Multiple Object Detection by Sequential Monte Carlo and Hierarchical Detection Network

Michal Sofka Jingdan Zhang S. Kevin Zhou
Dorin Comaniciu
Siemens Corporate Research
755 College Road East, Princeton, NJ 08540, USA
{michal.sofka, jingdan.zhang, shaohua.zhou, dorin.comaniciu}@siemens.com

Abstract

In this paper, we propose a novel framework for detecting multiple objects in 2D and 3D images. Since a joint multi-object model is difficult to obtain in most practical situations, we focus here on detecting the objects sequentially, one-by-one. The interdependence of object poses and strong prior information embedded in our domain of medical images results in better performance than detecting the objects individually. Our approach is based on Sequential Estimation techniques, frequently applied to visual tracking. Unlike in tracking, where the sequential order is naturally determined by the time sequence, the order of detection of multiple objects must be selected, leading to a Hierarchical Detection Network (HDN). We present an algorithm that optimally selects the order based on probability of states (object poses) within the ground truth region. The posterior distribution of the object pose is approximated at each step by sequential Monte Carlo. The samples are propagated within the sequence across multiple objects and hierarchical levels. We show on 2D ultrasound images of left atrium, that the automatically selected sequential order yields low mean detection error. We also quantitatively evaluate the hierarchical detection of fetal faces and three fetal brain structures in 3D ultrasound images.

1. Introduction

Multiple object detection has many applications in computer vision systems, for example in visual tracking [15], to initialize segmentation [20], or in medical imaging [2]. Figure 1 illustrates the two examples of multi-object detection we are interested in. State-of-the-art approaches for multi-object detection [5, 19, 9] rely on an individual detector for each object class followed by post-processing to prune spurious detections within and between classes. Detecting multiple objects jointly rather than individually has the advantage that the spatial relationships between objects can be exploited. Since obtaining a joint model of multiple objects is difficult in most practical situations, the multi-object detection task has been solved by multiple individual object detectors connected by a spatial model [4]. Relative locations of the objects provide constraints that help to make the system more robust by focusing the search in regions where the object is expected based on locations of the other objects. The most challenging aspect of these algorithms is designing detectors that are fast and robust, modeling the spatial relationships between objects, and determining the detection order. In this paper, we propose a multi-object detection system that addresses these challenges.

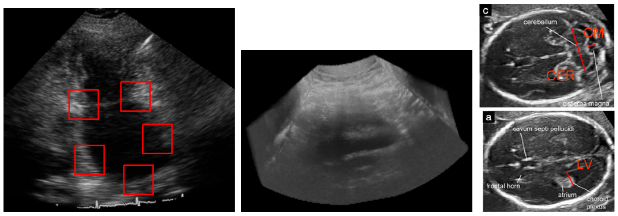


Figure 1. Examples of multi-object detection: five landmarks of left atrium (LA) apical two chamber (A2C) view (left) and 3D ultrasound volume of fetal brain with three anatomies (right).

The computational speed and robustness of our system is increased by hierarchical processing. In detection, one major problem is how to effectively propagate object candidates across the levels of the hierarchy. This typically involves defining a search range at a fine level where the candidates from the coarse level are refined. Incorrect selection of the search range leads to higher computational speeds, lower accuracy, or drift of the coarse candidates towards incorrect refinements. The search range in our technique is part of the model that is learned from the training data. The performance of our multi-object detection system is further improved by starting from objects that are easier to detect and constraining the detection of the other objects by exploiting object configurations. The difficulty of this strategy is selecting the order of detections such that the overall performance is maximized. Our detection schedule is designed to minimize the uncertainty of the detections. Using the same algorithm, we also obtain the optimal schedule of the hierarchical scales.

Our approach is motivated by Sequential Estimation techniques [8], frequently applied to visual tracking. In tracking, the goal is to estimate at time t the object state $\mathbf{x}_t$ (e.g. location and size) using observations $\mathbf{y}_{0:t}$ (object appearance in video frames). The computation requires a likelihood of a hypothesized state that gives rise to observations and a transition model that describes the way states are propagated between frames. Since the likelihood models in practical situations lead to intractable inference, approximation by Monte Carlo methods, also known as particle filtering, have been widely adopted. At each time step t, the estimation involves sampling from the proposal distribution $p(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{y}_{0:t})$ of the current state $\mathbf{x}_t$ conditioned on the history of states $\mathbf{x}_{0:t-1}$ up to time $t-1$ and the history of observations $\mathbf{y}_{0:t}$ up to time t.

We also use sequential Monte Carlo technique in multi-object detection. We sample from a sequence of probability distributions, but the sequence specifies a spatial order rather than a time order (Figure 2). The posterior distribution of each object pose (state) is estimated based on all observations so far. The observations are features computed from image neighborhoods surrounding the objects. The likelihood of a hypothesized state that gives rise to observations is based on a deterministic model learned using a large annotated database of images. The transition model that describes the way the poses of objects are related is Gaussian.

Most object detection algorithms have focused on a fixed set of object pose parameters that are tested in a binary classification system [17, 19]. Employing the sequential sampling model allows us to use fewer samples of the object pose and formally extend this class of algorithms to multiple objects. This saves computational time and increases accuracy since the samples are taken from the regions of high probability of the posterior distribution. Many ideas from the Sequential Sampling literature on visual tracking can likely be extended to multi-object detection. In Section 4, we will demonstrate the benefit of the sampling when detecting multiple landmarks in 2D images of the left atrium. Unlike in tracking, where the sequential order is naturally determined by the time progression, the order in multi-object detection must be selected. In our algorithm, the order is selected such that the uncertainty of the detections is minimized. So, instead of using the immediate precursor in the Markov process, the transition model could be based on any precursor, which is optimally selected. This leads to a Hierarchical Detection Network (HDN) 3. The likelihood of a hypothesized pose is computed using a trained detector. The detection scale is introduced as another parameter of the likelihood model and the hierarchical schedule is determined in the same way as the spatial schedule.

The paper is organized as follows. We give an overview of the background literature in Section 2. The sequential multi-object detection algorithm is proposed in Section 3. The algorithm is validated on a set of experiments presented in Section 4. We conclude the paper in Section 5.

2. Background

A discrete set of object poses is tested for an object presence with a binary classifier in many object detection algorithms [17, 19]. Unlike these algorithms, that typically sample the parameter space uniformly, we sample from a proposal distribution [14] that focuses on regions of high probability. This saves computational time as fewer samples are required and inreases robustness compared to the case, where the same number of samples would be drawn uniformly.

Multi-object detection techniques have focused on models that share features [16] or object parts [9]. This sharing results in stronger models, yet in recent literature, there has been a debate on how to model the object context in an effective way [7]. It has been shown that the local detectors can be improved by modeling the interdependence of objects using contextual [6, 13, 12] and semantic information [11]. In our Sequential Sampling framework, this interdependence is modeled by a transition distribution, that specifies the “transition” of a pose of one object to a pose of another object. This way, we make use of the strong prior information present in medical images of human body. The important questions are how to determine the size of the context region (detection scale) and which objects to detect first in an optimal way.

Multi-scale algorithms usually specify a fixed set of scales with predetermined parameters of the detection regions [1, 9]. Choosing the scale automatically has the advantage since objects have different sizes and the size of the context neighborhood is also different. We propose a multi-scale scheduling algorithm that is formulated in the same way as the detection order scheduling.

The order of detection has been specified by maximizing the information gain computed before and after the detection measurement is taken [21] and by minimizing the entropy of posterior belief distribution of observations [1]. Our scheduling criterion is based on probability of states (object poses) within the ground truth region. Other measures could be used as well thanks to the flexible nature of the Sequential Sampling framework.

3. Sequential Monte Carlo

The state (pose) of the modeled object t is denoted as $\boldsymbol{\theta}_t$ and the sequence of multiple object detections as $\boldsymbol{\theta}_{0:t} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t\}$. In our case, $\boldsymbol{\theta}_t = \{\mathbf{p}, \mathbf{r}, \mathbf{s}\}$ denotes the position $\mathbf{p}$, orientation $\mathbf{r}$, and size $\mathbf{s}$ of the object t. The set of observations for object t are obtained from the image neighborhood V_t. The neighborhood V_t is specified by the coordinates of a bounding box within an d-dimensional image V, $V : R^d \to [0, 1]$. The sequence of observations is denoted as $V_{0:t} = \{V_0, V_1, \ldots, V_t\}$. This is possible since there exists prior knowledge for determining the image neighborhoods $V_0, V_1, \ldots, V_t$. The image neighborhoods in the sequence $V_{0:t}$ might overlap and can have different sizes. An image neighborhood V_i might even be the entire volume V. The observations V_t with a marginal distribution $f(V_t|\boldsymbol{\theta}_t)$ describe the appearance of each object and are assumed conditionally independent given the state $\boldsymbol{\theta}_t$. The state dynamics, *i.e.* relationships between object poses, are modeled with an initial distribution $f(\boldsymbol{\theta}_0)$ and a transition distribution $f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1})$. Note that here we do not use the Markov transition $f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$.

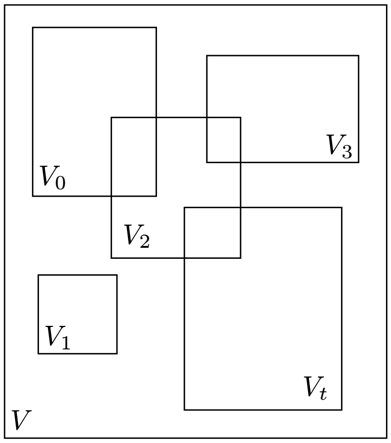


Figure 2. In multi-object detection, the set of observations is a sequence of image patches. The sequence specifies a spatial order rather than a time order. The latter is typically exploited in tracking applications.

The multi-object detection problem is solved by recursively applying *prediction* and *update* steps to obtain the posterior distribution $f(\boldsymbol{\theta}_{0:t}|V_{0:t})$. The prediction step computes the probability density of the state of the object t using the state of the previous object, $t-1$, and previous observations of all objects up to $t-1$:

$$f(\boldsymbol{\theta}_{0:t}|V_{0:t-1}) = f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}) f(\boldsymbol{\theta}_{0:t-1}|V_{0:t-1}). \tag{1}$$

When detecting object t, the observation V_t is used to compute the estimate during the update step as:

$$f(\boldsymbol{\theta}_{0:t}|V_{0:t}) = \frac{f(V_t|\boldsymbol{\theta}_t) f(\boldsymbol{\theta}_{0:t}|V_{0:t-1})}{f(V_t|V_{0:t-1})}, \tag{2}$$

where $f(V_t|V_{0:t-1})$ is the normalizing constant.

As simple as they seem these expressions do not have analytical solution in general. This problem is addressed by drawing m weighted samples $\{\boldsymbol{\theta}^j_{0:t}, w^j_t\}^m_{j=1}$ from the distribution $f(\boldsymbol{\theta}_{0:t}|V_{0:t})$, where $\{\boldsymbol{\theta}^j_{0:t}\}^m_{j=1}$ is a realization of state $\boldsymbol{\theta}_{0:t}$ with weight w^j_t.

In most practical situations, sampling directly from $f(\boldsymbol{\theta}_{0:t}|V_{0:t})$ is not feasible. The idea of importance sampling is to introduce a *proposal distribution* $p(\boldsymbol{\theta}_{0:t}|V_{0:t})$ which includes the support of $f(\boldsymbol{\theta}_{0:t}|V_{0:t})$.

In order for the samples to be proper [14], the weights are defined as

$$\begin{aligned} \tilde{w}^j_t &= \frac{f(V_{0:t}|\boldsymbol{\theta}^j_{0:t}) f(\boldsymbol{\theta}^j_{0:t})}{p(\boldsymbol{\theta}^j_{0:t}|V_{0:t})} \\ w^j_t &= \tilde{w}^j_t / \sum_{i=1}^m \tilde{w}^i_t. \end{aligned} \tag{3}$$

Since the current states do not depend on observations from other objects then

$$p(\boldsymbol{\theta}_{0:t}|V_{0:t}) = p(\boldsymbol{\theta}_{0:t-1}|V_{0:t-1}) p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}, V_{0:t}). \tag{4}$$

The states are computed as

$$f(\boldsymbol{\theta}_{0:t}) = f(\boldsymbol{\theta}_o) \prod_{j=1}^t f(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{0:j-1}). \tag{5}$$

Substituting (4) and (5) into (3), we have

$$\tilde{w}^j_t = \frac{f(V_{0:t}|\boldsymbol{\theta}^j_{0:t}) f(\boldsymbol{\theta}^j_{0:t})}{p(\boldsymbol{\theta}^j_{0:t-1}|V_{0:t-1}) p(\boldsymbol{\theta}^j_t|\boldsymbol{\theta}^j_{0:t-1}, V_{0:t})} \tag{6}$$

$$= \tilde{w}^j_{t-1} \frac{f(V_{0:t}|\boldsymbol{\theta}^j_{0:t}) f(\boldsymbol{\theta}^j_{0:t})}{f(V_{0:t-1}|\boldsymbol{\theta}^j_{0:t-1}) f(\boldsymbol{\theta}^j_{0:t-1}) p(\boldsymbol{\theta}^j_t|\boldsymbol{\theta}^j_{0:t-1}, V_{0:t})} \tag{7}$$

$$= \tilde{w}^j_{t-1} \frac{f(V_t|\boldsymbol{\theta}^j_t) f(\boldsymbol{\theta}^j_t|\boldsymbol{\theta}^j_{0:t-1})}{p(\boldsymbol{\theta}^j_t|\boldsymbol{\theta}^j_{0:t-1}, V_{0:t})}. \tag{8}$$

In this paper, we adopt the transition prior $f(\boldsymbol{\theta}^j_t|\boldsymbol{\theta}^j_{0:t-1})$ as the proposal distribution. Hence, the importance weights are calculated as:

$$\tilde{w}^j_t = \tilde{w}^j_{t-1} f(V_t|\boldsymbol{\theta}^j_t). \tag{9}$$

In future, we plan to design more sophisticated proposal distributions to leverage relations between multiple objects during detection.

When detecting each object, the sequential sampling produces the approximation of the posterior distribution $f(\boldsymbol{\theta}_{0:t}|V_{0:t})$ using the samples from the detection of the previous object as follows:

1. Obtain m samples from the proposal distribution, $\boldsymbol{\theta}_t^j \sim p(\boldsymbol{\theta}_t^j|\boldsymbol{\theta}_{0:t-1}^j)$.
2. Reweight each sample according to the importance ratio
$$\tilde{w}_t^j = \tilde{w}_{t-1}^j f(V_t|\boldsymbol{\theta}_t^j). \qquad (10)$$
Normalize the importance weights.
3. Resample the particles using their importance weights to obtain the unweighted approximation of $f(\boldsymbol{\theta}_{0:t}|V_{0:t})$:
$$f(\boldsymbol{\theta}_{0:t}|V_{0:t}) \approx \sum_{j=1}^{m} w_t^j \delta(\boldsymbol{\theta}_{0:t} - \boldsymbol{\theta}_{0:t}^j), \qquad (11)$$
where δ is the Dirac delta function.

3.1. The Observation and Transition Models

Let us now define a random variable $y \in \{-1,+1\}$, where $y = +1$ indicates the presence and $y = -1$ absence of the object. To leverage the power of a large annotated dataset, we use discriminative classifier (e.g. PBT [17]) in the observation model:

$$f(V_t|\boldsymbol{\theta}_t) = f(y_t = +1|\boldsymbol{\theta}_t, V_t), \qquad (12)$$

where $f(y_t = +1|\boldsymbol{\theta}_t, V_t)$ is posterior probability of object presence at $\boldsymbol{\theta}_t$ in V_t.

In tracking, often a Markov process is assumed for the transition kernel $f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}) = f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$, as time proceeds. However, this is too restrictive for multiple object detection. The best transition kernel might stem from an object different from the immediate precursor, depending on the anatomical context. In this paper, we use a pairwise dependency

$$f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1}) = f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_j), \quad j \in \{0, 1, \ldots, t-1\}. \qquad (13)$$

We model $f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{0:t-1})$ as a Gaussian distribution estimated from the training data. We will show how to select the best precursor j next.

3.2. Detection Order Selection

Unlike a video, where the observations arise in a naturally sequential fashion, the spatial order in multi-object detection must be selected. The goal is to select the order such that the posterior probability $P(\boldsymbol{\theta}_{0:t}|V_{0:t})$ is maximized. Since determining this order has exponential complexity in the number of objects, we adopt a greedy approach. We first split the training data into two sets. Using the first set, we train all object detectors individually to obtain posterior distributions $f(\boldsymbol{\theta}_0|V_0), f(\boldsymbol{\theta}_1|V_1), \ldots, f(\boldsymbol{\theta}_t|V_t)$. The second set is used for order selection as follows.

We aim to build a Hierarchical Detection Network (HDN) from the order selection. As shown in Figure 3, the HDN is a pairwise, feed-forward network. Note that the cascade is a special case of HDN.

Suppose that we find the ordered detectors up to $s-1$, $\theta_{(0)}, \theta_{(1)}, \ldots, \theta_{(s-1)}$. We aim to add to the network the best pair $[s,(j)]$ (or feed-forward path) that maximizes the expected value of the following score $S[s,(j)]$ over both s and (j) computed from the second training set:

$$S[s,(j)] = \qquad (14)$$
$$\int\limits_{\substack{\theta_s \in \Omega(\tilde{\theta}_s) \\ \theta_{(0:s-1)} \in \Omega(\tilde{\theta}_{(0:s-1)})}} f(\theta_{(0:s-1)}|V_{(0:s-1)}) f(\theta_s|\theta_{(j)}) f(V_s|\theta_s) d\theta_s d\theta_{(0:s-1)},$$

where $\Omega(\tilde{\boldsymbol{\theta}})$ is the neighborhood region around the ground truth $\tilde{\boldsymbol{\theta}}$. The expected value is approximated as the sample mean of the cost computed for all examples of the second training data set.

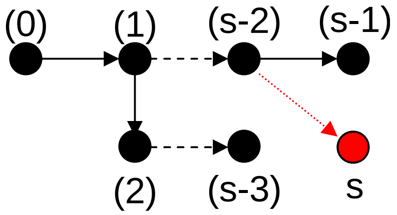


Figure 3. Illustration of the Hierarchical Detection Network (HDN) and order selection. See text for details.

3.3. Detection Scale Selection

Many previous object detection algorithms [17, 19] use a single size of image neighborhoods $\{V_i\}$. Typically, this size and corresponding search step need to be chosen a priori to balance the accuracy of the final detection result and computational speed [1]. We propose to solve this problem by hierarchical detection. During detection, larger object context is considered at coarser image resolutions resulting in robustness against noise, occlusions, and missing data. High detection accuracy is achieved by focusing the search in a smaller neighborhood at the finer resolutions. Denoting the scale parameter as λ in HDN, we treat the scale parameter λ as an extra parameter to $\boldsymbol{\theta}_s$ and use order selection to select λ as well.

4. Experiments

Our experiments are on 2D ultrasound images of left atrium and 3D ultrasound images of fetus. In both cases, we test the automatic detection order / scale selection (Section 3.2) and provide quantitative evaluation of the hierarchical detection (Section 3.3).

4.1. Sampling Strategy

In our first set of experiments, we detect five left atrium landmarks of the left atrium (LA) endocardial wall in the apical two chamber (A2C) view (Figure 1). The LA appearance is noisy since during imaging it is at the far end of the ultrasound probe. The expert annotated five landmarks in a total of 417 images. The size of the images is 120×120 pixels on average.

Three location detectors were trained independently using 281 images. The detection order for this experiment was fixed: 09 $\rightarrow$ 01 $\rightarrow$ 05 (see Figure 6 for landmark numbering). We test two different sampling strategies in detection within 136 unseen images. In the first strategy, we obtain N number of samples with the strongest weight. In the second strategy, we obtain up to $M = 2000$ samples with the strongest weight and perform k-means clustering to get N number of modes. After each landmark detection, these N samples are propagated to the next stage. The detected location is obtained by averaging the N samples for each landmark.

The number of samples, N, varies between 1 and 50. For each setting, the detection algorithm was run to obtain locations of the three landmarks. Mean of the 95% smallest errors was computed by comparing the detected locations to manual labeling. Figure 4 shows, that by using the k-means sampling strategy, the errors are lower for all number of samples. By focusing our representation on the modes of the distribution, we avoid the explosion in the number of samples that would otherwise be required [3, 18].

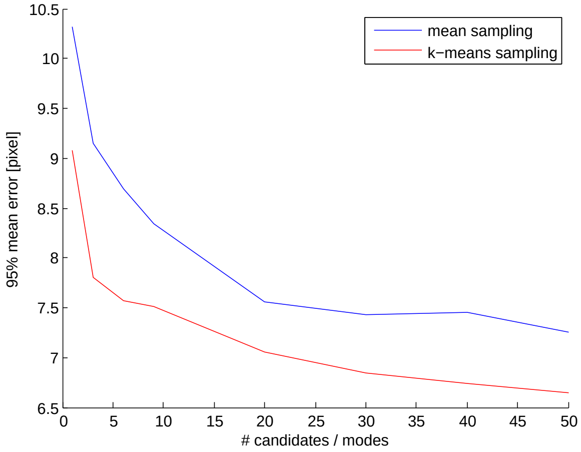


Figure 4. Sampling by obtaining N number of samples with the strongest weight or by using N strongest k-means. By focusing on the modes of the distribution, we can use small number of samples. The mean detection error is smaller.

4.2. Detection Order Selection

In the next experiment, we evaluate the automatic detection order strategy described in Section 3.2. The goal is to automatically determine the detection order of five left atrium landmarks (Figure 1). As before, the landmark detectors are trained independently using 281 annotated images. Total of 46 annotated images from the testing data set were used to obtain the detection order. The remaining 90 cases were used for detection and evaluation comparison.

Figure 5 shows the score value (normalized after each step) plotted for each stage of the 100 random cases and the automatically selected order. The greedy strategy selects order with the highest score value at each step. The final selection order or the HDN is shown in Figure 6.

The automatically selected sequential order is compared to 100 randomly generated orders. For each order, we record the final detection error averaged over all testing images and detected landmarks. We also compute score as the probability of states in the ground truth region (Eq. 15) for

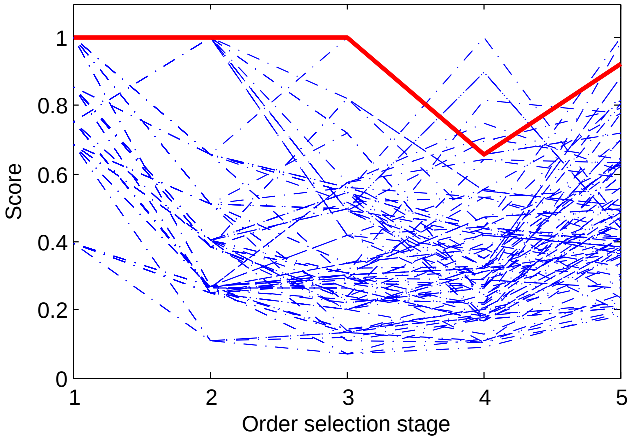


Figure 5. Selected order score values after each order selection stage. The selected order (red) has high score values across all stages. The two high score values in the final stage (see also Figure 7) have low scores at earlier stages. These detection orders were therefore not selected by the automatic algorithm.

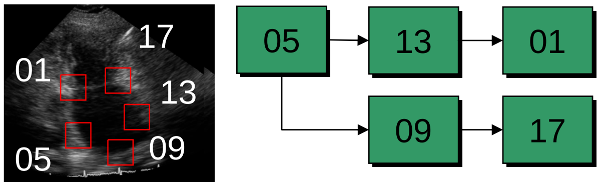


Figure 6. The final automatically selected detection order. At first, it might seem that landmarks 01 and 17 would be preferred over landmarks 5 and 13 due to the higher distinctiveness of the region. However, the high appearance variation of these landmarks causes preference of landmarks 05 and 13.

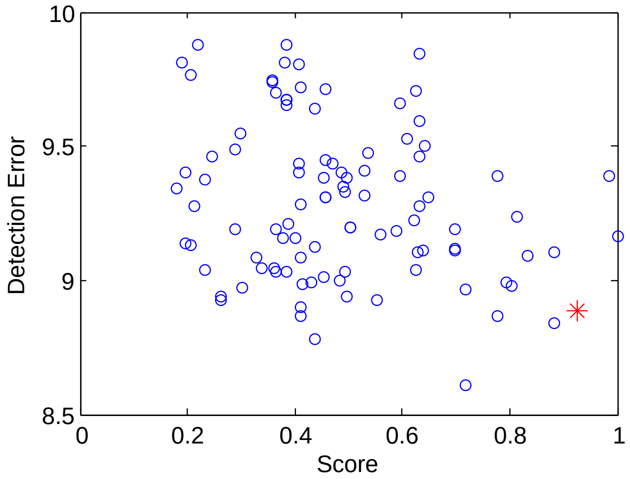


Figure 7. Comparing automatically selected order (red) to 100 randomly selected orders. The final detection errors were averaged over all testing images and detected landmarks. The score indicates preference of a particular order. The automatically selected order has a low mean detection error and a high score.

the final selection stage normalized by the maximum probability across all stages. The plot in Figure 7 shows, that the automatically selected order has low mean error (among the lowest when compared to the 100 random orders) and high probability (among the highest). The order with the highest probability was not selected due to the greedy strategy. This is because the probability of states near ground truth was low at earlier order selection. Since in real detection scenarios the ground truth is not available and sampling in low-probability regions is not reliable, these sequential orders are not preferred. Example detections are in Figure 10.

4.3. Brain Anatomies in 3d Ultrasound

Our next experiment is on detecting three fetal brain structures in 3d ultrasound data. The output of the system is a visualization of the plane with correct orientation and centering as well as biometric measurement of the anatomy. A total of 589 expert-annotated images were used for training and 295 for testing. The volumes have average size 250 × 200 × 150 mm. We use three resolutions in a hierarchical system shown in Figure 8.

Quantitative evaluation is in Table 1 and several examples of detected structures in Figure 11. The HDN average detection error 2.2 mm is lower compared to 4.8 mm error of a system without HDN.

4.4. Fetal Face in 3D Ultrasound

Our final experiment is on the detection of fetal face in 3d ultrasound volumes. A total of 962 images were used in training and 48 in testing. The gestational age of the fetus ranged from 21 to 40 weeks. The average size of the volumes is 157 × 154 × 104 mm. The major challenges of this data set include varying appearance of structures due to different developmental stage and changes in the face region caused by movement of the extremities and umbilical cord. The face was annotated by manually specifying mesh points on the face region [10]. Bounding box of the mesh specifies the pose that are automatically determined by the detection algorithm.

The system consists of three hierarchical levels with resolutions 4 mm, 2 mm, and 1 mm. The final training error was 5.48 mm and testing error 10.67 mm. The previous version of the system only operated on a single level of 1 mm which resulted in higer training and testing errors (6.90 mm and 14.10 mm respectively). Qualitative detection results are in Figure 9.

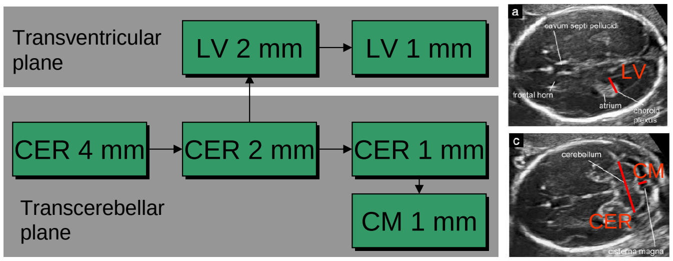


Figure 8. The detection order and the hierarchy of three brain structures: Cerebellum (CER), Cisterna Magna (CM), and Lateral Ventricles (LV). Scale selection is applied.

	mean	std	median	max	# train	# test
CER	2.289	0.884	2.213	4.197	589	295
CM	2.149	0.807	2.075	4.019	589	295
LV	2.245	0.817	2.154	3.891	589	295
CER	4.961	6.767	3.422	59.607	589	295
CM	4.989	6.832	3.519	68.679	589	295
LV	4.565	5.023	3.097	39.176	589	295

Table 1. Measurement errors of the hierarchical detection system (top part of the table) compared to an earlier system without the hierarchy [2]. Mean error, standard deviation, median error, and maximum error are computed. The system was trained using number of volumes specified in the 6th column and tested on the number of volumes specified in the 7th column. The average detection error using the hierarchy is 2.2 mm on data with 1 mm finest resolution. The average error of the system without the hierarchy is 4.8 mm.

5. Conclusion

We have presented a Sequential Monte Carlo based Hierarchical Detection Network (HDN) for detecting multiple objects. The order of detection is automatically determined by a greedy algorithm that puts the most reliable detections earlier in the detection sequence. The detectors are orga-

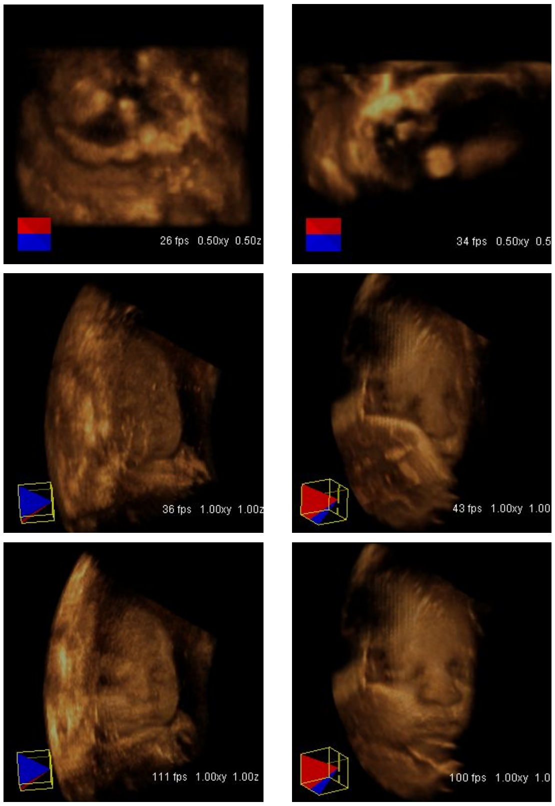


Figure 9. Example results of the fetal face detection using a hierarchy of three resolutions. Initial pose after loading the volume (top row), after automatic detection at the finest level (middle row), and after volume carving of the region in front of the face (bottom row).

nized in a multi-scale hierarchy with the scale parameter included in the order selection process. We have shown the effectiveness of the automatic order selection process on the detection of five left atrium landmarks in 2D ultrasound images. The multi-scale hierarchical detectors have higher detection accuracy than systems based on a single level as we demonstrated on detection of fetal face and three fetal brain structures in 3D ultrasound images.

The described framework opens up several possible avenues of future research. One area we are particularly interested in is how to include dependence on multiple objects at each detection stage. This will result in a stronger geometrical constraint and therefore improve performance on objects that are difficult to detect by exploiting only the pairwise dependence.

References

[1] N. J. Butko and J. R. Movellan. Optimal scanning for faster object detection. In *Proc. CVPR*, pages 2751–2758, Miami, FL, 20–25 June 2009. 2, 4

[2] G. Carneiro, F. Amat, B. Georgescu, S. Good, and D. Comaniciu. Semantic-based indexing of fetal anatomies from 3-D ultrasound data using global/semi-local context and sequential sampling. In *Proc. CVPR*, Anchorage, AK, 24–26 June 2008. 1, 6

[3] T.-J. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. In *Proc. CVPR*, volume 2, pages 239–245, 1999. 5

[4] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. CVPR*, volume 1, pages 10–17, 2005. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 1, pages 886–893, 2005. 1

[6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *Proc. ICCV*, 2009. 2

[7] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *Proc. CVPR*, pages 1271–1278, Miami, FL, 20–25 June 2009. 2

[8] A. Doucet, N. D. Freitas, and N. Gordon. *Sequential Monte Carlo methods in practice*. Birkhäuser, 2001. 2

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Anal. Machine Intell.*, 2010. To Appear. 1, 2

[10] S. Feng, S. Zhou, S. Good, and D. Comaniciu. Automatic fetal face detection from ultrasound volumes via learning 3d and 2d information. In *Proc. CVPR*, pages 2488–2495, Miami, FL, 20–25 June 2009. 6

[11] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *Proc. CVPR*, pages 1–8, Anchorage, AK, 24–26 June 2008. 2

[12] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, Oct. 2008. 10.1007/s11263-008-0137-5. 2

[13] S. Kumar and M. Hebert. Discriminative random fields: a discriminative framework for contextual interaction in classification. In *Proc. ICCV*, volume 2, pages 1150–1157, 2003. 2

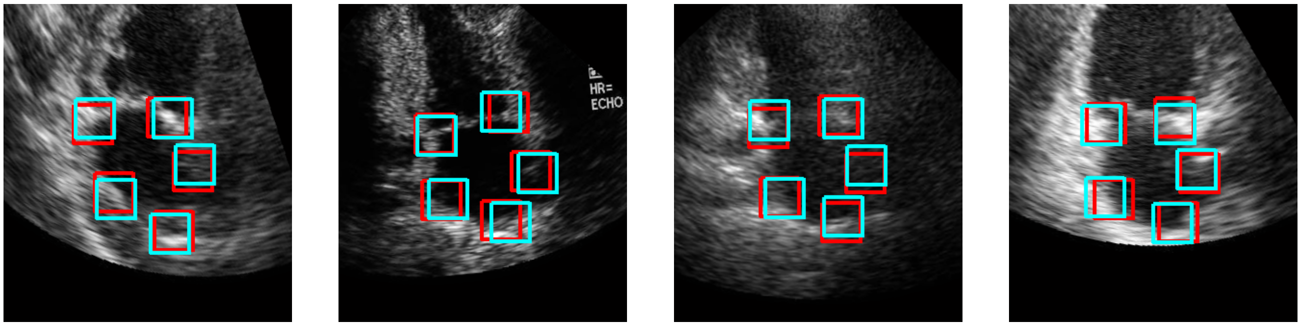
[14] J. S. Liu, R. Chen, and T. Logvinenko. A theoretical framework for sequential importance sampling with resampling. In A. Doucet, N. D. Freitas, and N. Gordon, editors, *Sequential Monte Carlo methods in practice*, pages 225–242. Birkhäuser, 2001. 2, 3

[15] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Proc. Eigth ECCV*, pages 28–39, 2004. 1

[16] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. CVPR*, volume 2, pages 762–769, 2004. 2

[17] Z. Tu. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In *Proc. ICCV*, volume 2, pages 1589–1596, 2005. 2, 4

[18] J. Vermaak, A. Doucet, and P. Perez. Maintaining multimodality through mixture tracking. In *Proc. CVPR*, volume 2, pages 1110–1116, 2003. 5



(6.39, 6.91, 4.64, 7.21, 6.26) (2.42, 6.84, 9.95, 7.33, 8.41) (7.94, 5.03, 7.03, 8.18, 5.00) (5.24, 9.83, 6.16, 5.12, 7.71)

Figure 10. Final sequential detection result (cyan) compared to ground truth (red). Notice that the landmarks are accurately detected despite the noise, high appearance and shape variations, and shadowing effects. The landmark detection errors (in pixels) are shown below each image in the left-bottom-right order.

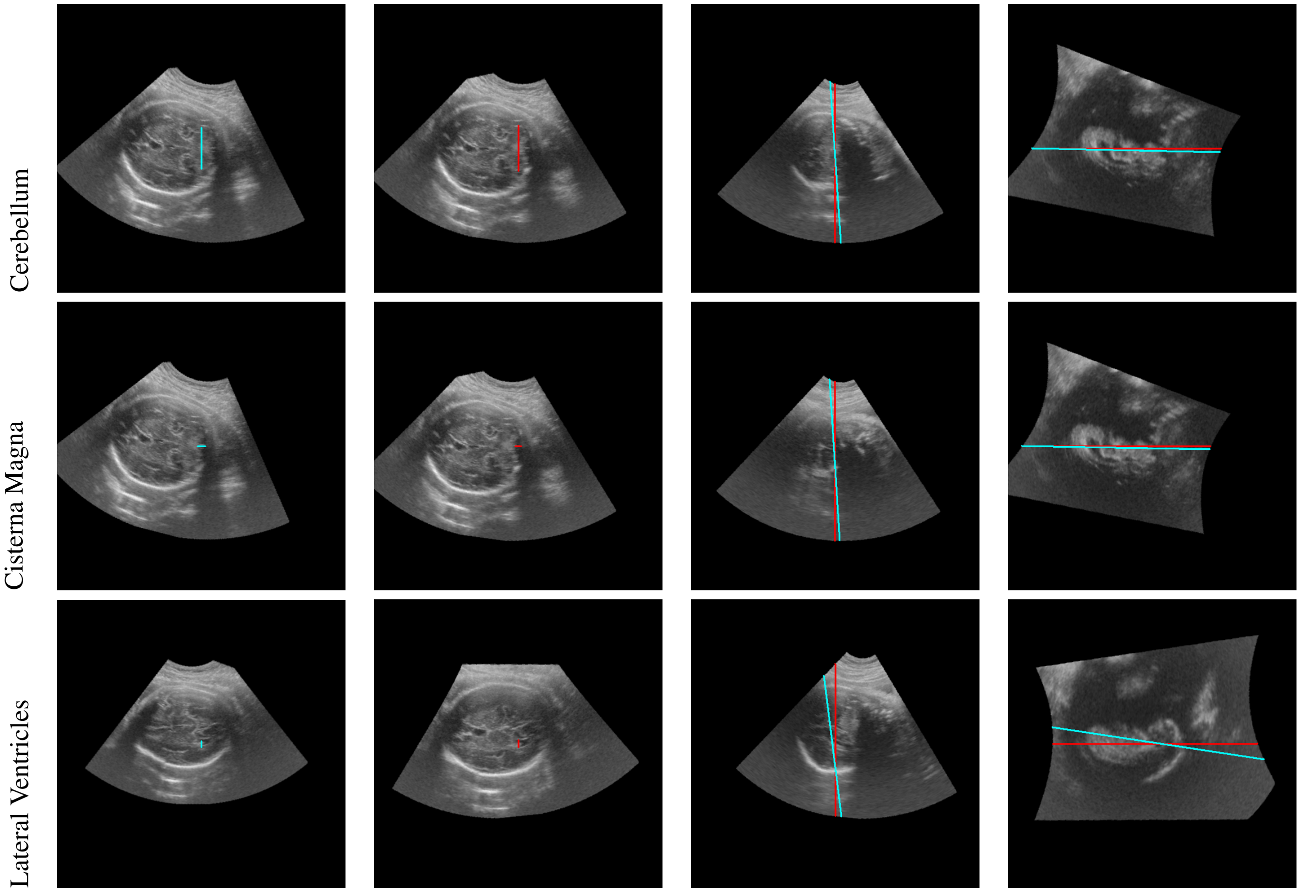


Figure 11. Final hierarchical detection (Figure 8) result (cyan) compared to ground truth (red). The last two columns show the agreement of the detection plane in the sagittal and coronal cross section.

[19] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comp. Vis.*, 57(2):137–154, 2004. 1, 2, 4

[20] B. Wu and R. Nevatia. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. *International Journal of Computer Vision*, 82(2):185–204, Apr. 2009. 1

[21] Y. Zhan, X. Zhou, Z. Peng, and A. Krishnan. Active scheduling of organ detection and segmentation in whole-body medical images. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2008*, pages 313–321, 2008. 2