Video Presentation Slide Alignment

Apurva Shah

Abstract

This paper presents an application of machine learning used to enhance slide recognition and alignment in online video presentations. I describe an approach based solely on image features, intended to be robust enough to handle the wide variety of online videos. Using a binary classifier to determine whether a given frame and slide match, coupled with a temporal Markov process, the alignment produced is better than using vanilla object recognition techniques. The classifier is trained on features derived from SIFT keypoint matches between frame slide pairs. This approach proves robust to partial occlusion and moderately tolerant to poor lighting conditions. Empirical results demonstrate that a large and diverse training set is necessary and that better features are necessary to handle the most ambiguous cases. In this paper, the videos used for training and testing were constrained to videos where the entire slide is always present in every frame, and the view point and zoom are mostly stationary. However, the techniques used can be expanded to work on a larger class of videos.

Introduction

The Internet has allowed for the wide distribution of recordings of live slide presentations making the content available to a much wider audience. However, the online video watching experience often suffers because of the poor image quality resulting from high compression. The slides, which are often critical to the presentation, commonly become completely indecipherable. Authors are aware of the issue, and often provide the associated slides with video. The goal of this project is to come up with an automated process for aligning the slides to a video. Having proper alignments has the potential to improve the viewing the experience by providing the user with a synchronized display of the digital format of the slides while watching either in a separate window. This alignment also has implications for information retrieval allowing for videos to be internally indexed based on the features of slides displayed at that given point and time.

In this project, I attempt to align videos and their corresponding slides by training a classifier that as input, takes in a video frame / slide pair, and detects whether the pair is an actual match based on the matching key points identified by the SIFT algorithm. I leverage the assumption that presentations are mostly monotonic and do not often have very large jumps from slide to slide. From my training data, I construct a naive Markov process that models the probability of slide transitions from one frame to the next. I induce a probability from the classifier, and combine it with the Markov process, and use a modified version of the Forward Algorithm for determining the most likely path.   Using SIFT keypoints, allows the alignment to be robust to the variable renderings power point slides and partial occlusion of the slides (perhaps caused by the presenter). This method also avoids the problem of finding the bounding box which can be tricky.

Related Work

Work from Foote *et al.*[1]   [1], proposes an algorithm for identifying when periods in videos where there is a presentation given. They achieve 94% accuracy, but do not attempt to do the slide identification, and the model

[1] Foote, J., Boreczsky, J., and Wilcox, L. 1999. Finding presentations in recorded meetings using audio and video features. In*Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE international Conference - Volume 06*(March 15 - 19, 1999). ICASSP. IEEE Computer Society, Washington, DC, 3029-3032. DOI= http://dx.doi.org/10.1109/ICASSP.1999.757479

relies on acoustic data from the presentation. Girgensohn et. al. promote discrete cosine transform and Hadamard Transforms [2] to classify presentation in to certain categories (e.g. "graphics", "speaker", "both", "long" shot"). Both of these approaches are very successful at their task, but do not address this specific issue.

General approaches to image recognition (e.g. SIFT) apply naturally to this problem but do not fully leverage the special circumstances of this problem. The temporal information can provide a strong prior for a potential match. Most videos have proper orientation and slide image in the video generally retain the same aspect ratios as the original slide. Also per construction, I have limited the task to videos where every frame contains a slide. This constraint is mainly imposed by the limited training data that I was able to acquire.

Data

For my training data and test data I extracted used videos and power point presentations from Programming Systems Seminar Series '06/'07 from Intel Reearch.[3] Despite coming from the same event the distortion in the videos vary quite a bit. For training, I extracted frames from 2.65 hours of video, relating to 86 slides. The test data was from one hour long video corresponding to 46 slides. Given the time to process and annotate the data, I did not have enough time to generate more. The videos were in wmv3 yuv420p compression, 428x240 pixels , and had 15.00 frames per second. I used FFMpeg[4] to extract one frame every 10 seconds and convert it to 8-bit PGM format. I used OpenOffice[5] to convert the slides to PGM format. I used SIFT Keypoint detector to extract features from v4[6] the images.

From the keypoints extracted from the frames and slides, I used a modified form of the standard SIFT matching algorithm to derive matching points between every frame / slide pair. SIFT accepts matches that have a euclidean distance a certain threshold better than the second best match. By default, SIFT requires that the closest match must be less than .6 times the distance of the second best match. I increase this threshold to .8 to compensate for the heavily distorted videos. This averaged to 1.94 matches per frame slide pair in the training data and 1.7 matches in the test data.

This led to 38,000 training samples for the classifier.

Alignment

Video Slide Pair Classification

My binary classifier was built to determine whether a given frame / slide pair is actually a true alignment. I trained it on every frame slide pair for a given video. This data set is biased towards negative samples, as there are M-1 negative examples for evey 1 positive example, where M is the number of slides, given that every frame only matches one slide. To account, for this I reweighed the training data to balance out the training data (using the Weka library's Cost Sensitive Classifier). Without this rebalancing, the classifiers which I experimented with, simply classified every thing as not a match.

I aggregated the matching points between a given frame / slide pair and constructed a feature vector with the following features:

Feature Category	Features	Intuition

[2] Girgensohn, A., Foote, J. Video classification using transform coefficients. Proc. ICASSP '99, VI, pp.3045-3048.
[3] http://irbseminars.intel-research.net/
[4] http://ffmpeg.org/
[5] http://www.openoffice.org/
[6] http://people.cs.ubc.ca/~lowe/keypoints/

Distance	• Average Distance • # of matches with distance between a given interals	The matches with the lowest distance are more likely
Scale	• Average scale of keypoint in the frame • # of matches with frame keypoint scale between given set of intervals	I figured that given the compression artifacts some high scale features would be completely noise
Orientation	• Average square difference between the orientation of the keypoints in the match • # of matches with the squared differences between given intervals	The orientation of matching keypoints should be the same given that frames and slides are always upright.
Matches	• Total number of matches • Number of matches proporti	More matches more likely indicate a stronger correspondence
Temporal	• Percentage of video completed at this frame • Percentage of slides before and include	Since slide presentations tend to monotonically increase this could be a basic indicator

Markov Process:

Using the training data, I computed the probability for transitioning between any two slides based on position. I counted number of times a given transition occurred a with distance (in slides) and direction (forward or backward) and divided it by the total number of transitions. The test data contained 900 transitions. I used Laplace Smoothing to have at least some probability for every transition given the sparseness of data.

Let $S{:}\{s_1,s_2,..s_n\}$ be the set of all slides. Let $F{:}\{f_1,f_2,..f_m\}$ be the set of all frames. The probability of the optimal alignment (optimal path through the Markov process) P_{op} that has frame f_t containing s_x is upper bounded by:

$$P_{op}(f_t = s_x) \le \max_y \left(P_{op}(f_{t-1} = s_y) * P_{transition}(y \rightarrow x)\right) * P_{image}(f_t = s_x)$$

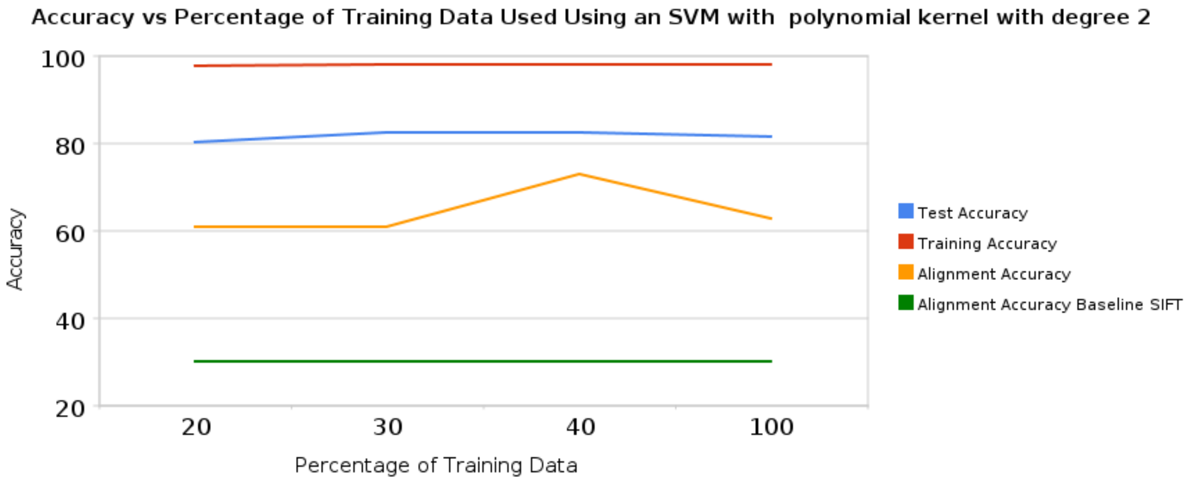
$P_{transition}$ is the transition probabilities that we computed from the transitions in our training data. P_{image} is the normalized probability induced from our classifier. The normalization is to ensure

$$Pimage(f_t) = \sum_{i=1}^{n} P_{image}(f_t = s_i) = 1$$

I then used the Forward algorithm to compute the most likely path, and use the path to derrive the alignment.

Experimentation Results

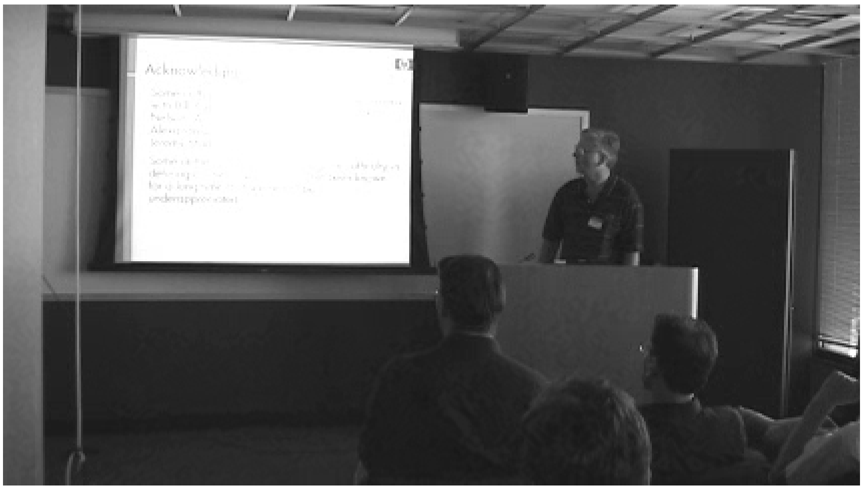
I first trained my classifier using the Weka library's SMO algorithm with a second degree polynomial. I used I then fit a logistic regression classifier to the resulting model to induce probabilities from the model. The baseline SIFT alignment, is computed for each frame by choosing the alignment for the slide that has the most matching points. This is not affected by the amount of training data.   This baseline is a little unfair, since it does not have my temporal model.



Analysis and Conclusions:

This result seems to indicate a combination of problems with this model. From a classification aspect, the fact that there is very low training error, and a bit higher test error seems to indicate that there is high variance, possibly indicating over fitting to my data. However, as the amount of data increased, there was no real improvement in this gap. This indicates that the training data needs to be much larger and more diverse to capture the variation in data. The video used for test set may just have some fundamental distortions that are unseen during training.

Looking over the misclassified frames in the alignmen and the corresponding keypoints it is apparent that there are certain regions of frames where SIFT is not sensitive enough. The ~60% accuracy is a big step up from the baseline, howerver for the alignment indicates that my objective function should include the final alignment. One reason for this could be that the penalty for a misclassification is not truly realized, given the magnified of affects of misclassification in the Markov process. Making the alignment the true objective function could force the classifier to give lower probability to its predictions when it is not completely confident.



A frame the method fails to align.