Detection of spatiotemporal stochastic motion in echocardiography videos

Anonymous Author(s)
Affiliation
Address
email

Abstract

A challenge in fetal echocardiography is the unpredictable relative motions between the fetus being assessed and the probe which manifests as a change in the viewing state of a relevant anatomy on the ultrasound video monitor. It is a difficult transition for automatic medical video analysis pipelines. We treat the detection of such spatiotemporally stochastic transitions as an anomaly detection case and use a novel statistical decision method over learnt spatiotemporal representations from convolutional LSTM models to encode partwise distance measures across local regions from multiple frames in video segments as discrete probability distributions compared by a Matusita coefficient score.This helps compute divergences between discrete probability distributions so obtained and allows segregation between normal and anomalous spatiotemporal representations out of video segments.

1 Introduction

There has been a considerable focus on automation of workflows in medical image analysis, but while most machine learning methods account for variability in spatial feature representations, there has been little effort to account for unforeseen variation in the spatiotemporal regime. Such random or unexpected motion is not uncommon in physiological cases, particularly in the case of the developing fetus where stochastic motion of the fetus relative to the uterus, or due to the patient and the probe, poses challenges in obtaining requisite standard viewing planes for anatomical observation in ultrasound sessions. Here, we attempt to identify such stochastic motions between the fetus and probe visible on the ultrasound video. Anomaly detection has been attempted for natural videos using techniques like multiple instance learning[1] and sparse coding[2]. Such methods are not well-posed for our task of inferring random motions in fetal ultrasound as durations of normal behaviour are not strictly confined to the beginnings or ends of input videos. Generative techniques using autoencoders[3] have also been attempted to perform novelty detection. In most anomaly or novelty detection problems with videos of natural scenes where an anomaly manifests as a significant spatial difference across a temporal length. Also, optical flow definition is relatively well-developed and sudden motion can be interpreted jointly with the actual frame level and flow based features. However, our task of assessing unpredictable, stochastic relative motions between the probe and the fetus from video sequences obtained in fetal echocardiography has significant differences which make it challenging.Here, arbitrary motions show up as sudden movements of the heart structure being viewed (akin to an unanticipated jerk in natural videos) as a complex combination of translations and rotations, a sudden change in the visibility of the heart or sudden lapses into the background due to unfavourable positioning of the fetus with respect to the ultrasound probe or an unplanned alteration in the viewing plane(our dataset has three standard planes [4] – four chamber(4C), three vessel(3V) and the left ventricular outflow tract(LVOT) view) due to the motion. Also, optical flow definition is poor in our ultrasound videos because of speckle and fine structural motion. This makes it difficult to use a two-stream CNN [5] for segregating unexpected motion patterns in fetal ultrasound.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

2 Methodology

The task of determining sudden motion of the fetus, an unexpected relative motion between the probe and the patient or a sudden change in the fetal cardiac views can be fundamentally treated as that of a spatiotemporal superposition of a stochastic motion pattern over that of the deterministic physiological motion – in our case the base cardiac rhythm in a standard fetal echocardiography sequence. As such, the separation of superimposed motion is a non-trivial problem with the complexity increased due to non-deterministic occurrence in fetal ultrasound sequences. So, the idea of determining the temporal instances of the onset and conclusion of such motion patterns and a quantification of the displacement of visible structures in the coordinate system fixed to a frame is of interest. The added complication is that the ultrasound modality has a strong influence of speckle and enhancements which are an extraneous influence on optical flows, and render the latter unsuitable for this task. With these constraints, we propose modelling the problem as one of supervised spatiotemporal anomaly detection, by relying on successive alterations in local spatial cues in a temporal sequence relative to sequences labelled as being normal. Essentially, we use a formulation of utilising distance based similarity measures and implementing them as probability distributions over localised patches and expanding over the temporal dimension. The idea is to enable the model to simultaneously learn to identify similar normal videos and to be able to discriminate between normal and anomalous video sequences. This is done by feeding in training samples grouping triplets of 8-frame video clips with two of them being normal and the third being an anomalous instance different from the other two. Over the individual video inputs in the group of three, we divide individual frames in a 8-frame video sequence into quadrant regions, and quadrants similarly located in the same frames of three comparator videos are fed to a triplet network which is trained to calculate a L1 distance between them. The distance metric so calculated is normalised using sigmoid activation, and a distribution of distances so computed is obtained for a given pair (so, this would be a discrete probability distribution with 32 instances if each frame in 8-frame videos is divided into 4 regions). Such probability distributions are compared using a Matusita metric [6]as it estimates overlap of discrete probability distributions without being impaired by the singularity problem that arises with more common measures like the Chi-squared metric in cases of similar distributions[7]. Our minimization objective is created as the Matusita measure between the two sets of normalised distances obtained in between the similar sequences and those in between the dissimilar sequences, in a training instance so chosen that two of them are labelled as similar and are different to the third instance. Sequences that are similar and therefore have similar probability distributions would have a small Matusita metric, whereas dissimilar distributions would have a larger value for the same. A labelled dataset where normal and abnormal sequences are identified can be used to train a triplet network with two branches' fully-connected layers used to calculate quadrant level distances, which are normalised, and used to feed to a layer evaluating the Matusita metric where similar distributions (labelled '0') and dissimilar distributions (labelled '1') are classified with a binary cross-entropy function.

Triplet architecture We implement modified Triplet Networks [8]with parallel strands accepting inputs as parts of triplets of 8-frame videos (from same or different class, i.e, both normal,identified as of 'same' class or 0, and two normal-one abnormal triplet, or a 'different' class or 1). The triplet legs process each input stream independently and jointly compute a L1 distance followed by a loss function using the L1 output sequences to calculate a Matusita coefficient and optimising it to corresponding labelled values of 0 or 1 by learning to classify 'same' and 'different' classes. Inputs are pre-processed as sequences of relevant quadrants to ensure efficient dataset curation and tractable distance computation. This implies that quadrants chosen for same locations in corresponding frames of three input sequences are processed by parallel strands of the triplet Network. This is done for all the quadrant pairs obtained for any two inputs (thus, two 8-frame videos are compared using a distribution of 32 normalised distances serving to input discrete probability distributions to the Matusita metric layers).A strand has 8 convolutional layers with 3x3 filters, ReLU activation and alternate max-pooling. This is followed by LSTM layers of 512 units to allow progression on the temporal video scale. Next is a 512-way fully-connected layer encoding representations for distance computation. Each input is separately computed upon using shared weights leading to three feature representations from each input pair example. We encode similarity between images by an L1 distance between the FC layers from pairwise parallel strands. This enables us to obtain two L1 distances for every input triplet.These two L1 distances are constrained between 0 and 1 with a sigmoid operation and used to compute a Matusita measure as the function to be optimised feeding into a binary cross-entropy classification to classify between 0 (similar) and 1 (different) classes.So

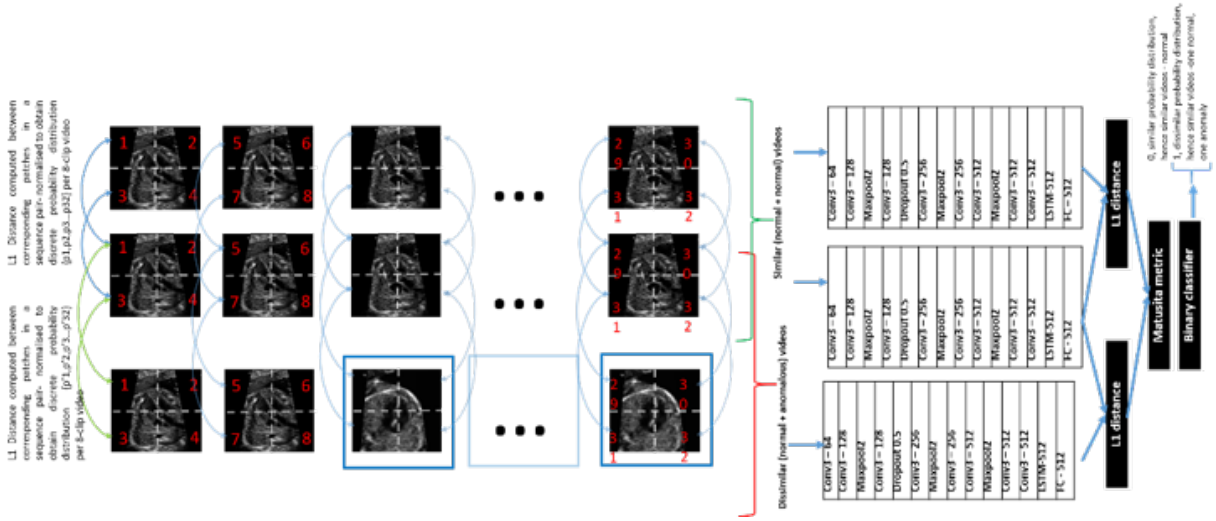

Figure 1: A triplet of videos is taken as an input instance – the triplet would include two video instances with normal motions and one with anomalous motion (onset marked with a blue rectangle on the left video – this is a sudden viewing plane change). Each frame divided into quadrants, patchwise evaluated in a triplet network to obtain patchwise L1 distances. This is repeated for a total of 32 patches (4 patches and 8 frames). The distribution of the normalised distances is interpreted as a discrete probability distribution to be compared using a Matusita metric. (expanded in Appendix)

we have a classification scheme by using the constraint that similar videos would have a distribution of local image level distances similar to each other due to the similar phases of spatial level changes over time and those with anomalies would show a different distribution due to non-deterministic changes in spatial features over time,i.e, given the deterministic nature of normal physiological motion,two videos starting at a similar landmark would evolve similarly in the absence of any abnormal motion and so local features would match.

Objective function For comparison of two discrete probability distributions, which is how we interpret the set of distance values of the divided quadrants from frames in the triplet of sequences, we model the divergence as a proxy for the dynamic similarity between the sequences. This is done using the Matusita measure, originally devised as a measure of risk in the analysis of statistical decision functions and defines the distance between two probability distributions as: $\sum_{i=1}^{N}(\sqrt{p(i)} - \sqrt{p'(i)})^2$

Here, N is the number of quadrant triplets to be considered for every training or test instance, i identifies the triplet worked upon, p(i) is the normalised distance between the first two quadrants in the triplet, and p'(i) is the normalised distance between the second and the third. The input to the Matusita loss function are the respective L1 distances activated using a sigmoid layer to be between 0 and 1. This essentially converts the normal versus anomalous video instances classification problem into a similarity measure optimisation task, with the objective being the class label of either 0 or 1.

Training We begin with 3624 8-frame echo clips,with 2424 clips tagged normal, and 1200 with one of the three types of motion anomalies studied here. The training to test data split is at a 80:20 ratio. The clips are pre-processed to include sequences of similarly positioned quadrants across the sequential video frames for each of the input videos, thereby leading to an effective number of four clips per input clip at runtime. These are fed together as parts of the same triplet so as to generate a combined discrete probability distribution per video pair. To create input triplets, we follow a strategy relying on the ordering of the dataset and choose a triplet such that two instances are normal and the third could be either normal or abnormal in a randomised selection, and correspondingly attach a label of 0 or 1. Here it is ensured that no more than half the total number of triplets are labelled 0 or 1 (or contain all normal or a normal-abnormal combination). Following the preparation of the input set, the three-way triplet network is trained for 50 epochs with a batch size of 20 and a learning rate of

Table 1: Results with overall anomaly data, and with subsets of the anomaly data considered

Classification Accuracy (percentage)				
Name	Overall	Visibility shift	Viewing plane shift	Sudden motion
Triplet Net	69.32	66.67	76.54	69.20
Autoencoder	54.86	61.17	68.50	67.35
Resnet-50	62.35	63.47	75.90	56.05
VGG-16	58.50	60.12	74.20	51.48

0.01. The experiments are conducted in four stages. First, all of the anomaly videos dataset is used for generating triplets. Following this, three other runs are performed with each of the classes of anomalies (Visibility shift, Viewing plane shift, Sudden motion) considered as the sole dataset for an anomalous case for triplet generation. We use baseline results derived using direct classification from a VGG-16 and Resnet-50 adapted(input of 8-frame clips at first layer[4]) with same training protocols as the triplet network. Also, an autoencoder baseline is used to represent reconstruction based anomaly detection, using 5 convolutional layers with 3x3 kernels in both upsampling and downsampling stages, trained to reconstruct non-anomalous patterns over classes. The classification is by sigmoid normalising the reconstruction error and estimating the class (normal, abnormal).

3 Results

The evaluation metric for our three-way distance encoded convolutional-LSTM formulation is the classification accuracy as normal and abnormal triplets are labelled. At test time, triplet inputs were designed that a test video was used with two normal videos,to have a triplet whose class label would be 0 (all three normal) or 1 (two normal, one anomalous video). The trained triplet network would output a label to be compared with the ground truth label of the input triplet. This was done for separate types of anomalies discernible in our videos – sudden fetal/probe motions, sudden changes in visibility and sudden changes in cardiac plane. In our data, unlike natural videos, the notion of spatiotemporal anomaly is not confined to cases of drastically different spatial cues emerging over short time instances. Rather, the nature of anomalies is such that overall spatial definitions across a single frame often stays constant, with an evolving variation in the orientation or linear shifts over temporal distances. This implies that a model relying on overall feature representation across frames is sub-optimal. So, using quadrants to compare distance based embeddings of local regions per frame of compared videos shows a performance gain and the method of comparison of probability distributions derived from consecutive local distances across pairs of videos is amenable to complexities of fetal and probe motion observation in the ultrasound videos considered. The relative difference between the classification accuracies in the case of Viewing Plane shift is the lowest overall compared across methods.In this class, there is a substantial change in frame level image description on the whole as the standard viewing plane of the fetus abruptly changes, causing a significant change in the global frame level features as to be amenable to an end-to-end training for classification. Contrarily, compared to other methods the Triplet Net does relatively much better on cases of sudden motion as these instances are characterised by overall frame level similarities but in the aggregate of multiple frames, manifest as a departure from the normal motion pattern. The average Matusita coefficient over the test data is reported for the test normal instances as 0.1125 and the test abnormal instances as 0.8944. Ideally, the Matusita coefficient would be 0 for normal and 1 for the abnormal instances. The overall digression from ideal values is indicative of a degree of error in modelling the similarity and dissimilarity of video instances, explained as a result of the loss of certain local level comparisons due to not going into any finer division than a quadrant level assessment. Note that the Overall accuracy here does not refer to an aggregate of the Visibility anomaly accuracies, the viewing plane anomaly accuracies and the sudden motion anomaly accuracies. This is because these are four different runs with the latter three conducted on separated anomaly video datasets. To conclude, we detect unanticipated motion in echocardiography videos by treating it as an anomaly over normal motion patterns, with a caveat that abrupt patterns may not lead to overall changes in global spatial features visible frame-to-frame but would manifest as discernible local changes. This scheme allows us to consider patchwise embeddings for a distance measure recast as a sequence of probabilities being compared on a triplet network thus overcoming limitations of simple CNNs caused by local pose and orientation invariance.

References

[1] Maron, Oded, and Tomás Lozano-Pérez. "A framework for multiple-instance learning." Advances in neural information processing systems. 1998.

[2] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In CVPR, pages 3313–3320, 2011.

[3] Wei, Qi, et al. "Anomaly detection for medical images based on a one-class classification." Medical Imaging 2018: Computer-Aided Diagnosis. Vol. 10575. International Society for Optics and Photonics, 2018.

[4] Patra, A., Huang, W., Noble, J. A. (2017). Learning Spatio-Temporal Aggregation for Fetal Heart Analysis in Ultrasound Video. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 276-284). Springer, Cham.

[5] Feichtenhofer, C., Pinz, A., Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1933-1941).

[6] Matusita, K. Decision Rules, Based on the Distance, for Problems of Fit, Two Samples, and Estimation, Ann. Math. Statist., Volume 26, Number 4 (1955), 631-640.

[7] Bruzzone, Lorenzo, Fabio Roli, and Sebastiano B. Serpico. "An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection." IEEE Transactions on Geoscience and Remote Sensing 33.6 (1995): 1318-1321.

[8] Hoffer, Elad, and Nir Ailon. "Deep metric learning using triplet network." International Workshop on Similarity-Based Pattern Recognition. Springer, Cham, 2015.

Appendix

Figure 2: A magnified version of figure 1. A triplet of videos is taken as an input instance – in half training and test data used the triplet would include two video instances with normal motions and one with anomalous motion (onset marked with a blue rectangle on the left video – this is a sudden viewing plane change). Each frame divided into quadrants, patchwise evaluated in a triplet network to obtain patchwise L1 distances. This is repeated for a total of 32 patches (4 patches and 8 frames). The distribution of the normalised distances is interpreted as a discrete probability distribution to be compared using a Matusita metric.