A Crowdsourced Approach to Student Engagement Recognition in e-Learning Environments

Aditya Kamath,* Aradhya Biswas*
Georgia Institute of Technology
{adityakamath,aradhya.biswas}@gatech.edu

Vineeth Balasubramanian
Indian Institute of Technology Hyderabad
vineethnb@iith.ac.in

Abstract

Massive Open Online Courses (MOOCs) have initiated a revolution in higher education by providing opportunities for interested students to learn from the comfort of their individual locations at their desired pace. However, an important and highly successful aspect of traditional classroom education/pedagogy, which is modulating content delivery based on understanding real-time student feedback, is conspicuously missing in such e-learning environments. We aim to bridge this gap by proposing a system for automatic recognition of students engagement levels during e-learning sessions, using a crowdsourced discriminative learning approach. The key contributions of this work include a custom dataset for engagement recognition that will be made publicly available with corresponding crowdsourced non-aggregated labels, as well as a novel instance-weighted Multiple Kernel Learning SVM method that can directly consider vote distributions from crowdsourcing platforms in the learning methodology. Our results showed a 14% improvement on the dataset against traditional methods, and a 46% improvement when the most ambiguous class from the dataset is ignored, corroborating the promise of the method.

1. Introduction

The exponential growth of Massive Open Online Courses (MOOCs) (eg., EdX, Coursera) has transformed the landscape of modern education by providing for personalized delivery of knowledge. Though this is a step in the right direction, a lot remains to be achieved by these online platforms in coming close to the experience of traditional classroom teaching. One of the important aspects that is currently missing from these online education platforms is the live feedback from the teacher depending on the reactions of the student during a lecture. For example, in the middle of a lecture, if a student is not engaged or is confused, the instructor may seek to ask the student on what’s wrong, and he/she might then choose to repeat the concept or request the particular student to meet him/her later in the day to clarify the doubt. Such situations currently cannot be handled by online MOOC platforms. The lack of feedback mechanisms in MOOC settings has resulted in a staggering dropout rate of over 93% [18].

Courses consist of static content and it is up to the student to find answers to doubts, either through the internet or through other forms of communication such as mailing the instructor or class forums which can often take longer to receive answers. One effective approach to bridge this gap between traditional classrooms and online MOOC platforms is to provide feedback to the system about the behavioral state of the students watching the lectures online. Behavioral states could include the state of engagement, confusion, frustration, etc. This information could either be relayed to the instructor, who could intervene as required; or the content could be created with sufficient supporting materials, which could be picked seamlessly during the course of a single lecture depending on the behavioral state of the student watching the lecture. The proliferation in the use of mobile devices, such as smartphones and tablets, as well as the increasing availability of cameras in such mobile devices, play an important part in enabling such feedback. MOOC platforms could utilize the front-facing web cameras of these devices to gauge the behavioral state of the student during the course of the lecture.

Among the various possible behavioral states, engagement can be considered the most fundamental state in a learning environment. Any other affective state, such as boredom, confusion, sleepiness, etc. gets reflected in the engagement levels of a student. Hence, this work contributes to the feedback system in e-learning environments by addressing the problem of engagement recognition of students while they watch online lectures by rating them on a scale of 3 engagement levels. One can conclude that the eye gaze can be associated as a indication of engagement, this approach to engagement determination has two major issues. First, the eye gaze determination is in itself an unsolved

*These authors contributed equally to this work.
problem. Second, it can be noted that eye gaze is not a necessary condition for engagement, for example, a student may be analyzing and not making eye contact with the instructor. We note here that while engagement is very useful in e-learning, this work can also be applied in other application spheres, such as in advertising where it may help understand users’ preferences better.

An important aspect of this problem is the lack of a publicly available dataset on which various methodologies can be tested and compared. As part of this work, we have developed a dataset, which mimics real-world conditions of MOOC environments, using crowdsourced labels for engagement recognition. We then propose a novel approach of engagement recognition which makes use of data from the crowdsourced labels to improve the generalized performance of the system. We then show the experimental results of our approach, which shows promise. Our proposed methodology shows an improvement in performance of over 14% accuracy. Furthermore, when the class label with the neutral engagement level (which we found confused labelers) was removed from the dataset, our approach shows an improvement of about 46%, which showed the promise of the proposed methodology.

The remainder of the paper is organized as follows. In Section 2, we describe existing related work in the field of engagement detection, not just in the case of students but also in other fields such as advertisements. We also study existing commercial products that claim to solve a similar problem. In Section 3 we describe the development of our dataset. In Section 4 we describe the proposed methodology for engagement recognition, and Section 5 describes the results on our dataset. We conclude with discussions in Section 6.

2. Related Work

Since the exponential growth in popularity of MOOCs over the last few years, e-learning has received widespread attention from many research communities. Initial efforts focused on using machine learning methods in the development of the personalization of curriculum, adaptive evaluation and recommendation system based on learner preferences and browsing behavior, as in [11], [24], [16] and [7]. Multiple systems have been proposed in this regard, as in [8] and [16]. There have also been efforts in developing real-time adaptive e-learning systems that use visual clues to provide a further level of personalization; for example, [21] uses eye-tracking for the purpose. Even though this domain has progressed greatly in recent years, there have been very limited efforts in real-time understanding of the learner’s visual behavioral states, such as engagement, in real-world e-learning settings.

Determining the affective state of an individual using computer vision and machine learning methods has been studied for several years now [25][33]. However, most existing work has focused on the six basic expressions (happiness, sadness, anger, disgust, fear, surprise) and the facial action units associated with them, for example, as in [32][26]. Although the last few years have seen the growth in addressing the recognition of subtler affective states, including modeling emotions in terms of dimensions such as valence and arousal[14], very limited work has been carried out in recognizing affective states of e-learning settings, as described below.

In [15], Hernandez et al. modeled the problem of determining the engagement of a TV viewer as a binary classification problem, using multiple geometric features extracted from the face and head and SVMs for the classification. Considering the lack of any public dataset, a custom dataset was created in this work. However, the dataset was small and was labeled by a single coder. Relying on a single coder in the case of a subtle affect such as engagement can introduce personal bias and influence generalizability. Another drawback of the study is that it ignores the Hawthorne effect (subject awareness of experiment objectives during the capture, described further in Section 3) in the creation of the dataset. Further, the dataset was under fairly controlled settings, and do not capture the real-world issues of MOOC environments. Lastly, their dataset is not available to the community for further research.

On similar lines as our work, Whitehill et al. [30] carried out initial work for automatic determination of engagement in images by using a custom dataset created for this purpose. On one hand, the dataset created in this work also had the same issues: ignoring the Hawthorne effect, not reflecting real-world challenges of e-learning environments, non-availability of the dataset for the community to carry out further research, as well as labeling by a limited set of coders who used rules to label the engagement level in a face image. In contrast to both these aforementioned efforts, we propose the following contributions in this work: (i) we create a new dataset that will be made publicly available, which are captured as subjects watch videos in the MOOC setting; (ii) carefully considering the Hawthorne effect in the design of our dataset capture sessions; (iii) using the crowd to label our data to broaden the scope of applicability of the results of this work; and (iv) a methodology to integrate the varying opinions in crowdsourced labels in our engagement determination system.

From another perspective, the importance of the need for user engagement determination and its real-world applications can be gauged by the increasing number of commercial applications that have surfaced in recent years to address the very question. Commercial applications, such as Emotient [2], Emovu[3] and SightCorp[5], provide a similar affective state (called attentiveness, for example in SightCorp) in their systems. However, our studies with these
systems showed that their performance on challenging real-world videos is not satisfactory. For example, applications such as [5] present user attention determined by the eye gaze of the subject(s) as the sole determinant of the engagement level. This correlation between eye gaze and attention may not always hold, especially in online classrooms. To further illustrate this issue, Figure 1 shows the engagement level of a subject from our dataset as determined by the demo application of one of the commercial systems [2]. Comparing a subjective assessment of the images on the graph and the actual attentiveness score clearly shows the shortcomings of such state-of-the-art systems, this highlighting the need for further research. Hence, in this work, we focus our efforts on developing a system for images/videos in challenging real-world environments typically used for e-learning towards the overall objective of engagement determination.

In summary, while limited work has been carried out earlier in this area, no dataset is available publicly to train models for the recognition of student/user engagement (or other affective states that are relevant in e-learning environments). As mentioned earlier, existing work (and datasets) in the domain of affective computing have broadly focused on the commonly known main expressions such as joy, surprise, anger, fear, disgust and sadness. However, emotions and affective states are far more complex [6], and success in recognizing subtler affective states such as engagement is still in a nascent stage. Few commercial systems that attempt to recognize user attentiveness work under constrained settings, where the user is expected to be in a controlled environment and is looking into the camera’s field-of-view most of the time - which is inorganic considering that e-learning environments require the user to be engaged for longer duration of time. Lastly, affective states such as engagement levels are subjectively understood, and concurrence on labels may not be straightforward.

Due to the subtlety of the expression of engagement and the intuitive nature of its perception by humans, the annotation of such a dataset requires a large number of labels for each image from various coders to be able to elicit common understanding. In order to address this need, crowdsourcing [28][23][10] has emerged as a solution. Utilizing Software-as-a-Service (Saas) platforms such as Amazon Mechanical Turk and Crowdflower provide options to achieve this objective. Many efforts in recent years [20][12][17] have attempted to control the quality of crowdsourced labels as well as study methods on aggregating labels in different problems. However, integrating the presence of multiple labels in the classification method has remained a challenge, and we propose a new methodology for this purpose in this work (Section 4). We now describe the dataset that was developed in this work.

3. Dataset Creation

As mentioned earlier, in this work, we have created a dataset that is: (i) made publicly available for further research and development; (ii) captures the nuances of real-world settings in an organic manner that doesn’t impose constraints on the subject; (iii) provides labels of engagement levels that are obtained using a wider voter base (to address the issue of subjectivity and label concurrence). Earlier work in this area [30][15] attempted to address some...
of these issues in a limited manner, but do not capture real-world challenges; moreover, the datasets are not publicly available for further research.

The primary objective of this dataset was to obtain images/videos of subjects watching lectures on a computer screen in real-world conditions. To this end, a custom framework was created which would play a lecture video to the subject. The lecture was about 10 minutes in duration and a video of the subject was recorded using a webcam placed suitably on the screen to capture the subject’s face and some part of the upper body. The subjects were allowed to scroll through the lecture video as they wished, which resulted in video captures that sometimes were longer than the 10-minute duration of the lectures. The subjects were also asked some simple questions before and after the video regarding the content of the lecture in the video. This data however is not used in this study. The subjects that participated in this study were undergraduate and graduate students in the age group of 18-24 and the dataset consisted of videos of 23 subjects in total.

The dataset was intended to be as close to real-world e-learning settings as possible. Hence, the aforementioned capture sessions were conducted in a research lab and no restrictions were placed on the other occupants of the lab. Hence, the captured videos have significant background noise and clutter with people frequently walking around in the background. This can be seen in the sample images in Fig 2. Off-the-shelf consumer-grade web cameras were used for the capture to, once again, mimic the typically available cameras on computing devices owned by students taking online courses.

An important aspect of such a video capture is the awareness of the subject of being recorded, which makes the subject act unnaturally and, often, in favorable ways to the result of the experiment. This is commonly referred to as Hawthorne Effect [4]. However, capturing videos of subjects without their knowledge would be a case of privacy breach. Hence, in our recording sessions, the video of the subjects was captured without their knowledge; however, at the end of the session, the subject was told about the video recording and his/her consent was taken to use the video for research purposes. If the subject declined to give consent, the captured video was immediately destroyed. This approach ensure that the subject’s rights were not violated while the Hawthorne effect was mitigated.

Whitehill et al. [30] studied the usefulness of the entire video sequence against just the images in the sequence for labeling and classification towards engagement recognition, and concluded that engagement is more a spatial concept, rather than a spatio-temporal one. Motivated by this conclusion, we used the images of the subjects obtained from the frames of the video, rather than the entire video clips themselves. Images were obtained by sampling the captured videos at a fixed rate and extracting the frames.

The extracted images, which comprise our dataset, were then uploaded to CrowdFlower[1], a widely used crowdsourcing SaaS platform, with a task of asking the raters to pick one of 3 possible engagement levels for the subject in the given image:

1. Not engaged
2. Nominally engaged
3. Very Engaged

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of Images</th>
<th>Vote Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>400</td>
<td>13.75%</td>
</tr>
<tr>
<td>2</td>
<td>2245</td>
<td>46.99%</td>
</tr>
<tr>
<td>3</td>
<td>1763</td>
<td>39.25%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>4408</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Dataset statistics
The tasks on Crowdflower were designed to be small to not get the rater too bored, which would impact the labeling. The raters were also compensated monetarily according to standard Crowdflower rates. We also used features provided by CrowdFlower to verify the veracity of the labels, and mitigate the possibility of bots completing the task. We obtained at least 25 annotations for each image, by a total of 76 different raters that completed this exercise on Crowdflower. The final dataset statistics\(^1\) are given in Table 1. In this work, we used the majority-vote aggregation technique to decide the final label of a given image. The vote distribution column describes the percentage of votes that each class received in the entire dataset.

### 4. Instance-Weighted Multiple Kernel Learning SVM for Engagement Recognition: Proposed Methodology

Before we describe the proposed methodology for engagement recognition, we first study the accuracy of classification using a standard linear Support Vector Machine (SVM) classifier (using a 1-vs-all approach to handle the multi-class setting) on the developed dataset. Using the Viola-Jones face detector [29] for face cropping, the Histogram of Oriented Gradients (HoG) [9] as feature representations, and 3-fold cross validation across the dataset, the average accuracy obtained was 35.0\%. Considering the highest accuracies reported by [30][15] in similar work are also 50 – 60\%, we clearly see the potential for improvement in performance, which we attempt using the proposed method. (We note that the highest reported accuracies in [30], viz. \(\approx 71\%\) are in single-class classification settings, rather than a multi-class setting). We also studied the use of kernels in the SVM classifier on our dataset, but did not observe any significant improvements in performance.

#### 4.1. Multiple Kernel Learning

As a first step of improving the performance, we studied the use of Multiple Kernel Learning (MKL) inside the SVM framework. Use of Kernel methods that map data to a higher-dimensional space in order to better classify non-linearly separable data, has grown immensely over the past decade. However, one concern in kernel methods in SVMs is the difficulty in choosing a kernel that provides good results for a problem. In order to address this issue, Lanckriet et al. [19] allows the SVM framework to choose a linear combination of kernels that maximize a cost function such as the maximum margin classifier, and learn the weights of the kernels in the final model. This approach is termed as MKL, and many such algorithms have been studied in [13] along with their complexities.

The traditional SVM formulation (given in Section 4.2) is in the primal space as a minimization problem. The same problem can be viewed in the dual space as a maximization problem as given below.

\[
\max_{\alpha} L(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y(i)y(j)\alpha_i\alpha_j \langle x(i), x(j) \rangle
\]

\[
\text{s.t} \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, m \\
\sum_{i=1}^{m} \alpha_i y(i) = 0
\]

The MKL formulation can be obtained by replacing the inner product with a linear combination of kernel Gram matrices, \(K_i\), as below:

\[
\min_{p \in P} \max_{\alpha \in Q} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} (\alpha \odot y) ^T (\sum_{i=1}^{m} p_i K_i) (\alpha \odot y)
\]

Here \(P = \{ p \in \mathbb{R}^m : p^T e = 1, 0 \leq p \leq 1 \}\) represents the set of kernel weights and \(Q = \{ \alpha \in \mathbb{R}^m : \alpha^T y = 0, \alpha \geq 0 \}\) represents the SVM dual variables, \(e\) is a vector of all ones, \(\{ K_i \}, i = 1, 2, \ldots, m\) is a group of kernel matrices that are defined on the projected data \(x_i\), and \(\odot\) denotes the vector dot product. This MKL formulation can be solved using semi-definite programming [19], but more recent approaches include Quadratically Constrained Quadratic Programming, Semi-Infinite Linear Programming and Sub-gradient Descent.

Given that MKL learns a more sophisticated kernel by using a linear combination of kernels, we expected improvements in the classification accuracy over the regular SVM. We trained a MKL classifier with a set of polynomial and RBF kernels across a range of parameters, and used the same training and testing set used as mentioned at the beginning of Section 4. We used the SimpleMKL library [22] for these experiments. The average accuracy obtained was 36.46\%, which is only a slight improvement from the traditional SVM, thus motivating an even better approach.

#### 4.2. Our Approach: Instance-Weighted MKL SVM

The traditional soft margin SVM can be formulated as follows:

\[
\min_{\gamma, w, b} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{m} \xi_i
\]

\[
\text{s.t} \quad y(i)(w^T x(i) + b) \geq 1 - \xi_i, i = 1, \ldots, m \\
\xi_i \geq 0, i = 1, \ldots, m
\]

In the above formulation \(x(i)\) is the feature vector of the \(i^{th}\) data point having label \(y(i) \in \{-1, 1\}\). \(\xi_i\) represents the

\(^1\)Comparing to 7574 images in the HBCU data set (primary data set used in [30]) and 14 videos of 25 minutes sessions in [15].
error allowable in determining the optimal decision boundary. The parameter $C$ controls the balance between making the objective $\|w\|^2$ minimum and also ensuring a margin of at least 1 for most samples. The issue, however with overlapping data, is that a single constant $C$ might not suffice to ensure both the above mentioned goals and ensure a good classification margin.

A simple solution to this as described by Wu and Srihari [31] is to allow $C$ to be different for each instance of the training data. Hence, instead of representing a single constant, we now have a vector of values which would control the importance of a particular point in the training data. The new SVM formulation, called the instance-weighted SVM, changes only slightly to incorporate this as given below. Notice that we now have $g(v_i)$ instead for each instance.

$$
\min_{\gamma,w,b} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} g(v_i)\xi_i
$$

s.t. $g^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, i = 1, \ldots, m$

$$
\xi_i \geq 0, i = 1, \ldots, m
$$

Here $g(v_i)$ is a monotonically increasing function on some parameter $v_i$ of the instance $i$ that is specified externally. It must be noted that each of the $v_i$s and the function $g(v_i)$ are to be provided and are not learnt as part of the optimization process. [31] also gives evidence to show that this approach gives better performance especially when the weights are chosen correctly.

The instance-weighted SVM and Multiple Kernel Learning methods are used as an integral part in determining the engagement level of a test image in our approach. From our observation of the crowdsourced data and the experiments we conducted, there was a large overlap of data across engagement levels, and this was much expected since there is no hard boundary in labeling an image. The most common approach to getting around this problem is to use soft-margin SVMs with kernels to make the data as linearly separable as possible. The slack variable in the soft margin SVM controls how much a point can deviate from the decision boundary on the opposite side. Using the instance-weighted approach, we can now control the importance of a particular point in the training set. In our approach, to account for the importance or slackness of a point during training, the following algorithm was employed. Each training sample had a distribution of votes across the three engagement levels as obtained from the crowd sourced data as $\pi_1, \pi_2, \pi_3$ where $\pi_j$ represents the votes the image got for engagement level $E_j$. (Each $\pi_j$ is 1-normalized). When training the SVM for engagement level $E_j$, we set the parameter $v_i = a \pi_j$, where $a$ is a normalization factor, and use the function $g$ as described in Equation 2. Both $\alpha$ and $\gamma$ are parameters in Equation 2.

$$
g(v_i) = \alpha e^{\gamma v_i}
$$

Section 5 details the values of the parameters used in the experiments. Such an approach to use the vote distribution for instance-weighted SVMs in crowdsourced datasets has not been done before, to the best of our knowledge. Besides, Equation 2 can also be interpreted as assigning an instance-weight, based on each label’s contribution to the overall entropy ($\sum_{i \in \text{labels}} p_i \log p_i$) of the vote distribution across the labels for a given data instance.

Multiple kernel learning allows us to automate the process of finding the best kernel (or a linear combination) of a set of kernels which are provided as input. Considering the amount of overlap in our dataset, MKL allows us to specify the kernels that could be useful for separating the data, and let the system decide the linear combination that would work best for the dataset. Hence, we use an instance-weighted MKL SVM in this work for engagement recognition, where the instance weights are given in Equation 2. A one-vs-all approach was used to make this method applicable for multi-class classification. We now describe the experimental results of the proposed method on our dataset.

5. Experiments and Results

Our experiments were carried out on the dataset captured as part of this work (Section 3). 3-fold cross-validation (with each fold obtained randomly from the entire dataset, but maintaining the same proportion of class labels as the overall dataset) was used to study the performance of the method, as in earlier experiments described so far. We used 1-fold for training and 2-folds for testing. For each input image, face detection was carried out using the Viola-Jones face detector [29] to obtain a bounding box for the face of the subject. The face is then cropped and the rest of the image is discarded. The cropped face is resized to $100 \times 100$ for normalization purposes. Figure 3 shows sample images from our dataset after face cropping. We then used the Histogram of Oriented Gradients (HOG) [9] as the feature representation for this face patch to obtain a final feature vector of dimension 31. This is obtained by considering the window to be the entire image and using a bucket size of 31 for the gradient vectors. VLFeat [27] was used to obtain the feature vectors. These features were then used to train the instance-weighted MKL SVM to build a model, which was then evaluated on the test portion of the data to measure the performance of the system. A set of polynomial and RBF kernels were specified for MKL. All our experiments were carried out on a HP ProLiant Server with Intel Xeon X5675 and 32GB of memory.

The average accuracy of these runs was 43.98% while the maximum accuracy obtained was 50.77%. While the
accuracy by itself is not high, this a considerable improvement over the accuracy obtained by using the traditional SVM and MKL approaches. Also, the challenging nature of the dataset makes it a very difficult problem, and this performance is a significant benchmark to begin with. We hope that making the dataset publicly available will further improve this benchmark in the near future. An interesting observation during our experiments was also that during training, both the plain SVM and MKL-SVM approaches usually terminated after reaching the maximum number of iterations specified for the optimization algorithm, whereas our approach with instance-weighted MKL terminated much earlier. This indicates that our approach was able to make the training set more linearly separable, hence allowing the optimizer to reach the provided error bound in fewer iterations.

**Variation of Accuracy with $\gamma$**

An important parameter in the training phase is the instance weight given to each instance. As mentioned in Equation 2, the values of $\alpha$ and $\gamma$ can alter the results drastically. However, based on initial experiments, we found $\alpha = 1$ to be the ideal choice. With this value of $\alpha$, we studied the variation of classification accuracy with respect to $\gamma$. Figure 4 shows these results. Based on these results, $\gamma = 20$ provides the best results. We also note that the value of $\gamma$ cannot be made too large, in which case we might hit an overflow in the calculation of the exponent.

**Experiments without Images from Engagement Level 2**

Another interesting insight we had in our dataset was that engagement level 2 in our setting was the most confusing class for the crowdsourced labelers, expectedly. In numbers, this meant that the vote distribution, although favored the label for the image to be engagement level 2, the votes in other classes were very close by and in some cases, the sum of votes of the other classes exceeded the votes obtained for engagement level 2. In order to better verify the usefulness of our proposed methodology without such inherent label domain issues, we ran our experiments without images in engagement level 2 and we obtained a classification accuracy of 75.77%. In comparison, running only a linear SVM on this set of images resulted in an accuracy of 18.15% and the MKL-SVM provided an accuracy of 29.24%. This experiment clearly corroborated the potential of the instance-weighted MKL SVM proposed in this work. These experiments were conducted with the best setting in our study which was obtained with $\gamma = 20$ and the same HOG feature vector.

**Accuracies of Individual SVMs**

Considering the way we define engagement levels (Section 3), we recall that we have solved the multi-class classification problem using a 1-vs-all SVM approach. However, considering the challenging nature of the dataset and the ambiguity in the crowdsourced labels, we studied which class among the three was the hardest to classify, which could give insight into future improvements of our system. Towards this end, we considered each SVM of the 1-vs-all setting individually and computed the classification accurac-
racies. Table 2 reports the individual accuracies of the best setting that we observed. Among the three, the SVM for engagement level 1 clearly has a much better accuracy than the rest, which is quite intuitive since its much easier to judge if a person is not engaged than how engaged a person is.

<table>
<thead>
<tr>
<th>SVM</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM for $E_1$</td>
<td>81.59</td>
</tr>
<tr>
<td>SVM for $E_2$</td>
<td>50.9</td>
</tr>
<tr>
<td>SVM for $E_3$</td>
<td>60.01</td>
</tr>
</tbody>
</table>

Table 2: Accuracy (%) of individual SVMs

6. Discussions

The proposed pipeline as described in Section 5 uses a face detector to crop the face from an image to be used for further processing. However, we found that most off-the-shelf face detectors were not sufficient for the challenging images in our dataset. Images in which the face was tilted or had obstructions to the face were the ones where the face detector would fail to detect a face. This was mostly true for the images in $E_1$. For our experiments, we manually cropped the images in which the face was not detected to make the image useful in our training and testing set. Another approach that seemed to work well was to use face tracking in the videos that we captured and then use the bounding boxes as reported by the tracker for the frames of interest. We plan to explore this further in the near future.

As noted in Section 5, many of the images in our dataset had a vote distribution which was very even and the majority vote was often decided by a small margin. There also seemed to be some inconsistencies with votes that similar images received. Figure 5 shows an example of two pairs of similar images which received a different label based on the majority vote aggregation approach. In our setting of crowdsourcing annotation, no specific instructions were given as to how the engagement level should be determined to allow for common knowledge to emerge, and our definition of $E_2$ seemed to be the most confusing class for most people. As seen in Section 5, our approach resulted in a much better accuracy when the images in $E_2$ were ignored. This would imply that images as those shown in Figure 5 prevent the classifier from arriving at a good decision boundary with minimal classification error which results in an overall poor accuracy. This is another direction of our future work.

7. Conclusions and Future Work

In summary, engagement recognition is an intuitive process to a human, and we have attempted to model this intuitive process in this work. In particular, our dataset for the purpose of engagement recognition was captured in a real-world setting, and was annotated via crowdsourcing instead of the previously followed approaches of using a small set of expert coders. We then propose a new approach of incorporating the crowdsourced labels into the training phase of a classifier, viz. the instance-weighted MKL SVM approach, and report the results of experiments conducted in this setting. The results show that our approach is promising; however, when using the multi-class formulation on the difficult dataset that we captured, the results allow room for further research. This observation, however, is also shared by similar failures of even commercial systems that claim to solve the same problem. However, the performance of the proposed method on the same dataset without the most ambiguous class gives impressive results, considering earlier results in this domain. We believe this approach can be adapted to other applications that use crowdsourced labels, and thus can be generalized. Our future directions of work include studies with video snippets for engagement determination (instead of images), and exploring methods to reduce ambiguity in the dataset by a deeper understanding of the vote distribution in the crowdsourced label space. The implementation of the proposed pipeline and dataset captured in this work is available at: https://github.com/e-drishti/wacv2016.

Acknowledgment The project was supported by the IBM Students for a Smarter Planet Award.

References


