categorization is the classification of natural text documents into a fixed number of predefined categories based on their content. Since a document can be assigned to more than one category this is not a multi-class classification problem, but can be viewed as a series of binary classification problems, one for each category. One of the standard representations of text for the purposes of information retrieval provides an ideal feature mapping for constructing a Mercer kernel. Indeed, the kernels somehow incorporate a similarity measure between instances, and it is reasonable to assume that experts working in the specific application domain have already identified valid similarity measures, particularly in areas such as information retrieval and generative models.

Traditional classification approaches perform poorly when working directly because of the high dimensionality of the data, but Support Vector Machines can avoid the pitfalls of very high dimensional representations. A very similar approach to the techniques described for text categorization can also be used for the task of image classification, and as in that case linear hard margin machines are frequently able to generalize well. The first real-world task on which Support Vector Machines were tested was the problem of hand-written character recognition. Furthermore, multi-class SVMs have been tested on these data. It is interesting not only to compare SVMs with other classifiers, but also to compare different SVMs amongst themselves [23]. They turn out to have approximately the same performance, and furthermore to share most of their support vectors, independently of the chosen kernel. The fact that SVM can perform as well as these systems without including any detailed prior knowledge is certainly remarkable [25].

8.12 Strength and Weakness of SVM:

The major strengths of SVM are the training is relatively easy. No local optimal, unlike in neural networks. It scales relatively well to high dimensional data and the trade-off between classifier complexity and error can be controlled explicitly. The weakness includes the need for a good kernel function.

CONCLUSION

Increasing teacher engagement has emerged as a key challenge for researchers, and educational institutions. Many of the current tools used to measure engagement – such as self-reports, teacher introspective evaluations, and checklists – are cumbersome, lack the temporal resolution needed to understand the interplay between engagement and learning, and in some cases capture teacher compliance rather than engagement. In this project we explored the development of realtime automated recognition of engagement from teachers' facial expressions. The motivating intuition was that teachers constantly evaluate the level of their students' engagement, and facial expressions play a key role in such evaluations. Thus, understanding and automating the process of how people judge teacher engagement from the face could have important applications.

Our work extends prior research on engagement recognition using computer vision and is arguably the most thorough study on this topic to date: We collected a dataset of teachers' facial expressions from online video lectures. We experimented with multiple approaches for human observers to assess teacher engagement. We found that inter-observer reliability is maximized when the length of the observed clips is approximately 10 seconds. Shorter clips do not provide enough context and reliability suffers. Longer clips tend to be harder to evaluate because they often mix different levels of engagement. When discriminating low v. high levels of engagement, inter-observer reliability was high (Cohen's _ = 0:96). We also found that the engagement judgments of 10-second clips could be reliably approximated (Pearson r = 0:85) by averaging single frame judgments over the 10 seconds. This indicates that static expressions contain the bulk of the information observers use to assess teacher engagement. Our results suggest that machine learning methods could be used to develop a real-time automatic engagement detector with comparable accuracy to that of human observers. We showed that both human and automatic engagement judgments correlate with task performance. In particular, teacher post-test performance was predicted just as accurately (and statistically significantly) by observing the face of the teacher during learning (r = 0:47) as from the pre-test scores (r = 0:44).However, a-posteriori statistical analysis suggests this may be due to ceiling effects and a fundamental limitation of short-term laboratory studies such as ours. In such studies, most teachers tend usually to be quite engaged, which is quite different from the long-term engagement or disengagement found in classrooms. This points to the importance of long-term studies that approximate the classroom ecology in which some teachers are engaged and others are chronically disengaged for days, months, and years.

While the progress made here is modest, it reinforces the idea that automatic recognition of teacher engagement is possible and could potentially revolutionize education as we know it. For example, using computer vision systems, a set of low-cost, high-resolution cameras could monitor engagement levels of entire classrooms, without the need for self-report or questionnaires. The temporal resolution of the technology could help understand when and why students get disengaged, and perhaps to take action before it is too late. Web based teachers could obtain real-time statistics of the level of engagement of their students across the globe. Educational videos could be improved based on the aggregate engagement signals provided by the viewers.

Such signals would indicate not only whether a video induces high or low engagement, but most importantly, which parts of the videos do so. Our work underlines the importance of focusing on long-term field studies in real-life classroom environments. Collecting data in such environments is critical to train more reliable and ecologically valid engagement recognition systems. More importantly, sustained, long-term studies in actual classrooms are needed to gain a better understanding of the interplay between engagement and learning in real life.

Accuracy

We use the SKLearn's metrics module to find the accuracy of our project. We use the accuracy_score function in python 3.5 to find the accuracy.

Accuracy of our project is approximately 0.64 using SKLearn's Support Vector Machine Classifier (with default kernels) with our online video lectures data. Whereas Human accuracy is approximately 0.69 and we are looking further for improvement.

PYTHON 3.5 SCRIPT

```
# DATA EXTRACTION

# MODULES REQUIRED_____________________________________:

from __future__ import print_function

from scipy import ndimage as ndi

from skimage.util import img_as_float

from skimage.filters import gabor_kernel

from python_speech_features import mfcc, logfbank

import scipy.io.wavfile as wav

import numpy as np

import cv2

import os

import csv



# Some variables for later use

num_of_feature_rows_per_vid = 200  # 20 features per image in 10 sec video and according to video 200 audio feature rows

num_of_thetas = 4  # thetas used in kernels for image features

frequencies = (0.10, 0.30, 0.50, 0.70, 0.90)  # frequencies used in kernels for image features

width_of_image = 100  # 100x100 dimension image is used for image features

Winlen = 0.1  # used in audio features. 0.1 means window length is 100ms

Winstep = 0.1  # used in audio features for adjusting audio frame settings


ALL_FEATS = []  # STORE ALL IMAGE AND AUDIO FEATURES


data_write_location = "./data_per_lec/"

data_file_extension = '.txt'
```

```
read_location = "./videos_and_audios_used_in_project/"
write_location = "./images_per_sec/"
video_extension = ".mp4"
image_extension = ".png"
audio_extension = ".wav"


# GET THE NAMES OF FILES FROM LOCATION OF YOUR DATA___:
videofiles_name = []
audiofiles_name = []
files_name = []
for unused1, unused2, files in os.walk(read_location):
    for file in files:
        if file.endswith(video_extension):
            videofiles_name.append(file)
        elif file.endswith(audio_extension):
            audiofiles_name.append(file)
        name = file.split(".")
        if name[0] not in files_name:
            files_name.append(name[0])

# Sort all names for similarity
videofiles_name.sort()
audiofiles_name.sort()
files_name.sort()

# IMPORT CLASSIFICATION LABELS AND COMBINE IT WITH DATA:
labels = {}
with open('LABELS.csv') as csvfile:
```

```
    reader = csv.DictReader(csvfile)
    for row in reader:
        labels = row
print (labels)

# EXTRACT AUDIO AND VIDEO FEATURES_____:
# Function for redefining the dimensions of the frame:
def resize_img(file_name, width):
    ratio = int(width) / file_name.shape[1]
    dim = (int(width), int(file_name.shape[0] * ratio))
    resized = cv2.resize(file_name, dim, interpolation=cv2.INTER_AREA)
    return resized

# Function for computing image features:
def compute_feats(image, kernels):
    feats = np.zeros((len(kernels), 2), dtype=np.double)
    for k, kernel in enumerate(kernels):
        filtered = ndi.convolve(image, kernel, mode='wrap')
        feats[k, 0] = filtered.mean()
        feats[k, 1] = filtered.var()
    return feats

# prepare filter bank kernels with 8 orientations and 5 frequencies for images:
kernels = []
for theta in range(num_of_thetas):
    theta = theta / 4. * np.pi
    for frequency in frequencies:
        kernel = np.real(gabor_kernel(frequency, theta=theta))
        kernels.append(kernel)
```

```
# Load haarcascade for face detection:
face_cascade = cv2.CascadeClassifier("haarcascade_frontalface_default.xml")


img_name = 0  # for naming cropped images
for i, name in enumerate(files_name):
    if videofiles_name[i].startswith(name) and audiofiles_name[i].startswith(name):
        image_feats_per_vid = []
        audio_feats_per_vid = []
        cap = cv2.VideoCapture(read_location+videofiles_name[i])  # Capture video from location
        FPS = int(cap.get(cv2.CAP_PROP_FPS))  # Count the FPS of given video
        print("Working on", name)
        print('Video FPS', FPS)
        per_vid_counter = 0  # for counting 10 images per video
        count = 0  # used with FPS
        while (cap.isOpened()):
            if per_vid_counter == 10:
                break
            ret, frame = cap.read()
            if ret == False:
                break
            gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
            faces = face_cascade.detectMultiScale(gray, 1.2, 5)
            for (x, y, w, h) in faces:
                face = resize_img(gray[y:y + h, x:x + w], width_of_image)  # Get faces from image
                if face.any() and (count % FPS == 0 or count % FPS == 1):
                    print("Extracting Faces", per_vid_counter)
                    cv2.imwrite(write_location+str(img_name)+image_extension, face)  # write the face image to given directory
```

```
            image = img_as_float(face)
            image_feats = compute_feats(image, kernels)  # 20 features per image
            image_feats_per_vid.append(image_feats)
            per_vid_counter += 1
            img_name += 1
            break
        count += 1
    cap.release()
    image_feats_per_vid = np.asarray(image_feats_per_vid)  # Convert the python list into numpy array
    image_feats_per_vid = np.reshape(image_feats_per_vid, (num_of_feature_rows_per_vid,-1))  # Converting 3d array to 2d array
    print ("IMAGE DATA SIZE PER VIDEO =", image_feats_per_vid.shape)


    # Getting audio features
    (rate, signal) = wav.read(read_location + audiofiles_name[i])  # Reading audio file
    mfcc_feat = mfcc(signal, rate, winlen=Winlen, winstep=Winstep)
    mfcc_feat = mfcc_feat[:num_of_feature_rows_per_vid, :]
    fbank_feat = logfbank(signal, rate, winlen=Winlen, winstep=Winstep)
    fbank_feat = fbank_feat[:num_of_feature_rows_per_vid, :]
    audio_feats_per_vid = mfcc_feat
    print ("AUDIO DATA SIZE PER VIDEO =", audio_feats_per_vid.shape)


    # Combine image and audio features with label into one list and attach it to file name using dictionary
    label = []
    for unused in range(num_of_feature_rows_per_vid):
        label.append(int(labels[name]))
    label = np.array(label)
```

```
        video_audio = np.column_stack((image_feats_per_vid,audio_feats_per_vid))
        full_feats = np.column_stack((video_audio, label))
        ALL_FEATS.append(full_feats)
        print("SHAPE OF FINAL DATA (with labels) =", ALL_FEATS[i].shape)


    # SAVE FEATURES TO A VIDEO NAMED FILE
    print("WRITING DATA TO FILE...")
    np.savetxt(data_write_location+name+data_file_extension, ALL_FEATS[i], delimiter=',')
    print("DONE WRITING")
    print("--" * 30)


# THIS LINE SHOULD BE AT BOTTOM ALWAYS
print("**"*10,"END OF PROGRAM","**"*10)


# APPLYING SVM
import os
import numpy as np
from sklearn import svm
from sklearn.cross_validation import train_test_split
from sklearn.metrics import accuracy_score
#np.set_printoptions(threshold=np.nan)  # For full array printing


# IMPORT DATA_____________________________________________:
data_read_location = "./data_per_lec/"
data_file_extension = ".txt"
```

```
filenames = []
for unused1, unused2, files in os.walk(data_read_location):
    for file in files:
        name = file.split(".")
        if name[0] not in filenames:
            filenames.append(name[0])
dataset = []
for i, filename in enumerate(filenames):
    dataset.append(np.loadtxt(data_read_location+filename+data_file_extension,
delimiter=","))
dataset = np.reshape(dataset, (-1,16))
print (dataset.shape)


# Separate features and output
X = dataset[:,:15]
y = np.ravel(dataset[:,15:])



# SPLIT TRAINING AND TESTING DATA_____________________:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30)



# APPLY SVM ON TRAINING SET_____________________________:
print("Training...")
SVM = svm.SVC()
SVM.fit(X_train,y_train)



# PREDICT WITH TESTING DATA_____________________________:
print ("predicting...")
```

```
y_pred = SVM.predict(X_test)



# CHECK THE ACCURACY OF YOUR MODEL_____:
accuracy = accuracy_score(y_test, y_pred)
print ("Accuracy =", accuracy)



print("**"*10,"END OF PROGRAM","**"*10)
```

REFERENCES

[1] Wikipedia Online. Http://en.wikipedia.org/wiki

[2] Tutorial slides by Andrew Moore. Http://www.cs.cmu.edu/~awm

[3] V. Vapnik. The Nature of Statistical Learning Theory. Springer, N.Y., 1995. ISBN

0-387-94559-8.

[4] Burges C., "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998, (Volume 2).

[5] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, pages 281– 287, Cambridge, MA, 1997. MIT Press.

[6] Theodoros Evgenuiu and Massimilliano Pontil, Statistical Learning Theory: a Primer 1998.

[7] Olivier Bousquet, Stephane Boucheron, and Gabor Lugosi, "Introduction to Statistical Learning Theory".

[8] Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.

[9] Image found on the web search for learning and generalization in svm following links given in the book above.

[10] David M Skapura, Building Neural Networks, ACM press, 1996.

[11] Tom Mitchell, Machine Learning, McGraw-Hill Computer science series, 1997.

[12] J.P.Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.

[13] Vapnik V., "Statistical Learning Theory", Wiley, New York, 1998.

[14] M. A. Aizerman, E. M. Braverman, and L. I. Rozono´er. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote

Control, 25:821–837, 1964.

[15] N. Aronszajn. Theory of reproducing kernels. Trans. Amer. Math. Soc., 686:337–404, 1950.

[16] C. Cortes and V. Vapnik. Support vector networks. Machine Learning, 20:273 – 297,

1995

[17] A. J. Smola. Regression estimation with support vector learning machines. Master's

thesis, Technische Universit¨at M¨unchen, 1996.

[18] N. Heckman. The theory and application of penalized least squares methods or reproducing kernel hilbert spaces made easy, 1997.

[19] Vapnik, V., Estimation of Dependencies Based on Empirical Data. Empirical Inference Science: Afterword of 2006, Springer, 2006

[20] http://www.enm.bris.ac.uk/teaching/projects/2004_05/dm1654/kernel.htm

[21] Duda R. and Hart P., "Pattern Classification and Scene Analysis", Wiley, New York 1973.

[22] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, Neural Networks for Signal Processing VII — Proceedings of the 1997 IEEE Workshop, pages 276 – 285, New York, 1997. IEEE.

[23] M. O. Stitson and J. A. E. Weston. Implementational issues of support vector machines. Technical Report CSD-TR-96-18, Computational Intelligence Group, Royal Holloway, University of London, 1996.

[24] Burges B.~Scholkopf, editor, "Advances in Kernel Methods--Support Vector Learning". MIT press, 1998.

[25] Osuna E., Freund R., and Girosi F., "Support Vector Machines: Training and Applications", A.I. Memo No. 1602, Artificial Intelligence Laboratory, MIT, 1997.

[26] Trafalis T., "Primal-dual optimization methods in neural networks and support vector machines training", ACAI99.

[27] Veropoulos K., Cristianini N., and Campbell C., "The Application of Support Vector Machines to Medical Decision Support: A Case Study", ACAI99

[28] A. Kapoor, S. Mota, and R. Picard. Towards a learning companion that recognizes affect. In AAAI Fall Symposium, 2001.

[29] A. Kapoor and R. Picard. Multimodal affect recognition in learning environments. In Proceedings of the 13th annual ACM international conference on Multimedia, pages 677–682, 2005.

[30] K. R. Koedinger and J. R. Anderson. Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8:30–43, 1997.

[31] M. Lades, J. C. Vorbr ¨ uggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. W¨ urtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computers, 42:300–311, 1993.

[32] R. Larson and M. Richards. Boredom in the middle school years: Blaming schools versu blaming students. American journal of education, 99:418–443, 1991.

[33] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. Image and Vision Computing, 24(6):615–625, 2006.

[34] G. Littlewort, M. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. Image and Vision Computing, 27(12):1797–1803, 2009.

[35] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. Computer expression recognition toolbox. In Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG'11), pages 298–305, 2011.

[36] R. Livingstone. The future in education. Cambridge University Press, 1941.

[37] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade dataset: A complete dataset for action unit and emotion-specied expression. In CVPR Workshop on Human-Communicative Behavior, pages 94–101, 2010.

[38] M. Mahmoud and P. Robinson. Interpreting hand-over-face gestures. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction, pages 248–255, 2011.

[39] S. Makeig, M. Westerfield, J. Townsend, T.-P. Jung, E. Courchesne, and T. J. Sejnowski. Functionally independent components of early event-related potentials in a visual spatial attention task. Philosophical Transactions of the Royal Society: Biological Science, 354:1135–44, 1999.

[40] S. Mason and A. Weigel. A generic forecast verification framework for administrative purposes. Monthly Weather Review,137:331–349, 2009.

[41] G. Matthews, S. Campbell, S. Falconer, L. Joyner, J. Huggins, and K. Gilliland. Fundamental dimensions of subjective state in performance settings: task engagement, distress, and worry.Emotion, 2(4):315–340, 2002.

[42] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser. Facial features for affective state detection in learning environments. In Proceedings of the 29th Annual Cognitive Science Society, pages 467–472, 2007.

[43] J. Mostow, A. Hauptmann, L. Chase, and S. Roth. Towards a reading coach that listens: Automated detection of oral reading errors. In Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93), pages 392–397, 1993.

[44] J. R. Movellan. Tutorial on gabor filters. Technical report, MPLab Tutorials, UCSD MPLab, 2005.

[45] H. O'Brien and E. Toms. The development and evaluation of a survey to measure user engagement. Journal of the American Society for Information Science and Technology, 61(1):50–69, 2010.

[46] J. Ocumpaugh, R. S. Baker, and M. M. T. Rodrigo. Baker-Rodrigo observation method protocol 1.0 training manual. Technical report, EdLab, Manila, Philippines, 2012.

[47] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. Systems, Man, and Cybernetics – Part B: Cybernetics, 36(2), 2006.

[48] J. Parsons and L. Taylor. Student engagement: What do we know and what should we do. Technical report, University of Alberta, 2011.

[49] A. Pope, E. Bogart, and D. Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. Biological Psychology, 40:187–195, 1995.

[50] K. Porayska-Pomsta, M. Mavrikis, S. D'Mello, C. Conati, and R. S. Baker. Knowledge elicitation methods for affect modelling in education. International Journal on Artificial Intelligence in Education, 2013.

[51] D. Shernof, M. Csikszentmihalyi, B. Schneider, and E. Shernoff. Student engagement in high school classrooms from the perspective of flow theory. School Psychology Quarterly, 18(2):158–176, 2003.

[52] K. VanLehn, C. Lynch, K. Schultz, J. Shapiro, R. Shelby, and L. Taylor. The Andes physics tutoring system: Lessons learned. International Journal of Artificial Intelligence and Education, 15(3):147– 204, 2005.

[53] P. Viola and M. Jones. Robust real-time object detection. International Journal of Computer Vision, 57(2):137–154, 2004.

[54] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan. Drowsy driver detection through facial movement analysis. In Proceedings of the IEEE International Conference on Human-Computer Interaction, pages 6–18, 2007.

[55] J. Whitehill, M. Bartlett, and J. R. Movellan. Automatic facial expression recognition for intelligent tutoring systems. In Proceedings of the CVPR 2008 Workshop on Human Communicative Behavior Analysis, pages 1–6, 2008.

[56] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. Transactions on Pattern Analysis and Machine Intelligence, 31(11):2106–2111, 2009.

[57] J. Whitehill and J. Movellan. A discriminative approach to frame-by-frame head pose tracking. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2008.

[58] J. Whitehill, Z. Serpell, A. Foster, Y.-C. Lin, B. Pearson, M. Bartlett, and J. Movellan. Towards an optimal affect-sensitive instructional system of cognitive skills. In Computer Vision and Pattern Recognition Workshop on Human Communicative Behavior, pages 20–25,
2011.

[59] 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, 4(3):129–164, 2009.

[60] T. Wu, N. Butko, P. Ruvolo, J. Whitehill, M. Bartlett, and J. Movellan. Multilayer architectures for facial action unit recognition. IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics, 42(4):1027–1038, 2012.

[61] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. Transactions on Pattern Analysis and Machine Intelligence, 29(6):915–
928, 2007.