Integrated Detection Network for Multiple Object Recognition

Michal Sofka[a,b]

[a] *Cisco Systems*
Charles Square Center, Karlovo namesti 10
120 00 Prague 2, Czech Republic
[b] *Czech Technical University*
Dept. of Computer Science, Karlovo namesti 13
121 35 Prague 2, Czech Republic

Abstract

Recognizing multiple objects involves two inter-dependent tasks, object localization and classification. The goal of the object localization is to accurately find the object pose parameters relative to an established reference, such as the origin of the image coordinate system. The object classification assigns class labels to the objects according to the pre-specified categories. Multi-object recognition has been previously solved by designing a set of individual single-object detectors or by training a combined multi-object detection and classification system. In the medical domain, these models can be further improved by relying on strong spatial prior information present in medical images of a human body. This chapter describes, how the spatial prior can be used to recognize multiple anatomical structures which results in the Integrated Detection Network. The structures are recognized sequentially, one-by-one, using optimal order such that the later recognitions can benefit from constraints provided by previously recognized structures. The recognition relies on Sequential Estimation techniques, with the posterior distribution of the structure pose and label being approximated at each step by sequential Monte Carlo. The samples are propagated within the sequence across multiple structures and hierarchical levels. The system is general and provides accurate recognitions of anatomical structures in 3D images of various modalities.

Keywords: object recognition, object detection, object classification, discriminative learning, sequential sampling

Email address: msofka@cisco.com (Michal Sofka)

1. Introduction

Recognizing multiple objects or anatomical structures has many applications in medical imaging systems, for example in multi-object visual tracking, to initialize segmentation of the structures, or to provide accurate measurements. The goal of the recognition is to find the pose parameters of all structures of interest relative to the origin of the image coordinate system, camera, or other structures and to assign the correct object class label to the structures. In the first step, a model is constructed from a set of training examples to capture the statistics of the object class. The parameters are then found during inference by applying the model on a new image. In generative models, the probability model consists of the object appearance variations conditioned on the pose and a probability model of appearance variations of the background along with the prior probabilities of each class. The inference is computed by evaluating the posterior probabilities using the Bayes' theorem. In practice, it is very hard to model all variations of the structures since the structure shape, shadows and occlusions, image characteristics, and acquisition parameters can vary significantly. Discriminative models are more tractable. The probability model consists of an object presence or class label conditioned on the appearance. The inference typically computes the posterior probability given a set of discrete parameter values. Multi-scale and sampling techniques are used to make the process efficient.

The robustness to the variations in the photometric appearance of the structures is achieved by features that are invariant to intensity transformations. One example of such features are Haar features that compare intensity statistics computed in adjacent rectangular windows of various configurations (Viola and Jones, 2004). Another popular feature types are Local Binary Patterns (LBP) that relate the intensity values of a pixel to that of its neighbors. When combined with Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005), invariance to small geometric transformations is achieved by computing local histograms of the features. Larger geometric transformations can be directly modeled by decomposing a structure into parts (Felzenszwalb et al., 2010) and considering the statistical relationships between the parts and the structure of interest. Local detectors can then be improved by modeling the interdependence of objects using contextual (Desai et al., 2011; Kumar and Hebert, 2006; Hoiem et al., 2008) and semantic information. Spatial information can also be disregarded altogether resulting in a bag-of-features model (Lampert et al., 2009). As an alter-

native to manually engineered features, hierarchical feature representations can be learned from large databases (Sermanet et al., 2014).

State-of-the-art approaches to multi-object detection (Viola and Jones, 2004; Felzenszwalb et al., 2010) 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. This is done implicitly in Deep Neural Networks, where the relationships are encoded by hidden layers (Sermanet et al., 2014). Obtaining a joint model of multiple objects involves the estimation of a large number of parameters which increases the requirements on the training time and the size of the annotation database. In situation, where this is not practical, the multi-object detection task has been solved by multiple individual object detectors connected by a spatial model. 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. Modeling long-range dependencies is straightforward in these models. The most challenging aspect of these algorithms is designing detectors that are fast and robust, modeling the spatial relationships between objects, and determining the order of object dependencies. In this chapter, we propose a multi-object recognition system that addresses these challenges.

The exposition starts in Section 2 by presenting a general framework for multi-object recognition without considering contextual dependencies between objects. The recognition is acomplished by a discriminative appearance model of each individual object. Section 3 then describes a Sequential Sampling framework, where the interdependence between objects is modeled by a transition distribution. The distribution specifies the "transition" of a pose of one object to a pose of another object as detailed in Section 3.1. This process relies on the strong prior information present in medical images of a human body. Together, all detectors and any associated processing form the Integrated Detection Network (IDN) presented in Section 3.2. Section 3.3 explains how to determine the size of the context region (detection scale) and which objects to detect first in an optimal way. The chapter concludes by highlighting examples of IDN applications in Section 4 and by final remarks in Section 5.

2. Independent Multi-Object Recognition

This section starts by discussing the independent recognition, where the spatial relationships between objects are not explicitly modeled. The next section then describes a technique that takes advantage of previously recognized objects to improve the recognition of a new object.

The state of the modeled object s is denoted as $\boldsymbol{\theta}_s$, where $\boldsymbol{\theta}_s = \{\boldsymbol{\pi}_s, y_s\}$. The first term, $\boldsymbol{\pi}_s$, denotes the pose $\boldsymbol{\pi}_s = \{\mathbf{p}, \mathbf{r}, \mathbf{s}\}$ with the position $\mathbf{p}$, orientation $\mathbf{r}$, and size $\mathbf{s}$ of the object s. The second term, y_s, denotes the object label. The set of observations for an object s is obtained from the image neighborhood V_s. The neighborhood V_s is specified by the coordinates of a bounding box within an N-dimensional image V, $V : R^N \rightarrow I$, where I is the image intensity. The images are typically two or three dimensional. The observations computed from V_s are features with a likelihood $f(V_s|\boldsymbol{\theta}_s)$. They represent the appearance of each object and are assumed conditionally independent given the state $\boldsymbol{\theta}_s$.

The task of object recognition consists of detecting the object instance inside the image V and identifying the object class both of which are accomplished by using observations computed from the image. The object s is detected by estimating the pose parameters $\boldsymbol{\pi}_s$ and classified by assigning the label y_s. The likelihood $f(V_s|\boldsymbol{\theta}_s)$ can be formulated as:

$$f(V_s|\boldsymbol{\theta}_s) = f(V_s|\boldsymbol{\pi}_s, y_s) = f(y_s, \boldsymbol{\pi}_s|V_s)\frac{f(V_s)}{f(y_s, \boldsymbol{\pi}_s)} = f(y_s|\boldsymbol{\pi}_s, V_s)f(\boldsymbol{\pi}_s|V_s)\frac{f(V_s)}{f(y_s, \boldsymbol{\pi}_s)}. \tag{1}$$

The term $f(y_s|\boldsymbol{\pi}_s, V_s)$ denotes the posterior of the object class label with object pose $\boldsymbol{\pi}_s$ given the observations from V_s. The term $f(\boldsymbol{\pi}_s|V_s)$ is the posterior of the pose given observations. The term $f(y_s, \boldsymbol{\pi}_s)$ denotes the prior on the labels and pose parameters and is estimated from the training data. The term $f(V_s)$ is set to a uniform distribution.

The object recognition can be accomplished by sliding a window, where the window defines the neighborhood V_s at each step. The observations from V_s are used to classify the window by assigning the class label. The object classifier is therefore represented by the model $f(y_s|V_s)$ which is the posterior of the object class within the image neighborhood V_s:

$$f(y_s|V_s) = \int_{\boldsymbol{\pi}_s} f(y_s|\boldsymbol{\pi}_s, V_s)f(\boldsymbol{\pi}_s|V_s)d\boldsymbol{\pi}_s. \tag{2}$$

In practice, the probability of the anatomical structure s being detected is evaluated using a discrete set of pose parameter values $\{\boldsymbol{\pi}_s\}$ and a binary (object vs. background) or multi-object classifier.

The set of best instance parameters $\hat{\boldsymbol{\theta}}_s = \{\hat{\boldsymbol{\pi}}_s, \hat{y}_s\}$ for each object s is then estimated using the observations from V_s:

$$\{\hat{\boldsymbol{\pi}}_s, \hat{y}_s\} = \arg\max_{y_s, \boldsymbol{\pi}_s} P(y_s|\boldsymbol{\pi}_s, V_s) P(\boldsymbol{\pi}_s|V_s). \tag{3}$$

To leverage the power of a large annotated dataset, discriminative classifier (PBT (Tu, 2005)) is used to best decide between positive and negative examples of the object. PBT combines a binary decision tree with boosting, letting each tree node be an AdaBoost classifier. This way, the miss-classified positive or negative examples early on can still be correctly classified by children nodes. Other classification approaches can be used as well.

3. Sequential Sampling for Multi-Object Recognition

Similarly to the individual detection of single objects, the goal of the multi-object detection is to estimate the likelihood of the observations given object parameters. The sequence of parameters of multiple objects is denoted as $\boldsymbol{\theta}_{0:s} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_s\}$ and the sequence of volumes to compute the observations as $V_{0:s} = \{V_0, V_1, \ldots, V_s\}$. It is possible to construct such sequence since there exists prior knowledge for determining the image neighborhoods $V_0, V_1, \ldots, V_s$. The image neighborhoods in the sequence $V_{0:s}$ might overlap and can have different sizes (Figure 1). An image neighborhood V_i might even be the entire volume V. The order in this sequence is determined manually based on the expert knowledge or automatically based on the posterior probability of object poses in the ground truth region (see more details below).

It is clear that the conditional likelihood model $P(V_{0:s}|\boldsymbol{\theta}_{0:s})$ is now much more complicated. The posterior of the object classes $f(y_{0:s}|\boldsymbol{\pi}_{0:s}, V_{0:s})$ involves the dependence of all instance labels jointly on all pose parameters and all observations. Such a large search space is computationally prohibitive both in training and in inference. Since the likelihood models in practical situations lead to intractable exact inference, approximation by Monte Carlo methods, also known as particle filtering or sequential estimation, has been widely adopted.

Sequential Estimation techniques (Doucet et al., 2001), estimate the object state $\boldsymbol{\theta}_s$ using observations from $V_{0:s}$ in a sequential spatial order. This way, the posterior distribution of the parameters (state) of each anatomical structure is estimated based on all observations so far. This concept is used to solve the multi-object detection problem by recursively applying *prediction* and *update* steps to obtain the posterior distribution $f(\boldsymbol{\theta}_{0:s}|V_{0:s})$. The

prediction step computes the probability density of the state of the object s using the state of the previous object, $s-1$, and previous observations of all objects up to $s-1$:

$$f(\boldsymbol{\theta}_{0:s}|V_{0:s-1}) = f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1})f(\boldsymbol{\theta}_{0:s-1}|V_{0:s-1}). \tag{4}$$

The state dynamics, *i.e.* relationships between object poses, are modeled with an initial distribution $f(\boldsymbol{\theta}_0)$ and a transition distribution $f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1})$. Note that here the first-order Markov transition $f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1})$ is not used since any detected object can depend on any other previously detected object. When detecting the object s, the observation V_s is used to compute the estimate during the update step as:

$$f(\boldsymbol{\theta}_{0:s}|V_{0:s}) = \frac{f(V_s|\boldsymbol{\theta}_s)f(\boldsymbol{\theta}_{0:s}|V_{0:s-1})}{f(V_s|V_{0:s-1})}, \tag{5}$$

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

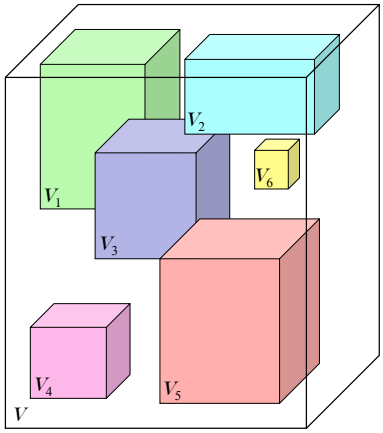

Figure 1: In multi-object detection, the set of observations is a sequence of image patches $\{V_s\}$. The sequence specifies a spatial order of structures. The structures are detected in this order which is automatically determined.

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

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

than sampling the parameter space uniformly (Tu, 2005; Viola and Jones, 2004), since sampling from a the proposal distribution (Liu et al., 2001) focuses on regions of high probability. This saves computational time as fewer samples are required and increases robustness compared to the case, where the same number of samples would be drawn uniformly.

It is now useful to discuss the concept of weighted samples. A set of weighted random samples $\{\boldsymbol{\theta}^j_{0:s}, w^j_s\}^m_{j=1}$ is called *proper* with respect to f, if for any square integrable function $h(\cdot)$ (Doucet et al., 2001)

$$E[h(\boldsymbol{\theta}^j_{0:s})w^j_s] = cE_f h(\boldsymbol{\theta}_{0:s}), \tag{6}$$

where c is a normalizing constant common to all m samples. Note, that the $\boldsymbol{\theta}$ estimated as $\sum^m_{j=1} w^j_s h(\boldsymbol{\theta}^j_{0:s})$ does not depend on the normalizing constant of f, i.e. c does not need to be known. In order for the samples from the proposal $p(\boldsymbol{\theta}_{0:s}|V_{0:s})$ to be proper, the weights are defined as

$$\begin{aligned} \tilde{w}^j_s &= \frac{f(V_{0:s}|\boldsymbol{\theta}^j_{0:s})f(\boldsymbol{\theta}^j_{0:s})}{p(\boldsymbol{\theta}^j_{0:s}|V_{0:s})} \\ w^j_s &= \tilde{w}^j_s / \sum_{i=1}^{m} \tilde{w}^i_s. \end{aligned} \tag{7}$$

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

$$p(\boldsymbol{\theta}_{0:s}|V_{0:s}) = p(\boldsymbol{\theta}_{0:s-1}|V_{0:s-1})p(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1}, V_{0:s}). \tag{8}$$

Note, that V_s was left out of the first term since the states in the sequence $\boldsymbol{\theta}_{0:s-1}$ do not depend on it. The states are computed as

$$f(\boldsymbol{\theta}_{0:s}) = f(\boldsymbol{\theta}_o)\prod_{j=1}^{s} f(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{0:j-1}). \tag{9}$$

Substituting (8) and (9) into (7), we have

$$\tilde{w}^j_s = \frac{f(V_{0:s}|\boldsymbol{\theta}^j_{0:s})f(\boldsymbol{\theta}^j_{0:s})}{p(\boldsymbol{\theta}^j_{0:s-1}|V_{0:s-1})p(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1}, V_{0:s})} \tag{10}$$

$$= \tilde{w}^j_{s-1}\frac{f(V_{0:s}|\boldsymbol{\theta}^j_{0:s})f(\boldsymbol{\theta}^j_{0:s})}{f(V_{0:s-1}|\boldsymbol{\theta}^j_{0:s-1})f(\boldsymbol{\theta}^j_{0:s-1})p(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1}, V_{0:s})} \tag{11}$$

$$= \tilde{w}^j_{s-1}\frac{f(V_s|\boldsymbol{\theta}^j_s)f(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1})}{p(\boldsymbol{\theta}^j_s|\boldsymbol{\theta}^j_{0:s-1}, V_{0:s})}. \tag{12}$$

In this chapter, the transition prior $f(\boldsymbol{\theta}_s^j|\boldsymbol{\theta}_{0:s-1}^j)$ is adopted as the proposal distribution. Compared to the more general proposal, $p(\boldsymbol{\theta}_s^j|\boldsymbol{\theta}_{0:s-1}^j, V_{0:s})$, the most recent observation is missing. In practice, this does not pose a problem in detection since the predicted samples are near the likelihood peaks. The importance weights are then calculated as:

$$\tilde{w}_s^j = \tilde{w}_{s-1}^j f(V_s|\boldsymbol{\theta}_s^j). \tag{13}$$

Other proposal distributions to leverage relations between multiple objects can also be designed.

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

1. Obtain m samples from the proposal distribution, $\boldsymbol{\theta}_s^j \sim p(\boldsymbol{\theta}_s^j|\boldsymbol{\theta}_{0:s-1}^j)$.
2. Reweight each sample according to the importance ratio

$$\tilde{w}_s^j = \tilde{w}_{s-1}^j f(V_s|\boldsymbol{\theta}_s^j). \tag{14}$$

   Normalize the importance weights.
3. Resample the particles using their importance weights to obtain more particles in the peaks of the distribution. Finally, compute the approximation of $f(\boldsymbol{\theta}_{0:s}|V_{0:s})$:

$$f(\boldsymbol{\theta}_{0:s}|V_{0:s}) \approx \sum_{j=1}^{m} w_s^j \delta(\boldsymbol{\theta}_{0:s} - \boldsymbol{\theta}_{0:s}^j), \tag{15}$$

   where δ is the Dirac delta function.

3.1. The Observation and Transition Models

The key components of the sequential sampling framework are the *observation* and *transition* models. The observation model $f(V_s|\boldsymbol{\theta}_s)$ in the update step describes the appearance of each object and is obtained from Eq. 1. This corresponds to the likelihood of a hypothesized state that gives rise to observations. As mentioned earlier, the model is based on a deterministic model learned using a large annotated database of images. The transition model in the prediction step describes the way states are propagated between the image neighborhoods. Relying on the anatomical context the transition kernel is based on a pairwise dependency

$$f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{0:s-1}) = f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_j), \quad j \in \{0, 1, \ldots, s-1\}. \tag{16}$$

Please note that a state of *any* previously detected object is used to compute the transition. This is less restrictive than a Markovian process, $f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s-1})$, which would always use the immediate precursor. The distribution $f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_j)$ is modeled as a Gaussian estimated from the training data. The statistical model captures spatial relationships between the structures while ignoring abnormal configurations that may be caused by a disease progression. During detection, the predictions are used as the best available estimates even for abnormal cases.

3.2. Integrated Detection Network (IDN)

The computational speed and robustness of the recognition system is increased by hierarchical processing. Further performance improvements are obtained by starting from structures that are easier to detect and constraining the detection of the other structures by exploiting spatial configurations. This design results in a large number of observation and transition models such that multiple structures can be efficiently recognized. The models and any intermediate processing are managed by the Integrated Detection Network (IDN). As shown in Figure 2(left), IDN is a pairwise, feed-forward network. IDN consists of *nodes* that perform operations on the input *data* and produce zero or more output data. The operations, such as candidate sample detection, propagation, and aggregation, are only related to each other through data connections. This makes it possible to easily add new nodes and data types to an existing network.

In detection, one major problem is how to effectively propagate detection candidate samples 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 cost, lower accuracy, or drift of the coarse candidates towards incorrect refinements. The search range in IDN is part of the model that is learned from the training data. One difficulty of sequential processing of multiple structures is in selecting the order of detections such that the overall performance is maximized. IDN detection schedule is designed to minimize the uncertainty of the detections as described in the next section.

3.3. Detection Order Selection

The spatial order of detections in IDN is automatically determined during training. The goal is to select the order such that the posterior probability $P(\boldsymbol{\theta}_{0:s}|V_{0:s})$ is maximized in the neighborhood region around the ground truth. Since determining this order has exponential complexity in the number of objects, a greedy approach is adopted. The training data is first

split into two sets. Using the first set, all object detectors are trained individually to obtain posterior distributions $f(\boldsymbol{\theta}_0|V_0), f(\boldsymbol{\theta}_1|V_1), \ldots, f(\boldsymbol{\theta}_s|V_s)$. The second set is used for order selection as illustrated in Figure 2(right) as follows.

Suppose that the detection order is determined up to $s-1$, $\boldsymbol{\theta}_{(0)}, \boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(s-1)}$. The order selection aims 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:

$$S[s,(j)] = \int\limits_{\substack{\boldsymbol{\theta}_s \in \Omega(\tilde{\boldsymbol{\theta}}_s) \\ \boldsymbol{\theta}_{(0:s-1)} \in \Omega(\tilde{\boldsymbol{\theta}}_{(0:s-1)})}} f(\boldsymbol{\theta}_{(0:s-1)}|V_{(0:s-1)}) f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{(j)}) f(V_s|\boldsymbol{\theta}_s) d\boldsymbol{\theta}_s d\boldsymbol{\theta}_{(0:s-1)}, \tag{17}$$

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.

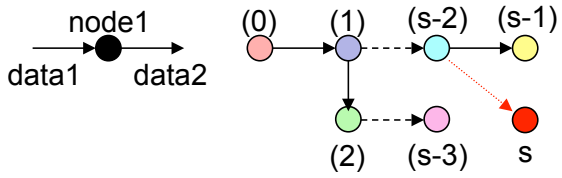

Figure 2: Integrated Detection Network (IDN) consists of *nodes* that operate on *data* (left). Illustration of the Integrated Detection Network (IDN) and order selection (right). See text for details.

During hierarchical 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. The resolution level and the size of the image neighborhoods $\{V_i\}$ can be selected using the same mechanism as the order selection by introducing additional parameters (Sofka et al., 2014). Choosing the scale automatically is advantageous since objects have different sizes and the size of the context neighborhood is also different.

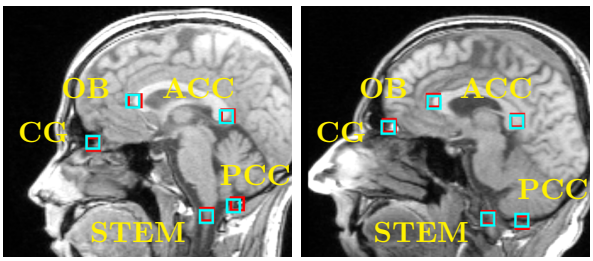

Figure 3: Automatic recognition results (blue) and ground truth reference (red) of five landmarks in two brain MRI scans: Crista galli (CG), occipital bone (OB), the anterior of the corpus callosum (ACC), the posterior of the corpus callosum (PCC), and the brain stem (STEM).

4. Applications

This section highlights applications where IDN is effective in recognizing multiple anatomical structures. The first application is to recognize landmarks in brain MRI scans (Sofka et al., 2012). A total of 384 volumes were used training and 127 for testing with the average volume size of $130 \times 130 \times 101$ voxels after resampling to 2 mm isotropic resolution. In each volume, the system detects crista galli (CG), occipital bone (OB), the anterior of the corpus callosum (ACC), the posterior of the corpus callosum (PCC), and the brain stem (STEM). The average detection error is 2.37 mm. Example detection are shown in Figure 3.

The second application shows how to automatically detect and measure anatomical structures in fetal head ultrasound volumes (Sofka et al., 2014). A total of 1982 volumes were used for training and 107 for testing. The average volume size was $186 \times 123 \times 155$ voxels after resampling to 1 mm isotropic resolution. The IDN produced a standardized visualization plane with correct orientation and centering as well as the biometric measurement of the anatomy. The plane parameters and the measurement were derived from the pose of the anatomical structure. The following measurements were obtained (Figure 4): Cerebellum, Cisterna Magna, Lateral Ventricles, Occipitofrontal Diameter, Biparietal Diameter, and Head Circumference. The average measurement error was below 2 mm and within the inter-user variability.

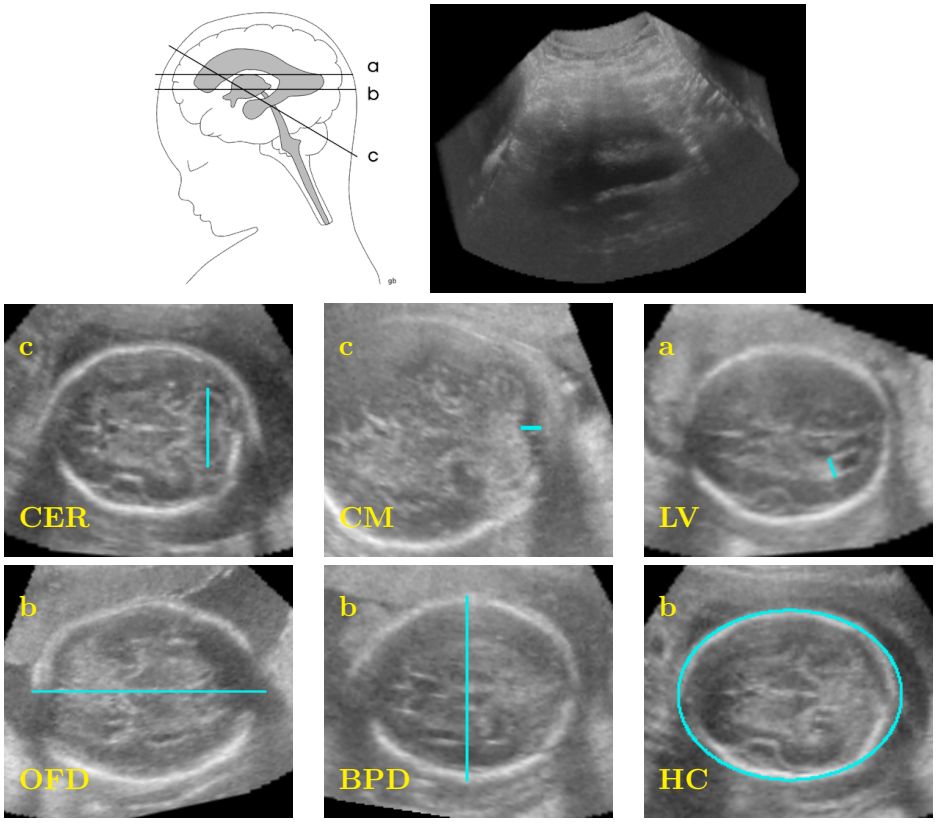

Figure 4: Fetal Head and Brain (AFHB) system provides automatic measurements at three standardized planes: Ventricular (a), Thalamic (b), and Cerebellar (c) from a 3D ultrasound volume. Shown are example results for Cerebellum (CER), Cisterna Magna (CM), Lateral Ventricles (LV), Occipitofrontal Diameter (OFD), Biparietal Diameter (BPD), and Head Circumference (HC).

5. Conclusions

This chapter presented the Integrated Detection Network (IDN) for recognizing multiple objects by exploiting their relative spatial configurations. Modeling interdependence of objects introduces additional constraints that make it possible to achieve high localization accuracy. The approach is motivated by Sequential Estimation techniques that estimate a spatial order of probability distributions for a sequence of objects. The computation requires a likelihood of a hypothesized state (object pose and label) that gives rise to observations and a transition model that describes the way the states

are propagated between objects. Sampling techniques have been used to approximate the posterior distribution and make the modeling tractable. At each step, the prediction step involves sampling from the proposal distribution of the current state conditioned on the history of states and the history of observations. The posterior distribution of the pose (state) of each anatomical structure is then estimated during the update step based on the prediction and all observations so far. The observations are features computed from image neighborhoods surrounding the anatomies. 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 anatomical structures are related is Gaussian.

The modular nature of the IDN makes it straightforward to adopt different observation and transition models. These models can capture more intricate object relationships (e.g. context from multiple previously detected objects) or introduce application-specific constraints. All these properties contribute to the improved detection and classification accuracy which makes the IDN attractive choice for multi-object recognition tasks.

References

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proc. CVPR. Vol. 1. pp. 886–893.

Desai, C., Ramanan, D., Fowlkes, C., 2011. Discriminative models for multi-class object layout. International Journal of Computer Vision 95 (1), 1–12.

Doucet, A., Freitas, N. D., Gordon, N., 2001. Sequential Monte Carlo methods in practice. Birkhäuser.

Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D., Sep. 2010. Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Machine Intell. 32 (9), 1627–1645.

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

Kumar, S., Hebert, M., 2006. Discriminative random fields. International Journal of Computer Vision 68 (2), 179–201.

Lampert, C. H., Blaschko, M., Hofmann, T., Dec 2009. Efficient subwindow search: A branch and bound framework for object localization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (12), 2129–2142.

Liu, J. S., Chen, R., Logvinenko, T., 2001. A theoretical framework for sequential importance sampling with resampling. In: Doucet, A., Freitas, N. D., Gordon, N. (Eds.), Sequential Monte Carlo methods in practice. Birkhäuser, pp. 225–242.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., April 2014. Overfeat: Integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (ICLR 2014).

Sofka, M., Ralovich, K., Zhang, J., Zhou, S., Comaniciu, D., 2012. Progressive data transmission for anatomical landmark detection in a cloud. Methods of Information in Medicine 51 (3), 268–278.

Sofka, M., Zhang, J., Good, S., Zhou, S. K., Comaniciu, D., May 2014. Automatic detection and measurement of structures in fetal head ultrasound volumes using Sequential Estimation and Integrated Detection Network (IDN). IEEE Transactions on Medical Imaging 33 (5), 1054–1070.

Tu, Z., 2005. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In: Proc. ICCV. Vol. 2. pp. 1589–1596.

Viola, P., Jones, M. J., 2004. Robust real-time face detection. International Journal of Computer Vision 57 (2), 137–154.