Étude comparative de méthodes de vision par ordinateur pour la détection de quasi-doublons d'images de coraux

Alissa Chahine[1,2], Juliette Van Poulle[1,2]

[1] Université de Technologie de Compiègne, Département Génie Informatique
[2] Aubay, Lab'Innov

Résumé

La qualité du jeu de données est une préoccupation récurrente en vision par ordinateur. Cette étude évalue et compare différentes méthodes, classiques et profondes, de détection de quasi-doublons afin d'améliorer la qualité d'un jeu de données d'images de coraux. Les quasi-doublons sont définis comme des images distinctes présentant de faibles variations et représentant un même individu de corail. Leur présence dans un jeu de données réduit l'efficacité de l'apprentissage des modèles, peut biaiser l'évaluation de leur performance et augmente inutilement les besoins en stockage et en calcul. De plus, les images sous-marines présentent des altérations importantes. Ce facteur, combiné aux motifs spécifiques des coraux, confère à cette problématique un besoin important de robustesse et d'adaptation des méthodes de détection existantes.

Mots-clés

quasi-doublons, vision par ordinateur, hachage perceptuel, NetVLAD, ORB, DELF

Abstract

Dataset quality is a recurring concern in computer vision. This study evaluates and compares different near-duplicate detection methods, both classical and deep learning-based, to improve the quality of a coral image dataset. Near-duplicates are defined as distinct images exhibiting minor variations while representing the same coral individual. Their presence in a dataset reduces the effectiveness of model training, can bias performance evaluation, and unnecessarily increases storage and computational requirements. In addition, underwater images often present significant alterations. This factor, combined with the specific visual patterns of corals, highlights the need for robustness and adaptation of existing detection methods to address this challenge.

Keywords

near-duplicates, computer vision, perceptual hashing, NetVLAD, ORB, DELF

1 Introduction

Les quasi-doublons sont des images représentant le même contenu mais avec de légères variations, telles que des changements de rotation, d'échelle, de recadrage, de luminosité, d'équilibre des couleurs ou de point de vue. Elles ne sont pas identiques pixel par pixel, mais elles représentent visuellement la même scène ou le même objet.

La présence de ces images au sein d'un jeu de données peut poser problème, selon le contexte dans lequel elle se manifeste. Il faut donc distinguer deux situations fondamentalement différentes. D'une part, l'augmentation de données est une pratique standard et volontaire en vision par ordinateur : elle consiste à introduire délibérément des variantes transformées d'images existantes afin d'améliorer la capacité de généralisation des modèles. D'autre part, la redondance non maîtrisée au sein d'un corpus, c'est-à-dire la présence de quasi-doublons non identifiés, répond à une tout autre logique et peut, elle, introduire des biais indésirables. Lorsque des quasi-doublons apparaissent de manière non contrôlée à la fois dans les ensembles d'entraînement et de test, la performance du modèle peut être artificiellement gonflée : au lieu de généraliser à de nouvelles observations, le modèle retrouve une image quasi identique à celle vue lors de l'entraînement. Cela conduit à des métriques de précision trompeuses et affaiblit la robustesse réelle du système. Par ailleurs, une telle redondance peut causer une surreprésentation non souhaitée de certaines classes et entraîner un besoin inutile de stockage et de calcul. Cette problématique est critique dans le contexte de jeux de données de surveillance environnementale, où les images sont souvent acquises en séquence dans des conditions similaires, favorisant l'émergence naturelle de quasi-doublons. Cette étude a été réalisée suite à la demande de la Fondation Science4Reefs, qui œuvre à la préservation des récifs coralliens en Polynésie française, auprès de la société Aubay. Son objectif est d'évaluer la faisabilité de la détection automatique de telles images au sein du jeu de données d'images de coraux fourni par la Fondation Science4Reefs. L'approche proposée adapte et compare trois catégories de méthodes : le hachage perceptuel, les descripteurs globaux d'images (NetVLAD), et les descripteurs locaux reposant sur la détection de points-clés (ORB et DELF). La combinaison de ces approches offre une analyse comparative des compromis entre précision, robustesse aux transformations et coût computationnel. Cette étude se concentre ainsi sur la capacité de chaque méthode à retrouver des quasi-doublons sous des transformations réalistes.

2 Contexte

Cette section pose les bases de l'étude comparative. On y présente d'abord les travaux existants dans la littérature, en détaillant quatre familles de méthodes d'extraction de descripteurs d'images retenues : le hachage perceptuel, NetVLAD, ORB et DELF. On décrit ensuite le jeu de données fourni par la Fondation Science4Reefs, composé de photographies sous-marines de coraux de Moorea, puis les métriques d'évaluation (Recall@K, mAP, Recall et Pureté) utilisées pour mesurer les performances des approches étudiée sont définies, avant de présenter la méthodologie mise en place pour l'étude comparative.

2.1 Travaux de la littérature

La détection de quasi-doublons est particulièrement étudiée en vision par ordinateur, bien que peu documentée pour les images aquatiques. Elle repose sur l'extraction de descripteurs d'images, étape centrale, et le calcul de leur similarité pour identifier les correspondances. Ainsi, cette section présente quatre méthodes d'extraction de descripteurs d'images retenues pour cette étude comparative : le hachage perceptuel, les descripteurs globaux et les descripteurs locaux. Leur sélection s'appuie un sur travail exploratoire de recensement et d'analyse critique de publications concernant la détection de quasi-doublons et la reconnaissance visuelle de lieux. Un benchmark a permis l'analyse comparative de ces méthodes à l'aide de plusieurs critères, notamment l'adéquation avec les ressources disponibles et leur robustesse face aux variations attendues.

2.1.1 Hachage perceptuel

Le hachage perceptuel est une technique qui consiste à transformer une image en une empreinte numérique de taille fixe, appelée hash. Cette empreinte résume les caractéristiques visuelles principales de l'image et permet de comparer efficacement des images entre elles sans avoir à les comparer pixel par pixel. Contrairement aux fonctions de hachage cryptographique classiques, qui produisent des empreintes totalement différentes dès qu'un seul pixel est modifié, le hachage perceptuel a pour propriété de préserver la similarité visuelle : deux images proches visuellement produisent des hashes proches, mesurables par la distance de Hamming. Plusieurs variantes existent dans la littérature, parmi lesquelles aHash (average Hash), dHash (difference Hash) et pHash (perceptual Hash) [8]. Ces méthodes se distinguent par la manière dont elles résument le contenu de l'image, respectivement par la moyenne des niveaux de gris, les gradients horizontaux locaux, et les coefficients de basse fréquence d'une transformée en cosinus discrète (DCT). Leur principal avantage est leur faible coût computationnel, qui les rend adaptées au traitement de jeux de données à grande échelle. En revanche, leur robustesse reste limitée face à des transformations géométriques importantes telles que les rotations ou les recadrages agressifs.

2.1.2 NetVLAD

NetVLAD [1] est une méthode d'apprentissage profond d'extraction de caractéristiques globales d'images. Elle est basée sur un réseau de neurones convolutifs (CNN) intégrant une couche d'agrégation de type Vector of Locally Aggregated Descriptors (VLAD) qui permet un apprentissage de bout en bout. Bien que le terme « NetVLAD » corresponde à la couche d'agrégation, il désigne, dans cette étude, la méthode complète. Le CNN est tronqué au niveau de sa dernière couche convolutive et sert d'extracteur de descripteurs locaux, robustes aux variations d'éclairage et d'angle tout en conservant leurs caractéristiques spatiales. La couche NetVLAD permet d'agréger les descripteurs locaux extraits par le CNN en une représentation globale et compacte de l'image. Il s'agit d'une généralisation de la couche VLAD classique, qui y introduit un mécanisme d'attribution plus souple. Les descripteurs d'images globaux produits sont adaptés à une reconnaissance visuelle de lieux sous faible supervision et restent robustes face à des transformations courantes dans les milieux aquatiques, ce qui rend la méthode fortement pertinente pour cette étude.

2.1.3 ORB

ORB (Oriented FAST and Rotated BRIEF) [6] est un descripteur local rapide et robuste, conçu comme une alternative libre de droits à SIFT [4] et SURF [2]. Il repose sur le détecteur de points-clés FAST pour la localisation et sur le descripteur binaire BRIEF pour la description, augmenté d'une composante d'orientation afin de garantir l'invariance à la rotation. La comparaison de descripteurs s'effectue via la distance de Hamming, ce qui le rend particulièrement efficace en termes de temps de calcul. Dans le cadre de la détection de quasi-doublons, ORB permet d'établir des correspondances locales entre paires d'images, même en présence de transformations géométriques modérées telles que la rotation ou le recadrage partiel. Toutefois, sa robustesse reste limitée face aux changements d'échelle importants et aux variations d'illumination prononcées. Son principal avantage réside dans son faible coût computationnel, le rendant adapté à un traitement à grande échelle sur des jeux de données de taille importante.

2.1.4 DELF

DELF (DEep Local Features) [5] est une méthode d'extraction de descripteurs locaux basée sur un CNN entraîné sur de grandes collections d'images. Contrairement aux descripteurs classiques tels qu'ORB ou SIFT, DELF apprend des représentations locales directement à partir des données, ce qui lui confère une bien meilleure robustesse aux changements de point de vue, d'échelle et de conditions d'illumination. Un mécanisme d'attention permet de sélectionner automatiquement les points-clés les plus discriminants pour la tâche de mise en correspondance. DELF a initialement été développé pour la reconnaissance visuelle de lieux et s'est imposé comme une méthode de référence dans ce domaine. Son application à la détection de quasi-doublons est naturelle : les correspondances denses et précises qu'il établit entre images permettent de détecter des similarités même sous des transformations complexes. Son principal inconvénient est son coût computationnel élevé, tant pour l'extraction des descripteurs que pour la mise en correspondance.

2.2 Jeu de données

Nous disposons d'un jeu de données fourni par la Fondation Science4Reefs. Ce dernier est composé de 5 967 photographies sous-marines de coraux de Moorea, en Polynésie française. Parmi celles-ci, 633 images réparties en 159 groupes de quasi-doublons ont été identifiés manuellement (cf. Figure 1). Cette quantité d'images reste limitée pour l'entraînement de modèles d'apprentissage profond. En pratique, les architectures utilisées en vision par ordinateur atteignent une meilleure stabilité et une capacité de généralisation satisfaisante à partir de plusieurs dizaines de milliers d'images, voire de centaines de milliers lorsque les variations visuelles sont importantes, comme c'est le cas en milieu sous-marin. Ainsi, ce jeu de données est utilisable dans cette étude comparative, mais ne permet pas d'exploiter pleinement les modèles étudiés sans risque de surapprentissage.

FIGURE 1 – Un exemple de quasi-doublons dans le jeu de données fourni par la fondation Science4Reefs

2.3 Métriques d'évaluation

Les méthodes de détection de quasi-doublons produisent généralement une liste d'images du jeu de données, classées par similarité décroissante avec l'image requête. Pour évaluer ces classements, deux métriques complémentaires sont couramment utilisées dans la littérature : le Recall@K et la mean Average Precision (mAP). Le Recall@K évalue la capacité du système à retrouver les bons résultats, tandis que la mAP mesure la qualité globale du classement. Ainsi, elles permettent une évaluation robuste des méthodes proposées.

Le **Recall@K** mesure la proportion de requêtes pour lesquelles au moins un near-duplicate apparaît parmi les K images les plus similaires retournées par le modèle. Il se calcule comme suit :

$$\text{Recall@K} = \frac{N_{\text{hit}}}{N_q} \tag{1}$$

où N_{hit} est le nombre de requêtes ayant au moins un near-duplicate dans les K images les plus similaires selectionnées par le modèle pour une requête donnée, et N_q le nombre total de requêtes.

La **mAP** évalue à la fois la pertinence et l'ordonnancement des résultats retournés par le système. Une mAP élevée indique que les images les plus similaires à la requête sont bien positionnées en tête du classement. Pour chaque requête, on calcule d'abord l'Average Precision (AP), qui mesure la qualité du classement des éléments pertinents dans la liste :

$$\text{AP} = \frac{1}{N} \sum_{i=1}^{k} P(i) \cdot \text{rel}(i) \tag{2}$$

où N est le nombre total d'éléments pertinents, $P(i)$ est la précision au rang i, et $\text{rel}(i)$ vaut 1 si l'élément au rang i est pertinent, 0 sinon. Cette valeur est ensuite moyennée sur l'ensemble des requêtes :

$$\text{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP}_q \tag{3}$$

où Q est le nombre total de requêtes.

De plus, deux métriques ont été considérées pour mieux caractériser le comportement du système lors de l'implémentation du clustering : le Recall et la Pureté.

Le **Recall** reflète la capacité du système à récupérer l'ensemble des quasi-doublons. Une valeur élevée indique que le système minimise les faux négatifs. Il se calcule ainsi :

$$\text{Recall} = \frac{N_{\text{correct}}}{N_{\text{ND}}} \tag{4}$$

où N_{correct} est le nombre de quasi-doublons correctement identifiés, et N_{ND} le nombre total de quasi-doublons dans le jeu de données.

La **Pureté** évalue la qualité des résultats de clustering en mesurant l'homogénéité des groupes générés. Un cluster est considéré pur si toutes ses images appartiennent au même groupe de near-duplicate tel que défini par les annotations manuelles. Cette métrique est utilisée à titre indicatif uniquement, les annotations manuelles n'étant pas considérées comme fiables. Elle se calcule comme suit :

$$\text{Pureté} = \frac{N_{\text{pur}}}{N_{\text{cluster}}} \tag{5}$$

où N_{pur} est le nombre d'images appartenant à des clusters purs, et N_{cluster} le nombre total d'images appartenant à des clusters de taille supérieure à 1.

2.4 Méthodologie

La détection de quasi-doublons est décomposée en cinq étapes séquentielles : l'extraction de descripteurs des images, le calcul d'une matrice de similarité entre paires d'images, la sélection d'un seuil de similarité, et une phase de clustering (cf. Figure 2). Les descripteurs permettent de comparer les images deux à deux via la matrice de similarité, le seuil transforme ces scores continus en décisions binaires similaire/non similaire, et le clustering agrège ces relations locales en groupes cohérents. Chaque étape conditionne directement la suivante, ce qui justifie une évaluation progressive et indépendante de chacune d'elles.

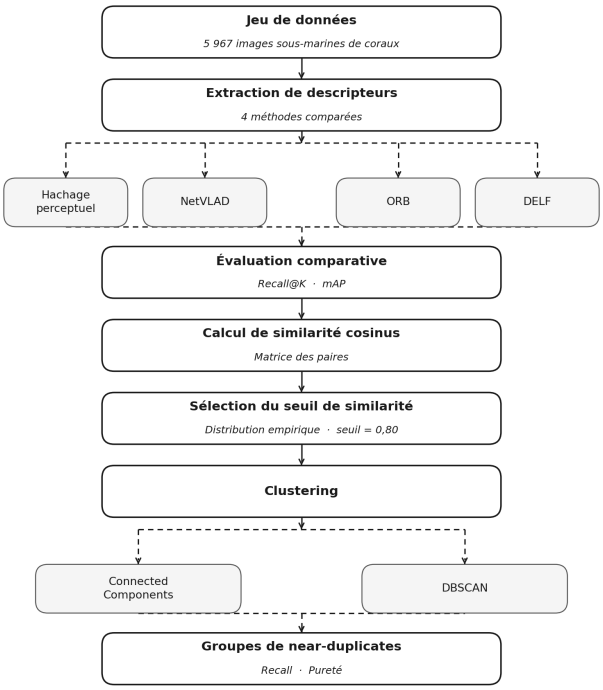

FIGURE 2 – Organigramme résumant étapes principales de la démarche.

3 Étude comparative

Cette section présente les résultats obtenus pour chacune des méthodes d'extraction de caractéristiques, en évaluant leur performance sur le jeu de données d'images de coraux. À partir des descripteurs produits par la méthode ayant obtenu les meilleures résultats, NetVLAD, une chaîne de traitement complète de détection de quasi-doublons est ensuite mise en place. Elle repose sur le calcul d'une matrice de similarité et le regroupement des images via un algorithme de clustering. Deux approches sont comparées (Connected Components et DBSCAN) en évaluant leur performance pour différents seuils de similarité.

3.1 Extractions de caractéristiques

3.1.1 Hachage perceptuel

Plusieurs méthodes de hachage perceptuel issues de la bibliothèque Python ImageHash [3] ont été étudiées. Cette bibliothèque propose différentes implémentations de hachage d'images, conçues pour mesurer la similarité visuelle tout en restant peu coûteuses en calcul. Une série d'expériences comparatives a été menée afin d'évaluer les performances de ces techniques. Les méthodes testées incluent aHash, dHash, pHash, wHash, colorHash et un crop-resistant hash. Cette phase expérimentale a permis de constater que le pHash, bien que largement utilisé dans la littérature, présentait une performance inférieure sur ce cas d'étude par rapport à d'autres méthodes telles que wHash ou aHash. Les performances ont été mesurées à l'aide de la métrique Recall@K, et les résultats obtenus sont résumés dans le tableau 1 :

	R@1	R@5	R@10
aHash	0.42	0.53	0.60
wHash	0.40	0.53	0.60
dHash	0.25	0.36	0.40
colorHash	0.24	0.46	0.56
crop resistant hash	0.21	0.29	0.34
pHash	0.18	0.28	0.34

TABLE 1 – Comparaison des méthodes de hachage perceptuel.

L'analyse qualitative des différentes fonctions met en évidence l'influence du type de représentation visuelle utilisé par les méthodes de hachage. Les approches basées sur l'apparence globale de l'image, comme le aHash, obtiennent de bons résultats car les quasi-doublons du jeu de données présentent des structures très similaires et peu de transformations complexes.
À l'inverse, le pHash, qui repose sur une représentation fréquentielle plus abstraite, perd une partie de l'information discriminante lorsque les différences entre images reposent sur des variations locales ou de légers changements de point de vue. Les méthodes sensibles aux détails locaux, comme le wHash et le dHash, obtiennent des performances intermédiaires.
Enfin, les méthodes basées uniquement sur la couleur ou conçues pour être très robustes aux transformations fortes montrent des résultats plus limités. Le colorHash peut rapprocher des images aux palettes similaires mais de contenu différent, tandis que les approches tolérant fortement les recadrages perdent une partie de leur pouvoir discriminant.
Dans l'ensemble, ces résultats suggèrent que, pour ce jeu de données, des représentations simples capturant directement les similarités visuelles globales ou locales sont plus adaptées que des méthodes plus abstraites ou fortement robustes aux transformations.

Hachage hybride couleur–structure. Afin de dépasser les limites des approches basées sur un seul type de hachage perceptuel, une stratégie de hachage hybride a été explorée. Elle consiste à combiner deux méthodes de hachage complémentaires à l'aide d'une distance pondérée :

$$d = \alpha \, | \, Hash_1^{\text{query}} - Hash_1^{\text{image}} \, | \\ + \beta \, | \, Hash_2^{\text{query}} - Hash_2^{\text{image}} \, | \qquad (6)$$

Les meilleures performances ont été obtenues avec la combinaison colorHash + aHash. Les résultats montrent qu'une pondération majoritaire du colorHash (80 %) améliore significativement les performances, avec un gain supérieur à 10 % sur la métrique Recall@10 par rapport aux méthodes individuelles.
Cette amélioration s'explique par la complémentarité des informations capturées. Le colorHash agit comme un descripteur global permettant de regrouper rapidement les images partageant une palette chromatique similaire, ce qui est particulièrement pertinent pour les images de coraux qui possèdent souvent des signatures de couleur caractéristiques. Toutefois, cette information seule reste insuffisante

aHash (α)	colorHash (β)	R@1	R@5	R@10
0,2	0,8	0,48	0,66	0,73
0,4	0,6	0,47	0,64	0,70
0,6	0,4	0,44	0,59	0,66
0,8	0,2	0,39	0,55	0,62

TABLE 2 – Performance du hachage hybride combinant aHash et colorHash selon différentes pondérations.

car des structures différentes peuvent présenter des couleurs proches. L'intégration du aHash introduit alors une contrainte géométrique simple basée sur la structure globale de l'image, qui agit comme un filtre permettant de réduire les faux positifs générés par la similarité de couleur seule. Ainsi, le colorHash assure une sélection large basée sur l'apparence chromatique, tandis que le aHash joue le rôle de filtre structurel permettant de vérifier la cohérence géométrique des images candidates. Cette complémentarité explique les gains observés et souligne l'intérêt d'associer des méthodes de hachage perceptuel capturant des propriétés visuelles différentes.

3.1.2 NetVLAD

La méthode NetVLAD permet d'agréger des descripteurs locaux d'images en un vecteur global compact. Après l'implémentation de la méthode, un fine-tuning a permis de l'adapter aux similarités fines entre images de coraux proches, tout en conservant les représentations générales apprises sur de larges jeux de données de référence. Le jeu de données annoté étant limité et fortement déséquilibré, l'entraînement complet du modèle n'était pas envisageable.

Pour le fine-tuning, une validation croisée à cinq folds, organisée par groupes de quasi-doublons, a été mise en place pour garantir une évaluation plus fiable et reflètant la capacité du modèle à généraliser sur de nouveaux coraux. Les différents hyperparamètres du modèle ont ensuite été ajustés afin de déterminer une configuration optimale. Le backbone choisi pour l'extraction des descripteurs est le CNN ResNet-50, préentraîné sur le jeu de données ImageNet, dont les couches intermédiaires ont été dégelées, permettant un apprentissage approfondi. Deux fonctions de perte ont été comparées : la MultiSimilarityLoss et la HardTripletLoss. Le choix s'est porté sur la MultySimilarityLoss qui fournit des gradients progressifs et favorise une convergence stable en début d'entraînement. La pondération des paires difficiles a permis de rapprocher les quasi-doublons et d'éloigner les images non similaires. Enfin, le taux d'apprentissage a été optimisé grâce à un ordonnanceur et à la mise en place de taux d'apprentissage distincts pour le CNN et pour la couche NetVLAD.

Une fois le fine-tuning réalisé sur les cinq folds de dix epochs, le modèle final a été sélectionné par une évaluation de type Leave-One-Out qui consiste à comparer chaque image requête, représentative d'un groupe de quasi-doublons, à toutes les images du même fold, à l'exception de celles de son propre groupe. Cette approche permet d'évaluer la qualité des descripteurs en évitant les biais liés à une évaluation sur l'ensemble complet des 5 967 images du jeu de données. Cependant, une telle évaluation a permis d'estimer les performances réelles du modèle en situation pratique. Pour cette évaluation, le jeu de données a été séparé en un jeu d'images requêtes et un jeu d'images candidates : le premier contient une image représentative par individu ayant des quasi-doublons, tandis que le second contient toutes les autres images. Cette configuration simule un cas réel de recherche de quasi-doublons et permet de mesurer les performances finales du modèle en termes de mAP et de Recall@k, assurant la généralisation du modèle NetVLAD. Le modèle final retenu correspond à la version du fold 4, epoch 10, ayant fourni, sur l'ensemble du jeu de données, la meilleure mAP (82,66%) et le troisième meilleur Recall@1 (86,79%), ainsi que des Recall@10, et plus, dépassant les 98% de récupération (cf. Figure 3).

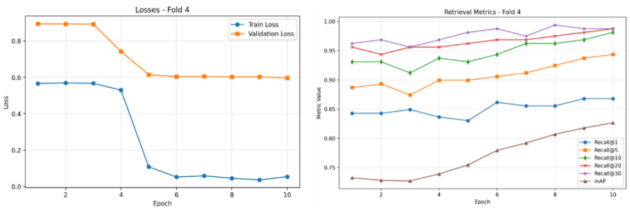

FIGURE 3 – Évolution de la perte en entraînement et en validation (gauche) et des métriques Recall@k et mAP (droite) pour le fold 4 en fonction des epochs

3.1.3 ORB

L'algorithme ORB est particulièrement prisé pour son excellente rapidité d'exécution et son absence de brevets, le rendant idéal pour les applications temps réel. ORB détecte principalement des coins et motifs structurés, et des zones à fort contraste. Sur les images de coraux, les textures sont souvent irrégulières et organiques, avec peu de coins nets, ce qui limite le nombre de point clé pertinents. Cela explique les performances limitées du modèle résumées dans le tableau 3.

R@1	R@5	R@10	R@20	R@30	mAP
0,19	0,38	0,43	0,49	0,55	0,12

TABLE 3 – Performance de la méthode ORB.

3.1.4 DELF

Afin de spécifier l'extraction des données au contexte spécifique des coraux, la piste d'un fine-tuning du modèle DELF a été identifiée comme une évolution pertinente. Cette approche consisterait à spécialiser le CNN pré-entraîné sur des images de récifs coralliens afin d'orienter les mécanismes d'attention vers les structures biologiques d'intérêt. Le fine-tuning de DELF s'est donc appuyé sur le dépôt DeLF-pytorch [7], qui a été cloné et adapté au contexte du projet. Ce dépôt fournit une implémentation modulaire de DELF ce qui permet le fine-tuning du modèle et l'extraction de descripteurs locaux. Deux configurations de fine-tuning ont donc été évaluées : le fine-tuning du backbone uniquement, ainsi que le fine-tuning conjoint du backbone et du

module de points-clés. Les hyperparamètres ont été ajustés entre les différents entraînements réalisés, et les résultats principaux sont résumés dans le tableau 4.

Méthode	R@1	R@5	R@10	mAP
Zéro-shot	0,73	0,79	0,82	0,25
Backbone, LR=1e-4	0,53	0,64	0,68	0,36
Backbone, LR=5e-5	0,30	0,44	0,48	0,27
Backbone, LR=1e-5	0,43	0,53	0,57	0,36
Full, LR=1e-4	0,79	0,87	0,90	0,59
Full, LR=5e-5	0,80	0,87	0,90	0,60
Full, LR=1e-5	0,81	0,88	0,90	0,61

TABLE 4 – Résultats du fine-tuning de DELF selon les méthodes et les paramètres.

Dans le cas du fine-tuning du backbone uniquement, les performances chutent par rapport au modèle zero-shot, quel que soit le taux d'apprentissage. Cela montre que l'adaptation des poids du CNN seule perturbe les représentations visuelles apprises lors du premier entraînement, sans permettre une réorganisation cohérente des points-clés. Le module d'attention, qui reste figé, continue de sélectionner des régions selon des critères inadaptés aux nouvelles caractéristiques extraites par le backbone. Cela crée un désalignement qui conduit à des descripteurs moins discriminants, ce qui explique la baisse des métriques. À l'inverse, le fine-tuning conjoint du backbone et du module d'attention conduit à des gains significatifs par rapport à la version zéro-shot. Cela s'explique par la capacité du modèle à adapter en même temps l'extraction des caractéristiques locales et le mécanisme d'attention qui sélectionne les régions pertinentes. Le module d'attention apprend ainsi à se concentrer sur des zones plus distinctives du corail comme les textures, plutôt que sur des éléments parasites ou trop génériques.

3.2 Regroupement des images

La méthode NetVLAD ayant fourni les meilleurs résultats, les descripteurs qu'elle a généré sont utilisés lors des étapes suivantes.

Calcul de similarité. Après l'extraction des descripteurs de chaque image du jeu de données, on calcule la similarité entre chaque paire de descripteurs pour déterminer si deux images sont des quasi-doublons. Cette comparaison repose sur la similarité cosinus afin de leur attribuer un score de similarité. Elle se calcule ainsi :

$$\text{Similarité cosinus}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} \qquad (7)$$

Deux images sont des quasi-doublons quand leur score de similarité dépasse un certain seuil. Une méthodologie s'appuyant sur les scores de similarité va alors permettre de le déterminer. Tout d'abord, les similarités entre toutes les paires d'images sont calculées afin de constituer une matrice de similarité. La mise en parallèle du jeu de données manuellement annoté avec les distributions de similarité (cf. Figure 4) révèle un score moyen de 0,7 pour les quasi-doublons et de 0,27 pour les singletons, ce qui confirme que le modèle distingue nettement les deux groupes. Cependant, un chevauchement subsiste entre 0,4 et 0,7. Plusieurs seuils ont donc été testés dans le cadre du clustering afin de déterminer un seuil satisfaisant.

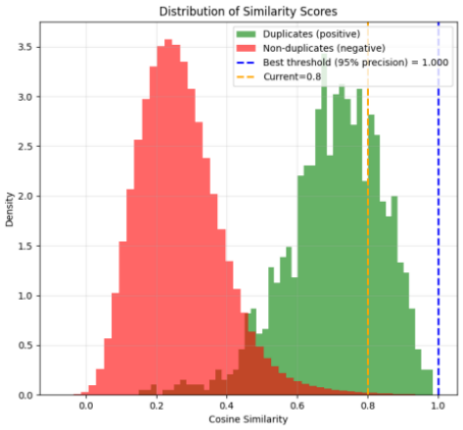

FIGURE 4 – Distribution des scores de similarité par paire d'images selon leur statut dans le jeu de données annoté : near-duplicate (vert) ou singleton (rouge)

Clustering. La phase de clustering est essentielle pour passer de la comparaison de similarité entre deux images à la détection de groupes de quasi-doublons, c'est-à-dire des clusters. Deux méthodes principales ont été étudiées : la méthode de chaînage transitif Connected Components (composantes connexes) et l'algorithme de partitionnement Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Connected Components forme un graphe de similarité non orienté, où chaque nœud représente une image et une arête relie deux images si leur score de similarité dépasse le seuil fixé. La méthode DBSCAN, quant à elle, regroupe automatiquement les images à partir de leur score de similarité en prenant en compte le seuil fixé, et un nombre d'image minimal pour former un groupe de quasi-doublons. Par expérimentation, la méthode retenue qui fournit de meilleurs résultats est Connected Components (cf. Figure 5).

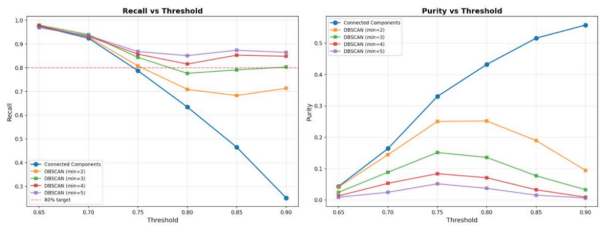

FIGURE 5 – Comparaison des approches Connected Components (bleu) et DBSCAN selon le Recall et la Pureté en fonction du seuil de similarité

Le choix du seuil a été guidé par un compromis : un seuil trop faible (0,75) génère des méga-clusters, tandis qu'un seuil trop strict (0,85) exclut de nombreux quasi-doublons.

Le seuil de 0,80 a été retenu, offrant un équilibre entre Recall (63%) et Pureté (43%). Cette valeur a été choisie de manière conservatrice afin de privilégier la précision et limiter les faux positifs.

4 Discussion

Cette étude s'est articulée autour d'une analyse exploratoire de publications de détection de quasi-doublons et de reconnaissance visuelle de lieux. Elle a permis d'identifier quatre approches pertinentes adaptées au milieu sous-marin et aux motifs complexes des coraux.

Les expérimentations, menées sur le jeu de données fourni par la Fondation Science4Reefs, montrent que la méthode ORB est peu adaptée aux besoins de cette étude, tandis que le hachage perceptuel et DELF offrent des performances satisfaisantes, indiquant un potentiel d'adaptation intéressant. Le gain significatif obtenu par le fine-tuning complet de DELF souligne que l'adaptation au contexte spécifique des images est essentielle pour la détection de quasi-doublons, qui s'appuie sur des motifs fins caractéristiques. NetVLAD, quant à elle, a surpassé les autres approches, atteignant un Recall@1 de 86,79% et une mAP de 82,66%. Ces résultats démontrent que, correctement fine-tunée, la méthode NetVLAD constitue un outil efficace pour détecter des quasi-doublons dans des images de coraux. La phase de clustering repose sur la méthode Connected Components et un compromis conservateur entre récupération et précision, pouvant conduire à la formation de méga-clusters. Étant donné que le clustering minimise les faux positifs, l'intégration d'une validation humaine peut s'avérer pertinente pour garantir la qualité des clusters.

La chaîne de traitement développée permet un nettoyage fiable du jeu de données, réduisant les besoins en stockage et en calcul, tout en ouvrant d'autres perspectives d'utilisation, telles que la détection d'images d'un même individu de corail, voire d'espèces spécifiques. Les pistes de poursuite de cette étude concernent notamment l'expérimentation de nouvelles méthodes d'extraction de caractéristiques d'images et de techniques de clustering afin d'améliorer les performances. Par ailleurs, le jeu de données fourni reste limité en nombre d'images, ce qui restreint le fine-tuning des méthodes de deep learning et leur capacité à apprendre des caractéristiques spécifiques aux coraux. Cette contrainte a guidé des choix méthodologiques pragmatiques, mais souligne l'importance de poursuivre ces travaux sur des jeux de données plus vastes et diversifiés afin d'augmenter la généralisation des modèles et leur robustesse en contexte réel. En effet, la méthodologie proposée, bien qu'appliquée ici à des images de coraux de Moorea, présente un potentiel de généralisation à des images de coraux issues d'autres régions géographiques, et à d'autres domaines caractérisés par des acquisitions séquentielles (faune marine, biodiversité terrestre). Cette généralisation est conditionnée à une adaptation au domaine cible, notamment concernant le paramétrage des modèles, selon la finesse des motifs à détecter.

5 Conclusion

L'objectif de cette étude est de comparer différentes méthodes de détection de quasi-doublons, appliquées à un jeu de données d'images de coraux, caractérisé par des motifs fins et les variations naturelles de l'environnement aquatique. Cette étude a permis d'adapter, d'évaluer et de comparer quatre approches d'extraction de caractéristiques d'images : le hachage perceptuel, les descripteurs globaux (NetVLAD), et les descripteurs locaux reposant sur la détection de points-clés (ORB et DELF). Elle a également permis de développer une méthode de clustering, complétant alors la chaîne de traitement de détection de quasi-doublons. Les travaux réalisés apportent ainsi une étude comparative détaillée de différentes méthodes de détection de quasi-doublons, adaptées au domaine corallien sous-marin, encore peu documenté en vision par ordinateur. Ainsi, ces travaux ouvrent la voie à des applications de suivi et de préservation de récifs coralliens.

Remerciements

Nous tenons, tout d'abord, à remercier Marie-Hélène Abel, professeure à l'UTC et membre de l'UMR CNRS Heudiasyc, pour sa grande disponibilité et la qualité de ses conseils. Nous tenons également à remercier Arjen Kraneveld, référent technique, et Lionel Mathieu, référent gestion de projet, au sein de la cellule d'innovation Lab'Innov de l'entreprise Aubay, pour leur accompagnement tout au long du développement de ces travaux. Enfin, nous tenons à remercier Laetitia Hédouin, directice de recherche au CNRS et présidente de la Fondation Science4Reefs, pour la qualité de nos échanges et pour son engagement dans ce projet dédié à la préservation des récifs coralliens.

Références

[1] Relja Arandjelovic , Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad : Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF : Speeded up robust features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.

[3] Johannes Buchner. ImageHash : A python perceptual image hashing module, 2013.

[4] David G. Lowe. Distinctive image features from scale-invariant keypoints. volume 60, pages 91–110. Springer, 2004.

[5] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3456–3465. IEEE, 2017.

[6] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB : An efficient alternative to SIFT

or SURF. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011.

[7] Minchul Shin. DeLF-pytorch : Deep local features in PyTorch, 2021. Dépôt GitHub, consulté en 2025.

[8] Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. In *Media Watermarking, Security, and Forensics III*, volume 7880. SPIE, 2010.