Published in final edited form as:

Mach Learn Med Imaging. 2015 October ; 9352: 321–329. doi:10.1007/978-3-319-24888-2_39.

Multi-source Information Gain for Random Forest: An Application to CT Image Prediction from MRI Data

Tri Huynh[1], **Yaozong Gao**[1,2], **Jiayin Kang**[1], **Li Wang**[1], **Pei Zhang**[1], **Dinggang Shen**[1,2], and **Alzheimer's Disease Neuroimaging Initiative (ADNI)**

[1]IDEA Lab, Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

Random forest has been widely recognized as one of the most powerful learning-based predictors in literature, with a broad range of applications in medical imaging. Notable efforts have been focused on enhancing the algorithm in multiple facets. In this paper, we present an original concept of *multi-source information gain* that escapes from the conventional notion inherent to random forest. We propose the idea of characterizing information gain in the training process by utilizing *multiple beneficial sources of information*, instead of the *sole governing of prediction targets* as conventionally known. We suggest the use of location and input image patches as the secondary sources of information for guiding the splitting process in random forest, and experiment on the challenging task of predicting CT images from MRI data. The experimentation is thoroughly analyzed in two datasets, i.e., human brain and prostate, with its performance further validated with the integration of auto-context model. Results prove that the *multi-source information gain* concept effectively helps better guide the training process with consistent improvement in prediction accuracy.

1 Introduction

Since introduced by Breiman [1] in 2001, random forest has become one of the most powerful learning-based predictors with state-of-the-art performance on a broad range of applications, ranging from detection, classification, to segmentation, etc.

With its increasing success, random forest has attracted growing efforts in improving the method from various facets. Menze et al. [2] proposed a supervised approach to define the optimal "*oblique*" split direction on the features, instead of the popular *orthogonal* split in the training process. This approach adapts more effectively to the nature of data and dramatically reduces the complexity of decision trees. Marin [3] et al. also adopted this idea with the use of Support Vector Machine in learning the splitting direction. While in [4], Robnik-Šikonja provided insight that using multiple attribute evaluation measures in

Correspondence to: Dinggang Shen.

different trees for split selection would improve the performance by decreasing the correlation among the trees. The author also showed significant improvement in deriving final prediction result by weighting the trees based on their performance on similar inputs. Recently, the most notable advancement of random forest is the introduction of structured random forest, which extends random forest from predicting scalar outputs to directly predicting structured outputs. Structured random forest can better preserve the neighborhood information in the structured outputs as a whole, which has shown preeminent performance [5] [6].

Although many enhancements have been proposed for random forest, all of the methods follow the same strategy in growing trees based solely on minimizing the variety of the prediction targets in each child node. The purity of prediction targets or labels is unarguably the most important factor in choosing the splits on data. However, prediction targets are not the only source of information that is beneficial to guide the process. In this paper, we, for the first time, explore the use of *multiple sources of information* as the splitting criteria in random forest. Specifically, we devise the general model for multi-source information gain, and suggest the use of location and input image patches (built upon the success of structured random forest) as other secondary sources of information to guide the splitting process. The method is then analyzed through the challenging problem of predicting computed tomography (CT) image from magnetic resonance (MR) image in two datasets, human brain and prostate region. The performance is also further thoroughly examined and validated with the integration of auto-context model. Results provide insights into the method as well as show that significant improvement could be gained by the proposed approach.

2 Random Forest

We first review the *classic* random forest, followed by *structured* random forest. They are the foundation for extending to *multi-source information gain* in Section 3.

2.1 Classic Random Forest

Random forest comprises of multiple decision trees. At each internal node of a tree, a feature is chosen to split the incoming training samples to maximize the information gain. A training sample consists of an *input* feature vector and its *output* target. Let $\boldsymbol{u} \in \boldsymbol{U} \subset \mathbb{R}^q$ be an input feature vector, and $v \in V \subset \mathbb{Z}$ be its corresponding prediction target in the classification problem. For a set of samples $S_j \subset \boldsymbol{U} \times V$ arriving at node j, the information gain achieved by choosing the k-th feature is computed by:

$$I_j^k = H\left(S_j\right) - \frac{\left|S_{j,\mathrm{L}}^k\right|}{\left|S_j\right|} H\left(S_{j,\mathrm{L}}^k\right) - \frac{\left|S_{j,\mathrm{R}}^k\right|}{\left|S_j\right|} H\left(S_{j,\mathrm{R}}^k\right), \quad (1)$$

$$H(S) = -\sum_{v} p_v \log\left(p_v\right), \quad (2)$$

where L and R denote the left and right child nodes, $S_{j,\mathrm{L}}^{k} = \left\{(\boldsymbol{u}, \nu) \in S_j \middle| \boldsymbol{u}^k < \theta_j^k\right\}$, $S_{j,\mathrm{R}}^{k} = S_j \backslash S_{j,\mathrm{L}}^{k}$, $\boldsymbol{u}^k$ is the k-th feature in $\boldsymbol{u}$, θ_j^k is the splitting threshold chosen to maximize the information gain I_j^k, and $|\cdot|$ is the cardinality of the set. $H(S)$ denotes the entropy of target values in S, with p_ν the fraction of elements in S having value ν. For regression problem ($V \subset \mathbb{R}$), the entropy is replaced by variance as follows:

$$H(S) = \sum_{\nu} p_\nu (\nu - \bar{\nu})^2, \quad (3)$$

$$\bar{\nu} = \sum_{\nu} p_\nu \nu, \quad (4)$$

2.2 Structured Random Forest

In the classic random forest, the output space is *either* a class label for the case of classification, *or* a real value for the case of regression. Recently, a few pioneering works [5] [6] advanced random forest into structured random forest, which directly predicts a structured patch instead of a single value, and achieved preeminent performance. The difference between *structured* random forest and *classic* random forest is illustrated in Fig. 1, using an example of predicting CT image from MRI data. *Structured* random forest helps preserve the neighborhood information in the predicted structured patch and further reduce the expected number of decision trees since a voxel now receives information from multiple neighboring patches.

When extending to *structured* random forest, the main issue is how to characterize the entropy of the structured patches. That is, how to efficiently capture the similarity of different target patches. One naïve way is formulating the similarity based on individual voxels inside the patches. However, the computation is highly expensive and the method is too sensitive to individual voxel changes rather than high-level patch structure. A more effective way is to find a mapping that can effectively capture the information from each image patch. In this paper, we characterize the image patches by principal component analysis (PCA), i.e., the mapped coefficients of $\boldsymbol{\nu}$ represent its first d PCA coefficients. This mapping has the advantage of being computationally efficient while effectively deriving the most significant information in each target patch. Suppose the prediction targets now are $\boldsymbol{\nu} \in \boldsymbol{V} \subset \mathbb{R}^g$, and $\boldsymbol{w} = \Pi(\boldsymbol{\nu})$ denotes the mapped coefficients of $\boldsymbol{\nu}$, where $\boldsymbol{w} \in \mathbb{C} \subset \mathbb{R}^d$, $d < g$. Then, the entropy $H(S)$ from Eqs. (3) and (4) can be computed as:

$$H(S) = \sum_{\boldsymbol{w}} p_{\boldsymbol{w}} \left\|\boldsymbol{w} - \bar{\boldsymbol{w}}\right\|_{\mathbf{2}}^{\mathbf{2}} \quad (5)$$

$$\bar{\boldsymbol{w}} = \sum_{\boldsymbol{w}} p_{\boldsymbol{w}} \boldsymbol{w} \quad (6)$$

By using *structured* random forest, the predicted neighboring patches can be fused, i.e., by averaging, to reconstruct a final predicted image.

3 Multi-source Information Gain

In random forest, the splitting maximizes the information gain by minimizing the variety of prediction targets in each child node (Eqs. 1–6). The rationale of this process is to find the best feature and threshold that have the highest discriminative power to categorize the samples. That discriminative power is measured solely by the convergence of prediction targets, which is intuitive and the prediction targets are unarguably the most important information to guide the learning procedure. However, they are not the only source of information that is helpful to guide the splitting process. For example, many applications hold the spatial constraint, in which certain output patterns usually appear in certain locations. In these cases, a split results in the grouping of similar output patterns and is more informative and robust in nearby locations, which should also carry higher information gain compared to the case of having those output patterns in highly scattered locations. In this case, location is another source of information that could contribute to the information gain obtained at each split. Therefore, we propose to include multiple indicators in guiding the splitting process, and devise the notion of multi-source information gain for this purpose.

Suppose we have N sources of information that we would like to integrate to the final information gain. The sources can be of various forms, from discrete labels, real values, or structured patches, with their individual entropies to be computed as in Eqs. 2, 3–4, and 5–6, respectively. Since the information gain from different sources could have very different ranges, we define the information gain ratio for one source as a variant of the information gain in Eq. 1 to normalize the gain from different sources:

$$R_j^k = \frac{I_j^k}{H\left(S_j\right)} \quad (7)$$

where R_j^k denotes the information gain ratio obtained from one source by choosing the *k-th* feature to split the samples at node j, I_j^k is the information gain as defined in Eq. 1, and $H(S_j)$ is the entropy of set S_j. Letting $R_j^k(n)$ denote the information gain ratio obtained from source $n \in \{1, \ldots, N\}$, we define the multi-source information gain as the weighted combination of the information gain ratios from all sources as follows:

$$M_j^k = \sum_{n=1}^{N} \alpha(n) G_j^k(n) \quad (8)$$

$$G_j^k(n) = \begin{cases} R_j^k(n) & R_j^k(n) \geq 0 \\ 0 & R_j^k(n) < 0 \end{cases} \quad (9)$$

where M_j^k denotes the multi-source information gain when choosing the k-th feature to split the samples at node j, and $\alpha(n)$ is the weighting factor which indicates the relative importance of different sources.

In this paper, we suggest and analyze the use of three different sources of information gain: 1) the prediction target patches, 2) the location of the samples, and 3) the corresponding input image patches. The contribution of location information has been previously discussed, while input image patches also can potentially provide helpful information to better guide the splitting process. In many problems where there are strong correlations between the input and output structures, we could expect that a similarity in input structures should also correspond to the similarity in output structures. Thus, information gain from input structures could potentially further enhance the confidence of similarity of output structures. This is especially helpful with the introduction of *structured* random forest, where the information gain from both input and output can be measured in the corresponding structured patches, thus better exploiting their correlation.

4 Experimental Analysis

4.1 Predicting CT Image from MR image

We apply the proposed method to the problem of predicting CT image from corresponding MR image. We choose to perform analysis on this problem because it is a challenging problem, with complex relationship between CT and MR images, and also location information could be exploited. These conditions allow us to best demonstrate the proposed multi-source information gain model. This task is highly important in performing attenuation correction (AC) for Positron emission tomography (PET) images in the PET/MRI system. AC is required to make PET images readily applicable for clinical diagnosis, which relies on the attenuation map obtained from CT images. Therefore, predicting CT image from MR image is crucial in PET/MRI system. Examples of the MR and CT image pairs are shown in Figs. 2–3. As can be seen in the figures, predicting CT image from MR image is very challenging, with the complex relationship between two modalities. The same range of MR intensity values can correspond to different ranges of CT values (Fig. 2), while multiple ranges of MR values can also correspond to the same CT value range (Fig. 3).

4.1.1 Datasets—We experiment on two datasets: 1) The brain data were acquired from 16 subjects with both MR and CT scans in the Alzheimer's Disease Neuroimaging Initiative data-base (adni.loni.usc.edu). 2) The prostate dataset is our in-house data, which has 22 subjects, each with the corresponding MR and CT scans.

4.1.2 Training Procedure

Pre-alignment: In order to learn the relationship between MR and CT images, we first need to perform the intra-alignment for the MR and CT image pair of each subject [8]. Afterwards, in order to utilize the spatial information, we perform inter-subject registration [7] to roughly bring all the subjects onto a common space.

Training: We utilize structured random forest to predict CT image from corresponding MR image, as discussed in Section 2.2. Different combinations from three sources of information gain are experimented: MR patches, target CT patches, and location. The following parameters were used - MR input patch size: 15×15×15; CT target patch size: 3×3×3; Number of PCA coefficients used in structured random forest: 10; Weighing factors for CT patches, MR patches, and locations in the multi-source information gain model are 1, 0.2, and 0.2, respectively.

4.1.3 Results—To provide thorough evaluation of the performance using different sources of information gain, we experiment on **four** different configurations: **1**) information gain from target CT patches alone (CT), **2**) from CT and MR patches (CT_MR), **3**) from CT patches and Locations of the patches (CT_LOC), and **4**) from MR, CT patches, and Locations of the patches (CT_MR_LOC). Leave-one-out cross validation was performed on both datasets using two popular metrics: Peak signal-to-noise ratio (PSNR) and normalized mean square error (NMSE). Results are provided in Figs. 6–7, with qualitative samples in Figs. 4–5. Following conclusions could be drawn:

- The information gain from **location** always notably improves the performance in both brain and prostate data (CT_LOC versus CT, and CT_MR_LOC versus CT_MR). The location information gain helps favoring the grouping of similar image patches in nearby locations, making the grouping more robust.
- The information gain from **MR patches** slightly improves the prediction performance in brain dataset, but degrades the prediction in prostate dataset (CT_MR versus CT, and CT_MR_LOC versus CT_LOC). One possible reason is due to the nature of datasets. In brain data, we have more one-to-multiple mappings from MR to CT images, where similar intensities from MR (e.g., air and bone) correspond to highly different CT intensities. Thus, the added refinement from MR patches helps better differentiate the CT patches. On the other hand, in prostate data, there are more multiple-to-one mappings from MR to CT images. Thus, the further information gain from MR patches does not help, and actually makes the grouping overfitting and leads to more wrong predictions.

To validate the confidence in improvement of the multi-source model, we also performed statistical tests with the obtained p-values well below 0.05 for both datasets.

4.2 Integration to Auto-context Model

To perform in-depth analysis of the method, we further experiment the multi-source information gain in the second layer of auto-context model (ACM) [9]. ACM utilizes the prediction result from the previously learned model as the added contextual information, and uses features extracted from this result together with features from the original input image to train a new refining model.

We use the *best predicted results* from the first layer (CT_MR_LOC for brain, CT_LOC for prostate data) as the context features to train the second layer random forest. The prediction performance of different sources of information gain is provided in Figs. 8–9. We can see that although the improvement has been more saturated in the second layer, adding more sources of information gain still has the same effect as in the first layer. Specifically, location information gain always improves the performance in both datasets, while information gain from MR image patches helps advance the prediction in brain dataset, but degrades in prostate dataset.

To further provide a complete comparison, we also show the performance of autocontext model using the *traditional* random forest (information gain based solely on CT patches) and our *multi-source information gain* based random forest (CT_MR_LOC for brain, and CT_LOC for prostate data). In this experiment, each method uses *its own predicted results* as context features. Results are presented in Figs. 10–11. From this experiment, we can clearly see the improvement of the proposed method compared to the traditional one. In both datasets, the performance of the multi-source model in the first layer almost reaches the result of the traditional model in the second layer.

5 Discussion and Conclusion

In this paper, we proposed the use of multiple sources of information in characterizing the information gain in random forest. A general model was proposed and the experimentation was carried out for the challenging task of predicting CT images from MRI data. Results clearly show that, when using appropriately, the information gain from other contributive sources besides the prediction targets consistently improves the prediction performance. This is the first time a multi-source information gain concept is proposed with promising results, which could open potentials for future shifts into this line of research. In the future, we would like to investigate the use of other sources of information gain that could also be taken into consideration.

References

1. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32.
2. Menze BH, Kelm B, Splitthoff DN, Koethe U, Hamprecht FA. On oblique random forests. In: Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M, editorsECML PKDD 2011, Part II LNCS. Vol. 6912. Springer; Heidelberg: 2011. 453–469.
3. Marin J, Vazquez D, Lopez AM, Amores J, Leibe B. Random forests of local experts for pedestrian detection. IEEE International Conference on Computer Vision (ICCV). 2013:2592–2599.
4. Robnik-Sikonja M. Improving random forests. In: Boulicaut J-F, Esposito F, Giannotti F, Pedreschi D, editorsECML 2004 LNCS (LNAI). Vol. 3201. Springer; Heidelberg: 2004. 359–370.

5. Kontschieder P, Bulò SR, Bischof H, Pelillo M. Structured class-labels in random forests for semantic image labeling. ICCV. 2011:2190–2197.

6. Dollar P, Zitnick CL. Structured forests for fast edge detection. ICCV. 2013:1841–1848.

7. Jenkinson M, Smith SM. A global optimisation method for robust affine registration of brain images. Medical Image Analysis. 2011; 5(2):143–156.

8. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: a toolbox for intensity based medical image registration. IEEE Transactions on Medical Imaging. 2010; 29(1):196–205. [PubMed: 19923044]

9. Tu Z. Auto-context and its application to high-level vision tasks. IEEE Conference on Computer Vision and Pattern Recognition. 2008:1–8.

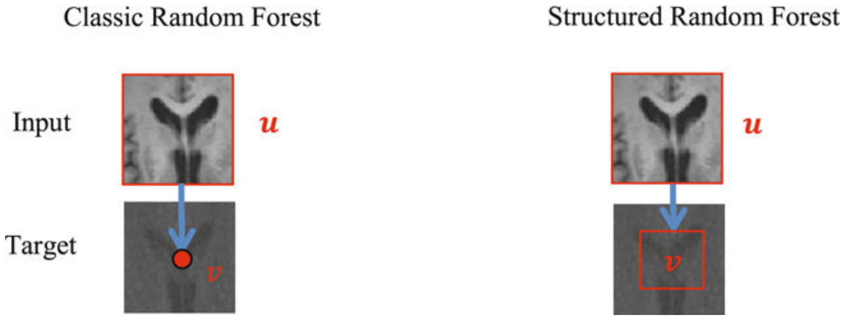

Fig. 1.

Illustration of *classic* random forest and *structured* random forest. In the classic random forest, the input feature vector $\boldsymbol{u}$ derived from MR image patch is used to predict a target value $\boldsymbol{v}$ for a voxel (red point) in the CT image, while, in the structured random forest, the same feature vector $\boldsymbol{u}$ is used to predict all values $\boldsymbol{v}$ in a target CT patch (red rectangle).

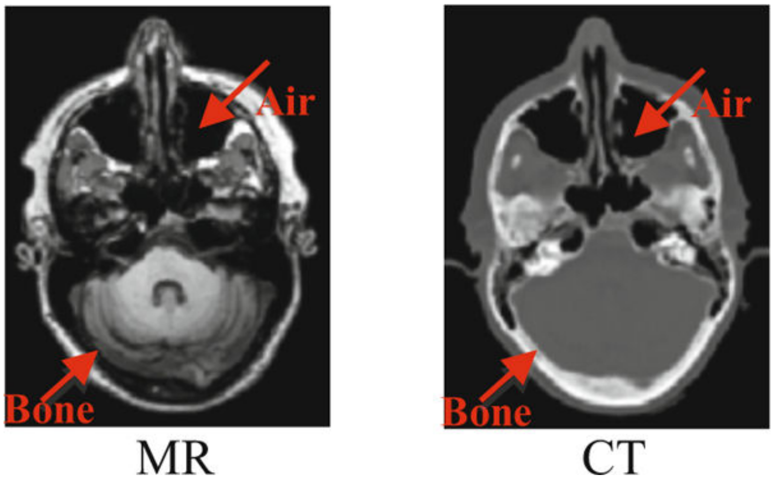

Fig. 2.

A pair of MR image and corresponding CT image from the same human brain. Example of “one-to-multiple” relationship: both air and bone have very low response in MR images, but can be highly differentiated in CT images.

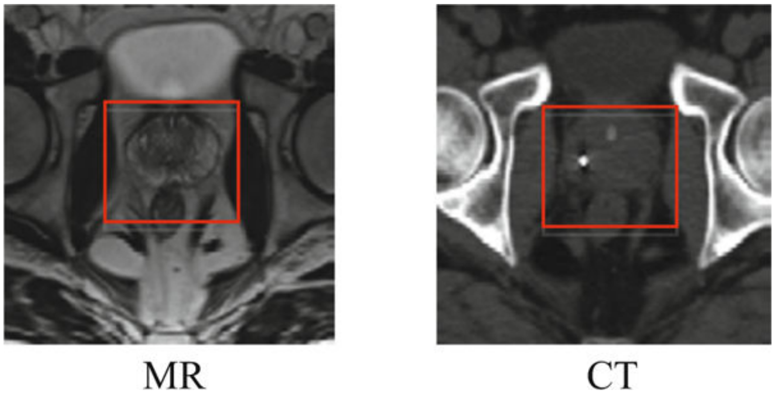

Fig. 3.

A pair of an MR image and its corresponding CT image around the prostate area. Example of “multiple-to-one” relationship: there are many intensity levels in MR image corresponding to the same intensity level in CT image (red rectangle).

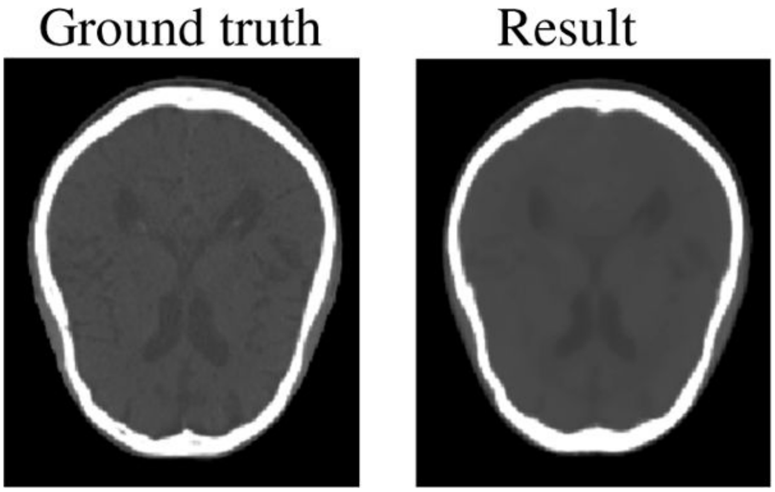

Fig. 4.
Sample result on brain data, using CT_MR_LOC configuration.

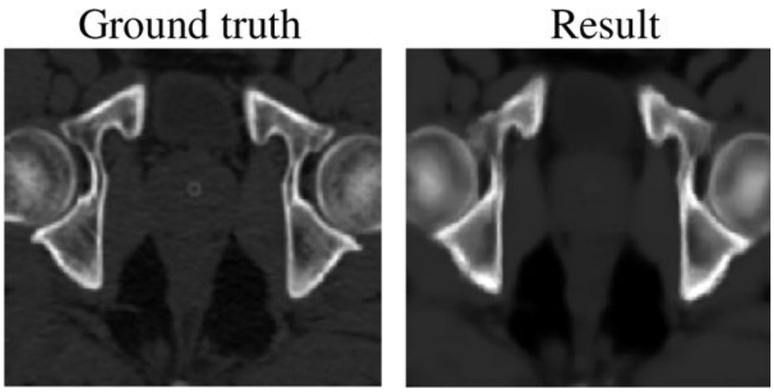

Fig. 5.

Sample result on prostate data, using CT _LOC configuration.

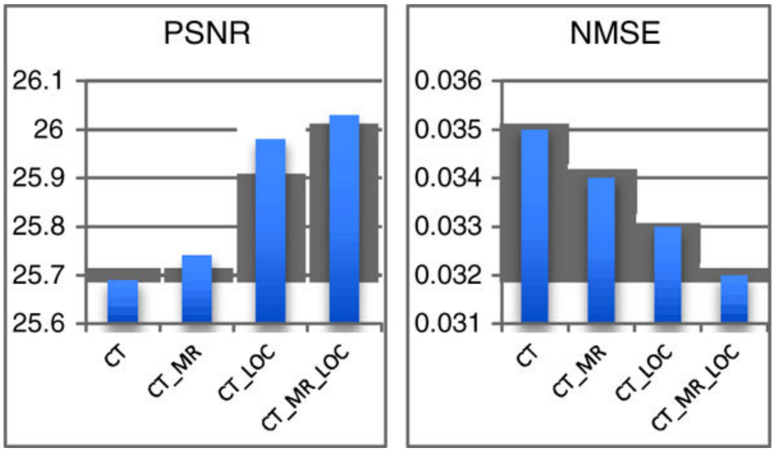

Fig. 6.

Prediction results on brain data.

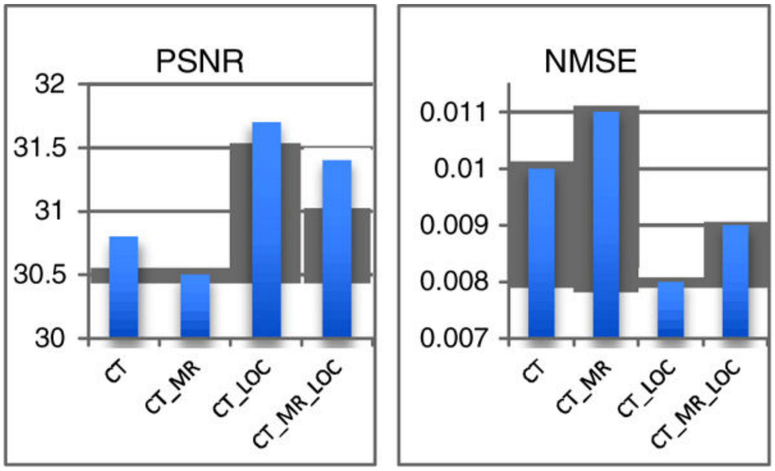

Fig. 7.

Prediction results on prostate data.

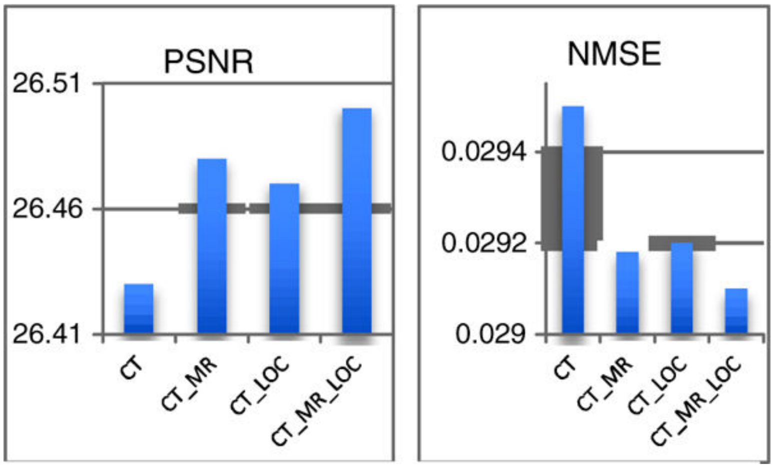

Fig. 8.

Prediction results of second layer ACM on brain data.

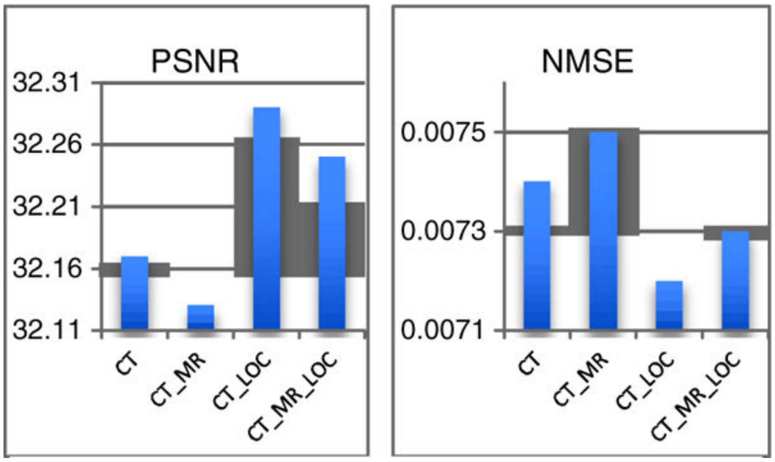

Fig. 9.

Prediction results of second layer ACM on prostate data.

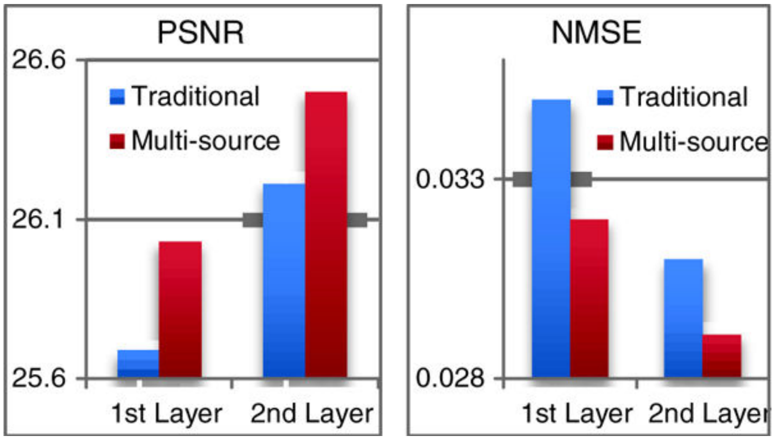

Fig. 10.

Prediction results on brain data, in different layers of ACM.

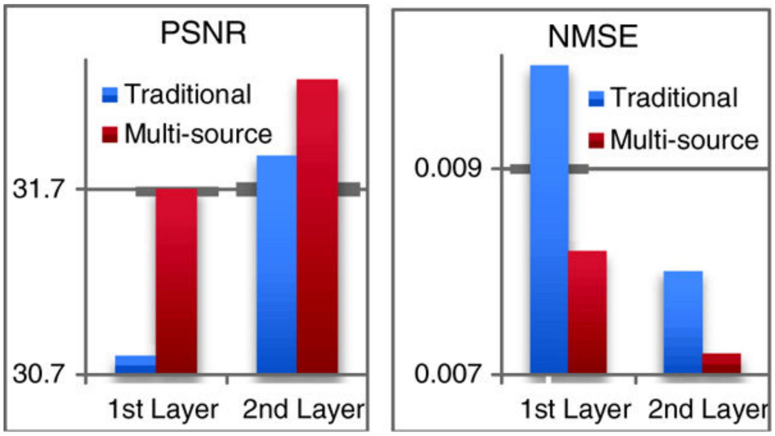

Fig. 11.

Prediction results on prostate data, in different layers of ACM.