

Evaluation de la qualité d'images IRM en 2D et 3D.

Lu Zhang¹, Christine Cavaro-Ménard¹, Patrick Le Callet²

¹LISA, Université d'Angers, 62 avenue Notre Dame du Lac, 49000 Angers - France

²IRCCyN, University of Nantes, Rue Christian Pauc, BP50609, 44306 Nantes cedex 3 - France

lu.ge@etud.univ-angers.fr

christine.menard@univ-angers.fr

Patrick.Le-Callet@univ-nantes.fr

Resumé

Nous présentons dans ce papier les protocoles d'expérimentation que nous avons mis en place pour évaluer la qualité des images médicales (2D et 3D). Ces expérimentations, proches des analyses réalisées en routine clinique, permettent de quantifier la performance des radiologues et sont nécessaires pour valider des observateurs numériques. Les méthodes statistiques d'analyse des données sont également présentées pour les différentes tâches diagnostiques de détection et de détection-localisation jointe.

Mots-clés: *Evaluation de la qualité, images médicales, IRM, ROC.*

1 Introduction

L'imagerie médicale est devenue incontournable dans la prise en charge d'un patient que ce soit au cours du dépistage ou dans la phase diagnostique et dans le suivi thérapeutique. L'évaluation de la qualité des images médicales est indispensable pour la comparaison, voire l'optimisation des systèmes d'imagerie, tels que les systèmes d'acquisition, les systèmes de visualisation et les algorithmes de post-traitement des images (compression, watermarking...).

L'objectif majeur des images médicales est d'aider les radiologues et cliniciens à rendre un diagnostic. Il est d'ailleurs largement admis que la qualité des images doit être optimale afin d'optimiser la décision diagnostique. En outre, la qualité d'une image médicale doit être évaluée dans le contexte d'une tâche diagnostique spécifique. Une telle approche est connue sous le nom d'évaluation basée sur la tâche (en anglais, *task-based assessment*) [1].

Dans ce cas, le résultat d'une ou plusieurs tâches diagnostiques peut être évalué soit par un observateur humain (radiologue), soit par un observateur numérique (modèle mathématique). L'observateur humain est l'utilisateur final des images médicales, il demeure alors l'agent principal des décisions diagnostiques.

Afin d'évaluer la qualité des images médicales (2D et 3D) dans le cadre du *task-based assessment*, il est nécessaire de bien caractériser le processus diagnostique. Une approche communément adoptée est de diviser le processus d'interprétation des images en trois tâches [2]: la détection, la localisation et la caractérisation.

- La tâche de détection appelle simplement à une note de confiance concernant la présence d'un signal (par exemple, une lésion).
- La tâche de localisation consiste à indiquer l'emplacement des lésions.

- La tâche de caractérisation est quant à elle plus complexe, car elle nécessite une analyse des différents éléments de la lésion (contour, texture, etc.) pour un diagnostic différentiel et implique normalement une réponse linguistique (anatomique, tissulaire, etc.).

Dans ce papier, nous présenterons tout d’abord les protocoles d’expérimentation que nous avons mis en place pour quantifier la performance des radiologues lors de la tâche de détection-localisation jointe. Nous illustrerons ensuite les méthodes statistiques de l’analyse des données.

2 Protocoles d’expérimentations

2.1 Pathologie et modalité d’étude

Nous avons choisi d’étudier la sclérose en plaques (SEP), car les lésions SEP sont subtiles et difficiles à détecter. De plus, la localisation des lésions SEP est un indicateur des troubles générés par la pathologie. La SEP est donc une pathologie appropriée pour analyser les tâches de détection et de localisation.

Le diagnostic de la SEP est largement basé sur l’analyse d’IRMs cérébrales. Dans certaines circonstances le diagnostic de la SEP peut être confirmé avec le seul examen d’IRM. Nous avons donc choisi l’IRM comme modalité d’étude.

2.2 Images utilisées

Les images IRM utilisées dans les expériences sont extraites de séquences FLAIR T2 de volontaires sains. Des lésions simulées sont ensuite ajoutées à ces images afin d’obtenir le “gold-standard” (vérité terrain). La pertinence des caractéristiques des lésions simulées a été vérifiée par un radiologue expert en neurologie.

Pour l’analyse des images 2D, la lésion de SEP est modélisée par une fonction gaussienne elliptique bidimensionnelle. Pour l’analyse des images 3D, la lésion de SEP est modélisée par une fonction gaussienne 3D :

$$[x]_p = \alpha \exp \left(-\frac{1}{2} (\mathbf{p} - \mathbf{q})^t \mathbf{A}^t \mathbf{D}^{-1} \mathbf{A} (\mathbf{p} - \mathbf{q}) \right) \quad (1)$$

où $[x]_p$ désigne la valeur de l’intensité du signal à l’emplacement défini par les coordonnées 3D \mathbf{p} . L’atténuation de l’intensité du signal est modélisée par une fonction gaussienne d’amplitude α , centrée sur \mathbf{q} . La matrice diagonale \mathbf{D} (Eq. (2)) spécifie trois paramètres, les rayons σ_1 , σ_2 et σ_3 selon respectivement les axes x , y et z , déterminant la forme de l’ellipsoïde. La matrice de rotation \mathbf{A} correspond à une rotation dans le sens contraire des aiguilles d’une montre dont les angles d’Euler sont ϕ , θ et ψ en utilisant la convention x - y - z , illustrée dans Eq. (3).

$$\mathbf{D} = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \quad (2)$$

$$\mathbf{A} = \begin{bmatrix} \cos \theta \cos \psi & -\cos \phi \sin \psi + \sin \phi \sin \theta \cos \psi & \sin \phi \sin \psi + \cos \phi \sin \theta \cos \psi \\ \cos \theta \sin \psi & \cos \phi \cos \psi + \sin \phi \sin \theta \sin \psi & -\sin \phi \cos \psi + \cos \phi \sin \theta \sin \psi \\ -\sin \theta & \sin \phi \cos \theta & \cos \phi \cos \theta \end{bmatrix} \quad (3)$$

Un exemple de lésion 3D dont $\phi = \theta = 0$, $\psi = \pi/4$ est représenté sous forme volumétrique, Figure 1.

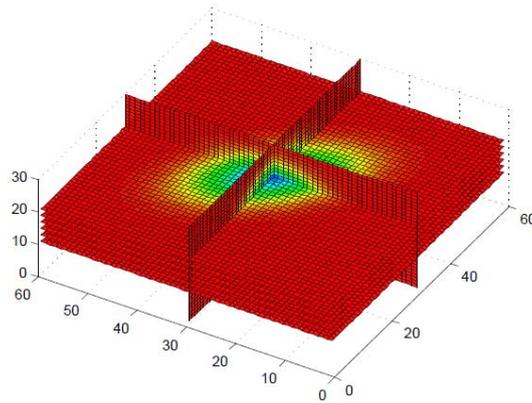


Figure 1: Un exemple de lésion 3D représenté sous forme volumétrique

2.3 But des expérimentations

Le but des expérimentations est d'estimer la performance des radiologues pour la détection et la localisation des lésions de SEP sur une coupe ou une séquence (volumétrique) d'images IRM cérébrales.

2.4 Interfaces graphiques

La Figure 2 présente l'interface graphique de l'expérimentation sur coupe 2D. La figure 3 et la figure 4 présentent l'interface graphique de l'expérimentation sur séquences 3D.

Lors de l'expérimentation sur coupe 2D, les images à interpréter sont soit normales (sans lésion), soit anormales avec lésion(s). Le nombre de lésions sur une image est aléatoire. Les participants ne connaissent ni le nombre de lésions, ni l'emplacement, l'intensité, l'orientation et l'échelle de celles-ci. Il est demandé aux participants de détecter et localiser toutes les lésions suspectes. En déplaçant la position de la souris, les participants peuvent marquer le centre d'une lésion suspecte par une petite croix (cf. Figure 2). Il est également demandé aux participants d'estimer la probabilité qu'à la lésion suspecte d'évoquer une lésion SEP (note de confiance).

Lors de l'expérimentation sur séquences 3D, les participants peuvent faire défiler les coupes en utilisant les touches du clavier (Flèche vers le haut et Flèche vers le bas) afin d'examiner toute l'information sur la séquence 3D. Après la détection d'une lésion suspecte, il est demandé aux participants de sélectionner la coupe centrale de la lésion 3D et de marquer le centre de la lésion sur cette coupe (cf. Figure 3). Une boîte de dialogue s'affiche ensuite pour la saisie de la note de confiance (cf. Figure 4).

Pour les deux expérimentations, le centre de la zone suspecte doit être marqué avec une bonne précision de par son influence sur l'analyse de l'étude. Les marques de localisation et la probabilité que la zone soit une lésion peuvent être désactivées pour ne pas gêner l'observateur.

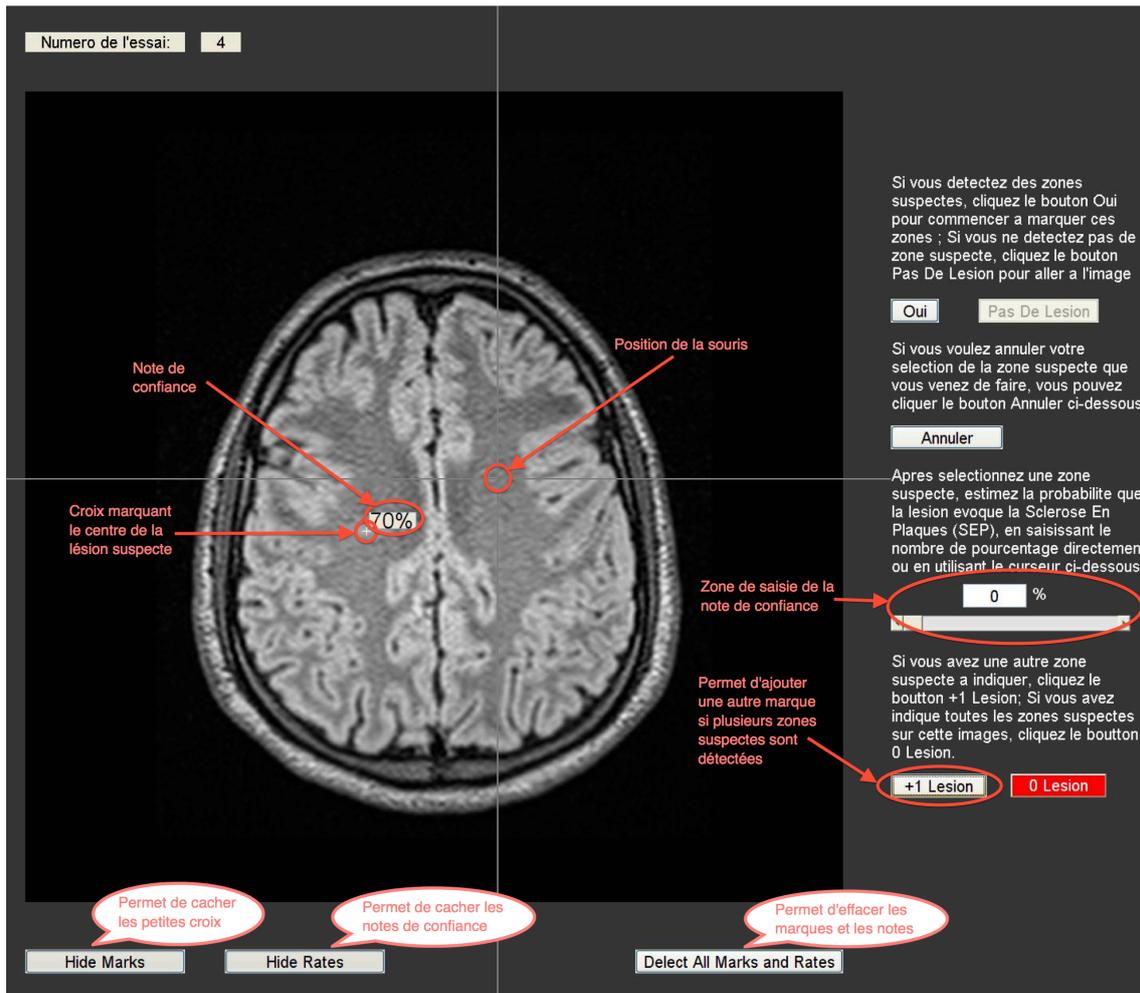


Figure 2: Interface graphique de l'expérience sur coupe 2D

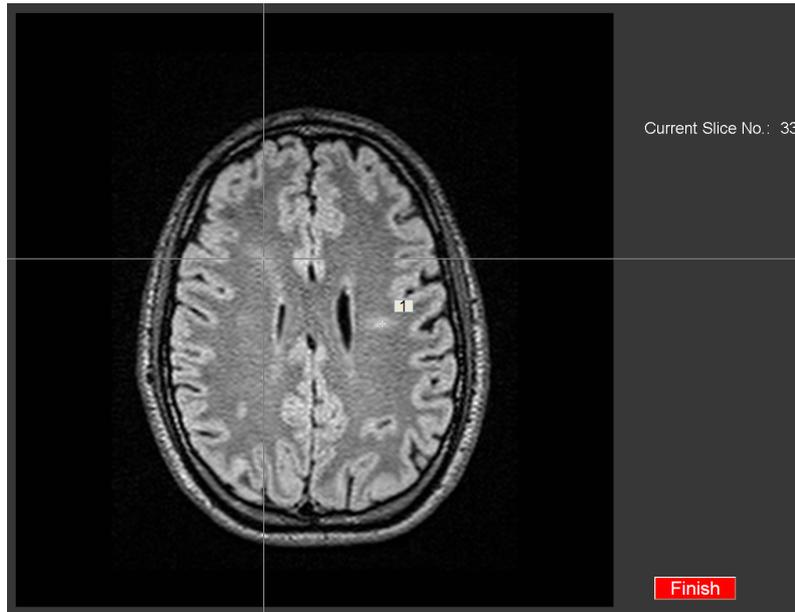


Figure 3: Interface graphique de l'expérience sur séquence volumétrique (3D)

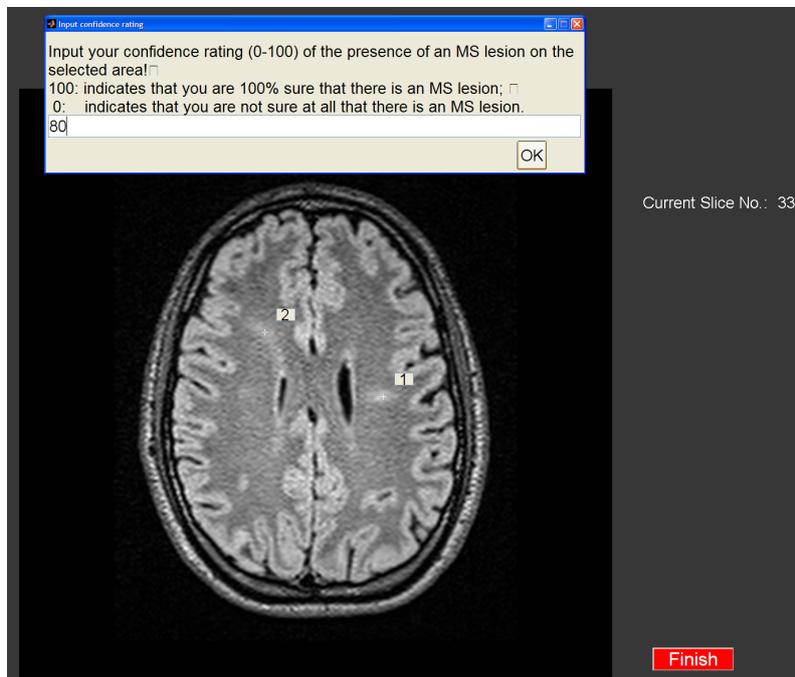


Figure 4: Interface graphique de l'expérience sur séquence volumétrique (3D), avec la boîte de saisie de la note de confiance

3 ROC et Variantes de ROC

Différentes figures de mérite ont été proposées pour caractériser les performances de différentes tâches diagnostiques, comme indiqué sur la Figure 5.

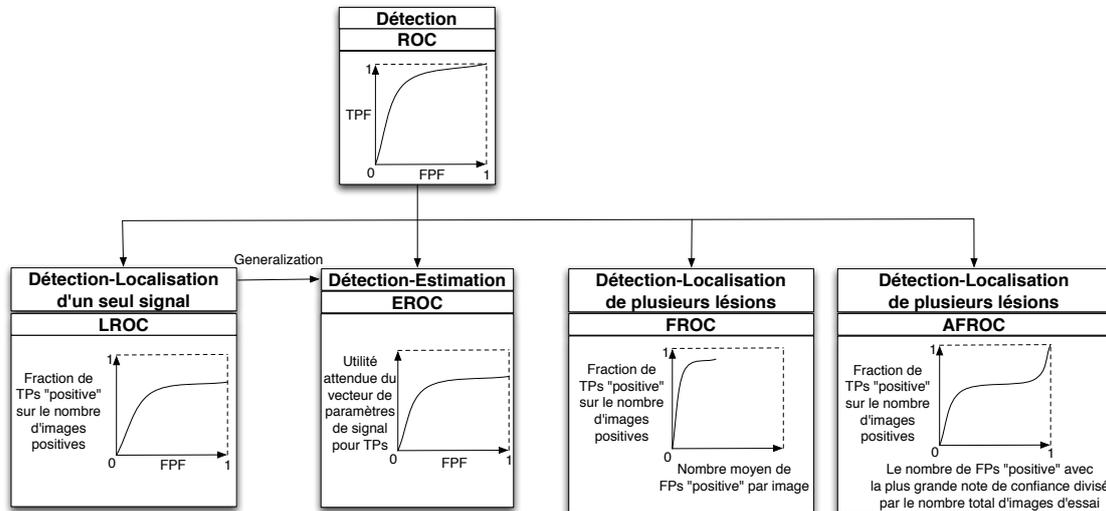


Figure 5: Figures de mérite pour différentes tâches

L'analyse ROC (Receiver Operating Characteristic ou ROC) permet d'évaluer la performance d'un système pour la tâche de détection lorsque le seuil de discrimination varie. Une décision est dite :

- *vrai positive* (TP) : lorsque la présence d'un signal est déclarée par l'observateur et qu'il est aussi présent selon le gold standard ;
- *faux positive* (FP) : lorsque la présence d'un signal est déclarée par l'observateur alors qu'il est absent selon le gold standard.

La courbe ROC donne le taux de vrais positifs (TPF) en fonction du taux de faux positifs (FPF), voir un exemple Figure 5. Nous pouvons aussi calculer l'aire sous cette courbe (AUC pour Area Under the Curve) afin d'obtenir une mesure scalaire, qui mesure la capacité d'un système à discriminer la présence ou l'absence d'une lésion.

L'analyse ROC caractérise uniquement la tâche de détection. Pour la tâche de détection-localisation, comme dans nos expérimentations, des variantes de ROC ont été développées :

- LROC pour Localization ROC, traite les cas où il y a un seul signal sur l'image. Pour inclure la précision de la localisation dans l'évaluation, l'ordonnée de la courbe LROC est le taux de vrais positifs avec localisation correcte.
- FROC pour Free ROC et AFROC pour Alternative Free ROC, traitent les cas où il y a plusieurs signaux sur une image. En conservant l'ordonnée de la courbe LROC, la courbe FROC utilise le nombre moyen de décision FP par image en abscisse. La seule différence entre les courbes AFROC et FROC est leur abscisse, l'abscisse de la courbe AFROC est le taux de décision de présence d'un signal ou de plusieurs signaux sur les images sans lésion sur le nombre total d'images sans lésion.

Il convient de noter que la tâche de localisation concerne un critère de proximité en présence d'un signal selon le gold standard, qui peut être un rayon d'acceptation ou un pourcentage de chevauchement de la zone détectée et du vrai signal.

4 Performances des radiologues dans les expérimentations 2D et 3D

Un radiologue expert en SEP et quatre radiologues généralistes ont participé aux deux expérimentations. Afin de quantifier les performances de détection-localisation des radiologues, nous avons utilisé la figure de mérite JAFROC1 (une mesure de l'aire sous la courbe AFROC). Dans la littérature, il a été montré que parmi les différentes figures de mérite pour quantifier la performance de la tâche de détection-localisation, JAFROC1 possède la plus grande puissance statistique [3].

Les tableaux 1 et 2 listent pour chaque radiologue la valeur de JAFROC1 avec son écart-type, le taux de détection (nombre de décisions TP divisé par le nombre total de lésions simulées), le nombre de décisions TP et le nombre de décisions FP, pour les expérimentations sur une coupe 2D et sur une séquence 3D, respectivement.

| | JAFROC1 | Ecart-type | Taux de détection | # TP | # FP |
|--------------|---------|------------|-------------------|------|------|
| Expert 1 | 0.7672 | 0.0326 | 0.8953 | 77 | 2 |
| Radiologue 1 | 0.4736 | 0.0378 | 0.8372 | 72 | 3 |
| Radiologue 2 | 0.4278 | 0.0399 | 0.9186 | 79 | 2 |
| Radiologue 3 | 0.4742 | 0.0419 | 0.7209 | 62 | 5 |
| Radiologue 4 | 0.4728 | 0.0369 | 0.6163 | 53 | 5 |

Table 1: Performances des radiologues dans l'expérimentation sur une coupe (2D)

| | JAFROC1 | Ecart-type | Taux de détection | # TP | # FP |
|--------------|---------|------------|-------------------|------|------|
| Expert 1 | 0.9267 | 0.0161 | 0.7241 | 105 | 22 |
| Radiologue 1 | 0.7837 | 0.0636 | 0.7448 | 108 | 112 |
| Radiologue 2 | 0.9192 | 0.0420 | 0.6690 | 97 | 107 |
| Radiologue 3 | 0.6547 | 0.0852 | 0.6897 | 100 | 110 |
| Radiologue 4 | 0.6320 | 0.0995 | 0.5172 | 75 | 76 |

Table 2: Performances des radiologues dans l'expérimentation sur une séquence (3D)

Nous observons que les performances des radiologues sont plus élevées dans le cas des séquences 3D. Grâce à l'information volumétrique, la performance des radiologues non experts en interprétation des images IRM de SEP augmente sensiblement et peut atteindre une performance proche de celle d'un expert (cas du Radiologue 2).

5 Conclusion

Dans ce papier, nous avons expliqué le principal de l'approche *task-based assessment* pour l'évaluation de la qualité des images médicales 2D et 3D en utilisant l'observateur humain. Nous avons également montré les résultats des performances de détection-localisation jointe des radiologues dans les deux cas en caractérisant leur performance avec la figure de mérite JAFROC1.

En comparant la performance des radiologues (observateurs humains) avec celle des observateurs numériques, nous pouvons également valider ces observateurs numériques.

Les observateurs numériques basés sur les approches *task-based assessment* pourront alors être appliqués lors de la comparaison des performances diagnostiques de systèmes d'imagerie (système d'acquisition, système de traitement d'image et système de visualisation) [4], comme indiqué dans la Figure 6.

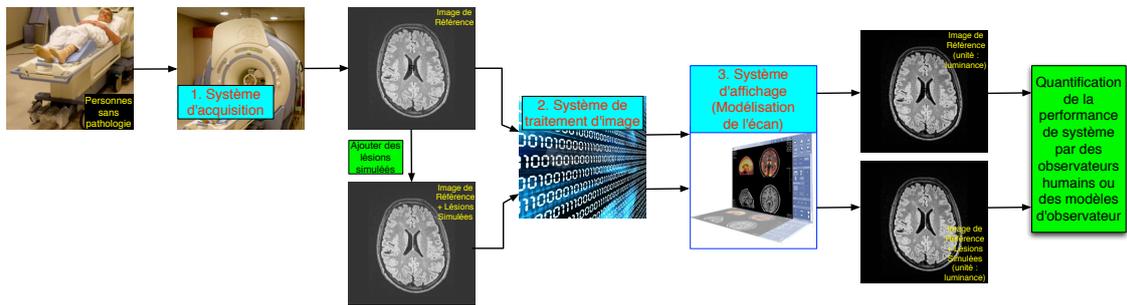


Figure 6: Comparaison des performances des systèmes d'imagerie

References

- [1] H.H. Barrett and K.J. Myer. *Foundations of Image Science*. John Wiley and Sons, Inc., Hoboken, New Jersey, USA, 2004.
- [2] C. Cavaro-Ménard, L. Zhang, and P. Le Callet. Diagnostic quality assessment of medical images: Challenges and trends. In *Visual Information Processing (EUVIP), 2010 2nd European Workshop on*, pages 277–284, July 2010.
- [3] D. P. Chakraborty. New developments in observer performance methodology in medical imaging. *Seminars in Nuclear Medicine*, 41:401–418, 2011.
- [4] L. Zhang, C. Cavaro-Ménard, P. Le Callet, and J. TANGUY. A perceptually relevant channelized joint observer (pcjo) for the detection-localization of parametric signals. *IEEE Transactions on Medical Imaging*, 31:1875–88, 2012.