Echocardiogram Video Summarization

Shahram Ebadollahi[*a], Shih-Fu Chang[a], Henry Wu[b], Shin Takoma[b]

[a]Department of Electrical Engineering, Columbia University, New York, NY. 10027
[b]Columbia University College of Physicians and Surgeons, New York, NY. 10034

ABSTRACT

This work aims at developing innovative algorithms and tools for summarizing echocardiogram videos. Specifically, we summarize the digital echocardiogram videos by temporally segmenting them into the constituent views and representing each view by the most informative frame. For the segmentation we take advantage of the well-defined spatio-temporal structure of the echocardiogram videos. Two different criteria are used: presence/absence of color and the shape of the region of interest (ROI) in each frame of the video. The change in the ROI is due to different modes of echocardiograms present in one study. The representative frame is defined to be the frame corresponding to the end-diastole of the heart cycle. To locate the end-diastole we track the ECG of each frame to find the exact time the time-marker on the ECG crosses the peak of the R-wave. The corresponding frame is chosen to be the key-frame. The entire echocardiogram video can be summarized into either a static summary, which is a storyboard type of summary and a dynamic summary, which is a concatenation of the selected segments of the echocardiogram video. To the best of our knowledge, this is the first automated system for summarizing the echocardiogram videos based on visual content.

Keywords: Echocardiogram videos, digital video libraries, PACS, video summarization, key-frame selection, scene boundary detection.

1. INTRODUCTION

Echocardiography[1] is a widely accepted tool for the qualitative and quantitative assessment of a patient's cardiac health. Its wide acceptance is more due to the fact that it is a non-invasive and cheap method for imaging the cardiac structure. Every year, a typical hospital acquires hundreds; if not thousands; of echocardiogram videos from the patients. Currently these videos are recorded on analog videotapes. Although videotape is a versatile and cost effective means of recording and storing echocardiogram videos, there are significant limitations with it. Several such limitations are as follows:

- A full echo study, which includes different forms of echocardiography may take 10 to 20 minutes. The clinician does not have the time and patience to view such long studies.
- Finding a particular echo study from a large archive of videotapes is very time consuming. Usually several studies are stored on a single videotape, which makes the problem even worse.
- Sharing the echocardiogram videos with other clinicians for the purpose of tele-echocardiology is very difficult if not impossible.

The new echocardiogram acquisition devices have the capability to record and store the echo studies in digital format. Also the existing echo videos that are captured on analog tapes can be encoded into digital format using video capture cards. The transition from analog to digital storage of the echocardiograms provides the means to apply the techniques of video processing to the echocardiogram videos and therefore equips the collection of such videos with efficient tools that solves the problems mentioned above.

The goal of our research is to automatically and fully index and annotate the content of the digital echocardiogram videos and link the elements of the visual content of the videos to the text of the diagnosis report and patients' records, which are usually available for each echocardiogram. In order to acquire enough information about the echo videos to be able to get a full semantic description of the videos and therefore be able to annotate them, we need to first pay attention to the syntax of the echo videos. The analysis of the syntactical structure of the echo videos not only helps to get more insight about the higher-level description of the echo videos but also provides us with means to summarize the echocardiograms into a format that is convenient for browsing the archives of the digital echocardiograms. In Figure 1 the block diagram of the system that is used

[*] Correspondence. Email: shahram@ctr.columbia.edu, WWW: http://www.ctr.columbia.edu/~shahram/, Telephone: +1 212 939 7155, Fax: +1 212 932 9421

for the automatic processing of the echocardiogram videos both on the syntactic and semantic levels is presented. In this paper we will only concentrate on the syntax of the echo videos and their summarization, which is embodied in the "Parser" and "Key-frame selector" blocks in the figure.

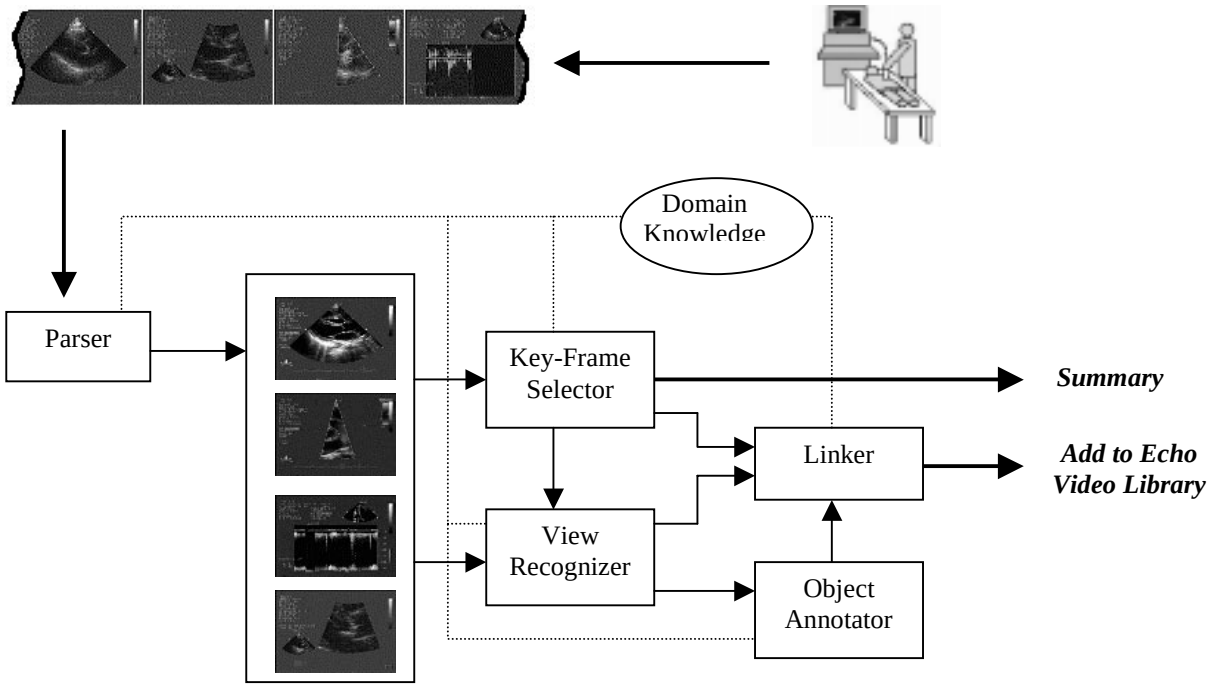


Figure 1. Block Diagram of the Echo Indexing System

Video summaries (abstractions) form the essential components of digital video libraries because they allow the user of such libraries to locate the segment of their interest in the video in a very time-efficient manner. They also provide more efficient and faster communication of the essential content of videos by communicating the video segments or representative images. Summarization is in fact mapping an entire video to a small number of representative images called key-frames.

The first step in summarizing a video is to temporally segment the sequence of frames in the video into smaller units called *shots* (The method described here is what usually is referred to as content-based sampling of the video. The other method is summarizing the video by uniformly sampling the video content. That method does not apply to what we will discuss here.) A shot is a series of frames of a video, which correspond to an uninterrupted camera operation. In the context of echocardiogram videos, a shot is a sequence of frames that correspond to a single position and angle of the ultrasound transducer. Therefore each shot shows the cardiac structure from a particular angle, and therefore it has a distinct characteristic from the other shots. From now on we call each shot in the echocardiogram video a *view* because through each shot we view a collection of different cardiac objects, or the same objects from a different angle. To distinguish between the shots in a video, one usually looks for the occurrence of discontinuity of some selected feature of the frames or a collection of such features with time. The type of change we are interested in is what is known as *cut* or abrupt change in the more common types of videos.

There are many algorithms available for shot boundary detection[2,3,4,5,6]. Each of those algorithms has its strength and weaknesses[7]. These techniques generally exploit the differences in feature values across two or more adjacent frames in a sequence. Later in this work we propose a method for view boundary detection in the echocardiograms based on the shape of the ROI. This method can be regarded as a special case of the edge tracking algorithm[6], because the change in the shape of the ROI is essentially the same as the change in the location of the edges.

The other component of video summarization is the selection of one or more frames to represent each video segment in the final summary. Different criteria have guided the selection of key-frames as reported in the literature[2,8,9]. The most intuitive

and primitive method for selecting key-frames is to choose the n-the frame of a video segment. Wolf[8] has suggested selecting the key-frames from the local minima of motion in a shot. Kender et al.[9] have suggested that the best representative frames in the home videos are the frames that are immediately after *motion and stop* operation of the camera, where motion refers to either pan or zoom.

The selection of the most informative key-frames for a video shot highly depends on the content of the video. Therefore different types of videos require different key-frame selection criteria to best capture the essence of the video. In the case of the echocardiogram videos we indirectly take advantage of the periodic structure of the cardiac motion. The method as described later in section (3) selects the key-frames corresponding to the local extrema of the cardiac periodic expansive/contractive motion in an indirect way. The time at which the cardiac motion changes from expansive to contractive corresponds to the end-diastole and the time at which the motion changes from contractive to expansive corresponds to end-systole. We use the ECG[10] that is available at the bottom of each frame of the echocardiogram video to detect the two extremes of the cardiac motion.

Having segmented the echocardiograms into their constituent views, and having selected the representative frames for each view, we construct a summary of the echo video. Two types of summaries can be imagined for the echocardiograms. The first type is what is called a *static summary*, which is essentially a collection of key-frames in the form of a storyboard. This type of the summary is very useful for browsing the content of the echo video. In fact the display space has been traded for time in this type of summary to provide a rapid overview of the entire echo. Using the static summary the user is provided by an efficient tool for randomly accessing the views of the echocardiogram, as opposed to the traditional way of using the VCR navigation tools (FF, RWD) to search the entire echo to find the view of his/her interest. The other type of summary is the *dynamic summary* or what is referred to as *video skim*[11]. This type of summary is called the *clinical summary* among the clinicians. The dynamic summary is a concatenation of small segments of the videos. Essentially from a clinical point of view one cycle (R-R segment) of the heart motion is sufficient to see all the different cardiac objects and their dynamics in any particular view. In the dynamic summary we extract one (or more based on the preference of the user or network conditions for transmission purposes) R-R section of each view and join the sections from the different views to construct the summary. This type of summary can be very useful for tele-echocardiology applications, or even for a rapid viewing of the entire echo before moving on to a more careful examination of the echo video.

The following section explains the temporal segmentation of the echo videos into different views. Section (3) covers the issue of key-frame selection for echo videos. In section (4) we will see an example of a static summary of an echo video, which can be very useful for browsing the content of such videos. Finally in section (5) we will conclude and raise some issues for future research directions.

2. ECHO VIDEO SEGMENTATION

As was mentioned before the shape of the ROI and color information is used to detect view changes in the echocardiograms. The shape of the ROI in each frame depends on the type of the echo and the spatial structure of the echocardiogram frames in each type. Therefore before starting the discussion on the temporal segmentation of echocardiogram videos, we would like to make some notes on the spatio-temporal structure of these videos.

An ultrasound interrogation of the cardiac structure is consisted of different forms of echocardiography such as: M-mode echo, 2D echo, and doppler echo. Among these different forms, 2D echocardiography is the backbone of cardiac ultrasound. Only after first having a reference from the 2D echocardiogram one obtains Doppler and M-mode echocardiograms[1]. Therefore the standard echocardiographic views were chosen from the 2D echocardiograms. These standard views were selected to visualize the chambers and valves in planes that lie orthogonal to their tissue- blood interfaces.

Echocardiogram videos have a certain temporal structure. In an echo study of a patient, the sonographer follows a sequence of transducer locations and angles. These correspond to a set of predefined 2D views. Occasionally for more accurate results the sonographer may zoom-in to a particular object or take the color Doppler echo for that view. These give rise to a probabilistic *transition graph* (Figure 2). Starting from a 2D state (the oval shapes in Figure 2), the echo video can proceed to a color Doppler, a zoom-in, an M-mode Doppler or to the next 2D view in the sequence with a certain probability for each transition. The transition graph is almost the same in different health care centers. This is due to the fact that in order to evaluate the cardiac condition, the clinician needs to view certain cardiac objects, which are only viewable from certain angles and locations of the transducer.

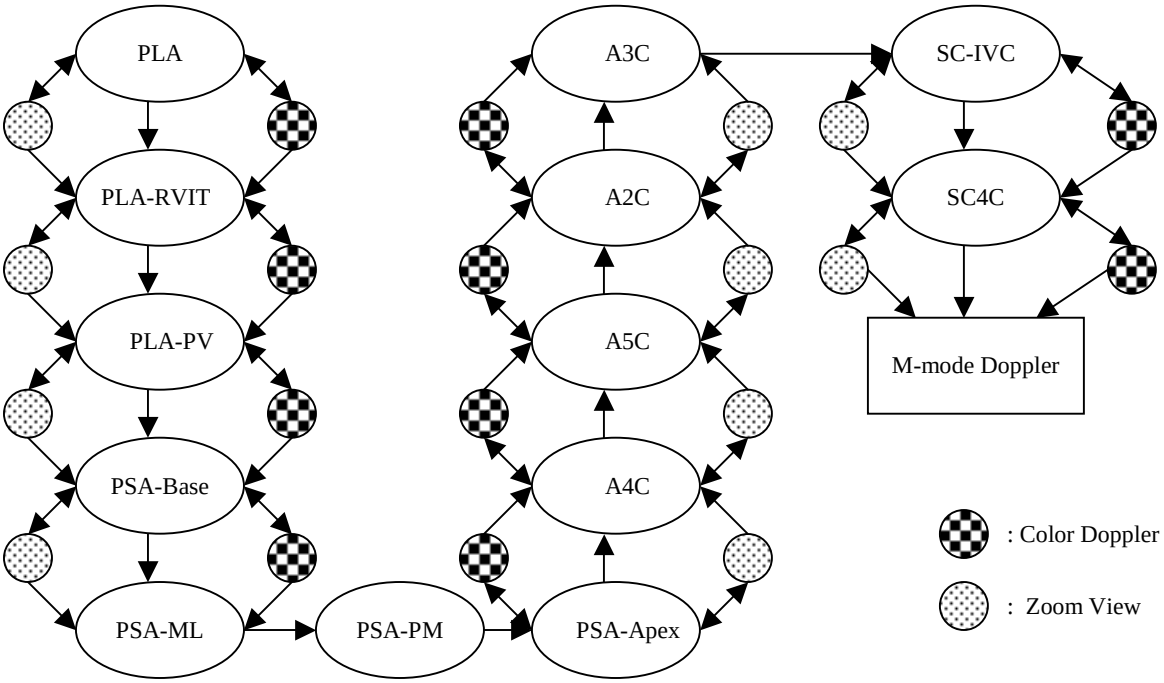


Figure 2. Echocardiogram View Transition Graph

Each frame of the echo video; depending on which state of the transition graph it belongs, has a certain spatial structure. More specifically, the shape of the ROI; which is the region that the ultrasound image appears in the frame (Figure 3); depends on the type of the echo and therefore on the state of the frame. For 2D echo, the ROI is triangular as is shown in Figure 3. For the zoom-in views the ROI is either trapezoidal or square. The M-mode Doppler view has a rectangular ROI. The ROI in the frames belonging to a color Doppler view are also triangular but have an extra distinguishing factor, which is the presence of the color overlay that shows the blood flow pattern through the cardiac valves.

The objective is to find the boundaries of the views or the times of transition from one state to the other in the *transition graph.* In order to reach this goal we exploit the spatio-temporal structure of the echo videos as mentioned above. For each frame of the echocardiogram video we extract features such as the color content and the shape of the ROI in the frame. At the time instances that either the shape of the ROI or the color content of the frame or both change from one frame to the other.

Each frame of the echo video is segmented into foreground objects and background. Foreground consists of the ROI, the text area the gray-scale bar that shows the dynamic range of the ultrasound machine and the ECG signal at the bottom of the frame (Figure 3). This is done by first segmenting the entire frame into five different levels of gray (this number is found heuristically) by using the K-means clustering algorithm and then merging the small regions together to form the foreground. Using morphological dilation we eliminate the discontinuities in the foreground objects. Hough transform for detection of lines is used on the edge map to find the combination of straight lines, which define the shape of the ROI. The process is shown in Figure 4. In addition to the shape of the ROI, the presence/absence of color is also evaluated for each frame to help in distinguishing between the color Doppler views and the others.

This is an effective approach for echo video segmentation due to the fact that in the *transition graph* (Figure 2), the probability of transition from a 2D echo view directly to another 2D echo is very small. The frames of a 2D view, with high probability, go through the zoom and color Doppler states before transiting to the next 2D echo state. Sometimes there is also a sequence of blank frames separating two views from each other. This is due to the fact that the sonographer has to turn off the recording of the echo video while repositioning the transducer on the patient's chest in order to get the best view before resuming. The method as described above fails in the case of the *Parasternal Short Axis*[1] (PSA) views, which follow one another in the *view transition graph* without any other type of frames or any blank frames in between. At this stage we do not

have any solution to this problem, which is threshold independent and fully automatic. We hope to solve this problem when we address the issue of the view recognition in the future.

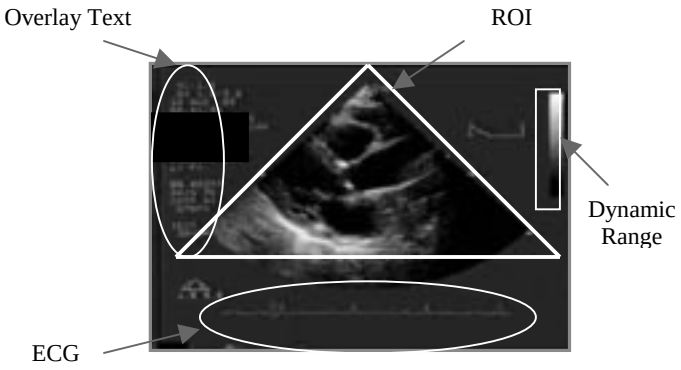


Figure 3. Spatial Structure of the Echocardiogram Frames

The view boundary detection algorithm as described above, accurately detects any change in the shape of the ROI or any change in the color content of the frames. This algorithm does not depend on any sort of thresholds and therefore is fully automatic. Overall it reaches on average %97 accuracy[*] in detecting all the view boundaries in a complete echo study. The failure rate is due to the sequence of the PSA views. Obviously the process of view change detection is slow due to the fact that each frame of the echo video has to be processed for extracting the shape and color information. To speed up the process we first select the key-frames as is described in the next section and then examine if there is a view transition between any two consecutive key-frames. This will result in substantial increase (almost 30 fold) in the speed of the boundary change detection.

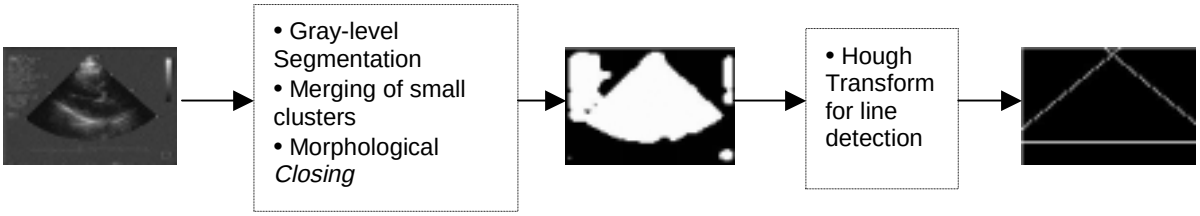
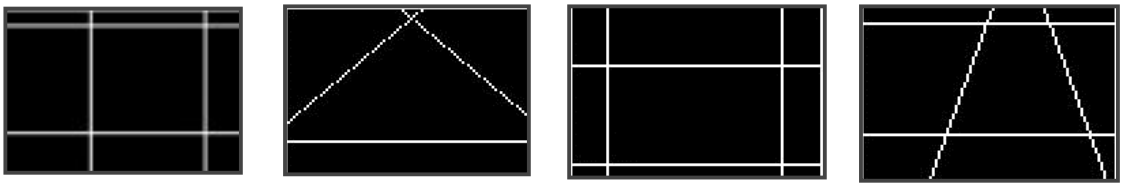


Figure 4. ROI shape extraction (top), Different types of ROI shapes, from left: square for zoom type,

Triangular for 2D, rectangular for M-mode, and trapezoidal for zoom. (bottom)



[*] This rate for accuracy is for an echo study where each 2D echo view is followed by a zoom and a color view before transiting to the next 2D view, except for the PSA views. There is a minor difference between the *view transition graphs* from one health care center to the other and therefore we can expect to get almost the same results for all echo videos.

3. KEY-FRAME SELECTION

Heart has a repetitive motion with a period equal to the duration of one heartbeat. For an echocardiogram encoded in 30 frames per second the duration of a heart beat is roughly between 15 to 30 frames depending on the subject of the study. During one cardiac cycle, heart goes through two phases. In one phase heart expands and is called the *diastole*[1], and in the next phase heart contracts, which is called *systole*[1]. The frames corresponding to the final portion of each phase represent the heart motion in its two extremes. These two extreme states are important from a clinical point of view because in certain views (e.g. PSA views) the clinicians can compare the status of the cardiac objects in the two states to determine certain abnormalities. The frames corresponding to the end-diastole of the heart have and additional importance because heart is most expanded and the cardiac objects have a good visibility.

Therefore in each view the frames corresponding to the end-diastole and the end-systole of the heart cycle can be chosen as the representative frames. Although these two frames by themselves do not capture the dynamics of the heart, but according to the experts in echocardiography they provide good visual cues for browsing the content of the digital echocardiograms. In this work we only choose the end-diastole frame as the representative frame of each view. Since as mentioned before, heart has a repetitive motion, there are several end-diastole frames per view. Therefore we only need to take one of them to represent the view.

Heart has a three dimensional motion in space, which in addition to translation and expansion/contraction it also twists during the motion. Following the motion of the heart in order to determine the extremum points of it is a very difficult job if not impossible, since the heart motion can not be formulated. Hence in order to find the frames corresponding to the end-diastole, we take advantage of the ECG that is available at the bottom of each frame of the echocardiogram video. Figure 5 shows the location of the end-diastole frame on the ECG. The end-diastole as we can see from the figure occur after the peak of the R-wave. Therefore in order to find the key-frames for each view, using the ECG we find the frames corresponding to the peaks of the R-wave.

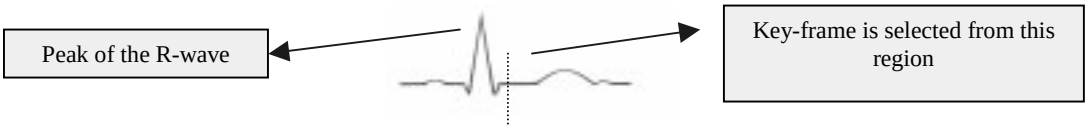


Figure 5. Location of the R-wave peak, and the key-frame

To find the location of the R-wave peak, we process the ECG to extract the location of the *time marker* (see Figure 6) and also to extract the number of peaks of the R-wave to the left of the time marker. Whenever the time marker crosses a R-wave peak, there will be a one-unit increase in the number of the R-wave peaks that are to the left of the time marker. Therefore for each frame, we calculate the number of left peaks and when there is an increase in that number we declare a R-wave peak detected. The reason to look only for the left-side R-wave peaks is that the time marker sweeps the ECG from left to right and the R-wave peaks to the right of it are old data, which belong to the previous sweep.

For finding the location of the time marker and the R-wave peaks we use morphology with special structuring elements. Figure 6 demonstrates the sequence of operations. Eroding the ECG in order to find the locations of the R-wave peaks will result in more than one point for each peak. We use K-means clustering algorithm to cluster these points together and take the mean of each cluster to represent the peaks. To cluster the points we use the following algorithm:

- Initialize the number of clusters to *N* the total number of points
- Find the distances between any two clusters
- If *(new_distance – old_distance) / new_distance > TH* stop, otherwise:
    - Merge clusters together and set its center to the mean of the two clusters
    - Set number of clusters to *(N – 1)*
    - Goto the 2-nd step

The value of the threshold (*TH*) is set by considering the typical intra and inter-cluster distances. The inter-cluster distance is equal to the distance from one R-wave peak to the other, and the maximum intra-cluster distance is the diameter of the circle enclosing all the points representing one R-wave peak.

The performance of this method for key-frame extraction in echocardiogram videos highly depends on the quality of the ECG. For a good quality ECG, the performance of the system as described in terms of the standard measures for information retrieval is as follows: (the values are taken from experiments on a few echo videos and do not represent a large collection of echo videos)

Precision = Number of correctly found R-wave peaks / Total number of peaks that were found = 65%

Recall = Number of correctly found R-wave peaks / Total number of R-wave peaks in the ECG = 100%

These measures mean that all of the R-wave peaks of the ECG have been detected, but almost 35% of the detected points do not correspond to the R-wave peaks. This is due to the irregularities in the appearance of the ECG from one frame to another. Since morphological operations are based on the shape and size of the structuring elements, any small change in the ECG image that is comparable to the size of the structuring element will affect the results. To improve the precision of the algorithm, we look into the neighborhood of the detected R-wave peaks. In a neighborhood defined by a window of length equal to *w*, we measure the number of detected peaks. If this number is more than 2 it means we have two or more R-wave peaks that have happened very close to each other in an interval of length *w*, so we know that this interval contains an invalid peak and therefore we delete the peak. The length *w* is chosen to be small with respect to the R-R distance. For a R-R distance of 24 frames we choose *w* to be equal to 5. After filtering the output of the R-wave peak detector, we get the following results: *Precision* = 94%, and *Recall* =88%. Obviously precision has improved a lot at the expense of recall. It is more desirable to have a high precision than recall because we want to have the correct R-wave peak locations to be sure that the selected key-frames are actually taken from the end-diastole of the cardiac cycle. Due to the repetitive nature of the heart motion, the missed R-wave peaks can be tolerated, because any R-wave peak in a view is good as the other ones in representing that view.

It is worth mentioning that the ECG that is available at the bottom of each frame does not have clinical value, and is only useful as a visual reference for timing purposes. Usually clinicians examine the 12-lead ECG to find out about the patient's cardiac health.

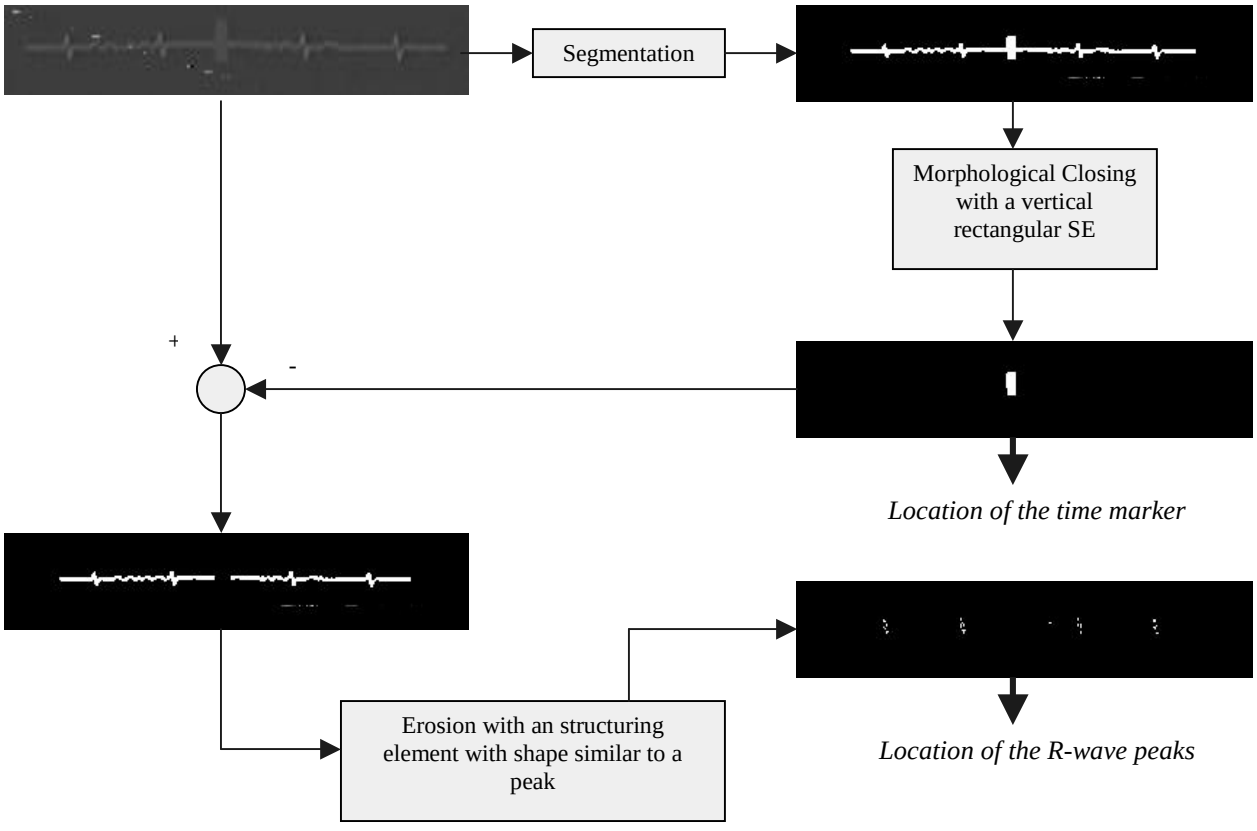


Figure 6. Extraction of the R-wave peaks and the location of the time marker for a frame of echo video

4. SUMMARIZATION

As was mentioned in the introduction part of this correspondence, we can imagine two types of summaries for the echocardiogram videos. In the static summary (Figure 7) every view of the echo video, is represented by a key-frame. This highly facilitates the browsing of the echocardiogram video. By selecting any of the images in the summary, the user would be able to view the video segment of that view only. Therefore instead of watching the entire video and using the serial navigation tools of an analog VCR, the clinician has random access to the different views.

To construct the dynamic summary, the sequence of frames between two consecutive R-wave peaks are extracted from each view and then concatenated together to form the summary. One can create dynamic summaries with different duration, or even different content. For example, since the 2D echo views are clinically the most important views, one can only create a summary that is consisted of 2D views only. Or even based on the preference of the user, the summary can have different length and instead of one R-R cycle for each view, one can include several R-R cycles in the summary. The dynamic summary is also named *clinical summary* because each R-R cycle of a view of the heart has sufficient information about that view.

5. CONCLUSION AND FUTURE WORK

In this paper we demonstrated the application of video summarization to the field of echocardiogram videos. With the advent of digital echocardiography machines and the gradual transition of health care centers to become fully digital in terms of managing their multimedia data, systems such as the one mentioned here are becoming necessary. However, there is still much work to be done before such systems could be useful in the health care centers. The most important work would be the automatic recognition of the views of the echocardiogram. This will enable us to fully annotate the archive of echocardiograms and also link the textual information that are usually available in the form of diagnosis report or patient information to the echo videos. Such archives of the echocardiogram videos and their related data could be integrated and added to the PACS system in the hospitals.

In order to incorporate digital echocardiogram videos into the PACS system (Cardio-PACS), one has to construct proper *information objects*[12] for the description of echo videos in the format of the DICOM standard, which is the standard that addresses the network exchange and media storage of medical images and videos in the PACS environment. The *information object* describing the echo video contains information about the different views and the position and orientation of the transducer in each view and the specific anatomic structure being imaged in an echo study. Therefore to properly integrate the digital echo videos with PACS one needs to automatically extract such information from the content of the captured echocardiograms[**]. Using the algorithms presented in this paper and adding the view recognition results, we would be able to construct such *information objects*.

REFERENCES

1. H. Feigenbaum, *Echocardiography*, LEA & FEBIGER, 1993.
2. B. Shahraray, "Scene Change Detection and Content-Based Sampling of Video Sequences", *Proc. SPIE*, vol. 2419, pp. 2-13, 1995.
3. F. Arman, A. Hsu, and M. Y. Chiu, "Image Processing on Compressed Data for Large Video Databases", *Proc. ACM Multimedia 93*, pp.267-272, August 1993.
4. H. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic Partitioning of Full-Motion Video", *Multimedia Systems*, 1(1), pp. 10-28, 1993.
5. H. Zhang, C. Y. Low, and S. W. Smoliar, "Video Parsing and Browsing Using Compressed Data", *Multimedia Tools and Applications*, 1, pp. 89-111, 1995.
6. R. Zabih, J. Miller, and K. Mai, "A Feature-Based Algorithm for Detecting and Classifying Scene Breaks", *Proc. ACM Multimedia 95*, San Fransisco, CA, pp. 189-200, November 1993.

[**] One can argue that the sonographer can key in such information during the acquisition phase. Although the new echo machines have such capability, such annotations are not made. This can't be done for the digital echo videos, which are converted to digital.

7. J. S. Boreczky and L. A. Rowe, “Comparison of Video Shot Boundary Detection Techniques”, *Storage and Retrieval for Image and Video Databases IV, Proc. Of IS&T/SPIE 1996 Int’l Symp. on Elec. Imaging: Science and Technology*, San Jose, CA., February 1996.
8. W. Wolf, “Key Frame Selection by Motion Analysis”, *Proc. ICASSP’96*, vol. 2, pp. 1228-1231, 1996.
9. J. Kender, “On the Structure and Analysis of Home Videos”, *Proc. ACCV*, 2000.
10. A. Bayes de Luna, *Clinical Electrochardiography: a text book*, Futura Pub., Armonk NY., 1998.
11. M. Smith, and T. Kanade, “Video Skimming for Quick Browsing Based on Audio and Image Characterization”, *Carnegie Mellon University, School of Computer Science Technical Report CMU-CS-95-186R*, Pittsburgh, PA, May 1996.
12. J. D. Thomas,”The Dicom Image Formatting Standard: What It Means for Echocardiographers”, *J. AM. SOC. ECHOCARDIOGR*, vol.8, pp. 319-327, 1995.

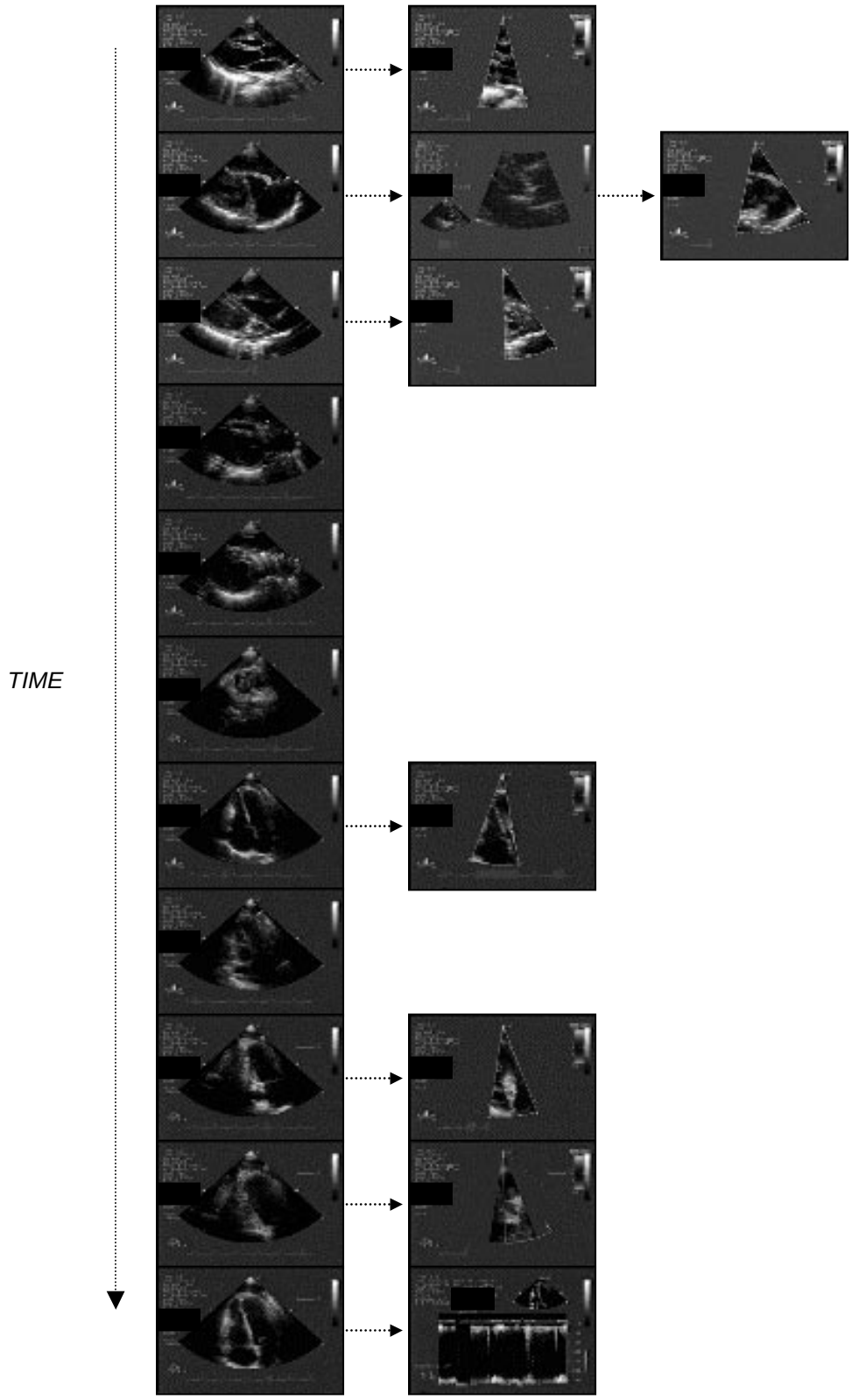


Figure 7. A typical static summary of an echocardiogram video