Modeling of Echocardiogram Video Based on Views and States

Aditi Roy[1], Shamik Sural[1], J. Mukherjee[2], and A.K. Majumdar[2]

[1] School of Information Technology
[2] Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur, India
{aditi.roy@sit, shamik@sit, joy@cse, akmj@cse}.iitkgp.ernet.in

Abstract. In this work we propose a hierarchical state-based model for representing an echocardiogram video using objects present and their dynamic behavior. The modeling is done on the basis of the different types of views like short axis view, long axis view, apical view, etc. For view classification, an artificial neural network is trained with the histogram of a *'region of interest'* of each video frame. A state transition diagram is used to represent the states of objects in different views and corresponding transition from one state to another. States are detected with the help of synthetic M-mode images. In contrast to traditional single M-mode approach, we propose a new approach named as '*Sweep M-mode*' for the detection of states.

1 Introduction

In the last few decades, medical imaging research has seen a rapid progress. Echocardiography is a common diagnostic imaging technique that uses ultrasound to analyze cardiac structures and function [1]. Present work is motivated by the grow-ing interest in managing medical image/video collections based on their content.

Some systems and standards such as PACS [2] and DICOM [3] are used in medical imaging centers to digitize, view, communicate and store medical images. How-ever, these do not take into account the characteristics of the content of the medical images or videos. In recent past, advances have been made in content based retrieval of medical images [4]. Research has also been done on the extraction of cardiac object boundaries from sequences of echocardiographic images [5]. Work on echo-cardiographic video summarization, temporal segmentation for interpretation, storage and content based retrieval of echo video has been reported [6]. This process is heavily dependent on the available domain knowledge which includes spatial structure of the echo video frames in terms of the '*Region of Interest' (ROI)*, where an ROI is the region in a frame containing only the echo image of heart. On the other hand, an approach towards semantic content based retrieval of video data using object state transition data model has been put forward in [7][8]. In these articles, the echo videos are segmented based on states of the heart object. A view-based model-ing approach has been reported in

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 397–408, 2006.

[9], which uses parts based representation for automatic view recognition. They represent the structure of heart by a constellation of its parts (chambers) under the different views. Statistical variations of the parts in the constellation and their spatial relationships are modeled using Markov Random Field. Support Vector Machine [SVM] is used for view recognition which fuses the assessments of a test image by all the view-models. A state based modeling approach [10] measures the relative changes in left ventricular cavity in echo video sequences to identify end diastolic and end systolic frames. This information is then used in conjunction with the statistical correlation between echo video frames to extract information about systole and diastole states of heart. Thus, view-based modeling and state-based modeling of echo video are done separately. But hierarchical state-based modeling, combining views and states, is a new problem which has not been addressed till now, to the best of our knowledge.

In our work, we segment an echo video hierarchically based on views and states of the heart object by exploiting specific structures of the video. The advantage of using this approach is that it allows storage and indexing of the echo video at different levels of abstraction based on semantic features of video objects.

For hierarchical state-based modeling we first segment the video based on views. To detect view boundary, we use traditional color histogram based comparison [11] and edge change ratio [12]. After detecting shot boundary, we apply a novel technique for automatic view classification of each shot which is based on the signal properties and their statistical variations for each view in echo video. We train an artificial neural network [17] with the histogram of 'region of interest' of each video frame for classification. For state detection, we propose a new method using single and sweep M-Mode.

The rest of the paper is organized as follows. In Section 2, we discuss echo video segmentation techniques based on views. First, we give a brief description of the shot detection techniques used for detecting various views. Then we introduce our method for view classification. In Section 3, we discuss state detection using single and sweep M-mode and finally, we conclude in Section 4 of the paper.

2 Echocardiogram Video Segmentation Based on View

For browsing and content based manipulations of echo video, the visual information must be structured and broken down into meaningful components. Shots are the basic structural building blocks for this and shot boundaries need to be determined without user intervention. In Figure 1, we show the basic block diagram of the system. After detecting each shot, we classify them using our classifier. Then, each view is further segmented based on the states of the heart, as specified by its state transition diagram.

2.1 Shot Detection

As mentioned above, the first step of video processing for hierarchical state-based modeling is segmentation of the input video into shots. A shot can be defined as a sequence of interrelated frames captured from the same camera

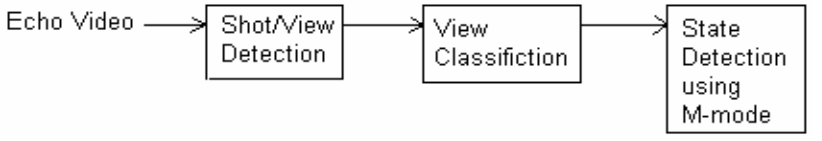


Fig. 1. Block diagram of the system for hierarchical state-based modeling of echo video

location that represents a continuous action in time and space. In echo video segmentation, traditional definition of shot is not applicable. An echo video is obtained by scanning the cardiac structure with an ultrasound device. Hence, depending on the location of the transducer, different views of echocardiogram video are obtained.

Echocardiogram Shot Definition. Echocardiogram images are used by cardiologists to analyze physiological and functional behaviors of cardiovascular components, e.g. heart chambers, valves, arteries, etc. In this process, the heart chamber can be viewed through different angles by changing the transducer position. Accordingly, the same chamber can be seen in different perspectives, known as *views*. In [6], four modes of echocardiography are identified i.e. two dimensional, Doppler, color Doppler, zoom-in. They define '*view*' as the sequence of frames corresponding to a single transducer location and mode of imaging.

But in this paper we use a different definition of view as mentioned in [1]. The views considered are:

1) Parasternal Long Axis View (LAX): Transducer placed parallel to the long axis of left ventricle and the ultrasound wave passes through the center of left ventricle chamber.
2) Parasternal Short Axis view (SAX): Transducer is rotated 90 in clockwise direction from the parasternal long axis view position.
3) Apical View: Transducer is placed in the cardiac apex.
4) Color Doppler: This uses color overlays on the reference frame sequence to show the blood flow in the heart based on the Doppler effect.
5) One dimensional: This is the gray-scale reference frame sequence of one dimensional signal locating anatomic structures from their echoes along a fixed axis of emission.

In this paper, we use the terms '*view*' and '*shot*' interchangeably. Transition from one view to another is always sharp.

Echocardiogram Shot Detection. Various automatic shot boundary detection algorithms for videos like movie, news, sports, etc., have been proposed in the literature. Due to the presence of high speckle noise in echo video, it is difficult to detect views in echo videos by applying these algorithms. We explore two methods to detect shot boundary, namely, histogram based comparison and edge change ratio. The main idea of these techniques is that if the difference between

two consecutive frames is larger than a threshold value, then a shot transition is assumed to exist at that frame position.

The first approach is global histogram based comparison method [14][15]. Here we compute color histogram for each frame using 192 color bins, where each bin contains the percentage of pixels from the whole frame. Color histograms of two consecutive frames are compared using a cosine similarity metric. When the similarity value is below the threshold, a shot boundary is detected. The main drawback of this method is that, if two image frames belonging to different shots have similar histograms, then the shot boundary may not get detected.

The second approach is based on edge change ratio. Here each frame is first turned into gray scale image and then the edges are detected. We use Sobel operator due to its smoothing effect which is important for noisy echocardiogram video. Then for two consecutive frames, edge ratio is computed in terms of the number of new edge pixels entering the frame and the number of old edge pixels leaving the frame[12][16]. Exit-ing pixels are identified by keeping pixels in the first frame but not the second, and the entering pixels are identified by keeping pixels in the second frame and not in the first. Using these results, the edge change ratio (ECR) is determined as follows.

$$ECR_i = MAX(\frac{E_i^{in}}{E_i}, \frac{E_{i-1}^{out}}{E_{i-1}}) \tag{1}$$

Here for the i^{th} frame, E_i^{in} is the number of entering edge pixels, E_i^{out} is the number of exiting edge pixels, and is E_i the total number of edge pixels. When the edge change ratio exceeds a threshold, a shot boundary is considered to exist. A global motion compensation based on Hausdroff distance is performed before the calculation of the ECR.

Between the two techniques described above, edge based method outperforms histogram based approach. The only drawback is that it cannot detect shot transition between apical view and color Doppler. In such situation, color histogram based comparison gives desired result. We combine these two methods using majority voting to detect shots in echocardiogram video and obtain 98% accuracy.

2.2 View Recognition

Automatic view recognition in echo video is a challenging task due to the presence of multiplicative noise and structural similarity among the constellations of the different views. Variations in the images captured under the same view but for different patients, make the problem even more difficult.

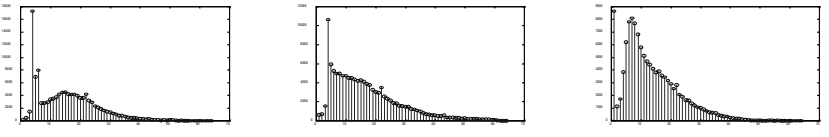
Here we classify three types of views i.e., long axis view, short axis view and apical view. Color Doppler view is classified at the time of detecting shots by the presence of color. One dimensional view is also identified during shot detection from their static nature.

For view classification we use the fact that, for each view, different sets of cardiac chambers are visible. The number of chambers present, their orientation

and the presence of heart muscles in each view, gives different patterns of histogram. We identify the views based on their unique histogram patterns. In our approach, we first define a '*region of interest*' for each frame to minimize the effect of noisy background. The ROI region is selected after performing experiments on a large number of different types of echo videos containing all the view types. The ROIs marked on three representative frames, one from each view, are shown in Figure 2. Next we generate a gray scale histogram for this ROI using 64 bins. For each view, the histogram pattern is unique as shown in Figure 3. We use a neural network [17] for classifying the views from the 64-dimensional normalized histogram vector of each frame.

(a) (b) (c)

Fig. 2. Frames with ROI: (a) Short Axis View; (b) Long Axis View; (c) Apical View



(a) (b) (c)

Fig. 3. Histogram of (a) Short Axis View, (b) Long Axis View, (c) Apical View for Fig. 2

We train the neural network with a total of 1260 frames, 400 each of short axis view and long axis view, and 460 of apical view. Every frame in the data set is manually labeled. We use a multilayer perceptron (MLP) with one input layer, one hidden layer and one output layer. The number of units in the input layer is 64, one for each histogram component. The number of units in the output layer is 3, one to represent long axis view, one for short axis view and one for apical view. The number of units in the hidden layer is empirically chosen as 80.

Table 1. View recognition results

True Class	Predicted Class		
	Short Axis View (no.of frames)	Long Axis View (no. of frames)	Apical View (no. of frames)
Short Axis View	128	2	1
Long Axis View	1	80	1
Apical View	3	12	140

We evaluated the performance of the classifier on a test data set of 365 frames. Table 1 shows view classification result in the form of a confusion matrix. An overall precision of 95.34% is obtained. The main source of misclassification is incorrect recognition of apical view frames as long axis view frames.

3 State Based Modeling of Echocardiogram Video

A state based video model is a means for extracting information contained in an un-structured video data and representing this information in order to support users' queries. A state stores information about the past, i.e. it reflects the input changes from the system start to the present moment. A transition indicates a state change and is described by a condition that needs to be fulfilled to enable transition. Action is an activity that is to be performed at a given moment. State transition diagram describes all the states that an object can have, the events or conditions under which an object changes state (transitions) and the activities undertaken during the life of an object (actions). In an echocardiogram video, the two states are *systole* and *diastole*.

Systole: During systole, the ventricles contract. The aortic and pulmonary valves open and blood is forcibly ejected from the ventricles into the pulmonary artery to be re-oxygenated in the lungs, and into the aorta for systemic distribution of oxygenated blood. At the same time, the mitral and tricuspid valves close to prevent backflow and the atria start to fill with blood again.

Diastole: During diastole, the ventricles relax. The pulmonary and aortic valves close and the mitral and tricuspid valves open. The ventricles then start to fill with blood again.

Figure 4 shows the state transition diagram of heart, where the two states are systole and diastole. When ventricles start expanding, transition from systole to diastole occurs. Similarly, transition occurs from diastole to systole with ventricular contraction.

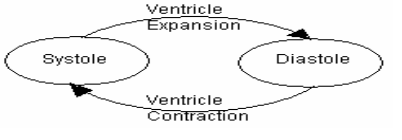


Fig. 4. State transition diagram of heart

3.1 Identification of States Using Single M-Mode

The signal in M-mode or motion mode echocardiography is obtained from the time sequence of a one-dimensional signal locating anatomic structures from their echoes along a fixed axis of emission. These measurements are usually represented as an image (see Figure 5(b)) in which the abscissa and the ordinate represent time and depth (or distance), respectively. The intensity of each pixel is a function of the reflected ultrasound energy. The cardiac borders have a

periodic translational movement with time. Due to high level of noise, automated detection of these borders is not a trivial task to be solved by standard image processing techniques.

To obtain synthetic M-mode from 2-D echocardiogram, we use short axis view. User draws a straight line perpendicular to the walls of left ventricle where the wall motion is maximum. M-mode can be of different types depending on the position of the straight line. If the line is vertical, then the generated M-mode is termed as vertical M-mode. If the line is horizontal or diagonal, the M-mode is named accordingly. In general, only vertical M-modes are computed for diagnosis. But it is observed that for some views, horizontal M-mode gives more useful information than vertical M-mode. We, therefore, compute horizontal M-mode to get state information more accurately. Figure 5(a) shows the horizontal line drawn on a short axis view frame. To compute M-mode image, we scan along this line for each frame in the short axis view segment of the video. The intensity value along the straight line is taken as ordinate and frame number is taken as abscissa as shown in Figure 5(b).

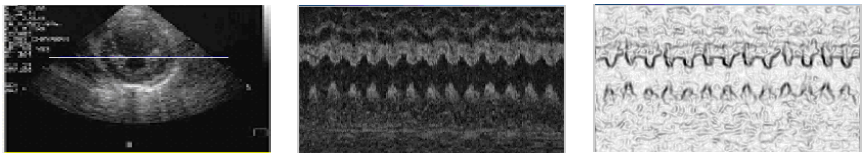
Pre-processing. Most echocardiograms have a relatively high noise level because of intrinsic limitation in the measurement device. Substantial noise reduction with minimal information loss is achieved by smoothing the image. We first use Gaussian filtering to remove noise from M-mode images and then apply Sobel operator for edge detection.

1) *Smoothing:* We use a convolution kernel that approximates a Gaussian with of 0.45. To compute a Gaussian smoothing with a large standard deviation, we convolve the image three times with a smaller Gaussian kernel [77]. The Gaussian outputs a 'weighted average' of each pixel's neighborhood, with the average weighted more towards the value of the central pixels. Because of this, a Gaussian provides gentler smoothing and preserves edges better than other types of smoothing filter. Thus it helps in getting better edge detection result.

2) *Edge Detection:* For edge detection, we use Sobel operator mask because of its higher noise suppression characteristics than other edge detection operators.

The resulting M-mode image after applying Gaussian and Sobel operator on the M-mode image of Figure 5(b) is shown in Figure 5(c).

Border Extraction. Border extraction is one of the most important steps in processing an echo video. Difficulties arise from the fact that the observed time



(a) (b) (c)

Fig. 5. (a) User drawn line on a chosen frame of video, (b) M-mode image, (c) Edge detected smoothed M-mode image

trajectories of cardiac borders sometimes present discontinuities and may not always correspond to well-defined edges. Here we extract the border by searching for optimal path along time axis. We use a maximum tracking procedure [13] whose performance is improved by using a local model to predict the position of the next border point.

The system described by Unser *et al.*[13] is based on a search algorithm that traces gray-scale maxima corresponding to each relevant cardiac internal structure along the horizontal time axis. Although this algorithm was initially designed to be applied to the data directly, it can also be used in our system in which tracking of cardiac border is based on the position of *minimum intensity* pixel on the edge detected M-mode image. A digitized M-mode echocardiogram is represented as a two dimensional image $\{X_{k,l}\}$, where 'k' represents the time variable and 'l' represents the distance along the axis of the straight line drawn on the view. A brief review of the algorithms is presented here.

1) *Basic Algorithm:* The basic procedure is based on the fact that the movement of cardiac borders from one time frame to another is restricted to a relatively narrow region. A starting point is first determined by the user. Then, assuming the present position of the cardiac border to be l, the algorithm searches for the point with mini-mal intensity in the next vertical line in a window centered around the previous position ($l \pm w$), where w is the window size. This point is then taken as the next position of the border. The procedure is iterated until all the frames have been considered. This simple approach follows a single path guided by locally optimizing the sum of the signal values along the trajectory. It usually detects the posterior wall epicardium satisfactorily, but generally fails in detecting the endocardium or the boundaries of the interventricular septum. We use this algorithm to detect the lower border, having lesser movement, shown in Figure 6(a).

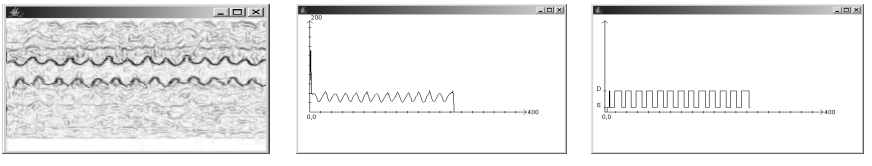
2) *Kuwahara Algorithm (KMTA):* Kuwahara *et al.* suggest searching for the cardiac borders by reference to the border that has already been detected. Their algorithm uses the same principle as the basic algorithm except that the position and extent of the search window at a given time k is also a function of the relative displacement of the reference structure $\triangle l_r(K) = l_r(K) - l_r(K-1)$, where l_r is the reference cardiac border position. The reference position is now given by $l(K-1) + \triangle l_r(K)$, where $l(K-1)$ denotes the previously detected position of the structure. Furthermore, the width of the search window is increased in an asymmetric way, depending on the sign of $\triangle l(K)$. Broadening the search window in the expected direction of movement is designed to compensate for the greater velocity of the endocardium in systole and early diastole. We use this algorithm to detect the upper border in Figure 6(a).

State Identification. For identifying the states we first compute the distance between the two endocardium borders in each frame, as shown in Figure 6(a). Figure 6(b) shows the variation of cardiac border distance with respect to time.

In order to obtain state information, we use the cardiac border distance. The bottom end points of Figure 6(b) indicate end systole points and the top

points indicate the end diastole points. During the time elapsed between end systole point to end diastole point, the heart is in diastole state and is in systole state from end diastole point to end systole point. In the diastole state left ventricle expands, thus the distance between the endocardium borders increases. Hence, the slope is positive. Similarly, in diastole, left ventricle contracts and the distance between the endocardium borders decreases which results in negative slope in the distance graph.

Using the distance information, we classify the echocardiogram video frames into two classes, namely, systole and diastole. In the echocardiogram video, the frames corresponding to positive slope are classified as diastolic state frames, while those corresponding to negative slope are classified as systolic. This completes the process of state detection from an echocardiogram video. The state transition graph obtained from Figure 6(b) is shown in Figure 6(c).



(a) (b) (c)

Fig. 6. (a) Border extracted M-mode image, (b) Variation cardiac border distance with time, (c) State transition graph

Table 2 shows the detailed result of state identification from 267 frames in short axis view using single M-mode. This method gives total misclassification error of 26.22%. Most of the misclassified or unclassified frames are those during which state transition occurs. The table also shows Sweep M-mode results as explained in the next sub-section.

3.2 State Identification Using Sweep M-Mode

In order to further improve the state detection accuracy, we propose a multiple M-mode generation method termed as '*Sweep M-mode*'. It is so named because multiple M-modes are generated by scanning the intensity value along a straight line as before, while the straight line is continuously swept by a specified interval in a direction normal to a fixed axis.

To obtain sweep M-mode, user draws a line perpendicular to the direction of left ventricular wall motion (see straight line in Figure 7). Sweep M-modes are created by scanning the intensity value along the straight line perpendicular to this vertical line, taking it as Y-axis, for each frame of the video considered as X-axis. The Sweep M-modes are generated along the horizontal broken straight lines as shown in Figure 7. Now, as explained in Section 3.1, tracking of the cardiac borders is done for each M-mode individually.

To start tracking, initial border points need to be determined. For this user draws the left ventricle cavity border freehand at the same time when he draws

the straight line in order to obtain sweep M-mode (freehand line in Figure 7). The intersection points of the cavity border and perpendicular straight lines are considered as the start-ing points for tracking the cardiac borders. Since border tracking of all the M-modes is not perfect due to inherent noise, we need to select only those few which are meaningful. It has been observed that in M-modes where tracking is perfect, the distance between the two cardiac borders never exceeds a threshold value. But if it does, the M-modes are surely mistracked. We use this as a heuristic to identify the proper M-modes. Thus, we select the most perfectly tracked M-modes as shown in Figure 8(a). The same method is followed for each individual M-mode to extract state information from a single M-mode.

The plot of the distance between two endocardium borders with respect to time for the selected M-modes is shown in Figure 8(b) and the state transition plot is shown in Figure 8(c). It is seen that the individual plots have many false state transitions. So we combine them using majority voting to identify the state information of each frame.

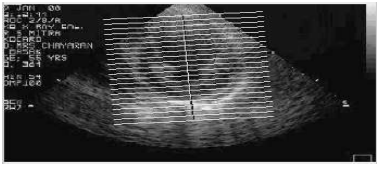


Fig. 7. User drawn straight line and cavity on a chosen frame of video

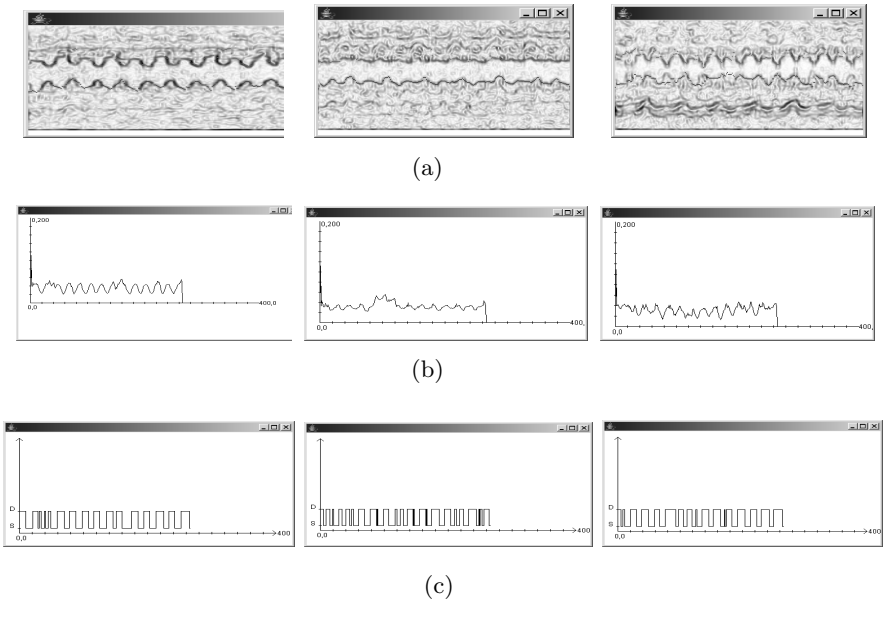


Fig. 8. Three selected M-modes: (a) Border extracted M-mode images, (b) Corresponding cardiac distance graph with time, (c) Corresponding state transition graph

We carried out experiments with the same echocardiogram video as before. The right hand part of Table 2 shows the detailed result obtained using sweep M-mode. It is seen that the misclassification error has come down to about 12.36%. Thus, use of sweep M-mode reduces the misclassification error as compared to single M-mode described in the previous section.

Table 2. Results of state identification using M-mode

True Class	Predicted Class					
	Single M-mode			Sweep M-mode		
	Systole	Diastole	Undetected	Systole	Diastole	Undetected
Systole	96	5	21	111	10	2
Diastole	26	101	18	18	123	3

4 Conclusion

In this paper we have proposed a new approach for hierarchical state-based modeling of echo video data by identifying the views and states of objects. We first detect the view boundaries using histogram based comparison and edge change ratio. Then we classify the views by considering signal properties of different views. Our technique gives a precision rate of 95.34%. To extract state information from each view, we use synthetic M-modes. At first, we apply single M-mode. But the misclassification error in identifying the states with the help of single M-mode is quite high (around 27%). So we introduce a new type of M-mode generation method named as sweep M-mode. Application of sweep M-mode reduces the misclassification error to about 13%. We have used this approach of hierarchical state-based modeling to develop an object relational video database for storage and retrieval of echocardiogram video segments. The proposed scheme is now being extended for finer (sub state) segmentation.

Acknowledgments. This work is partially supported by the Department of Science and Technology (DST), India, under Research Grant No. SR/S3/EECE/ 024/ 2003-SERC-Engg.

References

1. Feigenbaum, H.: *Echocardiography*, LEA & FEBIGER, 1997
2. Huang, H.K.: PACS: *Basic Principles and Applications*, Wiley, New York, 1999
3. DICOM: Http://Medical.nema.org/dicom.html
4. Shyu, C.R., Brodely, C.E., Kak, A.C, Osaka, A.: ASSERT: A physician-in-the loop content-based retrieval system for HRCT image databases. *Computer Vision and Image Understanding*, Vol. 75, July/August, pp. 111-132, 1999
5. Duncan, J.S., Ayache, N.: Medical image analysis: progress over two decades and the challenges ahead. *IEEE Trans. Pattern Analysis and Machine Intelli-gence*, Vol. 22, No. 1, January 2000, pp. 95-105, 2000

6. Ebadollahi, S., Chang, S.F., Wu, H.: Echocardiogram videos: summarization, temporal segmentation and browsing. *Proc. IEEE Int'l Conference on Image Processing*, (ICIP'02), Vol. 1, pp. 613-616, September 2002
7. Sing, P.K., Majumdar, A.K.: Semantic content based retrieval in a video data-base. *Proc. Int'l Workshop on Multimedia Data Mining*, (MDM/KDD'2001), pp. 50-57, August 2001
8. Acharya, B., Mukherjee, J., Majumdar, A.K.: Modeling dynamic objects in databases: a logic based approach. *ER 2001*, *LNCS 2224*, Springer Verlag, pp. 449-512
9. Ebadollahi, S., Chang, S.F., Wu, H.: Automatic view recognition in echocardiogram videos using parts-based representation. Proc. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (CVPR'04), Vol. 2, pp. II-II-9, June 2004
10. Acharya, B., Mukherjee, J., Majumdar, A.K.: An object state transition model for echocardiogram video data. *Proc. Conference of Multimedia Processing Systems*, pp. 103-106, 2000
11. Yu, H., Wolf, W.: A visual search system for video and image database. *Proc. IEEE Int'l Conference on Multimedia Computing and Systems*, pp. 517-524, June1997
12. Zabin, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classifying scene breaks. *Proc. ACM Multimedia*, pp. 189-200, 1995
13. Unser, M., Palle, G., Braun, P., Eden, M.: Automated extraction of serial myocardial borders from M-mode echocardiograms. I*EEE Trans. Medical Imaging*, Vol. 8, pp. 96-103, March 1989
14. Murphy, N., Marlows, S., O'Toole, C., Smeaton, A.: Evaluation of automatic shot boundary detection on a large video test suite. *The Challenges of Image Re-trieval - 2nd UK Conference on Image Retrieval*, 1999
15. Mas, J., Fernandez, G.: Video shot boundary detection based on color histogram. TREC 2003
16. Lienhart, P.: Reliable transition detection in videos: a survey and practioners guide. *Int'l Journal of Image and graphics*, Vol. 1, No. 3, pp. 469-486, 2001
17. Bishop, C.M.: *Neural networks for pattern recognition.* Oxford: Oxford Univer-sity Press.1997