# RECONSTRUCTION DE CARTES DE PROFONDEUR À PARTIR D'UNE CAPTATION RGB MONOCULAIRE

*Nicolas DEMAGNY[1], Enora ROLLAND[1], Théo PAUMARD[1]*

([1]) Université de Rennes, ESIR 3 IN

Représentant entreprise B-COM : *Maxime PAPIN* ( Maxime.PAPIN@b-com.com )

**ABSTRACT**

Pour estimer la profondeur d'une image monoculaire, il existe plusieurs modèles de réseau de neurones applicables à différents contextes. Nous avons choisi MiDAS pour obtenir une première estimation. Ensuite, grâce aux deux capteurs Aruco placés à différentes profondeurs connues dans la scène, nous appliquons une mise à l'échelle au résultat de MiDAS. Nous considérons ensuite une suppression des valeurs aberrantes par simples seuils. Ainsi, nous obtenons une carte de profondeur en mètres (m) répondant à une précision moyenne de 96% de points estimés à moins de 20 cm de la vérité terrain.

***Mots clés—*** carte de profondeur, intelligence artificielle, monoculaire, vision par ordinateur

## 1. INTRODUCTION

Ce projet a pour objectif l'élaboration d'une solution de reconstruction de cartes de profondeur à partir d'une captation RGB monoculaire. Autrement dit, l'identification de la profondeur des éléments qui composent une image issue d'une seule caméra.

Habituellement, les cartes de profondeur sont reconstruites à partir de plusieurs images d'une même scène ou acquises avec du matériel spécifique (ie. LIDAR ou infrarouge). Ainsi, ce projet consiste à trouver une alternative qui s'affranchit de ces contraintes. Des exemples d'estimation de cartes de profondeurs sont visibles sur la figure 1.

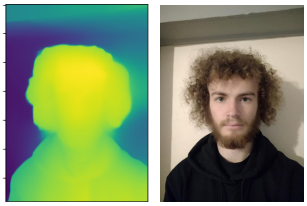

*Figure 1 : Exemple d'estimation de carte de profondeur.*

## 2. PRÉSENTATION DU CADRE

### 2.1. L'entreprise

B<>com est un Institut de Recherche Technologique privé fondé il y a 10 ans. Sa vocation est d'explorer, concevoir et fournir des innovations aux entreprises qui veulent développer leur compétitivité grâce au numérique. La société fonctionne sur un modèle de co-investissement entre la recherche et l'entreprise privée.

Le travail présenté dans ce papier a été fait sous la supervision de Maxime Papin, tuteur entreprise.

### 2.2. La problématique

Une caméra RGB-D fournit simultanément une image couleur et une carte de profondeur caractérisant la distance des objets vus dans l'image. B<>com les exploite et les interprète dans leurs applications de réalité augmentée appliquées à l'estimation de pause humaine..

Cependant le coût important et la faible démocratisation de ces outils présentent un désavantage majeur. L'enjeu des projets de réalité augmentée de l'IRT est de proposer des solutions facilement déployables et transportables (à l'hôpital ou à domicile par exemple). C'est pour cela que l'institut cherche à s'affranchir du recours à une caméra avec un capteur de profondeur.

Ce projet a donc pour problématique **la reconstruction de carte de profondeur à partir d'une caméra RGB monoculaire**. Il comprendra les étapes suivantes : (3.) Recherches et état de l'art sur les solutions de deep learning et leurs limites. (4.) Un rapport détaillé sur le fond et la forme des informations contenues par la carte de profondeur estimée. La constitution d'une échelle métrique fidèle et d'un nuage de points 3D. (5.) Une évaluation de la robustesse et des performances.

## 3. ETAT DE L'ART

### 3.1. Méthodes étudiées

La première mission qui nous est confiée consiste à étudier l'ensemble des techniques et modèles disponibles aujourd'hui qui permettent l'estimation de carte de profondeur à partir d'une unique image RGB. Les deux techniques les plus répandues actuellement sont basées sur des images stéréos et sur des architectures de réseaux de neurones convolutifs encodeur-décodeur. Comme précisé lors de l'introduction, nous n'avons qu'une seule image. Ainsi, c'est dans le domaine des modèles de réseaux de neurones que nous poussons notre étude.

Lors de nos recherches nous constatons que les différences qu'il y a entre les modèles sont plutôt subtiles. Par exemple, le modèle BTS [1] ajoute des couches de guidage planaire à la partie décodeur du réseau, le modèle AdaBins [2] ajoute un module nommé Adaptive Bins pour compléter le décodage.

Le premier problème rencontré est la constante évolution et l'amélioration des techniques de Deep Learning. Chaque mois un nouveau modèle encore plus performant est publié et d'autres sont améliorés. Nous orientons nos recherches vers un modèle stable, peu complexe pour une utilisation quasi en temps réel et dont la précision est suffisante pour notre projet.

Le second problème pour trouver le modèle idéal provient des datasets utilisés pour les entraîner. Les datasets les plus courants sont NYU [3] et KITTI [4]. Le premier est constitué d'images représentant des décors d'intérieurs (pièces, meubles), le second comprend des images de décors extérieurs captées à partir d'un véhicule (rues, villes). Ces bases de données contiennent énormément d'images ce qui est très bon pour un apprentissage solide. Cependant, notre tâche finale est très spécifique : on souhaite estimer la carte de profondeur de la pose d'une personne. Les datasets cités apportent peu de précision quant il s'agit d'estimer la profondeur d'un modèle humain. Ainsi, nous retenons le modèle MiDaS [5].

### 3.2. Modèle retenu

MiDaS est entraîné sur cinq datasets différents incluant des poses humaines dont un spécialement conçu pour l'occasion en se basant sur des scènes de film. Ce modèle utilise également NYU et KITTI pour son évaluation en plus d'autres datasets comme ETH3D [6] et DIW [7]. MiDaS est donc un modèle polyvalent avec des résultats qui semblent en accord avec nos objectifs. Ce modèle est libre d'exploitation avec une licence MIT en plus d'avoir un dépôt github public. De plus, il a l'avantage de proposer des modèles de différentes tailles pouvant s'exécuter sur un ordinateur comme sur un smartphone au détriment d'une dégradation de la précision et de la fiabilité des résultats d'après le tableau ci-dessous.

| Modèle MiDaS | small | hybrid | large |
|---|---|---|---|
| Zero-shot error | 13.43 | 8.69 | 8.32 |

*Tableau 1: robustesse des différents modèles de MiDaS tels qu'ils sont exprimés dans le papier, résultats calculés sur le dataset NYU*

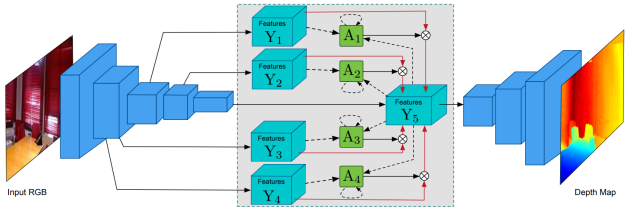


*Figure 2 : Illustration de l'architecture proposée par MiDaS pour l'estimation de la profondeur monoculaire.* [8]

La figure 2 ci-dessus illustre l'architecture proposée par MiDaS pour l'estimation de la profondeur monoculaire. Les blocs bleus indiquent le réseau de neurones convolutif frontal, qui dans leur implémentation est composé d'un encodeur et d'un décodeur. La boîte grise contient une représentation schématique du modèle, à l'intérieur, les cases vertes indiquent les données des cartes de profondeurs estimées et reconstruites avec le décodeur, tandis que les cases bleu clair représentent les caractéristiques de l'image retenues pour les évaluer (features). Les flèches indiquent les dépendances entre les variables.

L'autre option envisagée est d'entraîner notre propre modèle à partir d'une base de données constituée uniquement de poses humaines représentatives de notre contexte. Cependant, l'élaboration d'une telle base de données est longue et fastidieuse étant donné qu'il faut une caméra RGB-D pour la vérité terrain et plusieurs personnes de morphologies différentes pour varier l'entraînement. De plus, il est nécessaire d'avoir le matériel adapté à l'apprentissage d'un réseau de neurones sans pour autant avoir la garantie que ce modèle sera optimisé pour une utilisation en temps réel. En conclusion, il y a trop de paramètres contraignants et coûteux en temps à l'élaboration de notre propre modèle dans le cadre du projet industriel.

## 4. NOTRE DEMARCHE

### 4.1. Gestion managériale

Pour s'assurer du bon déroulement de ce projet, nous allons user de techniques managériales étudiées durant notre cursus.

Nous avons découpé notre projet en 3 tâches principales, et nous avons pour celles-ci évalué une valeur quantifiée de leur approbation selon la méthode SMART.

| Tâche | Outils | Format | Approbation |
|---|---|---|---|
| Etude de l'existant et approbation | Bibliographies et recherches | Rapport Compte-rendu | Evalué avec le pôle entreprise |
| Implémentation sur nos données | Données de test, Python | Exécutable | Robustesse, temps |
| Récupérer une carte de profondeur métrique | Capteurs Aruco, vérité terrain | Carte de profondeur métrique | Précision < 20 cm |

*Tableau 2: Division et analyse des principales tâches par la méthode SMART.*

Nous allons ensuite réaliser un tableau d'analyse de gestion des risques selon leur probabilité et leur gravité. L'objectif étant de prévoir des solutions au préalable afin de les anticiper.

| Risque | Gravité G (sur 4) | Probabilité P (sur 4) | Criticité (G*P) |
|---|---|---|---|
| Manquer de données de test. | 1 | 4 | 4 |
| Les résultats ne sont pas suffisants | 2 | 4 | 8 |
| Les modèles appris sont hors-contextes | 4 | 3 | 12 |
| Problème de droits d'auteur des modèles ou dataset utilisés | 4 | 2 | 8 |
| Les objectifs ne sont pas atteints | 4 | 1 | 4 |

*Tableau 3: Analyse de gestion des risques par niveau de gravité (1-faible, 4-critique), de probabilité P (1-faible, 4-très probable) et de la criticité du risque.*

Pour répondre à cette analyse des risques, nous l'avons complétée de réflexions et d'idées pour les prévenir :

- Assurer une **communication efficace** entre les différentes parties.
- **Planifier** et faire des **points réguliers** sur l'avancement du projet.
- Réaliser une **étude de l'existant** complète et rigoureuse, notamment à propos des licences d'utilisation.
- Assurer et accompagner la **montée en compétence.**

Enfin, avec l'ambition de respecter les deadlines et d'avoir une avancée effective tout au long du projet, nous avons réalisé une nouvelle division des tâches afin de les répartir précisément suivant un planning de Gantt.

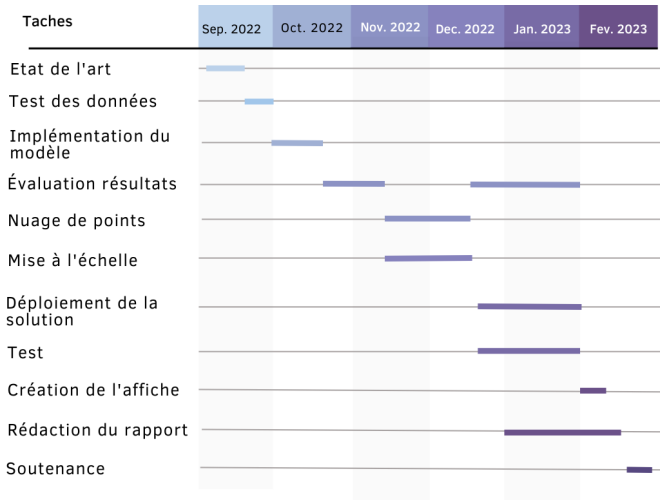


*Figure 3 : Planning de GANTT de notre projet.*

Nous nous sommes régulièrement réunis (au moins 1 fois par mois) avec notre tuteur entreprise pour faire part de nos avancées et pour prévoir nos nouveaux objectifs pour notre prochaine rencontre, tout en gardant en vue les deadlines et la comparaison de nos avancées avec notre planning prévisionnel de Gantt ci-dessus.

### 4.2. Notre méthode

#### *4.2.1. Acquisition de l'estimation*

Notre estimation est obtenue grâce à l'implémentation de MiDaS.

Nous lui donnons en entrée notre image RGB qui va déjà suivre quelques étapes de pré-traitement. Nous réduisons la résolution de l'image pour accélérer les calculs de prédictions. Plus la dimension est réduite, plus les calculs sont rapides. Cependant la qualité de la prédiction peut être altérée. Comme précisé dans la partie 3.3, MiDaS propose différents modèles. Dans ce projet, nous utilisons le modèle 'small' qui permet de grandement réduire le temps de calcul au détriment de la qualité de prédiction. Cependant cette dernière reste largement exploitable. Nous nous permettons

même de baisser la résolution de nos images jusqu'à 320x180.

Les estimations ainsi calculées par MiDaS seront à comparer avec la vérité terrain obtenue avec une caméra RGB-D pour évaluer la performance de cette méthode. Les estimations de MiDaS étant fournies en 16 bits, il faudra au préalable les convertir en format 8 bits pour avoir une comparaison juste. De plus, les valeurs de la prédiction ne sont pas à la même échelle que celles de la vérité terrain. Il nous faudra donc adapter notre prédiction à la vérité terrain pour pouvoir les comparer efficacement.

*4.2.2. Mise à l'échelle de la carte de profondeur avec les capteurs Aruco*

Dans le but de tester notre méthode dans une configuration représentative de la réalité. Nous avons acquis plusieurs séquences d'images de poses humaines avec une caméra RGB-D.

*Figure 4 : scène capturée avec une caméra RGB-D*

Dans notre scène, nous avons au préalable placé deux marqueurs Aruco [9] à deux profondeurs bien distinctes visibles sur la figure ci-dessus. De préférence, un premier capteur au premier plan devant l'acteur, et un deuxième capteur derrière lui en arrière-plan.

Cette distinction va nous permettre de réaliser une mise à l'échelle métrique qui sera d'autant plus précise pour les valeurs du personnage dans la scène entre les deux capteurs.

Pour ce faire, nous nous plaçons dans le cas idéal où la caméra est à l'horizontale et où les marqueurs Aruco sont orientés dans le plan de l'image. Ainsi, nous négligeons ici dans notre scène les effets de perspectives. Connaissant la taille des capteurs dans la scène réelle, nous pouvons obtenir avec exactitude leur distance à la caméra, et donc le rapport relatif à leur profondeur estimée par MiDaS.

*4.2.3. Mise à l'échelle*

Grâce aux données de vérité terrain obtenues avec le placement des deux marqueurs Aruco dans l'image, nous pouvons calculer le coefficient de mise à l'échelle entre les données que nous donne MiDaS et les valeurs métriques.

Soient Z1 et Z2 nos vérités terrains de profondeur pour nos deux marqueurs . $\tilde{z}1$ et $\tilde{z}2$ sont les prédictions de ces mêmes points réalisées par MiDaS. Nous supposons que la relation entre la profondeur de la vérité terrain et de l'estimation est affine. Nous recherchons les paramètres S et O tels que :

$$Z1 = \{\tilde{z}1 * S + O \qquad (1)$$
$$Z2 = \{\tilde{z}2 * S + O \qquad (2)$$

Soit selon (1) - (2) :

$$\{S = \frac{Z1 - O}{\tilde{z}1} \qquad (3)$$
$$\{Z1 - Z2 = S * (\tilde{z}1 - \tilde{z}2) \qquad (4)$$

Ainsi :

$$\{S = \frac{Z1 - Z2}{\tilde{z}1 - \tilde{z}2} \qquad (5)$$

Et nous pouvons ensuite déduire l'offset O grâce aux équations (1), (2) ou (3).

La mise à l'échelle est ensuite appliquée point par point sur notre première estimation de MiDaS. Nous obtenons finalement une carte 3D à l'unité métrique, avec laquelle nous allons pouvoir réaliser une étude de performance et évaluer la robustesse par rapport à notre vérité terrain.

*4.2.4. Résultat*

Grâce à notre implémentation précédente, nous pouvons afficher notre estimation sous forme de nuage de points, et la comparer à celui de la vérité terrain.

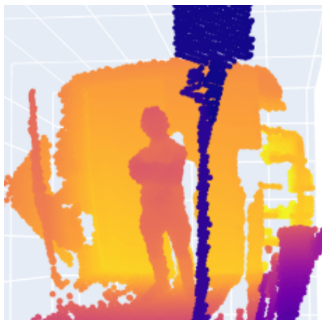

*Figure 5 : Nuage de points de la vérité terrain.*

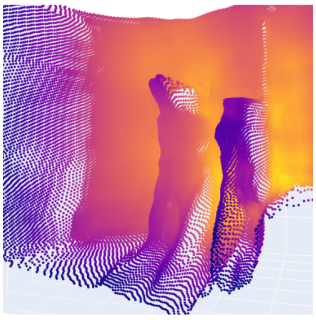

*Figure 6 : Nuage de point de notre prédiction.*

Nous remarquons sur notre estimation que la majorité des détails précis des formes et des corps sont perdus, mais il reste des protubérances au niveau du modèle humain et de l'emplacement du premier capteur, ce qui permet d'avoir une information sur leur profondeur cohérente.

## 5. ETUDE DE PERFORMANCE

### 5.1. Mesure de robustesse

*5.1.1 Première évaluation*

Pour évaluer les performances de la méthode et les résultats obtenus, nous calculons trois indicateurs de performances : une RMSE, une distance moyenne et une proportion de bonnes prédictions.

La RMSE, l'erreur quadratique moyenne est une métrique définie par l'écart entre la valeur de la prédiction (pred) et la valeur de la vérité terrain (truth). N est le nombre de pixels d'une image.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i,j}(pred_{ij} - truth_{ij})^2} \qquad (6)$$

La distance moyenne est calculée en faisant la moyenne des écarts pixels à pixels de la prédiction et de la vérité terrain.

$$DM = \frac{1}{N}\sum_{i,j}|pred_{ij} - truth_{ij}| \qquad (7)$$

La proportion de bonnes prédictions est une donnée en pourcentage qui montre la proportion de points bien placés à moins d'un seuil préalablement fixé à 20 cm. Un point bien placé vérifie la condition suivante :

$$|pred_{ij} - truth_{ij}| < seuil \qquad (8)$$

La RMSE est la métrique communément utilisée pour comparer l'efficacité de prédiction d'un réseau de neurones convolutif. Nous nous en servons pour étudier la robustesse de modèles avec nos propres données de test. La distance moyenne donne une valeur en millimètre comme la valeur des pixels de la vérité terrain. Nous obtenons donc l'erreur moyenne entre la prédiction et la vérité terrain ce qui rend la donnée plus facile à comprendre et à comparer avec d'autres issues de différentes images.

On calcule ces métriques à partir d'une séquence de 128 images. Les résultats sont présentés dans le tableau ci-dessous :

| | RMSE | DM | PS |
|---|---|---|---|
| Résultats | 70.498 | 5.5 | 54.59 |

*Tableau 4: résultats obtenus sur une séquence (RMSE : erreur quadratique moyenne (en mm sur 8bit), DM : distance moyenne (en m), PS : proportion de bonne prédiction selon un seuil de 20 cm (en %)*

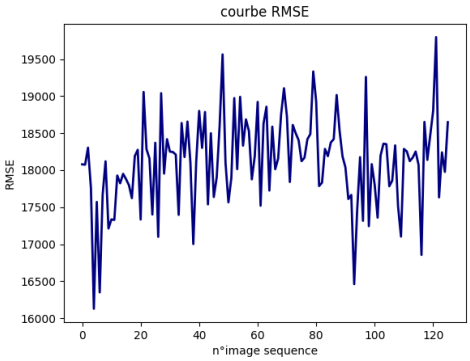


*Figure 7: Evolution de la RMSE (en mm, sur la carte de profondeur exprimée sur 16 bits) sur une évaluation de 128 images.*

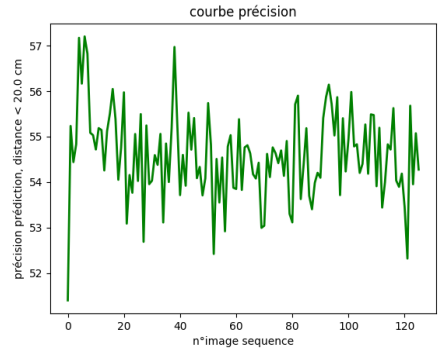


*Figure 8: Evolution de la précision (PS en %) sur une évaluation de 128 images.*

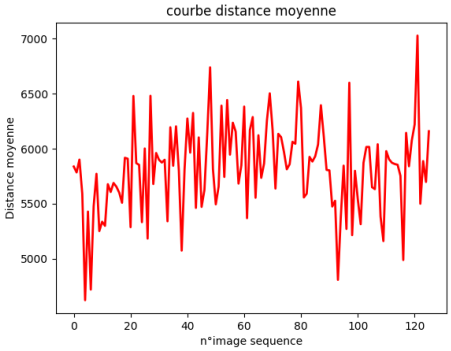


*Figure 9: Évolution de la distance moyenne (DM en mm) sur une évaluation de 128 images.*

La séquence de 128 images étudiée sur les figures 6, 7 et 8 est en réalité une séquence filmée d'une scène où rien ne bouge. Cette scène est identique à la scène montrée figure 3.

Comme nous pouvons l'observer sur la figure 6, 7 et 8 La RMSE, la DM et et la PS sont trois métriques étroitement corrélées. La précision est le symétrique de la RMSE qui est elle-même identique à la distance moyenne.

Les variations dans les courbes sont dues aux aléas de MiDaS, en effet, pour chaque image il effectue une nouvelle prédiction et chaque prédiction peut être différente de la précédente car le modèle n'est pas déterministe.

*5.1.2 Amélioration de la performance*

Nous avons distingué des valeurs aberrantes dans la prédiction de MiDaS causées par certaines images de test. Le contraste, la lumière, des objets de la scène ont pu perturber l'estimation. Sur la figure 6, dans la partie résultat de ce rapport, nous pouvons remarquer que les plus grandes erreurs d'estimation se trouvent sur les bords de l'image (là où se trouve le mur), à droite (où se trouve le lit). Dans ces cas, le modèle les estime comme beaucoup plus au premier plan qu'ils ne le sont réellement.

Nous décidons donc de calculer les métriques d'évaluations uniquement avec les pixels aux alentours de notre région d'intérêt. Pour cela, nous appliquons des seuils pour retirer les éléments situés devant et derrière le sujet. On estime ces seuils à 900 et 2700 mm. Ces valeurs permettent de retirer les étagères situées au fond de la scène et des parties du mur de gauche et du lit situé à droite. Ce post-traitement nous permet ainsi de mieux calculer la robustesse du modèle au niveau de notre zone d'intérêt autour du sujet.

Avec ce traitement, nous recalculons nos métriques. Les résultats sont présentés dans le tableau ci-dessous.

| | RMSE | DM | PS |
|---|---|---|---|
| Résultats | 2,723 | 0.549m | 96.38 |

*Tableau 5: résultats améliorés obtenus sur la même séquence (RMSE : Erreur quadratique moyenne (en mm), DM : distance moyenne, PS : proportion de bonne prédiction selon un seuil de 20 cm (%) )*

Nous constatons qu'en retirant près de la moitié des valeurs, et seulement les plus aberrantes, nos résultats deviennent bien meilleurs comme nous pouvons constater dans le tableau 5. Cela est dû aux décalages entre ce que prédit MiDaS pour les éléments du décor et leur réelle profondeur.

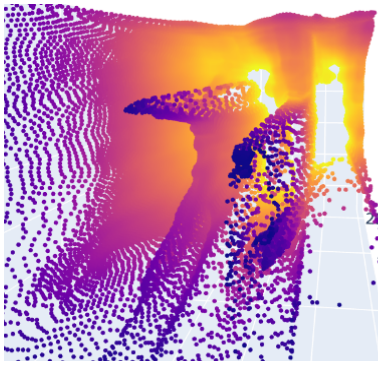

*Figure 10 : Nuage de point de notre prédiction avec amélioration*

Nous remarquons sur la figure 10 la disparition des éléments perturbants comme le lit (en bas à droite) et une partie du fond où se trouve des étagères. Ayant retiré les points les plus proches et les plus éloignés avec cette méthode, le nuage de points s'est agrandi et met en évidence la tête et les pieds du sujet.

Notre précision à évolué à 96.38%. Elle correspond à la probabilité qu'un point se trouve à moins une distance de 20 cm à ce même point dans la vérité terrain.

*5.1.3 Robustesse de la méthode*

Nous avons comparé notre RMSE à celles obtenues par MiDaS sur différents datasets. Étant donné que nos données sont en 16 bits, nous avons dû faire la conversion sur 8 bits afin de pouvoir les comparer.

| Datasets | RMSE | RMSE Log |
|---|---|---|
| KITTI | 2.573 | 0.092 |
| NYU | 0.357 | □ |
| DCM | 1.033 | 0.375 |
| eBDtheque | 1.416 | 0.659 |
| Notre test | 2,723 | 0.525 |

*Tableau 6: Comparaison des résultats de notre ensemble de test par rapport aux résultats de MiDaS sur différents datasets.*

Comme nous pouvons le constater, les résultats de MiDaS sur notre dataset de test ont une moins bonne RMSE que ceux obtenus par d'autres datasets comme KITTI ou NYU. Cela peut s'expliquer par de multiples facteurs notamment l'éloignement de nos données de test par rapport aux données d'entraînement du modèle. Nos images sont prises dans un contexte que le modèle n'a probablement jamais rencontré. Les données d'entraînement sur lesquelles il a appris ne sont pas en majorité des personnes en mouvement.

La luminosité de la pièce ainsi que les paramètres intrinsèques de la caméra peuvent perturber MiDaS. De plus, des détails sur l'image, tel que le contraste peuvent également fausser les résultats. Si une utilisation plus concrète de MiDaS était considérée, il serait alors judicieux de normaliser les images en entrée, et de trouver un traitement de contraste et de luminosité qui permet un résultat optimal pour l'estimation. D'une autre manière, il serait possible de faire un fine tuning sur le modèle utilisé pour compléter son entraînement avec des données représentatives de la réalité.

**5.2. Temps de calcul**

Afin d'évaluer la faisabilité de notre solution en temps réel, nous avons calculé le temps que cela nous prendrait de calculer MiDaS sur une séquence d'images (sur CPU intel® i5-7300hq 2.50ghz). Pour chacun de nos tests nous nous sommes basés sur le même modèle à savoir MiDaS-small.

Nous avons expérimenté avec différentes résolutions en images d'entrée.

1280x720

640x360

320x180

*Figure 11: Résultat de MiDaS calculé sur une même image avec différentes résolution*

| Résolution | Fréquence |
|---|---|
| **1280x720** | 5 images/s |
| **640x360** | 8 images/s |
| **320x180** | 9.5 images/s |

*Tableau 7: Table de performance de MiDaS en temps réel*

Comme nous pouvons l'observer dans le tableau 7, diviser la résolution des images en entrée par quatre permet d'accélérer le temps de traitement des images par deux. Pour chacun de ces cas, comme nous pouvons l'observer sur la figure 10, les résultats obtenus sont très semblables. Baisser la résolution des images en entrée n'a donc pas tellement d'incidence sur la robustesse et la précision de l'estimation. Or cela permet de fortement accélérer la vitesse de traitement d'après nos observations répertoriées dans le tableau 6.

## 6. CONCLUSION

L'objectif de notre période de recherche était de découvrir l'état de l'art dans le domaine de l'estimation de profondeur, l'amélioration et l'implémentation dans un projet industriel.

Nous avons donc découvert via un état de l'art les différents modèles de réseau de neurones existants répondant à l'estimation de cartes de profondeur à partir d'une image monoculaire. Nous les avons comparés, et nous nous sommes principalement intéressés à leur contexte d'utilisation très divers. Notre problème correspondant à une estimation de squelette humain dans une pièce, nous avons opté pour MiDaS.

Dans le but de répondre à la limitation de MiDaS qui ne fournit pas une carte en valeur métrique, nous cherchons à améliorer les résultats de l'estimation par une mise à l'échelle utilisant des connaissances sur la taille de capteurs présents dans la zone acquise. Nous réalisons ensuite une mise à l'échelle de notre prédiction MiDaS, et une suppression par seuil des valeurs aberrantes.

Lors de notre comparaison des performances de MiDaS associé au post-traitement du résultat, sur nos données comparées à une application classique sur d'autres datasets, nous obtenons une erreur (RMSE) similaire, légèrement supérieure. Nous obtenons une précision à 96% évaluée pour une distance au point inférieure à 20 cm de la vérité terrain.

Pour aller plus loin, nous pourrons améliorer la robustesse de notre algorithme de post-traitement en utilisant autre chose que de simples seuils. Il serait aussi possible d'automatiser la procédure de mise à l'échelle qui utilise actuellement l'existence de deux marqueurs Aruco dans notre scène, en se basant sur des a priori sur la taille d'un humain ou d'un objet présent dans la pièce par exemple. De plus, nous supposons que les capteurs Aruco sont placés parallèlement au plan de l'image, et que l'échelle utilisée est une échelle linéaire. Or, nous pouvons obtenir des informations concernant le plan des capteurs dans l'image, chose que nous avons actuellement négligée dans notre algorithme.

Concernant le modèle d'estimation utilisé, il aurait été intéressant, avec plus de temps et de moyens, de réaliser un entraînement supplémentaire (fine-tuning) sur des images prises dans notre contexte exact. Au niveau performance temporelle, il était envisagé un fonctionnement en temps réel. Pour se faire, il faudrait penser à utiliser des informations temporelles pour interpoler des images intermédiaires sans y réaliser l'estimation par MiDaS.

Les résultats actuellement observés sont convaincants, et il reste à voir s'ils sont suffisants pour une utilisation précise et robuste dans un cadre industriel. Il serait aussi possible d'étudier l'impact du pré-traitement des images en entrée de MiDaS pour voir s'il n'y a pas des transformations qui augmentent la qualité de l'estimation. Imaginons par exemple une normalisation ou une égalisation d'histogramme qui permettrait d'ajuster les contrastes et la luminosité de l'image, afin de rendre les estimations de MiDaS plus robustes et uniformes en toutes circonstances.

## 7. RÉFÉRENCES


[1] Lee J. H., Han M.K., Ko W. et Suh I.H. (2021). Titre de l'article. *From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation.*

[2] Bhat S. F., Alhashim I. et Wonka P. (2020). Titre de l'article. *Adabins: Depth Estimation using Adaptive Bins.*

[3] Silberman N., Hoiem D., Kohli P. et Fergus R. (2012). *Indoor segmentation and support inference from rgbd images. In Computer Vision – ECCV*, pp 746– 760.

[4] Geiger A., Lenz P., Stiller C. et Urtasun R. (2013). Titre de l'article. *Vision meets Robotics: The KITTI Dataset.*
[5] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. et Koltun, V. (2020). Titre de l'article. *Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer*.
[6] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, (2017) Titre de l'article. *A multi-view stereo benchmark with highresolution images and multi-camera videos*
[7] W. Chen, Z. Fu, D. Yang, and J. Deng. (2016), Titre de l'article. *Single-image depth perception in the wild*
[8] Dan X., Wei W., Hao T., Hong L., Nicu S., Elisa R. (2018). Titre de l'article. *Structured Attention Guided Convolutional Neural Fields for Monocular Depth Estimation*
[9] S. Garrido-Jurado, R. Munoz-Salinas, F.J Madrid-Cuevas, M.J. Marın-Jimenez Department of Computing and Numerical Analysis. *Automatic generation and detection of highly reliable fiducial markers under occlusion*