

# Progressive Data Transmission for Hierarchical Detection in a Cloud

Michal Sofka, Kristof Ralovich, Jingdan Zhang, S. Kevin Zhou,  
and Dorin Comaniciu

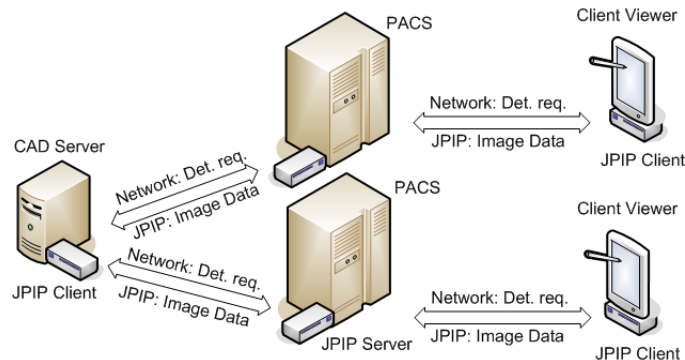
Siemens Corporate Research  
755 College Road East, Princeton, NJ 08540, USA

**Abstract.** In response to the growing need for image analysis services in the cloud computing environment, this paper proposes an automatic system for detecting landmarks in 3D volumes. The inherent problem of limited bandwidth between a (thin) client, Data Center (DC), and Data Analysis (DA) server is addressed by a hierarchical detection algorithm that obtains data by progressively transmitting only image regions required for processing. The client sends a request for a visualization of a specific landmark. The algorithm obtains a coarse level image from DC and outputs landmark location candidates. The coarse landmark location candidates are then used to obtain image neighborhood regions at a finer resolution level. The final location is computed as the robust mean of the strongest candidates after refinement at the subsequent resolution levels. The feedback about candidates detected at a coarser resolution makes it possible to only transmit image regions surrounding these candidates at a finer resolution rather than the entire images. Furthermore, the image regions are *lossy* compressed with JPEG 2000. Together, these properties amount to at least 30 times bandwidth reduction while achieving similar accuracy when compared to an algorithm using the original data.

## 1 Introduction

In this paper, we propose a system of algorithms for automatic detection of anatomical landmarks in 3D volumes in the cloud computing environment. The system, dubbed *Detection in a Cloud (DiC)*, is used by thin-client devices that request the display of an anatomical part for a specific patient. The patient data stored in a Data Center are transmitted to a high performance Data Analysis server that runs the detection algorithm. (In the medical domain and also in this paper, these servers are referred to as the Picture Archiving and Communication System (PACS) and Computer Aided Detection (CAD) server, respectively). The image with the anatomy highlighted is returned back to the client for display (Figure 1).

Inherent difficulties in designing such system are large image sizes (often hundreds of megabytes) and limited bandwidth between thin client, Data Analysis server (e.g. CAD server), and Data Center (e.g. PACS server). Depending on the bandwidth, the transmission of large datasets can take tens of seconds or even minutes. This complicates the workflow in interactive applications where the



**Fig. 1.** *Detection in a Cloud (DiC) system.* The client sends a request for the detection of an anatomy for a specific patient. The detection algorithm on the CAD server requests data from the PACS server. The detection results are efficiently visualized by the client via JPEG 2000 JPIP protocol.

results must be available immediately. Finally, limited memory and insufficient CPU power of the client necessitates remote data processing.

Since it is not possible to process the data on the client directly, an obvious solution is to efficiently transmit the data from the PACS server to the CAD server for processing. However, such procedure becomes prohibitive already when several detection requests are made simultaneously since this would require bandwidths of tens of GBits / second. This is true even when the data is compressed with the lossless JPEG 2000 (Section 4).

We propose an efficient hierarchical learning-based detection system to avoid the problem of transmitting large datasets. The system runs on the CAD server that obtains portions of the original dataset from the PACS server on demand. The algorithm starts detection on a downsampled low-resolution image that has been compressed and transmitted to the CAD server. The coarse landmark candidate positions define the regions in a finer resolution image, where the coarse candidates are refined. The refinement steps continue until all levels of the hierarchy have been processed. The final detection result is obtained by robustly combining strongest candidates from the finest level.

The amount of transmitted data is significantly reduced in the DiC system. First, the algorithm only processes candidate *regions* at finer resolutions rather than the *entire* images. Second, all image regions are compressed with a lossy compression. When combined, these properties result in an overall reduction of the original data size by a factor of 30 (CT data) and by a factor of 196 (MRI data). Our experiments show that the lossy compression does not hinder the final detection accuracy. The experiments also demonstrate the robustness and accuracy of the hierarchical algorithm and advantages of training on compressed images.

In summary, the paper makes three main contributions: (1) an overall system for landmark detection using remote datasets, (2) hierarchical detection

algorithm with a local refinement, and (3) evaluation of training and detection on images compressed with lossy 3D JPEG 2000.

## 2 Background

Previous discriminative approaches [1, 2] detect objects by testing entire images exhaustively at all locations. Hierarchical modeling has focused on exploiting multiple feature levels of different resolutions [3–5] and on part-based [6] or region-based [7] architectures. In our multi-resolution hierarchy, the position candidate hypotheses are propagated during training and detection. This results in a more robust detector than when the levels are trained independently [8] and can be extended to multiple objects.

JPEG 2000 standard [9] also includes client/server Interactive Protocol (JPIP) for transmitting image regions at desired resolutions using the least bandwidth required. The JPIP protocol is useful for visualizing large Dicom images remotely [9] and has a potential to be used in image analysis applications. The quality of JPEG 2000 images after lossy compression have been previously evaluated for reading radiology datasets [10]. In this paper, we evaluate the robustness of a learning-based algorithm using compressed images in training and detection.

Operating under bounded bandwidth and computational power has been previously addressed in visual surveillance applications [11, 12]. The extracted information (regions [11] and detected objects [12]) has much smaller size than the original images and can be transmitted efficiently over a wide-area network.

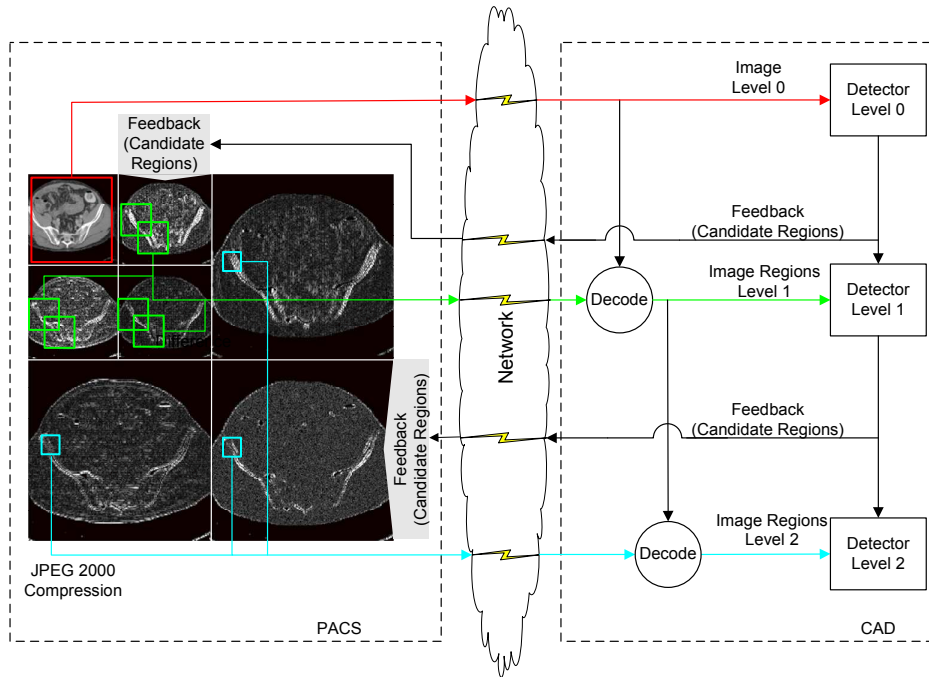
## 3 Algorithm

The core of the DiC system (Figure 2) is a hierarchical learning algorithm (Section 3.1) with one detector trained for each level (Section 3.2). At each level, the detector search region is defined by the image neighborhoods surrounding the highest probability landmark position candidates from the previous level. The image regions are progressively obtained over the network from the PACS server. Since they are encoded with a JPEG 2000 image compression, only high frequency wavelet components need to be transmitted at each subsequent level.

### 3.1 Discriminative Learning

At each hierarchical level, the input to the algorithm is an image volume  $V(r, q, R) : \mathbb{R}^3 \rightarrow [0, 1]$  of resolution  $r$ , quality  $q$  (such as measured by peak signal-to-noise ratio, pSNR), and size  $R$ . The quality  $q$  is lower for images with artifacts caused by image compression. The pSNR value is determined with respect to the uncompressed image, which has the highest quality  $\tilde{q}$ . Each landmark is represented by its position  $\boldsymbol{\theta} = (p_x, p_y, p_z)$ . The goal of the system is to automatically estimate the set of position parameters  $\hat{\boldsymbol{\theta}}$  using a volumetric context surrounding the landmark position:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|V), \quad (1)$$



**Fig. 2.** Overall DiC system diagram. The hierarchical detection algorithm progressively obtains image regions required for detection at each level.

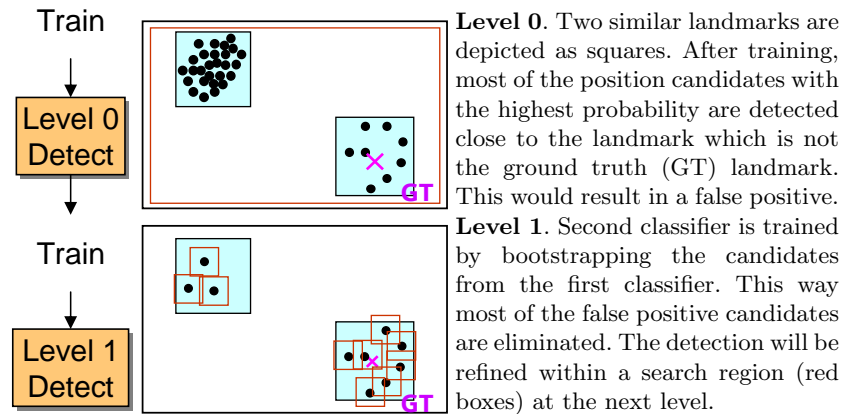
where  $P(\theta|V)$  is the probability of the parameters given the image volume. Let us now define a random variable  $y \in \{-1, +1\}$ , where  $y = +1$  indicates the presence and  $y = -1$  absence of the anatomy. We train a Probabilistic Boosting Tree classifier (PBT) [1] with nodes composed of AdaBoost classifiers trained to select features that best discriminate between positive and negative examples of the landmark. We can then evaluate the probability of a landmark being detected as  $P(y = +1|\theta, V)$ . Therefore we can rewrite Eq. 1

$$\hat{\theta} = \arg \max_{\theta} P(y = +1|\theta, V). \quad (2)$$

The robustness of the discriminative AdaBoost framework makes it possible to use images with compression artifacts in training and detection and still obtain highly accurate detection results. In Section 4 we will show that we can train the classifier adapted to different levels of 3D JPEG 2000 compression of the input images. This is better than training on the uncompressed data since the classifier can learn the consistent anatomical structures and ignore the compression artifacts.

### 3.2 Hierarchical Detection

During detection, the landmark position candidate hypotheses are propagated from the coarser levels to the finer levels as follows. At the coarsest resolution  $r_0$ , a classifier  $D(r_0, q)$  is trained using the volume region  $V(r_0, q, R_0)$  as described in the previous section. The size  $R_0$  of the region is the size of the whole image at resolution  $r_0$ . The detector is then used to obtain position candidate hypotheses at this level. The candidates with the highest probability are *bootstrapped* to train a detector  $D(r_1, q)$  at the next level with resolution  $r_1$ . The volume region  $V(r_1, q, R_1)$  is composed of the union of neighborhood regions of size  $R_1$  surrounding the position candidates. The bootstrapping procedure continues until all levels  $\{r_i\}$  have been processed. The schematic diagram of the hierarchical detection is in Figure 3.



**Fig. 3.** Schematic diagram showing the robustness of the hierarchical processing. The search regions are shown with red rectangles. Only two resolution levels are shown to demonstrate the concept.

The hierarchical processing has several advantages. First, the decreasing context region size helps to avoid local maxima of the probability distribution that would otherwise cause false positives. This results in a more robust and efficient algorithm that operates on datasets of reduced size. Second, the search step depends on the resolution at each level and does not need to be chosen to balance the accuracy of the final result and computational speed.

### 3.3 Progressive Data Transmission

The hierarchical detection algorithm allows for enormous bandwidth savings between the CAD and PACS servers. First, since the images are encoded with a *lossy* 3D JPEG 2000 compression, only high frequency wavelet components are transmitted at the higher resolution levels. Second, when we incorporate the bootstrapping of the candidates across levels, image at the coarsest resolution  $r_0$

is transmitted in its entirety and only image regions surrounding the candidates are used at subsequent levels.

The DiC system must compromise between data bandwidth and the detection accuracy. Denoting  $q_l$  image quality used at the level  $l$  with resolution  $r_l$ , the total size of all image regions for transmission is

$$S_{total} = S(V(r_0, q_0, R_0)) + S(V(r_1, q_1, R_1)) + \dots + S(V(r_n, q_n, R_n)), \quad (3)$$

where  $n$  is the total number of levels used. An algorithm only using the finest resolution  $r_n$  without the bootstrap feedback would require size  $S(V(r_n, q_n))$ . If we fix the quality at all levels,  $q_l = q'$ , the overall size of all image regions is bounded as

$$S(V(r_0, q')) = \frac{S(V(r_n, q'))}{(2^d)^{n-1}} < S_{total} \leq S(V(r_n, q')) \leq S(V(r_n, q_n)), \quad (4)$$

where  $d$  is the image dimension. The final detection score  $P_{final} \in [0, 1]$  is computed from the scores at each level

$$P_{overall} = P(D(r_0, q_0)) \cdot P(D(r_1, q_1)) \dots P(D(r_n, q_n)). \quad (5)$$

Using these definitions, we can now formulate the following optimization problems to determine the quality used at each resolution level. First, given a fixed size budget  $S'$ , the goal is to maximize the detection performance by choosing different quality values  $(q_0, \dots, q_n)$  for different resolutions:

$$(q_0, \dots, q_n) = \underset{S \leq S'}{\operatorname{argmax}} P_{overall}. \quad (6)$$

Second, given a required detection performance  $P'$ , the goal is to minimize the total size of all image regions by choosing different quality values  $(q_0, \dots, q_n)$  for different resolutions:

$$(q_0, \dots, q_n) = \underset{P_{overall} \geq P'}{\operatorname{argmin}} S_{total}. \quad (7)$$

Solutions to these problems are complicated by the fact that the choice of the quality at the level  $l$  not only directly influences the detection score  $D(r_l, q_l)$ , but also the detection score  $D(r_{l+1}, q_{l+1})$  and the selection of  $q_{l+1}$  at the next level,  $l + 1$ . In this paper, we adopt a simpler approach by setting the quality to the same pSNR value at each resolution. We study the effect of this selection on the average detection error and on the total size  $S_{total}$  of all image regions for transmission.

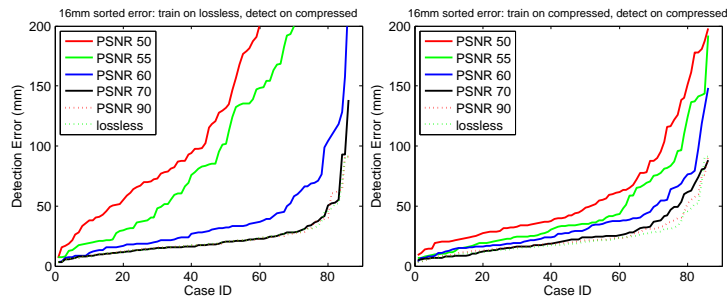
## 4 Experiments

Our experiments start by showing the advantage of training on compressed images, even though their quality (measured by pSNR) is worse than that of the original images. We will then show that hierarchical learning improves the robustness of the algorithm and loosens the bandwidth requirements. Finally, we

will present experiments comparing the full detection pipeline evaluating the hierarchical learning on compressed and uncompressed images.

Our first set of experiments is on 247 CT volumes (86 for training and 161 for testing) with average size  $97 \times 80 \times 165$  voxels after resampling to 4 mm isotropic resolution. The landmark of interest is the right hip bone landmark (Figure 7 left). Second set of experiments is on 511 MRI volumes (384 for training and 127 for testing) with average size  $130 \times 130 \times 101$  voxels after resampling to 2 mm isotropic resolution. In each volume, we detect the crista galli (CG) landmark of the brain (Figure 7 right). The training and testing datasets are disjoint and the volumes in each were chosen randomly.

In the first experiment, we test the detection on images compressed at different pSNR levels; see Figure 6 for examples. The detection error statistics were computed for images of different pSNR levels with classifiers trained on (a) images with the same pSNR level, and (b) uncompressed images. The plots in Figure 4 show, that we can obtain better detection performance when training on compressed images thanks to the classifier’s ability to adapt to the training data and ignore inconsistencies caused by the compression artifacts (see Section 3).

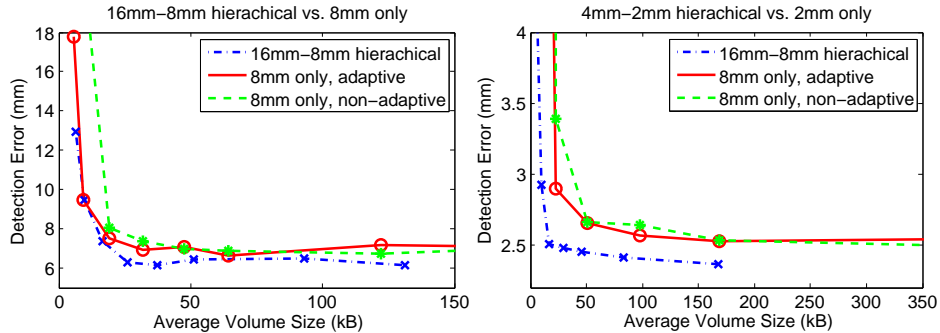


**Fig. 4.** Sorted detection error for different levels of compression in testing when trained on uncompressed (left) and compressed (right) images. By using compressed images in training, the classifier is more robust to the compression artifacts during detection.

Our second experiment demonstrates the robustness of the hierarchical detection. A single level classifier trained on CT images with 8 mm resolution is compared to the hierarchical classifier (Section 3.2 and Figure 3) trained on images with 16 mm and 8 mm resolution. This experiment is repeated for MR images with 2 mm resolution and a 4 mm-2 mm hierarchy. In Figure 5, the median of the 80% smallest errors<sup>1</sup> are plotted against the average volume size computed for each pSNR level. By training on compressed images we can achieve smaller detection errors for a given average volume size than by training on uncompressed images. The detection errors decrease further when using the hierarchical approach due to the robust search strategy.

The final experiment compares the overall hierarchical detection using uncompressed images and images compressed at pSNR 70. The hierarchical system

<sup>1</sup> The large errors can be easily rejected as outliers based on the detection score.



**Fig. 5.** Detection error vs. average volume size for hip bone landmark in CT (left) and crista galli landmark in brain MRI (right). The images were compressed in training and testing with the same pSNR level (adaptive) and uncompressed in training and compressed in testing (nonadaptive). The hierarchical processing results in lowest detection error through the focused coarse-to-fine search and training on compressed volumes. The average size of the uncompressed and lossless-compressed volumes is: 404 kB and 189 kB (8 mm CT), 3334 kB and 985 kB (2 mm MRI).

is also compared to a simpler algorithm operating on uncompressed images with a single resolution. The results summarized in Table 1 show that an algorithm trained on images compressed with lossy compression achieves data size reduction by a factor of 3.7 (9.9 for MRI) when compared to a hierarchical training on lossless-compressed images, by a factor of 12.7 (58.0 for MRI) when compared to an algorithm operating on a single resolution, and by a factor of 30.0 (196.2 for MRI) when the original (uncompressed) images are used. The median detection error is comparable for all three cases.

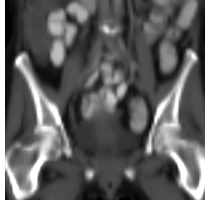
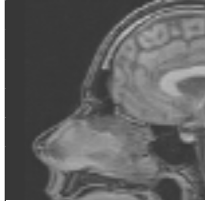


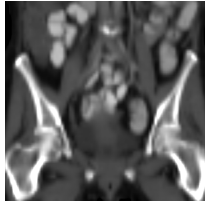
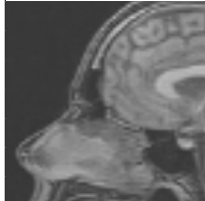
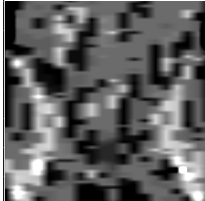


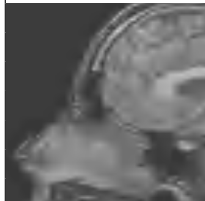
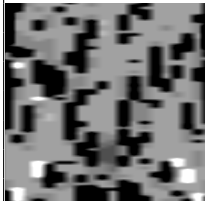

	CT			MRI		
	16-8-4 hier pSNR 70	16-8-4 hier lossless	4 mm lossless	4-2 hier pSNR 70	4-2 hier lossless	2 mm lossless
Error [mm]	3.87	3.54	3.98	2.50	2.37	2.27
Avg. Data Size [kB]	106.22	393.45	1345.96	16.99	168.07	984.76

**Table 1.** The median detection error of the hierarchical detection on images compressed at pSNR 70 (2nd and 5th column), on uncompressed images (3rd and 6th column), and on a single resolution losslessly-compressed images (4th and 7th column). The average size of uncompressed volumes is 3188 kB (4 mm CT) and 3334 kB (2 mm MRI). The hierarchical algorithm trained with images of pSNR 70 requires the least amount of data without sacrificing the detection accuracy.

## 5 Conclusion

We presented *Detection in a Cloud (DiC)* system for anatomical landmark detection in the cloud computing environment. At the core of the system is a hierarchical learning algorithm that propagates position candidate hypotheses across a hierarchy of classifiers during training and detection. The algorithm only requires image regions surrounding the candidates which results in less bandwidth for remote data access. Further bandwidth savings (without sacrificing the detection accuracy) are achieved by compressing the images regions with



		CT	MRI			CT	MRI
		4, 8, 16 mm	2, 4 mm			4, 8, 16 mm	2, 4 mm
uncompressed (lossless)		3187.5, 403.3, 50.1 (1346.0, 188.3, 26.6)	3333.8, 420.9 (984.8, 121.6)	pSNR 60		92.0, 19.1, 4.0	22.9, 4.6
							
		795.2, 122.2, 18.6	415.3, 53.0			49.6, 9.4, 2.2	16.6, 3.5
pSNR 90				pSNR 55			
		240.9, 47.4, 8.5	51.0, 10.1			35.7, 5.8, 1.2	11.6, 2.7
pSNR 70				pSNR 50			

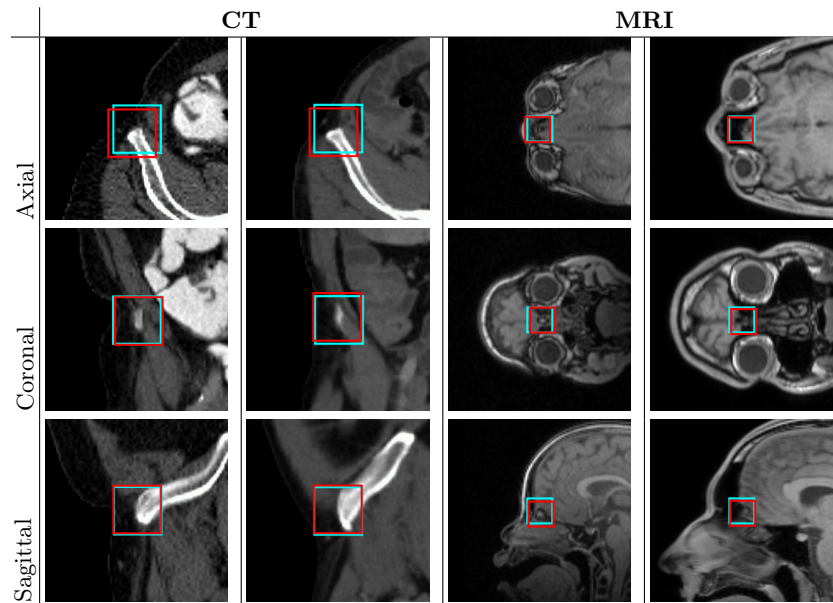
**Fig. 6.** Average compressed volume sizes (in kB) at different compression levels and resolutions. The first row also shows sizes after lossless compression (in parentheses). These statistics are computed over all testing volumes.

lossy JPEG 2000. The total bandwidth savings for retrieving remotely stored data amount to 30.0 times (CT data) and 196.2 times (MRI data) reduction when compared to the original data size and 12.7 times (CT) and 58.0 (MRI) when compared to data size after lossless compression.

The proposed approach makes it possible to shift the integration, maintenance, and software updates from the client to the CAD server. Therefore, when the classifiers are updated, they are immediately available to all clients. In the clinical environment, detected anatomical parts can be reviewed on the client devices remotely. The current system opens many exciting future research directions both on the algorithmic side as well as on the systems side. We are interested the most in building more complicated models with several landmarks of interest trained for different modalities. Such large scale systems will require coordination of multiple CAD servers possibly distributed in a wide-area network.

## References

1. Tu, Z.: Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In: CVPR. Volume 2. (2005) 1589–1596



**Fig. 7.** Example images with ground truth locations (red boxes) and detection results (cyan boxes). The CT and MRI images correspond to the right hip bone and crista galli landmarks respectively.

2. Viola, P., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: CVPR. Volume 1. (2001) 511–518
3. Schnitzspan, P., Fritz, M., Roth, S., Schiele, B.: Discriminative structure learning of hierarchical representations for object detection. In: CVPR. (2009) 2238–2245
4. Zhang, W., Zelinsky, G., Samaras, D.: Real-time accurate object detection using multiple resolutions. In: ICCV. (2007)
5. Zhu, L., Yuille, A.L.: A hierarchical compositional system for rapid object detection. In: NIPS. (2005) 1633–1640
6. Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S.: Describing visual scenes using transformed objects and parts. IJCV **77** (2008) 291–330
7. Vilaplana, V., Marques, F., Salembier, P.: Binary partition trees for object detection. TIP **17** (2008) 2201–2216
8. Butko, N.J., Movellan, J.R.: Optimal scanning for faster object detection. In: CVPR. (2009) 2751–2758
9. Schelkens, P., Skodras, A., Ebrahimi, T.: A comprehensive guide to the latest features in JPEG 2000. John Wiley and Sons (2009)
10. Ringl, H., Scherthaner, R., Sala, E., El-Rabadi, K., Weber, M., Schima, W., Herold, C.J.: Lossy 3D JPEG2000 compression of abdominal CT images in patients with acute abdominal complaints: Effect of compression ratio on diagnostic confidence and accuracy. Radiology **248** (2008) 476–484
11. Dufaux, F., Ebrahimi, T.: Scrambling for video surveillance with privacy. In: CVPR Workshop. (2006)
12. Fleck, S., Busch, F., Biber, P., Strasser, W.: 3D surveillance a distributed network of smart cameras for real-time. In: CVPR Workshop. (2006)