Multimodal Alignment of Scholarly Documents and Their Presentations

Bamdad Bahrani
Department of Computer Science
National University of Singapore
bamdad.bahrani@nus.edu.sg

Min-Yen Kan[1,2]
[1]Department of Computer Science
[2]NUS Interactive and Digital Media Institute
National University of Singapore
kanmy@comp.nus.edu.sg

ABSTRACT

We present a multimodal system for aligning scholarly documents to corresponding presentations in a fine-grained manner (i.e., per presentation slide and per paper section). Our method improves upon a state-of-the-art baseline that employs only textual similarity. Based on an analysis of baseline errors, we propose a three-pronged alignment system that combines textual, image, and ordering information to establish alignment. Our results show a statistically significant improvement of 25%, confirming the importance of visual content in improving alignment accuracy.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

Keywords

Digital library, fine-grained document alignment, slide presentation, slide image classification

1. INTRODUCTION

Scholars use publications to disseminate scientific results. In many fields, scholars also congregate at annual congresses to narrate their scientific discoveries through presentations. These two vehicles that document scientific findings are interesting in their complementarity; while they overlap in content, presentations are often aimed at an introductory level and may motivate one to take up the details in the more complete publication format.

As the presentation is often more visual and narrated by an expert, often it can be regarded as a summary of the salient points of a work, taken from the vantage point of the presenter. By itself, presentations may fufill information needs that do not require in-depth details or call for a non-technical perspective of the work (for laymen, as opposed to subject matter experts). In such cases, a useful function would be to link and present the vehicles in a fine-grained manner that would allow seamless navigation between both forms.

We follow the previous characterization of the *document-to-presentation alignment* problem [7]:

Given:	Presentation S: Slides $s_{1,\ldots,n}$
	Document D: Text units $d_{1,\ldots,m}$
Output:	Alignment $f(s) = (x, y)$

which maps each slide s_i to a contiguous set of document paragraphs d_x to d_y, or to *nil* (unaligned).

Previous work up to now has maintained a text matching approach to this task. However, some studies suggest that different methods for measuring the similarity between short segments of text (i.e search queries and tags) [9, 12]. However, our input data is neither as short as mentioned studies, nor long as usual text similarity studies. Therefore we use weighted TF*IDF for similarity measure as an baseline and then improve it from different aspect to gain better results. Hayama *et al.* first tackled this problem with Japanese technical papers and presentation sheets using a Hidden Markov Model [5]. In Kan's [7] SlideSeer digital library, the scope of the alignment work was enlarged to include the crawling of document-presentation pairs and the bi-modal (presentation- or document-centric) user interface. His work also recognized that certain slides should not be aligned – termed as *nil* alignment – when new material is introduced in the presentation that is not present in the document. More recently, Beamer and Girju [1] performed a detailed analysis of different similarity metrics' fitness for the alignment. All mentioned studies was not able not achieve high accuracy in their results. One thing that has always been missing in this task is taking advantage of visual content of the slides.

In related but separate vein of work, a few studies show the importance of proper slide structure identification: *i.e.*, differentiation between presentation body and title text, identification of graphical elements such as figures, charts and plots. Such structure is leveraged in downstream applications: slide reuse [6], retrieval [4, 8], and presentation generation from documents [10, 11]. In contrast, there has been minimal work to incorporate this information for document-presentation alignment.

We contribute to the state-of-the-art by addressing this weakness. Our system builds from existing text similarity baselines [7, 1], exploiting graphical information to specifically correct weaknesses of the text-only alignment when dealing with certain classes of presentation slides.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

Table 1: Demographics from Ephraim's 20-pair dataset [3].

Total # of slides	751
Average # of slides per presentation	37.5
Total # of sections	515
Average # of sections per document	25.75

2. CORPUS STUDY OF PRESENTATIONS

We take the publicly available document-presentation pair corpus from [3] as a starting point (Table 1). Their dataset consists of 20 pairs, drawn from papers in DBLP[1] (in .pdf) on databases and information retrieval, where an author-authored presentation in Microsoft Powerpoint format (.ppt) was found. The dataset is annotated with ground truth alignments, including non-alignable slides (*nil*).

To understand the weakness of previous work, we implemented a basic, text-only alignment system informed by the previous work. Employing standard textual similarity (cosine similarity with $tf.idf$ weighting), we aligned the sections of the document to each slide to observe its performance.

Taking a slide-centric approach to analysis, we observe several classes of slides, detailed below. Interestingly, we also find that the performance of text-only alignment also varies per slide class.

- **Text** slides usually form the bulk of content in presentations. These are often bulleted points, distilling the content from the paper into a pithy form.
- *Nil.* These can be title, example, or ending slides (*i.e.*, Q&A, references) or any other content not directly extracted from the paper. The previous work reports that classifying such slides correctly may improve alignment performance anywhere from 3 to 25% [1, 7].
- **Outline**. These are an important sub-class of *nil* slides, that we have separated from the main class. These slides exist solely to present or recap the presentation structure, to help sync the audience to the material being presented.
- **Image** slides consist almost solely of image(s). These are challenging to align for the baseline, since there is little or no textual evidence for alignment.
- **Table** slides are self-explanatory. We note that text extraction often functions to extract the textual strings within the table, which when extracted verbatim from the document, constitute strong evidence for textual alignment.
- **Drawing.** These slides consist of drawing elements: simple shapes, arrows, graphs and text boxes, authored within the presentation software.
- **Continued.** These slides continue information from the previous slide, used when the content overflows from the previous slide. These slides sometimes have textual cues ("cont'd"), and slides identified as such should be aligned as a block with the initial slide.

[1] `http://www.informatik.uni-trier.de/~ley/db/`

Table 2 gives the distribution of these slide classes in terms of prevalence in the dataset by presence/absence in the presentations, and by raw number. We also indicate error rate (raw and percentage), in the last column. Note that we omit **Text** slides that are well-suited to alignment by textual similarity. We see that both *nil* and **Image** classes are found in the large majority of presentations, and also constitute a large number of errors in the baseline. These are the classes of errors we target to ameliorate by our multimodal technique.

Table 2: Analyses of the dataset. Frequency and baseline alignment performance shown by slide class.

Slide Type	Present in # of presentations (out of 20)	# of slides (out of 751)	# (%age) of incorrect alignments
nil	19 (95%)	128 (17%)	83 (64%)
Outline	8 (40%)	36 (4.8%)	13 (36%)
Image	19 (95%)	90 (12%)	73 (81%)
Table	5 (25%)	8 (1%)	4 (50%)
Drawing	12 (60%)	65 (8.7%)	35 (53%)
Continued	9 (45%)	61 (8.1%)	24 (39%)

3. METHOD

Our automated multimodal alignment system presupposes extracted text from both the document and the presentation. As text extraction from documents is a noisy process (even for those born digitally, as in our dataset), we spent a fair amount of work in creating a pipeline to engineer relatively clean output data. For documents, we leverage the PDFX package[2], which outputs an XML format that largely preserves the paper's title and text, recognizing sections, figure and table captions. For presentations, we used custom Visual Basic code to extract its title and body content.

Given this pre-processing, our system then exploits the textual content and slide images to perform the alignment, as in Figure 1. We architect our system to consider textual similarity evidence and natural, linear ordering as probabilistic preferences in the alignment process, building a separate module for each of these two analyses. Given a slide s_j, the two modules output a vector v_i of length $|D|$ that represents the probability of aligning the slide to the particular document section d_i. In the final fusion phase (marked as $\otimes$), we fuse these vectors into the final alignment through heuristic rules that consider the visual slide image. If no module yields strong evidence for alignment (*i.e.*, low probability), then the slide is deemed *nil*.

Text Similarity. This serves as the primary basis to align **Text** slides. We compute a cosine similarity between the two extracted texts (slide and section), using $tf.idf$ weighting, as is recommended by the prior work [1, 7]. These similarity scores are summed to normalize the vector v_{Ti} to unity. We use this text similarity module alone as the baseline for comparison in our evaluation.

Linear Ordering. Studying our corpus, we find that most pairs' show that the ordering between slides and sections are monotonic. We thus output an alignment probability vector v_{Oi} that gives the linear mapping of M sections to N slides (i.e., for slide number i: $\lfloor \frac{i}{\lceil N/M \rceil} \rfloor$) the highest probability, and close neighbors a smaller probability. Doc-

[2] Available at `http://pdfx.cs.man.ac.uk/`

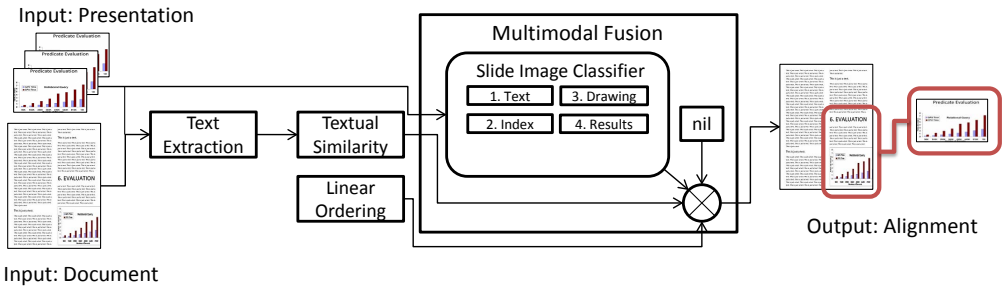

Figure 1: System Architecture.

ument sections not close to the natural alignment point are deemed to have zero probability.

3.1 Slide Image Classification-based Fusion

We observed that over 20% of our dataset are represented by visual cues, which attests to its importance in the alignment process. To enable a first attempt at using this information, we devised an slide image classification, aimed to distinguish four, easy-to-differentiate slide image classes. Note that this classification overlaps but is not identical to our earlier, baseline error-driven analysis.

The four classes covered by our classifier are: 1) *text*, 2) *outline*, 3) *drawing elements* and 4) *results*. *Results* slide images encompass charts, tables, and other visual objects that typically appear in the evaluation portion of a presentation.

Using 10-fold cross validation over a separate dataset of 750 manually annotated slides (by the first author), we trained a linear SVM using features representing the Histogram of Oriented Gradients (HOG) (as suggested in [2]) as computed over the slide image. HOGs are feature descriptors used in computer vision and image processing for the purpose of object detection [2]. The technique counts occurrences of gradient orientations in localized portions of an image. This method is similar to other vision techniques of edge orientation histograms, scale-invariant feature transform descriptors (SIFT) and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. HOG relies on both the gradient and edges in an image and is more robust than its predecessor features. Also, it has been recently shown that HOG can be applied for text detection/extraction from images [13]. We used a patch size and bin size for HOG were 9 and [32 32], respectively, after appropriate manual tuning. Pre-processing was performed using blurring filters following by normalization (to enforce a mean of 0 and standard deviation of 1). Overall, the resulting image classifier returned an acceptable 87% average accuracy over our cross-validation runs (Table 3). Figure 2 gives sample classification probabilities on slides from two classes – *results* and *outline*: for each class the probability that classifier has assigned to some example slides is shown.

Table 3: 10-fold slide image classifier performance.

Slide Type	Text	Outline	Drawing	Results	Average
Accuracy	86%	95%	83%	84%	87.2%

The resultant slide image classification tells us what sources of evidence to trust for the final alignment: text, ordering, or the possible *nil* alignment. We find that the *results* class is

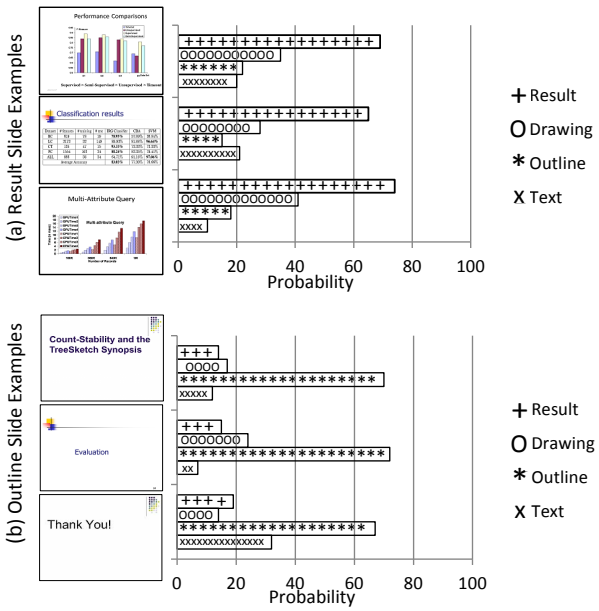

Figure 2: Image classification performance on two sets of example slides: (a) "Result" slide examples and (b) "Outline" slide examples.

accurate enough to use as an unconditional rule: if the image classifier deems a slide as a *result*, we attempt to locate the document section(s) marked by an appropriate header (*i.e.*, "Results", "Experiments", "Discussion"; actually determined by a regular expression over section headers); sample results are shown in Figure 3(a). For the other three classes, the text similarity, linear ordering alignment probability vectors are weighted and fused with a threshold for *nil* alignment to yield the final alignment.

— For *Text* slides, the system increases the weight for text alignment output by 2/3 of its current value, since text similarity measures are more accurate when sufficient amount of text is available in a slide.

— For *Outline* slides which are potentially *nil* slides, the system lowers both w_{Ti} and w_{Oi} to disfavor alignment, setting both as 1/3 of their current values. Figure 3(b) shows some examples of *Outline* slide thumbnails.

— For *Drawing* slides, no conclusion can be drawn from

the slide visual appearance. This kind of slides may relate to any section, thus our system gives uniform probability to all weights. Our system relies on the judgment of the other modalities for the alignment decision. We fuse the alignment vectors v_{Ti} and v_{Oi} and *nil*, re-weighting them with their coefficients as follows:

$$v_A i = argmax_i\{w_t(v_{Ti}) + w_o(v_{Oi}) + (1 - w_t - w_o)(nil)\} \quad (1)$$

The section i (or *nil*) that has the highest likelihood is thus deemed the correct alignment.

4. EXPERIMENTS AND RESULTS

We run through four experimental conditions to assess the performance of our methods.

Our first experiment starts with paragraph-to-slide alignment, as opposed to section-to-slide alignment, to make our evaluation comparable to Kan [7]. Here, we only used textual data to compute the probability vector. We achieve 52.1% accuracy, outperforming his results. We believe the main reasons for this improvement is due to 1) our preprocessing, which uses more accurate text extraction tools for both slides and papers, resulting in less noisy data; and 2) Kan's evaluation method uses a more complex weighted Jaccard accuracy, whereas in our proposed system, a slide is correctly aligned if the first suggested paragraph is correct. Thes best result achieved by Kan and our baseline system are shown in Table 4.

Aligning to coarser-grained sections instead of paragraphs will result in higher accuracy without any system chance. When we performed our bi-modal alignment to sections, accuracy improves by 8.6%. This simplifies the problem, but from a usability standpoint may be more useful, as a user interface is likely to show sections of a document rather than a single paragraph (the latter being too small a text unit to show individually).

Table 4: Alignment accuracy results for different experimental conditions.

Method	Accuracy
Kan (weighted Jaccard) [7]	41.2%
B&G (automated) [1]	50%
(1): Baseline	52.1%
(2): Section-to-slide	60.7%
(3): 2 + Ordering	66.8%
B&G (w/ manual *nil* removal) [1]	75%
(4): 3 + Image Classification	77.3%

A third condition adds in ordering alignment to the text-only baseline. In this multimodal alignment, we gave static, uniform weights to both probability vectors. We observe a real improvement of 6%, obtained by accounting for natural bias toward monotopic ordering.

Our final condition (Experiment 4) runs our entire system, adding in the proposed image classification system over the previous condition. Using the full functionality of multimodal method improves an additional 10.5%, for a aggregate accuracy of over 77%, a significant improvement over the baseline 52%. Our results also yield higher accuracies to that of Beamer and Girju's [1], even accounting for their manual removal of *nil* slides (a simplication of our problem).

5. CONCLUSION

We have shown that visual information constitutes an important source of evidence for document-presentation alignment. We design a simple, supervised image classifier that uses HOG features to distinguish slide images that are 1) primarily text, 2) outlines, 3) results, and consist of 4) drawing elements. This image classifier is used to properly weight image, text and ordering evidence in alignment. The results particularly help to identify non-alignable (*nil*) slides, improving accuracy substantially.

6. REFERENCES

[1] B. Beamer and R. Girju. Investigating automatic alignment methods for slide generation from academic papers. In *Proceedings of CoNLL*, page 111, 2009.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of CVPR*, volume 1, pages 886–893. IEEE, 2005.

[3] E. E. Ephraim. Presentation to document alignment. Undergraduate thesis, National University of Singapore, 2006.

[4] T. Hayama and S. Kunifuji. Relevant piece of information extraction from presentation slide page for slide information retrieval system. *Knowledge, Information, and Creativity Support Systems*, pages 22–31, 2011.

[5] T. Hayama, H. Nanba, and S. Kunifuji. Alignment between a technical paper and presentation sheets using a hidden markov model. In *Proceeding of Active Media Technology*, pages 102–106. IEEE, 2005.

[6] T. Hayama, H. Nanba, and S. Kunifuji. Structure extraction from presentation slide information. *Proceedings of PRICAI: Trends in Artificial Intelligence*, pages 678–687, 2008.

[7] M.-Y. Kan. Slideseer: A digital library of aligned document and presentation pairs. In *Proceedings of JCDL*, pages 81–90. ACM, 2007.

[8] G. Liew and M. Kan. Slide image retrieval: a preliminary study. In *Proceedings of JCDL*, pages 359–362. ACM, 2008.

[9] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. *Advances in Information Retrieval*, pages 16–27, 2007.

[10] M. Sravanthi, C. Chowdary, and P. Kumar. Slidesgen: Automatic generation of presentation slides for a technical paper using summarization. In *Proceeding of FLAIRS*, 2009.

[11] Y. Wang and K. Sumiya. Skeleton generation for presentation slides based on expression styles. *Intelligent Interactive Multimedia: Systems and Services*, pages 551–560, 2012.

[12] W. Yih and C. Meek. Improving similarity measures for short segments of text. In *Proceedings of Artificial Intelligence*, volume 22, page 1489. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[13] J. Zhang and R. Kasturi. Text detection using edge gradient and graph spectrum. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3979–3982. IEEE, 2010.