Automatic Alignment of Echocardiogram Sequences

Amani R. Maina-Kilaas
Harvey Mudd College
Claremont, CA, USA
amainakilaas@hmc.edu

Abstract

In order to apply dimension reduction techniques to echocardiogram videos, the videos must first be synchronized such that all hearts beat in unison. In this paper, we present an algorithm to automatically perform such alignment. We additionally provide a heuristic for measuring the performance of such algorithms.

1. Introduction

Echocardiography is a form of medical imaging of the heart that utilizes non-invasive ultrasound waves. The image sequences (videos) provided by echocardiography reveal critical information and can assist in diagnosing many cardiovascular diseases. Unfortunately, examining every part of a large number of videos is very time consuming. Several dimensionality reduction techniques exist for identifying the important parts of these videos, particularly for common cardiovascular abnormalities, which would greatly reduce the work required in diagnosis. However, such techniques require a large collection of properly-aligned videos. We here present an algorithm that automates this task of aligning echocardiogram videos, preparing the data for downstream use in other methods. The repository (without data for privacy) is available at https://github.com/amanirmk/ecg-alignment.

2. Related Work

Video synchronization is a difficult task in computer vision. Stein's method assumes static cameras and homography between images, which is true for our echocardiogram data [3]. Caspi and Irani present a method for aligning two sequences of the same dynamic scene, which may also work well for tasks like this one [1]. Tuytelaars and Van Gool's proposed algorithm is more general, allowing cameras to be independent [4].

3. Task Specification

We have been provided with data from a study containing echocardiogram videos of varying lengths for several patients, where each of these videos is taken from one of eight standard views. The full study contains 100 patients each with typically five to seven videos per standard view, but we design our model around a smaller subset. Each video is given as a DICOM, a standard medical imaging format, from which we can access the video as an array of colored images. An example image is shown in Figure 1.

The desired output of this algorithm is the same set of DICOM data, where the videos of each view are aligned with one another, such that any two videos from any patients (of the same view) play in sync with respect to the heartbeats. This involves potentially compressing or dilating the time of some videos.

After obtaining the desired output, we need a method by which to assess the quality of our results. As there are currently no ground-truth alignments, this is difficult. The main assessment of results will be visual human evaluation, but we aim to develop a numerical measure as well.

4. Methods

4.1. Approach

Each heart beats differently, they can be fast or slow, steady or irregular. Accordingly, synchronizing these videos requires the manipulation of time. One method is to use a reference video and align every other video to it, cutting off excess video. We approach the task differently, using a method for synchronization we call *normalization*. In normalization, we find the complete heartbeat cycles in a video and set them all to a fixed number of seconds. This has the benefit of each video's processing being entirely independent. As a result, the processing cannot be negatively affected by choice of reference or issues in other videos. The independence could also be exploited for parallelization. Another benefit is that videos do not need to be trimmed at the alignment step, and can be trimmed after the fact once the desired subset of videos is chosen.

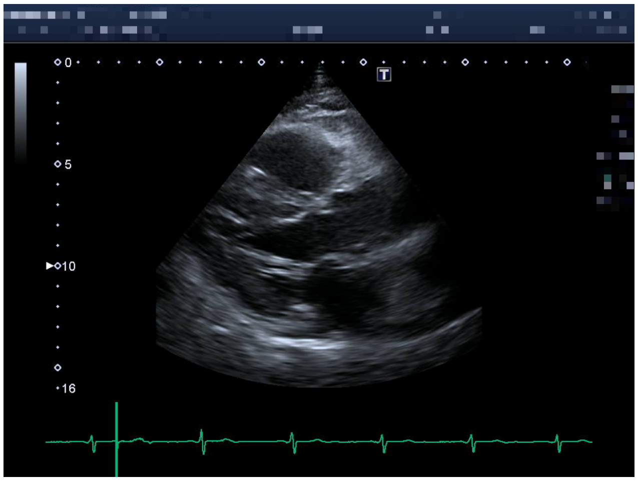

Figure 1. A frame of an echocardiogram video. A black and white image of the heart is in the center and a green heartbeat signal is shown at the bottom. Some parts of the image are pixelated to preserve privacy.

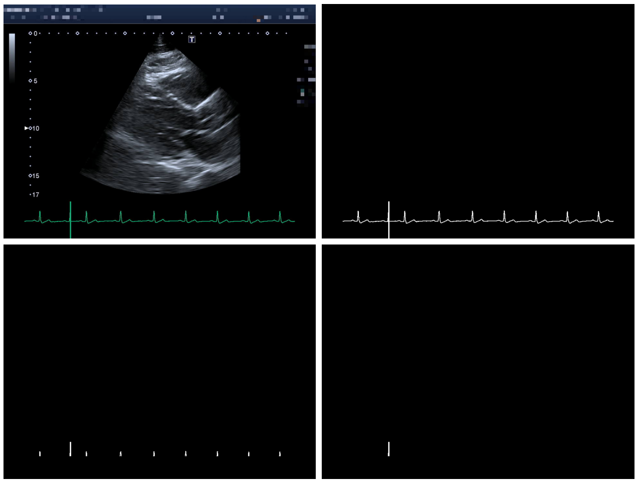

Figure 2. Top left: Original image. Top right: Isolated heartbeat signal with a mask. Bottom left: All but peaks are cropped. Bottom right: Largest bounding box is used to identify the vertical line indicating the location in the signal.

Our approach takes advantage the green heartbeat signal at the bottom of the video and ignores the actual heart footage. This may seem questionable, as it voluntarily throws out very pertinent information. However, the signal is easy to interpret with low ambiguity, allowing the process to utilize simple methods instead of complex learning techniques—which would require ground-truth training data that does not exist. This also has the benefit of being computationally inexpensive.

4.2. Algorithm

The algorithm is composed fundamentally of three steps: (1) identify which part of the signal a frame corresponds to, (2) determine whether the frame contains the start of a heartbeat, (3) segment video by complete heartbeat cycles and normalize. In this section we describe the technical process of these steps.

To identify which part of the signal corresponds to the given frame, we locate the large vertical line in the heartbeat signal which indicates what is currently being measured. First, a simple mask is applied to the image to obtain the green signal. At this point, the highest pixels in the image should correspond to the vertical line. However, we want our process to be robust to potential noise and employ further methods to isolate the line. Next, we remove pixels that are too low to be the line. Lastly, we find contours and use the largest bounding box to isolate the line. The results of this process are shown in Figure 2.

To determine whether the frame contains the start of a heartbeat, we analyze a stretch just before the vertical line that is five pixels wide. This segment is shown in Figure 3. We analyze the stretch in two ways. The first is the variance of the pixel's vertical positions. A low variance indicates a steady line while a high variance indicates a peak. We use a threshold variance of five pixels, which works for most signals. However, some signals have very high variance in general, hence our second analysis. We also compare the height of the peak to the typical peak height. We estimate the typical peak height by looking at the 5th and 95th percentiles of pixel heights and measuring their distance from the median. Once we know the typical height, we require that peaks also be at least 50% of this value.

From this, we have a set of frames that overlap with a heartbeat, but we specifically want the frames that begin the heartbeat. To fix this, we condense frames that are fewer than two frames apart down to just the first one.

Lastly, once each frame that starts a heartbeat is identified, we split the video into complete heartbeat cycles. Each of these segments is given their own frame rate so that the whole cycle lasts a fixed number of seconds. Then these segments are concatenated to obtain the full normalized video.

5. Results

To evaluate our synchronized videos, we rely largely on human assessment as there is no ground truth data. We find that our algorithm synchronizes the videos fairly well, although it is more clear for videos from the same patient. Figure 5 displays the original videos for a single patient side by side, while Figure 6 displays the aligned versions.

To try to measure our improvement, we use a simple heuristic. For each frame, we measure the standard deviation of a pixel's value across the videos. We then average these values for the pixels over each frame. The result of

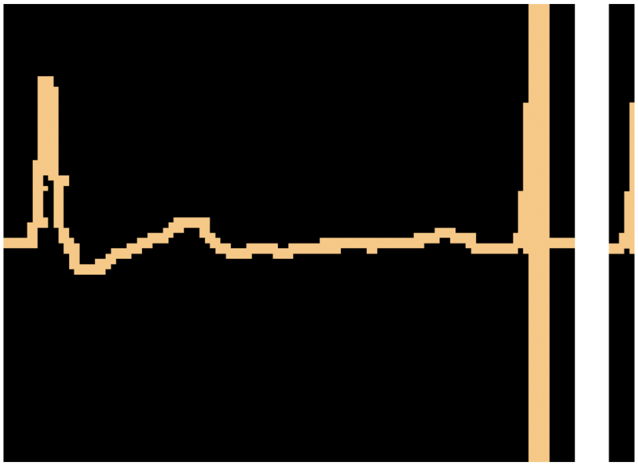

Figure 3. Left: A close up of the signal with the vertical line in view. Right: The stretch of signal analyzed for containing a heartbeat.

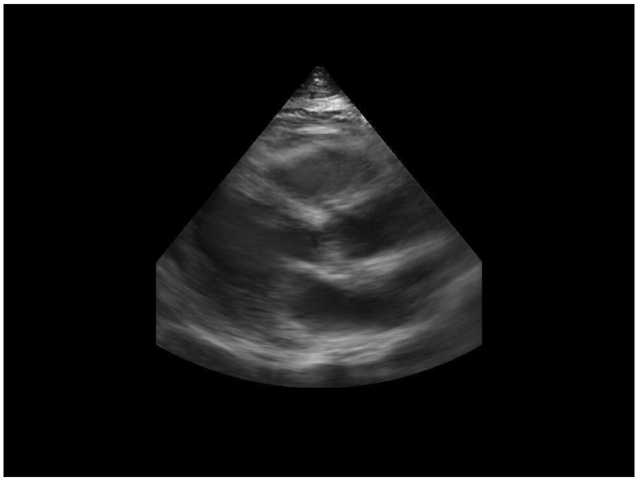

Figure 4. The standard deviation for a pixel across videos, averaged over frames.

this step can be visualized as an image like shown in Figure 4. Lastly, we average all of the pixels to obtain a single value. On two identical videos, this value should be zero.

We find that this measurement is fairly poor at highlighting the difference in synchronization, but does indeed capture the relationships we expect. As shown in Table 1, there is a slightly lower difference in the aligned videos than the original videos, and a moderately lower difference in the original videos than when the frames are shuffled.

6. Conclusion

In this paper, we presented an algorithm for synchronizing a specific view of echocardiogram videos for later use. We find that it performs reasonably well and propose a simple measure for comparing synchronized videos. However, the metric does not distinguish well. Future work could investigate utilizing video similarity measures such as ViSiL, introduced in [2]. Such similarity measures could also be utilized to optimize similairty when synchronizing. Other approaches could also involve deep learning techniques.

Data	Synchronization	Difference
9 Patients (23 videos)	Aligned	8.0265
	Original	8.0323
	Shuffled	8.1427
1 Patient (4 videos)	Aligned	4.5598
	Original	4.5859
	Shuffled	5.0271

Table 1. The differences calculated by the standard deviation metric. We compare the original videos, the aligned videos, and the original videos where frames are shuffled. The shuffled values are averaged over 50 trials.

Acknowledgments

We would like to acknowledge Jamie Haddock for providing the task and data, and Calden Wloka for his insightful advice.

References

[1] Yaron Caspi and Michal Irani. A step towards sequence-to-sequence alignment. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 682–689. IEEE, 2000. 1

[2] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Visil: Fine-grained spatio-temporal video similarity learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6360, 2019. 3

[3] Gideon P Stein. Tracking from multiple view points: Self-calibration of space and time. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 521–527. IEEE, 1999. 1

[4] T. Tuytelaars and L. Van Gool. Synchronizing video sequences. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, 2004. 1

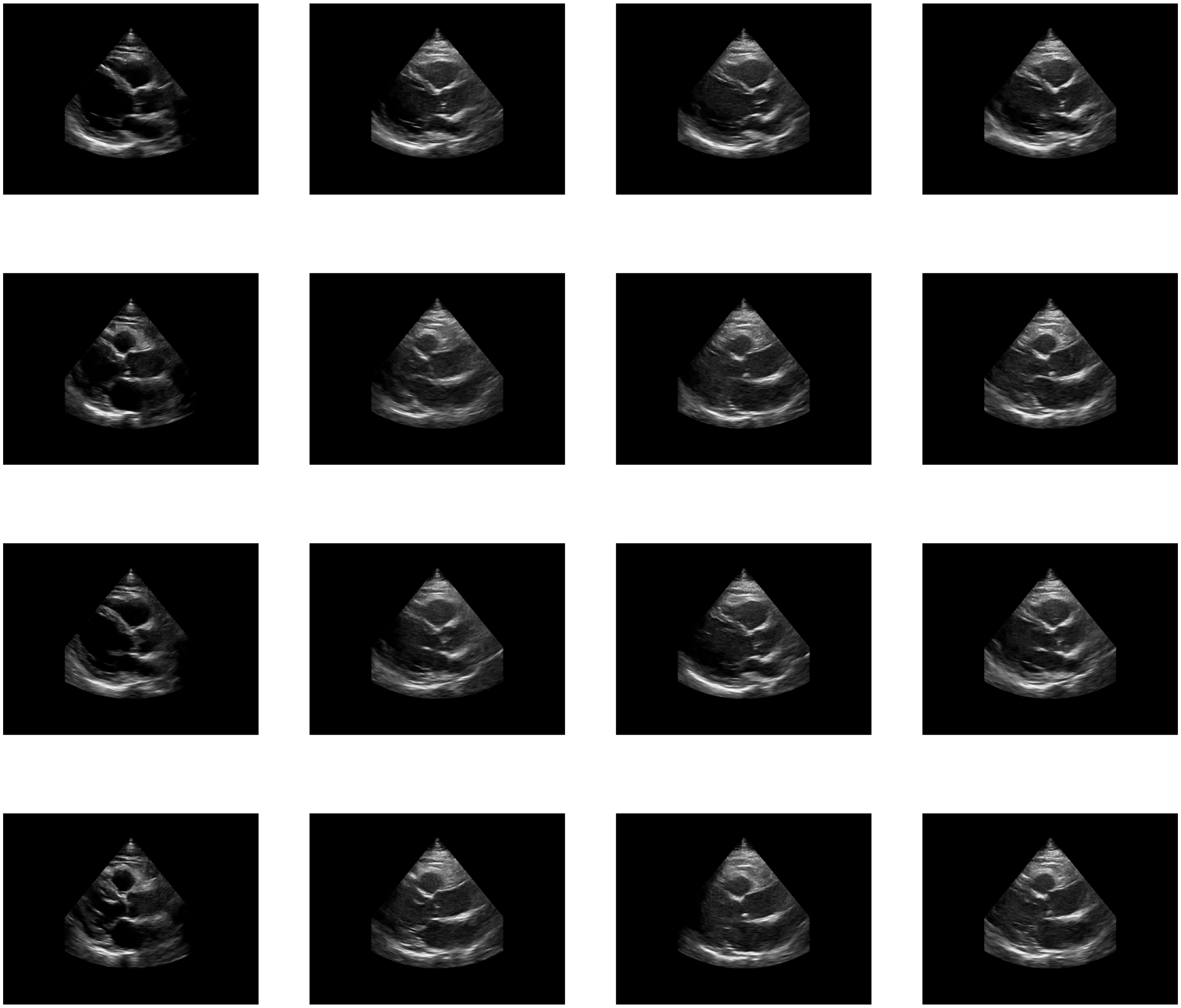

Figure 5. Original echocardiogram videos for a single patient. Each column holds a different video, where each row contains the same frame index. Each row is three frames apart, covering 12 frames total.

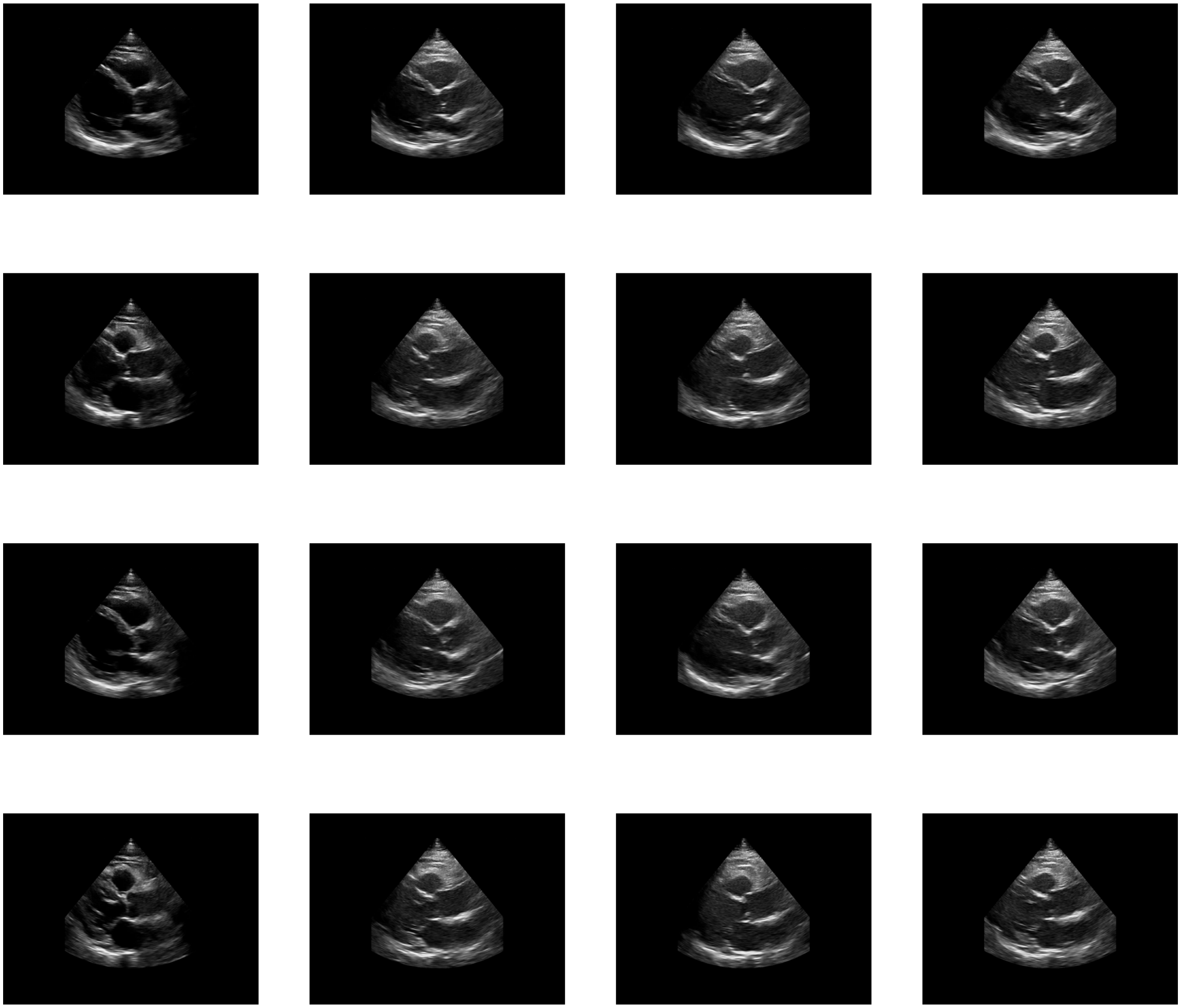

Figure 6. Aligned echocardiogram videos for a single patient. Each column holds a different video, where each row contains the same frame index. Each row is three frames apart, covering 12 frames total.