1. Introduction — semantic chunking and linking of educational video

Lecture materials are rapidly accumulating on the Web for student use. It is now common to provide video and audio records of academic classes and corporate training sessions in addition to providing traditional lecture notes and electronic slides. Some sites (e.g., [4]) provide links to a handful of video libraries in top schools. Others create and maintain lists of lectures in a specific area (e.g., [5, 6]). Of particular interest is the growing domain of distance learning. According to the latest survey [25] supported by the Alfred P. Sloan Foundation, over 4.6 million students (25.3% total enrollment) from over 2,500 colleges and universities were taking at least one online course during fall 2008 which is 16.9% growth over 3.9 million students (21.9%) in the Fall of 2007.

This trend suggests a future in which a much larger segment of people across the world will have access to excellent learning opportunities from on-line instructional video. However, taking full advantage of this data requires that the information within video is easily accessible. We are now accustomed to on-line search engines providing easy access to information on any topic available *within* text-oriented documents. However, there is no analogous way for students or other interested parties to find a video segment of a lecture on a particular topic or sub-topic. Lecture videos typically lack any textual description or metadata pointing to specific inner content. A typical video lecture may be listed as 'Introduction to CS - Lecture 5.' Without metadata, their content cannot be indexed and retrieved using popular video search engines that depend on surrounding textual meta-data provided by humans to index the videos.

E-learning has attracted much previous attention in multimedia research (e.g., [31, 67, 101, 131]). We characterize the state of affairs of current technology as having achieved some competence, **but definitely not mastery**, at providing seamless access to the rich content potentially available in presentation videos. Currently, searching for video on the web is largely limited to finding entire videos based on textual metadata that was created manually at some time and associated with the video. In particular, methods to access the inner content of video are largely limited to: 1) detailed and labor-intensive manual annotation; 2) making use of speech to text conversion of the audio track which is still error prone, and the stream of words is not structured as to importance (see, for example, [1, 26, 77] and the review [121]); 3) applying OCR to video frames which has similar problems; or 4) making browsing within video more efficient so that users can quickly exclude videos that are not of interest [1, 26], which is time consuming for the searcher and does not scale to finding content in large video collections.

We suggest that a simple idea can make a substantive difference to fine-grained video access. Electronic slides, which are typically used for instructional presentations, provide natural semantic handles into the video content. Slides can organize otherwise dense streams of content into semantic topics — chunks — which we posit are the right level of granularity as targets for searching and a base for further exploration. For example, in computer graphics courses, students learn about homogenous coordinates, a topic covered quite differently in math. We hypothesize that being able to align video chunks on a particular concept, across disciplinary boundaries, will provide students with two advantages. First, they can get to video on topics they want to study much more efficiently. Second, they can view alternative explanations and view points, which will promote comprehension. This is analogous to what many of us do with entries from resources like Wikipedia to learn more about concepts and topics we encounter while reading or studying. Ideally, video chunks will be available for similar use where one can easily jump around explanations of technical concepts, all with representations in multiple mode, including video that is particularly engaging for many learners.

How this vision can be achieved. Presentation slide use provides a segmentation of the lecture into semantic chunks. If the slide files are available, they make this segmentation easier, and provide for reliable extraction of index words, enhanced by their structured use (e.g., title, bullets). When slides are not available, word extraction will be less robust, but extracted words will still be suitable for our goals. Words extracted from the audio signal on a per-chunk basis can form part of the representation of the semantics of a chunk. Going further, the rich multimodal data provides further opportunities for characterizing the chunk such as laser pointer

usage, gestures, and speech expressiveness. In summary, analysis of video chunks defined by slide use can effectively open up educational video to fine grained browsing and searching.

In this project we propose to develop the algorithms and software to make this approach viable on a large scale. Specific technical contributions that will result include: 1) rapid and robust matching of video frames to presentation slides (§2.1); 2) extracting semantic chunks defined by slide use from videos where the presentation slides are not available in the original form (§2.2); 3) approaches for quickly and robustly identifying videos that use slides to support mining the web for video chunks defined by slide use (§3); 4) extracting semantics from video chunks (§4); and statistical modeling of semantic video chunks (§5).

To demonstrate the potential of our approach we will deploy a unified media web-portal for searching and browsing video chunks within a large body of instructional video (§6). We will use semantic web multimedia annotation technology to align extracted representations of component video and slide structure with existing open ontology standards. Search terms will point to chunks and retrieved results will be organized by clusters that can be further explored or excluded from consideration. Users will be able to select videos for viewing within an interface where the slides are synchronized with the video, and can be locally browsed based on their chunks. Further, these chunks will be linked to other chunks in other videos.

We will seed the demonstration web portal with hundreds of videos from dozens of classes and talks (§7). We will provide a simple web interface for **anyone to add videos** and slides. The system itself will be built from modular components whose **source code will be disseminated** to the research and educational community. Together with the sensitivity to standards already mentioned, others will be able to integrate, extend, and modify our work.

We will evaluate the efficacy of this new approach to structuring online educational data in authentic distance education delivery system (§8). In particular, we will study whether exposing the inner semantics of the videos is helpful and engaging for students in four on-line computer science classes compared with basic video access as is typically provided by on-line course management systems such as Desire2Learn (DTL) [2, 3].

Collaboration synergies. This project is thus a vertical integration of: 1) theoretical and algorithmic investigation; 2) deployment of a portal that demonstrates how the advances open up access to information within instructional video; and 3) studying how this new kind of access is used by students and whether it is effective. Ongoing collaboration across these three facets will lead to more focused algorithm and technology development, and unique opportunities for learning how students engage with the embodiment of these new ideas in on-line education delivery. Inter-twining the development of new approaches with the study of how they are used will thus make both activities more effective.

Broader impacts. This project will contribute to making instructional video content substantively more accessible. It will provide algorithmic advances to support the emergence of web portals where students and scholars can find multiple informative lecture segments on nearly any topic imaginable. The project will also demonstrate how statistical modeling can improve searching and browsing of the extracted information. Because the overall system will be developed in tight collaboration with an on-going on-line learning project, the approach will inform what works in real distance learning settings.

1.1. Improvements over our previous submission

We previously submitted a similar proposal to this venue. The panel rated the previous version as ‘competitive for funding’, and the reviews were very encouraging; hence this resubmission. One excellent suggestion was to increase the expected adoption of the system by incorporating standards and tools developed by the semantic web community. To help execute this suggestion, we have added co-PI Morrison who has significant experience in that area (as well as bringing other strengths to the team). We address improving the value of what we are doing by using ideas and standards from semantic web research in Section 6.3.

A second issue raised was some lack of clarity between what we have already done, and whether the challenge of what we propose to do was sufficient. We have substantially revised parts of the narrative with this comment in mind. Although we build on our existing work, we

cannot achieve our vision by simply applying what we have done so far. Specifically, while we can find slide images in good quality video captured under controlled conditions, we have also learned that good conditions are the exception. Second, the important case where the slides are not available remains undone. This case is difficult but addressable. Third, taking advantage of this particular kind of multi-modal data, including slide structure and speech, to guide the building of a large-scale information system, is uncharted terrain. Finally, the development of this kind of system in close proximity to educational users is clearly critical to validate our hypothesis that this concept will lead to better access to online presentations.

The reviews expressed concern as to whether speech recognition performance is sufficient for our needs. Speaker-independent speech recognition performance has increased significantly in the last decade, but it still has far to go. In Section 4 we clarify that we do not rely on speech transcripts alone, but we can exploit them when they are available. Further, in recent work we have shown that slides can be used to improve speech recognition [125].

1.2. Evaluation of the approach, the system, and its components

We will evaluate the efficacy of the general approach in authentic distance learning environments as developed in detail in Section 8. Further, algorithmic components will be evaluated throughout the project. PI Barnard has substantive experience evaluating computational methods and developing methodologies for doing so [33, 37, 41, 76, 120]. We argue that evaluation must be automated. It is not practical to evaluate the results of each experiment directly; rather, appropriate ground truth data must be built by human based annotation or other means so that evaluation can always be thorough. In the case of evaluating clustering and hierarchies whose correctness is defined by human use, this leads to challenges in defining the ground truth. We will rely on instrumenting the interface to collect data that serves as indicators of what students are finding useful. For key hypotheses, we may need to evaluate the results directly. Details that pertain to evaluating methods for each particular aim are discussed together with the proposed approaches to address them below.

2. Breaking instructional video into semantic chunks

Our premise is that slide use breaks instructional video into semantic chunks, even if the speaker does not follow the slide content in detail. Specifically, in this context, we define a semantic chunk to be the video segment where one slide is used. To extract these chunks we need to segment the video based on slide usage. Given a semantic chunk, we need to extract the semantics from it. For both segmentation and extraction, we consider two cases. **Case one**: If the presentation slides corresponding to the video are available in electronic format (e.g., PowerPoint, PDF handouts), then we can directly match video frames to the slides, and extract information from the slides, augmented by information from the speech transcript and video data (e.g., laser pointer use). **Case two**: If we only have the video, we will develop novel methods for segmenting out chunks defined by slide use and extracting information from those chunks.

2.1. Matching slide images to video frames (case one)

Previous work by others. Synchronizing slides with videos is recognized as an important capability. Several systems achieve this using a variety of means including manually editing time stamps [19, 112], methods based on recording time stamps during the presentation [21, 22], and simultaneously capture of slides and video [17, 18, 136, 137]. The limitations of these encumbered methods has led to attempts to automatically match slides with video content, generally by first extracting the slide region, and then identifying the particular slide. Several researchers have developed approaches to these two steps using a variety of image processing methods [66, 80, 100, 126].

Our matching approach. We have already developed a method for matching slides to video frames [68, 70] that combines image information and temporal information. Image similarity is based on scale invariant feature transformation (SIFT) keypoints [102] extracted from the video keyframes and slide images. These are matched under the assumption of a constrained homography [83] which is determined using random sample consensus (RANSAC) [74]. This method **simultaneously** identifies the slide and determines the geometric

mapping between the slide into the video frame. To improve the method in the case where the slide part of the video frame is small (zoomed out), we also match elements in the background to build a background model (Figure 1, next page), which then allows us to constrain where we look for the slide in the frame. We further reduce errors by integrating slide sequence statistics (e.g., advancing to the next slide is common, jumping to an out of order slide is rare), and camera change statistics (e.g., slide changes while zooming in are rare).

Taking all sources of information into account, we have achieved good accuracy (roughly 95%) [68] under good capture conditions. However, if the speaker uses lots of animation, stands in front of the slide image frequently, or the capture quality is poor, then the performance of the basic method suffers significantly. Further, ongoing data collection has informed us that studio quality capture is rare, and we observe that informal capture is becoming much more common. Hence we propose addressing the gap between ideal data and typical data.

We refer to our matching problem as "asymmetric" because we have detailed understanding of the slides (the model), but the target video frames are low resolution and have unknown and disparate noise processes which include: 1) out of focus imaging; 2) occlusion of the screen by the speaker, and screen motion when the speaker brushes against it; 3) lighting gradients and washed out areas out due to room lighting; and 4) incorrect color balance. On top of these problems, use of animation (e.g., one bullet is added at a time) means that temporally adjacent slides are often very similar.

A new method for asymmetric matching of slides to video frames. We propose a new approach as follows. First we establish the geometric mappings (homographies) between all video frames by aggregating homography information found by matching nearby frames. This can be done by linking SIFT features under consistent homography as we have done to construct the background model in previous work [68, 70] (Figure 1, next page). Of course, a dramatic slide change may prevent finding homographies solely by frame matching. However, we can assume that the camera operator does not change the camera settings (i.e., no pan or zoom) during a slide change. This assumption makes sense because the camera operator responds to the slide changes. Even if that is not strictly true, any reasonable panning or zooming rate will not change the homography much in the time it takes a slide to change. This means that when isolated successive frames cannot be matched under a homography constraint, we can assume that the homography is the identity mapping.

The next step is to apply our previous frame-slide matching approach to find some frame-slide matches. This only returns a match when the homography is very likely to be approximately correct. This leads to the following important observation. While the slides identified may be wrong, the homographies will generally be correct. We will then use these "anchor" homographies to determine all slide-frame homographies in the video.

At this point we know the geometric relationship between any hypothesized slide and any video frame. However, instead of matching slides directly, we will study matching pairwise differences between them. The intuition is simple. Consider two slides, a and b, which differ by an added bullet point. We can discriminate between them by focusing our attention on the added bullet point, whose frame location we can determine from the homographies established in the previous steps. We will develop methods to determine the relative probability that a frame, F, matches slide a versus slide b from such data. To identify the best match to F over all slides, we need to aggregate all the pairwise discriminative information.

One novel approach to do so is as follows. Let p_i be the sought after probability that slide i matches F, and let $\hat{p}_i$ be an estimate of it based only on i and F. For pairs of slides, a and b, let p_{ab} be the probability that slide a matches F, given that either slide a or slide b matches F, and let $\hat{p}_{ab}$ be the estimate of it based on the differences between slide a and b as motivated in the previous paragraph. Reversing the order of a and b (e.g., $\hat{p}_{ba}$) denotes the reverse. To link these estimates to our unknowns, p_a and p_b, we use $p_{ab} = p_a / (p_a + p_b) \approx \hat{p}_{ab}$, Rearranging, we

get $p_a\left(\hat{p}_{ab}-1\right)+p_b\hat{p}_{ab}\approx 0$, which are linear equations in the unknowns p_a and p_b. For each a and b we get two equations of this form because we estimate $\hat{p}_{ab}$ and $\hat{p}_{ba}$ independently. Finally, for all i, we have $0\le p_i\le 1$ and $\sum p_i = 1$. Thus we can turn the discriminative estimates into absolute probabilities by solving these equations in the least squares sense using quadratic programming. For stability, we need to introduce equations, $p_i \approx \lambda\hat{p}_i$, where λ is a positive smoothing constant that sets the balance between the two kinds of estimates, and ensures the equations have a solution. Additional accuracy should be achievable by weighting each equation based on error estimates of measured quantities.

Evaluating matching slides to frames. We have already built an efficient tool for human annotation of the segments of video that correspond to each slide (or no slide). This establishes a ground truth for automated evaluation of algorithms that match frames to slides. The tool makes creating ground truth for hundreds of videos practical. All ground truth data will be made available on-line.

2.2. Extracting slides directly from video (case two)

If a slide file is not available, then it will be necessary to extract the slides from the video. This is more challenging than the previous case because a representation of the visual characteristics of the slides must be learned during the analysis of the video. Despite the challenge, the benefit of doing so is great, and thus we will target this as a major enabling capability. If we are able to analyze even a modest subset of on-line educational videos, then we will be able to provide an incredibly rich educational resource.

We begin as in the previous case by determining the homographies between all the frames. From these we can construct a representation of all the areas that the camera looked at (Fig. 1). To simplify discussion, we assume that the speaker does not occlude the screen. Assuming that the slide boundary has been visible at some point, this representation encodes a static background area surrounding a slide area that can be mapped to a rectangle under a homography. Further, the slide area changes dramatically on occasion, but is typically relatively static. Together with appearance classification strategies, the slide should be identifiable in many videos. If the background is never visible, then the same procedure should work, except that only part of slide area might be accessible, and mapping the slide area to a rectangle to rectify it cannot be done at this stage.

Extracting slide image groups. Ignoring fast changes (slides shown briefly) and small changes (noise) should provide sets of images that correspond to a possible slide. These sets will be used to extract slide words and structure as described below. We process the group of images together because multiple images of each slide linked by a known transformation can be exploited to compensate for the low resolution and relatively high noise of video capture data.

Evaluation of extracting slides from video. We will evaluate our approach using the ground truth data mentioned above for case one (§2.1). This ground truth identifies which video frames belong together based on slide use. We will develop a measure for extraction efficiency based on this measure.

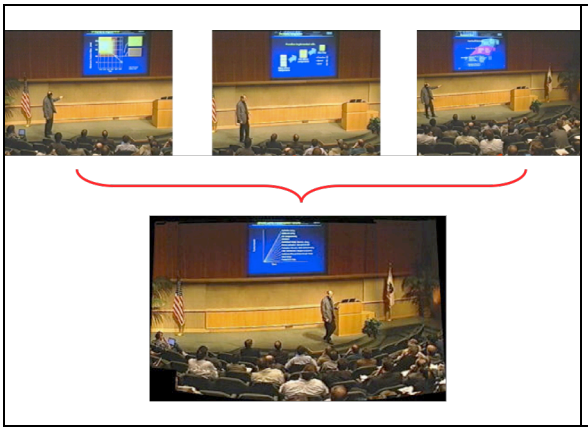

Figure 1. Illustration of how panning and zooming can both be accounted for and exploited. Here all the frames in the span of the three frames above are combined using AutoStitch [54] to form a model of the entire background scene, shown with a new slide below. In practice, this model is used implicitly —the form shown here for illustration. Notice that the bottom image has more of the scene than any of the frames that were used to construct it. Such linking, even when there is no background, will form the first step of both the new method for matching slides to frames (case one) and the proposed method to extract slides from video directly (case two).

3. Identifying presentations with significant slide use

We envision that capability to crawl the web for videos that make significant use of slides, and extracting the inner semantic content, thereby rapidly expanding the coverage of the repository. In this scenario, slides will typically not be available (i.e., case two will dominate), but slide decks that are hypothesized to be associated with video could be checked. Regardless, the first step in the processing of videos is to determine whether they make significant use of slides.

Notice that the process for extracting slides from videos will collaterally identify whether a video uses slides and whether information can be extracted from it on that assumption. However, this is an expensive way to identify videos of instructional content that use slides, and hence we will experiment with using simple heuristics (e.g., ones based on file names, domain names, and compressibility), and pattern classification methods to identify likely candidates quickly. Classifiers could be based on a spatiotemporal texture model. Classifiers can be trained on video that has associated slide files (e.g., case one), that has been matched, either algorithmically, or using the ground truth data set.

Slide analogs. Elements of our approach has potential for presentations that do not use standard slides (e.g., PowerPoint) but have similar icons that anchor content such as black board / white board use (erasing or switching boards changes slides) and flip charts. Although extracting semantics from hand written content is beyond the scope of the proposed work, semantic chunks from videos that we can segment based on slide analogs can be organized to some extent from the speech transcript.

4. Extracting semantics from video chunks

Interpreting slides used in video (case one). Having the slides used in a video is helpful because information can be more easily and reliably extracted from slide files than video frames. We have already developed good strategies for extracting words, their grouping into titles and bullet points, and their location in PowerPoint and PDF files. In what follows, we will assume that we have, at a minimum, the slide words, tagged as being part of the title or text blocks, numbered from top to bottom. To broaden search responses and to support clustering as described below (§5) we will experiment with mapping words to sense disambiguated WordNet [71, 72, 104] terms, and extending the vocabulary with synonyms and hypernyms as we have done in other work [35, 43].

Using speech. We will also extract text from the audio signal associated with the semantic chunk. Here there is no difference between the cases where slides are available and where they are not, other than the accuracy to which the boundary of the chunk can be determined, and we proceed similarly in both cases. As part of our collaboration with IBM we have had access to high quality speech recognition software (now available to researchers on a trial basis [15]). While state of the art speech-to-text still cannot produce human quality transcripts, they can extract words sufficiently well to support good search [77]. What the speaker says carries additional information not visible on slides, and is thus provides an alternative representation of the video chunks and what is useful and interesting about them.

Extracting semantics from video frames showing slides (case two). We will extract slide words in spatial groups such as title blocks, bullets, and other text blocks. To extract text, we will first experiment with making use of optical character recognition (OCR) technology. Wang et al. [132] have recently demonstrated that OCR can be successfully used for extracting text from lectures video. Preliminary work, using a readily available OCR package, FreeOCR, also suggests that this is a promising approach. We will test and compare FreeOCR with the open source OCR package, OCRopus [12], which deals with many of the intricate issues in OCR and document analysis, and it makes sense to take advantage of this and similar efforts. Given that the framework is open, small adjustments will be feasible if needed. We emphasize that reasonable OCR results are sufficient for the goals of this project.

Using the speech transcript to improve the OCR output. We have already developed a statistical method to align the speech transcript with slide words [125]. This provides better captioning for hearing impaired students. It also allows for correcting the transcript because the speech recognizer is likely to have difficulty with rare and technical words, precisely the ones

that are likely to occur on slides, and likely some of the more important ones to the transcript user. Here we propose that the same reasoning can be applied to help extract the slide words. When choosing among alternatives, we favor ones whose phoneme decomposition matches the phoneme string from the speech transcript. Further, if the OCR system (and speech recognizer) has tunable parameters, we could tune the tool based on maximizing the concordance of slide and transcript words.

Evaluation. We can construct a ground truth for slide words that should be found in video frames from the frame-slide matching ground truth data set because the words from the slides can be reliably extracted. Hence we can run the extraction process on frames without the benefit of the slides, and compare the result with this ground truth. Evaluating speech recognition will be based on ground truth transcripts.

5. Statistical clustering of semantic video chunks

The next challenge is to provide efficient access to an arbitrarily large number of video chunks and corresponding structured multimodal information. Since part of the information we extract is embodied as text, one might consider the problem solved by keyword search. However, this misses the point that users provide keywords to express semantic concepts. “Explanation of Dijkstra's algorithm” is a concept that is suggested by, but is not necessarily tied to, the particular keywords used in the example. Furthermore, what this concept relates to is even less obviously exposed by the words in the query. Hence, to make the semantic chunks that we have extracted fully accessible, we need to go beyond simple keyword search.

We will build a statistical model for semantic concepts that are associated with multiple video chunks. In the last decade, statistical approaches have been widely advocated for improving access to text data (e.g. [50, 61, 87-89, 110]), and PI Barnard has contributed significantly to this field in the case of multi-modal data [36, 38]. Statistical approaches are effective because they specifically account for variance among examples (e.g., a particular speaker’s word choice) that is not relevant to the topic at a particular level of granularity represented by the cluster. We will further use the power of these approaches to organize the clusters into a hierarchy. With respect to the project goals, the benefits of this general approach are:

- Queries are “soft,” being interpreted relative to semantics encoded in the clustering model.
- Search results can be organized by retrieved concepts, each represented by video chunk, allowing for simplified display and feedback from the user about which is relevant.
- The entire data set, or parts thereof, can be browsed based on hierarchical structure, which tends to be natural for human users, especially with very large data sets.

5.1. A statistical model for video semantic chunk clusters

We will develop generative statistical models for the data extracted from chunks (e.g., title block words, speech transcription term). For a given chunk, we consider the observed items as samples from a distribution whose parameters, θ, are associated with a cluster, *c*. We will study a number of representations of chunks and ways that they can be modeled statistically. For concreteness, to illustrate our general approach, we outline a very basic model.

Here, we represent the occurrence of words in the title block (T), bullet blocks (B), and the speech transcript (S), ignoring duplicated words. Then the observed data $O = \{T, B, S\}$, where $T = \{t_j\}$, $B = \{b_j\}$, and $S = \{s_j\}$ are binary vectors of length, V, the vocabulary size, indicating the presence of word j. The parameters $\theta = \{\theta_c^T, \theta_c^B, \theta_c^S\}$ are then vectors of the probabilities of occurrences of each word in each place for each cluster c. We assume that all observations are conditionally independent given the clusters, and thus, for a chunk, i,

$$p(O_i) = \sum_c p(c) \prod_j p(t_j|c) \prod_j p(b_j|c) \prod_j p(s_j|c) \quad (1)$$

This formula can easily be augmented to weight the three kinds of data differently. The parameters specifying the word probabilities can be learned while fitting the clustering model to data using EM (expectation maximization) [62] or MCMC (Markov chain Monte Carlo) sampling

[28, 99, 108]. We further would impose a smoothing prior so that each word has some probability of being in a cluster. We do not know in advance if this Bernoulli model for the clusters will be better than an approach based on multinomial distributions. We and others have successfully used both approaches for image annotation [36, 38, 63, 73, 91, 97, 98], and we will compare methods based on both of these building blocks for this application.

We expect good mileage from the general assumption that the chunk data can be broken into parts (e.g., as in $\{T, B, S\}$ above) that are conditionally independent given the cluster. However, we will also experiment with simple Markov models to determine if capturing some ordering of the text blocks is useful.

5.2. Learning a hierarchy for video semantic chunk clusters

We propose to develop generative statistical models that organize the video semantic chunk clusters into a hierarchy, so that they can be grouped for browsing and visualizing at various levels of detail. To add hierarchical structure to the clustering model just described (§5.1), we allow the distributions for observations (e.g., title words) to be composed from multiple latent "aspects" [90] that are shared among various sets of clusters. The sharing is arranged hierarchically as illustrated in Figure 2. The intuition behind the organization is that aspects that are shared among more video segments are semantically more general.

We define the mixing of the latent aspects, l, by $p(l \mid c)$, where only aspects that occur on the path from the cluster leaf node to the root node are considered. The model (1) becomes

$$p(O_i) = \sum_c p(c) \sum_l p(l|c) \prod_j p(t_j|l) \prod_j p(b_j|l) \prod_j p(s_j|l) \quad . \tag{2}$$

Notice that the clusters are multi-modal in that the distributions for the various components (e.g. title block words, transcript words) are linked at the aspect level, meaning that they will reinforce each other in defining semantic concepts for the chunk clusters. Given a fixed topology, we can learn the parameters using EM as we have done for images with associated text [36, 38, 63], or using MCMC sampling approaches as we have done learning structure from images [114-116].

Model selection. A research direction that we are particularly interested in pursuing is learning a good topology for the hierarchy from data. Alternative hypothesis can be compared using reversible jump MCMC [81, 82] as we have done in several other problems [114-116], provided that the posterior function sensibly penalizes over fitting. One approach for this is to use a model complexity penalty based on parameter counts such as AIC [24] or BIC [49, 118]. However, in our experience these are not very helpful for this kind of problem. Instead, we will investigate achieving model selection by considering the match between the probability distribution for the proposed model and the data. While the data must be relatively likely under the model for a good fit, choosing among models does not usually consider whether the proposed model tends to fit data that is very different from what is observed. We suggest that a good symptom of overfitting is the extent to which the model represents structure that is likely fictitious. Being careful with this criteria has been very helpful to us when fitting models to images [114], and here proposed fictitious structure is easier to identify. For clustering, this line of thinking is related to work on using Kullback-Leibler divergence to guide model selection while fitting Gaussian mixtures [135], but we are not aware of an applicable general approach.

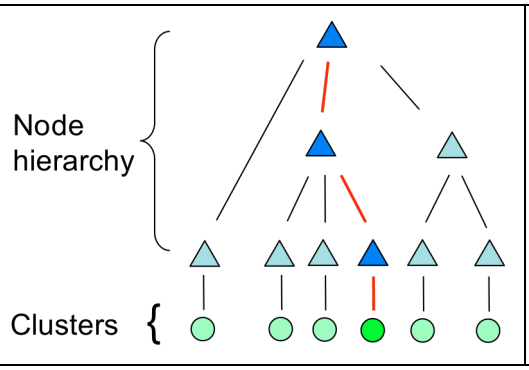

Figure 2. A statistical hierarchical model for video chunks. We visualize sharing of statistics through nodes (triangles) denoting aspects that contribute to the statistics. Clusters (circles) are associated with the path from the leaf to the root which defines the nodes whose parameters contribute to the cluster distributional parameters, as illustrated by the highlighted cluster (fourth from the left). Upper level nodes carry statistics about more common and more general words. The hierarchy can be rigid as we used for clustering images with text [36, 38], or more flexible as shown here.

5.3. Evaluation of the statistical models

The quality of clusters and hierarchies are difficult to evaluate directly because it is defined by the usefulness to users. For example, a better hierarchy may lead to user finding what they are after faster. We will evaluate the system with tasks such as these, which are very close to the goal of improving access to video data. However, as discussed above, we also need proxy measures that can be applied cheaply and continuously during development. For example, since our approach uses predictive models, we can compare alternative models on a large scale using cross validation across modalities. Taking this approach we can measure how data from some of the modes (e.g., speech transcript words) can predict other modes (e.g., title words) in data held out from training. Other proxy measures will be derived from recording system use. For example, if a link from one chunk (showing one solution to a problem) leads the user to another chunk (showing a different solution), and that the user views the second chunk for some, we might assume that the link was useful and valid and that the chunks should in fact be in the same clusters. Recording such use will provide data that can be used for proximal measures of cluster utility that can be run automatically. Of course, such proxy measures need to be checked against human performance more directly related to the goal of better access to video data.

5.4. Supporting searching and browsing

The statistical model for chunk semantics provides the basis for exposing chunk semantics to the user. Every chunk belongs to each cluster to a varying degree. If we refer a chunk as belonging to a cluster, we simply mean the one that it is maximally within. For chunks used to build the model the cluster weights are already known; for others they are easily computed. Keyword searches can now rank chunks, returning the top slice, based on how likely the words are to be associated with the chunks or their cluster membership. Such search is soft because not all keywords need to occur in the retrieved results. Search results can also be organized (hierarchically) by clusters instead of simply as a ranked list. Here each set of results grouped together by their cluster can be identified with the best chunk for the search. The group can then be expanded to reveal its contents under user control.

Similarly, the model can be used to define semantic similarity. A simple, computationally efficient approach is to compute the probability that the two chunks come from the same cluster. This does not distinguish between chunks that are essentially only in one cluster. Thus close matches will be refined using relatively standard text similarity across components, where we have additional information about the statistics of the words in that group, and can, for example, weight matches on rare words more strongly.

6. A Portal into instructional video content

To demonstrate, disseminate, and test our approach for making fine-grained, high-resolution semantic content of instructional video accessible, we will develop an on-line interface for easy navigation of the concept space of semantic video chunks through browsing and searching. We will integrate this into our existing system for browsing video chunks within a given video. We will also instrument the system for capturing details about how people use it, both for general data collection about system use, and for specific studies of its use as part of distance education delivery systems as detailed below (§8).

Browsing data from many courses and talks will be open to all. However we will support restricting particular content both on the basis of login credentials or IP address. This capability will encourage use by instructors who do not wish to make their materials broadly available at a particular time, and simplify tracking portal use for specific courses over a semester.

The system will be built from modular components whose source code will be disseminated to the research and educational community. We will adopt semantic web standards for representing the chunk structure and semantic topics extracted from and associated with the video and slide documents we process. This will allow us to use a variety of existing open source tools for supporting complex queries and browsing, and also make our data more accessible to other semantic web-enabled projects. Some details follow.

6.1. The SLIC System — Current State

A screen shot of our current system (http://slic.arizona.edu/) for browsing within a particular educational video is shown in Figure 3. Here the video viewer (left box) and large slide view (right box) are synchronized based on the current slide shown in the video. Smaller slide images along the bottom provide the users with ability to select a point of interest in the video by choosing a corresponding slide. Once a slide is selected, the video player plays the video portion that is associated with a selected slide. Similar synchronization is now also used by some on-line resources (e.g., [8]), but synchronization is not automated. In the SLIC system the user can also search within the video for all segments whose slide has query words.

Exploiting slide-frame homography. Our method for matching frames to slides collaterally determines the geometric transformation (homography) between the slide file and the slide in the video frame. Knowing this homography supports novel interface capabilities including: 1) mouse clicks on the video image can be translated into slide locations which we have exploited to provide a frame magnifier that uses slide image data; 2) slide images can be 'backprojected' into the video frames to improve video quality; 3) bandwidth can be reduced for mobile browsing [133]; and 4) laser tracks can be interpreted with respect to the slide image. However, further development of these techniques is beyond the scope of the current proposal.

6.2. Searching and browsing video semantic chunks.

Our goal is to make browsing and searching videos transparent within and across boundaries. We will begin by designing two integrated parallel interfaces, one that retains most of the ideas from the current system described above (*viewing*), and one that provides searching and browsing over the entire space of video semantic chunks based on clusters (*navigation*). Particular snippets will be easily selected for viewing in the video interface, either immediately, or to form part of a group, much like using an on-line shopping cart.

Viewing an individual chunk will translate into viewing the video containing the chunk using the current interface, starting at the chunk location. As the video containing the chunk is browsed, it will possible to grab other videos containing chunks related to, and thus linked to ones encountered while viewing. These grabbed videos can be navigated to directly, added to the viewing group, or used as a starting point in the navigation interface.

For viewing a group we will implement a novel idea where multiple videos are synchronized based on their closest semantic chunk to the one being directly viewed. This will add additional rows of slides at the bottom of interface shown in Figure 3. One row will clearly be the current video being viewed. This will allow an immediate visualization of alternative presentations of similar material. The navigation interface will be based upon the notion that a cluster can be represented by a canonical example or metadata extracted from cluster parameters. Thus a single conceptual unit can represent a large number of video chunks. Navigating among chunks will take advantage of the learned hierarchical structure described earlier (§5.2). In particular search results will be organized by cluster, and the clusters will be organized using the hierarchy. Conversely, browsing the results can restrict the search space. This kind of

Figure 3. Screen shot of the current state of the SLIC interface. On the left is a box into which the video is streamed. To the right is the slide file version (e.g., image extracted from PowerPoint or PDF) of the current slide in the video, automatically synchronized with the video and vice versa. The lower bar provides quick, slide-oriented access into the presentation. Nominally the miniature slide sequence is centered around the slide being addressed in the video, but the position is under user control through the slider at the bottom. Selecting one of the miniature slides jumps the upper displays to that point.

interaction between searching and browsing was found to be effective a decade ago in the scatter/gather work by Hearst and others [84, 85, 134].

6.3. Semantic web multimedia annotation

Up to this point, we have described our method for extracting chunks of video that are aligned with how slides are presented, and for building probabilistic models of semantic topics for clustering and hierarchical representation. In order to make the structure and content that we extract machine readable and available beyond our web portal, we will make use of established semantic web multimedia annotation tools. This provides two key advantages: (1) we can make use of existing semantic web search tools [13, 14, 86, 103] for rapidly prototyping different search and browsing methods, and (2) our extracted educational video database can be made available to the larger semantic web community, for example connecting it to the Linked Data cloud [9]. We will use the MPEG-7 [11] multimedia markup framework; this provides a rich language for describing audio-visual media, including annotation of regions within video frames and over time and describing compositional structure. This will serve as the base framework to anchor the outputs of our extraction algorithms in a consistent and flexible representation. However, as has been pointed out by a number of authors [30, 124, 129], MPEG-7 alone is often not sufficient to ensure interoperability of annotations. For this reason, we will also make use of the Core Ontology for Multimedia (COMM) [30, 75]. COMM was specifically designed to augment MPEG-7 and provide richer support for describing the semantics of multimedia annotations, ranging from media content and structure, to describing *how* annotations were extracted [30]. Finally, as we mentioned above, we will use the WordNet lexical database to link extracted text and topics to WordNet terms. We will associate WordNet annotations with the metadata of extracted video chunks, again making the results of the automated video curation process widely available.

7. Building the collection.

We will assimilate sufficient instructional video to achieve the following goals:

- Demonstrate and disseminate our approach and its benefits.
- Study the capability provided in both distance education delivery systems as well as a resource supplementing classroom attendance.
- Develop and test our models and algorithms on a sufficient scale to set the stage for deploying web portals to expose the semantics of the instructional video on the web.

We will provide an easy to use web interface for collaborating instructors and other contributors to place videos and slides (or links to them) into the pipeline for automated processing and integration into the system. For U. Arizona SISTA/CS classes, collection will be overseen by students working with co-PI Westbrook to put all our courses on-line, including the courses targeted for the study described later (§8). Sufficient video for the above goals will be available from existing UA SISTA/CS course video, and planned collection throughout the duration of the project. We estimate having videos for a minimum of 900 lectures assimilated by the time the study planned for Year 3 begins, providing the infrastructure for broad searching and browsing into multiple treatments of the same and related material.

The collection will be enriched with video from other sources such as the video of talks hosted by the iPlant collaborative [20], and on-line sites that target making instructional video publicly available (e.g., [4, 8]). Finally, dissemination of our approach will be served by encouraging instructors everywhere to contribute video. As already mentioned, providing simple access controls will remove one obstacle to participation. In addition, processed materials can be stored on our site, but do not need to be, mitigating some concerns about copyright. To further facilitate building the collection, we will implement simple scripts that can take a set of videos and slide decks found in a particular location (e.g., a collaborating instructor's web page) and match videos to slide decks automatically.

8. Evaluation of the system for on-line study

We hypothesize that providing students with clusters of instructional videos that can be easily browsed and accessed in meaningful chunks will be associated with increased engagement,

more extended viewing and, in turn, with better mastery of the course content. This claim reflects recent theories of how people learn, including the importance of encountering key principles, facts and relations in multiple contexts, and the need for users to monitor their evolving understanding and review material as needed to improve their comprehension [51, 122]. In addition, researchers argue that learning involves attentional and motivational processes as well as cognitive processes, and that instructional activities that provide the learner with control tend to be more motivating than highly directed activities [47, 59].

The evaluation plan includes both quantitative and qualitative approaches to assess the hypothesis. The evaluation test bed will include six classes from four University of Arizona Computer Science courses: Introduction to Computer Science (250 students), Program Design and Development (100 students), Computer Organization (100 students), Programming Languages (100 students). The courses are integrated with Desire2Learn (D2L), a popular course management system that allows students to view videos of lectures. Students will be randomly assigned to either the traditional D2L site, which will allow students to view the instructor video, or a parallel site enhanced with the SLIC portal. The total sample will be an estimated 550 students, with 275 students per condition.

Students will be asked at initial log in to consent to their course activity data being used for the evaluation, as well as access to grades in prerequisite courses. Students who consent will be assigned a unique identifier that will be used to organize data from multiple sources. Identifiers will be linked to names in a secured file stored separately, with the link being discarded after all data are collected and integrated and the analyses have been completed. All recruitment and consent materials will be reviewed and approved by the U. Arizona's Institutional Review Board.

The SLIC and D2L course portals will be instrumented to record students' viewing of the available video clips, including information about the day and time that the videos were accessed, the time the video was played (as indicated by starting and then either reaching the end, actively stopping, or a click on another video or area of the portal), and the sequence of choices made by the user in that session (e.g., viewing the course instructor's video for the topic, and then moving on to a segment from another course, or perhaps going back to a video on a previous topic). Dr. Beal has extensive experience with the design of systems for collecting such user interaction data in the context of tutoring systems and discussion boards for math, science and engineering courses, and with data mining approaches to the evaluation of instructional systems [45, 46, 57, 93]. The user interaction data will be queried to address questions in two primary areas:

First, we will investigate temporal patterns in the use of video during an instructional period such as a course term: Are viewing patterns related to course events such as topic changes and exams? Do viewing patterns differ for students in D2L and SLIC groups, and by course topic? If the SLIC portal offers the hypothesized benefits, students' use of video should be more frequent and more sustained than for students who use only the intact instructor video accessible via D2L.

Second, we will investigate the relations, if any, between students' use of video resources accessible only via D2L versus through SLIC and performance in the course. One of the project claims is that access to semantically meaningful groups of video segments will engage learners, allow them to locate video resources that will improve their understanding of the course material, and in turn have demonstrable benefits for learning outcomes. If this hypothesis is accurate, then students in the SLIC-enhanced condition should not only show behavior suggestive of greater video viewing for learning but also higher course performance than those who can only view intact video through the relatively cumbersome traditional course management system. We will establish that students in the SLIC and D2L groups are comparable at the start of the term by examining grades in prerequisite courses. Learning outcome measures will include performance on exams and final course grades. Although our primary hypothesis is that enhanced accessibility to rich video resources will be linked to stronger learning outcomes, an alternative hypothesis is that the higher degree of choice involved with the SLIC interface might be distracting to students with weaker skills. One criticism

of distance learning and web-based activities is that students can become “lost” in a seemingly endless web of materials, and lose track of their learning objective [56]. Thus, it is at least plausible that access to rich sets of video may not necessarily be helpful for some students. Comparing the D2L and SLIC groups will provide some initial evidence to distinguish these alternative possibilities.

In addition to obtaining quantitative data to evaluate the impact of the SLIC portal on student performance, we will also solicit students' perceptions of the video resources through surveys about their experiences in the courses. These surveys will include multiple items to assess students' satisfaction with the video resources available in the course, the ease of access to these resources, technical issues encountered, the extent to which they attempted to locate video resources from locations outside the course web tree, their awareness that different students had access to different video resources and any attempts to access materials across condition, and their perception of the value of the video resources for their understanding of the course content. Likert-type ratings will serve as the primary response mode, with the addition of optional open-ended comment response boxes [47]. Students in each course will be encouraged to complete the survey through a lottery for a modest prize such as a gift certificate to the campus bookstore. Responses will be summed and averaged for the target constructs (satisfaction with resources; perceived value; technical issues; ease of access and use; etc.) and then compared to the D2L and SLIC groups. In addition, students' patterns of access to video resources during the term will be examined as predictors of satisfaction and perceived value. In particular, we want to establish that students who report high or low satisfaction with video resources for learning actually used the resources during the term.

The results from the evaluation will significantly extend our understanding of how students use video resources, and whether providing students with rich, semantically organized sets of video materials that are designed for easy access will be associated with improved learning outcomes and higher levels of student satisfaction with digital resources for learning.

9. Broader Impacts

This project will produce a robust approach to extracting fine grained semantics from presentation videos where slides are used. The methods that we will develop will set the stage for building an organized repository of video chunks for a substantive fraction of instructional video available on the web. We are now accustomed to easily finding textural information on any topic, but the same does not apply to instructional video. Hence this project has great potential for transformative impact by providing this capability to instructional video. Further, this project will produce an approach for integrated searching and browsing of such a repository based on statistically models of the extracted fine-grained semantic chunks. These contributions will enable significant leveraging of this rich information source to the benefit of distance learners and anyone wanting to find informative video on any topic.

9.1. Research student training

The bulk of the requested funds is to support students, and their training will be a significant consequence of the proposed activities. Currently we are a diverse group that includes two female grad students in the computational sciences. Further, recruitment of additional students will carefully consider under-represented groups in the computational sciences.

9.1.1. Increasing undergraduate participation in research

Several excellent undergraduates including Andrew Winslow, Juhani Torkkola, Michael Thompson, and Daniel Mathis have made substantive contributions to the project so far, and we are committed to maintaining and extending the inclusion of undergraduates. Hence, we will seek REU supplemental funding to increase our ability to involve undergraduates in future work.

Prior results. The team has a strong record in mentoring undergraduate research students. In the last four years, the PI's have mentored many undergraduate research students that have contributed to a number of publications [39, 43, 44, 65, 116, 133] and on-line demos [27, 123], and who have received a number of honors including finalist in the North America wide Computing Research Association (CRA) undergraduate research award (Kate Taralova), a U.A.

College of Science outstanding senior (Kate Taralova), two departmental outstanding seniors (Kate Taralova and Andrew Winslow), and a Galileo Circle Scholar [16] (Matt Johnson).

Undergraduate researchers working with the project are, and will continue to be, integrated into all lab activities, and have desks in the same area as the graduate students. Collectively, such measures have been found to make a substantive difference by faculty in our department [58] and elsewhere [48, 78, 79, 127], and are consistent with multiple recommendations for increasing participation by women [60, 96, 109].

Evaluation. Indicators of success include 1) undergraduates as authors of publications; 2) acceptances into prestigious graduate programs; and 3) job offers where research experience can often be identified as a factor.

9.2. Dissemination

This project will lead to many conference presentations and journal papers. We are also committed to broadly disseminating other research products. PI Barnard has a strong track record in the distribution of carefully calibrated data sets [32, 37, 42, 92, 119] and providing open source software together with input files which reproduce results in journal papers [40]. We will continue to share in this project. We will put demonstrations on-line as they are developed, make data available where permitted, and release software source code for academic use. We will make it possible for others to push data through our pipeline so that they are automatically integrated into the system. The output of our system will include a packaging of video and slide documents in well-established standards for multimedia annotation, including MPEG-7 and use of the COMM ontology. We will also provide WordNet annotations to the extracted text. While we will document and make accessible a web-based portal to submit to and search within the database, the data itself will also be accessible and thus available for other semantic web-enabled tools.

10. Schedule of Work

The following plan of work is organized as main tasks for students focused on 1) the computer vision aspects of the project (S1); 2) statistical modeling of multimedia data (S2); and 3) interfaces for browsing and searching large collections and their applications to distance learning (S3). See the Collaboration Plan (§12) for further information on project management.

Year one	S1	1) Implement and test the novel method for matching frames to frames (§2.1); 2) Link frames in videos without using slides to establish homography changes and implement filtering strategies to deal with noise and occlusion of the screen by the speaker (§2.2).
	S2	1) Establish robust extraction of slide structure information based on preliminary work (§4); 2) Initial (non-hierarchical) clustering experiments for video chunks (§5.1).
	S3	1) Design and implement video and slide drop site to automate collection building, and establish a diverse seed collection (§8); 2) Initial design of video chunk browsing and searching interface (§6.2); 3) Incorporate MPEG-7 and COMM annotation standards
Year two	S1	1) Develop the approach to extract video chunks based on slide use where we do not have the slide deck (§2.2); 2) Develop the approach to extract text from chunks using OCR (§4).
	S2	1) Develop statistical models for semantic video chunks with a fixed hierarchy (§5.2); 2) Develop new approaches for learning hierarchies (§5.2).
	S3	1) Complete video chunk browsing and searching interface (§6.2); 2) Initial usability testing.
Year three	S1	1) Complete handling of videos using slides where we do not have the slide deck (no planned year three testing depends on this activity (§2.2); 2) Develop heuristics for establishing if videos might use slides; 3) Validate video analysis software and documentation thereof for dissemination.
	S2	1) Finish model selection work from year two (§6.2); 2) Finalize interface with semantic web (§6.3); 3) Validate clustering software and documentation thereof for dissemination.
	S3	1) Fall term. Conduct an evaluation of the technology with students in courses who use the interface to find and view clusters of video snippets, recording detailed data about how students use the technology (§8); 2) Spring term. Analyze the data collected in the fall, and replicate the data collection; 3) Summer. Complete analysis and prepare reports.

11. Results From Prior N.S.F. Support (last 5 years)

PI Barnard is funded by CAREER grant #0747511, Learning Models for Object Structure. Here we developed a method for learning the category topology of furniture objects from single view images of multiple examples [114] (2nd paper under review), and using such models to help infer scene structure and robots learning about objects (papers under review). This award has contributed greatly to the PHD awarded to Joseph Schlecht, and is now helping support new PHD students Jinyan Guan and Luc del Pero. This award also supports our Integration of Science and Computing (ISC) Summer Camp for middle school students [7]. Barnard is also co-PI on IOS# 0723421, Characterization of Pollen Tube Repulsion in *Arabidopsis thaliana*, lead Ravi Palanivelu. Here we identified the gene (LORELEI, AT4g26466) that was defective in a mutant we previously isolated in a screen [128]. LORELEI encodes a putative GPI-anchor containing protein with unknown function. We demonstrated that LORELEI has a role in pollen tube repulsion, pollen tube reception, double fertilization and early seed development. Barnard's role is the development of statistical models for tracking multiple pollen tubes in vitro, and for pollen tube / ovule interaction (two manuscripts on-line [52] [53]). PI Barnard serves as a faculty advisor for the NSF funded iPlant Collaborative [20]. **Barnard and co-PI Efrat** have also worked on a subcontract to an NSF grant for the Large Synoptic Survey Telescope project (LSST) [10] (AST #0551161; PI William Smith) on tracking asteroids [34, 95, 111].

Co-PI Efrat is funded by CAREER grant #0348000, Pattern Matching, Realistic Input Models and Sensor Placement, Useful Algorithms in Computational Geometry, and by recently awarded CNS #1017115, TC: Small: Collaborative Research: Protecting Networks from Large-Scale Physical Attacks and Disasters. Most relevant to this proposal, his CAREER grant has contributed to several SLIC project publications. Support for undergraduate Andrew Winslow contributed to showing how the SLIC system can be exploited by users of hand held devices [133]. Support for undergraduate Michael Thompson contributed using slide words to improve speech transcripts [125]. Finally, summer support for Efrat contributed to [64]. The two grants listed above have also lead to several recent publications dealing with different network problems [23, 29, 113]. Funding from these two grants has supported the work of three PhD students (Jesus Arango, Swaminathan Sankararaman, and Javad Taheri).

Co-PI Beal has directed several NSF projects focusing on curriculum-aligned online learning for STEM topics. The AnimalWatch algebra readiness project has been funded by the National Science Foundation (9555737, 9714757,0725917, 0903441) to support participation by students from groups traditionally under represented in STEM, including females, English Learners, and students with visual impairments, and is now supported by the Institute of Education Sciences (R305K05086; R305K090197). Dr. Beal also directed the Wayang Outpost project, which provided online geometry tutoring with the goal of improving participation for students from groups traditionally under represented in science (0120809; 0429125, 0411886). Dr. Beal also participated in a project to increase retention of first year engineering students through enhancements to online course discussion tools and resources (0618859).

Co-PI Westbrook is PI on an NSF CPATH II grant, Computational Thinking as a Foundation for Interdisciplinary Undergraduate Education (CNS-0938763, $800,001) for the development of interdisciplinary classes for the School of Information: Science, Technology, and Arts at the University of Arizona. From this work, seven new undergraduate courses have been developed and at least four more are in planning for the next year. Dr. Westbrook also was Co-PI on an NSF funded grant to support a workshop for K-12 computing teachers: ABI and CSTA Collaboration to Reach K-12 Teachers from Under-represented Communities at the 2009 Grace Hopper Celebration (DRL-0936706, $79,899, 8/15/2009 – 7/31/2010). This event served approximately 100 K-12 teachers from across the United States. Dr. Westbrook has also served as a faculty advisor for the NSF funded iPlant Collaborative [20], working on activities integrating computational thinking in plant biology education and research (K12 and university).

Co-PI Morrison has not been supported by NSF grants in the past 5 years. His funding in the past 5 years has been from DARPA, ONR, AFOSR, and DCIA.

12. Collaboration Plan.

All participants are at the University of Arizona and have a history of working together. PI Barnard and co-PI Efrat have a six year history of fruitful collaboration with a number of joint publications [34, 68-70, 94, 95, 111, 133]. Barnard and co-PI Westbrook have collaborated on iPlant [20] management. Westbrook has collaborated for a number of years with Efrat and Barnard on the U. Arizona computer science curriculum. PI Barnard and co-PI Morrison have also collaborated on computer science curriculum and have had ongoing research discussions for several years. Co-PI Morrison and co-PI Beal have collaborated on a number of projects, most recently B2E2 [105, 130], which bridges computer science, biology and education

All PI/co-PIs will work together to manage the project, with PI Barnard having final responsibility for ensuring that the project goals are being met according to the timeline set in the proposal (§10). In addition, we have budgeted sufficient salary to allow co-PI Morrison to take a very active role in project software development management. Morrison has extensive experience doing so on several large DARPA-funded team projects that included managing teams of graduate students and programmers, the development of software modules and integration with large software systems, as well as technical contributions and reporting. The PI/co-PIs will lead the intellectual efforts in their own domains, jointly where they overlap. The roles of the key personnel follow.

12.1. Participants and their roles.

Kobus Barnard, associate professor, computer science and ECE. PI Barnard will have primary responsibility for ensuring that the schedule is being met, and working with the co-PIs to make any needed adjustments. He will implement monthly meetings with key personnel (bi-weekly in year one), and weekly project meetings where students will join. He will organize additional meetings as needed. He will ensure that meeting agendas are distributed, and that meeting minutes are made available on our project Wiki. He will also ensure that activities that involve students working on different aspects of the project remain coordinated. He will work closely with co-PI Morrison on managing the software development associated with the video processing pipeline and the portal. In addition, he will work with co-PI Westbrook to ensure that video data and classroom use is proceeding smoothly.

Barnard will take the lead for the algorithm and software development pertaining to video processing and semantic chunk clustering. He will supervise students working on these topics, jointly with co-PI Efrat as appropriate. He will also co-supervise a student, designated as S3 in the schedule of work, likely to be new PHD student Yekaterina Kharitonova, who will work on the interface between the system and students on other users, instrument the system to collect data on how is used, and analyzing and interpreting this data, together with co-PI Beal.

Clayton Morrison, assistant research professor, computer science. Co-PI Morrison will take the lead role in managing the software life cycle for this project as he has successfully done in other projects, such as the DARPA Evidence Extraction and Link Discovery, Integrated Learning, and Bootstrapped Learning projects. Most critically, he will ensure that the system makes best use of existing semantic web multimedia representation standards such as combining MPEG-7 metadata annotation with the COMM multimedia ontology. He also has significant experience using ontologies and semantic web technologies from past projects, including the DARPA Integrated Learning project [55], which developed a system for learning to compose semantic web services based on expert demonstration [106]. Co-PI Morrison will also (co)supervise students working on the extraction of semantics from video chunks.

Alon Efrat, associate professor, computer science. Co-PI Efrat will work with co-PI Barnard on the development of video processing algorithms and clustering models, continuing a long-term collaboration already in place in theses areas. He will jointly supervise students working in this area as appropriate. In particular, Efrat is the primary advisor for PHD student Qiyam Tung who has already contributed to the preliminary work, and who is a likely student to recruit into this project.

Carole Beal, professor, cognitive science and computer science. Co-PI Beal will also contribute to project management. In Years 1 and 2 she will be responsible for the development and implementation of the database and user tracking tools that will be required in Year 3 to obtain data for the evaluation. In Year 3, she will also take primary responsibility for data extraction and analysis as part of the evaluation study.

Suzanne Westbrook, associate director of SISTA[1]. Co-PI Westbrook will form a bridge between this project and a significant funded effort to put U. Arizona SISTA and CS classes on-line. She will work with PIs Barnard and Morrison to ensure that videos from these classes are being assimilated into the system. She will also help with communication between the project and students and instructors using the portal for these classes, helping to ensure that possible opportunities for feedback are not missed. Finally, for the third year, Westbrook will play a key role in working with the instructors of the four CS classes that will be used for the on-line learner study to help recruit participants, resolve data collection problems, and gather feedback.

Arnon Amir, senior scientist, IBM Almaden. Arnon Amir has collaborated with us on this research direction from the very beginning, and he will continue to do so in this project. Arnon provides a wealth of knowledge and experience on making multimedia data accessible.

Sandiway Fong, associate professor of linguistics and computer science. Sandiway Fong collaborates with the project on issues involving speech and language (e.g., [125]).

12.2. Managing the project.

Because all personnel (other than collaborator Amir) are at the University of Arizona, managing the project is simplified. The faculty and students that have worked together on previous components of the SLIC project are already used to meeting weekly to review project progress and discuss how to approach the scientific problems. In this project, all project personnel will again participate in weekly meetings. Project meetings will have agendas prepared in advance and minutes will be recorded and posted on the project Wiki. Access through Skype and/or telecom will be available for project members that are traveling. Throughout the project duration we expect that our industry collaborator, Arnon Amir, will regularly join us remotely as he does currently.

We will also have monthly meetings (bi-weekly for the first year) for key personnel with the specific purpose of reviewing project planning and progress. Agendas and minutes for these meetings will be circulated through E-mail and archived appropriately.

12.3. Multi-disciplinary issues.

Compared with our previous work that has set the stage for this project, this work will be even more interdisciplinary. The PI and co-PIs all have substantive experience with interdisciplinary projects. Barnard and Efrat have contributed to the LSST project [34, 95, 111] and analyzing images of brain neurons [94, 107]. Barnard has worked on modeling biological structure [115-117] and is a faculty adviser for iPlant [20]. Beal has experience directing interdisciplinary projects that bridge engineering, computer science and education, as well as with both quantitative and qualitative approaches to project evaluation. Co-Pi Westbrook is also a faculty adviser for iPlant [20]. Co-PI Morrison also has experience leading interdisciplinary projects, such as B2E2 project [105, 130], which bridges computer science, biology and education.

From this experience we know the need for --- and the value of --- informing each other about the language and science of what is happening in the project parts outside our range of expertise. Importantly, this needs to occur at the right level, with clear emphasis on non-technical understanding and resolving issues with domain specific terminology, and domain specific use of terminology. Hence we will reserve one meeting per month for presentations by faculty and students on the science behind the components that they are working on, with a key specification being that a significant portion of the presentation is accessible across disciplines. Having students participate in this will be an educational benefit to them.

[1] SISTA is the University of Arizona School of Information Science, Technology, and Arts. At the University of Arizona, computer science is within SISTA.