Automatic View Recognition in Echocardiogram Videos using Parts-Based Representation

Shahram Ebadollahi, Shih-Fu Chang
Department of Electrical Engineering
Columbia University
New York, NY
<shahram,sfchang>@ee.columbia.edu

Henry Wu
College of Phyisicians and Suregons
Columbia University
New York, NY
hdw1@columbia.edu

Abstract

Indexing echocardiogram videos at different levels of structure is essential for providing efficient access to their content for browsing and retrieval purposes.

We present a novel approach for the automatic identification of the views of the heart from the content of the echocardiogram videos. In this approach the structure of the heart is represented by the constellation of its parts (chambers) under the different views. The statistical variations of the parts in the constellation and their spatial relationships are modeled using Markov Random Field models. A discriminative method is then used for view recognition which fuses the assessments of a test image by all the view-models.

To the best of our knowledge, this is the first work addressing the analysis of the echocardiogram videos for the purpose of indexing their content. The method presented could be used for multiple-object recognition when the objects are represented by their parts and there are structural similarities between them.

1. Introduction

Echocardiography is a common diagnostic imaging modality that uses ultrasound to capture the structure and function of the heart [7]. A comprehensive evaluation entails imaging the heart in several planes (aspects) by placing the ultrasound transducer at various locations on the patient's chest wall. The recorded image sequence (*echocardiogram video*, or *echo video* for short) therefore displays the 3-dimensional heart from a sequence of different 2-dimensional cross sections (views). Under different views, different sets of cardiac cavities (objects) are visible. The spatial arrangement of those objects is unique to each view. Figure 1 shows the representative images taken from two different views.

Our goal is to automatically index the content of the echocardiogram videos at the view and object levels using the characteristics of the spatial arrangement of the cardiac chambers in the different views.

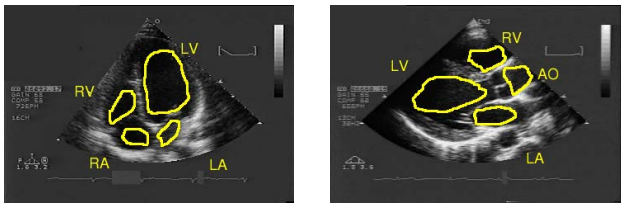

Figure 1. Spatial arrangement of caridac chambers for apical four chamber (left), and parasternal long axis (right) views. (LV: Left Ventricle, LA: Left Atrium, RV: Right Ventricle, RA: Right Atrium, AO: Aorta) [7]

A view-centered modeling approach is employed in which the spatial arrangement of the cardiac cavities and the statistical variations of their properties are modeled for each view. A test image represented by the constellation of its parts (cardiac chambers) is classified into one of the view classes by *fusing* the assessments of it by all the view-models. The fusion framework is employed to resolve ambiguities and make correct recognition. The ambiguities are due to the structural similarities between the constellation of cardiac chambers in the different views of the echo video.

1.1. Challenges

There are several reasons to make the problem of view recognition in echo videos a difficult task. The appearance of the images captured under the same view of the heart for the same patient and among different patients will always be subject to a high degree of variations. This is because of the following two reasons:

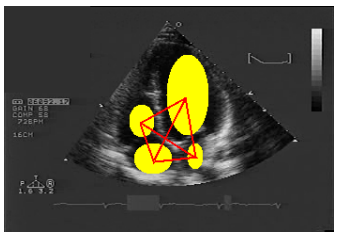

Figure 2. Constellation of caridac cavities for the apical four chamber view [7]. Chambers are represented by the blobs, and the lines indicate the relationships between them.

1. Different patients have slightly different heart structures based on their physical characteristics.
2. There are no natural markers available for placing the ultrasound transducer on the patient body for imaging.

Therefore, the general category of the appearance-based methods which have successfully been applied to some recognition problems [12] can not be employed for the task of view recognition in the echo videos. For this reason, we resort to the category of the part-based approaches. Here, one can model the variations in the properties of the parts, which makes it more suitable for modeling the structure of the heart although it faces its own set of problems.

Because echo videos are the result of the ultrasound interrogation of the structure of the heart, they are highly degraded by multiplicative noise. Therefore, automatic detection of the cardiac cavities results in missed and false cavities which correspond to *occlusion* and *clutter* in the language of object recognition. In addition, there is a great degree of *uncertainty* in the properties of the chambers in the constellation, which is the result of the same artefacts mentioned above.

Even if all of the cardiac cavities could be detected correctly and the false chambers avoided, one still faces the high degree of *structural similarity* among the constellations of the different views. The ambiguities because of the structural similarity between the views will even become worse when one adds the possibility of missing and false chambers to the scenario. In summary, the challenging task is to be able to distinguish between the instances of the different views of the echo video in the presence of occlusion (missing parts), clutter (false parts), uncertainty in the features of the parts, and structural similarity between the different views.

1.2. Prior Work

The part based representation has been widely used in the literature for object recognition. The intuition behind this approach is the fact that an instance of an object could be identified when correct parts are observed in correct spatial configuration [13]. In this type of modeling the features (parts) are distinct and specific detectors for each of the different types of parts are available. The goal is then to choose a set of foreground parts in the correct spatial configuration in order to be able to identify the objects of interest.

The method that we use to model the constellation of cardiac chambers is between the modeling method mentioned above and another category of models which is referred to as *pictorial structures*. In a pictorial structure, the parts are described using appearance models and their constellation is described using a spring-like model [8, 3]. The image is then searched for the optimal position of the parts based on their match with their appearance model and their spatial configuration.

Although all of the above mentioned approaches and similar ones focus on object/clutter discrimination, they are not meant to distinguish between the different objects with high structural similarity. Boshra and Bhanu [2] addressed the issues of recognition performance and its dependence on object model similarity and the effects of distortions such as *occlusion* and *clutter*. They also obtained bounds on the performance of the recognition in the presence of the distortion factors for the case where the uncertainties of the parts properties were expressed using uniform distributions and finding the correspondence between the model and the scene was performed using a voting scheme.

1.3. Our Approach and Contribution

We use a *generic* cardiac chamber detector on the images obtained from the echo videos to locate the chambers. The spatial arrangement and the properties of the cardiac cavities are captured using an *attributed relational structure*. Figure 2 shows the relational structure for the constellation of parts of an image taken from one of the views. The model of the spatial arrangement of the cardiac cavities and the distribution of their properties are also expressed as a relational structure.

Finding the correspondence between the two relational structures expressing the model of a view and that of the observed constellation is posed as a search for the optimal configuration of a *random field* defined on the parts of the observed constellation. The optimal *configuration* is the one that minimizes the overall *posterior energy* of the field.

Markov Random Field [9] is a statistical modelling method in which one could efficiently expresses the contextual constraints using locally defined dependencies between the interacting entities. Through the *Hammersly-Clifford* theorem these local dependencies lead to the encapsulation of the joint global characteristics. *MRF* has been

applied to high-level vision problems such as image interpretation before [9].

In order to disambiguate between the constellations taken from the different views and to be able to recognize the correct view-label for them, we use a discriminative approach. Methods such as Support Vector Machines (SVM) [5] have been shown to perform well in discriminating between the instance of the different classes in general pattern recognition applications.

For each observed constellation, we obtain the optimal labeling of its parts according to each view-model. The vector of the energies assigned to the observed constellation at the optimal configuration of the random field according to the different view-models is used to classify the constellation into the correct view-class. By using this method, as a matter of fact, we fuse the assessments of the observed constellation by the different view-models in order to correctly identify its originating view.

The optimal energy assigned to the constellation by the different view-models projects the constellation of parts into a point in the *energy space* as shown in Figure 4. The SVM classifier learns the decision boundaries to correctly classify the instances of the different views in this space. This approach is similar in nature to the one proposed in [10] in the way that both use all the models to disambiguate the object; however, there are some fundamental differences between them. See Figure 5 for an illustration of the idea of disamiguation.

This work is to the best of our knowledge, the first attempt to automatically analyze the content of the echo videos for indexing and annotation purposes. The proposed method uses the discriminative classifiers on the results of the application of the generative view-models to classify images taken from echo videos into correct view classes. The use of the reported approach and results are not limited to the application presented here and could be applied to multiple object recognition problems when the objects are represented by the constellation of their parts.

2. Part-Based Representation of Echocardiogram Views

The spatial arrangment of the chambers of the heart was proven effective by Tagare *et al.* [11] in finding similar image planes in tomographic images. The assumption was that if two images contain the same chambers and each chamber is surrounded similarly by other chambers, then those two images should have been taken from the same view of the heart. This is intuitive when one looks at the images as cross sections of the imaging plane and the 3-dimensional structure of the heart at different angles. In that application, the chambers of the heart were manually located and labeled by an expert. The focus of the work was on expressing the spatial arrangements and assessing the similarity between such representations.

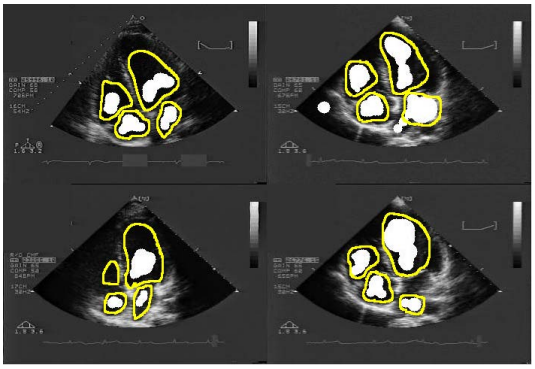

Figure 3. GSAT applied to four different key-frames in the apical four chamber view. The automatically detected chambers are shown with solid blobs. The boundary lines are drawn by hand to show the actual location of the chambers. The top-right image has a false positive, whereas in the bottom-left image a chamber is missing.

In light of that work, one could assume that the spatial arrangement of cardiac cavities is a good indicator for identifying the different views in the echo videos. We have to make a note here that the physicians do not rely solely on the spatial arrangement of the chambers for identifying the different views and use additional contextual information in doing so. This is also confirmed in [11], where the results showed that retrieval of similar image planes was slightly better by an expert compared to the machine which was only relying on the spatial relationships between the chambers.

2.1. Automatic Chamber Detection

The cardiac chamber segmentation has been an active field of research. There are various semi-automatic and automatic methods proposed in the literature. We use the method proposed by Bailes [1] which is based on the *Gray-Level Symmetric Axis Tranform (GSAT)* to detect the cardiac cavities. This method assumes that there exists a distinct cavity in the image for each chamber of the heart. It then tries to locate the deepest cavities in the image. We avoid the details of the *GSAT* method and refer the reader to [1]. Figure 3 shows the application of the *GSAT* method to key-frames taken from a view of the echo video.

Before proceeding, we need to mention that we represent each temporal segment of the echo video corresonding to a view of the heart by a set of representative frames or *key-frames* which are sampled from the content at semanti-

cally meaningful time instances (see our previous work on content-based sampling of the echocadiograms: [6]). The location of each key-frame corresponds to the structural state of the heart where the heart is most expanded (*End-Diastole* [7]).

The reason for using only images of each view at this structurally unique state is that during each heart cycle heart goes through different phases of activity, where in each phase the number of parts, their spatial relationships, and their properties change. Such structural variability will hamper our ability to learn a model of the heart for representing the constellation of its parts in the different views.

3. Modelling the Constellation of Parts

We express both the models of the constellation of parts in each view and the observed constellations from a given image using *Relational Structures (RS)*. Each node in such structure represents a cavity (see Figure 2). Both the nodes and the edges in the structure have attributes. The attributes of the nodes are the properties of the cavities and those of the edges are the relationship between two neighboring nodes. For handling the missed and false cavities in the scene (observed constellation), we let all the nodes be related to each other in the relational structure.

Let the parts in the observed constellation be indexed by the set $S = \{1, 2, \ldots, N\}$ (we follow the notation presented in [9] in the following), and let $d = [d_1(i), d_2(i, i')]$, where $i, i' \in S$ be the *unary* and *binary* properties of the parts. We can represent the relational structure of the observed constellation as the attibuted graph: $\mathcal{G} = (S, \mathcal{N}, d)$, where $\mathcal{N}$ represents the fully connected neighborhood structure in the relational structure. Likewise, the model of a view with M parts having labels $\mathcal{L} = \{l_1, l_2, \ldots, l_M\}$, and the parts properties $D = [D_1(I), D_2(I, I')]$, where $I \in \mathcal{L}$, could be expressed as the attributed graph: $\mathcal{G}' = (\mathcal{L}, \mathcal{N}', D)$, in which $\mathcal{N}'$ denotes the full connectivity between the parts in the model. Note that in the ideal case where there are no missing and false cavities in the observed constellation, one would have $M = N$.

The task of labeling the observed constellation using the model of a certain view is then posed as finding the best mapping between the nodes of the two attibuted graphs: $f : \mathcal{G} \rightarrow \mathcal{G}'$. We define a set of random variables $f = \{f_1, f_2, \ldots, f_N\}$ (*Random Field*) on the nodes of the attributed graph of the scene, where each of the random variables take values in the set of labels $f_i \in \mathcal{L}$ in the attributed graph of the model of the view. Using this framework, the problem of finding the optimal mapping between the two attributed graphs could be posed as a search for the optimal configuration of the random field f in the space of possible configurations $\Omega = \mathcal{L}^N$.

The optimal configuration of the random field is the one that results in maximum posterior probability $P(f|d)$ (*MAP-MRF*) or equivalently minimizes the posterior energy [9]; *i.e.*,

$$f^* = \underset{f \in \Omega}{argmin}(U(f|d)) = \underset{f \in \Omega}{argmin}(U(f) + U(d|f)) \quad (1)$$

where $U(f)$ is the prior energy and $U(d|f)$ the likelihood energy. The prior energy of a configuration is defined as (note that we only consider cliques of size up to two);

$$U(f) = \sum_{i \in S} V_1(f_i) + \sum_{i \in S, i' \in S - \{i\}} V_2(f_i, f_{i'}) \quad (2)$$

where the potentials are defined as follows:

$$V_1(f_i) = \begin{cases} v_1 & \text{if } f_i = 0, \\ 0 & \text{otherwise.} \end{cases}$$
$$V_2(f_i, f_{i'}) = \begin{cases} v_2 & \text{if } f_i = f_{i'} \text{ and } f_i \neq 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

where both v_1 and v_2 are positive values. The intuition behind this definition is that a *NULL* label assignment (hereafter *NULL* and $f_i = 0$ will be used interchangabely) to a site incurs a positive energy. This is to avoid all sites in the field to be labeld *NULL*. A *NULL* label only will be assigned to a site if the relative penalty of it is smaller than the the likelihood of the site having any other label. The pair-site prior potential encourages the sites to have distinct labels. This is derived from the fact that in the constellation of the heart chambers a chamber can not appear more than once.

The energy likelihood is defined as;

$$U(d|f) = \sum_{i \in S} V_1(d_1(i)|f_i) + \sum_{i \in S, i' \in S - \{i\}} V_2(d_2(i, i')|f_i, f_{i'}) \quad (4)$$

where the likelihood potentials on single cliques and double cliques are defined as:

$$V_1(d_1|f_i) = \begin{cases} \sum_{k=1}^{K_1} (d_1^{(k)}(i) - D_1^{(k)}(f_i))/\{2\sigma_k^2(f_i)\} \\ \text{if } f_i \neq 0, \\ 0 \end{cases}.$$
$$V_2(d_2|f_i, f_{i'}) = \begin{cases} \sum_{k=1}^{K_2} (d_2^{(k)}(i, i') - D_2^{(k)}(f_i, f_{i'}))/\{2\sigma_k^2(f_i, f_{i'})\} \\ \text{if } f_i \neq f_{i'}, f_i \neq 0, f_{i'} \neq 0, \\ 0 \end{cases}.$$

K_1 and K_2 are the total number of unary and binary features defined on single and pair-site cliques. In the present work, we use *location*, *area*, and *directionality* as the properties of each individual part, and *distance* and *angle* between a pair of parts as their joint properties. The properties

of the single and double site cliques are considered to be distributed according to Gaussian distributions with means defined by $D_j^{(k)},\ j = 1, 2$.

We obtain the maximum likelihood estimate of the parameters of the potential functions from the manually labeled training data. The bounds on the prior parameters v_1 and v_2 are estimated such that the training data are embedded at the minimum energy configurations of the configuration space.

To embed the training constellations at the minimum energy configuration in the configuration space, we note that any perturbation of the configuration of the labels on the training set should result in non-optimal configuration. By changing the *non-NULL* labeled parts to *NULL* and vice versa and obtaining the energy of the new configuration, we therefore obtain a set of inequalities which provides us with the bounds on the prior parameters. Equations 6 and 7 illustrate the conditions to obtain the lower and the upper bounds for the prior parameters.

$$\begin{aligned} &\forall i \in S, f_i^* \neq 0 \rightarrow f_i = 0, \\ &E(f_i|f^*_{S-\{i\}}, v_1, v_2) - E(f_i^*|f^*_{S-\{i\}}, v_1, v_2) > 0, \\ &\Rightarrow v_1 > E(f_i^*|f^*_{S-\{i\}}, v_1, v_2) \end{aligned} \tag{6}$$

$$\begin{aligned} &\forall i \in S, f_i^* = 0 \rightarrow f_i \neq 0, \\ &E(f_i|f^*_{S-\{i\}}, v_1, v_2) - E(f_i^*|f^*_{S-\{i\}}, v_1, v_2) > 0, \\ &\Rightarrow E(f_i|f^*_{S-\{i\}}, v_1, v_2) > (v_1 - \alpha \times v_2) \end{aligned} \tag{7}$$

where α is zero if the label assigned to the previously *NULL* site is a label that is missing from the observed constellation and is one if that label already exists in the constellation; and therefore, we incur a penalty by having two sites with the same label. We apply this estimation of the bounds of the prior parameters to each constellation in the training data and take the common value over all those constellations.

The value of the *local energy* resulting from assigning the label f_i to a site *i* is defined as:

$$E(f_i|f_{S-\{i\}}) \triangleq E_1(f_i) + \sum_{i' \in S-\{i\}} E_2(f_i, f_{i'}) \tag{8}$$

where;

$$E_1(f_i) = \begin{cases} V_1(d_1(i)|f_i) & \text{if } f_i \neq 0, \\ v_1 & \text{otherwise.} \end{cases}$$

$$E_2(f_i, f_{i'}) = \begin{cases} V_2(d_2(i,i')|f_i, f_{i'}) & \text{if } f_i \neq 0,\ f_{i'} \neq 0, \\ v_2 & \text{if } f_i = f_{i'} \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

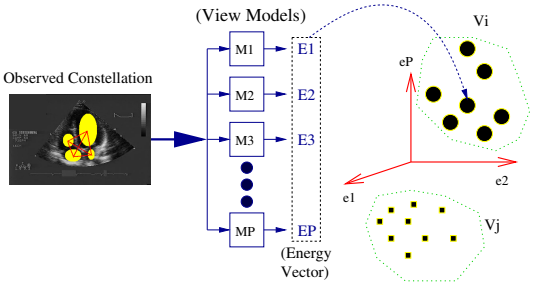

Figure 4. A constellation is shown in its energy space representation.

4. View Recognition

So far, we have learned the model of the constellation of the cardiac chambers for each view of the echo video. Given the constellation $\mathcal{C}$ obtained from a test image $\mathcal{I}$, we want to determine its correct view-label V^*

The chamber constellation $\mathcal{C}$ is matched against the model of each view $\mathcal{M}_k$, where $k = 1, \ldots, K$ (total number of views in our case is $K = 10$). The optimal labeling of the chambers in the observed constellation according to each of the models is inferred by minimizing the posterior energy of the configuration of the labels such that the resulting configuration is both consistent with the prior knowledge and the evidence. The *HCF* method proposed by Chou and Brown [4], which is a deterministic algorithm for combinatorial optimization, is used to obtain the optimal configuration of the random field. The main advantage of this method is its low complexity $O(n)$ (*n*= number of sites).

Say labeling the constellation $\mathcal{C}$ against all the K models results in the set of optimal configurations and their corresponding energy values,

$$\begin{aligned} \mathcal{F}^*(\mathcal{C}) &= \{F_k^*(\mathcal{C}) \mid k = 1, \ldots, K\}, \\ \mathcal{E}^*(\mathcal{C}) &= \{E_k^*(\mathcal{C}) \mid k = 1, \ldots, K\} \end{aligned} \tag{10}$$

where, $F_k^*(\mathcal{C})$ is the optimal configuration of the random field defined on the sites in the constellation $\mathcal{C}$ according to the model of the $k-th$ view, and $E_k^*(\mathcal{C})$ is its corresponding energy. Each element of the enregy vector is an indicator of how the corresponding model matches the constellation. In other words, it quantifies how each model *"sees"* the constellation $\mathcal{C}$. Figure 5 illustrates this idea.

For now, we assume that the constellation $\mathcal{C}$ is *complete*, meaning that it does not have any false parts and there are no chambers missing from it. In this case if one slightly purturbes the properties of the nodes and edges in the attributed relational structure of $\mathcal{C}$ the corresponding optimal energy vector will change to $\mathcal{E}^*(\mathcal{C}) + \delta\mathcal{E}(\mathcal{C})$.

Therefore, the energy vectors for all the *complete* constellations taken from the same view populate the same region in the energy space (see Figure 4 for illustration). One can use discriminative techniques such as SVM [5] to find the best classifier for discriminating between the constellations taken from the different views. Therefore, the results of matching the observed constellation $\mathcal{C}$ against all the models are fused together in order to decide the correct view label.

Two views (classes) will become indistinguishable from each other in this framework when the energy vector of the constellations taken from them populate the same region in the energy space. This would happen if none of the view-models can distinguish the identities of the two views. In this case the representation used and the models of the constellation of the parts in the different views do not possess enough complexity.

If we now allow missing parts (still no false parts observed), the distributions of the energy vectors of the constellations of the different views will tend to spread out and have more overlap with the ones belonging to the other models. The reason is that based on which combination of the parts is missing, the observed constellation could potentially become more similar to the typical constellations in another view. Still one could learn a classifier to separate the constellations of the different views. Naturally, one would expect higher error rates in the recognition results.

The most challenging case would be the one when both missed and false parts are allowed in the constellations. The effect of the false positives will completely deviate the energy vector associated with the observed constellation from its normal distribution. This is because the introduction of the false positives will potentially contribute to higher a degree of similarity between the two models, based on the properties of those false parts and their numbers.

In this case, we form a *multi-hypothesis* testing case. For each possible view $k \in 1, \ldots, K$, we assume that the observed constellation was taken from that view, *i.e.* $\mathcal{C} \in \mathcal{V}_k$, where $\mathcal{V}_k$ has the model $\mathcal{M}_k$. For each such assumption, we label the constellation $\mathcal{C}$ according to the model $\mathcal{M}_k$, and find the optimal label of the parts. We then delete the parts labeled as false according to this labeling and then obtain the energy vector for the *filtered* constellation. By doing this, if the observed constellation $\mathcal{C}$ was truly taken from view $\mathcal{V}_k$, it would be correctly classified by the classifier learned for the case where only missing parts were allowed. We now apply the same process to the observed constellation according to all the different view assumptions and find the corresonding classification results for each such obtained energy vectors. One could then decide on the correct view label for the observed constellation by comparing the confidence scores of the different classification results under the different view assumptions and fuse the decisions.

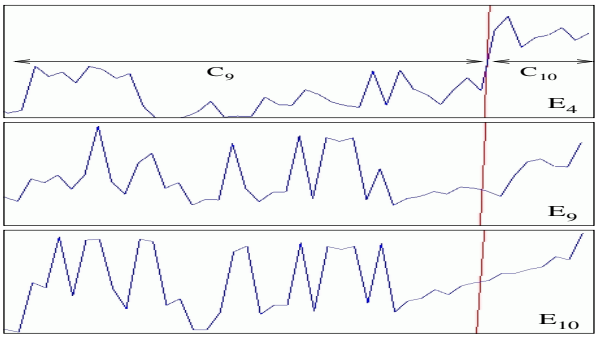

Figure 5. The images from top to bottom show the energies assigned to constellations taken from the views 9 and 10 (C_9, C_{10}) by the models of views 4, 9, and 10 (E_4, E_9, E_{10}). It is evident from the figures that neither the model of view 9, nor is view 10 able to distinguish between the constellations. However, the model of view 4 could be used to distinguish the constellations taken from those two views. In other words, the model of view 4 *"sees"* the constellations taken from the views 9 and 10 differently.

5. Experiments and Results

In this section, we report the results of our experiments for view recognition in echo videos using the proposed method. Our data set is consisted of $N_1 = 15$ echo videos of normal, and $N_2 = 6$ echos of abnormal cases. The echos of the normal heart with a total of 2657 key-frames are used both for training and testing (in *leave-one-out* fashion), while those of the abnormal echo videos with total of 552 key-frames are only used for testing purposes. Every key-frame in the data set is manually labeled by an expert. There are $K = 10$ different views present in these echo videos, taken from the *Parasternal Long Axis* (2 views), *Parasternal Short Axis* (4 views), and *Apical* (4 views) angles of imaging [7].

We conduct the experiments in a *leave-one-out* scheme, where in each round (15 rounds total) one echo videos is left out as the test data and we learn the models from the remaining 14 echos. For each of the views, we learn the priors and the paramters of the Gaussian distributions of the single and double cliques from the hand-labelled training data. The SVM classifiers are also learned for each round of experiment. The experiments and results are reported for four different cases below.

CASE 1 *(Normal Echos-Complete Constellations)*. In

the first set of experiments, we want to investigate the performance of the view recognizer for the case where the constellations are *complete*; *i.e.*, they do not have false parts and none of the parts are missing. In this case, the uncertainties in the properties of the parts in the model of each view and the closeness of the geometry of the constellations creates ambiguities between the constellations of the different views.

From the available training data, only the key-frames which contain complete constellations and maybe false positive chambers are selected. If the constellation contains false parts, those parts are removed from the constellation (*filtered constellations*). Each filtered constellation is then labeled according to the models of the different views, and the corresponding energy vector at the optimal configurations is found.

Using the energy vectors obtained for all training data, a multi-class *SVM* classifier is then trained for each round. The energy vectors of the test key-frames (also filtered constellations) are then classified using the learned multi-class SVM classifier. The top two images in Figure 6 show the Hinton diagram of the confusion matrix for this case with and without considering *clinical similarities*. We define clinically similar views as the ones that even the human expert could not discriminate and for clinical purposes they are regarded as identical. In the present case, the views in each of the following sets $\{4, 5, 6\}$, $\{7, 8\}$, and $\{9, 10\}$ are considered clinically similar.

The average precision over all rounds of experiment and all views is 67.8% without and 88.35% with taking into account the clinical similarities. As shown in Figure 6, for the original case the set of views that are often confused with each other are $\{4, 5, 6\}$, $\{1, 3, 7\}$, and $\{2, 10\}$. In the first set, all views have a single part (degenerate case) with highly overlapping property distributions; in the second case all views have constellations with four parts; and in the last case, all have two part constellations.

It is worth mentioning that the direct comparison of the energy values obtained from comparing the test images with the different view-models results in an average precision of 20% for this case.

CASE 2 *(Normal Echos-Complete Constellations with False Chambers)*. In this case, the introduction of the false chambers contributes to the similarity between the constellations taken from the different views. The false chambers only occur at certain regions of the image based on the view. However, because of the Gaussin model used for the cliques in the random field modeling, they still could be labeled non-Null under certain view assumptions based on their properties in the image.

In order to do the view recognition in this case, we use the multi-hypothesis approach as explained in Section 4 and use the same SVM classifiers learned for the complete case.

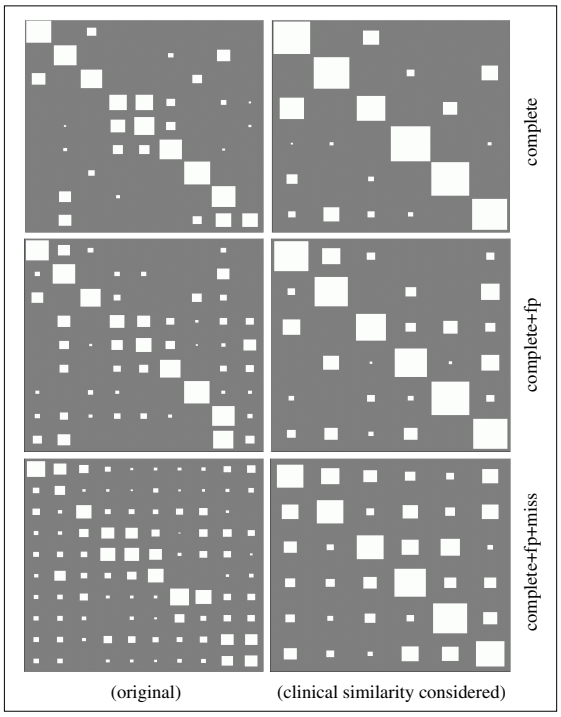

Figure 6. Hinton diagrams of the confusion matrices for cases $1 \sim 3$ from top to bottom respectively. Ideally, one should only see full white squares on the diagonal. On the left, the confusion matrices for the original set of views are shown, and on the right are the ones for the case where the clinical similarities are taken into consideration.

The Hinton diagrams of the confusion matrices for the cases with and without taking into account the clinical similarities are shown in the middle row of Figure 6. The average precisions of 54.1% and 74.34% are obtained for the original and the clinically similar cases respectively.

CASE 3 *(Normal Echos-Constellations with Missed and False Chambers)*. For this case, we learn the SVM classifiers for each round of experiment using constellations with their false chambers removed. Then, the multi-hypothesis approach was used to classify each real constellation (both false chamber and missed chambers exist).

As seen from the confusion matrices shown in the bottom row of Figure 6, the ambiguity between the views naturally becomes worse than the case that only false positives were allowed. The average precision in this case drops to 34% without and 52% with the clinical similarity taken into account.

It is useful to compare these results to the random guessing of the views with non-uniform priors. Given that the

prior probability of correctly guessing the view labels for the constellations taken from the view k is p_k, the overall rate of error becomes; $\epsilon = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$. The average rate of error of randomly guessing the view labels with non-uniform priors over the 15 rounds of experiments became $\epsilon = 88.13\%$, that is average rate of correct recognition by guessing is 11.87%. Therefore, the automatic recognition is almost three times better than random guessing in the worst case.

CASE 4 *(Abnormal Echos)*. In this experiment, we want to see how the models learned from the key-frames taken from the normal echos perform on the test images taken from the abnormal ones. There are 6 abnormal echo videos on which we test the performance of the view recognizer. We only consider the complete constellation case in this experiment.

The confusion matrix of this case is very similar to that of case (1). The average precision is 56% without and 78% with taking into account the clinical similarities. Compared to the similar case for the normal echos, the precision has slightly dropped.

6. Conclusion and Future Work

In this paper, we addressed the issue of automatic view recognition in echocardiogram videos and its application to the indexing of the content of these videos. The automatic indexing process could potentially be applied to both the analog echo videos (after digitization) and the digital ones. It is very expensive to manually index the content of the analog echo videos. For the digital echos although the new acquisition devices provide online annotation tools their use is burdonsome because the focus during the acquisition process is on capturing the best and the most clear images rather than annotating the content.

We used a part-based representation for the constellation of the heart chambers in the different views and used *Markov Random Fields* to model such represenations. The collection of the energies obtained from comparing a test image to the models of the different views was then used as the input to a SVM classifier to find the view label. This fusion scheme helped to disambiguate the constellations taken from the views with structural similarities.

The results of the automatic recognition of the view labels could be improved if one uses more complex distributions for the properties of the parts. Also, the training of the MRF model of the constellations of the cardiac chambers and the SVM classifiers were performed separately. One can combine the two to obtain better results.

In our approach, the models of the different views of the echocardiogram video were learned from the key-frames extracted from the content of the videos. We believe that by using the spatio-temporal characteristics of the constellation of the chambers of the heart, we not only can improve the results but also potentially recover the different phases of activity of the heart during each cycle of its activity. We are currently investigating this issue.

Although the approach presented in this paper was applied to the echo images, it could potentially be applied to any multiple object recognition problem when the objects are represented by the constellation of their parts and there are ambiguities due to similarity in their structures.

References

[1] D. R. Bailes. The Use of the Gray Level SAT to Find the Salient Cavities in Echocardiograms. *Journal of Visual Communication and Image Representation*, 7(2):169–195, June 1996.

[2] M. Boshra and B. Bhanu. Predicting performance of object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):956–969, September 2000.

[3] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *European Conference on Computer Vision*, volume 2, pages 628–641, 1998.

[4] P. B. Chou and C. M. Brown. The theory and practice of bayesian image labeling. *International Journal of Computer Vision*, 4:185–210, 1990.

[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

[6] S. Ebadollahi, S.-F. Chang, and H. Wu. Echocardiogram Videos: Summarization, Temporal Segmentation and Browsing. In *IEEE International Conference on Image Processing, (ICIP'02)*, volume 1, pages 613–616, September 2002.

[7] H. Feigenbaum. *Echocardiography*. Lea Febiger, 1993.

[8] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 66–73, 2000.

[9] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer-Verlag, 2001.

[10] A. Selinger and R. C. Nelson. Appearance-based object recognition using multiple views. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume I, pages 905–911, 2001.

[11] H. D. Tagare, F. M. Vos, C. C. Jaffe, and J. S. Duncan. Arrangement: A Spatial Relation between Parts for evaluating Similarity of tomographic section. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 17(9):880–893, September 1995.

[12] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, June 1991.

[13] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *6th European Conference on Computer Vision, (ECCV'00)*, June 2000.