

# Disease-Specific Extraction of Text from Cardiac Echo Videos for Decision Support

Tanveer Syeda-Mahmood    David Beymer    Arnon Amir  
IBM Almaden Research Center  
650 Harry Road, San Jose, CA 95120 USA  
stf@almaden.ibm.com

## Abstract

*Echo videos are an important modality for cardiac decision support. In addition to describing the shape and motion of the heart, they capture important diagnostic measurements as textual feature-value pairs that are good indicators of the underlying disease. In this paper, we describe reliable extraction of such textual information through selective image processing and region extraction prior to using an OCR engine. We then use tabular layout analysis rules to recover measurement attribute value pairs from the recognized text in videos. The measurement feature-value pairs are used to retrieve matching videos from a database. A ranked list of matching diseases is then obtained through collaborative filtering. Results are demonstrated on a large echo video database of patients with various diseases.*

## 1 Introduction

Cardiac echo videos are an important source of diagnostic information in cardiac decision support. Captured during an echocardiographic exam, these videos depict shape and motion of the heart from different viewpoints. In addition, they capture important measurements such as area of the left ventricle, the velocity of the doppler flow jet, and mitral valve area as text feature-value pairs. Figure 1 shows a text-only page from an echocardiogram. As can be seen, relevant features and their values (eg. Left ventricular (LV) Area, LV Volume) are grouped by cardiac regions (left ventricle (LV), mitral valve (MV)), and appear in an implicit tabular format with rows corresponding to specific measurements, and columns capturing 1) temporal points in the heart cycle (systolic, diastolic), 2) different viewpoints (Apical 4-chamber, Parasternal long-axis) or 3) computational method for estimation (“method of disks”, “area length”). Cardiologists use these measurements to make conclusions about the disease. For example, the ejection fraction (LV EF) recorded is 45.2% indicating a left ventricular dysfunction.

Thus, sufficient information is available from these features to infer similarity between cardiac echo videos for purposes of decision support.

In this paper, we present a new approach to finding similar cardiac echo videos based on feature-value pairs extracted from embedded text regions in videos. To reliably extract the embedded text, we use image processing operations based on mathematical morphology, edge detection and rank filtering. This is then followed by grouping and connected components analysis to prepare the text regions for an OCR engine. Next, we detect the tabular layout structure within the video frames that depict measurements to capture the pairings of features with their values. Each echo video is thus represented as a set of feature-value pairs. Similar echo videos are obtained by ranking the database of videos using a similarity metric that captures the fraction of overlap between features with similar values. Finally, we demonstrate an application of similarity retrieval of videos for clinical decision support.

The rest of the paper describes our approach to the automatic extraction of these feature-value pairs of measurements and their use in similarity retrieval of videos for decision support. In Section 2, we review related work. In Section 3, we describe our image pre-processing prior to the application of an OCR engine. In Section 4, we present a top-down approach to understand the tabular structure of these pages for a reliable extraction of feature-value measurements. In Section 5, we present our document retrieval algorithm that uses the set of measurements returned by the tabular analysis. Finally, in Section 6, we present results of using automatically extracted feature-value pairs for text-based document retrieval on a large database of videos.

## 2 Previous work

While we are not aware of any work that uses disease-specific feature-value pairs from embedded text in videos, there is considerable work on both optical character recognition (OCR) in videos[6], and in document layout analysis.



Figure 1. Sectioning and tabular structure of a sample text-only echo page.

Extraction of text from videos has been an active research problem in the last decade [9]. The accuracy of such systems has not reached the performance obtained by more developed OCR engines for documents such as Tesseract [10] due to the complex backgrounds present in videos.

Considerable work has also been done in tabular structure analysis in documents[2]. Most common are bottom-up approaches that build the table in a data driven manner exploiting general a priori knowledge about table formatting. Frequently, it involves locating horizontal and vertical ruled lines bounding the table cells [3], grouping the cell text into rows and columns using vertical and horizontal projections [1], building an adjacency graph on text regions [7], or modeling as a tree structure[12]. Due to the constrained nature of the text-only pages in the videos, a top-down approach based on templates from exemplars is more reasonable for our work. Top-down templates have also been suggested earlier to handle billing statements[8], and detecting tables by their headers [13, 4]. Finally, the retrieval of similar videos for cardiac decision support using video-derived features has been reported [11], where a limited number of diseases could be distinguished due to the small number of features used. Textual measurements, on the other hand, can give enough features to capture a larger complexity of cases where patients have more than one disease, making them practical for decision support.

### 3 Video pre-processing for text recognition

Given the collection of echo clips from one patient visit, we first identify text-only frames and separate them from the rest of the video. We then extract text-containing regions and assemble them into a single image for optical character

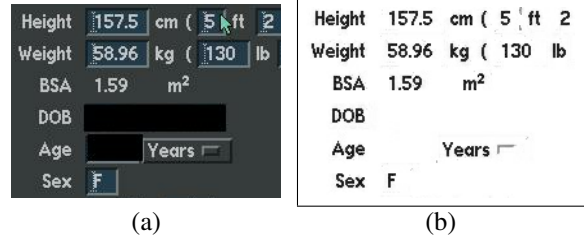


Figure 2. Sample portion of a text-only frame. In the original image (a) boxes, lines, carets and cursor are visible. In the cleaned image (b) these have been removed.

recognition as described below.

### 3.1 Text-only page recognition

A set of product logo templates were made for various echo machine models (eg. Siemens Sequoia) and normalized correlation in pre-determined regions in successive video frames made it possible to recognize the machine model in a given video. Once the machine type is known, a pre-determined template of text-only frames is used to separate the text-only frames from the rest of the video.

### 3.2 Extraction of text regions

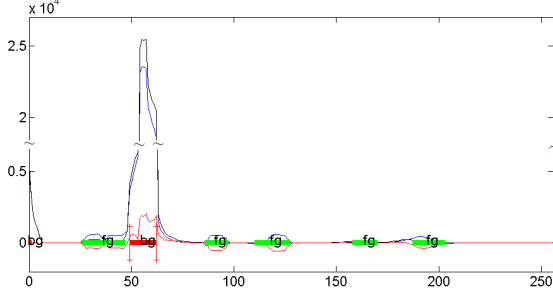
Echocardiogram frames contain various image, graphical and textual information in a compact visual presentation. An example is shown in Fig. 2(a). The text may appear in several different colors and gray levels. Some parts are highlighted and rendered in inverse colors (black on white). The frame may contain input fields, bordered with lines. Each input field contains a caret (cursor), often connected to one of the characters. Before such frame can be effectively processed by an OCR engine, any non-textual content has to be removed and the text appearance should be made homogeneous, for optimal OCR performance.

#### 3.2.1 Foreground-background segmentation

For optimal OCR recognition, regions of text of different color or gray level are detected and normalized. Text and graphics in echo videos are often associated with designated colors or gray level bands, such as the white and gray lines observed in Fig. 2(a).

If those gray level bands are present, our processing automatically detects the bands, labels them as foreground or background, processes the content of each foreground band, and removes lines and other non-textual graphics before normalizing the text gray levels.

To detect and classify gray level bands in a text-only video frame, we first compute two gray-level histograms; one of the original frame, shown by a black curve in Fig. 3,



**Figure 3. Histogram analysis and automatic classification of histogram bands associated with figure (green bars) and background regions (red bars).**

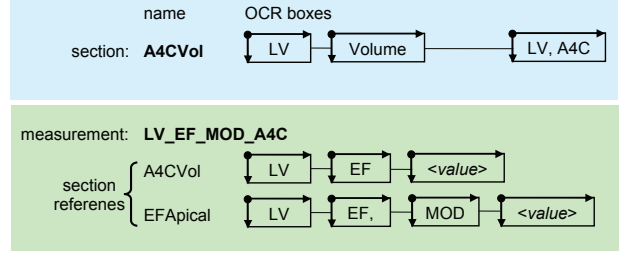
and the other on the frame after local median filtering, shown in blue. We then look at the difference between the two histogram (red curve). Foreground bands are associated with negative difference, while background bands are associated with a positive difference. Once bands are separated the image foreground objects can be easily separated and processed. This involves line detection and removal, histogram stretching of bands associated with text, and replacing gray background regions with darker background values, to create a uniform background color. Once the regions are obtained, we apply a conventional OCR engine [10] to extract the words in various text regions.

## 4 Analyzing tabular layout

We now discuss the extraction of tabular structure in the text-only pages.

### 4.1 Template generation

The generation of templates for section and measurements within sections is echo machine-specific. In our collection, we have page samples from Siemens models Sequoia, Cypress, and Aspen. To form the section template per machine model, we use the geometric coordinates of words returned by the OCR engine as well as the recognized text on sample training text-only video frames as shown in Fig. 4. The templates for measurements within sections are similarly made as shown in Fig. 4. Since the same measurement can be in more than one section, we retain a list of section occurrences for each measurement. Each section occurrence captures the name of the measurement label per row in the section. It also captures the geometric location of the measurement label and its value.



**Figure 4. Templates for defining sections and echo measurements.**

#### 4.1.1 Table analysis

We now describe how the section and measurement templates are matched to the text-only pages to extract measurement values.

**Section recognition** Given the OCR word boxes, the first step to layout analysis is to locate the variable sections on the text-only pages. We detect variable sections by matching a section definition,  $sec$ , to the page OCR output,  $ocr$ . Further, let  $ocr_i$  be the page OCR starting at the  $i$ th word. The best match is found by “sliding”  $sec$  over the page  $ocr$ , testing the match at each OCR word  $ocr_i$ . Let  $e_i$  be a text match measure for offset  $i$

$$e_i = (|sec| - \text{editdist}(ocr_i, sec)) / |sec| \quad (1)$$

where  $|sec|$  is the length of  $sec$  in characters and normalizes  $e_i$ . Let  $d_i$  be a geometrical match between word bounding boxes in  $ocr_i$  and  $sec$ , where  $sec$  is translated in  $y$  to match the  $y$ -coordinate of the first word of  $ocr_i$ ,  $y_i$ , with the  $y$ -coordinate of the first word of  $sec$ ,  $y_{sec}$

$$d_i = \text{bbox-error}(ocr_i, \text{translate}(sec, y_i - y_{sec})) \quad (2)$$

and  $\text{bbox-error}$ , depicted in Fig. 5 for one pair of words, is computed for all pairs of words in  $sec$  and their corresponding words in  $ocr_i$ , and finally maximized over words. The match between  $ocr$  and  $sec$  is finally given by

$$m(ocr, sec) = \arg \max_i \{e_i | (e_i > 0.9) \& (d_i < 10)\} \quad (3)$$

where we impose hard constraints on  $e_i$  and  $d_i$  and maximize  $e_i$  over the surviving  $i$ . Locating section definitions on the text-only pages parses the page vertically; the page OCR is partitioned appropriately according to section,  $ocr_{sec}$ .

**Measurement extraction** Measurements list all sections they can appear in,  $sec_j$ , along with the appearance of the variable label,  $ocr_j$

$$meas = \{(sec_j, ocr_j)\}.$$

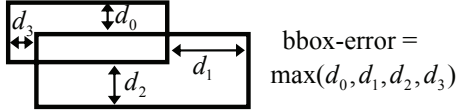


Figure 5. Measuring the geometrical distance between two OCR bounding boxes.

Section detection has detected some number of *meas*'s sections on the current text-only page. For each such detected section,  $sec_j$ , call the OCR partition of this section  $ocr_{sec_j}$ . Measurements are extracted by matching  $sec_j$ 's OCR block  $ocr_j$  against the detected partition of  $sec_j$ ,  $ocr_{sec_j}$

$\forall j$ , IF ( $sec_j$  detected) &  $m(ocr_{sec_j}, ocr_j)$  THEN  
 $value_{meas} = \text{word in } ocr_{sec_j} \text{ matching } \langle value \rangle \text{ in } ocr_j$

Thus, a measurement value can be located if it occurs in any of its constituent sections,  $sec_j$ . All text only pages of a patient visit are processed to extract measurements and the union of all measurements are reported for patient retrieval.

## 5 Retrieval of similar patient records

We now discuss similarity retrieval of echo videos based on extracted feature-value pairs of measurements. If we denote the set of all feature names found across all text-only pages of videos by  $F = \{f_j\}$ , then an echo video  $D_i$  can be characterized by a feature list

$$H_i = \{(f_j, v_j), f_j \in F_i\} \quad (4)$$

where  $f_j$  is the feature and  $v_i$  is the value of the feature and  $F_i \subseteq F$ . Two features  $f_{li} \in F_i$  of video  $D_i$  and  $f_{mj} \in F_j$  match provided  $f_{li} = f_{mj}$ , and  $|v_{li} - v_{mj}| < \tau$ , where  $\tau$  is a suitably chosen threshold based on the knowledge of feature semantics. The American Heart Association has defined guidelines for tolerance ranges on cardiac measurements to signal diseases, which were used to serve as feature-specific thresholds. Let the set of matching features between two videos  $D_i$  and  $D_j$  be denoted by  $M_{ij} = \{(f_{li}, f_{mj}), |v_{li} - v_{mj}| < \tau\}$  Then the extent of match of two cardiac videos  $D_i$  and  $D_j$  is given by

$$d(D_i, D_j) = \frac{|M_{ij}|}{|F_i| + |F_j|} \quad (5)$$

Given a query video of a patient, the videos in the database are ranked based on the above metric and the best matches retained are those with a  $d(D_q, D_j) > T$  for a threshold  $T$ . The value of  $T$  is varied to achieve desired precision and recall as described below.

## 5.1 Decision support through similarity retrieval

So far we have kept the discussion general from the point of video retrieval. We now discuss its use in clinical decision support. In clinical decision support, we are primarily interested in inferring the distribution of diseases from the similar videos retrieved. To offer decision support, we retain  $K$  matching videos that are within a distance threshold  $T$  from the query video in a style reminiscent of collaborative filtering[5]. The choice of  $K$  and  $T$  is derived through standard cross-validation experiments as done in collaborative filtering[5]. Let the matches to the query document thus retained be denoted by  $D_M = \{D_1, D_2, \dots, D_K\}$ .

Let the disease labels associated with the video  $D_i$  be  $E_i = \{e_1, e_2, \dots, e_k\}$ . Then the ranked list of disease matches is obtained by taking the histogram of hits for the disease labels corresponding to the matching videos. Thus a rank of a disease label  $e_j$  is given by

$$R(e_j) = |\{E_i | e_j \in E_i\}| / K \quad (6)$$

The above ranking reflects the principle of collaborative filtering. Thus a disease label with a high rank means it is a disease label that is voted by the majority of matches, thus increasing its probability of being the correct disease label for the query video. All diseases with a rank  $R(e_j) > \delta$  are then retained as valid diseases labels for the query.

## 6 Results

To test the system, we experimented with a large data set of 972 patients with a total of 1876 echo studies. All echo studies were pre-diagnosed so that the disease labels associated with the studies were known. Each echo study was processed to extract text-only pages resulting in a total of 19366 pages. Each video in the database was then represented by the set of feature value pairs  $H_i$ . We now report results of our experiments on this dataset.

### 6.1 Text extraction performance

To validate the techniques used in our paper, we manually defined ground truth of (*meas*, *value*) pairs for a subset of 11 patients, 114 text-only pages, and 1719 total measurements. For the validation set, our system extracted 99.7% of the measurements correctly, with the remaining errors caused by the numeric *value* words being split by OCR into multiple words (e.g. 100 being read as "1" and "00").

To evaluate the additional performance brought to the system by the OCR preprocessing module, we ran the same validation test set with preprocessing turned off. The recognition rate dropped to 94.9%, so preprocessing adds 4.8% to measurement extraction. Of the error cases, 69% of the errors involve mistakes in a numeric word, and in the remaining 31%, the measurement was missed altogether.

## 6.2 Evaluation of decision support

To test the validity of the decision support approach in this paper, we used each echo video of the database as query and retained the top list of matching diseases as captured in Equation 6. Since the disease label of the query video was known a priori, we could measure precision and recall. Consider the query video  $Q$  and its ground truth disease label  $E_Q$ . Let the top disease labels retained after rank filtering be denoted by  $E_M$  where

$$E_M = \{e_j | e_j \in E_i, 1 \leq i \leq K, R(e_j) > \delta\} \quad (7)$$

Then we measure recall performance in decision support by

$$Recall = |E_Q \cap E_M| / |E_Q| \quad (8)$$

Thus a 100% recall is when all the query diseases have a rank  $R(e_i) > \delta$ . We measure the precision as a validation accuracy by recording the average rank of the query matches as

$$precision = \sum_{e_i \in E_Q} R(e_i) / |E_Q| \quad (9)$$

Thus a high average rank indicates that the query labels are indicated as the more likely disease labels and are thus validated by the data. The retrieval performance of our algorithm for the entire database as query videos is shown in Figure 6 for two choices of top K values. As can be seen, over 76% of the queries have their disease labels perfectly matched in the returned documents when  $K=20$  and over 95% of the queries have at least 50% of their disease labels match in the retrieved documents. By averaging the rank scores of all query matches for all queries above, we obtained an average precision score of 85%. Both these results demonstrated the utility of measurement-based cardiac echo video retrieval for diagnosis validation during decision support.

## 7 Conclusion

In this paper, we have presented a new method of retrieving similar cardiac echo videos using automatically extracted feature-value pairs from embedded text in videos. An application of similarity retrieval of cardiac videos for cardiac decision support has been described.

## References

- [1] S. Chandran and R. Kasturi. Structural recognition of tabulated data. In *ICDAR*, pages 516–519. IEEE Computer Society, 1993.
- [2] D. W. Embley, M. Hurst, D. Lopresti, and G. Nagy. Table-processing paradigms: a research survey. *International Journal of Document Analysis*, 8(2):66–86, 2006.
- [3] O. Hori and D. S. Doermann. Robust Table-form Structure Analysis Based on Box-driven Reasoning. In *ICDAR*, pages 218–221. IEEE Computer Society, 1995.
- [4] B. Klein, S. Gokkus, T. Kieninger, and A. Dengel. Three Approaches to “Industrial” Table Spotting. In *ICDAR*, pages 513–517. IEEE Computer Society, 2001.
- [5] W. S. Lee. Collaborative Learning and Recommender Systems. In *International Conference on Machine Learning*, pages 314–321, 2001.
- [6] R. Lienhart. Video OCR: A Survey and Practitioner’s Guide. In *Video Mining*, page Chapter 6. Springer, 2003.
- [7] J. Ramel, M. Crucianu, N. Vincent, and C. Faure. Detection, Extraction and Representation of Tables. In *ICDAR*, pages 374–378. IEEE Computer Society, 2003.
- [8] J. H. Shamilian, H. S. Baird, and T. L. Wood. A Retargetable Table Reader. In *ICDAR*, pages 158–163. IEEE Computer Society, 1997.
- [9] J.-C. Shim, C. Dorai, and R. Bolle. Automatic Text Extraction from Video for Content-Based Annotation and Retrieval. In *ICPR ’98*, pages 618–621. IEEE Computer Society, 1998.
- [10] R. Smith. An Overview of the Tesseract OCR Engine. In *ICDAR*, pages 629–633. IEEE Computer Society, 2007.
- [11] T. Syeda-Mahmood, D. Ponceleon, and J. Yang. Validating cardiac echo diagnosis through video similarity. In *ACM MULTIMEDIA ’05*, pages 527–530. ACM, 2005.
- [12] T. Watanabe, Q. Luo, and N. Sugie. Layout Recognition of Multi-Kinds of Table-Form Documents. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(4):432–445, 1995.
- [13] C. Wenzel and W. Tersteegen. Precise Table Recognition by Making Use of Reference Tables. In *DAS ’98: Third IAPR Workshop on Document Analysis Systems*, pages 283–294. London, UK, 1999. Springer-Verlag.

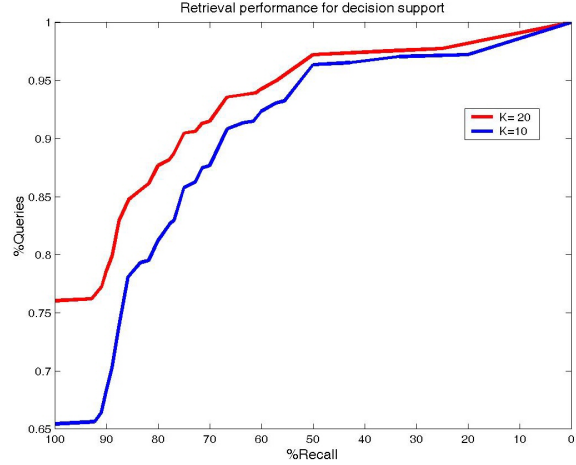


Figure 6. Retrieval performance for our cardiac decision support application.