Detecting and Categorizing Indices in Lecture Video using Supervised Machine Learning

Christopher Brooks, G. Scott Johnston, Craig Thompson, and Jim Greer

University of Saskatchewan, Department of Computer Science
110 Science Place, Saskatoon, SK
cab938@mail.usask.ca, {g.scott.j,craig.thompson,jim.greer}@usask.ca

Abstract. This work reports on the evaluation of detecting scene transitions in lecture video through supervised machine learning. It expands on previous work by gathering training data from multiple human raters. We include a robust evaluation that compares predictions against the entire set of expert classifications in disagreement. Finally, we explore some of the issues around constructing training data from multiple human experts, specifically emphasizing that evaluation strategies should be carefully considered when using aggregated training data.

1 Introduction

Computer-based lecture capture and media systems, such as Opencast Matterhorn[1] and echo360[2], provide the ability to record video of classroom lectures, including projector content such as PowerPoint slides or other desktop activity. Our interest in these technologies is to enable fast, accurate navigation through content by way of thumbnails that represent the start of new segments in a lecture. These segments can be thought of as roughly corresponding to new slides in a PowerPoint presentation, though our intent is to work with broader forms of presentation and not rely on any particular technology or lecture paradigm. Unlike previous work which has used static algorithms [1] or learning algorithms trained on data from a single human rater [2] for determining these kinds of indices, we provide a method to compare algorithms based on multiple raters that are in disagreement. The result is an increase in the quality of indexing compared to non-trained algorithms, and a method that considers training data from multiple raters.

One contribution of this work is an exploration of the effects of considering multiple raters when annotating data for supervised machine learning. In previous work [2], algorithms trained with data from a single rater were shown to produce better results than static algorithms. We go further and demonstrate *a*) the challenges in achieving agreement between multiple raters in this domain, and *b*) how aggregates for training can be formed on these conflicting ratings, and how these aggregates compare with the current state-of-the-practice.

[1] http://www.opencastproject.org
[2] http://echo360.com/

This paper is organized as follows: Section 2 presents a case study, the kind of data we are interested in, and how human raters perform the categorization task. Section 3 compares our approach to other popular methods. Finally, the paper closes in Section 4 with a description of ongoing work in how supervised learning can be used to classify segments of video and other avenues of further research.

2 Human Indexing in Lecture Video: A Case Study

We performed a case study to collect training data for our supervised approach and to measure how well different people agree on indexing in educational video. In this study we have a set of videos captured from the data projector for nine lectures from a single undergraduate course in Computer Science. These videos are roughly 80 minutes each, and are broken into still images at one frame per second, resulting in a total of 43,770 video frames. These video frames were shown to six study participants who were selected to control for gender (three males and three females) and educational experience (they all had taken a university level course within the last two years). In contrast to our previous work [2], we explicitly sought out participants who were neither graduate students nor instructors, as we wanted to examine how non-pedagogues would perform educational video content indexing. While a group of six participants may be too small to make generalizations about how the general population indexes video, it is large enough to illustrate some of the problems encountered and to compare with unsupervised algorithms.

Study participants were instructed to mark index points in each video using a tool that allowed them to navigate through the video on a frame-by-frame basis. A purpose-based goal was used as a motivator for this task: learners were to "...mark all transitions as if [they] were building the left hand navigation window for a lecture video player," and were shown an image of a production lecture capture system to help better understand the task. Based on previous experiments, learners were dissuaded from semantically analyzing the content, and were asked to "mark transitions based on visual changes that [they thought] would be helpful for this lecture and other similar lectures." Finally, subjects were asked to limit their index choices to between fifteen and thirty indices per lecture video. Participants were able to perform this task at their own pace, and were free to navigate back and forth through the video, and select/unselect index points as they saw fit.

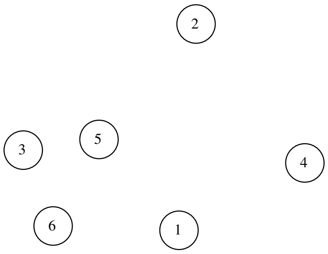
Fleiss' κ [3] is a measure of interrater reliability[3] useful for determining the agreement between pairs of people or within a larger group. We use it here to examine the relationship between our human participants; and later to evaluate the accuracy of our models. It ranges from -1.0 to 1.0. κ is chance-corrected:

[3] We use two vocabularies to refer to our study subjects: participants and raters. This is because the κ literature refers to raters that provide disagreeing ratings of instances; when discussing κ in general, we default to a rater/rating vocabulary to describe participants and their classifications of instances.

thus a score of 0 indicates only chance agreement, and positive scores indicate agreement beyond what would be expected by chance alone.[4] An issue with using κ in groups is that as the number of total raters increases, the amount of change any single rater can make to the overall level of agreement within the group decreases. To avoid ignoring differences between human raters, it is therefore useful to measure all pairwise κs (for our study, Figure 1a) to understand the significance of each individual's ratings and to observe any outliers.

	2	3	4	5	6
1	.532	.581	.520	.657	.609
2		.436	.481	.461	.456
3			.421	.774	.777
4				.432	.395
5					.734

(a)



(b)

Fig. 1: (a) contains pairwise κ between raters. These were calculated with the standard formula for κ; the input data is a series of yes/no classifications of each video image by two participants. Two simultaneous yeses or noes is an agreement. (b) is an agreement map for 6 raters where distances between nodes are approximately inversely proportional to κ (raters closer together have higher agreement).

Using terminology from Landis and Koch [4], the κs between pairs of our participants (Figure 1a) range from "fair" ($\kappa = 0.395$) to "substantial" ($\kappa = 0.777$). These values vary from the group κ of the participants; in fact, their group κ is 0.577, and the mean of their pairwise κ is 0.551. This would indicate that the whole of their agreement is better than the sum of the parts of their agreement. Figure 1b visualizes the agreement between raters using a spring model. Each node represents an individual rater, and the Euclidian distance between nodes approximates the level of agreement between two raters. In the figure, raters two and four have the largest distance between themselves and other raters, indicating that their ratings of indices shared less is common with the other raters. Raters three, five, and six form a more tightly knit group, a reflection of their high level of agreement relative to the others ($\kappa \geq 0.734$ from Figure 1a).

After the first phase of the study, we had a consensus building activity with the participants that included a verbal discussion of the strategies each individual

[4] κ is chance-corrected, as it is a ratio of the agreement achieved in a group, $\bar{P} - \bar{P}_e$, to the agreement that would have been achieved by chance alone, $1 - \bar{P}_e$. The details of its calculation are omitted for space.

used to generated indices. We found no evidence that the human experts could predict who they were most similar to or different from, further illuminating the complexity inherent in ill-defined domains. Beyond the collection of human rater data, the lack of agreement between subjects is an important takeaway for our machine indexing task. This suggests that forming training data based on multiple raters will lead to more robust and less over-trained results. This is explored in the next section, with a focus on quantifying the effects of using multiple raters in disagreement.

3 Training Strategies from Ratings in Disagreement

As our goal is to mark significant slide transitions, our approach used attributes that each represent some measure of the difference between two images that are adjacent in time. Thus, an image at time t has a feature vector of attributes, each based on the difference between it and the image at time $t-1$. For tractability, the images were taken from videos at one second intervals.

In total, nine different high level approaches were used to generate our attributes. The attribute groups are mostly taken unmodified from [2], or, were either minimally modified from [2] or first used in this work. In total, our dataset consisted of 134 different attributes for each sequential pair of images. For example, one attribute first used in this work was a "form" attribute, where our goal was to replicate a preceding phase that determined a semantic tag for images. To represent this attribute, we hand-coded each image as one of four such tags, and used a decision tree's classifications as an attribute in our main classification task. While some of our attributes are novel and worth reporting on, we leave the specifics of attribute formation and evaluation for future work, and instead focus on the overall process of gathering training data from multiple human raters and appropriately evaluating results from this type of data.

We generated models for classifying pairs of images as scene transitions, using our attributes and the J48 decision tree algorithm, as implemented in the WEKA Machine Learning toolkit.[5] As this algorithm requires a single classification per training instance, and because there was a lack of consensus from our human raters, we trained six J48 decision trees ($T_1, \ldots, T_6$) using six different aggregations of the participants' ratings. To aggregate the training data for each T_i, i is the minimum number of experts who must agree in order for an instance to be marked as a transition in the training set. Thus, the aggregate for T_1 required only one individual to indicate an index for a given instance, and T_6 required all six to agree; the T_1 aggregate therefore had the most positive classifications of indices, and T_6 had the fewest (approximately 30). All algorithms were evaluated using ten-fold cross-validation.

For comparison, we also evaluated the accuracy of three non-learning algorithms. The first of these, *Time*, is a naïvealgorithm that would select an index every 180 seconds into the video regardless of the content of video frames. The

[5] http://www.cs.waikato.ac.nz/ml/weka/

second, *Opencast*, was the method used in production code by the Opencast Matterhorn project. This algorithm uses differences in RGB intensities between frames to detect changes. The final algorithm we considered, *Dickson*, from [1], is a multi-pass image processing function that examines both pixel and block characteristics of video to determine stable events. All of these algorithms have seen real-world deployment in lecture capture systems, though few studies have been done as to the quality of these methods.

It is important to note that our usage of κ in this section differs from its typical usage in the supervised machine learning literature. Often, κ is used to compare the agreement between the model that results from a machine learning algorithm, and a test dataset; in general, κ is used to measure agreement between one algorithm and one training set, or within a group of humans. However, our goal is not to evaluate the ability of the algorithm to develop an accurate model of the aggregated training data, as we are using existing algorithms that have been demonstrated to be capable. Instead, our goal is to evaluate the ability of the resulting model itself (not the algorithm that generated a model). Specifically, we are evaluating a model's ability to replicate the pattern of decision making that our human raters used when they classified our training data. Thus our goal in using κ is to evaluate the ability of a trained model to match the unaggregated human raters' opinions. Future work to evaluate the suitability of the algorithm should also consider other metrics of evaluation (such as accuracy and precision/recall curves).

As κ can evaluate the agreement between multiple raters simultaneously, we can use κ to evaluate how well our model's predictions agree with the set of all six human raters at the same time, rather than an aggregation. This allows us to compare our six aggregation strategies to each other more fairly, by consistently measuring the agreement of T_1 through T_6 with all six human raters, rather than comparing each one to its own training aggregate. Recall that we are interested in finding the aggregation strategy that produces a model that best agrees with the human raters, not the model that best agrees with its own training aggregate. This distinction is subtle, but important.

We therefore believe it is more appropriate to use κ to compare the learned model with the entire group of expert raters, rather than comparing the learned model to the training set that was aggregated from the group of expert raters. The results presented in this section thus represent the ability of a model (T_1 through T_6, *Time*, *Opencast*, *Dickson*) to agree with the opinions of human raters, rather than the model's ability to agree with its training data.

However, using κ for groups instead of pairs requires a different method of interpreting the values of κ. This is because the group of human raters is large in size compared with the addition of one new rater, and because the group of human raters are often largely in agreement. A group κ includes the human raters' agreement, and is only somewhat affected by the addition of a rater. To more clearly see the effect of any particular model, we should contextualize the κs with how it changes based on one rater's choices. To do this, we calculate an upper and lower bound on the possible κ by computing κ when an artificial rater is added to the group. Since it is more likely that a set of raters will disagree partially on any given instance than that they will agree unanimously, the maximum agreement an added artificial set of ratings could achieve is determined by adding a rater who always agrees with the majority; the same is true for the minimum level of agreement, obtained by adding a rater who disagrees with the majority. We can construct these two maximally agreeing and minimally agreeing ratings for any group of raters, and their group κs with the original group of raters, are the upper and lower bounds of possible κ values for the given group and any additional rater.

Table 1 presents these lower and upper bounds on κ given our group of human raters, and shows that any given algorithm can at best raise the group κ to 0.626, or at worst lower the group κ to -0.15. The results of adding ratings by the different comparison algorithms (*Time*, *Opencast*, *Dickson*) to the human rater values, as well as the results of our trained algorithms (T_1 through T_6) are also shown in Table 1.

κ Bounds		Comparison Algorithms			Our Trained Algorithms					
upper	*lower*	*Time*	*Opencast*	*Dickson*	T_1	T_2	T_3	T_4	T_5	T_6
0.626	-0.15	0.391	0.370	0.448	0.574	0.565	0.565	0.537	0.530	0.487

Table 1: Group κ between raters and algorithms. *upper* and *lower* are the max and min values any algorithm could provide for κ. Recall that κ between the six expert raters without an algorithm was 0.577.

To interpret the κs of a system's trained models versus comparison algorithms, we can compare the κs with the group κ of the six participants together but without an algorithm (0.577). We note that each comparison algorithm lowers κ more than any of our trained algorithms. We can interpret this as saying that our training strategies provide better indexing results when compared with human experts than the automated algorithms.

In the case of a study that does not have comparison algorithms, because of the expense of reimplementing other works or for other reasons, the value of the κs of the trained algorithms can still be compared versus the calculated bounds. While Landis and Koch [4]'s subjective categorizations would indicate that κs near 0.5 show only "moderate agreement," the κs achieved here are reasonably close to the maximum possible once you compare with the entire group of human raters. Further, our model T_1 provided nearly the same level of agreement as the group of expert raters alone ($\kappa = 0.574$ versus $\kappa = 0.577$); this suggests that our model T_1 finds index points as reliably as any one of our human experts.

It is also interesting to note that as we require more consensus from our training set (e.g. T_6 instead of T_1), κ decreases. We believe the increasing sparseness of positive instances decreases the models' accuracy. This further casts doubt on generic aggregation strategies in the case of significant class imbalance.

4 Conclusions and Future Work

This paper advances the state of the art in generating indices from video using supervised machine learning. This is a subjective and ill-defined task, requiring human raters to mark points within a video to be used for navigation. The task results in a subjective dataset that contains disagreement between raters that should not be dismissed as noise. In particular, we have provided two contributions to the fields of machine learning and user modelling.

1. Often, in the field of machine learning, predictive accuracy of a model is quantified by comparing the model's predictions with true values with a given evaluation metric. We disagree with this strategy in the case of gathering human participants' opinions for training an applied system; here, the goal of evaluation should be to judge the model's ability to predict the participants' opinions, not the ability of an algorithm to create a model that agrees with its training data. We draw this conclusion directly from our pairwise use of κ in Section 2: although participants given the same task produce reasonable aggregated training data, the process of aggregation removes nuances of their disagreement, which can influence model accuracy. In the future, we hope to continue examining how individual raters' opinions are better predicted by specific attribute types, and how to better aggregate multiple ratings as training data for algorithms that accept only one class attribute.
2. We demonstrated how bounds on κ could be determined so that κ can be used effectively for comparing models with groups of human raters. Although κ can be used to compare a model's predictions with test data, by comparing an algorithm with humans that represent a target audience for a practical system, we can directly evaluate the algorithm's usefulness at emulating opinions humans would find useful. Interpretation of group κs with bounds is important because of how κ is normally reported as significant for values larger than 0.0, as in [4]; but we demonstrate that even naïve algorithms such as *Time* achieve "significant" agreement because of the inherent agreement in the group itself. Instead, we suggest determining whether a κ measured with a group of humans *a*) is close to the upper bound on achievable κ, and *b*) is close to the contribution to κ that each disagreeing human rater makes.

References

1. Dickson, P., Adrion, W., Hanson, A.: Automatic Capture of Significant Points in a Computer Based Presentation. In: Eighth IEEE International Symposium on Multimedia (ISM'06). (2006) 921–926
2. Brooks, C., Amundson, K.: Detecting Significant Events in Lecture Video using Supervised Machine Learning. In: 2009 Conference on Artificial Intelligence in Education. (2009)
3. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5) (1971) 378–382
4. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics **33**(1) (1977) 159–174