Ré-identification de chiens à partir de vidéos en environnement non-contrôlé

Cyril BARRELET[1] Eugenio DIAS RIBEIRO NETO[1] Marc CHAUMONT[1,3] Gérard SUBSOL[1] Etienne LOIRE[2] Michel DE GARINE-WICHATITSKY[2]

[1]Equipe ICAR, LIRMM, Univ Montpellier, CNRS, Montpellier, France

[2]UMR-ASTRE, CIRAD, INRAE, Univ Montpellier, Montpellier, France

[3]Univ Nîmes, France

Résumé – Cet article aborde le problème de la ré-identification (ReID) de chiens à partir de pièges photographiques qui fournissent des vidéos de mauvaise qualité et des petits ensembles de données. Notre objectif est de définir un extracteur de caractéristiques (EC) robuste à ces conditions difficiles. Nous revisitons l'optimisation de la fonction de coût par triplet par la définition d'exemples difficiles lors de trois apprentissages consécutifs. Nous avons mené des expériences sur deux ensembles de données que nous avons rendus publics, l'ensemble de données YT-BB-Dog et SEAdogSEA. Les résultats de précision top-k dans le scénario de ReID sur l'ensemble de données YT-BB-Dog étiqueté sont bons, compte tenu des conditions difficiles. Nous proposons une chaîne de traitement complète utilisant une détection et un suivi standard et une méthode de ReID fondée sur notre EC robuste. Pour finir, nous donnons une évaluation qualitative sur l'ensemble de données SEAdogSEA.

Abstract – This paper addresses the problem of re-identifying (ReID) dogs from camera traps which give poor-quality videos, and small datasets. We aim to define a feature extractor that is robust to these challenging conditions. We thus revisit the triplet-loss function optimization by defining difficult examples in three consecutive training sessions. We conducted experiments on two publicly available datasets, the YT-BB-Dog dataset, and the SEAdogSEA dataset. Considering the challenging conditions, the top-k accuracy results in the ReID scenario on the labeled YT-BB-Dog dataset are good. We propose a complete processing chain using out-of-the-box detection and tracking as well as a ReID method based on our robust featrue extractor. Finally, we give a qualitative evaluation of the unlabelled SEAdogSEA dataset.

1 Introduction

Cet article traite de la détection d'objets dans des vidéos et plus précisément de leur réidentification (ReID). Par ReID, nous entendons qu'étant donné une imagette ou une séquence d'imagettes d'un objet, l'algorithme doit retrouver les k imagettes ou séquences d'imagettes les plus similaires, classées par ordre de similarité décroissante (le top-k). L'objectif est d'être capable de retrouver la même instance de l'objet et non pas juste un objet de la même classe.

L'algorithme fondamental de la ReID est celui qui donne une signature visuelle compacte (« hachage visuel ») de l'imagette. Ce problème a été largement étudié pour la ReID de personne avec l'émergence de l'apprentissage profond [14]. Globalement, la plupart des approches reposent sur des réseaux neuronaux et l'utilisation d'une fonction de coût par triplet [13].

La ReID d'animaux reste cependant peu étudiée, induisant un retard dans les performances. Cela s'explique par la petite taille des bases de données annotées disponibles (ne contenant souvent que 50 à 100 individus), mais aussi par la variabilité de la forme des animaux qui est bien supérieure à celle des humains ou la diversité des angles de prises de vue [8].

Il y a très peu de travaux pour la ReID des chiens. Moreira *et coll.* [6] ont travaillé avec des imagettes de têtes de chien de très bonne résolution (taille d'image de 200×200 et 250×250), avec une acquisition de face et des yeux alignés et deux ensembles de données très limités de seulement 18 et 21 individus. DogFaceNet [7] se concentre également sur le visage du chien avec des imagettes de bonne qualité acquises de face et seulement 48 individus. Enfin, Jang *et coll.* [1] se concentrent uniquement sur la truffe du chien pour l'identifier.

En conclusion, les articles relatifs à la ReID de chiens sont peu nombreux, ne traitent pas du problème général dans des conditions naturelles avec des vidéo de qualité limitée et n'utilisent que de petits ensembles de données. Dans cet article, nous nous concentrons sur la ReID de chiens dans des vidéos acquises par une configuration multi-caméras de qualité très standard, sans point de vue prédéfini, et sans aucune information sur l'identité ou les caractéristiques des chiens observés.

Étant donné que les vidéos sont acquises par diverses caméras à différents endroits sur une période donnée, le problème peut être divisé en deux parties. Tout d'abord, il consiste à localiser le(s) chien(s) dans chaque image et à le(s) suivre dans la séquence. Nous obtenons ainsi une Séquence d'Imagettes du Même Chien (SIMC) pour chaque chien suivi. Ensuite, il s'agit de trouver pour chaque SIMC si le même chien apparaît sous la forme d'une SIMC dans une autres vidéo. Cette deuxième partie est le problème à proprement parler de la ReID et permet de regrouper un ensemble de SIMC appartenant au même individu-chien.

La détection n'est pas un problème nouveau, et de nombreuses propositions existent pour localiser des objets spécifiques et les suivre au cours du temps. Dans cet article, nous utiliserons une approche récente et prête à l'emploi. En revanche, nous développerons la ReID des chiens en considérant

l'animal dans son ensemble. C'est un aspect essentiel car avec des conditions d'acquisition diverses et de mauvaise qualité, il est impossible d'avoir des informations discriminantes en utilisant seulement le visage.

Dans cet article, nous proposons a. La création de la base de données YT-BB-Dog, dérivé de l'ensemble de données YouTube Bounding Box [9] contenant 2 723 SIMC, chacune correspondant à un chien différent ; b. La création de la base de données SEAdogSEA[1][2] qui contient 227 SIMC de chiens errants acquises dans deux villes asiatiques par plusieurs pièges photographiques ; c. Un extracteur de caractéristiques visuelles qui utilise l'image complète de l'animal et pas seulement son visage, tout en ayant de bonnes propriétés discriminantes entre individus-chiens différents pour la ReID ; d. L'utilisation de séquences de vecteurs caractéristiques pour calculer des distances lors de la ReID ; et finalement, e. Une chaîne complète de traitement pour la ReID de plusieurs individus-chiens dans plusieurs vidéos.
YT-BB-Dog et SEAdogSEA ont été rendues publics et peuvent être téléchargées[3]

2 Apprendre un extracteur de caractéristiques efficace et générique

2.1 ReID et fonction de coût par triplet

La plupart des approches modernes reposent sur l'utilisation d'une fonction de coût par triplet introduit dans [11]. Elle permet de construire un extracteur de caractéristiques (EC) qui différencie les individus en augmentant la distance entre deux vecteurs lorsque les individus correspondants sont différents. L'apprentissage consiste à raisonner sur 3 vecteurs, chacun étant issu de l'EC appliqué à une imagette. Ces 3 imagettes sont appelées l'ancre, qui est l'imagette montrant un chien, le positif qui est une autre imagette montrant le même chien, et le négatif qui est une imagette montrant un autre individu-chien. La fonction objectif qui doit être minimisée est la somme de tous les coûts par triplet définis par :

$$\mathcal{L}_{\mathbf{a},\mathbf{p},\mathbf{n}} = \max(0, dist(\mathbf{a},\mathbf{p}) - dist(\mathbf{a},\mathbf{n}) + k) \qquad (1)$$

avec $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{p} \in \mathbb{R}^d$, et $\mathbf{n} \in \mathbb{R}^d$, les 3 vecteurs de caractéristiques correspondant aux imagettes ancre, positif et négatif, d étant la dimension des vecteurs, k un scalaire définissant la marge, $dist$ une fonction de distance telle que L2, et max la fonction maximum. En minimisant $\mathcal{L}_{\mathbf{a},\mathbf{p},\mathbf{n}}$, nous forçons les positifs à être proches des ancres et les négatifs à en être plus éloignés.

La sélection de ces triplets joue donc un rôle majeur pour la ReID. L'ancre et le positif étant généralement très proches, nous nous concentrerons sur le choix de l'exemple négatif, dont la difficulté influencera directement l'acuité de l'EC.

2.2 De YouTube Bounding Box à YT-BB-Dog

Le jeu de données *Youtube-BoundingBoxes* (YT-BB) [9] contient 380 000 séquences vidéo de 15 à 20 secondes extraites de 240 000 vidéos publiques de YouTube qui sont annotées suivant 23 classes qui constituent un sous-ensemble des classes du jeu de données COCO [3]. YT-BB est fourni avec les boîtes englobantes des différents objets présents ce qui permet de construire les imagettes.

À partir de ce jeu de données, nous avons automatiquement extrait les SIMC en utilisant la classe Dog et supprimé celles qui comportaient moins de 5 imagettes ou dont la résolution était inférieure à 50×50. Nous obtenons 9 411 SIMC (correspondant donc à des chiens différents), pour un total de 89 772 imagettes.

Afin d'affiner un peu plus la base de données et d'accroître les performances de l'EC, nous avons supprimé les SIMC dont les imagettes comprenaient moins de 10 points clés détectés par l'estimateur de postures MMPose [5]. Cet ultime filtrage mène au jeu de données YT-BB-Dog qui contient 2 723 SIMC, tout en proposant une grande variété d'images, de types de caméras, de conditions d'acquisition et de races de chiens.

2.3 Entraînement en 2 étapes pour une bonne gestion des exemples difficiles

L'EC doit renvoyer un vecteur très similaire pour différentes vignettes d'un même chien et un vecteur très distant quant il s'agit de chiens différents (même s'ils sont ressemblants). Nous proposons de l'entraîner avec une fonction de coût par triplet en 2 étapes comme le montre la figure 1.

Tout d'abord, nous utilisons comme EC initial $f^{(0)}$ un réseau ConvNext-Small [4] pré-appris sur ImageNet-1K et amputé de ses couches de classification. Appliqué à une imagette, $f^{(0)}$ renvoie un vecteur de dimension $d = 768$ qui n'est pas trop grand et qui est reconnu pour sa bonne représentativité.

Pour choisir les triplets nécessaires à l'entraînement de $f^{(0)}$, nous nous appuyons sur un réseau "auxiliaire" (représenté en gris dans la figure 1) EfficientNet entraîné à reconnaître les races de chiens sur la base de données Stanford Dogs Dataset [2]. Pour une ancre donnée, ce réseau nous permet de sélectionner un chien de même race (ou très similaire) comme exemple négatif. Nous obtenons alors un nouvel EC, $f^{(1)}$.

Cependant, nous avons remarqué que $f^{(1)}$ se focalisait principalement sur l'arrière plan. Il serait donc intéressant d'étudier la variabilité de ce dernier dans Stanford Dogs Dataset. A l'instar du premier réseau "auxiliaire", nous avons utilisé $f^{(1)}$ pour sélectionner des exemples difficiles. Cette fois-ci, pour une ancre donnée, l'imagette négative a tendance à présenter un chien de même race avec un arrière plan similaire, comme le montre les exemples de droite de la figure 1.

Finalement, $f^{(1)}$ est entraîné avec ce nouveau choix de triplets donnant $f^{(2)}$, qui obtient de bonnes propriétés de ReID.

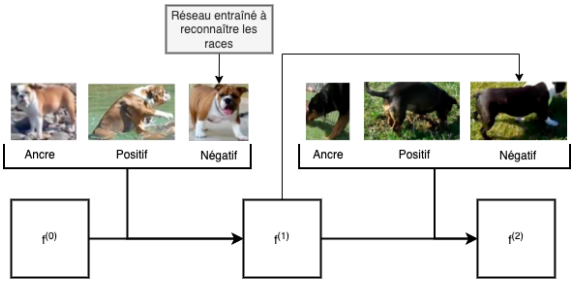

FIGURE 1 : Chaîne d'entraînement en 2 étapes.

[1] https://anr.fr/Projet-ANR-19-ASIE-0002

[2] Avec des remerciements à nos collègues thaïlandais Barandi SAPTA WIDARTONO, Najib ARUNG PETANA, Wayan T. ARTAMA, et Sopheak SORN

[3] https://www.lirmm.fr/~chaumont/YT-BB-Dog_SEAdogSEA.html

3 Expériences, résultats et discussion

3.1 Protocole expérimental

Les deux entraînements du réseau ConvNext-Small sur l'ensemble de données YT-BB-Dog ont été réalisés en utilisant l'optimiseur AdamW, le même que celui utilisé pour le pré-entrainement. Nous avons divisé l'ensemble de données YT-BB-Dog en 2 000 chiens pour l'entraînement et 723 chiens pour les tests, et nous avons effectué l'apprentissage pendant 40 époques.

Nous utilisons la précision top-k pour évaluer l'EC. Étant donné une imagette de requête d'un chien, nous marquons un score de un si au moins une autre imagette de ce chien est présente dans les k imagettes les plus proches (en fonction d'une distance entre vecteurs prédéfinie). Cette vérification est effectuée pour un ensemble de N requêtes, et le score cumulatif définit la précision top-k :

$$\frac{\sum_i^N [S^{(i)} \cap S_k^{(i)} \neq \emptyset]}{N} \quad (2)$$

avec [] la notation d'Iverson, i étant l'indice de l'imagette de requête, $S^{(i)}$ étant l'ensemble des indices (excepté i) des imagettes montrant le même chien que l'imagette de requête et $S_k^{(i)}$ étant l'ensemble d'indices des k imagettes les plus proches de l'imagette de requête.
Pour rappel, nous nous concentrons uniquement sur la partie de ReID étant donné notre approche en deux étapes : détection puis ReID.

3.2 Ré-identification sur le jeu de données YouTube Bounding Box filtré et limité aux chiens

Dans le tableau 3.2, nous rapportons les résultats obtenus dans un scénario dit "ensemble ouvert". Dans ce cas, l'EC est appris sur l'ensemble d'apprentissage YT-BB-Dog tel que décrit dans la section 2.3, et les précisions top-k sont calculées en utilisant la distance cosinus sur l'ensemble de test YT-BB-Dog avec d'une part les imagettes de 723 SIMC (soit 723 individus-chiens) et d'autre part un sous-ensemble de 100 SIMC (soit une recherche dans 100 individus-chiens possibles seulement).

	723 SIMC ≠ (ens. ouvert)	100 SIMC ≠ (ens. ouvert)
k = 1	43%	73 %
k = 5	58%	90 %
k = 200	89%	100%

TABLE 1 : Précision top-k sur l'ensemble de test YT-BB-Dog avec l'utilisation de la distance cosinus et l'EC appris sur l'ensemble d'apprentissage YT-BB-Dog en scénario ensemble ouvert

Nous avons obtenu une précision de 73% pour le top-1 sur l'ensemble de test de 100 SIMC. Il convient de noter qu'à notre connaissance, aucun article ne décrit l'évaluation en scénario ensemble ouvert pour la ReID de chiens. En comparaison, les auteurs de [10] rapportent une précision de top-1 de 86% pour la ReID de tigres dans le même scénario. Cependant, cette tâche est beaucoup plus simple compte tenu de la bonne résolution des images, des caractéristiques très distinguables chez les tigres et du nombre d'individus dans l'ensemble de test de seulement 18 individus différents.

Il est intéressant de noter que le top-5 donne un score de précision de 90% sur 100 SIMC, ce qui indique que l'EC donne une représentation pertinente des imagettes. Pour autant, les résultats obtenus sur les 723 différents individus ne sont pas aussi bons. Il est donc nécessaire de poursuivre l'effort de recherche pour rendre la méthode plus évolutive et utilisable lorsqu'il y a un grand nombre d'individus.

3.3 Evaluation qualitative sur SeaDogSea

A l'instar de YT-BB-Dog, SeaDogSea est composé de SIMC filtrés de la même façon. Au total, ce jeu de données non étiqueté comporte 228 SIMC pour 1 788 imagettes.

Afin de visualiser le pouvoir discriminant de notre EC pour l'application de surveillance des chiens errants par piège photographiques, nous avons utilisé un algorithme de clustering agglomératif fondé sur l'utilisation de la distance de Hausdorff pour définir une distance entre deux SIMC.

Au lieu de fixer le nombre de clusters, nous avons lancé plusieurs processus de clustering en faisant varier la distance seuil pour fusionner deux clusters de 0 à 0,5. Nous avons conservé le résultat de cluster dont le coefficient de silhouette est minimum. Nous avons trouvé 72 clusters pour un coefficient de silhouette de 0,32 et un seuil de 0,37. Trois exemples de clusters sont présentés dans la figure 2. On peut voir que le cluster A regroupe bien des chiens d'apparence similaire, probablement le même individu. Cependant, le clustering n'est pas parfait car on peut constater que des individus similaires qui correspondent probablement au même chien sont répartis entre les clusters B et C. Ces résultats permettent néanmoins de vérifier qualitativement le bon comportement de l'EC.

Pour une utilisation pratique, des informations supplémentaires telles que la position GPS des caméras ou l'heure d'acquisition pourraient être utilisées afin d'augmenter la précision des résultats et une vérification visuelle pourrait être nécessaire.

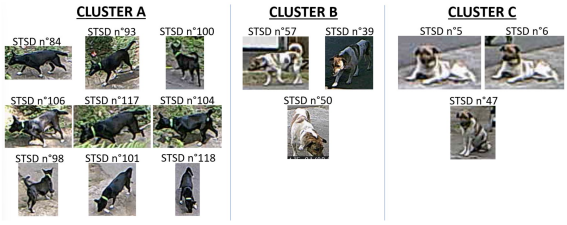

FIGURE 2 : Résultats du clustering agglomératif en utilisant la distance de Hausdorff. Trois clusters sur les 72 sont montrés.

4 Conclusions

Dans cet article, nous présentons une méthode pour ré-identifier les chiens lorsque le contenu des images est variable et de mauvaise qualité. Ce papier se démarque des autres publications par l'utilisation de jeux de données plus grands et plus diversifiés, mais également parce que nous utilisons l'apparence complète du chien et non seulement son visage. De plus, nous travaillons avec des Séquence d'Imagettes du

Même Chien (SIMC) et non pas des imagettes uniques. Dans ce contexte, nous présentons deux jeux de données dédiés à la ReID des chiens : les ensembles de données YT-BB-Dog et SEAdogSEA.

Notre proposition algorithmique consiste à construire un EC capable de discriminer entre différents individus similaires dans un scénario d'ensemble ouvert. Nous utilisons un fonction de coût par triplet et proposons une méthode pour définir automatiquement des exemples négatifs difficiles. Grâce à un apprentissage en deux étapes, l'EC se concentre sur les particularités de chaque chien et n'est pas influencé par le contenu de l'arrière-plan.

Sur l'ensemble de données YT-BB-Dog, nous atteignons une précision de top-1 de 73% sur un ensemble de test de 100 individus. Il s'agit d'un résultat très intéressant en tenant en compte de la difficulté de l'ensemble de données. Nous illustrons également la faisabilité de la chaîne de traitement (détection, suivi et ReID) sur l'ensemble de données SEAdogSEA, qui comprend des vidéos provenant de 28 pièges photographiques. La détection et le suivi sont effectués avec des algorithmes classiques (YoloV5 et StrongSORT) sans réapprentissage. La ReID est en suite effectuée à l'aide de l'EC appris sur YT-BB-Dog. Les résultats qualitatifs sont également prometteurs avec ces vidéos acquises dans un environnement non-contrôlé.

Pour améliorer ces résultats, les perspectives sont multiples, mais les plus intéressantes sont l'utilisation de la perte contrastive InfoNCE [12], l'utilisation de l'adaptation de domaine pour adapter l'EC à un ensemble de données non étiqueté dédié tel que SEAdogSEA, l'utilisation des méta-données, etc. Nous pourrions également parfaire la partie détection en apprenant sur YT-BB-Dog.

Références

[1] Dong-Hwa JANG, Kyeong-Seok KWON, Jung-Kon KIM, Ka-Young YANG et Jong-Bok KIM : Dog Identification Method Based on Muzzle Pattern Image. *Applied Sciences*, 10(24), 2020.

[2] Aditya KHOSLA, Nityananda JAYADEVAPRAKASH, Bangpeng YAO et Li FEI-FEI : Novel Dataset for Fine-Grained Image Categorization. *In Proceedings of the First Workshop on Fine-Grained Visual Categorization, FGVC'2011, IEEE Conference on Computer Vision and Pattern Recognition, CCVPR'2011*, Colorado Springs, CO, juin 2011.

[3] Tsung-Yi LIN, Michael MAIRE, Serge BELONGIE, James HAYS, Pietro PERONA, Deva RAMANAN, Piotr DOLLÁR et C. Lawrence" ZITNICK : Microsoft COCO : Common Objects in Context. *In Proceedings of the European Conference on Computer Vision, ECCV'2014*, pages 740–755, Zurich, Switzerland, septembre 2014. Springer International Publishing.

[4] Zhuang LIU, Hanzi MAO, Chao-Yuan WU, Christoph FEICHTENHOFER, Trevor DARRELL et Saining XIE : A ConvNet for the 2020s. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR'2022*, pages 11966–11976, New Orleans, LA, USA, juin 2022. IEEE.

[5] MMPOSE CONTRIBUTORS : Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

[6] Thierry Pinheiro MOREIRA, Mauricio Lisboa PEREZ, Rafael de OLIVEIRA WERNECK et Eduardo VALLE : Where is my puppy ? Retrieving lost dogs by facial features. *Multimedia Tools and Applications*, 76: 15325–15340, 2017.

[7] Guillaume MOUGEOT, Dewei LI et Shuai JIA : A Deep Learning Approach for Dog Face Verification and Recognition. *In Proceedings of the 16th Pacific Rim International Conference on Artificial Intelligence, PRICAI'2019*, pages 418–430, Cuvu, Yanuca Island, Fij, août 2019. Springer International Publishing.

[8] Prashanth C. RAVOOR et Sudarshan T.S.B. : Deep Learning Methods for Multi-Species Animal Re-identification and Tracking – a Survey. *Computer Science Review, Elsevier*, 38, 2020.

[9] Esteban REAL, Jonathon SHLENS, Stefano MAZZOCCHI, Xin PAN et Vincent VANHOUCKE : YouTube-BoundingBoxes : A Large High-Precision Human-Annotated Data Set for Object Detection in Video. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'2017*, pages 7464–7473, Honolulu, Hawaii, juillet 2017.

[10] Stefan SCHNEIDER, Graham W. TAYLOR, Stefan LINQUIST et Stefan C. KREMER : Similarity Learning Networks for Animal Individual Re-Identification – Beyond the Capabilities of a Human Observer. *In Conference : 2020 IEEE Winter Applications of Computer Vision Workshops, WACVW'2022*, Snowmass Village, CO, USA, mars 2020.

[11] Florian SCHROFF, Dmitry KALENICHENKO et James PHILBIN : FaceNet : A Unified Embedding for Face Recognition and Clustering. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,CVPR'2015*, pages 815–823, Boston, MA, USA, juin 2015.

[12] Aäron van den OORD, Yazhe LI et Oriol VINYALS : Representation Learning with Contrastive Predictive Coding, 2018.

[13] Di WU, Si-Jia ZHENG, Xiao-Ping ZHANG, Chang-An YUAN, Fei CHENG, Yang ZHAO, Yong-Jun LIN, Zhong-Qiu ZHAO, Yong-Li JIANG et De-Shuang HUANG : Deep Learning-based Methods for Person Re-Identification : A Comprehensive Review. *Neurocomputing*, 337:354–371, 2019.

[14] M. YE, J. SHEN, G. LIN, T. XIANG, L. SHAO et S. H. HOI : Deep Learning for Person Re-Identification : A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(06):2872–2893, juin 2022.