ACCESSMATH: INDEXING AND RETRIEVING VIDEO SEGMENTS CONTAINING MATH EXPRESSIONS BASED ON VISUAL SIMILARITY

Kenny Davila [1], *Anurag Agarwal* [2], *Roger Gaborski* [1], *Richard Zanibbi* [1], *Stephanie Ludi* [3]

[1] Department of Computer Science, Rochester Institute of Technology, Rochester, NY 14623
[2] School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623
[3] Software Engineering, Rochester Institute of Technology, Rochester, NY 14623

ABSTRACT

AccessMath project is a work in progress oriented toward helping visually impaired students in and out of the classroom. The system works with videos from math lectures. For each lecture, videos of the whiteboard content from two different sources are provided. An application for extraction and retrieval of that content is presented. After the content has been indexed, the user can select a portion of the whiteboard content found in a video frame and use it as a query to find segments of video with similar content. Graphs of neighboring connected components are used to describe both the query and the candidate regions, and the results of a query are ranked using the recall of matched graph edges between the graph of the query and the graph of each candidate. This is a recognition-free method and belongs to the field of sketch-based image retrieval.

Index Terms— Math Retrieval, Content-Based Image Retrieval, Sketch-Based Image Retrieval

1. INTRODUCTION

The AccessMath project will be a complete system aimed to help visually impaired students both in and out of the classroom. While the project has many components, the focus of this work is a retrieval procedure of the content found on videos of math lectures. Given a section of a frame of such videos, the retrieval procedure must return a ranked set of frames representing video segments with related content. This procedure requires an automated way of indexing the content of the videos, and also a method for similarity measurement between any given pair of regions of whiteboard content. This problem falls into the categories of content-based image retrieval (CBIR) and math information retrieval (MIR). Since the proposed solution is recognition-free and treats the math formulas as handwritten sketches, the solution becomes a sketch-based image retrieval (SBIR) system.

Figure 1 illustrates the two sources of video provided per lecture, the first one is a camera in the classroom and the second one is the software of a Mimio device. Additional details about these are provided in Section 3. These video sources are combined using image processing techniques to finally extract and index their content. To find the similarity between a given query and the content stored in the index, a similarity measurement based on both local and global features is used. At the local level, OCR-like features are applied to determine the similarity between two given connected components (CC). At the global level, a Graph of Neighboring Connected Components (GNCC) is built, and similarity is measured using the recall of matched graph edges between the GNCC of the query and the GNCC of the candidate regions. The indexing and retrieval process is discussed in Section 4.

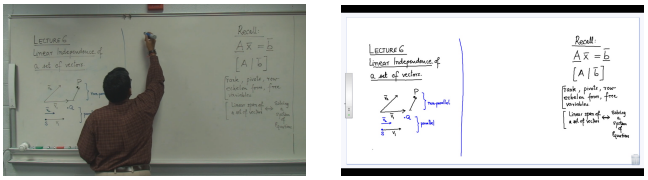
Supported by the National Science Foundation Award HCC-1218801

2. BACKGROUND

Retrieval of video segments is a multi-modal problem. Currently we retrieve content using only visual information, without explicit recognition of symbols.

Detection of changes in whiteboard content over a sequence of images is a requirement for content extraction. The ReBoard system [1] detects changes within cells of low and high resolution pixel grids. Also, the whiteboard capture system developed by Microsoft [2] uses a similar approach based on a pixel grid and classification of cells as whiteboard, foreground object or stroke. This classification is later refined using spatial and temporal information.

The retrieval of the math content is the most important issue to address on this application. There is previous work in math recognition and retrieval that aims to retrieve math for-



(a) Still Camera Video (b) Mimio Software Video

Fig. 1. Current video sources: (a) Main, (b) Auxiliary.

mulas found in images. A survey can be found in the work by Zanibbi and Blostein [3]. However, most of these approaches only work for printed math formulas. Also, they usually rely on optical character recognition (OCR) which requires all symbols that will be used to be known before-hand, which is not practical for handwriting on the whiteboard.

Retrieval of visually structured content found in images can be done through SBIR. A frequent idea on this field is that sketches are built using sets of primitives that are spatially interrelated. For applications like math retrieval, these spatial relationships play an important role in the semantics of the drawings. A common problem among SBIR systems is how to represent these spatial relationships, and a common solution is the use of graphs with vertices representing each primitive and edges representing spatial relation between pairs of primitives. For example, the work by Leung [4] uses hierarchy trees to represent inclusion relationships between strokes.

The measurement of similarity between sketches is important because it determines the performance of the SBIR system both in running time and quality of results. Efficient matching of similarity between graphs is an open problem, and different SBIR systems apply various graph-similarity metrics. Certain works use approximations of graph isomorphism, like for example Cordella et. al [5] on their application for retrieval of technical drawings. However, pure isomorphism can be used to tell whether two structures are equivalent or not, but not as a measurement of similarity. Other approaches use combinations of heuristic rules, such as the work by Leung [4] which applies different similarity measurements and combines them into a single value. Also, some use explicit graph embedding methods [6] for similarity. Additional examples of this method can be found on works that use graph spectra for similarity measurement [7] [8].

There are complete systems for retrieval of content written on the whiteboard that are relevant to our application. The system by Liwicki and Bunke [9] applies OCR over On-line and Off-line data of whiteboard notes and indexes their content. Another application is the Thor system [10] that indexes whiteboard content in images. Finally, the work by Leung [4] uses on-line data of traces for retrieval of drawings based on visual similarity.

3. DATASET

Our dataset is a small collection of videos of linear algebra lectures recorded at Rochester Institute of Technology. Currently, there are only six lectures on the collection, but it is growing and will be larger in the future. For each lecture, a still camera has been set in the classroom to record exactly one whiteboard and its content. Also, each recording comes with an auxiliary video of the strokes of the board captured using a Mimio Capture device and the Mimio software [1].

[1] http://www.mimio.com/en-NA/

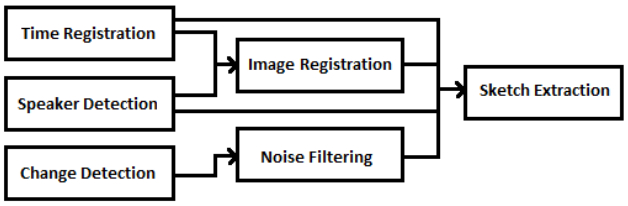


Fig. 2. Sketch Extraction Process.

The video coming from the still camera is the main source and has a resolution of 1440x1080 pixels. Auto focusing is turned off to avoid change in focus affecting the quality of the strokes. This video has some important drawbacks, the first one is the presence of the speaker blocking parts of the whiteboard content, and the second is the constant changes of illumination on the scene. For these reasons, an auxiliary video coming from the Mimio software is attached. Mimio Capture works using special marker sleeves that when the user writes they emit radio frequencies which allow identifying both position and color of the current marker. This information is sent in real-time to the Mimio software where it is recorded on a screen-captured video which means that its quality will be relative to the resolution of the screen where the Mimio software is displayed. It has the great advantage of the absence of the writer, but due to sensor errors it is usually very noisy to the point that it is not a reliable source of content. However, it is easier to detect changes using the Mimio video. Examples of frames extracted from these videos are shown in Figure 1.

4. METHODOLOGY

There are many processes involved in the extraction, indexing and retrieval of content of the whiteboard on videos from math lectures. Figure 2 shows a diagram of the most important procedures in the sketch extraction process which are briefly described on this section. For detailed descriptions please refer to our technical report [11].

Registration: Time and image registration are required to match the content between the main and the auxiliary videos of each lecture. Time registration is performed using features of the audio stream of each video. Motionless frames are selected from the main video and their corresponding frames from the auxiliary video are extracted. Then, image registration between pairs of corresponding frames is performed using Speeded Up Robust Features (SURF) [12].

Speaker/Change Detection : Frame differencing is applied over a sub-sampled version of the main video in order to detect motion and estimate the speaker location. Note that while sophisticated techniques could detect the speaker at pixel level, estimation at the region level is more than enough to ensure extraction of non-blocked content. Frame differencing is also applied over the auxiliary video to find pixels

with large changes in luminosity. A grid is created to group changed pixels into cells, and these cells are grouped to form sketches which are the basic units for retrieval.

Sketch Extraction: Using the results of the previous operations, the system extracts the images of the sketches from the main video. Then, edge detection is combined with morphological operations and CC labeling for extraction of the primitives of each sketch.

Sketch Description: Sketches are described at two different levels: local and structural. These descriptions can be indexed and used for similarity comparisons. At local level, each CC is normalized to a predefined size without losing the original aspect ratio, and different features are computed per CC. The first is the normalized aspect ratio. Second are the mean, the covariance matrix and a 2D histogram of foreground pixels locations. Finally, using horizontal and vertical lines at predefined positions, the intersections between the CC and these lines are computed, and three values are added per line: first, last, and count of intersections. More detailed descriptions about these features can be found in [11]. At structural level, two fully-connected graphs are created per sketch where each vertex represents a CC, and each edge is weighted using a distance metric between CC. The first graph uses distance between centers of CC while the second uses smallest distance between borders of their bounding boxes. A minimum spanning tree is calculated for each graph, and all surviving edges are combined to form a single GNCC where each CC is connected only to its closest neighbors. Thanks to the division of content, these graphs are usually small.

Sketch Grouping: Modification and deletion times of each sketch are used to generate special groups of sketches. These groups can be considered the Key Frames of the video that divide it into segments and are used as secondary units for retrieval.

Retrieval: Based on a similarity function that produces a value which can be used to rank the candidate results. Currently, this similarity function is implemented using Recall of matched pairs on the GNCC of the sketches. A pair $p = (u, v, \theta)$ on a GNCC represents two CC u and v connected by an edge on the graph. The angle θ represents the orientation of the line that connects their centers, and has a value on the range $\left[-\frac{\pi}{4}, \frac{3\pi}{4}\right)$. Two given pairs $p_1 = (u_1, v_1, \theta_1)$ and $p_2 = (u_2, v_2, \theta_2)$ can be matched using the function M defined in equation 2.

$$S(x, y) = ||F(x) - F(y)|| \leq \alpha \tag{1}$$

$$M(p_1, p_2) = S(u_1, u_2) \wedge S(v_1, v_2) \wedge (|\theta_1 - \theta_2| < \beta) \tag{2}$$

Where F is a function that returns the feature vector of a primitive, α represents a threshold of similarity between CC and β represents a threshold of similarity between orientations. The empirically chosen values for these constants are $\alpha = 3.5$ and $\beta = \frac{\pi}{8}$. Also, the ordering of the CC is always important as we want to compare the CC on the corresponding sides of the edges. Finally, suppose that two GNCC G_q and G_c and their corresponding sets of pairs P_q and P_c are given, then the function $Recall$ defined in equation 4 is the measurement of similarity between them.

$$H(p_i, P_c) = \begin{cases} 1, \text{if } \exists p_j \in P_c \mid M(p_i, p_j) = 1 \\ 0, \text{otherwise} \end{cases} \tag{3}$$

$$Recall(G_q, G_c) = \frac{\sum_{p_i \in P_q} H(p_i, P_c)}{|P_q|} \tag{4}$$

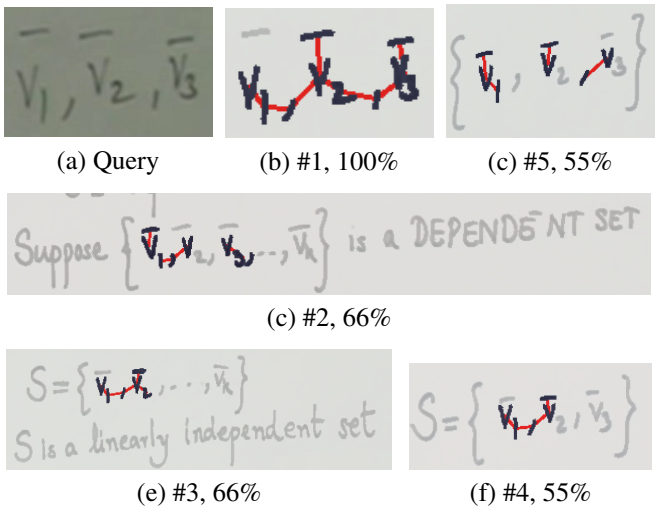
In equation 4, the recall of matched pairs is obtained by dividing the number of matches by the total of pairs on the query graph. In this sense, we could obtain precision by just swapping the parameters of this function. However, these matches are not unique using the current functions which means that precision might be measured with a different numerator. For this reason, the F-measure that combines precision and recall cannot be used until the matches are guaranteed to be unique. This can be solved using the Hungarian method [13], but it is computationally expensive and therefore a sub-optimal greedy matching is preferred.

Note that there are no further spatial restrictions applied between matched edges, and as a result it can be the case that two edges that have a vertex in common on the query could match two edges with no vertex in common on the candidate sketch. However, we could observe in our tests that this method works better than just matching CC individually without any structural restriction on the matches.

5. PRELIMINARY RESULTS

This is a work in progress and no benchmarking experiments have been performed yet. However, in our tests we could observe that using our method we usually obtain many relevant results in the top 10. Figure 3 shows an example of a query and the kind of sketches that were retrieved using the current method. The query simply contains three vectors. Note that all top 5 matches also contain vectors, and even if they have different arrangements, they can still be considered as valid matches. Our method is effective for queries like this one because vector notation is an example of a subexpression of two elements contained in a query that many users would expect to find in the results.

Currently there are many confusion errors and drawbacks in the matching process. Further refinement of parameters involved in the matching could reduce the confusion errors. However, in figure 3.(b) we can observe that even when a query is matched against itself, multiple edges of the GNCC of the query can be matched with a single edge of the GNCC of the candidate allowing graphs that are smaller than the query to achieve 100% recall. Another drawback is that the



(a) Query (b) #1, 100% (c) #5, 55%

(c) #2, 66%

(e) #3, 66% (f) #4, 55%

Fig. 3. Query executed using the recall of matched pairs on GNCC. The bold CC and red edges represent matched pairs

method is very sensitive to touching symbols that become a single primitive instead of making a pair. Usually, false positives are regions that contain many partial matches for the query, but are unrelated when considered as a whole. Still, even with the current limitations, the system achieves the retrieval of related content in the top 10 results for many queries and it needs to be tested on a larger scale.

6. CONCLUSION

Extracting information from videos is a challenging task prone to errors at many steps of the process, and even if all information is extracted perfectly, the measurement of similarity is critical in the production of relevant results for every query. Of course, this measurement also needs to be fast enough to handle queries on reasonable times, and usually special index structures can reduce these times. Our system would benefit from improvements in handling of noise, similarity measurement, and index structure. Also, experiments involving many users are required to identify additional areas of improvement for our method.

Different task have been identified as open for improvement. The first one is the retrieval task which could be improved with more sophisticated matching methods that consider additional spatial restrictions. Also, a fast method for 1-to-1 matching of pairs is required to allow us to apply the F-measure for ranking of results instead of just recall of matched pairs. In addition, a better set of local features would increase the overall quality of results. Subdivisions of CC will be required for partial matching for handling of cases with cursive writing and touching symbols. Finally, the index structure is currently just storage of pre-computed features, but a better alternative could be found that would speed-up the system by reducing the number of initial candidate sketches based on some general features of the sketches.

7. REFERENCES

[1] Gene Golovchinsky, Scott Carter, and Jacob Biehl, "Beyond the drawing board: Toward more effective use of whiteboard content," Tech. Rep., FX Palo Alto Laboratory, 2009.

[2] Li-wei He, Zicheng Liu, and Zhengyou Zhang, "Why take notes? Use the whiteboard capture system," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 5, pp. 776–779.

[3] Richard Zanibbi and Dorothea Blostein, "Recognition and retrieval of mathematical expressions," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 15, no. 4, pp. 331–357, 2012.

[4] Howard Wing Ho Leung, *Representations, feature extraction, matching and relevance feedback for sketch retrieval*, Ph.D. thesis, Carnegie Mellon University, 2003.

[5] Luigi P Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento, "A (sub) graph isomorphism algorithm for matching large graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1367–1372, 2004.

[6] Muhammad Muzzamil Luqman, Jean-Yves Ramel, Josep Lladós, and Thierry Brouard, "Fuzzy multilevel graph embedding," *Pattern Recognition*, vol. 46, no. 2, pp. 551–565, 2013.

[7] Pedro Sousa and Manuel J Fonseca, "Sketch-based retrieval of drawings using spatial proximity," *Journal of Visual Languages & Computing*, vol. 21, no. 2, pp. 69–80, 2010.

[8] Shuang Liang and Zhengxing Sun, "Sketch retrieval and relevance feedback with biased SVM classification," *Pattern Recognition Letters*, vol. 29, no. 12, pp. 1733–1741, 2008.

[9] Marcus Liwicki and Horst Bunke, *Recognition of Whiteboard Notes: On-line, Off-line, and Combination*, vol. 71 of *Machine Perception and Artificial Intelligence*, World Scientific, 2008.

[10] Mihai Parparita and Szymon Rusinkiewicz, "Thor: Efficient whiteboard capture and indexing," Tech. Rep., Princeton University, 2004.

[11] Kenny Davila, "Math expression retrieval using symbol pairs in layout trees," M.S. thesis, Rochester Institute of Technology, 2013.

[12] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*, pp. 404–417 Springer, 2006.

[13] Harold W Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.